CAMBRIDGE
UNIVERSITY PRESS

RESEARCH ARTICLE

Enhancing the accuracy of COVID-19 incidence and mortality predictions using Google Trends data across the 50 US states and the District of Columbia

Aleksandr Shishkin[1], Kiril Kuzmin[2], Gerardo Chowell[1], Yuriy Gankin[3], Alexander Perez Tchernov[4], Pavel Skums[5] and Alexander Kirpich[1]

[1]Department of Population Health Sciences, School of Public Health, Georgia State University, Atlanta, GA, USA
[2]Department of Computer Science, Georgia State University, Atlanta, GA, USA
[3]Quantori, Cambridge, MA, USA
[4]Faculty of Mechanics and Mathematics, Belarusian State University, Minsk, Belarus
[5]School of Computing, University of Connecticut, Storrs, CT, USA
Corresponding author: Alexander Kirpich; Email: akirpich@gsu.edu

Abstract

Quick and accurate forecasts of incidence and mortality trends for the near future are particularly useful for the immediate allocation of available public health resources, as well as for understanding the long-term course of the pandemic. The surveillance data used for predictions, however, may come with some reporting delays. Consequently, auxiliary data sources that are available immediately can provide valuable additional information for recent time periods for which surveillance data have not yet become fully available. In this work, a set of Google search queries by individual users related to COVID-19 incidence and mortality is collected and analyzed. The information from these queries aims to improve quick forecasts. Initially, the identified search query keywords were ranked according to their predictive abilities with reported incidence and mortality. After that, the ARIMA, Prophet, and XGBoost models were fitted to generate forecasts using only the available reported incidence and mortality (baseline model) or together with combinations of searched keywords identified based on their predictive abilities (predictors model). In summary, the inclusion of top-ranked keywords as predictors significantly enhanced prediction accuracy for the majority of scenarios in the range from 50% to 90% across all considered models and is recommended for future use. The inclusion of low-ranked keywords did not provide such an improvement. In general, the ranking of predictors and the corresponding forecast improvements were more pronounced for incidence, while the results were less pronounced for mortality.

Policy Significance Statement

The forecasting of infectious disease trends presents multiple challenges for near-future decisions and related policies due to reporting delays, data limitations, and rapidly changing epidemiological dynamics. In this article, the use of Google Trends search query data are studied to improve short-term forecasts of COVID-19 incidence and mortality, comparing traditional models with those incorporating search trends. The models that incorporate

---

Google Trends show substantial improvements in prediction accuracy for incidence forecasts when relevant search terms are included, while the impact on mortality predictions is moderate. The findings highlight the potential of digital surveillance tools to enhance epidemic forecasting, offering a valuable resource for timely public health decision making.

1. Introduction

The COVID-19 pandemic, caused by the SARS-CoV-2 virus, began in late 2019 and resulted in approximately 6.6 million lost lives (WHO Coronavirus (COVID-19) Dashboard, 2025). The pandemic revealed many public health challenges, including the imperfections in the available reporting systems for incidence and mortality (Hanaei et al., 2021), which are common in epidemics and outbreaks in general (Hoffman and Silverberg, 2018). Consequently, even though COVID-19-related incidence and mortality data had been widely monitored, collected, and reported, the quality of reporting varied substantially by country and within regions of each country. These variations arose from multiple factors and imperfections in the surveillance system (Osmani, 2020; Reis et al., 2020; Albani et al., 2021; Balashov et al., 2021). In particular, the *delays* in reporting can occur due to several factors needed for sample processing, such as the unavailability of test kits, shortages of hospital personnel, hospital overloads during pandemic peaks, and difficulties in case registration without a centralized database (Lau et al., 2021). These delays may also not be homogeneous over time. They can be exacerbated during peak periods when facilities and medical personnel are most overloaded.

The imperfections in the available reporting systems may make decision making challenging, and any auxiliary data sources about the pandemic that are available without any delays may complement the existing knowledge and improve the quality of future predictions. One such auxiliary source is the information that people are searching for on the Internet (Liu et al., 2016; Ciaffi et al., 2020; Sulyok et al., 2021; Amusa et al., 2022; Kirpich et al., 2022; Shishkin et al., 2023).

Unlike case reporting, which requires rapid availability of healthcare reports and a healthcare system that is functioning without any issues in the related infrastructure, Internet Google Trends search query frequencies are reported automatically and are available to public instantaneously (Google Trends, 2025). Researchers can use Google Trends data for multiple reasons, such as filling the gaps in the reporting by nowcasting incidence for those gaps (Mohebbi et al., 2011), as well as providing rapid forecasts in the epidemic trends for the near future when reported data are either not immediately available or not completely available (e.g., weekends and long holidays). This can be done by carefully selecting the appropriate search queries (aka keywords and key phrases) and using them as exogenous variables (i.e., predictors) in the model individually or in various combinations. The rules and recommendations for selecting predictors, along with the choice of the final set of predictors to be used for analysis, remain an open question (Nuti et al., 2014; Cervellin et al., 2017).

Many studies have already examined the potential of Google Trends data in understanding and predicting COVID-19 incidence patterns. Previous work in this area often focused on analyzing the frequency of a single search keyword, such as "COVID," about reported cases and disease spread (Kurian et al., 2020; Mavragani and Gkillas, 2020; Walker et al., 2020; Rovetta, 2021; Sulyok et al., 2021). Other studies expanded their scope to small sets of related search keywords, seeking to capture a broader array of public interest indicators (Ayyoubzadeh et al., 2020; Kurian et al., 2020; Nguyen et al., 2020; Sousa-Pinto et al., 2020; Springer et al., 2020; Prasanth et al., 2021; Amusa et al., 2022; Zayed et al., 2023). Most of these studies emphasized the correlation between search frequencies and reported COVID-19 case counts, illustrating the close relationship between online search behavior and real-world epidemiological trends (Kurian et al., 2020; Mavragani and Gkillas, 2020; Sousa-Pinto et al., 2020; Walker et al., 2020; Rovetta, 2021). In addition to correlational studies, several predictive models have been utilized to leverage Google Trends data in COVID-19 forecasting. These include traditional approaches like regression models (Nguyen et al., 2020) and ARIMA (Amusa et al., 2022), as well as more complex models such as

generalized additive models (GAMs) (Sulyok et al., 2021) and machine learning algorithms, including recurrent neural networks (Ayyoubzadeh et al., 2020; Nguyen et al., 2020; Prasanth et al., 2021; Rodríguez et al., 2024).

Compared to prior studies that typically analyzed limited keyword sets or focused on national-level predictions, this study aims to build upon and extend existing research by exploring a broader set of search keywords across all 50 states of the United States of America and the District of Columbia. We investigate how Google Trends data can improve model performance across different predictive frameworks while proposing a standardized methodology for incorporating Google Trends into epidemiological forecasting. This approach is intended to support more geographically granular and methodologically consistent analyses, with the potential to yield more accurate, real-time insights into epidemic trajectories and inform public health responses through timely, data-driven predictions.

Several studies that have already been published have attempted to evaluate the utility of Google Trends data in modeling COVID-19 incidence in various countries. However, these efforts were generally limited in geographic scope, keyword selection, or modeling depth. One study analyzed data at the national level, but considered only case counts (Ayyoubzadeh et al., 2020). Another relied primarily on a narrow set of keywords and employed a single ARIMA model (Amusa et al., 2022). A third study provided a comprehensive comparison of incidence and mortality across the US states but did not incorporate temporal dynamics (Mavragani and Gkillas, 2020). Other works, while applying more advanced modeling techniques, were constrained by limited keyword sets and lower geographic resolution (Nguyen et al., 2020; Prasanth et al., 2021). A closely related study on COVID-19 focused on how factors such as anxiety, worry, and perceived lack of knowledge motivated people to seek online health information, particularly regarding vaccination, including searches related to vaccine intent (Moffett et al., 2024).

In contrast to these studies, the present work aims to broaden both the scope and methodological rigor by analyzing a wider range of search keywords across all 50 US states and the District of Columbia, thus incorporating geographic components and the complete temporal span of the pandemic, thereby incorporating the temporal complement. The main goal of the presented work is to evaluate the utility of aggregated Google Trends search data as a potential supplementary predictor of COVID-19 incidence and mortality trends. In other words, the goal is to assess whether certain empirically and intuitively selected search terms, regardless of their specific motivational underpinnings, can provide useful signals for public health surveillance and forecasting. The secondary goals are to compare the performances of three distinct forecasting models, including ARIMA, Prophet, and XGBoost, within a standardized framework, in order to systematically assess the contribution of Google Trends data across different geographies, time spans, and model structures. Unlike most prior research, which often emphasized correlation or a single modeling paradigm, the predictive accuracy and robustness is evaluated under multiple model types, locations, and calibration window lengths. The policy goals are not only to justify the enhancement in model performance, but also to establish methodological guidance for future digital epidemiology studies in the field.

This study focused on models well suited for time series data, which both exhibit highly irregular behavior and support the straightforward inclusion of multiple external predictors. Since the analysis aimed to evaluate changes in model performance with the addition of predictors derived from Google Trends data, classical epidemiological compartmental models such as the *SIR* and *SIRS* frameworks were not considered. Although these traditional compartmental models and their variants are powerful and versatile, they pose challenges both in modeling multiwave dynamics and in incorporating external predictors. Each of these topics lacks a standardized approach and constitutes its own area of ongoing research.

The presented study aimed to assess the feasibility and performance of Google Trends data, readily available in real time, as a standalone supplementary source for predictive modeling. Although numerous alternative data sources can serve as valuable predictors, they are often challenging to compile quickly and, as a result, are typically not available in real time. These limitations reduce their utility for the rapid analysis required in epidemic forecasting. Therefore, the presented analysis

was intentionally focused on search interest data to isolate and better understand its unique contribution, independent of additional predictors that, unlike search data, may be difficult to obtain in real time. The overarching long-term goals are to "prove the principles" and to illustrate the approach's utility for the reported incidence of COVID-19 and related mortality from COVID-19 in all 50 states of the United States of America, together with the District of Columbia. In particular, the goals are to (a) investigate the set of predictors that can be used for COVID-19 incidence and related mortality predictions based on nonstatistical (e.g., epidemiological, medical, or any other relevant) reasoning; (b) select the most prominent (i.e., appropriate) subsets of predictors from the originally identified set of predictors; (c) identify how those selected predictors and their combinations affect the overall performance of the selected models in terms of the accuracy of predictions; and (d) compare different modeling frameworks in terms of their performance and prediction abilities based on both the selected sets of predictors, geographic regions, as well as the length (i.e., number of reported time slots) of the calibration period, which may be especially crucial for the early stages of the pandemics.

2. Methods

2.1. Data sources

The incidence and mortality data for COVID-19 in the United States were obtained from the Johns Hopkins University Data Repository for COVID-19 (COVID-19 Data Repository by the Center for Systems Science and Engineering (CSSE) at Johns Hopkins University, 2025) that collected aggregated data from multiple sources for multiple countries. For the USA specifically, these data were accumulated from health departments nationwide. Data counts were available daily at the county and state levels. At the same time, the reporting relied on the case definition issued by the CDC and was updated twice throughout the course of the pandemic (Coronavirus Disease 2019 (COVID-19), 2020). The dataset included both COVID-19 cumulative incidence and cumulative related mortality. The reported cases were available from January 22, 2020 until March 9, 2023. The available daily COVID-19 incidence and mortality data were aggregated into monthly counts for two main reasons: to reduce numerical instabilities, noise, and outliers caused by irregular reporting (such as sporadic delays or gaps on weekends and holidays, which vary between states); and, more importantly, because Google Trends data covering multiple years are only available at a monthly frequency. It should be noted, however, that the methods presented and utilized can be applied to any equally spaced time intervals, such as days or weeks, provided that reliable data are available for the early stages of future pandemics. The application of these methods remains identical. The only potentially necessary adjustment for shorter time intervals involves modifying the length of the lags used in the cross-correlations, as discussed further.

The Google Trends search queries data are publicly available and can be downloaded from the corresponding website (Google Trends, 2025). These data are anonymized (ensuring the absence of personal data), categorized (assigning topics to search queries), and aggregated (grouped) (FAQ about Google Trends data - Trends Help, 2025). In particular, Google Trends data do not contain complete individual searches but instead contain information about random samples from all search queries within a given period. While the sample sizes used for such summaries and the details of the sampling method are not disclosed, the Google Trends summaries are claimed to be sufficient for a selected geographic region, i.e. states in the presented analysis, along with Washington, DC. For each preselected keyword or key phrase, the resulting data are available as a scaled time series with a time interval length dependent on the length of the requested timeframe chosen to generate the series. In particular, for requested reporting periods longer than twelve months, the reporting time interval duration is one month. The numbers returned by Google Trends for a given search keyword or key phrase are already scaled to a range of 0 to 100, with 100 representing the highest value for the selected timeframe and 0 representing the lowest (Rogers, 2016). In this study, for each of the preselected keywords and key phrases, for each geographic location, the resulting time series is used as an

auxiliary predictor variable, which is used to improve the prediction of incidence rates and mortality counts. In particular, the Google Trends data were downloaded monthly between January 2017 and January 2023 to cover 5 years. However, only monthly records from January 2020 corresponded with the dates when COVID-19 reporting in the United States started. Data were downloaded using modified scraper code from Emiliano Mastragostino's GitHub repository (Scraper for extracting data from Google Trends, 2025). The modified scraper code has been made available on GitHub, as well as the rest of the project code (GitHub repository for the Google Trends Study, 2025). The source Google Trends can be redownloaded from the official website (Google Trends, 2025) using this scraper. The complete lists of the keywords selected for the study for incidence and mortality predictions are summarized in Supplementary Tables S1 and S2.

2.2. Keywords selection and time series models

2.2.1. Relationships between incidence, mortality, and corresponding keywords

One of the key challenges with keyword usage is that the terms an individual searches for may not necessarily be directly linked to experiencing the relevant symptoms of COVID-19 or an attempt to confirm the infection, either directly or indirectly. The reasons behind these searches can vary widely. For instance, some searches may stem from a general interest in the topic rather than from individuals who believe they have COVID-19 (Butler, 2013; Nghiem et al., 2016; Mavragani et al., 2019; Ming et al., 2021; Satpathy et al., 2023). Distinguishing between these (and potentially many other) reasons is impossible, as the exact keywords are often used for multiple purposes.

As a result, the realistic goal should be to identify phrases that people search for most strongly associated with the actual (i.e., reported) incidence. In other words, these phrases should serve as the best "proxies" or indicators of the number of cases. In this context, the exact underlying reasons for the search are less relevant. It does not matter whether the search keywords are directly related to symptoms, treatment, or simply to the general interest in COVID-19 sparked by media coverage (Butler, 2013; Limilia and Pratamawaty, 2020; Ming et al., 2021; Satpathy et al., 2023), as long as they serve as reasonable proxies that can be used to improve predictions.

To summarize, the first recommended step is to select a sufficient number of relevant keywords and key phrases that cover various aspects of the disease being studied (Nuti et al., 2014). These should include terms related to the disease name(s) (e.g., "COVID," "novel coronavirus," "coronavirus disease," "SARS-CoV-2"), symptoms (e.g., "loss of smell," "shortness of breath," "fatigue"), as well as diagnosis and treatment (e.g., "COVID-19 test," "COVID-19 medications"), and possibly other related topics. The selection of keywords was based on previously published studies that demonstrated improvements in forecasting performance (Ayyoubzadeh et al., 2020; Kurian et al., 2020; Mavragani and Gkillas, 2020; Nguyen et al., 2020; Sousa-Pinto et al., 2020; Springer et al., 2020; Walker et al., 2020; Cinarka et al., 2021; Lampos et al., 2021; Prasanth et al., 2021; Rovetta, 2021; Sulyok et al., 2021; Amusa et al., 2022; Chen et al., 2022; Ma and Yang, 2022; Wang et al., 2022; Marty et al., 2023; Zayed et al., 2023; Chu et al., 2025; Nia et al., 2025). In addition, new keywords and phrases deemed suitable were also included to improve the analysis.

The exact same approach should be used for COVID-19-related mortality, where the search keywords should be related to various funeral terminology, such as "burial," "cemetery," "coffin," "cremation," "funeral," "flowers," and "memorial service."

2.2.2. Granger causality test

The first approach to selecting a keyword pool was based on the Granger causality test. The Granger causality test is a statistical technique used to investigate how helpful a one time series is for predicting another time series (Granger, 1969). This test does not test or justify any causal relationships. Instead, it provides the utility of one time series for predicting the other series. The test is based on the idea that if the values of series $X(t)$ significantly (using some criteria) improve the prediction of the values of series $Y(t)$ beyond the information contained in the lagged values of series $Y(t)$ alone, then $X(t)$ helps predict $Y(t)$.

The test has been implemented via the function Granger test from the R package lmtest. The Granger causality test is performed as a two-step procedure. In the first step, autoregressive models are fitted individually for search frequency and COVID-19 incidence or mortality. These models only include lagged values of the respective variable as predictors (Granger, 1969):

$$X(t) = \sum_{i=1}^{p} \phi_{X,i} X(t-i) + \varepsilon_X(t) \tag{1}$$

$$Y(t) = \sum_{i=1}^{p} \phi_{Y,i} Y(t-i) + \varepsilon_Y(t) \tag{2}$$

where p represents the number of lagged terms included in the model, $\phi_{X,i}$ and $\phi_{Y,i}$ are the estimated autoregressive coefficients, and $\varepsilon_X(t)$ and $\varepsilon_Y(t)$ are the error terms.

In the second step, the corresponding autoregressive models are extended by including the lagged values of the other trend, which is either the search trend or the incidence or mortality series, as predictors (Granger, 1969):

$$X(t) = \sum_{i=1}^{p} \phi_{X,i} X(t-i) + \sum_{i=1}^{q} \theta_{X,i} Y(t-i) + \varepsilon_X(t) \tag{3}$$

$$Y(t) = \sum_{i=1}^{p} \phi_{Y,i} Y(t-i) + \sum_{i=1}^{q} \theta_{Y,i} X(t-i) + \varepsilon_Y(t) \tag{4}$$

where q represents the number of lagged terms of the other series included in the model, $\theta_{X,i}$ and $\theta_{Y,i}$ are the coefficients representing the influence of the lagged values of search frequency and incidence or mortality, respectively.

The significance of the Granger causality test is evaluated via the corresponding hypothesis testing (Granger, 1969; Hamilton, 2020). More specifically, the null hypothesis (H_0) states that Google Trend frequency does not "Granger cause" the corresponding COVID-19 incidence or mortality, that is, it does not help predict COVID-19 incidence and mortality, implying that the additional information provided by the trend does not improve the prediction of incidence (or mortality) beyond the lagged values of incidence (or mortality) alone. The alternative hypothesis suggests the presence of Granger causality, that is, that the corresponding Google Trend frequency helps predict COVID-19 incidence or mortality. The significance is evaluated using the *F*-test. The *F*-statistic is computed by comparing the squared residuals from the restricted model (without the additional predictors from the second series) to the sum of squared residuals from the extended model (with the additional predictors from the second series). The corresponding test result has been considered statistically significant, that is, the presence of Granger causality, if the corresponding test *p*-value was below $\alpha = 0.05$ to determine the presence of statistically significant Granger causality. A *p*-value below this threshold indicated that search frequency had a significant causal influence on COVID-19 incidence. No multiple testing adjustment was performed since the goal was to include the most predictors possible in the available set of potentially useful ones. This work compares the response series $Y(t)$ with multiple Google Trends series $X(t)$ using pairwise Granger causality tests.

2.2.3. Cross-correlation coefficient
The second approach for selecting a pool of keywords involved calculating cross-correlation coefficient values to assess the correlation between the corresponding time series.

The cross-correlation coefficient is a statistic used to assess the strength of the relationship between two time series for a specific lag value (Box et al., 2015). More precisely, the cross-correlation function at an

integer-valued time lag of τ, denoted as $\rho_{XY}(\tau)$, represents the correlation coefficient between the series $X(t)$ and $Y(t)$. The estimate of the cross-correlation coefficient $\hat{\rho}_{XY}(\tau)$ is calculated as (Box et al., 2015):

$$\hat{\rho}_{XY}(\tau) = \frac{1}{n-|\tau|} \sum_{t=1}^{n-|\tau|} \left[X(t) - \overline{X}(t)\right] \cdot \left[Y(t+\tau) - \overline{Y}(t+\tau)\right] \tag{5}$$

where n is the number of time slots in the original series, $X(t)$ and $Y(t)$ are the values of Google Trend and COVID-19 incidence at time t, τ is the integer time lag, and $\overline{X}(t)$ and $\overline{Y}(t+\tau)$ are the means of the original time series and the version shifted by the value of τ. For a lag value of $\tau > 0$, the number of time slots in the original series used to compute the correlation coefficient decreases by τ, respectively.

In the presented study, the cross-correlation coefficients were computed between pairs of series for different values of the lag τ to investigate the strength of the association between selected Google Trend keywords and COVID-19 incidence and mortality. Multiple lag values τ were considered based on the hypothesis that changes in search frequency might precede or follow corresponding changes in incidence and mortality. Since the correlation values range from -1 to 1, the goal is to focus on those whose trends' absolute values are close to 1, and subsequently use those keywords as predictors.

2.2.4. Data rescaling for comparisons

Since the time series were on different scales (i.e., actual counts for incidence and mortality data versus re-scaled Google Trends values on a scale from 0 to 100), all the series used in the analysis were rescaled and normalized. Suppose exact predictions are needed for practical purposes. In that case, the one-to-one backward transformation can be applied to the data resulting from predictions made by the models fitted to the rescaled and normalized data. In particular, the following normalization was used:

$$Y'(t) = \frac{Y(t)}{\overline{Y}} \tag{6}$$

where $Y'(t)$ is the rescaled value at time point t, $Y(t)$ is the original value at time point t, and $\overline{Y}$ is the average value of that time series across the available time slots. Since both considered time series (incidence or mortality vs. Google Trends) are non-negative, the presented standardization method allows for the scaling of the series without causing any values to become negative.

2.2.5. Principal component analysis

As an alternative to directly selecting keywords and their subsets, the utility of dimensionality reduction for the set of predictors was evaluated. To achieve this, principal component analysis (PCA) was applied to the initially selected pool of keywords, separately for incidence and mortality keywords. The subsets of transformed components that explained 90% or more of the variability in the original keyword sets were used as multiple external predictors in the models (Jolliffe, 2002; Hastie, 2009). Similar to the other models, these selected sets of transformed components were added as external predictors to each considered model type and prediction improvement was evaluated against the same model type without external predictors.

2.3. Prediction models and performance metrics

The processed and standardized data were used for prediction models with and without Google Trends as external predictors. In particular, time series methods were used, including Autoregressive Integrated Moving Average with Exogenous Variables (sometimes referred to as either ARIMAX to emphasize the inclusion of predictors or, just as the generic name ARIMA) and Prophet predictive models. In addition, the eXtreme Gradient Boosting method (XGBoost model), based on decision trees, has been adopted for time series analysis. Although numerous methods are available for estimation and prediction, the three models evaluated

here enable an assessment of predictor robustness across distinct established modeling frameworks, including likelihood-based (ARIMA), Bayesian (Prophet), and machine learning (XGBoost) approaches.

To evaluate and compare the models' prediction performances, they were fitted to specific calibration periods, and predictions were considered for the 2 months following these calibration periods. The predictions for those two months were compared with the actual historical data for those two months, which were not used for model calibration. The predictions were compared with the data using the mean absolute error (MAE). The length of the calibration period gradually increased, starting from 12 months and going up to 34 months (i.e., 36 months of data – 2 months of prediction), where 36 months corresponds to the total number of months in the dataset. The increased calibration period mirrors real-life settings, where the calibration window expands as more data about the ongoing epidemic typically become available. The smallest number of calibration points was specifically selected to be 12 to allow the models to capture the inherent annual seasonality of COVID-19.

More specifically, to generate predictions, the number of data points (i.e., months) for each calibration period gradually increased for each model and state combination. The dataset contained 36 months of data. Initially, a 2-month forecast was made using 12 months of historical data to capture a year of possible seasonality. An additional month of data were incorporated for each subsequent prediction, with the final prediction using 34 data points. This process yielded 23 two-month forecasts for each model and state combination. Prediction accuracy was then assessed by comparing the predicted values to the actual observed data for each length of data used for model fitting. Finally, median values were calculated across all states for each model and calibration length, and summaries were produced.

In summary, the analysis directly incorporates evaluations of the robustness of the presented approach both temporally (using multiple training sets ranging from 12 to 34 months with 1-month increments) and geographically (across all 50 states and the District of Columbia, each with distinct COVID-19 dynamics), leveraging a rich dataset.

Specifically, to address temporal robustness, models have been trained on progressively expanding monthly time series and then used to predict the subsequent months. This approach allowed thorough evaluation of model performance over multiple forecast horizons and different phases of the pandemic, including the early stages, which the authors consider among the most crucial from an epidemiological perspective.

To address geographic robustness, each expanding dataset has been evaluated across all 50 US states to account for geographic heterogeneity and to examine the consistency of trends across regions.

In total, the number of evaluations was 51 (locations) × (34 – 12 + 1) (time intervals) = 1173 for each keyword combination considered in the analysis for each incidence and mortality. These evaluations were repeated for 15 selected keyword combinations, resulting in 1173 × 15 = 17,595 model fits performed separately for incidence and mortality outcomes, as well as for each of the three models considered. For the PCA-transformed predictors, the number of comparisons was also 1173 model fits for each model type separately for incidence and mortality outcomes.

2.4. Autoregressive integrated moving average with exogenous variables (ARIMAX) model

The autoregressive integrated moving average (ARIMA) model is a likelihood-based model for analyzing time series data, used for predictions by incorporating past observations and moving averages. More specifically, the ARIMA model is characterized by three primary components: autoregression (AR), differencing or integration (I), and moving averages (MA) (Shumway and Stoffer, 2017). The ARIMA model is formulated as:

$$y(t) = c + \sum_{i=1}^{p} \phi_i y(t-i) + \sum_{j=1}^{q} \theta_j \epsilon_{t-j} + \epsilon_t. \tag{7}$$

In formula (7), values $y(t)$ denote the series values at time t (possibly after differencing transformations to ensure stationarity, if needed), and ϕ_i are the corresponding autoregressive coefficients, which are estimated. The coefficients θ_j are the moving average parameters, which are also estimated. At the same

time, ϵ_t are the residual errors from the moving average model at time t. The lowercase values $y(t)$ emphasize the potential preprocessing differencing transformations of the original values $Y(t)$, where for $d = 0$ there are no transformations, that is, $y(t) = Y(t)$; for $d = 1$, the differenced observations are $y(t) = Y(t) - Y(t - 1)$, and so on, with higher orders applying the same transformation to the values obtained in the previously applied transformation.

When additional explanatory variables (exogenous variables, which are Google Trends time series in this work) are included in the ARIMA framework, the model is extended to an ARIMAX (autoregressive integrated moving average with exogenous variables) model. The model has the form:

$$y(t) = c + \sum_{i=1}^{p} \phi_i y(t - i) + \sum_{j=1}^{q} \theta_j \epsilon_{t-j} + \sum_{k=1}^{m} \beta_k X_k(t - \tau) + \epsilon_t, \tag{8}$$

where $X_k(t - \tau)$ denotes the value of the k-th predictors series value at time t with the integer lag value $\tau \geq 0$, and β_k for $k = 1, 2, \ldots, m$ is the set of corresponding estimated coefficients for those predictors.

The presented formulation allows the ARIMAX model (8) to account for external predictors, improving the prediction of the studied time series. For this study, the auto.arima function from the R package forecast was used. This function automatically selects the best-fitting parameters of the model based on the Akaike information criterion (AIC). It can incorporate external predictors to fit the ARIMAX model. The numbers of parameters p and q were searched with the maximum number corresponding to the length of the calibration period, which was twelve points. Box–Cox transformations were also used to improve the stability of the model input data.

On top of that, to evaluate the robustness of the ARIMA-type models, the autoregressive fractionally integrated moving average (ARFIMA) model was also considered using the arfima function to assess potential differences in the model results.

2.5 Prophet model

The Prophet model was developed by Meta-Facebook's Core Data Science team for business forecasting (Taylor and Letham, 2018). It is a GAM variant with nonlinear smoothers applied to the regressors. This tool was designed for the user without expert knowledge about time series and their specific methods. It can handle multiple seasonalities with high magnitude, irregular influences like holidays, missing data, outliers, and changes in general trends. While it was created with business tasks in mind, it proved its usefulness in public health research as well (Navratil and Kolkova, 2019; Samal et al., 2019; Satrio et al., 2021; Kirpich et al., 2022; Shishkin et al., 2023; Bleichrodt et al., 2024). Prophet model is formulated as the sum of the following components:

$$Y(t) = g(t) + s(t) + h(t) + \epsilon(t), \tag{9}$$

where $Y(t)$ is the COVID cases or death at time point t, $g(t)$ is the trend component, $s(t)$ is the periodical seasonal component, $h(t)$ is the irregular component (like holidays, for example), and $\epsilon(t)$ is modeling error at time t. Trend component $g(t)$ is a logistic growth model that incorporates change points in which growth is allowed to change. Those change points could be specified by the model user or selected automatically. The periodical seasonal component $s(t)$ is a Fourier series, with parameters selected using AIC procedure. Irregular component $h(t)$ directly adds value at a particular time. The model also has embedded uncertainty estimation.

In this article, Prophet function from the package with the same name for R programming language was used. This implementation can use exogenous variables, so we incorporated Google Trends data in the model. The following model parameters were set: *yearly.seasonality* to 4, and *weekly.seasonality* as well as *daily.seasonality* disabled. Other parameters were not set, so the function used default values described in the package documentation.

2.6. XGBoost model

XGBoost (eXtreme Gradient Boosting) is an implementation of gradient-boosted decision trees (GBDT), widely used for multiple tasks, including classification, regression, and (most relevant for this project) time series forecasting, including epidemiologic forecasting (Chen and Guestrin, 2016; Wang and Guo, 2020; Rahman et al., 2022). The model sequentially builds weak learners, typically decision trees (Gehrke and Ye, 2003), to correct the errors of previous weak learners and to minimize a specified loss function.

In brief, for a set of n records and m predictors per each record represented by the set:

$$\mathcal{D} = \{(\boldsymbol{x}_i, y_i) \text{ where } |\mathcal{D}| = n, \boldsymbol{x}_i \in \mathbb{R}^m, y_i \in \mathbb{R}\}, \tag{10}$$

a tree ensemble model that predicts an output as a sum of K functions has the form:

$$\hat{y}_i = \phi(\boldsymbol{x}_i) = \sum_{k=1}^{K} f_k(\boldsymbol{x}_i), \text{where} f_k \in \mathcal{F}. \tag{11}$$

The set $\mathcal{F} = \{f(x) = w_{q(x)} \text{ where } q : \mathbb{R}^m \to T, w \in \mathbb{R}^T\}$ represents the set of regression trees, where q is the structure of each tree that maps the input x to the corresponding leaf index, and T is the number of leaves in the tree. Each function f_k corresponds to a specific tree structure q and leaf weights w. To identify the set of functions for the model, the following regularized objective function is minimized:

$$\mathcal{L} = \sum_i l(\hat{y}_i, y_i) + \sum_k \Omega(f_k), \tag{12}$$

where $\Omega(f) = \gamma T + \frac{1}{2}\lambda\|w\|^2$ is the regularization term described in detail later, which represents a penalty for the complexity of the model. In equation (12), l represents a differentiable convex function that measures the disagreement between the predictions $\hat{y}_i$ and the data points y_i. It is important to note that the tree model in equation (12) includes functions as parameters and cannot be minimized directly in Euclidean space using traditional optimization methods. Instead, the model is trained in an additive form, using multiple iterations.

More formally, if $\hat{y}_i^{(t)}$ is the prediction of the i-th function of the model at iteration t, this approach called boosting, ensures that each subsequent tree in the ensemble (called a forest) corrects the classification error of the previous tree. This algorithm minimizes the following regularized objective function:

$$\mathcal{L}^{(t)} = \sum_{i=1}^{n} \ell\left(y_i, \hat{y}_i^{(t-1)} + f_t(\boldsymbol{x}_i)\right) + \Omega(f_t), \tag{13}$$

where ℓ is the differentiable convex loss function that represents the difference between the prediction $\hat{y}_i^{(t)}$ and the real data y_i, and $\Omega(f_t)$ is the regularization term defined in the same way as in (12), having the form:

$$\Omega(f_t) = \gamma T + \frac{1}{2}\lambda\|w\|^2 = \gamma T + \frac{1}{2}\lambda\sum_{j=1}^{T} w_j^2. \tag{14}$$

In equation (14), T represents the number of leaves in the tree, while w_j denotes the weight assigned to the j-th leaf. The parameter γ controls the complexity penalty for each leaf, and λ is the L^2 regularization term that helps prevent overfitting by constraining the leaf weights. This objective function promotes the creation of simpler decision trees or, in the context of this model, weaker learners.

For optimization, the XGBoost model utilizes both first-order and second-order gradient information. The model updates the objective function by approximating it with a second-order Taylor expansion:

$$\mathcal{L}^{(t)} \approx \sum_{i=1}^{n} \left[g_i f_t(x_i) + \frac{1}{2} h_i f_t(x_i)^2 \right] + \Omega(f_t), \tag{15}$$

where $g_i = \frac{\partial \ell(y_i, \hat{y}_i)}{\partial \hat{y}_i}$ and $h_i = \frac{\partial^2 \ell(y_i, \hat{y}_i)}{\partial \hat{y}_i^2}$ are the first and second derivatives of the loss function concerning the prediction $\hat{y}_i$.

The XGBoost model is typically used for classification or regression tasks, where each row in the dataset is treated independently of the others. However, time series data have a fundamentally different structure, where the value at any given time point typically depends on values from both earlier and future time points. Therefore, to apply XGBoost to time series forecasting, the original time series must be preprocessed into a new dataset with a specific format that captures these temporal dependencies and makes the XGBoost framework applicable.

In the presented analysis, this has been implemented by creating "lagged features," that is, dataset columns that contain the values of the target variable from previous time points (Sharma, 2024). For example, to predict the number of cases for the given month, the model uses the number of cases from one month before, two months before, and so on, up to a fixed, predefined number of previous months for a given month as input features. Each row in this new dataset then corresponds to one point in time and includes its own set of lagged values of a fixed length, that is, a fixed number of previous points. This transformation allows the model to learn patterns over time, such as trends or cycles, based on recent history for each time point. In the presented model, the lag value was set to six previous time points (months). This choice was based on the available length of the time series, which was limited to 2 years, and the assumption that a typical wave of incidence develops within this timeframe, providing the model with sufficient context to identify meaningful patterns and make accurate forecasts.

The other limitation of decision tree-based models such as XGBoost is that they can only predict values within the range of the training data. This can cause prediction instabilities for nonstationary time series data, that is, those where the average level or variability changes over time, because the model may fail to capture upward or downward trends and instead produce flat (plateaued) predictions. To address this issue and to avoid such instabilities, relative changes in values, that is, how much a value increased or decreased compared to the previous time point, were modeled instead of using the absolute values (Sharma, 2024). This helps to stabilize the data, and such an approach is very similar to how ARIMA-type models utilize input data differencing transformations to ensure the stationarity assumption for the modeled time series. More specifically, for the analyzed data, the model was trained to predict whether the value would increase or decrease and by how much, rather than predict the exact absolute count.

This lag calculation approach was applied to both the target variable (e.g., incidence and mortality increment counts) and the external predictors from Google Trends. In this setup, for each keyword considered as a predictor, six additional columns corresponding to its lagged values were added to the resulting data frame to create the desired set of predictive features.

This study used the R implementation of the XGBoost model from the xgboost package. The training parameter values were: *nrounds* set to 100, *max_depth* set to 10, and *eta* set to 0.1.

3. Results

3.1. *Visual time series comparisons*

The preliminary visual comparison of the standardized incidence and mortality time series with the respective standardized Google Trends series can provide a good starting point for assessing the prediction abilities of the trends. For illustration purposes, the summaries for the four most correlated search keywords ("omicron," "covid test," "covid testing," and "covid treatment") across the three most populous states in the United States (California, Texas, and Florida) are presented in Figure 1 for incidence and in Supplementary Figure S6 for mortality. The results are provided with the corresponding curve

smoothers. In particular, light blue dots show the original incidence data after standardization, while light red dots are the original Google Trends data after standardization, so that the results are visually on the same scale. The corresponding blue line drawn through the light blue dots is the smoothed incidence data using the loess function in R, and, analogously, the corresponding red line drawn through the light dots is the smoothed Google Trends data using the same loess function. The smoothers were provided together with the standardized data to visually inspect and illustrate the similarities in the corresponding trends and patterns between the different data types. In the Figure 1 and Supplementary Figure S6 columns represent search keywords, and rows represent states. The visual summaries were comparable across all three states. The visual summaries indicate more substantial visual alignment for incidence rather than for mortality trends. In particular, one can see that for "covid testing," "covid treatment," and "covid test" smoothed search interest curve is almost identical to the smoothed incidence curve. While "covid omicron" has flat segment of search interest before this variant has emerged. For the mortality rates and search data, the similarity is less pronounced, however, spikes of search interest is still simultaneous to the mortality spikes.

3.2. Keywords selection summaries

The visualizations of the search keywords screening summaries for incidence and mortality, based on the cross-correlation coefficient with no lag (i.e., $\tau = 0$), are summarized in heatmaps in Figure 2 for incidence and in Supplementary Figure S7 for mortality, respectively. The keywords were ranked according to the absolute values of the average cross-correlation coefficients computed across all 50 states and DC, thereby ensuring the inclusion of strong negative correlations. For incidence search keywords, the strongest correlations were observed for "covid test/testing," "covid treatment," "covid," "symptoms covid/covid symptoms," "covid medication," "body aches," "loss of smell," "headache," "omicron variant," and "SARS-CoV-2." It is worth noting that for incidence, the heatmap summaries for Google Trends time series mainly were in agreement and well correlated with the reported incidence time series across all 50 states and DC. For mortality, the strongest correlations were observed for "funeral," "obituary," and "mortuary." It is worth noting that for mortality, the heatmap summaries for Google Trends time series were less in agreement and poorly correlated with the reported mortality time series across the 50 states and DC.

While cross-correlation heatmap summaries provide good visual overviews and allow ranking of search keywords based on the strength of cross-correlation, Granger test *p*-values (with or without adjustments for multiple comparisons) complement the cross-correlation summaries and offer alternative selection procedures. The corresponding Granger test *p*-values are summarized in Supplementary Tables S3 and S4 for incidence and mortality, respectively. The results obtained from Granger's test were in agreement with cross-correlation summaries for incidence, but less so for mortality. If the selection criteria for the set of search keywords rely solely on the standard statistical significance of the test (e.g., the rule of thumb $p < 0.05$), this approach is generally stricter than simply looking at the cross-correlation values above a certain threshold. This is the case even without the multiple testing adjustment for *p*-values to ensure broader screening of keywords with a possibility of more false positives, as used in this work. In practice, cross-correlation can be used together with Granger tests and (potentially other more elaborate) screening tools to identify predictors with the best predictive abilities.

3.3. Models performance

To summarize the performance improvements from incorporating one or more Google Trends predictors into the baseline models, the mean absolute error (MAE) values of models with predictors were compared to those without (baseline models) for each prediction interval. The results were visualized as heatmaps, illustrating increases and decreases in the median MAE across 50 states and DC for each calibration period length, with subsequent 2-month prediction errors summarized by MAE. The baseline model was not included in the heatmaps, as it served as the reference. The rows of the heatmaps corresponded to different
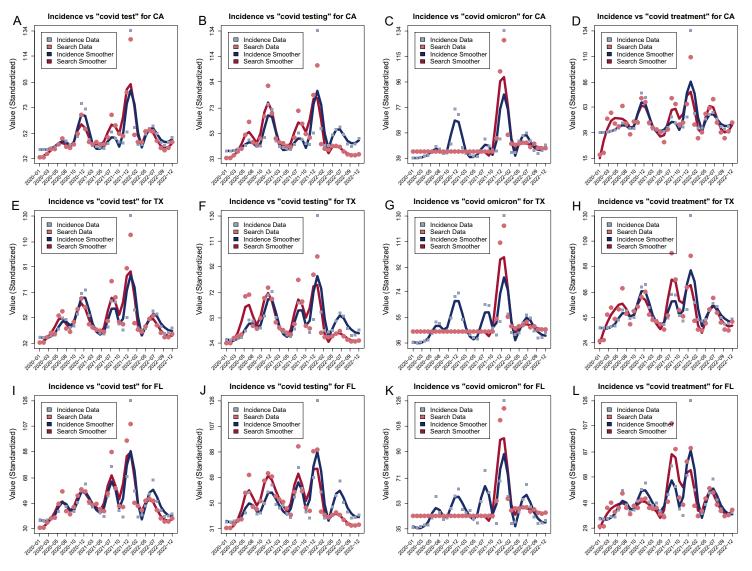
Figure 1. *Monthly incidence for California, Texas, and Florida, along with selected Google Trends time series from January 1, 2020 to December 31, 2022, are presented, together with the corresponding smoothers. The smoothers were produced using the loess smoothing method.*
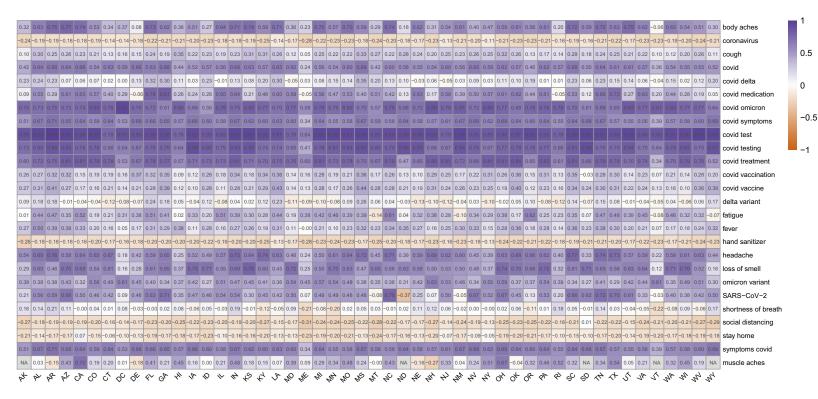
Figure 2. *The heatmap of cross-correlation coefficients between incidence rates and studied search keywords for Google Trends with a τ = 0 month lag (comparing time series for the same months).*

models based on the selected set of external predictors. At the same time, the columns represented the length of the calibration dataset in months. The calibration dataset length ranged from 12 to 34 months, reflecting real-life scenarios.

The first four rows represent the "best" search keywords, ranked according to the corresponding cross-correlation values. In comparison, the following four rows display the "worst" search keywords based on their respective cross-correlation values. The following eight rows are structured similarly, with the first four displaying summaries for models developed by progressively incorporating the "best" search phrases —starting with the top two, followed by the top three, and then the top four. The remaining four rows present summaries for models developed by progressively adding the "worst" search phrases, beginning with the best two of the "worst" four, followed by the best three, and finally, all four of the "worst" phrases. The last row represents summaries for models where transformed PCA components of the complete set of keywords, explaining 90% of the variability, were used as predictors. The average number of selected components across all locations was 2.14 (min: 2, max: 3) for incidence and 2.64 (min: 2, max: 4) for mortality. Additionally, the model's search keywords, which were incorporated as regressions, along with the model's name, are summarized on the right in each item.

For the ARIMA models, the heatmaps are presented in Figure 3 and Supplementary Figure S8, respectively. For the Prophet models, the heatmaps are presented in Figure 4 and Supplementary Figure S9, respectively. For the XGBoost models, the heatmaps are presented in Figure 5 and Supplementary Figure S10, respectively. Overall, the performance improvements for the ARIMA, Prophet, and XGBoost models were consistent across all three models. In particular, better-correlated keywords and their combinations showed consistent improvements in prediction performance. For single predictors, the improvements were most pronounced for the best-correlated search keywords, where the most correlated search keywords resulted in the largest improvements in terms of MAE. The improvements gradually decreased for the second and third most correlated search keywords. ARFIMA, as model alternative to ARIMA, produced results identical to ARIMA.

For incidence, the inclusion of the single worst-correlated search keywords resulted in consistent decreases in performance in terms of MAE. The decrease in performance was most pronounced for the worst-correlated search keywords, while the decreases were comparable among the "worst" search keywords. For incidence, the results for the multiple "best" search keywords mirrored the single "best" search keywords' performance, and similar trends were observed. However, the inclusion of more "best" search keywords for incidence improved the predictions, and the improvement increased as the number of search keywords grew. For incidence, the results for the multiple "worst" search keywords mirrored the performance of the single "worst" search keywords, but the performance was highly inconsistent. In particular, sometimes the performance worsened significantly, while in other scenarios, it improved slightly. Overall, the inclusion of a single search keywords, whether "best" or "worst," had the most dramatic impact on the model's performance. Adding more search keywords continued this trend but with relatively minor changes. The use of principal components as exogenous regressors for each model yielded results that incorporated a combination of the entire set of predictors, including both the "best" and "worst" keywords. In particular, it did not lead to better predictive performance compared to using a single keyword selected via cross-correlation, and in some cases, performance was even worse, as the transformed components indirectly incorporated less informative or poorly performing keywords.

For incidence within each model, the performance generally increased with longer calibration period histories, with the exception of specific periods that the parametric models, such as ARIMA and Prophet could not effectively capture. As a result, the left columns in the table, which correspond to the "worst" search keywords and their combinations, showed the worst performance in relation to the baseline model. In contrast, the right columns, corresponding to the "best" search keywords and their combinations, showed the best performance in relation to the baseline model. While the performance of the ARIMA, Prophet, and XGBoost models for independence predictions was consistent, there were some differences, which are discussed further.

For mortality, the results were comparable but much less pronounced for all ARIMA, Prophet, and XGBoost models, which was expected due to the fact that the observed cross-correlation estimates were
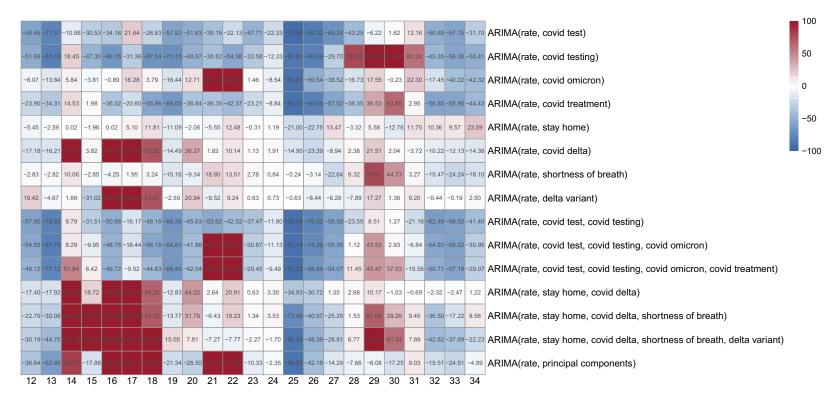
Figure 3. *The percentage improvement in MAE for ARIMA models with Google Trends exogenous regressors compared to the baseline ARIMA model without those regressors applied to incidence data. The columns represent the number of data points used for prediction, and the rows represent the models. The time series for the following search keywords showed the highest correlations: "covid test," "covid testing," "covid omicron," and "covid treatment." Other search keywords had the least correlation coefficients. The first four rows represent the "best" search keywords, ranked by their corresponding cross-correlation values. In contrast, the next four rows display the "worst" search keywords, based on their lower cross-correlation values. The following eight rows follow a similar structure: the first four summarize models built by progressively adding the "best" search phrases, starting with the top two, then the top three, and finally all four. The remaining four rows summarize models developed by similarly adding the "worst" search phrases, beginning with the best two among the bottom four, then the best three, and ultimately all four of the "worst" phrases. The last row represents summaries for models where transformed PCA components of the complete set of keywords, explaining 90% of the variability, were used as predictors.*
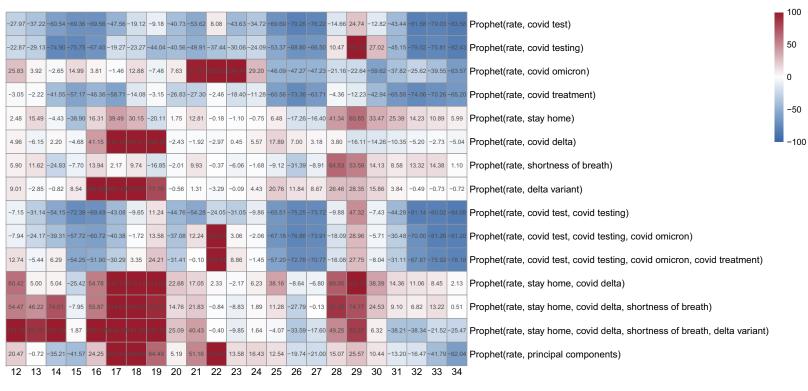
Figure 4. *The percentage improvement in MAE for Prophet models with Google Trends exogenous regressors compared to the baseline Prophet model without those regressors applied to incidence data. The columns represent the number of data points used for prediction, and the rows represent the models. The time series for the following search keywords showed the highest correlations: "covid test," "covid testing," "covid omicron," and "covid treatment." Other search keywords had the least correlation coefficients. The first four rows represent the "best" search keywords, ranked by their corresponding cross-correlation values. In contrast, the next four rows display the "worst" search keywords, based on their lower cross-correlation values. The following eight rows follow a similar structure: the first four summarize models built by progressively adding the "best" search phrases, starting with the top two, then the top three, and finally all four. The remaining four rows summarize models developed by similarly adding the "worst" search phrases, beginning with the best two among the bottom four, then the best three, and ultimately all four of the "worst" phrases. The last row represents summaries for models where transformed PCA components of the complete set of keywords, explaining 90% of the variability, were used as predictors.*
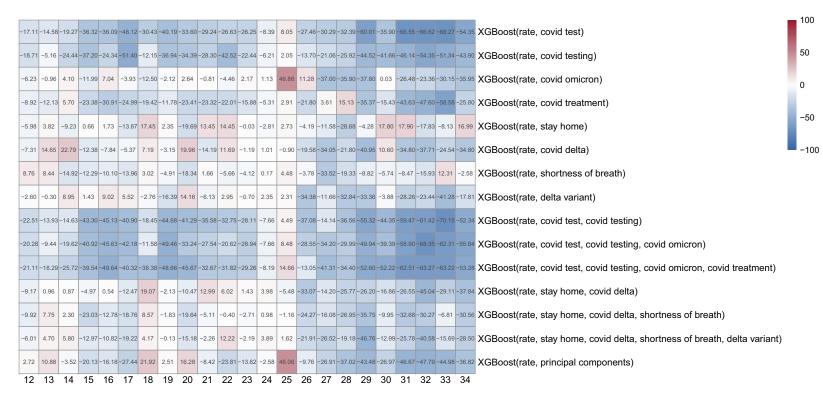
Figure 5. *The percentage improvement in MAE for XGBoost models with Google Trends exogenous regressors compared to the baseline XGBoost model without those regressors applied to incidence data. The columns represent the number of data points used for prediction, and the rows represent the models. The time series for the following search keywords showed the highest correlations: "covid test," "covid testing," "covid omicron," and "covid treatment." Other search keywords had the least correlation coefficients. The first four rows represent the "best" search keywords, ranked by their corresponding cross-correlation values. In contrast, the next four rows display the "worst" search keywords, based on their lower cross-correlation values. The following eight rows follow a similar structure: the first four summarize models built by progressively adding the "best" search phrases, starting with the top two, then the top three, and finally all four. The remaining four rows summarize models developed by similarly adding the "worst" search phrases, beginning with the best two among the bottom four, then the best three, and ultimately all four of the "worst" phrases. The last row represents summaries for models where transformed PCA components of the complete set of keywords, explaining 90% of the variability, were used as predictors.*

substantially smaller, as illustrated in Supplementary Figure S7. While the performance of the ARIMA, Prophet, and XGBoost models for mortality predictions was consistent, there were some differences, which are discussed further.

Specifically for incidence, for the ARIMA model, the top search keywords "covid test" significantly improved the median prediction metric for almost all calibration periods compared to the baseline model. The search keywords "covid omicron," despite having a high cross-correlation coefficient, significantly increased the median prediction error in both the single search keywords predictor model and the multiple search keywords predictor models, even for shorter calibration period lengths. Models with multiple search keywords predictors generally performed similarly to the single search keywords predictor models.

Overall, the metrics for the Prophet models applied to the incidence data were very similar to the ARIMA results. In summary, all Prophet models with highly correlated search keywords, except for the "covid omicron" search keywords, generally showed a decrease in median MAE. However, the barely correlated search phrases did not significantly improve performance and, in some cases, even worsened it.

The Prophet models applied to the mortality data slightly outperformed the ARIMA model, with moderate yet consistent improvements primarily observed for highly correlated key phrases and their combinations. These improvements were only noticeable when the calibration period was 28 weeks or longer. Shorter calibration periods did not yield significant improvements and, in many instances, even led to a decline in performance. Overall, the Prophet model results were more consistent, though still comparable to those of the ARIMA model. In summary, all Prophet models for incidence and mortality with highly correlated exogenous regressors, except for the "covid omicron" regressor, typically showed a decrease in median MAE. Meanwhile, the correlated search phrases did not significantly enhance performance and, in some cases, actually worsened it. Again, these results were anticipated, given that the observed cross-correlation estimates were notably smaller, as shown in Supplementary Figure S7.

Finally, the results for the XGBoost incidence models differed significantly from those of the ARIMA and Prophet incidence models, likely due to the distinct nature of these machine-learning models. The XGBoost incidence models demonstrated the greatest consistency, with a steady decrease in MAE for the highly correlated search keywords and their combinations across most calibration periods. Surprisingly, even small improvements were observed for search keywords that were not strongly correlated.

The results for the XGBoost mortality models also differed significantly from those of the ARIMA and Prophet incidence models, which was unexpected. The XGBoost mortality models showed the greatest inconsistency and poor performance for longer calibration periods. For shorter calibration periods, some improvements in MAE were observed. However, the differences in results for highly correlated search keywords and poorly correlated search keywords were much smaller compared to the differences between various calibration period lengths. This suggests that mortality search keywords were not strong predictors of mortality, likely due to substantially smaller correlation estimates, as shown in Supplementary Figure S7.

4. Discussion

The presented study illustrated how Google Trends data, which can be obtained rapidly, can enhance COVID-19 incidence predictions, especially in situations where such predictions are needed quickly. Our findings indicate that in cases of data reporting delays, and provided that appropriate (i.e., highly correlated) search keywords can be identified, incorporating Google Trends data for these phrases can significantly improve predictions compared to models that rely solely on incidence trends. The results were also mostly in agreement across the three very different models analyzed in this study. This study aims to establish the best practices for utilizing Google Trends data in public health research. To facilitate this, the authors made the analysis code in R publicly available and developed and published an RMarkdown notebook on GitHub (GitHub repository for the Google Trends Study, 2025), which can be easily adapted in other research projects.

While incidence forecasting models demonstrated substantial improvement with the inclusion of search data, the enhancement observed for mortality prediction was notably weaker. Several factors may

contribute to this discrepancy. First, search data related to mortality are likely to be noisier because the number of deaths is much lower than the number of infections, resulting in a weaker signal. In addition, mortality-related search terms may be influenced by external factors unrelated to actual deaths. For instance, searches for terms like "coffin" can spike around cultural events such as Halloween, potentially introducing noise into the data.

Another plausible explanation is behavioral. In the United States and similar settings, families typically receive direct guidance from healthcare providers, coroners, or funeral service companies shortly after a death. These services often initiate contact or are recommended through institutional channels, which reduces the need for individuals to search online for postmortem procedures. As a result, mortality-related search activity may be both less common and more contextually variable, limiting its utility as a predictor in forecasting models.

To improve prediction performance, as illustrated, one must be able to identify appropriate search keywords that closely correlate with disease data. For example, during the COVID-19 pandemic, search keywords reflecting immediate public responses to illness, such as "COVID test" or "COVID treatment," were shown to be the most effective. In general, automated procedures for identifying the most correlated queries can be utilized (Mohebbi et al., 2011).

We note that the screening methods used, such as cross-correlation and Granger causality analysis, have well-known limitations, especially in capturing multivariate dependencies. These techniques are inherently constrained in their capacity to infer true causal relationships, especially in the presence of potential confounders or complex interdependencies among variables. In the presented analysis, however, this methodological choice was intentional. Our goal was not to model or disentangle complex multivariate structures, nor to control for confounding factors. Rather, the aim was to prove the principle, that is, to illustrate that even simple screening tools, when applied to a small and conveniently selected subset of keywords, can yield insights potentially useful for epidemic prediction.

This study does not attempt to establish a direct causal relationship between search interest in specific topics and COVID-19 incidence or mortality. Any observed improvement, decline, or lack of change in model performance following the inclusion of particular search terms should not be interpreted as evidence of causality. Search trends can be shaped by a variety of external influences, including media coverage, social media discussions, public health campaigns, and other contemporaneous events. We evaluated search terms related to symptoms, vaccine hesitancy, and side effects for their collective predictive value, without attempting to draw conclusions about the behavioral motivations behind them. All terms were treated as candidate predictors within a consistent modeling framework. Other studies have already addressed behavioral aspects of the pandemic, such as perceptions of vaccine safety and the drivers of hesitancy, which were beyond the scope of this analysis (Moffett et al., 2024).

We emphasize that the automatic identification of predictors solely relying on cross-correlation coefficients or the Granger causality test to identify relevant search keywords may not always be reliable or ideal. This was demonstrated in this study through the use of the search keywords "covid omicron." Despite showing a strong cross-correlation value and a significant Granger causality test *p*-value, this term worsened prediction accuracy. This occurred because the search keywords emerged much later in the pandemic than others, and its Google Trends pattern differed from the incidence trend during the first half of the period. As a result, it is crucial to execute automatic search keywords selection using techniques like cross-correlation with caution and to complement such methods with external knowledge about the search keywords, along with possible additional checks, to ensure that the selected search keywords align well with the incidence data.

Regarding the number of external predictors, our results suggest that using a single search keyword as an exogenous predictor variable led to comparable improvements in prediction accuracy as combining multiple predictor terms. As a result, there is no indication that using multiple predictors significantly improved performance over a single predictor. This aligns with the importance of selecting appropriate predictors for achieving substantial and accurate improvements.

For model selection, we observed that all three modeling approaches, though different in structure and underlying assumptions, showed improved performance when relevant Google Trends search terms were included. In contrast, the inclusion of less appropriate or unrelated search keywords generally led to a deterioration in predictive accuracy across most models and scenarios. The use of PCA for dimensionality reduction and using the resulting components did not lead to better predictive performance compared to using a single keyword selected via cross-correlation, and in some cases, performance was even worse, as the transformed components indirectly incorporated less informative or poorly performing keywords. Moreover, the PCA-derived components were significantly more difficult to interpret.

As an illustration, our team proposes the following algorithm for incidence prediction based on the study results. First, researchers should identify search phrases that are unique and informative about the disease of interest. Most likely, these would be a colloquial disease name, symptoms, or a combination of the disease name and the word "test." The next step would be acquiring Google Trends data for those search keywords. Researchers should then visually inspect these time series (or at least a subset if there are many) and calculate the cross-correlation coefficient against their incidence data, selecting the most strongly correlated search keywords. The incidence and Google Trends time series should be scaled using *z*-score transformation or by dividing each series by its mean. Finally, the researcher can predict incidence using the search frequency time series as an exogenous regressor and scale the predictions back. After predicting the reporting gap up to the last available Google Trends data, the researcher can use a selected model to make predictions further into the future, provided the Google Trends data are available for those dates.

This study has several limitations. First, Google does not fully disclose the sampling methodology or the sample size, and the mechanism behind generating Google Trends data remains proprietary. The analysis primarily focused on the frequency of search keywords in English and within the United States. While the approach is expected to be broadly applicable to other languages and regions, the results have not been verified for contexts outside of this scope, and different patterns may emerge in other languages or geographical areas.

Another limitation is that Google Trends data must be available for the time slots when predictions are made. This is not a significant issue, as COVID-19 reporting often lags behind or temporarily stops during periods like weekends or holidays. Thus, the method proposed in this study can effectively fill these gaps. Additionally, while this study focused on COVID-19 data, we believe the approach presented here is broadly applicable and can be used for other diseases that experience reporting delays or even beyond.

One more limitation stems from the temporal and spatial resolution of the data used in this study. The authors utilized monthly data resolution for all time series, which was dictated by the resolution available for Google Trends data over 2020–2023. More frequent Google Trends data are available for shorter periods. However, only the most recent periods were selected for analysis. The authors hypothesize that using higher resolution data could reveal more significant fluctuations, but may also introduce more noise. Regarding spatial resolution, the smallest and most consistent reporting unit for Google Trends data in the United States is at the state level. No county-level data are available, and only major cities are reported for units smaller than a state. However, the authors do not expect these limitations to affect the overall results substantially.

In that regard, another limitation of this study is its exclusive focus on Google Trends search data. It is worth noting that more specialized Google datasets exist, such as the COVID-19 Vaccination Search Insights database (Google COVID-19 Vaccination Search Insights, 2025; Boulanger et al., 2021). While our team briefly considered this resource, the study was deliberately designed to span a longer temporal interval, beginning in early 2020. Since COVID-19 vaccines only became available in late 2020 and early 2021, the Vaccination Search Insights dataset, which begins in January 2021, does not include the critical early phases of the pandemic. Incorporating it would have required omitting a substantial portion of our study period.

A few words should be said about the choice of the models for the presented study. The set of selected models is by no means comprehensive. For example, there are multiple variations of the ARIMA model alone, such as ARFIMA (Kraft et al., 2015; Veenstra et al., 2015), which the authors considered briefly in the presented analysis and which provided very close results to the classic ARIMA the authors eventually ended

up using. The goal was to pick three models from very different modeling paradigms, such as a likelihood-based model (ARIMA), a Bayesian model (Prophet), and a machine learning model (XGBoost), to prove the utility of Google Trends and illustrate the approach across different modeling frameworks. Other models can be explored and adopted by the scientific community in application to Google Trends if needed.

The authors believe that future research should focus on the following areas. First, examining Google Trends correlations in other countries and languages could provide valuable insights into search phrase selection techniques. Second, evaluating the performance of various forecasting models that can incorporate exogenous regressors may lead to even better predictive models for this task. Third, other predictors such as wastewater data could also enhance COVID-19 forecasting models. Future research should explore how combining these with Google Trends data can further improve predictive performance (Rankin et al., 2025). Finally, applying these methods to study other diseases would help generalize this approach for future outbreaks. This would enable public health professionals to identify the most effective phrases and models for monitoring outbreaks globally.

References

Albani V, **Loria J**, **Massad E and Zubelli J** (2021) COVID-19 underreporting and its impact on vaccination strategies. *BMC Infectious Diseases 21*, 1–13.

Amusa LB, **Twinomurinzi H and Okonkwo CW** (2022) Modeling COVID-19 incidence with Google Trends. *Frontiers in Research Metrics and Analytics 7*, 1003972.

Ayyoubzadeh SM, **Ayyoubzadeh SM**, **Zahedi H**, **Ahmadi M and Kalhori SRN** (2020) Predicting COVID-19 incidence through analysis of Google Trends data in Iran: data mining and deep learning pilot study. *JMIR Public Health and Surveillance 6* (2), e18828.

Balashov VS, **Yan Y and Zhu X** (2021) Using the Newcomb–Benford law to study the association between a country's COVID-19 reporting accuracy and its development. *Scientific Reports 11* (1), 22914.

Bleichrodt A, **Luo R**, **Kirpich A and Chowell G** (2024) Evaluating the forecasting performance of ensemble sub-epidemic frameworks and other time series models for the 2022–2023 mpox epidemic. *Royal Society Open Science 11* (7), 240248.

Boulanger A, **Kumok A**, **Patankar A**, **Ghazi B**, **Miller B**, **Kamath C**, **Stanton C**, **Scott C**, **Desfontaines D**, **Kraft D**, **Gabrilovich E**, **Wellenius GA**, **Gupta JP**, **Davis JS**, **Smith KL**, **Gadepalli KK**, **Young M,Perera M**, **Sun M**, **Manurangsi P**, **Kumar R**, **Bavadekar S**, **Griffith T**, **Shekel T and Mayer Y** (2021) *Google COVID-19 Vaccination Search Insights: Anonymization Process Description.* Preprint on arXiv. Available at https://arxiv.org/abs/2107.01179 (accessed 27 June 2025).

Box GE, **Jenkins GM**, **Reinsel GC and Ljung GM** (2015) *Time Series Analysis: Forecasting and Control*. USA: John Wiley & Sons.

Butler D (2013) When Google got flu wrong: US outbreak foxes a leading web-based method for tracking seasonal flu. *Nature 494* (7436), 155–157.

Cervellin G, **Comelli I and Lippi G** (2017) Is Google Trends a reliable tool for digital epidemiology? Insights from different clinical settings. *Journal of Epidemiology and Global Health 7* (3), 185–189.

Chen J, **Mi H**, **Fu J**, **Zheng H**, **Zhao H**, **Yuan R**, **Guo H**, **Zhu K**, **Zhang Y**, **Lyu H**, **Zhang Y**, **She N and Ren X** (2022) Construction and validation of a COVID-19 pandemic trend forecast model based on Google Trends data for smell and taste loss. *Frontiers in Public Health 10*, 1025658.

Chen T and Guestrin C (2016) "XGBoost: A scalable tree boosting system". In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, New York, NY: ACM, pp. 785–794.

Chu AM, **Tsang JT**, **Chan SS**, **Chan LS and So MK** (2025) Utilizing Google Trends data to enhance forecasts and monitor long COVID prevalence. *Communications Medicine 5* (1), 1–16.

Ciaffi J, **Meliconi R**, **Landini MP and Ursini F** (2020) Google Trends and COVID-19 in Italy: could we brace for impact? *Internal and Emergency Medicine 15*, 1555–1559.

Cinarka H, **Uysal MA**, **Cifter A**, **Niksarlioglu EY and Çarkoğlu A** (2021) The relationship between Google search interest for pulmonary symptoms and COVID-19 cases using dynamic conditional correlation analysis. *Scientific Reports 11* (1), 14387.

Coronavirus Disease 2019 (COVID-19) (2020) *2020 Interim Case Definition, Approved April 5, 2020 | CDC*. Available at: https://ndc.services.cdc.gov/case-definitions/coronavirus-disease-2019-2020/ (accessed 20 April 2023).

CSSEGISandData (2025) - *COVID-19 Data Repository by the Center for Systems Science and Engineering (CSSE) at Johns Hopkins University*. Available at: https://github.com/CSSEGISandData/COVID-19 (accessed 20 April 2023).

FAQ about Google Trends Data (2025) *FAQ about Google Trends data - Trends Help*. Available at: https://support.google.com/trends/answer/4365533?hl=en&ref_topic=6248052 (accessed 11 April 2022).

Gehrke J and Ye N (2003) Decision tree. In *The Handbook of Data Mining*. Boca Raton, FL: CRC Press Publisher, pp. 3–23.

Google Trends *Google Trends*. Available at: https://trends.google.com/trends/ (accessed 20 April 2023).

Google Trends (2025) - *Google Trends - US*. Available at: https://trends.google.com/trends/?geo=US (accessed 07 November 2022).

Google Trends Scraper (2025) - *Scraper for extracting data from Google Trends*. Available at: https://github.com/emastra/actor-google-trends-scraper (accessed 20 April 2025).

Google Trends Study (2025) - *GitHub repository for the Google Trends Study*. Available at: https://github.com/keder/google_trends (accessed 20 April 2025).

Granger CW (1969) Investigating causal relations by econometric models and cross-spectral methods. *Econometrica: Journal of the Econometric Society 37*, 424–438.

Hamilton JD (2020) *Time Series Analysis*. Princeton university press.

Hanaei S, **Mohebi F**, **Moradi-Lakeh M**, **Jabbari P**, **Mehta SK**, **Kryvenko LS**, **Luongo L**, **Dupré L and Rezaei N** (2021) The epidemiologic aspects of COVID-19 outbreak: spreading beyond expectations. In: *Coronavirus Disease-COVID-19*. Springer, pp. 61–79.

Hastie T (2009) *The Elements of Statistical Learning: data Mining, Inference, and Prediction*.

Hoffman SJ and Silverberg SL (2018) Delays in global disease outbreak responses: lessons from H1N1, Ebola, and zika. *American Journal of Public Health 108* (3), 329–333.

Jolliffe IT (2002) *Principal Component Analysis*. Springer.

Kirpich A, **Shishkin A**, **Weppelmann TA**, **Tchernov AP**, **Skums P and Gankin Y** (2022) Excess mortality in Belarus during the COVID-19 pandemic as the case study of a country with limited non-pharmaceutical interventions and limited reporting. *Scientific Reports 12*(1), 5475.

Kraft P, **Weber C, and Lebo M** (2015, January) *The ArfimaMLM Package for R*. Available at: https://cran.r-project.org/web/packages/ArfimaMLM/ArfimaMLM.pdf.

Kurian SJ, **Bhatti AR**, **Alvi MA**, **Ting HH**, **Storlie C**, **Wilson PM**, **Shah ND**, **Liu H and Bydon M** (2020) Correlations between COVID-19 cases and Google Trends Data in the United States: A state-by-state analysis. *Mayo Clinic Proceedings*. Vol. *95*. New York, NY: Elsevier, pp. 2370–2381.

Lampos V, **Majumder MS**, **Yom-Tov E**, **Edelstein M**, **Moura S**, **Hamada Y**, **Rangaka MX**, **McKendry RA and Cox IJ** (2021) Tracking COVID-19 using online search. *npj Digital Medicine 4* (1), 17.

Lau H, **Khosrawipour T**, **Kocbach P**, **Ichii H**, **Bania J and Khosrawipour V** (2021) Evaluating the massive underreporting and undertesting of COVID-19 cases in multiple global epicenters. *Pulmonology 27* (2), 110–115.

Limilia P and Pratamawaty BB (2020) Google Trends and information seeking trend of COVID-19 in Indonesia. *Jurnal Aspikom 5* (2), 188–205.

Liu K, **Wang T**, **Yang Z**, **Huang X**, **Milinovich GJ**, **Lu Y**, **Jing Q**, **Xia Y**, **Zhao Z**, **Yang Y**, **Tong S**, **Hu W and Lu J** (2016) Using Baidu search index to predict dengue outbreak in China. *Scientific Reports 6* (1), 38040.

Ma S and Yang S (2022) Covid-19 forecasts using internet search information in the United States. *Scientific Reports 12* (1), 11539.

Marty R, **Ramos-Maqueda M**, **Khan N and Reichert A** (2023) The evolution of the COVID-19 pandemic through the lens of Google searches. *Scientific Reports 13* (1), 19843.

Mavragani A and Gkillas K (2020) COVID-19 predictability in the United States using Google Trends time series. *Scientific Reports 10* (1), 20693.

Mavragani A, **Ochoa G** (2019) Google trends in infodemiology and infoveillance: methodology framework. *JMIR Public Health and Surveillance 5* (2), e13439.

Ming WK, **Huang F**, **Chen Q**, **Liang B**, **Jiao A**, **Liu T**, **Wu H**, **Akinwunmi B**, **Li J**, **Liu G**, **Zhang CJP**, **Huang J and Liu Q** (2021) Understanding health communication through Google Trends and news coverage for COVID-19: multinational study in eight countries. *JMIR Public Health and Surveillance 7* (12), e26644.

Moffett KW, **Marshall MC**, **Kim JEC**, **Dahlen H**, **Denison B**, **Kranzler EC**, **Meaney M**, **Hoffman B**, **Pavisic I and Hoffman L** (2024) Analyzing Google COVID-19 vaccine intent search trends and vaccine readiness in the United States: Panel data study. *Online Journal of Public Health Informatics 16* (1), e55422.

Mohebbi M, **Vanderkam D**, **Kodysh J**, **Schonberger R**, **Choi H and Kumar S** (2011) *Google Correlate Whitepaper.* https://static.googleusercontent.com/media/research.google.com/en//pubs/archive/41695.pdf. Draft Date: June 9, 2011.

Navratil M and Kolkova A (2019) Decomposition and forecasting time series in the business economy using prophet forecasting model. *Central European Business Review 8*(4), 26.

Nghiem LT, **Papworth SK**, **Lim FK and Carrasco LR** (2016) Analysis of the capacity of Google Trends to measure interest in conservation topics and the role of online news. *PLoS One 11* (3), e0152802.

Nguyen HL, **Pan Z**, **Abu-gellban H**, **Jin F and Zhang Y** (2020) *Google Trends Analysis of COVID-19 arXiv preprint arXiv: 2011.03847.*

Nia ZM, **Seyyed-Kalantari L**, **Goitom M**, **Mellado B**, **Ahmadi A**, **Asgary A**, **Orbinski J**, **Wu J and Kong J** (2025) Leveraging deep-learning and unconventional data for real-time surveillance, forecasting, and early warning of respiratory pathogens outbreak. *Artificial Intelligence in Medicine 103*, 103076.

Nuti SV, **Wayda B**, **Ranasinghe I**, **Wang S**, **Dreyer RP**, **Chen SI and Murugiah K** (2014) The use of Google trends in health care research: a systematic review. *PLoS One 9*(10), e109583.

Osmani F (2020) Problems with the accurate reporting of COVID-19 statistics in Iran. *Gastroenterology and hepatology from bed to bench 13* (4), 275.

Prasanth S, **Singh U**, **Kumar A**, **Tikkiwal VA and Chong PH** (2021) Forecasting spread of COVID-19 using Google Trends: A hybrid GWO-deep learning approach. *Chaos, Solitons & Fractals 142*, 110336.

Rahman MS, **Chowdhury AH and Amrin M** (2022) Accuracy comparison of ARIMA and XGBoost forecasting models in predicting the incidence of COVID-19 in Bangladesh. *PLOS Global Public Health 2*(5), e0000495.

Rankin N, **Saiyed S**, **Du H and Gardner LM** (2025) A multi-city COVID-19 categorical forecasting model utilizing wastewater-based epidemiology. *Science of the Total Environment 960*, 178172.

Reis RF, **Melo Quintela B de**, **Oliveira Campos J de**, **Gomes JM**, **Rocha BM**, **Lobosco M and Dos Santos RW** (2020) Characterization of the COVID-19 pandemic and the impact of uncertainties, mitigation strategies, and underreporting of cases in South Korea, Italy, and Brazil. *Chaos, Solitons & Fractals 136*, 109888.

Rodríguez A, **Kamarthi H**, **Agarwal P**, **Ho J**, **Patel M**, **Sapre S and Prakash BA** (2024) Machine learning for data-centric epidemic forecasting. *Nature Machine Intelligence 6* (10), 1122–1131.

Rogers S (2016, July 1) *What Is Google Trends Data – and What Does it Mean?*. Available at: https://medium.com/google-news-lab/what-is-google-trends-data-and-what-does-itmean-b48f07342ee8 (accessed 11 April 2022).

Rovetta A (2021) Reliability of Google Trends: analysis of the limits and potential of web infoveillance during COVID-19 pandemic and for future research. *Frontiers in Research Metrics and Analytics 6*, 670226.

Samal KKR, **Babu KS**, **Das SK and Acharaya A** (2019) Time series based air pollution forecasting using SARIMA and prophet model. In: *Proceedings of the 2019 International Conference on Information Technology and Computer Communications*. Singapore: ACM, pp. 80–85.

Satpathy P, **Kumar S and Prasad P** (2023) Suitability of Google Trends™ for digital surveillance during ongoing COVID-19 epidemic: a case study from India. *Disaster Medicine and Public Health Preparedness 17*, e28.

Satrio CBA, **Darmawan W**, **Nadia BU and Hanafiah N** (2021) Time series analysis and forecasting of coronavirus disease in Indonesia using ARIMA model and PROPHET. *Procedia Computer Science 179*, 524–532.

Sharma N (2024) *How to Use XGBoost for Time-Series Forecasting?* Available at: https://www.analyticsvidhya.com/blog/2024/01/xgboost-for-time-series-forecasting/ (accessed 8 April 2024).

Shishkin A (2025) *keder/google_trends: Publication.* https://doi.org/10.5281/zenodo.17162598.

Shishkin A, **Lhewa P**, **Yang C**, **Gankin Y**, **Chowell G**, **Norris M**, **Skums P and Kirpich A** (2023) Excess mortality in Ukraine during the course of COVID-19 pandemic in 2020–2021. *Scientific Reports 13* (1), 6917.

Shumway RH and Stoffer DS (2017) ARIMA models. In: *Time Series Analysis and its Applications*. Springer, pp. 75–163.

Sousa-Pinto B, **Anto A**, **Czarlewski W**, **Anto JM**, **Fonseca JA and Bousquet J** (2020) Assessment of the impact of media coverage on COVID-19–related google trends data: Infodemiology study. *Journal of Medical Internet Research 22*(8), e19611.

Springer S, **Menzel LM and Zieger M** (2020) Google Trends provides a tool to monitor population concerns and information needs during COVID-19 pandemic. *Brain, Behavior, and Immunity 87*, 109.

Sulyok M, **Ferenci T and Walker M** (2021) Google Trends data and COVID-19 in Europe: Correlations and model enhancement are European wide. *Transboundary and Emerging Diseases 68* (4), 2610–2615.

Taylor SJ and Letham B (2018) Forecasting at scale. *The American Statistician 72* (1), 37–45.

Vaccination Search Insights (2025) *Google COVID-19 Vaccination Search Insights.* Available at: https://googleresearch.github.io/vaccination-search-insights (accessed 27 June 2025).

Veenstra JQJ, **McLeod AI and Veenstra MJQ** (2015) *Package 'Arfima'.*

Walker A, **Hopkins C, and Surda P** (2020) Use of Google Trends to investigate loss-of-smell–related searches during the COVID-19 outbreak. In: *International Forum of Allergy & Rhinology*, Vol. *10*. Hoboken, NJ: Wiley Online Library, pp. 839–847.

Wang T, *et al.* (2022) COVID-19 hospitalizations forecasts using internet search data. *Scientific Reports 12*(1), 9661.

Wang Y and Guo Y (2020) Forecasting method of stock market volatility in time series data based on mixed model of ARIMA and XGBoost. *China Communications 17*(3), 205–221.

WHO Coronavirus (2025) *WHO Coronavirus (COVID-19) Dashboard*. Available at: https://covid19.who.int (accessed 29 November 2022).

Zayed BA, **Talaia AM**, **Gaaboobah MA**, **Amer SM and Mansour FR** (2023) Google Trends as a predictive tool in the era of COVID-19: a scoping review. *Postgraduate Medical Journal 99*(1175), 962–975.