# Forecasting the COVID-19 epidemic integrating symptom search behavior: an infoveillance study

Alessandro Rabiolo, Eugenio Alladio, Esteban Morales, Andrew Ian McNaught, Francesco Bandello, Abdelmonem A Afifi, Alessandro Marchese

Submitted to: Journal of Medical Internet Research
on: March 17, 2021

# *Table of Contents*

# Forecasting the COVID-19 epidemic integrating symptom search behavior: an infoveillance study

Alessandro Rabiolo[1, 2] MD, FEBO; Eugenio Alladio[3] PhD; Esteban Morales[4] MSc; Andrew Ian McNaught[1, 5] MD, FRCOphth; Francesco Bandello[6] MD, FEBO; Abdelmonem A Afifi[7] PhD; Alessandro Marchese[6] MD, FEBO

[1]Department of Ophthalmology Gloucestershire Hospitals NHS Foundation Trust Cheltenham GB
[2]Department of Chemistry University of Turin Turin IT
[3]Jules Stein Eye Institute David Geffen School of Medicine University of California Los Angles (UCLA) Los Angeles US
[4]School of Health Professions Faculty of Health University of Plymouth Plymouth GB
[5]Department of Ophthalmology Vita-Salute University, IRCCS Ospedale San Raffaele Scientific Institute Milan IT
[6]Department of Biostatistics Fielding School of Public Health University of California Los Angeles (UCLA) Los Angeles US

**Corresponding Author:**
Alessandro Rabiolo MD, FEBO

## *Abstract*

**Background:** Previous studies have suggested associations between trends of web searches and COVID-19 traditional metrics. It remains unclear whether models incorporating trends of digital searches lead to better predictions.

**Objective:** To investigated the relationship between Google Trends searches of symptoms associated with COVID-19 and confirmed COVID-19 cases and deaths. To develop predictive models to forecast COVID-19 epidemic based on the combination of Google Trends searches of symptoms and conventional COVID-19 metrics.

**Methods:** An open-access web application was developed to evaluate Google Trends and traditional COVID-19 metrics via an interactive framework based on principal components analysis (PCA) and time series modelling. The app facilitates the analysis of symptom search behavior associated with COVID-19 disease in 188 countries. In this study, we selected data of eight countries as case studies to represent all continents. PCA was used to perform data dimensionality reduction, and three different time series models (Error Trend Seasonality, Autoregressive integrated moving average, and feed-forward neural network autoregression) were used to predict COVID-19 metrics in the upcoming 14 days. The models were compared in terms of prediction ability using the root-mean-square error (RMSE) of the first principal component (PC1). Predictive ability of models generated with both Google Trends data and conventional COVID-19 metrics were compared with those fitted with conventional COVID-19 metrics only.

**Results:** The degree of correlation and the best time-lag varied as a function of the selected country and topic searched; in general, the optimal time-lag was within 15 days. Overall, predictions of PC1 based on both searched termed and COVID-19 traditional metrics performed better than those not including Google searches (median [IQR]: 1.43 [0.74-2.36] vs. 1.78 [0.95-2.88], respectively), but the improvement in prediction varied as a function of the selected country and timeframe. The best model varied as a function of country, time range, and period of time selected. Models based on a 7-day moving average led to considerably smaller RMSE values as opposed to those calculated with raw data (median [IQR]: 0.74 [0.47-1.22] vs. 2.15 [1.55-3.89], respectively).

**Conclusions:** The inclusion of digital online searches in statistical models may improve the nowcasting and forecasting of COVID-19 epidemic, and could be used as one of the surveillance systems of COVID-19 disease. We provide a free web-application operating with nearly real-time data that can be used by any user to make predictions of outbreaks, improve estimates of dynamics of ongoing epidemics, and anticipate future or rebound waves.

(JMIR Preprints 17/03/2021:28876)
DOI: https://doi.org/10.2196/preprints.28876

## Preprint Settings

1) Would you like to publish your submitted manuscript as preprint?
Please make my preprint PDF available to anyone at any time (recommended).

Please make my preprint PDF available only to logged-in users; I understand that my title and abstract will remain visible to all users.
Only make the preprint title and abstract visible.

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

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

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

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

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

# Original Manuscript

**- ORIGINAL PAPER -**

# Forecasting the COVID-19 epidemic integrating symptom search behavior: an infoveillance study

**Authors:** Alessandro Rabiolo, MD*[1,2]; Eugenio Alladio, PhD*[3]; Esteban Morales, MSc[4]; Andrew I McNaught, MD[1,5]; Francesco Bandello, MD[2]; Abdelmonem A Afifi, PhD[6]; Alessandro Marchese, MD.[2]

**Authors' affiliation:**

(1) Department of Ophthalmology, Gloucestershire Hospitals NHS Foundation Trust, Cheltenham, UK

(2) Department of Ophthalmology, Vita-Salute University, IRCCS Ospedale San Raffaele Scientific Institute, Milan, Italy

(3) Department of Chemistry, University of Turin, Turin, Italy

(4) Jules Stein Eye Institute, David Geffen School of Medicine, UCLA, Los Angeles, USA

(5) School of Health Professions (Faculty of Health), Plymouth, UK

(6) Department of Biostatistics, Fielding School of Public Health, UCLA, Los Angeles, USA

* Contributed equally

**Corresponding author:** Alessandro Rabiolo, Department of Ophthalmology, Gloucestershire Hospitals NHS Foundation Trust, Sandford Rd, Cheltenham GL53 7AN, United Kingdom; rabiolo.alessandro@gmail.com; telephone +44300 422 2527

## ABSTRACT

Background: Previous studies have suggested associations between trends of web searches and COVID-19 traditional metrics. It remains unclear whether models incorporating trends of digital searches lead to better predictions.

Objective: To investigate the relationship between Google Trends searches of symptoms associated with COVID-19 and confirmed COVID-19 cases and deaths. To develop predictive models to forecast the COVID-19 epidemic based on the combination of Google Trends searches of symptoms and conventional COVID-19 metrics.

Methods: An open-access web application was developed to evaluate Google Trends and traditional COVID-19 metrics via an interactive framework based on principal components analysis (PCA) and time series modeling. The app facilitates the analysis of symptom search behavior associated with COVID-19 disease in 188 countries. In this study, we selected data of eight countries as case studies to represent all continents. PCA was used to perform data dimensionality reduction, and three different time series models (Error Trend Seasonality, Autoregressive integrated moving average, and feed-forward neural network autoregression) were used to predict COVID-19 metrics in the upcoming 14 days. The models were compared in terms of prediction ability using the root-mean-square error (RMSE) of the first principal component (PC1). The predictive abilities of models generated with both Google Trends data and conventional COVID-19 metrics were compared with those fitted with conventional COVID-19 metrics only.

Results: The degree of correlation and the best time-lag varied as a function of the selected country and topic searched; in general, the optimal time lag was within 15 days. Overall, predictions of PC1 based on both searched termed and COVID-19 traditional metrics performed better than those not including Google searches (median [IQR]: 1.56 [0.90-2.49]

vs. 1.87 [1.09-2.95], respectively), but the improvement in prediction varied as a function of the selected country and timeframe. The best model varied as a function of country, time range, and period of time selected. Models based on a 7-day moving average led to considerably smaller RMSE values as opposed to those calculated with raw data (median [IQR]: 0.90 [0.50-1.53] vs. 2.27 [1.62-3.74], respectively).

Conclusions: The inclusion of digital online searches in statistical models may improve the nowcasting and forecasting of the COVID-19 epidemic and could be used as one of the surveillance systems of COVID-19 disease. We provide a free web application operating with nearly real-time data that anyone can use to make predictions of outbreaks, improve estimates of dynamics of ongoing epidemics, and anticipate future or rebound waves.

**Keywords:** Google Trends; Symptoms; Coronavirus; SARS-CoV-2; Big Data; Time Series; Predictive Models; Shiny Web-Application; Infodemiology; Infoveillance.

## INTRODUCTION

COVID-19 is a new entity, and the dynamics of its propagation are difficult to predict. In the absence of compelling evidence, health and political decisions have been strongly driven by a wide variety of statistical models and simulation scenarios to forecast the COVID-19 epidemic. Still, large variations exist among the different models with respect to the predicted number of infected people, time to reach a peak of new cases, course of the epidemic, and identification of outbreaks.[1] One key limitation of such models is that they rely heavily on the number of confirmed infected subjects who usually seek medical attention due to moderate to severe symptoms. However, confirmed cases are most likely only a small proportion of the true number of cases as the vast majority of infected individuals have an asymptomatic or mildly symptomatic disease.[2]

There is increasing interest in the potential of ‘big data’ analysis to predict future areas of COVID-19 outbreaks and incidence of cases based on symptom search behaviors. In the past, search query data have been used to facilitate early detection and near real-time estimates of flu and Dengue.[3] A few studies have shown a correlation between Google Trends of medical terms searches and COVID-19 metrics,[4] suggesting that incorporating Google Trends data into conventional metrics could lead to better nowcasting and forecasting of the COVID-19 epidemic.

In this study, we systematically evaluate patterns of web queries for COVID-19 clinical manifestations and develop an open-access web application for exploring their correlations with COVID-19 propagation. We implement models integrating conventional COVID-19 metrics with Google Trends data and compare them to those not containing Google Trends data. The aim of this study is to present a framework for digital surveillance of COVID-19 using open-access big data of Google searches of symptoms associated with COVID-19.

## METHODS

### Data Collection

COVID-19 daily new confirmed, cumulative number, and number per million of cases and deaths for all available world countries were exported from the COVID-19 Data Repository by the Center for System Science and Engineering at Johns Hopkins University.[5] The selected countries used as case studies are given in the results section below. Countries choice was arbitrary, and the following principles were adopted: representation of the five continents; inclusion of countries where the COVID-19 epidemic had different levels of severity and different evolutions over time; inclusion of countries where Google is the preferred search engine; exclusion of countries with limited access to the internet; exclusion of countries where one or more Google Trends topic had only zero or missing values in the selected time frame; exclusion of countries whose reliability in terms of data reporting has been questioned. As data were fully anonymized and publicly available, no ethical approval was required.

Google Trends API was used to extract trends of Google searches for the most common COVID-19 signs and symptoms in those countries.[6] For each search term, geographic region, and time frame selected, Google Trends outputs an ‘interest-over-time’ (IOT) index, which estimates the relative search volume on a normalized scale from 0 (no searches) to 100 (search term popularity peak). Twenty topics were identified on the basis of the most frequent signs and symptoms of COVID-19 and included: abdominal pain, ageusia, anorexia, anosmia, bone pain, chills, conjunctivitis, cough, diarrhea, eye pain, fatigue, fever, headache, myalgia, nasal congestion, nausea, rhinorrhea, shortness of breath, sore throat, tearing.[7-11] Google Trends queries were carried out with the «topic» function, which includes all the related terms sharing the same concept in different languages. This approach ensures that the frequency of searches for closely related symptom types are

appropriately grouped together.

For each country and search term, data were automatically exported as csv files for two pre-specified timeframes: (i) five years weekly data from 1/Jul/2015 to 1/Jul/2020 to study the long-term pattern of searched term, and (ii) daily data from 22/Jan/2020 to 20/Dec/2020. As Google Trends allows daily data exportation up to nine months, daily data were reconstructed by means of an overlapping method.[12]

**Data Analysis**

IOT values for the five-year interval were used to distinguish topics with a significant deviation from their long-term pattern from the onset of the COVID-19 epidemic. For seasonal queries, trends were isolated from seasonal and random components with an additive decomposition method (Supplementary Figure 1); for non-seasonal queries, trends were extracted by smoothing the time series with a one-year moving average. Decomposition plots were visually inspected, and topics with no clear change in their five-year trends from January 2020 were excluded from the subsequent analyses.

The relationship between the daily IOT values for the selected topics and COVID-19 confirmed deaths and new cases were investigated in the shorter time frame indicated above. Relationships between IOT values of each topic and the number of new daily confirmed cases or deaths per million were visually assessed with line graphs. Changes in IOT values over time were visually assessed with streamgraphs. To smooth daily fluctuations in both IOT values and number of new cases, plots were generated using a 7-day moving average.

Time-lagged cross-correlations between COVID-19 new cases and each topic were calculated, using a 7-day moving average of both IOT values and COVID-19 confirmed cases and deaths to blunt the day-by-day fluctuation.

**Model Development and Assessment**

Principal Components Analysis (PCA) was used to perform data dimensionality reduction, decrease the number of input variables, and filter out noisy or redundant information. Two PCA models were implemented for each country: one using unprocessed data and the other using a 7-day moving average smoothing. PCA was applied to standardized data (i.e., with zero mean and unit variance). The PCA model was graphically inspected through PCA score and loading plots. PCA was assessed via 5-fold cross-validation, and the results obtained in each test sample were averaged. The amount of variance explained by each principal component (PC) in the model was inspected with scree plots, and, based on the elbow and Kaiser rules, the first two PCs (PC1 and PC2) were subsequently used for time-series modeling.[13]

Three different time series models were fitted on PC1 and PC2 values: Error Trend Seasonality (ETS), Autoregressive integrated moving average (ARIMA), and a feed-forward neural network autoregression (NNAR) model with one hidden layer.[14] Models were fitted on a 30-day window and used to predict future PC1 and PC2 values up to 14 days. The fourteenth predicted day was aligned to the peak and base of each wave. The new data scores predicted with the time series models were then reinserted into the model as input variables.

For each country, the three models were compared in terms of ability to predict the PC1 and PC2 using the root-mean-square error (RMSE) of the predicted values. For each time-series model, the predictive abilities of the model generated with raw data and the one generated with 7-day moving averages were compared. To further assess the PCA models based on both Google Trends data and conventional COVID-19 metrics, we also generated predictive models based on conventional COVID-19 metrics only; we then compared the

predictive ability of models with and without Google Trends data by means of RMSE for each country.

### Web-application

An open-source web application was developed in the R Shiny.[15] Data are collected, imported, and updated daily for 188 countries from the sources mentioned above.

The web application allows users to generate line graphs and streamgraphs to visualize IOT values and COVID-19 metrics and view worldwide trends over time in a choropleth map. Relationships between the variables at the various lags can be explored with cross-correlations. The web application allows fitting and evaluating PCA models, fitting a time-series model (either ETS or ARIMA), predicting PC components or any of the input variable of the model (including numbers of new cases and deaths), and evaluating the model performance graphically and with various metrics, such as RMSE and mean absolute error. The user has operational control on several model features, including the subset of variables to build the PCA model, the time window to fit the time-series model, and the time interval to predict.

## RESULTS

Three European countries (Italy, UK, and France), one Asian country (India), one Oceanian country (Australia), one North American country (US), one South American country (Brazil), one African country (South African), and one Middle Eastern country (Iran) were chosen as case studies. The cumulative numbers of cases and deaths in the selected countries are illustrated in Figure 1.

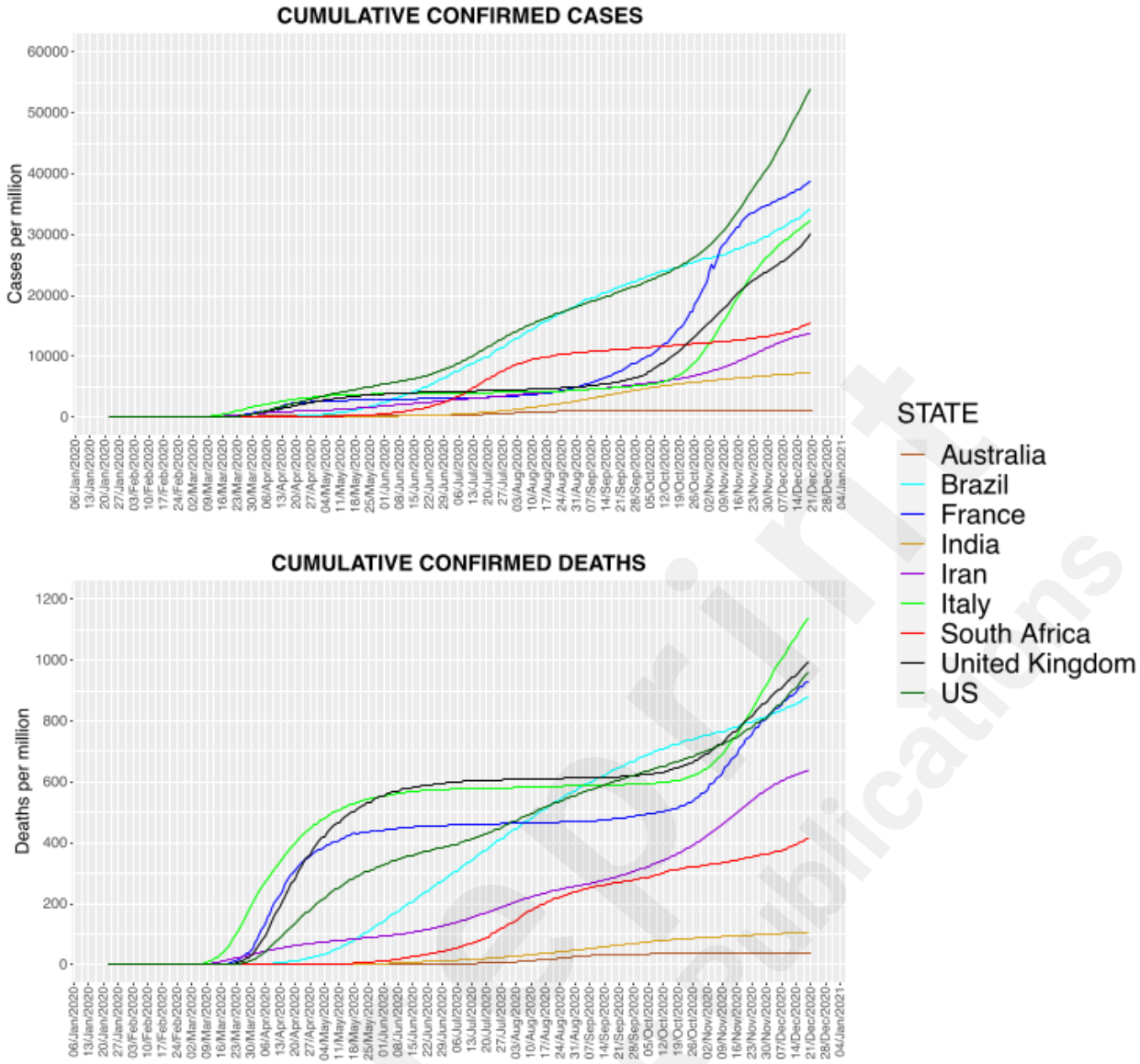



Table 1 summarizes information of the five-year analysis. Among the 20 screened topics, 13 showed seasonality, while the remaining were non-seasonal. Overall, eleven searched topics (Supplementary Figures 2-20) showed a clear deviation from their five-year trend: ageusia, anosmia, chills, cough, eye pain, fever, headache, nasal congestion, rhinorrhea, shortness of breath, sore throat.

**Table 1**. Symptoms screened at the 5-year analysis

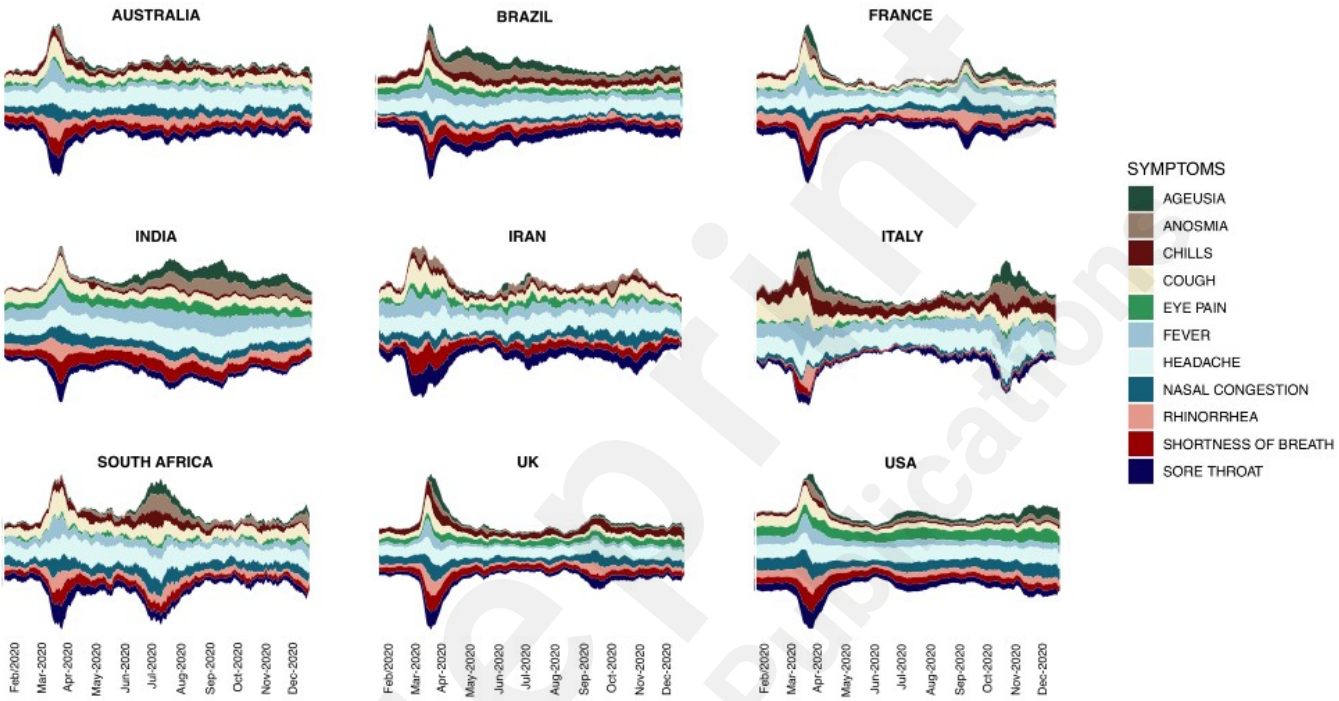
| Topic | Seasonality | Deviation from 5-year trend |
|---|---|---|
| Abdominal Pain | Nonseasonal | No |
| Ageusia | Nonseasonal | Yes |
| Anorexia | Seasonal | No |
| Anosmia | Nonseasonal | Yes |

| | | |
|---|---|---|
| Bone Pain | Nonseasonal | No |
| Chills | Seasonal | Yes |
| Conjunctivitis | Seasonal | No |
| Cough | Seasonal | Yes |
| Diarrhea | Seasonal | No |
| Eye Pain | Nonseasonal | Yes |
| Fatigue | Seasonal | No |
| Fever | Seasonal | Yes |
| Headache | Seasonal | Yes |
| Myalgia | Seasonal | No |
| Nasal Congestion | Seasonal | Yes |
| Nausea | Nonseasonal | No |
| Rhinorrhea | Seasonal | Yes |
| Shortness of Breath | Seasonal | Yes |
| Sore Throat | Seasonal | Yes |
| Tearing | Nonseasonal | No |
| Topic categorization into deviating and not deviating from their 5-year trend was determined on the visual inspection of decomposition plots. | | |

The relationships between the number of new cases and each searched topic are illustrated in Supplementary Figures 21-29. Several symptoms, including ageusia, anosmia, cough, rhinorrhea, and sore throat were aligned with the COVID-19 epidemic in most countries and were searched on Google well before the number of COVID-19 confirmed cases peaked. On the other hand, other topics showed less evident variations (chills, eye pain) or deviated from their trend only during the first wave (headache, shortness of breath). Also, the peak of interest of all symptoms (except eye pain) anticipated that of confirmed COVID-19 cases in most countries, and topics increasing earlier reached their highest IOT value before those growing later. Similar patterns were observed for IOT of searched terms when compared to the number of newly confirmed deaths (Supplementary Figures 30-37).

The IOT change for all the topics is illustrated in Figure 2. Overall, the IOT values of the selected topics had a peak in March in all the selected countries. In Italy, France, South Africa, and, to a lesser extent, Iran, the UK, and the US, there was a decrease in the searched terms after the first peak, followed by a second peak. In Iran, a third peak in

searches was seen, corresponding to the third COVID-19 wave. In India and Brazil, searches of medical terms remained high after the first peak, and no second peak was seen. In Australia, the IOT values of the selected topics returned to the pre-peak values soon after the first peak in March and remained low and stable.




Cross-correlations between each topic and the number of confirmed COVID-19 cases are reported in Supplementary Tables 1-9. Overall, ageusia, anosmia, and headache were most consistently correlated with COVID-19 cases across the selected countries. The degree of correlation and the best time lag largely varied as a function of the selected country and topic, but, overall, the optimal time lag was 15 days.

The scores and loadings plots for the PCA models are given in Figures 3 (Italy, USA and Australia) and Supplementary Figures 38-43 (remaining countries). The scores plot represents a summary of the collected data trends over time, while the loadings plot shows how strong each variable influences a PC. In the month of March 2020, all the selected countries deviated considerably from their previous scores and moved toward the PCs

directions of the loadings of the Google searched terms, anticipating the increment in the number of searches of the symptoms related to COVID-19. The latest 14 days show a similar pattern for all the selected countries, with the latest days pointing toward the loadings' directions of deaths and new cases. Specifically, for Italy, France, and South Africa, the latest scores are in the same area of loading plots corresponding to deaths and new cases, indicating a stable trend in these metrics. On the other hand, the UK, the US, Brazil, and India followed a worsening pattern, as their scores kept moving toward the direction of new deaths and cases. Australia and Iran were the only selected countries showing an improving trend, with the scores points moving away from the loadings of COVID-19 deaths and new cases.

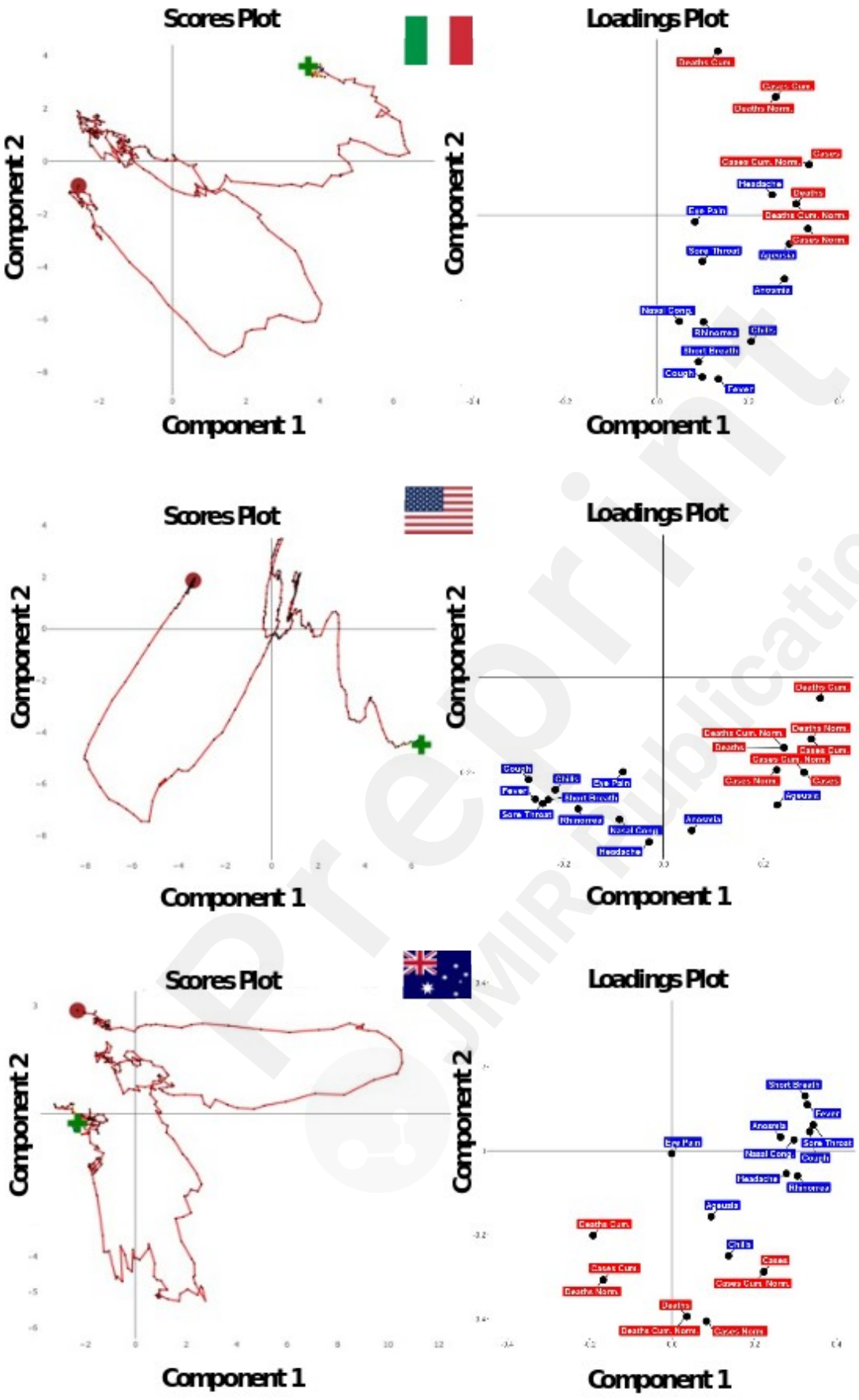
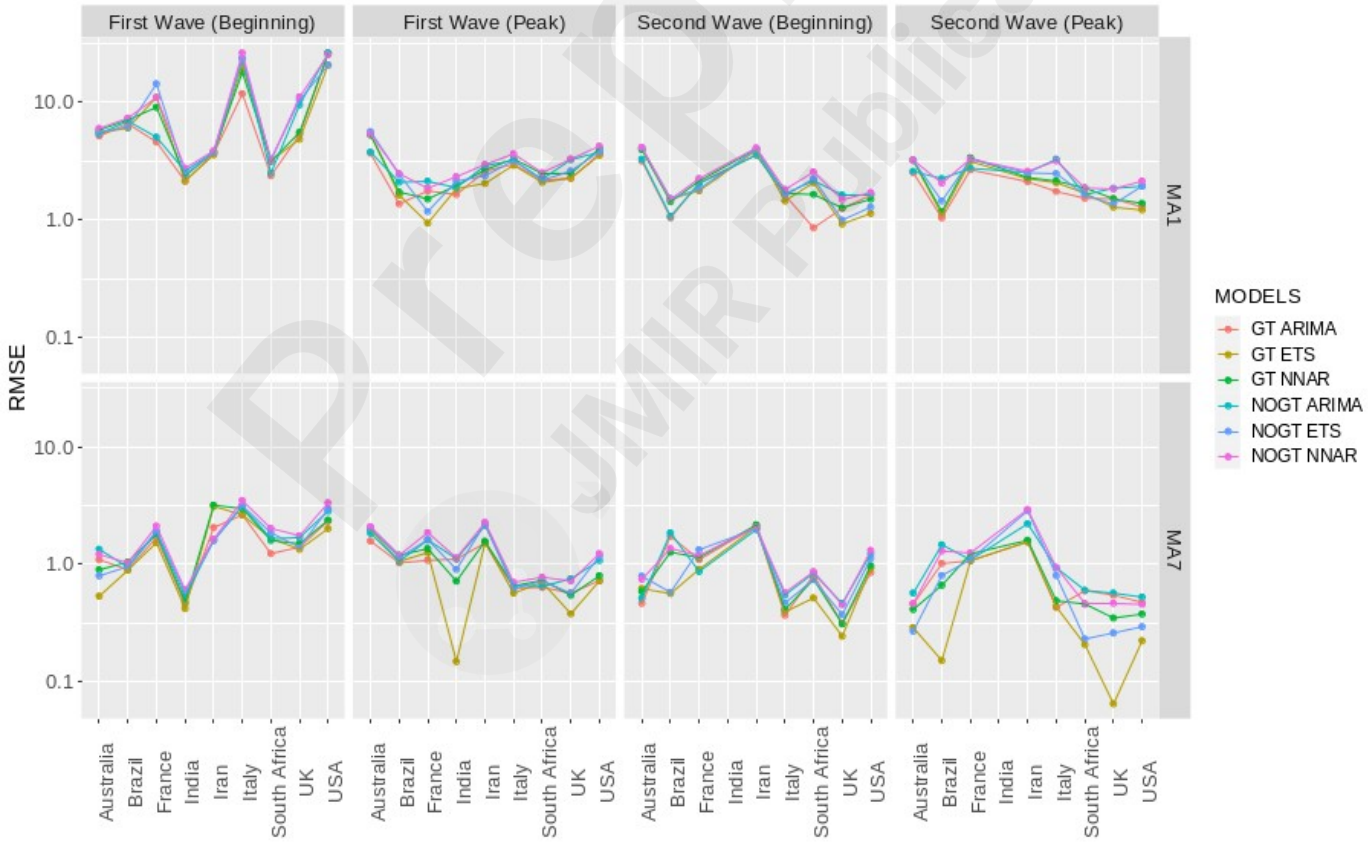



Figure 4 illustrates the RMSE for the prediction of PC1 values with the three time-series

models in the various countries, using raw data and a 7-day moving average. Models based on the 7-day moving average lead to considerably smaller RMSE values as opposed to those calculated with raw data (median [IQR]: 0.90 [0.50-1.53] vs. 2.27 [1.62-3.74], respectively). Overall, predictions based on both searched terms and COVID-19 conventional metrics performed better than those not including Google searches (median [IQR]: 1.56 [0.90-2.49] vs. 1.87 [1.09-2.95], respectively), but the improvement in prediction varied as a function of the selected country and timeframe. Although ETS (median [IQR]: 1.62 [0.87-2.7]) led to slightly smaller RMSE than ARIMA (median [IQR]: 1.65 [1.04-2.58]) and NNAR (median [IQR]: 1.82 [1.15-3.15]) models, none of the tested time-series models clearly outperformed the remaining two, and the best model varied as a function of country, time range, and period of time selected.




Similar results were obtained when trying to predict PC2 (Supplementary Figure 44).

## DISCUSSION

In this study, we investigated the relationship between Google Trends searches of symptoms associated with COVID-19 and confirmed COVID-19 cases and deaths. We found that some of the searched terms showed an unusually high recent online interest that deviated considerably from their expected behavior and anticipated the peak of confirmed COVID-19 cases by days to weeks. This pattern was consistent across different countries and of similar magnitude. We developed and validated predictive models to forecast the COVID-19 epidemic based on the combination of Google Trends searches of symptoms associated with COVID-19 and traditional COVID-19 metrics. We found that models incorporating Google Trends data performed generally better than those based solely on traditional COVID-19 metrics. We also developed a web application (https://predictpandemic.org) to translate our approach into action.

Our study identified patterns of Google searches of several symptoms and signs associated with COVID-19 in a consistent way across the studied countries. Overall, Google searches of COVID-19 symptoms followed a similar trend to that of the COVID-19 epidemic and anticipated traditional COVID-19 metrics. This behavior can contribute to the early recognition of new waves and epidemic peaks.

The interpretation of symptom search behavior during COVID-19 outbreaks should be carefully considered. Dynamics of online searches may show atypical patterns during pandemics where major restrictions occur, including shutdowns of economic activities, movement restrictions, and healthcare overload.[16] Constant media attention may contribute to raising the interest for some of the studied topics.[17] COVID-19 received extensive coverage that might have precipitated unusually high interest during lockdowns.[18] Our findings on online search behavior might be secondary to general media interest in specific COVID-19 symptoms, rather than a primary, and possibly predictive, consequence

of COVID-19 sufferers researching their own symptoms in real-time. All the selected countries had a peak in searches of medical terms in or around March and April 2020, including those countries with low numbers of cases at that time, such as South Africa and India. This pattern may indicate that curiosity and media clamor toward the new pandemic can explain part of the first peak in searched terms, in agreement with a previous study.[19] After the first peak, however, Google search behavior followed different patterns across countries and resembled the course of the COVID-19 epidemic. In those countries having a second wave, such as Italy, France, Iran, the UK, or the US, the number of Google searches had a second peak; the height of the second peak of the searches was lower than that of the first peak, despite a higher number of reported cases and deaths, suggesting that the individual curiosity toward the new pandemic could have inflated the first peak in searched terms. Iran also had a third peak in searches between October and December 2020, when the country had a third COVID-19 wave of its outbreak. South Africa had a second peak in its searches in July 2020, when the country had its first wave. In India, the first peak in searches was followed by a steady increase from June 2020, remained stable until October 2020, and then decreased gradually, resembling the shape of the COVID-19 epidemic in this country. In Australia, which effectively managed the COVID-19 epidemic and had among the lowest infection and death rates in the world, the IOT of the various search terms after the first peak remained low and comparable to pre-peak values.

We observe that not all the selected topics reached their peak searches simultaneously, but they had different time patterns, which were fairly consistent across all countries. We believe that the intense and simultaneous media coverage of all the selected topics should have had the same effect at the same date if the media influence entirely caused this search behavior.[20] Ageusia and anosmia showed the highest correlations when lagged by a few days, while cough, fever, nasal congestion, sore throat, rhinorrhea, and shortness of

breath anticipated COVID-19 by up to two weeks. This finding is consistent with the clinical course of COVID-19; in a large multicenter European study, olfactory and gustatory dysfunctions were among the latest and first manifestations in approximately 65% and 12% of patients, respectively.[9]

Besides describing how Google search terms changed over time in different countries and investigating their relationship with the numbers of cases and deaths, we also developed models combining IOT values of searches of COVID-19 symptoms with conventional metrics (e.g., number of new cases, number of new deaths) to predict the course of COVID-19 epidemic, and we compared the prediction ability of these models against that of models based only on conventional metrics. The PCA approach allowed us to reduce dimensionality, summarize information into 2 PCs, and filter out the noisy or redundant information. Another advantage of PCA was to provide visual representations of data patterns, similarity trends, and outliers. The PCA approach is highly flexible and potentially allows accommodating new variables of interest in future versions of our application. As the PCA itself does not make any predictions, we processed the PC computed values with different time-series models, and new data scores predicted with the time series models were reinserted into the PCA model. Our approach allows extracting the predicted values of any input variable, including the number of new cases and deaths. Models integrating IOT values of the searched topics and COVID-19 traditional metrics generally outperformed models based solely on confirmed cases and deaths, leading to improved predictions. There was no single best model in this study, and the best performing time series model varied as a function of the country, time frame, and moving average. Predictions were more accurate, leading to considerably smaller RMSE when obtained on a 7-day moving average, rather than on daily data. This result is not surprising as Google Trends data have high daily fluctuations, and COVID-19 reported cases greatly oscillate, reflecting testing and

reporting practices and contingencies.[21, 22]

To translate our results into practice so that the scientific community, agencies, and even curious users could potentially use them, we developed a web application, freely available at https://predictpandemic.org. The application is interactive and updates the data daily, so it operates in nearly real-time. It allows the user to visualize data for 188 world countries, choosing any time frame. Also, COVID-19 traditional metrics and google search terms IOT can be visualized globally on different graphs. The user can explore cross-correlations among selected keywords, generate predictive models with default or a user-selected subset of variables, and check model performance.

The present study has limitations. The Google Trends algorithm is a ‘black box’, and the exact calculation formula for the interest over time and raw data has never been made public. Searched results may differ slightly when download by different computers or on different days. However, we conducted search-research reliability, which showed excellent reliability for most of the topics included in this study (data not shown). The exclusion of those symptoms with no significant deviation from their five-year trends reduced the possibility of spurious correlations, but it was not possible to account for seasonality in the selected topics. In other words, a small proportion of the increasing trend in some topics might be explained by their usual seasonal variations. The results of this study may not apply to countries where Google is not a popular search engine or where Google is censored or limited in its use. However, this approach can be applied to other search engines (e.g., Baidu, Yahoo, Naver), as was done in previous studies on different diseases and on COVID-19 in Hubei province, China.[23, 24] Previous studies have shown that self-reported symptoms on social media, such as Twitter, can provide useful information to track the COVID-19 pandemic and can be used for infoveillance along with search engine data.[25-27] Our approach could, in principle, be applied to social media as well. Geographical

areas and groups of people (elders and children) with scarce Internet access cannot be studied with this strategy, and our results may not apply to largely rural countries. This study included only the most common clinical manifestations of COVID-19, and only a few selected countries were included as a case study. However, information and models for every country can be found on our web application.

Future work will include increased data granularity allowing to have information and make predictions at a regional level. Other metrics of interest, such as hospitalizations, will be included in our analysis as outcomes. Finally, we plan to allow the user to generate a one-page report for each country, summarizing the most relevant information.

In conclusion, the results of this study show that Google Trends searches during the COVID-19 pandemic may anticipate outbreaks by up to two weeks. The inclusion of digital online searches in statistical models may improve the nowcasting and forecasting of the COVID-19 epidemic and could be used as one of the surveillance systems employed by government agencies and supranational organizations to refine their monitoring of COVID-19 disease. We provide a free web application operating with nearly real-time data that anyone can use to make predictions of outbreaks, improve estimates of dynamics of ongoing epidemics, and anticipate future or rebound waves.

**Contributors:** AR and AM contributed to study conception, design, data acquisition, data analysis, data interpretation, manuscript drafting, and manuscript revision. EA contributed to data analysis, data interpretation, manuscript drafting, and manuscript revision. EM contributed to data acquisition, data analysis, and manuscript revision. AA, AIM, and FB contributed to data interpretation and manuscript revision. All authors had full access to all the data in the study. All authors revised the manuscript and approved the final version before submission. The corresponding author had final responsibility for the decision to submit for publication.

**Declaration of interests:** AIM reports personal fees and consultancy fees from Allergan, Pfizer, Novartis, Zeiss, Easyscan and Visufarma, outside the submitted work. FB reports consultancy fees from Allergan, Bayer, Boehringer-Ingelheim, Fidia Sooft, Hofmann La Roche, Novartis, NTC Pharma, Sifi, Thrombogenics, and Zeiss. All the authors declare no competing interests.

**Data sharing:** Raw data used in this study are publicly available on COVID-19 Data Repository by the Center for System Science and Engineering at Johns Hopkins University (https://github.com/CSSEGISandData/COVID-19) and Google Trends webpage (https://trends.google.com). All processed data used in this study can be accessed and reproduced on PredictPandemic website (https://predictpandemic.org). The code for the current version of the web application is open-source and freely available (https://zenodo.org/record/4603713#.YE4sfC2l28U)

**Funding:** This study was supported by the EOSCsecretariat.eu, which has received funding from the European Union's Horizon Programme call H2020-INFRAEOSC-05-2018-2019, grant Agreement number 831644.

**Acknowledgments:** None

## REFERENCES

1. Roda WC, Varughese MB, Han D, Li MY. Why is it difficult to accurately predict the COVID-19 epidemic? Infect Dis Model. 2020;5:271-81. PMID: 32289100. doi: 10.1016/j.idm.2020.03.001.

2. Fu L, Wang B, Yuan T, Chen X, Ao Y, Fitzpatrick T, et al. Clinical characteristics of coronavirus disease 2019 (COVID-19) in China: A systematic review and meta-analysis. J Infect. 2020 Apr 10. PMID: 32283155. doi: 10.1016/j.jinf.2020.03.041.

3. Ginsberg J, Mohebbi MH, Patel RS, Brammer L, Smolinski MS, Brilliant L. Detecting influenza epidemics using search engine query data. Nature. 2009 Feb 19;457(7232):1012-4. PMID: 19020500. doi: 10.1038/nature07634.

4. Walker A, Hopkins C, Surda P. Use of Google Trends to investigate loss-of-smell-related searches during the COVID-19 outbreak. Int Forum Allergy Rhinol. 2020 Jul;10(7):839-47. PMID: 32279437. doi: 10.1002/alr.22580.

5. Dong E, Du H, Gardner L. An interactive web-based dashboard to track COVID-19 in real time. Lancet Infect Dis. 2020 May;20(5):533-4. PMID: 32087114. doi: 10.1016/S1473-3099(20)30120-1.

6. Worldometers. Confirmed cases and deaths by country, territory, or conveyance. https://www.worldometers.info/coronavirus/#countries (accessed April 10, 2020).

7. Guan WJ, Ni ZY, Hu Y, Liang WH, Ou CQ, He JX, et al. Clinical Characteristics of Coronavirus Disease 2019 in China. N Engl J Med. 2020 Feb 28. PMID: 32109013. doi: 10.1056/NEJMoa2002032.

8. Fu L, Wang B, Yuan T, Chen X, Ao Y, Fitzpatrick T, et al. Clinical characteristics of coronavirus disease 2019 (COVID-19) in China: A systematic review and meta-analysis. J Infect. 2020 Apr 10. PMID: 32283155. doi: 10.1016/j.jinf.2020.03.041.

9. Lechien JR, Chiesa-Estomba CM, De Siati DR, Horoi M, Le Bon SD, Rodriguez A, et

al. Olfactory and gustatory dysfunctions as a clinical presentation of mild-to-moderate forms of the coronavirus disease (COVID-19): a multicenter European study. Eur Arch Otorhinolaryngol. 2020 Apr 6. PMID: 32253535. doi: 10.1007/s00405-020-05965-1.

10. Wu P, Duan F, Luo C, Liu Q, Qu X, Liang L, et al. Characteristics of Ocular Findings of Patients With Coronavirus Disease 2019 (COVID-19) in Hubei Province, China. JAMA Ophthalmol. 2020 Mar 31. PMID: 32232433. doi: 10.1001/jamaophthalmol.2020.1291.

11. Ngo CT, Mokete B. A suture technique for leaking sclerotomies. Retina Today, July/August 2012, pp. 69–70.

12. Tseng Q. Reconstruct Google Trends Daily Data for Extended Period. 2019 [10/12/2020]; Available from: https://towardsdatascience.com/reconstruct-google-trends-daily-data-for-extended-period-75b6ca1d3420.

13. Afifi A, May S, Donatello RA, Clark VA. Practical Multivariate Analisys (6th ed.). Boca Raton, FL: Chapman & Hall/CRC; 2020. ISBN: 9781315203737.

14. Hyndman RJ, Khandakar Y. Automatic Time Series Forecasting: The forecast Package for R. Journal of Statistical Software; Vol 1, Issue 3 (2008). 2008 07/29/.

15. Rabiolo A, Alladio E, Morales E, Marchese A. PredictPandemic.org. 2021 [17/03/2021]; Available from: https://predictpandemic.org.

16. Adebayo G, Neumark Y, Gesser-Edelsburg A, Abu Ahmad W, Levine H. Zika pandemic online trends, incidence and health risk communication: a time trend study. BMJ Glob Health. 2017;2(3):e000296. PMID: 29082006. doi: 10.1136/bmjgh-2017-000296.

17. Cervellin G, Comelli I, Lippi G. Is Google Trends a reliable tool for digital epidemiology? Insights from different clinical settings. J Epidemiol Glob Health. 2017 Sep;7(3):185-9. PMID: 28756828. doi: 10.1016/j.jegh.2017.06.001.

18. Effenberger M, Kronbichler A, Shin JI, Mayer G, Tilg H, Perco P. Association of the COVID-19 pandemic with Internet Search Volumes: A Google Trends(TM) Analysis. Int J

Infect Dis. 2020 Apr 16. PMID: 32305520. doi: 10.1016/j.ijid.2020.04.033.

19. Szmuda T, Ali S, Hetzger TV, Rosvall P, Sloniewski P. Are online searches for the novel coronavirus (COVID-19) related to media or epidemiology? A cross-sectional study. Int J Infect Dis. 2020 Aug;97:386-90. PMID: 32535297. doi: 10.1016/j.ijid.2020.06.028.

20. Adawi M, Bragazzi NL, Watad A, Sharif K, Amital H, Mahroum N. Discrepancies Between Classic and Digital Epidemiology in Searching for the Mayaro Virus: Preliminary Qualitative and Quantitative Analysis of Google Trends. JMIR Public Health Surveill. 2017 Dec 1;3(4):e93. PMID: 29196278. doi: 10.2196/publichealth.9136.

21. Bergman A, Sella Y, Agre P, Casadevall A. Oscillations in U.S. COVID-19 Incidence and Mortality Data Reflect Diagnostic and Reporting Factors. mSystems. 2020 Jul 14;5(4). PMID: 32665331. doi: 10.1128/mSystems.00544-20.

22. Rovetta A. Reliability of Google Trends: Analysis of the Limits and Potential of Web Infoveillance During COVID-19 Pandemic and for Future Research. medRxiv. 2021:2020.12.29.20248969. doi: 10.1101/2020.12.29.20248969.

23. Yuan Q, Nsoesie EO, Lv B, Peng G, Chunara R, Brownstein JS. Monitoring influenza epidemics in china with search query from baidu. PLoS One. 2013;8(5):e64323. PMID: 23750192. doi: 10.1371/journal.pone.0064323.

24. Qiu HJ, Yuan LX, Huang XK, Zhou YQ, Wu QW, Zheng R, et al. [Using the big data ofinternet to understand coronavirus disease 2019's symptom characteristics: a big data study]. Zhonghua Er Bi Yan Hou Tou Jing Wai Ke Za Zhi. 2020 Mar 18;55(0):E004. PMID: 32186171. doi: 10.3760/cma.j.cn115330-20200225-00128.

25. Klein AZ, Magge A, O'Connor K, Flores Amaro JI, Weissenbacher D, Gonzalez Hernandez G. Toward Using Twitter for Tracking COVID-19: A Natural Language Processing Pipeline and Exploratory Data Set. J Med Internet Res. 2021 Jan 22;23(1):e25314. PMID: 33449904. doi: 10.2196/25314.

26. Mackey T, Purushothaman V, Li J, Shah N, Nali M, Bardier C, et al. Machine Learning to Detect Self-Reporting of Symptoms, Testing Access, and Recovery Associated With COVID-19 on Twitter: Retrospective Big Data Infoveillance Study. JMIR Public Health Surveill. 2020 Jun 8;6(2):e19509. PMID: 32490846. doi: 10.2196/19509.

27. Panuganti BA, Jafari A, MacDonald B, DeConde AS. Predicting COVID-19 Incidence Using Anosmia and Other COVID-19 Symptomatology: Preliminary Analysis Using Google and Twitter. Otolaryngol Head Neck Surg. 2020 Sep;163(3):491-7. PMID: 32484425. doi: 10.1177/0194599820932128.

## ABBREVIATIONS

ARIMA: Autoregressive integrated moving average, ETS: Error Trend Seasonality; IOT: interest over time; IQR: interquartile range; NNAR: neural network autoregression; PC: principal component; PCA: principal components analysis; RMSE: root-mean-square error.

## FIGURE LEGENDS

**Figure 1. Cumulative number of confirmed cases (top panel) and deaths (bottom panel) per million for each country over time.**

**Figure 2. Streamgraphs of interest over time (IOT) index for each individual country.** X-axis is given in months. IOT values were plotted as seven-day moving average.

**Figure 3. Principal component analysis scores and loadings plot for Italy (top panel), USA (middle panel), and Australia (bottom panel).** For each country, the data starts from 22/Jan/2020 (red dot) to 20/Dec/2020 (Green cross); black dots over red lines indicate weeks.

**Figure 4. Root-mean-square errors (RMSE) of the prediction error for the principal component 2 of the various models for the selected countries**. MA1 and MA7 indicate analyses performed on moving average of data 1 day (i.e., original data) and 7 days, respectively. GT indicates models based on both traditional COVID-19 metrics and Google Trends data, while NOGT models based on COVID-19 metrics only. ARIMA: Autoregressive integrated moving average; ETS: Error Trend Seasonality; NNAR: feed-forward neural network autoregression.

# Supplementary Files

# Multimedia Appendixes

Supplementary Appendix.
URL: http://asset.jmir.pub/assets/91920d83a0610446ec4ed9b595a6b295.docx