communications medicine

A Nature Portfolio journal

Article

https://doi.org/10.1038/s43856-025-00896-6

Utilizing Google Trends data to enhance forecasts and monitor long COVID prevalence

Amanda M. Y. Chu[1], Jenny T. Y. Tsang[2], Sophia S. C. Chan[3], Lupe S. H. Chan[4] & Mike K. P. So[4]

Abstract

Background Long COVID, the persistent illness following COVID-19 infection, has emerged as a major public health concern since the outbreak of the pandemic. Effective disease surveillance is crucial for policymaking and resource allocation.

Methods We investigated the potential of utilizing Google Trends data to enhance long COVID symptoms surveillance. Though Google Trends provides freely available search popularity data, limitations in data normalization and retrieval restrictions have hindered its predictive capabilities. In our study, we carefully selected 33 search terms and 20 related topics from the long COVID symptoms list provided by the Centers for Disease Control and Prevention and the database "scite", and calculated their merged search volumes from Google Trends data using our developed statistical method for analysis.

Results We identify four related topics (ageusia, anosmia, chest pain, and headaches) that consistently exhibit increased search popularity before that of "long COVID." Additionally, nine related topics (aching muscle pain, anxiety, chest pain, clouding of consciousness, dizziness, fatigue, myalgia, shortness of breath, and hypochondriasis) show increased search popularity following that of "long COVID." We demonstrate that the merged search volume (MSV), derived from the relative search volume data downloaded from Google, can be used to forecast the prevalence of long COVID in a prediction study, supporting the use of the methodology in risk management regarding the prevalence of long COVID.

Conclusions By utilizing a comprehensive list of search terms and sophisticated statistical analytics, our study contributes to exploring the potential of Google Trends data for forecasting and monitoring long COVID prevalence. These findings and methodologies can be used as prior knowledge to inform future infodemiological and epidemiological investigations.

Plain language summary

Long COVID is a persistent illness that follows COVID-19 infection. It has emerged as a significant public health concern since the outbreak of the pandemic. Effective disease surveillance is crucial for policy making and resource allocation. We investigate the potential of using the number of searches of long COVID symptoms in Google to enhance surveillance and improve the predictability of long COVID prevalence. We found searches for several specific symptoms increased both before and after searches for long COVID, demonstrating that numbers of searches can predict long COVID prevalence. Google search results could therefore be used to monitor disease prevalence.

The coronavirus disease (COVID-19) has become the most devastating pandemic in recent history[1]. Caused by the SARS-CoV-2 viral infection, the highly transmissible disease has resulted in substantial morbidity and mortality worldwide[2,3]. As of July 7, 2024, the number of confirmed COVID-19 cases has surpassed 776 million, claiming the lives of seven million individuals[4]. The emergence of long COVID, characterized by persistent illness following recovery from an acute COVID-19 infection, has further exacerbated the severity of the ongoing outbreak[5]. With the continued global spread of COVID-19, long COVID has increasingly been recognized as a public health concern. To address this issue, accurate real-time surveillance for long COVID is essential in facilitating policymaking, the timely implementation of health measures, and resource allocation to promote recovery.

During acute COVID-19 infection, symptoms commonly manifest within 4–5 days. These symptoms, which typically include a fever, sore throat, cough, muscle aches, loss of taste/smell, and diarrhea, can range from mild to severe[6]. Most patients appear to recover from the acute illness within four weeks[7]. However, a substantial proportion of acute COVID-19

[1]Department of Social Sciences and Policy Studies, The Education University of Hong Kong, Hong Kong, China. [2]School of Nursing, Tung Wah College, Hong Kong, China. [3]School of Public Health, The University of Hong Kong, Hong Kong, China. [4]Department of Information Systems, Business Statistics and Operations Management, The Hong Kong University of Science and Technology, Hong Kong, China. e-mail: immkpso@ust.hk

survivors develop persistent or new symptoms that can last for months or even years, in a condition now known as "long COVID" (also known as "post COVID-19 condition" or "post-Acute Sequelae of SARS-CoV-2 infection (PASC)").

Long COVID is considered to be the sequelae of COVID-19, although its precise definition is still under discussion[8]. Research on long COVID was initially sparked by Paul Garner, an infectious disease professional, who highlighted the lingering symptoms following acute COVID-19 infection[9]. It is now known that long COVID is a multisystem disorder with a spectrum of mild to severe illness[10,11]. Symptoms of long COVID vary widely from person to person and can be confusing for patients and healthcare professionals. Individuals with long COVID usually experience one or more persistent symptoms. Common physical symptoms include fatigue, shortness of breath, dyspnea, palpitations, muscle pain, chest pain, chest tightness, headaches, dizziness, joint pain, and loss of taste and smell[7]. In addition to physical ailments, psychosocial health issues are commonly observed, such as brain fog, difficulty concentrating, confusion, cognitive impairment, poor memory, insomnia, and even mental illnesses such as anxiety, depression, and post-traumatic stress disorder[7,11–15]. It is believed that these psychosocial and mental symptoms are likely related to immune system alteration after acute COVID-19, inducing systemic inflammation and changing the cortical thickness of the brain[16,17]. Long COVID patients commonly experience exercise intolerance, impaired daily function, and lowered quality of life. Long COVID can also increase the risk of organ damage[18]. In severe cases, it often disrupts individuals' ability to work and leads to disability[19].

The estimated prevalence of long COVID varies widely across studies[20]. According to the WHO, 10% to 20% of individuals who have recovered from acute COVID-19 develop long COVID[21]. However, the UK Office for National Statistics reports a lower prevalence of 3% for long COVID in the UK[22]. A study conducted in Italy found that 87% of hospitalized COVID-19 patients exhibited at least one persistent symptom even 60 days after recovering from an acute infection[23]. Furthermore, in a meta-analysis encompassing 194 studies worldwide, involving 735,006 participants with an average follow-up of 126 days, it was estimated that 45% of COVID-19 survivors experienced at least one unresolved symptom[24]. Yao et al.[25] reviewed the findings reported by O'Mahoney et al.[24] and found considerable discrepancies between the data in the systematic review by O'Mahoney et al.[24] and the original studies; they argue that the rate of long COVID is ~20%. Despite the considerable variation in estimating long COVID cases, the available information highlights a substantial portion of individuals suffering from long COVID. Therefore, it is crucial to explore and implement effective disease monitoring and management strategies.

Previous studies that have aimed to track long COVID rely heavily on medical records containing self-reported symptoms from patients[23,26,27]. However, these records are limited to individuals who sought medical attention, posing a challenge in accurately determining the true prevalence of long COVID. It is important to note that the prevention and control of COVID-19 has moved to a phase whereby patients do not need to be hospitalized (for isolation) unless they have severe symptoms. Hence, patients with mild symptoms may not seek medical care and may instead adopt a self-management approach, further complicating the estimation of the prevalence of long COVID. Additionally, the diagnosis of long COVID presents difficulties. Patients with long COVID exhibit a wide diversity of symptoms, ranging in severity from mild to severe. Currently, there are no specific biomarkers or diagnostic tests that can reliably confirm a long COVID diagnosis. Furthermore, the lack of a consistent definition of long COVID adds to this challenge. The WHO defines long COVID as a condition that occurs in individuals with a history of probable or confirmed SARS-CoV-2 infection, typically manifesting three months after the onset of COVID-19 symptoms and lasting for at least two months, with no alternative explanation for the symptoms[21]. In contrast, the Centers for Disease Control and Prevention[7] defines long COVID as an umbrella term for a range of health consequences that persist for four or more weeks after COVID-19 infection. More definitions of long COVID by different public health bodies have been summarized in a comprehensive interdisciplinary review by Greenhalgh et al.[18]. These challenges contribute to the considerable variation in long COVID prevalence rates observed across studies, emphasizing the way in which relying solely on limited medical records might not provide accurate information in terms of monitoring long COVID. Therefore, it is crucial to explore complementary approaches to enhancing long COVID surveillance.

Internet use has become a prevailing human behavior, as is evident from the rapidly growing number of internet and social media users worldwide[28]. This widespread adoption has given rise to the development of an innovative approach in public health surveillance, referred to as infodemiological studies, which leverage information obtained from the internet to track various health issues[29]. The internet has emerged as an important source of healthcare information in recent years, with individuals frequently turning to online searches to gather information about their symptoms and illnesses before seeking medical care[30]. Consequently, data from internet searching activities have become a valuable resource in monitoring disease prevalence at the community level.

Google Trends[31] is a website that provides insights into everyday searches made on Google, which is the most popular search engine in the world, capturing nearly 92% of the global search market[32]. Google Trends aggregates data from multiple sources within the Google search system, including web searches, image searches, news searches, Google shopping, and YouTube searches. This comprehensive data collection allows for the analysis of search query popularity and provides real-time search patterns of internet users worldwide[31]. Over the past decade, there has been increasing interest in the use of Google Trends for public health and epidemiological research[33]. Previous studies have demonstrated the reliability of analyzing search volumes of internet queries in order to forecast and monitor various pathogenic infections, including Ebola[34], Middle East respiratory syndrome[35], and Dengue[36]. During the COVID-19 outbreak, Google Trends proved to be a valuable tool for monitoring the population's health concerns and forecasting COVID-19 prevalence[37], indicating the potential use of Google Trends data for long COVID surveillance.

Although Google Trends provides free access to data, there are certain limitations associated with retrieving information. Instead of providing the exact number of searches for a specific search term, Google Trends presents "interest over time", using relative search volume (RSV) time series data, which are normalized on a scale ranging from zero to 100. While this normalization facilitates easy comparisons, it may restrict researchers' ability to accurately track actual search behaviors. To overcome this limitation, in our previous research, we devised a statistical methodology that enhances the extraction of data from Google Trends[38]. This approach improves the resolution of the RSV, enabling a more accurate reflection of search trends as they manifest in the population.

This research aims to analyze the online search patterns of the public in regard to information about long COVID using a sophisticated statistical analysis of Google Trends data. The goal is to leverage this data to enhance long COVID surveillance. We identified the commonly used search terms (keywords) and related topics for online searches related to long COVID. The search popularity data of these search terms and related topics was retrieved from Google Trends. These original Google Trends data were then analyzed using statistical methods developed in our previous research, in order to better understand actual long COVID-related search behaviors and explore the potential use of an infodemiological approach to complementing long COVID surveillance in the community. We also conducted a prediction study to explore the usefulness of forecasting long COVID prevalence using the search volumes of the symptoms' related topics. We also studied the evolution of long COVID symptoms using the estimated parameters in the prediction model, as the symptoms may have changed over time, as evidenced by a study conducted by Wynberg et al.[39] that analyzed the evolution of symptoms listed by The International Severe Acute Respiratory and Emerging Infection Consortium[40] by following a cohort of patients for 12 months.

In this study, we address the limitations of resolution and scope in the RSV data by calculating the merged search volume (MSV) for our analysis.

We identify topics that consistently show increased search popularity before and after the topic of "long COVID" by comparing their time series plots and polar projection plots. In a prediction study, we demonstrate that these MSVs can be used to forecast the prevalence of long COVID. We support the use of this methodology in risk management related to the prevalence of long COVID based on our findings.

Methods

Search terms and topic discovery

Accurately identifying the search terms related to long COVID is crucial in understanding the search activity associated with this condition. Research suggests that individuals often use symptoms as search terms when seeking information about diseases online, and previous studies have shown an increase in the search popularity of terms representing long COVID symptoms as awareness of the condition grows[41]. Therefore, we focused on identifying search terms based on reported long COVID symptoms. To compile a list of search terms related to long COVID that internet users may enter into the search bar, we referred to the long COVID symptoms list provided by the Centers for Disease Control and Prevention[7]. We also determined possible search terms by searching the database "scite"[42], which is an AI-powered platform for discovering and evaluating scientific articles. Twenty-five peer-reviewed publications were found by asking the database for "literature related to 'long COVID'". Through this process, we established a list of 33 long COVID-related search terms. The literature referenced and the extracted search terms are listed in Supplementary Data 1. Subsequently, these search terms were entered into Google Trends to analyze their respective search popularity.

Upon entering the search terms into Google Trends, there were two options in terms of specifying the search term: "related queries" and "related topics". Related queries are words or phrases related to a search term, whereas related topics are pre-categorized sets of search terms recommended by Google Trends based on the entered search terms. Figure 1

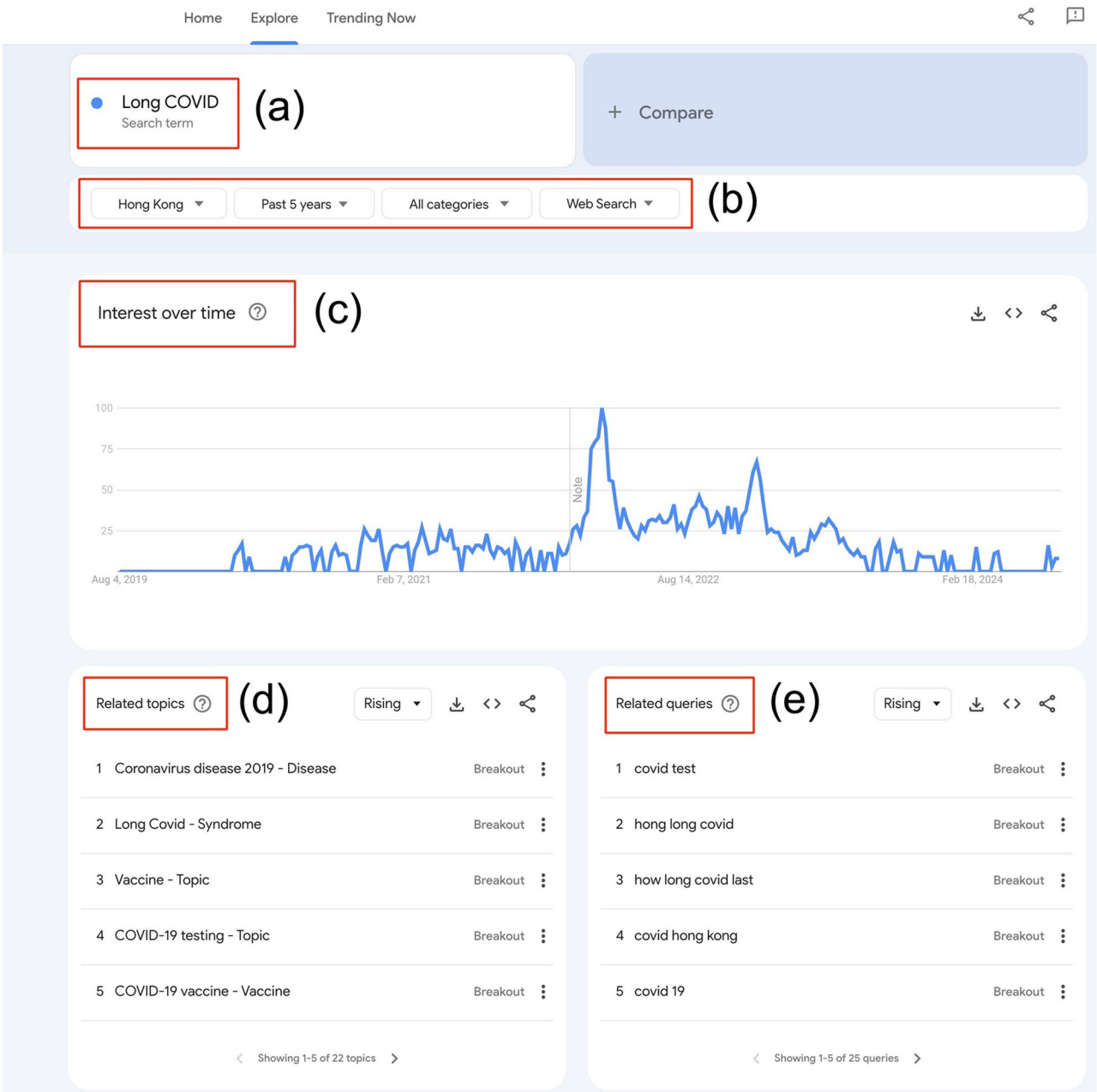

Fig. 1 | A screenshot of the Google Trends webpage (Google, 2024). The parts marked with **a**, **b**, **c**, **d**, and **e** are, respectively, the search term, the four additional parameters used to specify our analysis preferences, interest over time, related topics, and related queries.

contains an example of a search on Google Trends using the search word "COVID-19" (marked (a) in Fig. 1). The related topics (marked (d) in Fig. 1) contain broad categories related to COVID-19; in contrast, the related queries (marked (e) in Fig. 1) contain only text or phrases associated with COVID-19. While previous studies have often used "related queries" to broaden their search coverage, we chose to utilize "related topics" in our analysis. Each topic is assigned a Google Trends internal tag and represents a specific target, such as a person, object, company, symptom, or disease. The webpage for a specific topic can be reached by searching the corresponding internal tag on Google Trends. Selecting related topics is more suitable for our study due to their unambiguous characteristics and the inclusion of translations of the same concept in other languages. For example, when studying the search volume for Apple Inc., the technology company, entering the search term "Apple" alone would include the search popularity of everything related to the term "apple," such as the fruit. To specifically focus on the technology company, researchers can choose the appropriate related topics, such as "Apple (Technology Company)", which filters out irrelevant information. In our study, we selected 20 related topics to specify the 33 search terms to be compared against the topic "long COVID". All 20 related topics and their Google Trends internal tags are listed in Supplementary Data 1.

After entering the search terms and related topics into Google Trends, we input four additional parameters to specify our analysis preferences (marked (b) in Fig. 1). As our focus was on a global analysis, we selected "worldwide" as the location of interest. We specified the search period from the outbreak of the COVID-19 pandemic (January 1, 2020 to December 31, 2023). To encompass diverse categories of interest, such as health and people and society, we chose "All categories." Additionally, we selected "Web search" as the type of online search of interest.

In conducting this research, we affirm that our study did not involve human participants or animals; therefore, IRB approval was not required.

Data from Google Trends

After entering the search terms, selecting the appropriate related topics, and specifying the four parameters using geographical location, search duration, search category, and search engine, Google Trends initiated the analysis and provided us with time series data known as "interest over time" (marked (c) in Fig. 1). We downloaded this data as the relative search volume (RSV) for further analysis. RSVs can be downloaded for free from Google Trends. The values of interest over time, ranging from zero to 100, represent the search interest relative to the highest point for the defined filtering region and period. It is important to note that, according to Google's definition, a value of 50 indicates that the search term is half as popular, while a value of zero signifies a low search volume (i.e., less than 1% of the searches attributed to the most searched term in the search list[31]). Thus, the interest over time can be considered as a transformation of the actual search volume. Since the data are heavily normalized, this introduces some degree of noise.

Empirical evaluations, as well as results from other studies, have highlighted limitations of Google Trends data. These limitations include resolution limitations (e.g., data available only in weekly or even monthly form, instead of daily form, when requesting prolonged data, such as in the work of Olson et al.[43], and Borup and Schütte[44]) and scope limitation (e.g., high-resolution data only available when requesting data for a short duration, as in the work of Li et al.[45] and Mavragani and Ochoa[46]). To address these limitations, we employed our previously developed statistical method[38] to calculate MSV based on the RSV data obtained from Google Trends. This approach helped us mitigate the hindrances posed by the resolution and scope limitations.

Calculating the merged search volume (MSV)

To calculate the MSV using the RSV data, we downloaded the RSV data from Google Trends, which are the "interest over time" data for each related topics from January 1, 2020 to December 31, 2023 ($T = 1461$ days). Following the approach used in Chu et al.[38], let $z_{i,s}^{(t)}$ be the RSV of the i th related topics on day s, normalized using the correction factor $C_{i,t}$ (to be introduced later), for $i = 1, \ldots, 21$, where $i = 1$ corresponds to the related topics "long COVID" and $i = 2, \ldots, 21$ correspond to the related topics of the symptoms, and s and t can take values from $1, 2, \ldots, T$. Using a window size of n days, let

$$\mathbf{z}_{i,t} = \left(z_{i,t-n+1}^{(t)}, z_{i,t-n+2}^{(t)}, \ldots, z_{i,t-1}^{(t)}, z_{i,t}^{(t)}\right)^T \tag{1}$$

be a vector containing the most recent n days of RSV for the i th search term, normalized using the common correction factor $C_{i,t}$. We further let $Z_{i,t}$ be the actual search volume of the i th search term on day t, and

$$\mathbf{Z}_{i,t} = \left(Z_{i,t-n+1}, Z_{i,t-n+2}, \ldots, Z_{i,t-1}, Z_{i,t}\right)^T. \tag{2}$$

By assuming that Google Trends applied a common correction factor $C_{i,t}$ for normalizing the i th search term, $\mathbf{z}_{i,t}$ and $\mathbf{Z}_{i,t}$ are related via

$$\mathbf{z}_{i,t} = \left(\frac{Z_{i,t-n+1}}{C_{i,t}}, \frac{Z_{i,t-n+2}}{C_{i,t}}, \ldots, \frac{Z_{i,t-1}}{C_{i,t}}, \frac{Z_{i,t}}{C_{i,t}}\right)^T = \frac{1}{C_{i,t}}\mathbf{Z}_{i,t}, \tag{3}$$

which implies

$$\mathbf{Z}_{i,t} = C_{i,t}\mathbf{z}_{i,t}. \tag{4}$$

Consider $\mathbf{Z}_{i,t+1}$, the vector containing actual search volumes of the most recent n days, as of day $t+1$. The actual search volume on day t should be the same in $\mathbf{Z}_{i,t}$ and $\mathbf{Z}_{i,t+1}$, and thus we have

$$C_{i,t} z_{i,t}^{(t)} = C_{i,t+1} z_{i,t}^{(t+1)} \tag{5}$$

and thus

$$\frac{z_{i,t}^{(t)}}{z_{i,t}^{(t+1)}} = \frac{C_{i,t+1}}{C_{i,t}} := \mathbb{C}_{i,t+1|t}, \tag{6}$$

where $\mathbb{C}_{i,t+1|t}$ is defined as the adjustment factor. An estimator of $\mathbb{C}_{i,t+1|t}$ is obtained through the moment estimation, using the data on overlapping $(n-1)$ days in $\mathbf{z}_{i,t}$ and $\mathbf{z}_{i,t+1}$:

$$\hat{\mathbb{C}}_{i,t+1|t} = \frac{1}{n-1}\sum_{\tau=0}^{n-2} \frac{z_{i,t-\tau}^{(t)}}{z_{i,t-\tau}^{(t+1)}}, \tag{7}$$

where $z_{i,t-\tau}^{(t)}/z_{i,t-\tau}^{(t+1)}$ is set to 1 if the denominator is zero. Then, we adjust the RSV on day $t+1$ by

$$\hat{z}_{i,t+1}^{(t)} = \hat{\mathbb{C}}_{i,t+1|t} \cdot z_{i,t+1}^{(t+1)}. \tag{8}$$

In general, we adjust the RSV on day $t+k$ by

$$\hat{z}_{i,t+k}^{(t)} = \hat{\mathbb{C}}_{i,t+k|t} \cdot z_{i,t+k}^{(t+k)}, \tag{9}$$

where

$$\hat{\mathbb{C}}_{i,t+k|t} = \frac{1}{n-1}\sum_{\tau=0}^{n-2} \frac{\hat{z}_{i,t+k-1-\tau}^{(t)}}{z_{i,t+k-1-\tau}^{(t+k)}}, \tag{10}$$

for integer $k \geq 1$, where $\hat{z}_{i,t+k-1-\tau}^{(t)}/z_{i,t+k-1-\tau}^{(t+k)}$ is set to 1 if the denominator is zero, and $\hat{z}_{i,\tau}^{(t)} = z_{i,\tau}^{(t)}$ for $\tau = 1, \ldots, t$.

We choose day n as the baseline day. By iteratively applying the estimation starting at $t = n$, we can obtain a merged time series for the i th

search term

$$\hat{\mathbf{z}}_{i\cdot} = \left(\hat{z}_{i,1}^{(n)}, \hat{z}_{i,2}^{(n)}, \ldots, \hat{z}_{i,n}^{(n)}, \hat{z}_{i,n+1}^{(n)}, \hat{z}_{i,n+2}^{(n)}, \ldots, \hat{z}_{i,T}^{(n)}\right)^T, \quad (11)$$

where $\hat{z}_{i,\tau}^{(n)} = z_{i,\tau}^{(n)}$ for $\tau = 1, \ldots, n$. We pick $n = 29$ in this paper. We write $\hat{z}_{i,t}^{(n)} = \hat{z}_{i,t}$ to further simplify the formula.

Assessing the reliability of the data

There are several known problems with Google Trends, including RSV stability associated with repeated requests and Google improvements[47], and potential confounding factors[48].

RSV stability associated with repeated requests

According to Rovetta[47], repeating requests on Google Trends at different times or from different addresses could return different time series. To assess the RSV stability associated with repeated requests, similar to Rovetta[47], for each related topic, we repeatedly downloaded ten RSVs over the interval from January 1, 2020, to December 31, 2023, each obtained every 36 h. We calculated the intra-correlation residuals (IC) from the ten RSVs, (i.e., the correlation coefficients between all possible pairs of the ten RSV time series). An IC between any two time series smaller than 1 indicates the deviation between the two time series. Detailed results can be found in Supplementary Note 1.

RSV stability associated with Google improvements

Google Trends made three improvements in January 2011, 2016, and 2022, which changed the geographical allocation and data collection systems, making the RSVs before and after the improvements incomparable[47]. Following Rovetta[47], we calculated the percentage coefficients of variation in the RSVs before and after January 2022 to explore the impact of the improvement made on 1 January 2022. The detailed results are contained in Supplementary Note 1.

Confounding factors

Following Sato et al.[48], we conducted confounding factor tests between the MSVs of each symptom and "long COVID". The potential confounding factors include mass media influence, pharmacological interventions (for example, vaccination policies), non-pharmacological interventions (for example, closure policies), and local communication policies. We downloaded financial news from Reuters and counted the daily numbers of news stories containing the search terms for each related topic. The financial news data is licensed and requires a subscription for download. The numbers were used as proxies for testing if mass media influence is a significant confounding factor. We also used data from the Oxford COVID−19 Government Response Tracker (OxCGRT)[49], which contains information on which and when pandemic response measures were enacted by governments over the world, as proxies of pharmacological and non-pharmacological interventions, and local communication policies. The OxCGRT dataset is available for free download from the website of the corresponding author. The methods used to test confounding factors are detailed in Supplementary Note 2. Additionally, the results of the analyses regarding the confounding effects of local communication policies, pharmacological interventions, and non-pharmacological interventions are presented in Supplementary Note 3, while the findings related to media coverage can be found in Supplementary Note 4.

Visualization of the MSVs in polar projection plots

To better visualize and evaluate the trends of the MSVs, we constructed polar projection plots by projecting each MSV inside the unit circle and observing the projected shapes and the corresponding centroids. The methodology of polar projection plots is discussed in Supplementary Note 5.

Methodology for predicting the prevalence of long COVID and analysis of the evolution of symptoms

In this section, we illustrate that we can use Google Trends data to obtain accurate forecasts for the prevalence of long COVID. Following the approach used in Kumar and Susan[50], which modeled and evaluated the prediction of the evolution of the COVID−19 outbreak using an autoregressive integrated moving average (ARIMA) time series forecasting model, we evaluated the predictive performance, based on the root mean square errors and the mean absolute percentage errors, of two candidate models: (1) the ARIMA(p, d, q) model[51], which predicts the prevalence of long COVID using only the past MSVs of "long COVID" and (2) the vector autoregressive with a LASSO penalty[52], including the past MSVs of the symptoms as the predictors.

In addition, we analyzed the estimated parameters in the lasso-VAR over time. A non-zero parameter for a related topic in the lasso-VAR indicates that the MSV of the related topic is useful in explaining the variation of the MSV of "long COVID". Therefore, we can consider the related topic as a potential symptom of long COVID. Details on the methods used and the results can be found in Supplementary Note 6. Figure 2 presents an overall methodology flowchart for this paper.

Statistics and reproducibility

Python (Python 3) and R (version 4.2.2) were employed for all analyses. Python was used for the analyses contained in the main text while R was used for the additional analyses contained in the Supplementary Information.

First, we collected the RSVs through web scraping. We employed a moving-window approach, beginning with the download of RSVs from January 1, 2020, to January 29, 2020 ($n = 29$ days). We then proceeded to download RSVs from January 2, 2020, to January 30, 2020. This process continued iteratively until we reached the end date of December 31, 2023. Given the substantial volume of the RSVs, the downloaded data was stored in MongoDB (Banker et al., 2016). We then calculated the MSVs from the downloaded RSVs. For analytical purposes, we produced time series plots and polar projection plots of the MSVs.

The additional analyses in the Supplementary Information were conducted using R. We first divided the interval from January 1, 2020 to December 31, 2023 into 17 windows, each containing 3 months of RSV data. We assessed the stability of the Google Trends by repeatedly downloading the RSVs 10 times in each window. Then, we computed the intra-correlation residuals for each related topic. These intra-correlations were visualized using boxplots. To assess the effects of the Google improvements implemented on January 1, 2022, we calculated the coefficients of variations before and after the Google improvements for comparison.

Granger causality tests were conducted to test for confounding factors between the MSVs of long COVID and the Symptoms. We use the financial news and the Oxford COVID−19 Government Response Tracker data[49] as proxies for respectively the mass media influence and governmental interventions. Before the Granger causality test, the stationarity of the time series of these proxies and MSVs was tested using the augmented Dicky-Fuller test. If a time series was non-stationary, we stabilized it by differencing it.

To show that we can use Google Trends data to forecast the prevalence of the Long COVID, we fitted the MSVs of long COVID using the MSVs of the symptoms as predictors in vector autoregressive regression models with LASSO penalty. The models were fitted using a moving-window approach with a window size of 180 days. We evaluated the h-day ahead predictive performance ($h = 1, \ldots, 21$ days) using root mean square errors and mean absolute percentage errors. We also visualized the non-zero parameters in a heatmap to assess the dynamics of potential symptoms of long COVID.

Results

Reliability of the data

The instability of the RSV and the existence of confounding factors can make data unreliable. The stability of the RSV is crucial for ensuring that the data accurately reflect consistent trends over time. Increased variability may lead to inaccurate results, thus making it difficult to draw meaningful conclusions or to compare trends across different periods. On the other hand, the existence of confounding factors may create spurious correlations between the search volumes of the symptoms and the prevalence of long

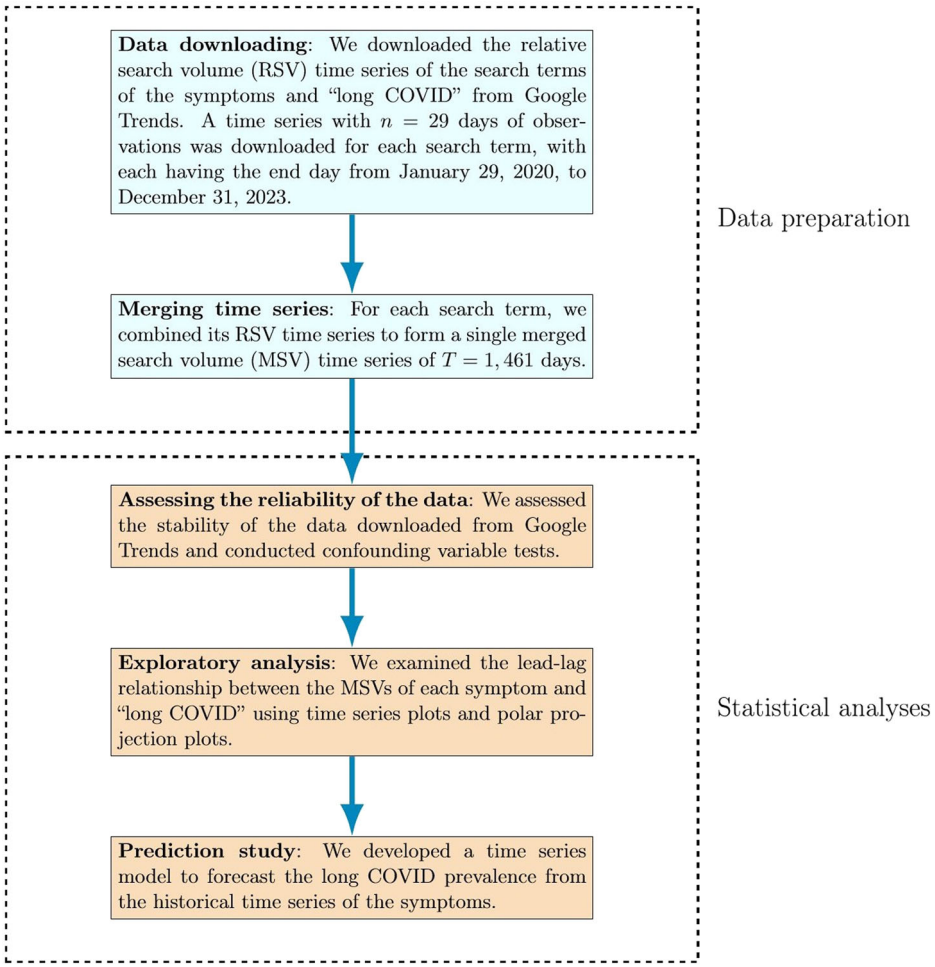

Fig. 2 | An overall methodology flowchart for this paper. The flowchart illustrates the progression of our research. We start by preparing the RSV data, which is then merged with the MSV data. Next, we conduct statistical analyses. In this phase, we first assess the reliability of the data, followed by an exploratory analysis, and finally, we proceed to a prediction study.

COVID, and then may result in false positive findings. Therefore, it is important to assess the stability of RSV and account for confounding factors to enhance the reliability of the data.

We conducted analyses of RSV stability associated with repeated requests and Google improvements[47] and tested for potential confounding factors[48]. We provide the analysis results in this section. Detailed methodologies and results for the analyses below can be found in Supplementary Notes 1 to 4.

Regarding the RSV stability associated with repeated requests, we provide a detailed analysis in Supplementary Note 1. The analysis indicates that three of the related topics, "aching muscle pain", "clouding of consciousness", and "hypochondriasis" have low ICs, indicating the instability in repeat requests of these related topics. We note that we need to be careful in interpreting the results based on these three related topics. The results suggest that most RSVs are stable and thus reliable for the analysis.

The analysis results of the RSV stability associated with Google improvements is also provided in Supplementary Note 1. The results show that, for most of the related topics, the CVs before and after the improvements are quite different and have non-overlapping confidence intervals. This indicates that the improvements can induce extra uncertainty in the long-term analysis of related topics. Although we do not know the severity of the impact this has had on the data, we further assess the usefulness of the RSVs in explaining and predicting the long COVID prevalence in "Methodology for predicting the prevalence of long COVID and analysis of the evolution of symptoms" section for the methodology and "Results for predicting the prevalence of long COVID and analysis of the evolution of symptoms" section for the results in this paper.

Furthermore, we present the confounding factor test results in Supplementary Notes 3 and 4. There are seven related topics with confounding factors that significantly affect the Granger-causality test results between the related topics and "long COVID". The related topics "clouding of consciousness", "hypochondriasis", and "lightheadedness" have confounding factors from both governmental policies and interventions, and media coverage. "Ageusia" and "chest pain" have confounding factors from governmental policies and interventions. "Headache" and "sleep disorder" have confounding factors from media coverage. Extra caution has been taken when analyzing the relationship between these related topics and "long COVID".

As the confounding factors are identifiable, the issue can be addressed by analyzing the filtered MSVs (i.e., the MSVs that control for the effects of the confounding factors), and thus are manageable. To maintain the focus of this paper and streamline the analysis, we do not consider the use of filtered MSVs in this paper. The analysis in this section instead has identified the confounding factors and can guide future research toward a more comprehensive analysis that takes the confounding factors into account.

Close-up inspection of the search popularity of "long COVID"

To initiate our analysis, we conducted a thorough examination of the merged search volumes for the focal topic of our study, which is "long COVID." The MSV time series for "long COVID" is depicted in Fig. 3. During the observation period, we observed many different patterns in the related topics "long COVID." However, we noticed two substantial peaks in particular that stand out, namely: (1) a gradually converging trend from November 20, 2021 to February 6, 2022, which closely resembles the outbreak of the COVID-19 Omicron variant[53] globally and locally; and (2) a rise and a fall from June 20, 2022 to October 1, 2022, which correlate with the popularization of the Omicron variant and the BA.5 and BA.6 subvariants of COVID−19[54,55], as well as increases in confirmed COVID-19 cases across many European countries[56]. Because of the substantial volume changes and the correlations with real-world events, we focus our study on these two periods.

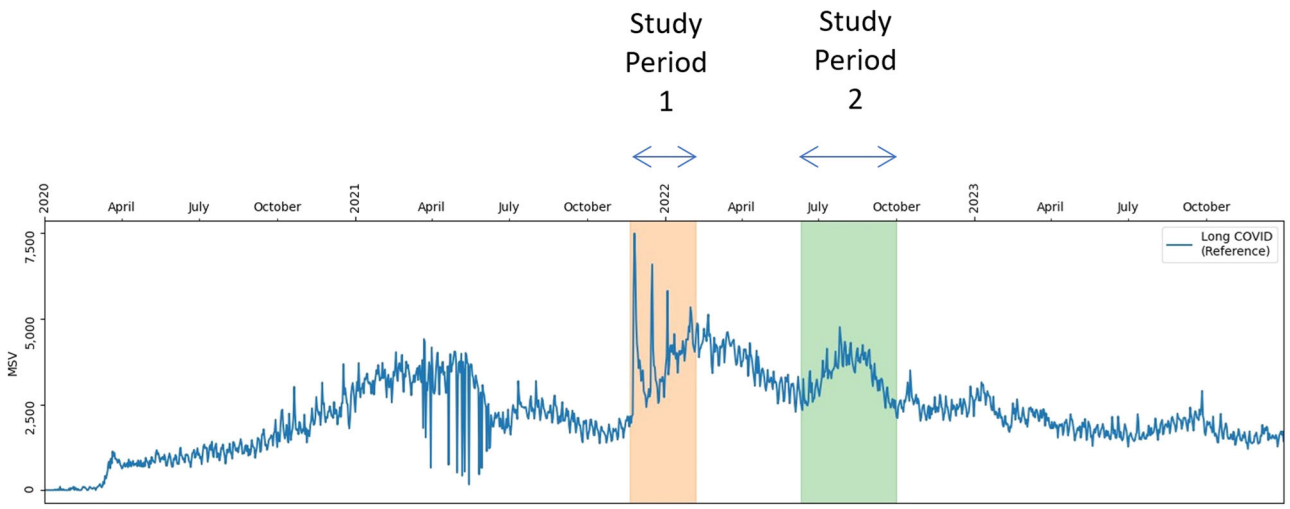

Fig. 3 | The merged search volume (MSV) of "long COVID" during the COVID-19 pandemic. The MSV of the related topics "long COVID" was plotted against time. The sharp and sudden spike in MSV occurred in late November 2021 and gradually converged until early February 2022. This section of MSV is highlighted in orange and is identified as the first study period (marked "Study Period 1"). Additionally, a parabolic peak in the "long COVID" MSV was also observed from late June 2022 to early October 2022, and is highlighted in green. We identified it as the second study period (marked "Study Period 2").

Search popularity of "long COVID" and the other related topics: the first study period

The MSV time series for the selected related topics were compared with that of "long COVID" during the first study period. related topics that exhibited a sharp MSV peak before, during, or after the peak observed in "long COVID" were classified as Class 1. After the comparison, 16 related topics were identified as belonging to Class 1. The time series plots displaying the MSV time series for each of these related topics and that of "long COVID" are presented in Fig. 4. Class 1 related topics demonstrated a similar MSV time series pattern to that of the related topic "long COVID" during the first study period by sharing at least one of the sharp peaks in close proximity to the three characteristic peaks of the related topic "long COVID"; hence, the related topics indicate their potential to be used in infodemiological studies of long COVID. Note that the related topics "aching muscle pain", "clouding of consciousness", and "hypochondriasis" were shown to be unstable on Google Trends, and may be subject to larger uncertainties. Caution is warranted in interpreting these related topics.

Search popularity of "long COVID" and the other related topics: The second study periodComparisons were conducted against the "long COVID" MSV to identify related topics in any comparable time series patterns during the second study period. However, we noticed that the raw MSVs of all the related topics, including "long COVID", fluctuate quite rapidly in the second study period. For the sake of clarity, we have included the one-week moving average (MA) of the MSVs of the related topics and "long COVID" in Figs. 5 and 6. The rolling window width was determined experimentally. After three moving average smoothings were applied to the data, with durations of one week, two weeks, and three weeks, we observed no drastic difference between the smoothed data. Hence, we pick the smallest MA duration of one week.

The results indicate that 17 related topics demonstrated a similar MSV pattern to that of "long COVID" during the second study period. Notably, all of these related topics also displayed a similar MSV time series pattern to that of "long COVID" during the first study period. These 17 related topics were classified as Class 2. Among these related topics, some have their MSV peak leading (preceding) the MSV peak of "long COVID," while others have their MSV peak lagging behind (following) it. The eight related topics with their MSV peaks leading that of "long COVID" are classified as Class 2a, which includes ageusia, anosmia, chest tightness, headaches, hip pain, mental health, insomnia, and major depressive disorder. The time series plots for these related topics are presented in Fig. 5. On the other hand, the nine related topics in Fig. 6 with an MSV peak lagging behind that of "long COVID" are classified as Class 2b, which includes aching muscle pain, anxiety, chest pain, clouding of consciousness, dizziness, fatigue, myalgia, shortness of breath, and hypochondriasis.

We noticed that Class 2a contains two related topics of the most prominent COVID-19 symptoms, ageusia and anosmia, as well as containing psychosocial wellness-oriented related topics—namely, insomnia, major depressive disorder, and mental health. Nonetheless, we also noticed several related topics for non-characteristic symptoms, such as chest tightness, headaches, and hip pain. On the other hand, related topics in Class 2b are dominated by search terms related to COVID-19 and long COVID symptoms, except for the related topic "anxiety", which is not specific to COVID-19/long COVID.

Based on the observations between Class 2a and Class 2b, there may be a change in the public's search priorities and thus in interest. We believe the public showed collective interest in symptoms related to COVID-19/long COVID, such as related topics in Class 2a, for self-diagnosis purposes. However, after noticing the spread of the Omicron sub-variants, the public also regained their interest in COVID-19.

We emphasize that the symptom-related topics in Class 2a do not show a drastic drop in search volume after the spike in "long COVID", which further supports our interest-focusing theory.

Polar projection and centroids

To further examine the MSV pattern within the entire COVID-19 pandemic period, we projected the four-year MSV time series from January 2020 to December 2023 onto a polar plane within a unit circle. This method, unlike the one used in "Search popularity of "long COVID" and the other related topics" section, which cross-compares the MSV pattern against the movement of "long COVID", enables us to compare the MSV fluctuations among related topics. The polar projection of "long COVID" and its projection centroid is shown in Fig. 7. Similar to the MSV time series plot in Fig. 3, the projections of "long COVID" show increases during the time of the first study period and the second study period. Additionally, the centroid of the projection was deduced using the method described in "Visualization of the MSVs in polar projection plots" section to indicate the overall public attention paid to "long COVID" over the four-year period. The polar projection and centroid plots for the other 19 related topics are also shown in Fig. 8, following the same method.

By comparing the shapes and positions of the polar projections of each related topic with that of "long COVID", we observe that nine related topics had their first projection increase around April 2020, which is much earlier

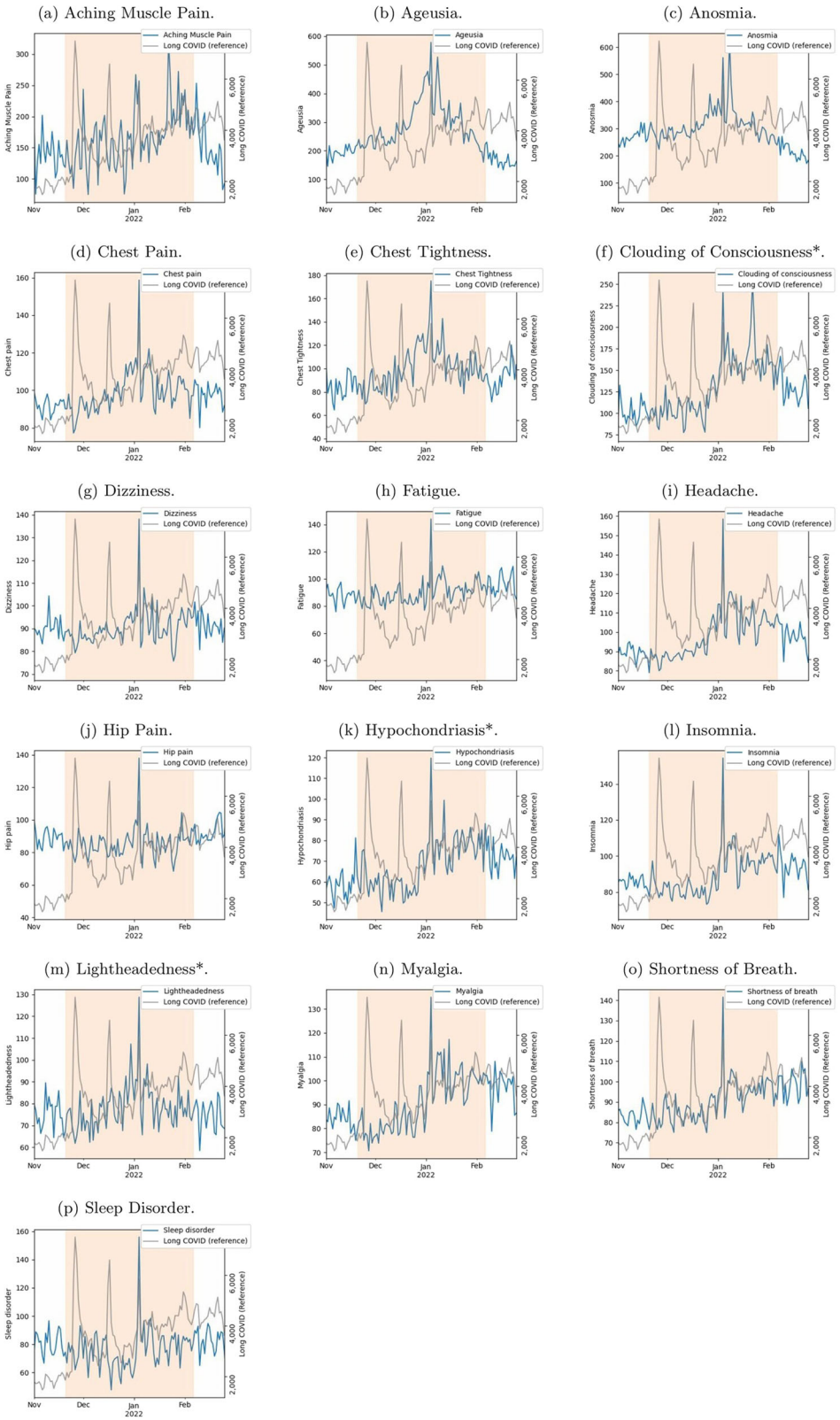

Fig. 4 | The merged search volume (MSV) time series plots of "long COVID" and the related topics classified as Class 1, from late November 2021 to early February 2022, with the first study period highlighted in orange. Subfigures **a**–**p** present time series plots of the MSVs for the related topics indicated in the legends and that for "long COVID" for comparison. The blue and gray lines respectively represent the MSV time series for the symptoms and "long COVID". The symptoms marked with * indicate that the RSVs were shown to be unstable on Google Trends, and may be subject to larger uncertainties.

than that of "long COVID." The projections and centroids of these related topics are shown in Fig. 8. These nine related topics were classified as Class 3a. Among them, headaches, ageusia, anosmia, and chest tightness were included in both Class 2a and 3a, which further indicates that the search popularity of these related topics increased ahead of "long COVID".

On the other hand, the polar projection plots in Fig. 9 show that the three related topics had their first projection increase at the same time or after that of "long COVID". These three related topics were classified as Class 3b. Among them, anxiety was included in both Class 2b and 3b, indicating that the search popularity for this related topics increased after that of "long COVID".

Finally, we summarized the MSVs of "long COVID" and the other 20 related topics by plotting their centroids on the same graph, as shown in Fig. 10. Note that the centroid for "long COVID" is near January 2022. This

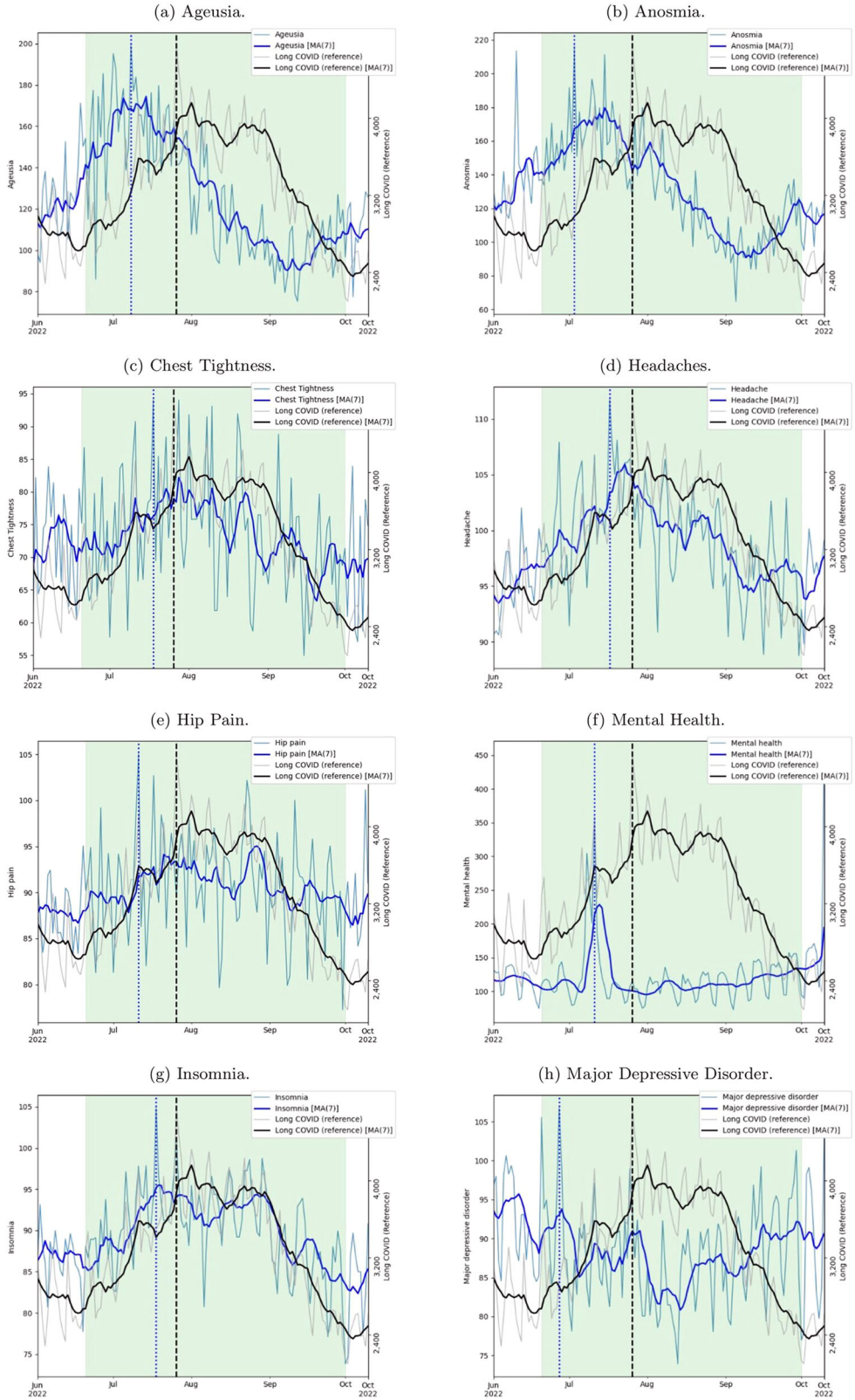

Fig. 5 | The merged search volume (MSV) time series plots of "long COVID" and the related topics classified as Class 2a. Subfigures **a**–**h** present time series plots of the MSVs for the related topics indicated in the legends and that for "long COVID" for comparison. The light blue and gray lines respectively represent the MSV time series for the symptoms and "long COVID", and the blue and black lines respectively represent the seven-day moving average time series of the MSVs of the symptoms and "long COVID", to better enable visualization. The vertical blue dotted lines show the period maxima of the MSVs of the corresponding related topics, while the black vertical dashed lines show the period maxima of the "long COVID" MSV.

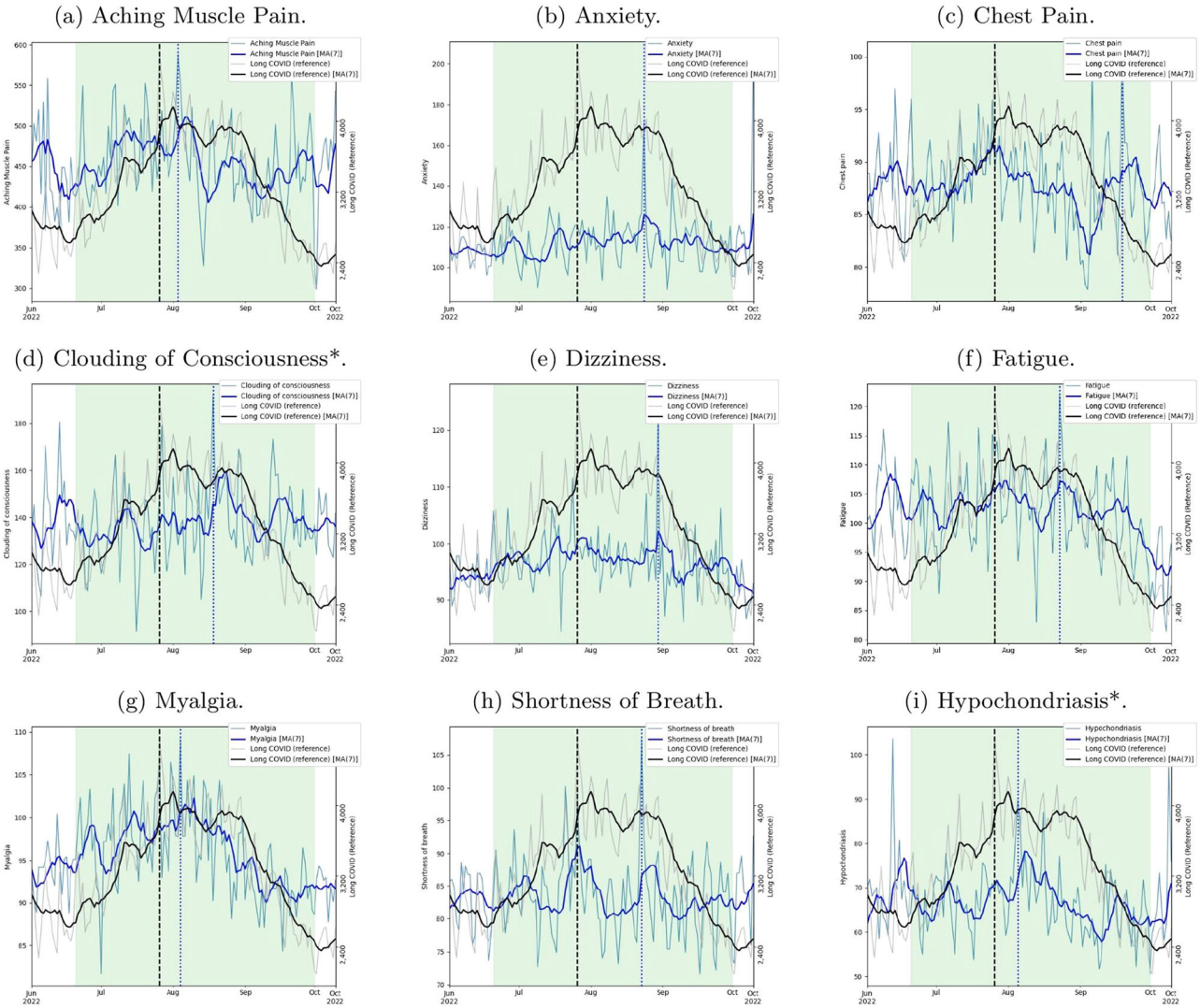

Fig. 6 | The merged search volume (MSV) time series plots of "long COVID" and the related topics classified as Class 2b. Subfigures **a–i** present time series plots of the MSVs for the related topics indicated in the legends and that for "long COVID" for comparison. The light blue and gray lines respectively represent the MSV time series for the symptoms and "long COVID", and the blue and black lines respectively represent the seven-day moving average time series of the MSVs of the symptoms and "long COVID", to better enable visualization. The vertical blue dotted lines show the period maxima of the MSVs of the corresponding related topics, while the black vertical dashed lines show the period maxima of the "long COVID" MSV. The symptoms marked with * indicate that the RSVs were shown to be unstable on Google Trends, and may be subject to larger uncertainties.

observation suggests that the search popularity of "long COVID" was strongest at that time, which correlates with the outbreak of the COVID-19 Omicron variant. Certain related topics, such as ageusia, anosmia, chest pain, headaches, light-headedness, chest tightness, palpitations, and shortness of breath, had their highest search popularity appear earlier than that of "long COVID". Conversely, other related topics, including aching muscle pain, clouding of consciousness, anxiety, mental health, major depressive disorder, hip pain, and insomnia, had their highest search popularity appear after that of "long COVID". This indicates a relationship between the search popularity of different related topics and that of "long COVID."

Results for predicting the prevalence of long COVID and analysis of the evolution of symptoms

We evaluated whether the search volumes of the symptoms downloaded from Google Trends can help to improve the predictive performance of the prevalence of COVID-19. We considered two candidate models: (1) the ARIMA(p, d, q) model[51], which predicts the prevalence of long COVID using only the past MSVs of "long COVID"; and (2) the vector autoregressive with LASSO penalty (lasso-VAR)[52], including the past MSVs of the symptoms as the predictors. We use the lasso-VAR model because there are many predictors (21 MSVs of symptoms), and the least absolute shrinkage and selection operator (LASSO) can help to select important variables for prediction. The predictive performances were evaluated based on the root mean square error (RMSE) and the mean absolute percentage error (MAPE), for $h = 1, \ldots 21$ day-ahead predictions. More details about the methodology can be found in Supplementary Note 6.

Supplementary Table 8 shows the RMSEs and the MAPEs of two candidate models. The RMSEs and MAPEs of the lasso-VAR are all smaller than those using ARIMA(p, d, q). Furthermore, the RMSEs and MAPEs of the lasso-VAR remain stable (RMSEs <500 and MAPEs are all smaller than 10%) when the prediction horizon h increases up to 21 days, whereas those of the ARIMA(p, d, q) dramatically increase as h increases.

The results suggest that including the MSVs of the symptoms can help to improve the predictive performance of the prevalence of long COVID. Google Trends data can be used to forecast and monitor the future prevalence of long COVID.

We also analyzed the estimated parameters in the lasso-VAR over time. A non-zero parameter for a related topic in the lasso-VAR indicates that the

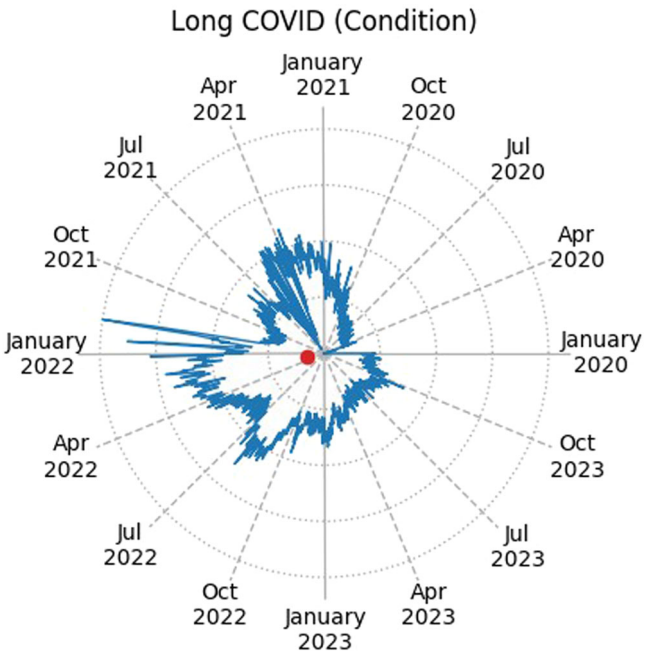

Fig. 7 | The rescaled polar projection of the merged search volume (MSV) for the topic "long COVID". The projection pattern is shown with a blue line, while the centroid is marked with a red dot.

MSV of the related topic is useful in explaining the variation of the MSV of "long COVID". Therefore, we can consider the related topic as a potential symptom of long COVID. The results are presented in Supplementary Note 6.

Discussion

The diverse manifestations of long COVID and the absence of definitive diagnostic tests to confirm the condition often lead to frustrations among long COVID patients when seeking help from healthcare providers[57,58]. This frustration can reduce the likelihood of individuals seeking medical assistance for their long COVID symptoms[57,59], further exacerbating the limitations and potential errors associated with relying solely on medical records to estimate long COVID prevalence. The feelings of "medical gaslighting"[60], as well as medical imparity[61], are detrimental to the trusting relationship between patients and medical professionals, and further hamper patients' health. Consequently, many individuals turn to online searches to find information about long COVID and potential solutions for their condition[62]. Therefore, tracking online search activity can serve as an alternative method to reveal the prevalence of long COVID in the population.

The proposed methodology in this study aims not to complicate nor replace the current practice of ascribing ill-defined symptoms to long COVID. Instead, we aim to provide an alternative method for monitoring long COVID. This methodology can be employed for risk management regarding the prevalence of long COVID. The findings from Google Trends can also help educate and raise awareness among the wider community about the prevalence of various symptoms. Future research can apply clinical methods to validate these findings. Furthermore, our methodology can inform risk assessments regarding long COVID, prompting individuals to seek medical confirmation and treatment, as well as to develop plans for health rehabilitation.

Note that, as the prevalence of long COVID would be affected by the list of symptoms used, in addition to the CDC guidelines, it is possible to extend the list of symptoms with the symptoms listed in the National Institute for Health and Care Excellence (NICE)[63] and the WHO guidelines[64] to obtain a more comprehensive result. We compare the proposed 33 search terms in this

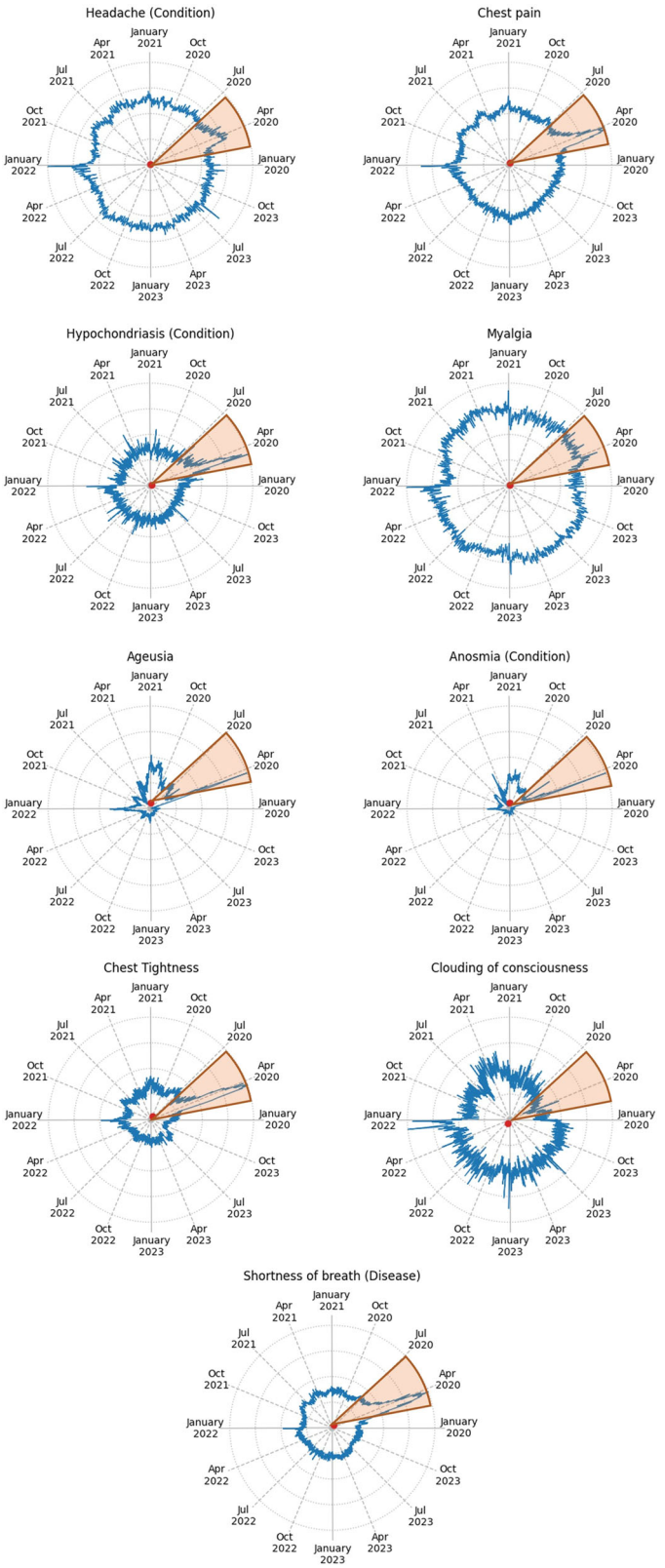

Fig. 8 | Polar projection plots of "long COVID" and the related topics classified as Class 3a. For each polar projection plot, the projection pattern is shown with a blue line, the centroid is marked with a red dot, and the peak is highlighted in orange. The related topics "clouding of consciousness" and "hypochondriasis" were shown to be unstable on Google Trends, and may be subject to larger uncertainties.

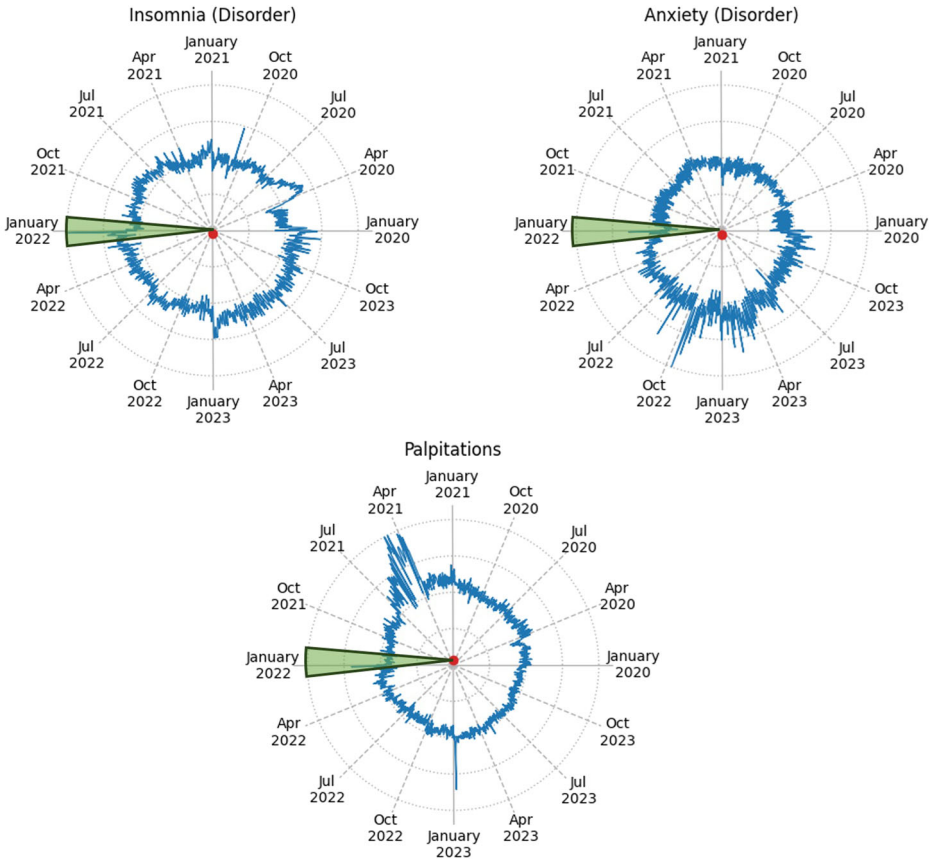

Fig. 9 | Polar projection plots of "long COVID" and the related topics classified as Class 3b. For each polar projection plot, the projection pattern is shown with a blue line, the centroid is marked with a red dot, and the peak is highlighted in green.

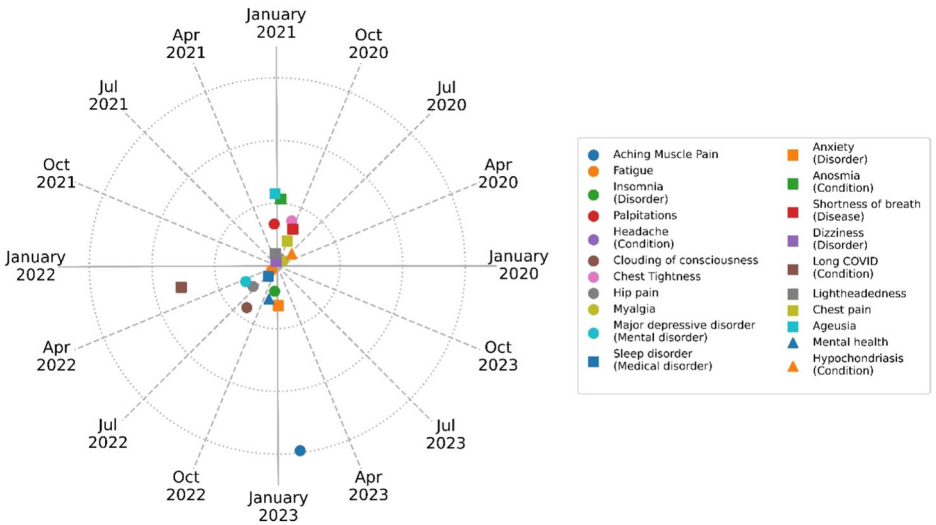

Fig. 10 | Plot of centroids of "long COVID" and the 20 related topics related to long COVID. The related topics "aching muscle pain", "clouding of consciousness", and "hypochondriasis" were shown to be unstable on Google Trends, and may be subject to larger uncertainties.

paper with the NICE and the WHO guidelines in Supplementary Note 7. We observed that, whereas there were some overlaps between the search terms proposed in this paper and the NICE and the WHO guidelines, there were some symptoms we did not include. The 32 search terms selected in this paper are based on the CDC guidelines and 25 highly relevant publications identified from the database "scite", representing a set of most common symptoms for analysis. This approach aims to streamline the analysis and focus on the most relevant symptoms. Although we presented a methodology for utilizing Google Trends data to forecast and monitor long COVID prevalence using this set of most common symptoms, we also acknowledge the potential to extend the analysis by including additional symptoms listed in Supplementary Data 2.

The analysis of MSV time series patterns of "long COVID" revealed two distinct study periods of increased search popularity. The first study period occurred from late November 2021 to early February 2022, while the second study period occurred from late June 2022 to early October 2022. These peaks in search popularity were found to be associated with the prevalence of acute COVID-19 cases. Specifically, the first study period of increased "long COVID" search popularity coincided with the fifth wave of acute COVID-19 cases, while the second study period occurred during the rise of the Omicron sub-variants[54,55] as well as notable surges in COVID-19 cases across Europe[56]. Previous studies have documented that a substantial proportion of individuals develop long COVID following acute COVID-19 infection[21,24]. Therefore, it is reasonable to observe an increase in search

popularity of "long COVID" following the peak of acute COVID-19 infection. These findings align with previous observations of information-seeking behavior in relation to various health conditions[38,65–67].

However, relying solely on the search popularity of "long COVID" may not provide a comprehensive representation of long COVID surveillance, as the condition is still not well-defined in terms of its definition and clinical manifestations. To enhance the infodemiological study of long COVID, it is crucial to investigate the search popularity of symptoms related to long COVID. By exploring the search popularity of these specific symptoms, a more comprehensive understanding of long COVID can be achieved.

We investigated the lead-lag relationship between the search popularities of "long COVID" and related topics in "Search popularity of "long COVID" and the other related topics" and "Polar projection and centroids" sections under the Results section by comparing the time series plots and polar coordinates plots of the MSVs. These findings suggest that physical symptoms garnered the most attention when long COVID initially emerged. On the other hand, as the awareness of long COVID grew, there was an increased focus on psychosocial and mental health conditions associated with long COVID. Monitoring the search popularity of all of these related topics can provide valuable information for long COVID surveillance, enabling a comprehensive understanding of the evolving landscape of symptoms and public concerns related to long COVID. We conducted a prediction study in "Results for predicting the prevalence of long COVID and analysis of the evolution of symptoms" section and analyzed the patterns of the fitted parameters to assess the evolution of the symptoms. The study shows that the related topic is useful in improving the predictability of the MSV of "long COVID", and thus can be used to accurately assess the possible future prevalence of long COVID for risk management purposes. The prediction can provide valuable insights into emerging trends and shifts in the population's perception of long COVID, enabling healthcare providers and policymakers to respond proactively.

Several limitations are present in our study. First, while the selected search terms related to long COVID symptoms were commonly used by individuals seeking information online and were chosen through a comprehensive search of the CDC guidelines and existing literature, it is possible that our list of selected search terms may not encompass all of the search terms related to long COVID, or are not what stakeholders related to long COVID, such as patients and their family, will search for. For instance, search terms related to medications used to treat COVID-19, such as ritonavir, were not included. Furthermore, we can extend the list of symptoms with the symptoms listed in the National Institute for Health and Care Excellence[63] and the WHO guidelines[64] to obtain more comprehensive results.

Even when the list of related topics was built based on a comprehensive literature review, some were shown to be unstable in relation to Google Trends, making it difficult to accurately analyze the relationship between these related topics and the prevalence of long COVID. Discarding these related topics simply due to the instability of the search volumes, however, may not be a rigorous approach. More research is needed to build methodologies in order to obtain stable search volumes for these related topics. Furthermore, there is a need to distinguish between searches made by people with the symptoms and those made by others who are merely interested in knowing more about the disease or for other reasons that do not necessarily mean they have the symptoms. It is also possible that the impact of other conditions and situations may be confused with long COVID by patients, including undesired effects of pharmacological and non-pharmacological interventions, such as vaccinations and lockdown policies. Although we cannot directly identify searches made by people with long COVID symptoms in comparison with people who are simply interested in knowing more about the disease for other reasons, we can filter the search volume using a linear regression. More specifically, our current methodology can be extended by considering the filtered MSVs (i.e., the residuals in the regression models regressing search volumes on the potential confounding factors). By repeating a similar analysis as that presented in this paper using the filtered MSVs, we can compare the two results to provide more comprehensive findings.

In addition, while Google is the dominant search engine used worldwide, with approximately 92% of the global search market share[32], its popularity and availability vary across different regions around the world. As a result, data obtained from areas where Google is more popular may be more representative of these areas, compared to regions such as China, where other search engines are commonly used. This discrepancy in search engine usage can impact the generalizability of the findings, highlighting the need for caution when interpreting and extrapolating results from infodemiological studies conducted primarily using Google search data.

Conclusion

In this study, we examined the potential of using online search data to track the prevalence of long COVID symptoms. Analysis of search activities for 33 search terms and 20 related topics revealed potential findings. Eight related topics (ageusia, anosmia, chest tightness, headaches, hip pain, insomnia, mental health, and major depressive disorder) consistently preceded "long COVID" in search popularity, suggesting their potential as predictors for long COVID surveillance. Conversely, nine related topics (aching muscle pain, anxiety, chest pain, clouding of consciousness, dizziness, fatigue, myalgia, shortness of breath, and hypochondriasis) appeared after "long COVID", representing long-lasting symptoms and suggesting their potential for monitoring long COVID effects. Polar projections and centroids analysis highlighted the increased attention paid to psychosocial and mental illnesses as long COVID persists. We also conducted a prediction study, showing that the prediction of the prevalence of long COVID can be improved by including the MSVs of the symptoms in a regression model. The fitted parameters in the prediction models can also be used to assess the evolution of the symptoms of long COVID. Monitoring the search activities of these related topics can help us to understand trends in long COVID-19 symptoms surveillance and in turn predict long COVID prevalence, which will provide necessary data for the public, especially in the context of the government's primary healthcare reforms focusing on prevention.

Data availability

The relative search volume data used in this study are freely available on the Google Trends website. The generated data, including the merged search volumes, are available in the Zenodo repository with the identifier https://doi.org/10.5281/zenodo.14997721[68]. The Oxford COVID-19 Government Response Tracker dataset is available online from the repository[69] as referenced in the research conducted by Hale et al.[49]. However, the financial news data used in this study are not publicly available due to the license of these datasets. A subscription to Reuters is required to access the data.

Code availability

The source code files used in this study are available in the Zenodo repository with the identifier https://doi.org/10.5281/zenodo.14997721[68].

Received: 13 August 2024; Accepted: 6 May 2025;
Published online: 16 May 2025

References

1. Rocklöv, J., Wilder-Smith, A., Gayle, A. A. & Liu, Y. The reproductive number of COVID-19 is higher compared to SARS coronavirus. *J. Travel Med.* **27**, https://doi.org/10.1093/jtm/taaa021 (2020).
2. Centers for Disease Control and Prevention. *COVID-19 Data Review: Update on COVID-19-Releated Mortality*. Retrieved from https://www.cdc.gov/coronavirus/2019-ncov/science/data-review/index.html (Centers for Disease Control and Prevention, 2023).
3. Zhou, Y. et al. Comorbidities and the risk of severe or fatal outcomes associated with coronavirus disease 2019: a systematic review and meta-analysis. *Int. J. Infect. Dis.* **99**, 47–56 (2020).
4. World Health Organization. *WHO COVID-19 Dashboard*. Retrieved 18 Feb from https://data.who.int/dashboards/covid19/cases?n=c (World Health Organization, 2024).

5. Robineau, O. et al. Persistent symptoms after the first wave of COVID-19 in relation to SARS-CoV−2 serology and experience of acute symptoms: a nested survey in a population-based cohort. *Lancet Reg Health Eur.* **17**, https://doi.org/10.1016/j.lanepe.2022.100363 (2022).
6. Docherty, A. B. et al. Features of 20,133 UK patients in hospital with COVID-19 using the ISARIC WHO clinical characterisation protocol: prospective observational cohort study. *BMJ*. https://doi.org/10.1136/bmj.m1985 (2020).
7. Centers for Disease Control and Prevention. *Post-COVID Conditions: Information for Healthcare Providers*. Retrieved from https://www.cdc.gov/coronavirus/2019-ncov/science/data-review/index.html (Centers for Disease Control and Prevention, 2024).
8. Soriano, J. B., Murthy, S., Marshall, J. C., Relan, P. & Diaz, J. V. A clinical case definition of post-COVID-19 condition by a Delphi consensus. *Lancet Infect. Dis.* **22**, e102–e107 (2022).
9. Garner, P. Paul Garner: for 7 weeks I have been through a roller coaster of ill health, extreme emotions, and utter exhaustion. https://blogs.bmj.com/bmj/2020/05/05/paul-garner-people-who-have-a-more-protracted-illness-need-help-to-understand-and-cope-with-the-constantly-shifting-bizarre-symptoms/ (2020).
10. Davis, H. E., McCorkell, L., Vogel, J. M. & Topol, E. J. Long COVID: major findings, mechanisms and recommendations. *Nat. Rev. Microbiol.* **21**, 133–146 (2023).
11. Raveendran, A. V., Jayadevan, R. & Sashidharan, S. Long COVID: an overview. *Diab. Metab. Syndrome Clin. Res. Rev.* **15**, 869–875 (2021).
12. Davis, H. E. et al. Characterizing long COVID in an international cohort: 7 months of symptoms and their impact. *eClinicalMedicine* **38**, https://doi.org/10.1016/j.eclinm.2021.101019 (2021)
13. Greenhalgh, T., Knight, M., A'Court, C., Buxton, M. & Husain, L. Management of post-acute COVID-19 in primary care. *BMJ*. https://doi.org/10.1136/bmj.m3026 (2020).
14. Wu, Y. et al. Nervous system involvement after infection with COVID −19 and other coronaviruses. *Brain Behav. Immun.* **87**, 18–22 (2020).
15. So, M. K. P., Chu, A. M. Y. & Tiwari, A. Persistent symptoms after SARS-CoV-2 infection: long-term implications for health and quality of life. *Lancet Reg. Health–Eur.* **17**, https://doi.org/10.1016/j.lanepe.2022.100373 (2022).
16. Besteher, B. et al. Cortical thickness alterations and systemic inflammation define long-COVID patients with cognitive impairment. *Brain Behav. Immun.* **116**, 175–184 (2024).
17. Greene, C. et al. Blood–brain barrier disruption and sustained systemic inflammation in individuals with long COVID-associated cognitive impairment. *Nat. Neurosci.* **27**, 421–432 (2024).
18. Greenhalgh, T., Sivan, M., Perlowski, A. & Nikolich, J. Ž. Long COVID: a clinical update. *Lancet* **404**, 707–724 (2024).
19. Lai, C.-C. et al. Long COVID: an inevitable sequela of SARS-CoV-2 infection. *J. Microbiol. Immunol. Infect.* **56**, 1–9 (2023).
20. Cabrera Martimbianco, A. L., Pacheco, R. L., Bagattini, Â. M. & Riera, R. Frequency, signs and symptoms, and criteria adopted for long COVID-19: a systematic review. *Int. J. Clin. Pract.* **75**, https://doi.org/10.1111/ijcp.14357 (2021).
21. World Health Organization. Coronavirus disease (COVID−19): Post COVID-19 condition. Retrieved from https://www.who.int/news-room/questions-and-answers/item/coronavirus-disease-(covid-19)-post-covid-19-condition (2023).
22. Office for National Statistics. Prevalence of ongoing symptoms following coronavirus (COVID-19) infection in the UK: 5 January 2023. Retrieved from, https://www.ons.gov.uk/peoplepopulationandcommunity/healthandsocialcare/conditionsanddiseases (2023).
23. Carfì, A., Bernabei, R. & Landi, F. Persistent symptoms in patients after acute COVID-19. *JAMA*, **324**, https://doi.org/10.1001/jama.2020.12603 (2020).
24. O'Mahoney, L. L. et al. The prevalence and long-term health effects of Long Covid among hospitalised and non-hospitalised populations: a systematic review and meta-analysis. *eClinicalMedicine*, **55**, https://doi.org/10.1016/j.eclinm.2022.101762 (2023).
25. Yao, L. et al. Was the rate of Long Covid as high as 45%—a scary report with flaw. *eClinicalMedicine* **59**, 101949 (2023).
26. Pfaff, E. R. et al. Identifying who has long COVID in the USA: a machine learning approach using N3C data. *Lancet Digit. Health* **4**, e532–e541 (2022).
27. Rao, S. et al. Clinical features and burden of postacute sequelae of SARS-CoV-2 infection in children and adolescents. *JAMA Pediatr.* **176**, https://doi.org/10.1001/jamapediatrics.2022.2800 (2022).
28. Kemp, S. *Digital 2022: Global Overview Report*. https://datareportal.com/reports/digital-2022-global-overview-report (2022).
29. Eysenbach, G. Infodemiology and Infoveillance: framework for an emerging set of public health informatics methods to analyze search, communication and publication behavior on the internet. *J. Med. Internet Res.* **11**, https://doi.org/10.2196/jmir.1157 (2009).
30. Chu, J. T. W. et al. How, when and why people seek health information online: qualitative study in Hong Kong. *Interact. J. Med. Res.* **6**, https://doi.org/10.2196/ijmr.7000 (2017).
31. Google. *Google Trends*. Google. Retrieved 1 Jun from https://trends.google.com/ (2024).
32. Rumage, J. *16 Top Search Engines to Try*. builtin. Retrieved from https://builtin.com/consumer-tech/search-engines-list (2023).
33. Jun, S.-P., Yoo, H. S. & Choi, S. Ten years of research change using Google Trends: from the perspective of big data utilizations and applications. *Technol. Forecast. Soc. Change* **130**, 69–87 (2018).
34. Alicino, C. et al. Assessing Ebola-related web search behaviour: insights and implications from an analytical study of Google Trends-based query volumes. *Infect. Dis. Poverty* **4**, https://doi.org/10.1186/s40249-015-0090-9 (2015).
35. Shin, S.-Y. et al. High correlation of Middle East respiratory syndrome spread with Google search and Twitter trends in Korea. *Sci. Rep.* **6**, https://doi.org/10.1038/srep32920 (2016).
36. Syamsuddin, M., Fakhruddin, M., Sahetapy-Engel, J. T. M. & Soewono, E. Causality analysis of Google Trends and dengue incidence in Bandung, Indonesia with linkage of digital data modeling: longitudinal observational study. *J. Med. Internet Res.* **22**, https://doi.org/10.2196/17633 (2020).
37. Springer, S., Menzel, L. M. & Zieger, M. Google Trends provides a tool to monitor population concerns and information needs during COVID-19 pandemic. *Brain Behav. Immun.* **87**, 109–110 (2020).
38. Chu, A. M. Y., Chong, A. C. Y., Lai, N. H. T., Tiwari, A. & So, M. K. P. Enhancing the predictive power of Google Trends data through network analysis: infodemiology study of COVID-19. *JMIR Public Health Surveill.* **9**, e42446 (2023).
39. Wynberg, E. et al. Evolution of coronavirus disease 2019 (COVID-19) symptoms during the first 12 months after illness onset. *Clin. Infect. Dis.* **75**, e482–e490 (2022).
40. ISARIC. Clinical Data Collection—the COVID-19 Case Report Forms (CRFs). https://isaric.org/research/covid-19-clinical-research-resources/covid-19-crf/ (2023).
41. Strzelecki, A., Azevedo, A. & Albuquerque, A. Correlation between the spread of COVID-19 and the interest in personal protective measures in Poland and Portugal. *Healthcare* **8**, 203. https://doi.org/10.3390/healthcare8030203 (2020).
42. Nicholson, J. M. et al. scite: a smart citation index that displays the context of citations and classifies their intent using deep learning. *Quant. Sci. Stud.* **2**, 882–898 (2021).
43. Olson, D. R. et al. Could Google Trends be used to predict methamphetamine-related crime? An analysis of search volume data

in Switzerland, Germany, and Austria. *Plos One* **11**, https://doi.org/10.1371/journal.pone.0166566 (2016).
44. Borup, D. & Schütte, E. C. M. In search of a job: forecasting employment growth using Google Trends. *J. Bus. Econ. Stat.* **40**, 186–200 (2020).
45. Li, X., Ma, J., Wang, S. & Zhang, X. How does Google search affect trader positions and crude oil prices? *Econ. Model.* **49**, 162–171 (2015).
46. Mavragani, A. & Ochoa, G. Google Trends in infodemiology and infoveillance: methodology framework. *JMIR Public Health Surveill.* **5**, https://doi.org/10.2196/13439 (2019).
47. Rovetta, A. Google trends in infodemiology: methodological steps to avoid irreproducible results and invalid conclusions. *Int. J. Med. Inform.* **190**, 105563 (2024).
48. Sato, K., Mano, T., Iwata, A. & Toda, T. Need of care in interpreting Google Trends-based COVID-19 infodemiological study results: potential risk of false-positivity. *BMC Med. Res. Methodol.* **21**, 147 (2021).
49. Hale, T. et al. A global panel database of pandemic policies (Oxford COVID-19 Government Response Tracker). *Nat. Hum. Behav.* **5**, 529–538 (2021).
50. Kumar, N. & Susan, S. COVID-19 pandemic prediction using time series forecasting models. In: *Proc. 11th International Conference on Computing, Communication and Networking Technologies (ICCCNT)*, Kharagpur, India, pp. 1–7. https://doi.org/10.1109/ICCCNT49239.2020.9225319 (2020).
51. Box, G. E., Jenkins, G. M., Reinsel, G. C. & Ljung, G. M. *Time Series Analysis: Forecasting and Control* (John Wiley & Sons, 2015).
52. Hsu, N. J., Hung, H. L. & Chang, Y. M. Subset selection for vector autoregressive processes using LASSO. *Comput. Stat. Data Anal.* **52**, 3645–3657 (2008).
53. Chatterjee, S., Bhattacharya, M., Nag, S., Dhama, K. & Chakraborty, C. A detailed overview of SARS-CoV-2 Omicron: its sub-variants, mutations and pathophysiology, clinical characteristics, immunological landscape, immune escape, and therapies. *Viruses* **15**, https://doi.org/10.3390/v15010167 (2023).
54. Elbe, S. & Buckland-Merrett, G. Data, disease and diplomacy: GISAID's innovative contribution to global health. *Glob. Chall.* **1**, 33–46 (2017).
55. Hodcroft, E. B. *CoVariants: SARS-CoV-2 mutations and variants of interest*. Retrieved May 24 from https://covariants.org/ (2021).
56. Mathieu, E. et al. *Coronavirus Pandemic (COVID-19)*. OurWorldInData.org. Retrieved 4 Jan 2024 from https://ourworldindata.org/coronavirus (2020).
57. Au, L., Capotescu, C., Eyal, G. & Finestone, G. Long covid and medical gaslighting: dismissal, delayed diagnosis, and deferred treatment. *SSM Qual. Res. Health*, *2*. https://doi.org/10.1016/j.ssmqr.2022.100167 (2022).
58. Callard, F. & Perego, E. How and why patients made Long Covid. *Soc. Sci. Med.* **268**, https://doi.org/10.1016/j.socscimed.2020.113426 (2021).
59. McNabb, K. C. et al. "It was almost like it's set up for people to fail": a qualitative analysis of experiences and unmet supportive needs of people with Long COVID. *BMC Public Health* **23**, https://doi.org/10.1186/s12889-023-17033-4 (2023).
60. Barnes, E. Trust, distrust, and 'medical gaslighting. *Philos. Q.* **73**, 649–676 (2023).
61. Shabnam, S. et al. Socioeconomic inequalities of long COVID: a retrospective population-based cohort study in the United Kingdom. *J. R. Soc. Med.* **116**, 263–273 (2023).
62. Russell, D. et al. Support amid uncertainty: long COVID illness experiences and the role of online communities. *SSM Qual. Res. Health* **2**, https://doi.org/10.1016/j.ssmqr.2022.100177 (2022).
63. National Institute for Health and Care Excellence (NICE). 9 common symptoms: COVID-19 rapid guideline: managing the long-term effects of COVID-19: guidance. https://www.nice.org.uk/guidance/ng188/chapter/9-Common-symptoms (2024).
64. Rajan, S. et al. *POLICY BRIEF 39: In the Wake of the Pandemic, Preparing for Long COVID*. https://apps.who.int/iris/bitstream/handle/10665/339629/Policy-brief-39-1997-8073-eng.pdf (WHO Regional Office for Europe, 2021).
65. Beck, F. et al. Use of the internet as a health information resource among French young adults: results from a nationally representative survey. *J. Med. Internet Res.* **16**, https://doi.org/10.2196/jmir.2934 (2014).
66. Kalichman, S. C. et al. Internet use among people living with HIV/AIDS: coping and health-related correlates. *AIDS Patient Care STDs* **19**, 439–448 (2005).
67. Mangono, T. et al. Information-seeking patterns during the COVID-19 pandemic across the United States: longitudinal analysis of Google Trends data. *J. Med. Internet Res.* **23**, https://doi.org/10.2196/22933 (2021).
68. Chu, A. M. Y., Tsang, J. T. Y., Chan, S. S. C., Chan, L. S. H. & So, M. K. P. Source code for "Utilizing Google trends data to enhance forecasts and monitor Long COVID prevalence". Zenodo. https://doi.org/10.5281/zenodo.14997721 (2025).
69. Phillips, T. Oxford Covid-19 Government Response Tracker (OxCGRT). Retrieved from https://github.com/OxCGRT/covid-policy-dataset (2023).

Acknowledgements

The work was supported by the research grants from the Department of Social Sciences and Policy Studies, The Education University of Hong Kong and, and The Hong Kong University of Science and Technology research grant "Risk Analytics and Applications" (grant SBMDF21BM07). The funders had no role in the study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Author contributions

A.M.Y.C. and M.K.P.S. conceptualized the study. A.M.Y.C. and J.T.Y.T. set up the research questions. M.K.P.S. and L.S.H.C. collected and analyzed the data with statistical modeling. A.M.Y.C., J.T.Y.T., S.S.C.C., L.S.H.C., and M.K.P.S. interpreted the results. A.M.Y.C., J.T.Y.T., S.S.C.C., and L.S.H.C. drafted the manuscript. A.M.Y.C., M.K.P.S., and S.S.C.C. finalized the manuscript. All authors reviewed the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s43856-025-00896-6.

Correspondence and requests for materials should be addressed to Mike K. P. So.

Peer review information *Communications Medicine* thanks Chiranjib Chakraborty, Alessandro Rovetta and Nyarie Sithole for their contribution to the peer review of this work.