COVID-19 symptom Google search surges, precede local incidence surges: evidence from Spain

Alberto Jimenez, Rosa M. Estevez-Reboredo, Miguel A. Santed, Victoria Ramos

Submitted to: Journal of Medical Internet Research
on: August 14, 2020

Preprint
JMIR Publications

Disclaimer: © The authors. All rights reserved. This is a privileged document currently under peer-review/community review. Authors have provided JMIR Publications with an exclusive license to publish this preprint on it's website for review purposes only. While the final peer-reviewed paper may be licensed under a CC BY license on publication, at this stage authors and publisher expressively prohibit redistribution of this draft paper other than for review purposes.

Table of Contents

Preprint JMIR Publications

COVID-19 symptom Google search surges, precede local incidence surges: evidence from Spain

Alberto Jimenez[1*] MA; Rosa M. Estevez-Reboredo[1*] PhD; Miguel A. Santed[2*] PhD; Victoria Ramos[1*] PhD

[1]Instituto de Salud Carlos III Madrid ES
[2]The National Distance Education University (UNED) Madrid ES
[*]these authors contributed equally

Corresponding Author:
Miguel A. Santed PhD
The National Distance Education University (UNED)
Calle de Juan del Rosal, 10
Madrid
ES

Abstract

Background: COVID 19 is the first pandemic that has led to a global health crisis. This study is a small contribution that tries to find contrasted formulas to alleviate this global suffering and guarantee a more manageable future.

Objective: In this study, a statistical approach has been proposed that forecasts the incidence of the COVID 19 epidemic in Spain by means of correlation test and using information from search data provided by Google Trends.

Methods: Our method consists of the linear correlation between the Google Trends search data and the data provided by the National Center of Epidemiology in Spain -dependent on the Instituto de Salud Carlos III- of cases of COVID 19 reported with a certain time lag, enabling the identification of anticipatory patterns.

Results: In response to the ongoing outbreak, our results demonstrate that using this correlation test the evolution of the COVID 19 pandemic can be predicted in Spain, up to 11 days in advance.

Conclusions: During the epidemic, Google Trends offers the possibility to preempt health care decisions in real time by tracking people's concerns through their search patterns. This can be of great help given the critical (if not dramatic) need for complementary monitoring approaches, which can work on a population level, and inform public health decisions in real time.

The study of Google search patterns motivated by the fears of individuals in the face of a pandemic can be useful in anticipating its development.

(JMIR Preprints 14/08/2020:23518)
DOI: https://doi.org/10.2196/preprints.23518

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

Preprint JMIR Publications

Original Manuscript

Preprint JMIR Publications

Original Paper

Alberto Jimenez[1], MA ; Rosa M Estevez-Reboredo[2], PhD ; Miguel A Santed[3], PhD ; Victoria Ramos[4], PhD

[1]Instituto de Salud Carlos III, ISCIII, Information and Communication Technologies Unit, Monforte de Lemos, 5; 28029 Madrid, Spain.

[2]Instituto de Salud Carlos III, ISCIII, National Epidemiology Centre, Monforte de Lemos, 5; 28029 Madrid, Spain.

[3]The National Distance Education University (UNED), Madrid, Spain

[4]Instituto de Salud Carlos III, ISCIII, Telemedicine and Digital Health Research Unit, Monforte de Lemos, 5; 28029 Madrid, Spain.

Corresponding Author:

Miguel-Angel Santed-Germán, PhD
The National Distance Education University (UNED)
UNED - Facultad de Psicología
Calle de Juan del Rosal, 10
Madrid, 28040
Spain
Phone: 34 646517577
Email: msanted@psi.uned.es

COVID-19 symptom Google search surges, precede local incidence surges: evidence from Spain

Abstract

Background: COVID 19 is one of the biggest pandemics in human history along with other diseases such as Spanish flu, plaque and smallpox. This study is a small contribution that tries to find contrasted formulas to alleviate this global suffering and guarantee a more manageable future.

Objective: In this study, a statistical approach has been proposed to study the correlation between the incidence of the COVID 19 epidemic in Spain and search data provided by Google Trends.

Methods: Our method consists of the linear correlation between the Google Trends search data and the data provided by the National Center of Epidemiology in Spain -dependent on the Instituto de Salud Carlos III- of cases of COVID 19 reported with a certain time lag, enabling the identification of anticipatory patterns.

Results: In response to the ongoing outbreak, our results demonstrate that using this correlation test the evolution of the COVID 19 pandemic can be predicted in Spain, up to 11 days in advance.

Conclusions: During the epidemic, Google Trends offers the possibility to preempt health care decisions in real time by tracking people's concerns through their search patterns. This can be of great help given the critical (if not dramatic) need for complementary monitoring approaches, which can work on a population level, and inform public health decisions in real time.

The study of Google search patterns motivated by the fears of individuals in the face of a pandemic can be useful in anticipating its development.

Keywords: behavioral epidemiology; big data; smart data; tracking;

Preprint JMIR Publications

nowcasting; forecast; predict; infosurveillance; infodemiology; COVID-19.

Introduction

During the 2020 Chinese Lunar New Year, massive measures to reduce the spread of new 2019 coronavirus disease, now known as COVID 19, were first enacted by the Chinese authorities [1]. The first reported case of a SARS-CoV-2 infection appeared in late 2019 [2]. Subsequently, the WHO (World Health Organization) announced COVID 19 outbreak as a pandemic on 11 March 2020 [3]. The epidemiological characteristics of COVID 19 have not yet been fully understood, nor has its high transmissibility capacity, virulence, presence of asymptomatic carriers or those showing only mild symptoms. After seven months of the epidemic outbreak in March 2020, we are close to reaching one million deaths worldwide [4].

Besides Europe, Spain has the fifth highest number of detected cases in the world, behind the United States, Brazil, Russia and the United Kingdom [4].

As a consequence, developing a forecasting tool to predict the spread of the epidemic has become critical. This information can help us understand the evolution of how this coronavirus affects our health, and even be useful in preparing for future possible waves or even pandemics.

Outbreak detection

As Chu & Qureshi [5] pose:

> *[...] predicting the potential spread of COVID 19 is difficult task because we do not have many epidemiological data, such as the transmission mechanism, its mutation patterns, or the contagiousness of the virus, as well as other human factors such as the level of compliance with social distancing measures. Many models recently developed by infectious disease scientists (eg [6,7]) can produce vastly different predictions as they are constructed based on various assumptions that may not be close to reality (such as the actual level of compliance with social distancing may be much higher than what is assumed in the model, or the infection rates can vary across different regions and groups of people, which cannot be easily captured by any model).*

Google Trends (GT) offers a new approach to the possibility of predicting the pandemic by tracking individuals' concerns through their searches in these times. The work of Eysenbach G [8] is one of the pioneers studies to use Google Trends with such an approach in health. Later Ginsberg et al. [9] found a high correlation between the pattern of web search queries and the percentage of patients with influenza-like symptoms, also confirming that GT can detect -at that time- influenza expansion one or two weeks earlier than the CDC (Centers for Disease Control and Prevention).

Also, within the discipline of Behavioral Epidemiology there are articles that study fear in the development of epidemics, (eg, [10]) In our study, behavioral factors can be summarized through Google searches and then used as a correlation variable to identify patterns in the evolution of epidemics.

In recent years, the numbers of different search engines that deal with infodemiology issues -'*which studies the determinants and distribution of health information for public health purposes*' [8]- are increasing and GT is being tested as a useful tool for tracking social trends [11]. Until the end of May, during the COVID 19 pandemic, we found 7 articles that raised the

Preprint JMIR Publications

possibility of predicting the development of the disease (Table 1).

Table 1: articles that raised the possibility of predicting the development of COVID 19.

Refer.	Search Engine	Territory	Terms	Time Lag
[12]	GT, Baidu Index and Sina Weibo Index	China	Coronavirus, Pneumonia	6 to 8 days
[13]	GT	Taiwan	Handwashing and Mask Related Information	1 to 3 days
[14]	GT	China, Republic of Korea, Japan, Iran, Italy, Austria, Germany, UK, US, Egypt, Australia, and Brazil	'Coronavirus (Virus)'	11.5 days
[15]	GT	US	Sore Throat, Fever and Cough	1 to 2 weeks
[16]	GT	US	'COVID Pneumonia,' and 'COVID Heart'	12 days approx.
[17]	GT, Baidu Index	China, worldwide data, Italy and Spain, and the US states of New York and Washington	'Shortness of Breath', Anosmia, Dysgeusia and Ageusia, Headache, 'Chest Pain', and Sneezing Diarrhea, Fever, Cough, 'Nasal Obstruction', and Rhinorrhea	12 days
[18]	GT	32 countries	'Coronavirus Symptoms', 'Coronavirus Test', Fever, Cough, Coronavirus, 'Runny Nose', 'Dry Cough', 'Sore Throat', Chills, 'Shortness of Breath'	18 to 22 days

Preprint JMIR Publications

Objectives

In this study, a statistical approach has been proposed in order to test the correlation of the COVID 19 epidemic in Spain, using search data provided by GT.

In this paper, we study whether the GT data collected for searches using many different keywords (which the public entered into Google's Internet search engine during the coronavirus outbreak period) can predict the number of cases reported by the National Center of Epidemiology in Spain (Centro Nacional de Epidemiología, CNE).

Methods

Study design

Our null hypothesis is:

H0: There is no statistically significant relationship between the variables.

And the proposed alternative hypothesis:

H1: The obtained correlation coefficient comes from a population whose correlation coefficient is significant.

For the aforementioned objective, we analyzed search data obtained from GT and the official data on the number of daily cases registered by the CNE, during the period between February 20, 2020 and May 20, 2020.

The rationale behind choosing this time frame for data analysis has been that the CNE, carried out this work with a different method until May 20 2020, the current counting system began to be used from May 11. The CNE is the government body responsible for collecting and standardizing the data of the 17 autonomous communities that make up the Spanish state. Further analysis of this second data set is planned to extend the study.

Group of variables for GT Search Terms

In neither of the two groups of variables have we been able to obtain data for sex nor of gender. These are the two datasets to be correlated. Our methodology does not enjoy of explicit participants, in this may lie its debility but also its strength since this way lends itself to agile and timely interpretations.

Google Trend searches were carried out looking for COVID 19 symptoms, as well as search terms for COVID 19 synonyms. GT provides an index of time series data on the volume of queries users entered into Google within a given geographic area. Google calculates the number of searches for a given term as a proportion of the total number of searches in each location at any given time. These calculations are normalized in a Google Trends Relative Search Volume (RSV) index between 0 and 100, where an RSV index of 100 designates the date with the highest amount of search activity for that given term [19]. In the previous article [20] we established the mathematical formulation of how Google Trends calculates its monthly RSV for a particular term.

The terms searched were the equivalent in Spanish for:

- fatigue = cansancio
- coronavirus, COVID 19, COVID 19, COVID19 (here referenced as coronavirus+)
- diarrhea = diarrea
- sore throat = dolor de garganta
- fever = fiebre
- pneumonia = neumonia (without accent for being more relevant)
- lost sense of smell = perdida de olfato (without accent)
- cough = tos

Possible spurious terms that GT pointed out in its ‘related queries’ were eliminated in the search strings, putting the negation operator ‘-’ before the spurious term.

Full original GT searches -with its negatives queries- are shown in Textbox 1.

Textbox 1: Full original Google Trends searches, which deal with searches for the symptoms of COVID 19 (defined at that moment) in Spain for the period from February 20 to May 20 2020.

https://trends.google.com/trends/explore?date=2020-02-20%202020-05-20&geo=ES&q=cansancio%20-sociedad

https://trends.google.com/trends/explore?date=2020-02-20%202020-05-20&geo=ES&q=coronavirus,COVID 19,covid%2019,covid19

https://trends.google.com/trends/explore?date=2020-02-20%202020-05-20&geo=ES&q=diarrea

https://trends.google.com/trends/explore?date=2020-02-20%202020-05-20&geo=ES&q=dolor%20de%20garganta

https://trends.google.com/trends/explore?date=2020-02-20%202020-05-20&geo=ES&q=fiebre

https://trends.google.com/trends/explore?date=2020-02-20%202020-05-20&geo=ES&q=neumonia

https://trends.google.com/trends/explore?date=2020-02-20%202020-05-20&geo=ES&q=perdida%20olfato

https://trends.google.com/trends/explore?date=2020-02-20%202020-05-20&geo=ES&q=tos%20-opensigma%20-rap

Information provided by CNE

The CNE is the official Spanish center that collects and centralizes all the epidemiological information in the country. The accuracy of these data depends mainly on the agencies that supply it. In this case, it comes mainly from the Autonomous Communities that occupy the second administrative level within the Spanish Public Administration system.

The CNE itself has warned of a certain lack of homogenization in the data coming from the source of origin, the Autonomous Communities, in a first period. This gives rise to certain inconsistencies in the data received and is a problem that we have been able to verify ourselves when transforming the aggregate data of the CNE into daily values. Here we have obtained one negative value, but we have not normalized these data, since we consider that they are indeed representative of the state of emergency that we have gone through. In addition, even with this evidently erroneous data, the correlations obtained are very good.

This first original data for COVID 19 outbreak in Spain was found in CNE´s web page at May 24, 2020 under the tab ‘Documentación y Datos’ [21], CNE´s COVID 19 official data offered at that time aggregate numbers for:

- Polymerase Chain Reaction positive results (PCR+)
- Number of people hospitalized
- Admissions in Intensive Care Units (ICU)
- Deceased persons

This dataset is no longer available, and the one in its place is the result of a second data collection method since May 11. Our intention is to study this data set in future research using the same methodology as in the current study.

In the first part of our analysis we have used PCR + data, while for our final graphs (Figure 2) we have used the 4 data sets separately in order to draw the daily delay graph for each symptom or searched term.

Preprint JMIR Publications

Bias, study size and participants

The potential source of bias may result from the choice of GT search terms as COVID 19 symptoms differ slightly in Spanish vocabulary. Study size pretends to be the one of the universe as GT shows data for all the searches, and CNE´s data makes also a surveillance of all cases.

Statistical analysis

Pearson's correlation coefficient has been used to study the linear relationship between the two continuous variables (each of the symptoms searched for in GT versus the number of daily positive PCRs). This is a parametric test, it infers its results to the real population, which makes it necessary for the distribution of the sample to resemble the real distribution, and therefore, for normality to exist. In this way, the data, drawn randomly from a population in which the correlated variables are normally distributed, must be validated.

Given that the sample size is less than 50, the appropriate test for contrasting the goodness of fit to a normal distribution is the Shapiro-Wilk test, whose null hypothesis implies that the data are normally distributed.

This coefficient enables us to understand the intensity and direction of the relationship between them. It is a symmetric measure, that is, the correlation between X_i and Y_i is the same as between Y_i and X_i.

Time lag correlations were used to assess whether increases in GT data had a correlation with the evolution of the pandemic.

We will use a *P* < .05 threshold for significance.

Results

Table 2 presents the Pearson correlation coefficients for each of the symptoms categorized in the searches, from the day of the initial search until 21 days later. For each symptom, the day with the highest correlation is shaded.

Table 2: Pearson correlation coefficients and P values for each of the symptoms with time lags with respect to the incidence data of COVID 19

Lag in days	Sore throat	Coronavirus +	Fever	Cough	Diarrhea	Pneumonia	Lost sense of smell	Lag in days	Fatigue
1	0,4848 < .05	0,5090 < .05	0,4009 < .05	0,4149 < .05	0,5870 < .05	0,6551 < .05	0,6770 < .05	22	0,3926 < .05
2	0,5549 < .05	0,5730 < .05	0,4770 < .05	0,4964 < .05	0,6446 < .05	0,7163 < .05	0,7151 < .05	23	0,3782 < .05
3	0,6121 < .05	0,6273 < .05	0,5437 < .05	0,5741 < .05	0,6815 < .05	0,7654 < .05	0,7489 < .05	24	0,3632 < .05
4	0,6708 < .05	0,6795 < .05	0,6155 < .05	0,6443 < .05	0,7654 < .05	0,8239 < .05	0,7846 < .05	25	0,4947 < .05
5	0,7369 < .05	0,7256 < .05	0,6771 < .05	0,6991 < .05	0,7743 < .05	0,8544 < .05	0,7847 < .05	26	0,5296 < .05
6	0,7679 < .05	0,7521 < .05	0,7238 < .05	0,7439 < .05	0,7782 < .05	0,8639 < .05	0,7806 < .05	27	0,5171 < .05
7	0,8055 < .05	0,7740 < .05	0,7656 < .05	0,7794 < .05	0,7780 < .05	0,8756 < .05	0,7726 < .05	28	0,5480 < .05
8	0,8358	0,8118	0,8100	0,8201	0,7922	0,8593	0,7319	29	0,5253

Preprint JMIR Publications

	< .05	*< .05*	*< .05*	*< .05*	*< .05*	*< .05*	*< .05*		*< .05*
9	0,8608 *< .05*	0,8507 *< .05*	0,8434 *< .05*	0,8584 *< .05*	0,7907 *< .05*	0,8604 *< .05*	0,6814 *< .05*	30	0,4720 *< .05*
10	0,8743 *< .05*	0,8766 *< .05*	0,8751 *< .05*	0,8822 *< .05*	0,8031 *< .05*	0,8501 *< .05*	0,6356 *< .05*	31	0,5342 *< .05*
11	0,8799 *< .05*	0,8999 *< .05*	0,9086 *< .05*	0,9015 *< .05*	0,8117 *< .05*	0,8585 *< .05*	0,6111 *< .05*	32	0,5016 *< .05*
12	0,8924 *< .05*	0,8944 *< .05*	0,9039 *< .05*	0,8965 *< .05*	0,7858 *< .05*	0,8484 *< .05*	0,5592 *< .05*	33	0,5427 *< .05*
13	0,8672 *< .05*	0,8468 *< .05*	0,8788 *< .05*	0,8681 *< .05*	0,7001 *< .05*	0,8127 *< .05*	0,4968 *< .05*	34	0,5521 *< .05*
14	0,8279 *< .05*	0,8065 *< .05*	0,8319 *< .05*	0,8296 *< .05*	0,6326 *< .05*	0,7668 *< .05*	0,4419 *< .05*	35	0,5664 *< .05*
15	0,7664 *< .05*	0,7443 *< .05*	0,7743 *< .05*	0,7839 *< .05*	0,5913 *< .05*	0,7099 *< .05*	0,3803 *< .05*	36	0,6350 *< .05*
16	0,7214 *< .05*	0,6811 *< .05*	0,7234 *< .05*	0,7448 *< .05*	0,5192 *< .05*	0,6415 *< .05*	0,3259 *< .05*	37	0,4981 *< .05*
17	0,6720 *< .05*	0,6214 *< .05*	0,6827 *< .05*	0,7030 *< .05*	0,4733 *< .05*	0,5844 *< .05*	0,2524 *< .05*	38	0,4711 *< .05*
18	0,6093 *< .05*	0,5517 *< .05*	0,6330 *< .05*	0,6654 *< .05*	0,4271 *< .05*	0,5467 *< .05*	0,2053 *0,0815*	39	0,4388 *< .05*
19	0,5788 *< .05*	0,4838 *< .05*	0,5810 *< .05*	0,6161 *< .05*	0,3607 *< .05*	0,5142 *< .05*	0,1690 *0,1559*	40	0,4631 *< .05*
20	0,5192 *< .05*	0,4083 *< .05*	0,5164 *< .05*	0,5598 *< .05*	0,2845 *< .05*	0,4694 *< .05*	0,1140 *0,3438*	41	0,4915 *< .05*
21	0,4314 *< .05*	0,3208 *< .05*	0,4349 *< .05*	0,4921 *< .05*	0,1975 *0,1012*	0,3812 *< .05*	0,0630 *0,6042*	42	0,5325 *< .05*

Pearson correlation coefficients and P values for each of the symptoms categorized in GT searches and daily Polymerase Chain Reaction positive results (PCR+) cases. Best values for the column are shaded.

Note 1: lag in days are the days of lag between the two variables to be correlated.

Note 2: The searches for 'fatigue' correlate less strongly, but they do so with a lead of 36 days, so we represent them at the end of the table with a different scale of days.

According to the data shown in table 2, it is tested whether the variables follow a normal distribution.

Table 3: Normality test.

	Statistic	*P* value
Coronavirus	.945	.278
Pneumonia	.861	.007
Fever	.946	.290
Cough	.944	.261
Lost sense of smell	.885	.018
Sore throat	.929	.134
Diarrhea	.877	.014
Fatigue	.952	.364

Note 1: Critical region = .908 for sample size of 21 and significance level α = .05, obtained from the table of critical values of $W_{n,a}$ for the Shapiro-Wilk test.

According to the Table of Normality Tests with the Shapiro-Wilk test, we see that the variables fulfill a normal distribution. For searches for the terms "pneumonia", "loss of smell" and "diarrhea, which have a value lower than the critical region (= .908) for the .95 confidence level and a *P* value lower than .05, the null hypothesis of normality could be rejected; however, these symptoms do follow a linear trend, as can be seen in the graphs in Figure 1. These values are

Preprint JMIR Publications

close to the critical region, so can also be considered to follow a normal distribution, as seen in the first image of Figure 2 (Time-lagged correlation: PCR +).

Outbreak control measures

The number of daily cases begins to correlate from the first day that searches started for all evaluated symptoms. Since the *P* values for the correlations between daily cases and symptoms are below the .05 level of significance, it can be stated that the correlation coefficients are significant, which justifies rejecting the null hypothesis. For the majority of the symptoms that were searched for, day 11 has the highest correlation to the detection of new cases.

For the search 'fatigue' the data begin to correlate from the 3rd week (day 22); for the previous weeks the measures have been eliminated applying the Chauvenet criterion; according to this criterion, coefficients outside the confidence interval of 0.3565 and 1.000 can be discarded.

Table 4 shows the coefficient of determination between 'cases / day' and the rest of the search variables in GT, with a critical level (P<.001) lower than the established level of significance, generally P<.05.

Table 4: coefficient of determination between 'cases / day' and the rest of the search variables in Google Trends, with a critical level (P<.001)

	R^2	*P* values
Coronavirus	0.8098	0.0000000
Pneumonia	0.7666	0.0000000
Fever	0.8256	0.0000000
Cough	0.8128	0.0000000
Loss of smell	0.6157	0.0000000
Sore throat	0.7964	0.0000000
Diarrhea	0.6588	0.0000000
Fatigue	0.4032	0.0000002

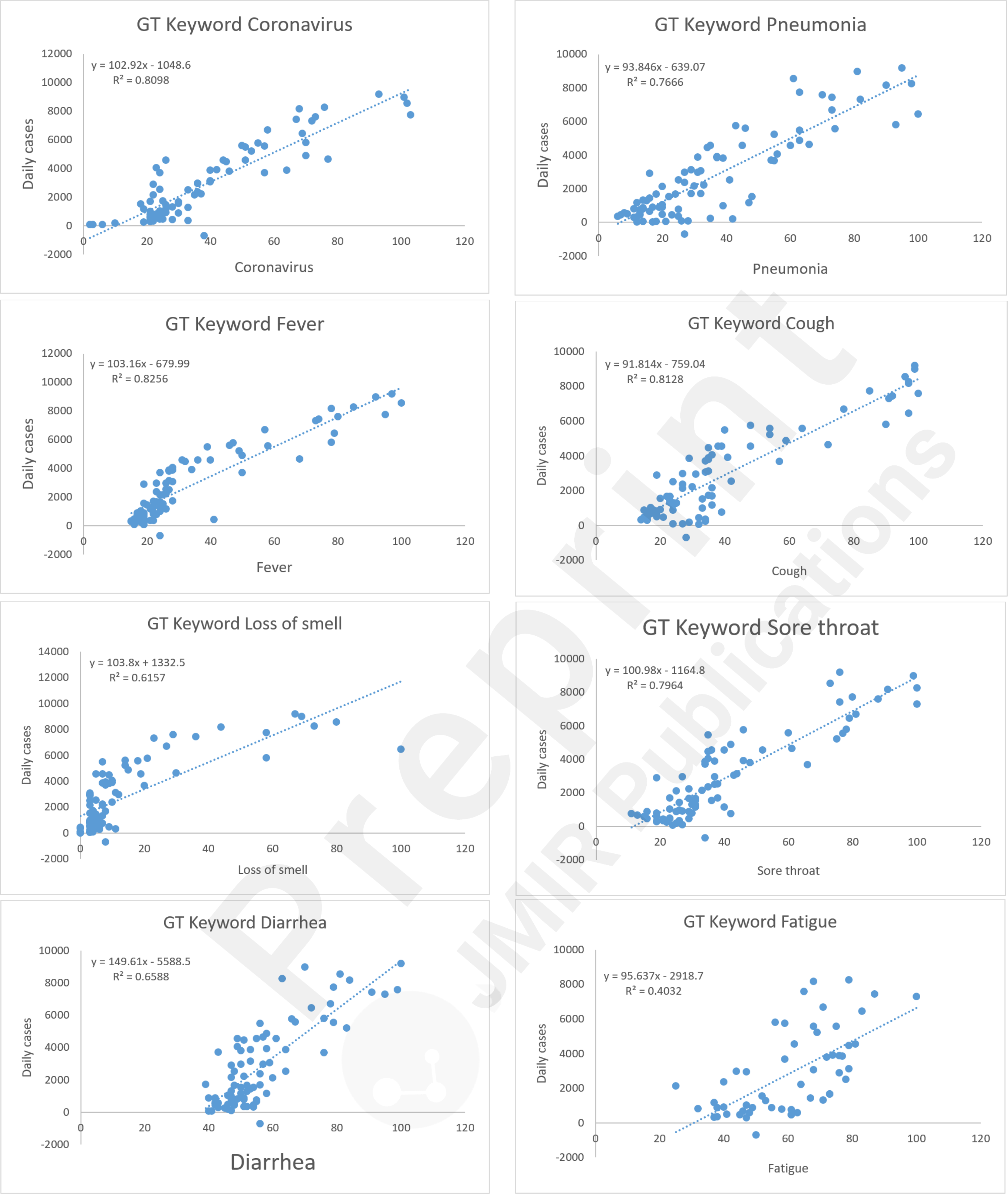
Coefficient of determination and significance level between daily Polymerase Chain Reaction positive results (PCR+) and the rest of the variables in searches by Google Trends.

Therefore, the coefficient is significantly different to zero. As a result, in this case, the null hypothesis is falsified (following Popper´s methodology [22]), so for the moment we can affirm that there is a significant positive linear relationship between daily cases and the searches for coronavirus and its symptoms in Google Trends, this suggest that the incidence of COVID 19 could be predicted 11 days in advance.

We used graphic procedures to verify the linearity of the study. The graphical representation of each of the symptoms (eight scatter graphs in Figure 1) shows the existence of a certain linear trend in each of the relationships. With the regression lines that were generated, a good follow-up of the data can be deduced. This then can be translated into the number of positive cases, thus verifying the correlation between GT searches and the incidence of COVID 19 in Spain.

Figure 1: linear trend in each of the relationships

Preprint JMIR Publications



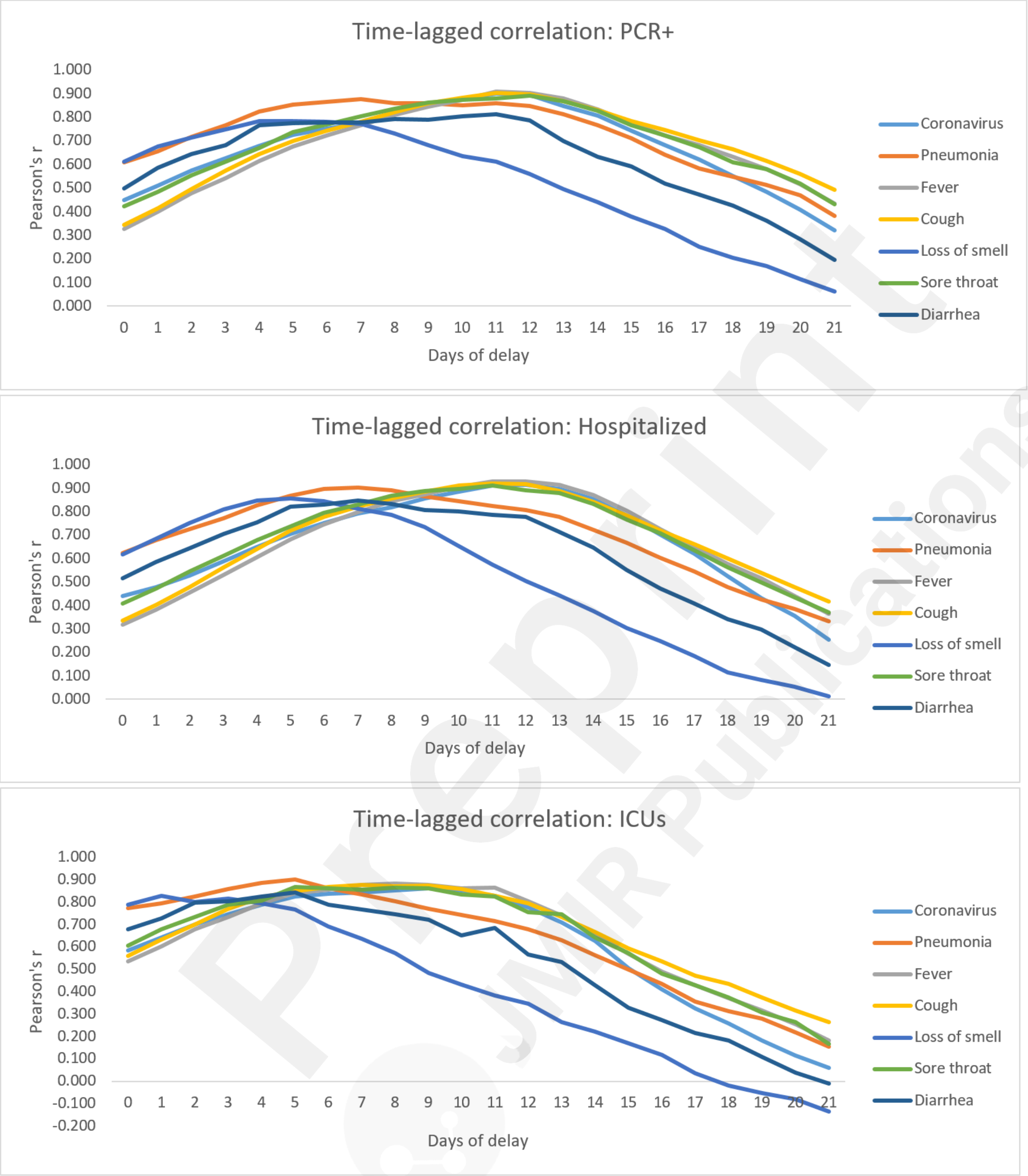
Correlations between new Polymerase Chain Reaction positive results (PCR+) confirmed daily cases of COVID 19 and data from Google Trends (GT) for COVID 19 related keywords, Spain, february-may 2020

The following graphs (Figure 2) represent the day with the best correlation between both variables (symptom keyword-positive case). They show the positive relationship between daily

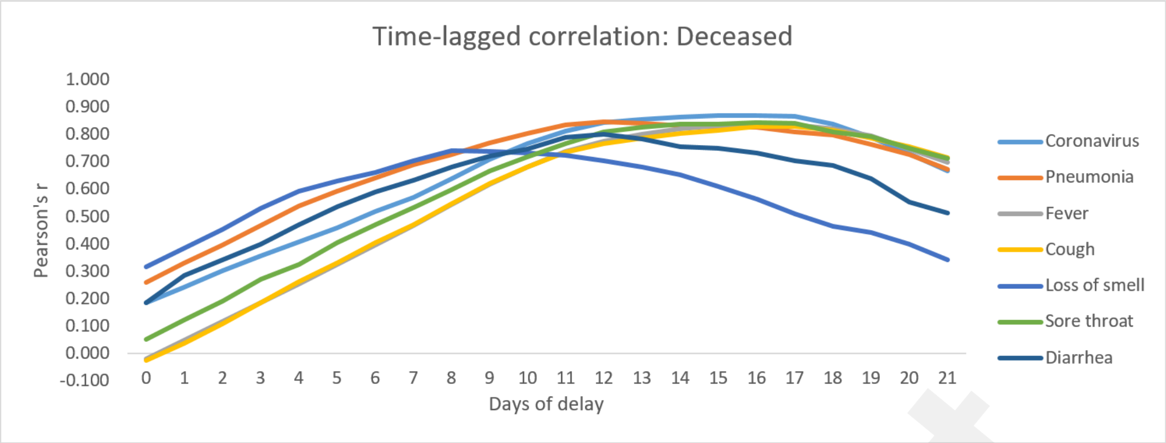
Preprint JMIR Publications

cases and the Google Trends searches for coronavirus and associated symptoms.

Figure 2: day with the best correlation between symptom keyword-positive case



Preprint JMIR Publications



Day when Polymerase Chain Reaction positive results (PCR+) daily cases of COVID 19 and data from Google Trends for COVID 19 related keywords best correlate.

Note: ICU, Intensive Care Units.

Discussion

Much like the influenza virus, SARS-CoV-2 causes ailments with certain flu-like symptoms, such as cough, fever and fatigue what in some cases can complicate the differential diagnosis. Exploring research using nontraditional data sources has several implications. We explore the use of search engines as help mitigate the impact of COVID 19. Our results demonstrated a potential application for the use of Google as a complementary tool to aid in understanding online search behavior, which could help mitigate the adverse effects of this pandemic and expedite the recovery process.

We found that Internet search patterns reveal a robust temporal pattern of disease progression for COVID-19. This study shows that Internet search patterns can be used to reveal the detailed clinical course of a disease. These data can be used to track and predict the local spread of COVID-19 before widespread laboratory testing becomes available, helping to guide the current public health response.

While laboratory testing serves as an important gauge of epidemic spread, it suffers from a number of important limitations. Alternative surveillance approaches are needed to overcome these limitations and serve as a complement to laboratory testing, especially during the critical early stages of a pandemic. Aggregated de-identified Internet search patterns have been used to track a wide range of health phenomena, and are a potential alternative source of information for surveilling pandemic spread.

When harnessed appropriately, Internet search patterns possess a number of powerful advantages relative to laboratory testing: 1. Surveillance data are available immediately when a new pandemic emerges. 2. Data are available at population-scale in countries with sufficient Internet access levels. 3. Delays are minimal, with search data available the same day. 4. There is no need for individuals to travel to a testing location; people can stay at home, avoiding increased exposure to themselves and to healthcare workers. 5. No physical intervention is required. 6. The data are available for free, independent of the scale of surveillance

Next work could be focused on checking the progression of symptom-related search terms over time in order to characterize the clinical course of COVID-19 by means of examine a range of possible search-term-based definitions for initial symptom onset. This should be based on various combinations of the earliest-peaking search terms and a detailed understanding of the stage of illness and the manifestations of the disease in the local environment and over time. Studies have indicated that the spread and severity of the disease can be affected by local conditions and search volume data can be a valuable complementary tool in studying potential local variations in disease presentation. Given the numerous limitations of laboratory testing, search data are a valuable complementary source for population-scale tracking of pandemics in real time.

Principal Results

This study showed that the data obtained from Google Trends, on keyword searches in Spanish related to COVID 19, 'coronavirus', 'neumonía', 'fiebre', 'tos', 'pérdida de olfato', 'dolor de garganta' and 'diarrea', correlated with data published by the CNE on the daily incidence of laboratory-confirmed cases by PCR, hospitalization, ICU admissions, and deaths from COVID 19. Going from R = 0.636 for 'fatigue' to a maximum of R = 0.908 for 'fever'. We also found that the GT data correlated with the daily incidence of COVID 19 with an 11 day time lag.

It should be noted that for ‘fatigue', the day with the highest correlation is day 36 (the sixth week after the search). Statistically, it is moderately relevant, but considering the high variability in the number of days in incubation, pathogenesis, and generation of an immune response in this disease, it may not be so evident when evaluating possible future positive cases. Consequently, it is possible that fatigue should may not be considered as a symptom to assess and predict a positive case using Google Trends.

Although we used correlations to examine the possible linear association between search queries and daily incidence, it should be noted that use of a search engine is voluntary and self-initiated search queries represent the users who are truly curious or worried about a situation. Thus, we believe that the unobtrusive search behavior of netizens may have resulted in an increase in search volume. The analysis and methods used in this study could aid public health and communication agencies. It would be crucial to study this same case in the rest of Europe since other countries, such as Italy, Great Britain and France have been affected by the COVID 19 pandemic, and new waves of it are foreseeable as long as social distancing measures get relax and the winter cold reenter.

This work presents the need for a detailed survey that provides both view of the COVID-19 in terms of its clinical features, prevention strategies, and the technological solutions including search engines data have been at the forefront. Findings from this study validate and extend previously published works that used Google keywords [5, 6, 8] and we demonstrate it potential for monitoring and prediction. Using Google Trends, the present study identified that there is a growing interest in COVID 19 globally and in countries with a higher incidence of the virus.

Limitations

Our study used Google Trends, which only provides the search behavior of people using the Google search engine. Future studies should consider studying the same topic on other search engines platforms to capture a more diverse population of users. The use of an automated program [24] can improve the accuracy of the data collected and analyzed in countries with a higher incidence of the virus.

The selection of keywords plays a very important role in ensuring the validity of the result.

Preprint JMIR Publications

Taking into account that this field of research is relatively new, there is no standard way of reporting, resulting in the same meaning of different terms, different meanings of the same term, and different abbreviations.

Search data may be subject to socio-economic, geographic, or other biases inherent in the local digital divide.

Lastly, Google Trends do not provide information about the methods used to generate search data and algorithms. Other search engines should be investigated. The transfer of conclusions to countries with a low level of internet access, should be done with caution.

Comparison with Prior Work

Using Google Trends, the present study identified that there is a growing interest in COVID 19 globally (and in countries with a higher incidence of the virus). The present work is consistent with previous studies such as those listed in Table 1 since all of them find a positive correlation between searches related to COVID 19 and the evolution of the pandemic. Furthermore, the correlation lag mode off the series is within the range of our findings.

Conclusions

Further research would be necessary to determine if this lag detected in our study has something to do with the results of clinical studies that postulate that 97.5% of symptomatic cases develop within 11.5 days after exposure [23].

This 11.5-day adjustment is an improvement over the initial case adjustment date of 15 days. In fact, in the second wave of the pandemic, 10-day quarantine has been considered sufficient in many places. For visualization purposes, a 10-day moving average provided slightly clearer plots.

Another priority in the early stages of an emergent pandemic such as COVID 19 is to characterize the clinical course of symptoms in affected individuals. It would thus be beneficial to pandemic tracking, case diagnosis and treatment if these clinical patterns could be ascertained earlier and at population-scale. We then investigated whether Internet search data can be used to characterize the clinical course of COVID 19 symptoms over time, providing a search-data-based view of the clinical course of illness.

Besides, for future studies could be useful to use Pytrends [24], a simple interface for automating downloading of reports from Google Trends.

Beyond the first attempts developed [11], a methodology must be advanced to approach this type of study in a systematic way.

For countries where the inflection curve has not yet occurred, this type of approach can be most useful if governments monitor the evolution of Google queries in their country to foresee the use of their hospital systems.

Others

Acknowledgments

The authors would like to acknowledge the contributions by César Martín Pérez in the study. And highly value the collaboration and support of Amparo Larrauri, Marina Pollán, Pedro Arias,

Preprint JMIR Publications

Rosa Cano and other colleagues from the CNE, who with their hard work in adverse conditions, have provided data on the prevalence of the epidemic in Spain and especially to all the workers of the National Health System of Spain who have saved so many lives and who have facilitated control over this pandemic. This work was supported in part by the project PI19CIII/00056 – TMPY 508/19, funding from Sub-Directorate-General for Research Assessment and Health Promotion in Spain (Instituto de Salud Carlos III).

The statements made herein are solely the responsibility of the authors.

Conflicts of Interest

None.

Abbreviations

CDC: Centers for Disease Control and Prevention.

CNE: Centro Nacional de Epidemiología (National Center of Epidemiology).

GT: Google Trends.

ICU: Intensive Care Units.

PCR+: Polymerase Chain Reaction positive results.

RSV: Google Trends Relative Search Volume.

UNED: Universidad Nacional de Educación a Distancia (The National Distance Education University).

WHO: World Health Organization.

References

1. Chen S, Yang J, Yang W, Wang C, & Bärnighausen T, 2020. COVID-19 control in China during mass population movements at New Year. The Lancet, 395(10226), 764-766. DOI: 10.1016/S0140-6736(20)30421-9. PMID: 32105609.

2. World Health Organization, 2020. WHO Coronavirus disease 2019 (COVID-19) Situation Report – 94. https://www.who.int/docs/default-source/coronaviruse/situation-reports/20200423-sitrep-94-covid-19.pdf. Accessed on September, 6, 2020.

3. World Health Organization, 2020. WHO Coronavirus disease (COVID-19) pandemic. https://www.euro.who.int/en/health-topics/health-emergencies/coronavirus-covid-19. Accessed on September 6, 2020.

4. World Health Organization, 2020. WHO Coronavirus Disease (COVID-19) Dashboard. https://covid19.who.int/. Accessed on June, 10, 2020.

5. Chu B & Qureshi S, 2020. Predicting the COVID-19 Pandemic in Canada and the US. Carleton Economic Papers 20-05, Carleton University, Department of Economics. https://ideas.repec.org/p/car/carecp/20-05.html. Accessed on 30 Jul 2020.

6. Imai N, Dorigatti I, Cori A, Riley S & Ferguson N, 2020. Report 1: Estimating the potential total number of novel coronavirus (2019-nCoV) cases in Wuhan City, China. URL

preprint JMIR Publications

https://www.imperial.ac.uk/media/imperial-college/medicine/sph/ide/gida-fellowships/Imperial-College-COVID19-update-epidemic-size-22-01-2020.pdf. Accessed on September 12, 2020.

7. Hannan A, Niemi J, House K, Reich NG, Youyang G, Shanghong X et al., 2020. 'The COVID Forecast Hub,' Web application. https://github.com/reichlab/covid19-forecast-hub. Accessed on September 12, 2020.

8. Eysenbach G, 2006. Infodemiology: tracking flu-related searches on the web for syndromic surveillance, in: AMIA ... Annual Symposium Proceedings / AMIA Symposium. AMIA Symposium. American Medical Informatics Association, pp. 244–248. PMID: 17238340.

9. Ginsberg J, Mohebbi MH, Patel RS, Brammer L, Smolinski MS, Brilliant, L, 2009. Detecting influenza epidemics using search engine query data. Nature 457, 1012–1014. DOI: 10.1038/nature07634. PMID: 19020500.

10. Epstein JM, Parker J, Cummings D, Hammond RA, 2008. Coupled contagion dynamics of fear and disease: Mathematical and computational explorations. PLoS One 3. DOI: 10.1371/journal.pone.0003955. PMID: 19079607.

11. Mavragani A, Ochoa G, 2019. Google trends in infodemiology and infoveillance: Methodology framework. J. Med. Internet Res. 21, e13439. DOI: 10.2196/13439. PMID: 31144671.

12. Li C, Chen LJ, Chen X, Zhang M, Pang CP, Chen H, 2020. Retrospective analysis of the possibility of predicting the COVID-19 outbreak from Internet searches and social media data, China, 2020. Eurosurveillance 25, 2000199. DOI: 10.2807/1560-7917.ES.2020.25.10.2000199. PMID: 32183935.

13. Husnayain A, Fuad A, Su ECY, 2020. Applications of Google Search Trends for risk communication in infectious disease management: A case study of the COVID-19 outbreak in Taiwan. Int. J. Infect. Dis. 95, 221–223. DOI: 10.1016/j.ijid.2020.03.021. PMID: 32173572.

14. Effenberger M, Kronbichler A, Shin JIl, Mayer G, Tilg H, Perco P, 2020. Association of the COVID-19 pandemic with Internet Search Volumes: A Google TrendsTM Analysis. Int. J. Infect. Dis. 95, 192–197. DOI: 10.1016/j.ijid.2020.04.033. PMID: 32305520.

15. Pekoz EA, Smith A, Tucker A, Zheng Z, 2020. COVID-19 Symptom Web Search Surges Precede Local Hospitalization Surges. SSRN Electron. J. DOI: 10.2139/ssrn.3585532.

16. Yuan X, Xu J, Hussain S, Wang H, Gao N, Zhang L, 2020. Trends and Prediction in Daily New Cases and Deaths of COVID-19 in the United States: An Internet Search-Interest Based Model. Explor. Res. Hypothesis Med. 000, 1–6. DOI: 10.14218/erhm.2020.00023. PMID: 32348380.

17. Higgins TS, Wu AW, Sharma D, Illing EA, Rubel K, Ting JY, 2020. Correlations of Online Search Engine Trends with Coronavirus disease (COVID-19) Incidence: Infodemiology Study (Preprint). JMIR Public Heal. Surveill. 6, e19702. DOI: 10.2196/19702. PMID: 32401211.

18. Lu T, Reis BY, 2020. Internet Search Patterns Reveal Clinical Course of Disease Progression for COVID-19 and Predict Pandemic Spread in 32 Countries. medRxiv 2020.05.01.20087858. DOI: 10.1101/2020.05.01.20087858.

Preprint JMIR Publications

19. Google. FAQ about Google Trends data. https://support.google.com/trends/answer/4365533?hl=en. Accessed on September, 6, 2020.

20. Jimenez A, Santed-Germán MA, Ramos V, 2020. Google Searches and Suicide Rates in Spain, 2004-2013: Correlation Study. JMIR Public Heal. Surveill. 6, e10919. DOI: 10.2196/10919. PMID: 32281540.

21. Instituto de Salud Carlos III, 2020. CNE COVID-19 Dashboard. https://cnecovid.isciii.es/covid19/#documentaci%C3%B3n-y-datos. Accessed on May, 24, 2020.

22. Wilkinson M, 2013. Testing the null hypothesis: The forgotten legacy of Karl Popper? J. Sports Sci. 31, 919–920. DOI: 10.1080/02640414.2012.753636. PMID: 23249368.

23. Sanitarias, C. de C. de A. y E, 2020. Información científica-técnica, enfermedad por coronavirus, COVID-19. https://www.mscbs.gob.es/profesionales/saludPublica/ccayes/alertasActual/nCov-China/documentos/ITCoronavirus.pdf. Accessed on June, 10, 2020.

24. Hogue J, Nescio N et al. 2020. GeneralMills/Pytrends. Web application. https://github.com/GeneralMills/pytrends. Accessed on September, 12, 2020.

Preprint JMIR Publications

Supplementary Files

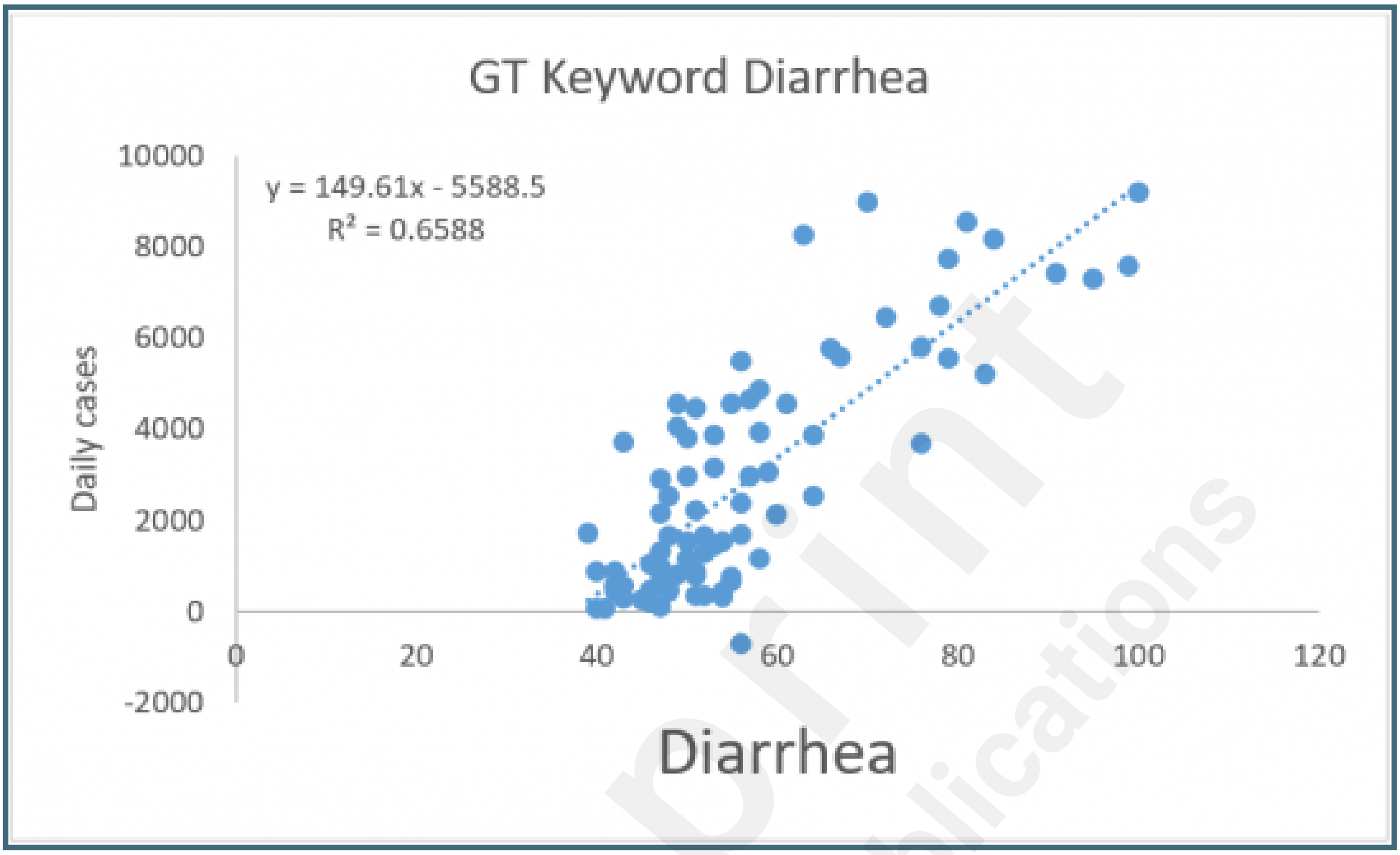
Untitled.
URL: https://asset.jmir.pub/assets/c0ea6e3ae1b33455d5fe991b6abcd6ff.xlsx

Preprint
JMIR Publications

Figures

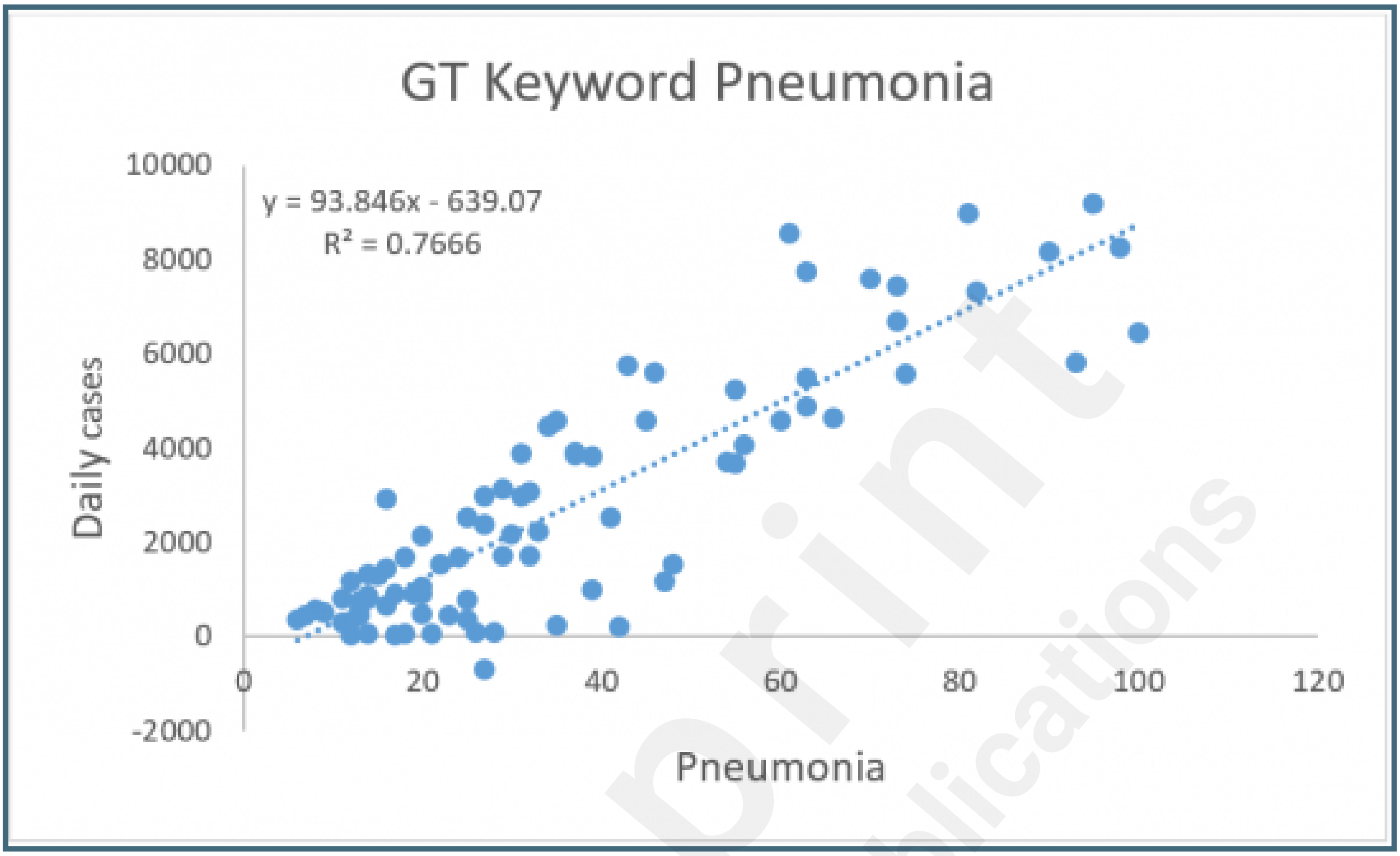
Preprint JMIR Publications

Linear trend in each of the relationships 7/8.



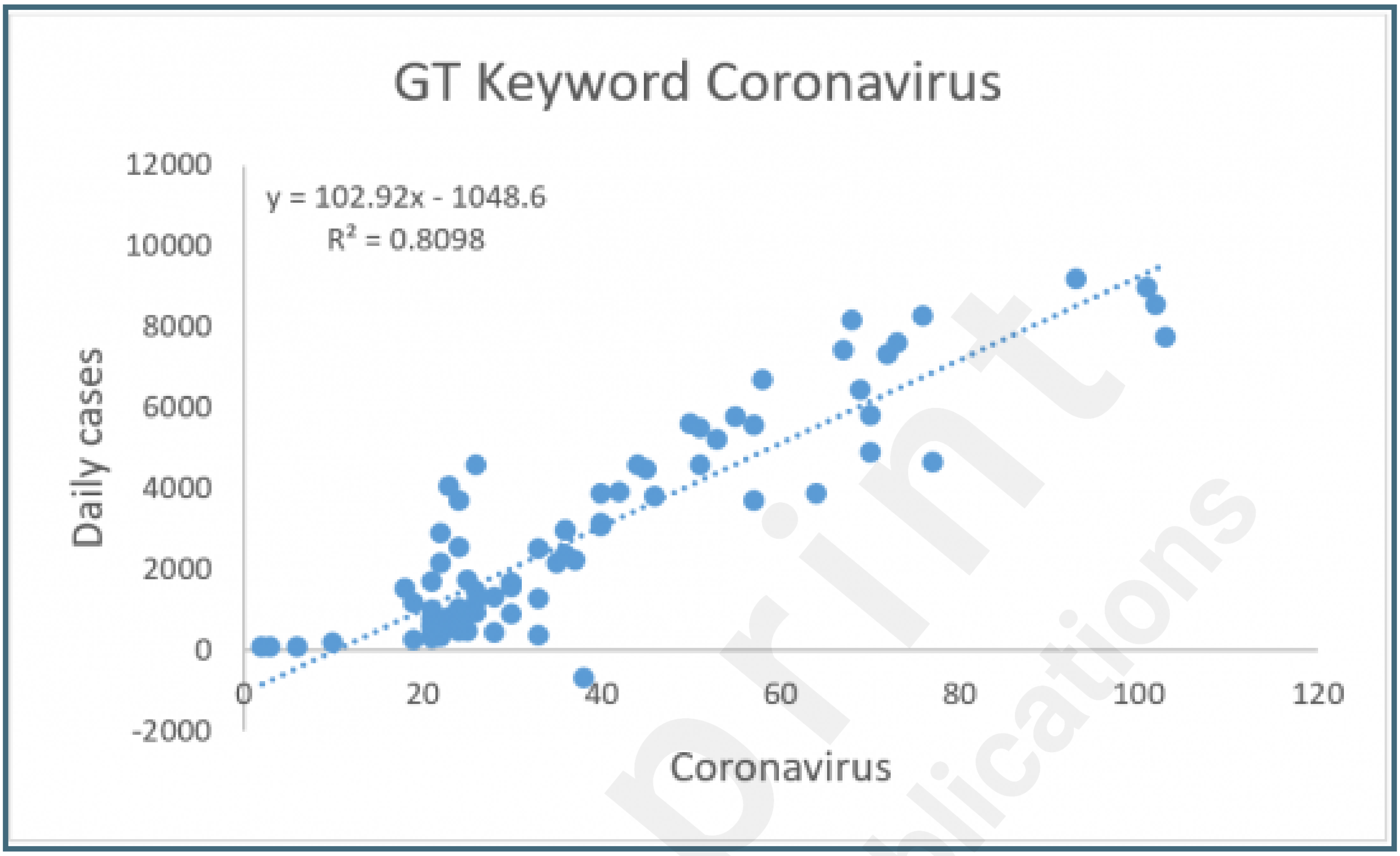
Preprint JMIR Publications

Linear trend in each of the relationships 2/8.



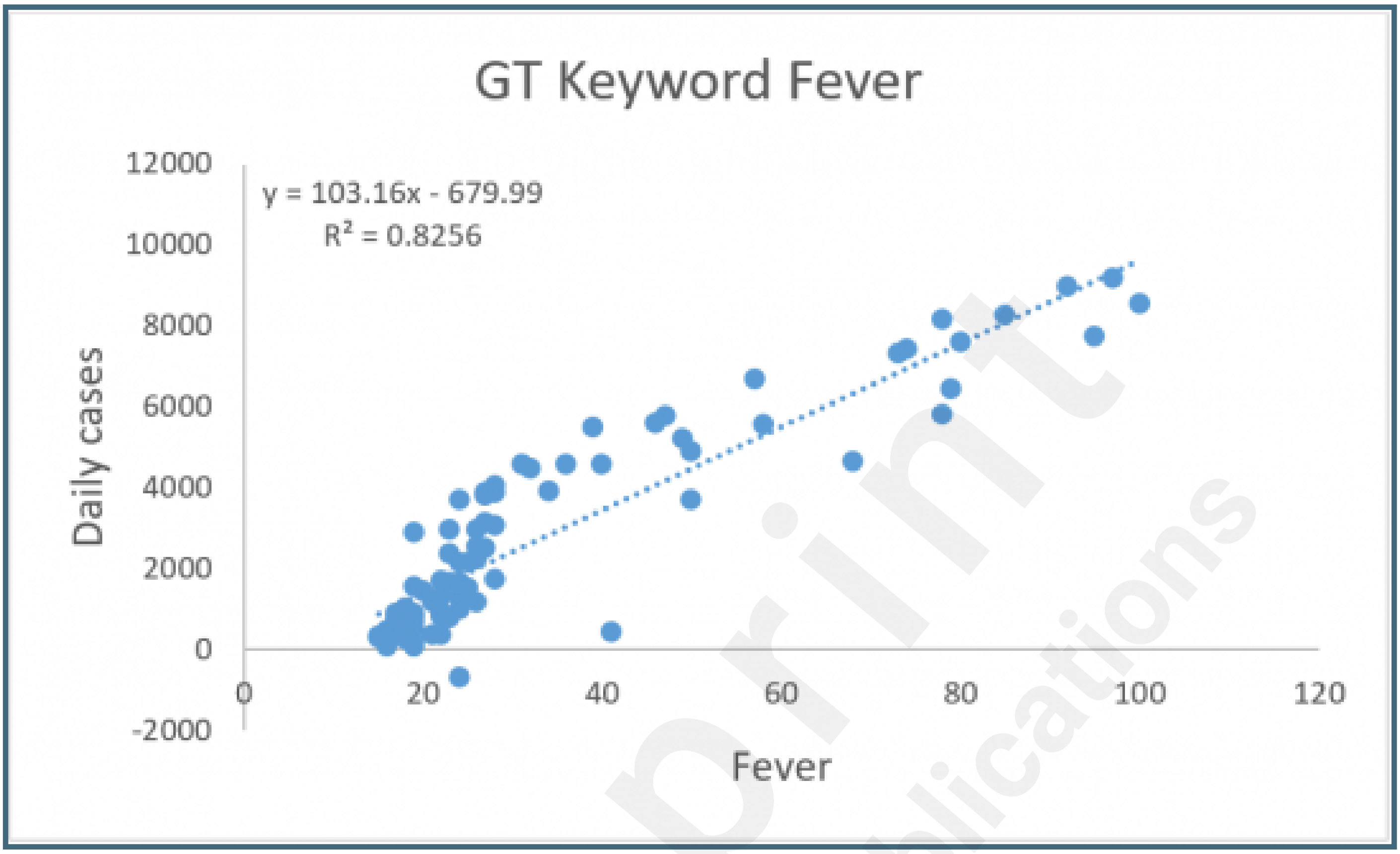
Preprint JMIR Publications

Linear trend in each of the relationships 1/8.



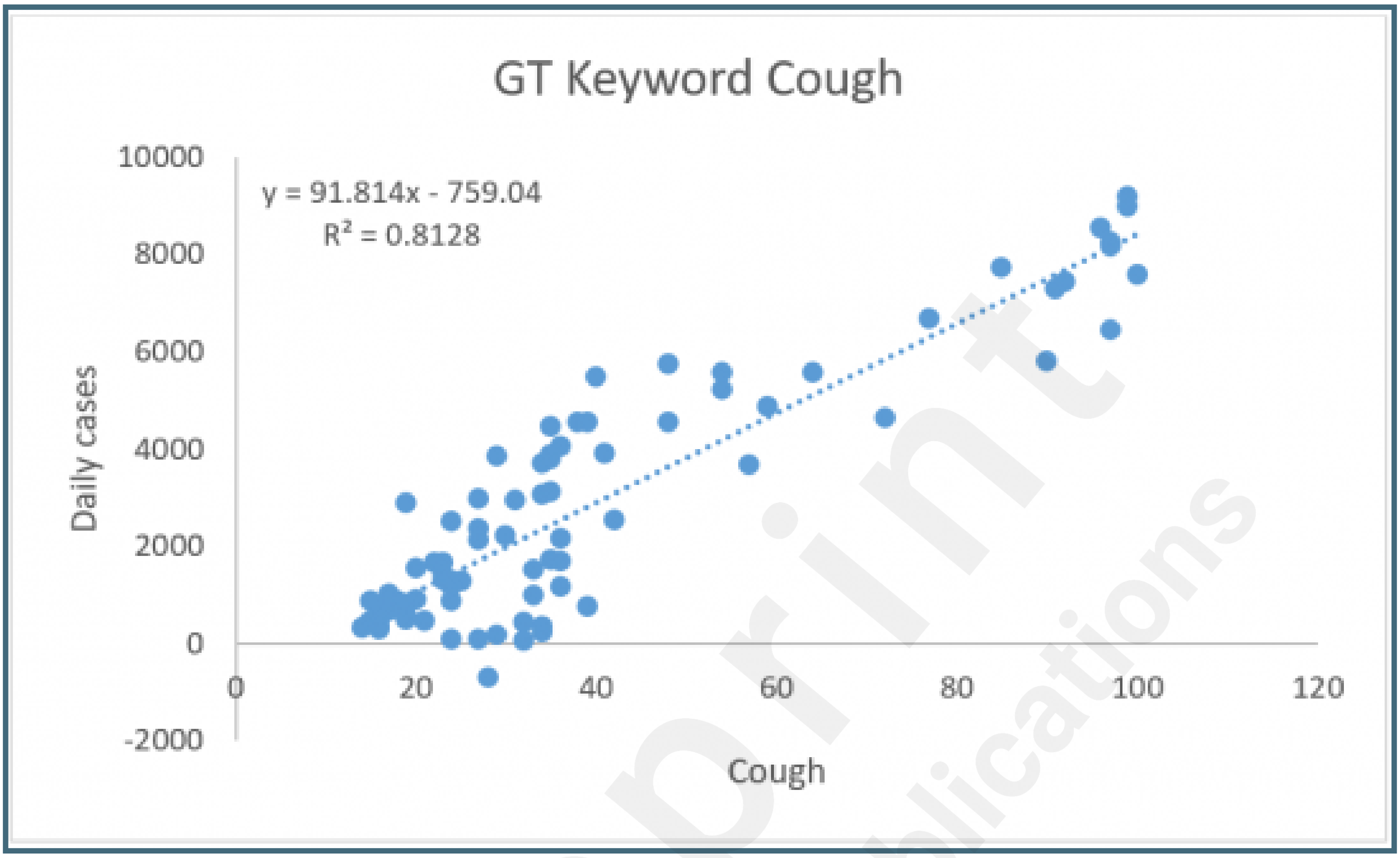
Preprint JMIR Publications

Linear trend in each of the relationships 3/8.



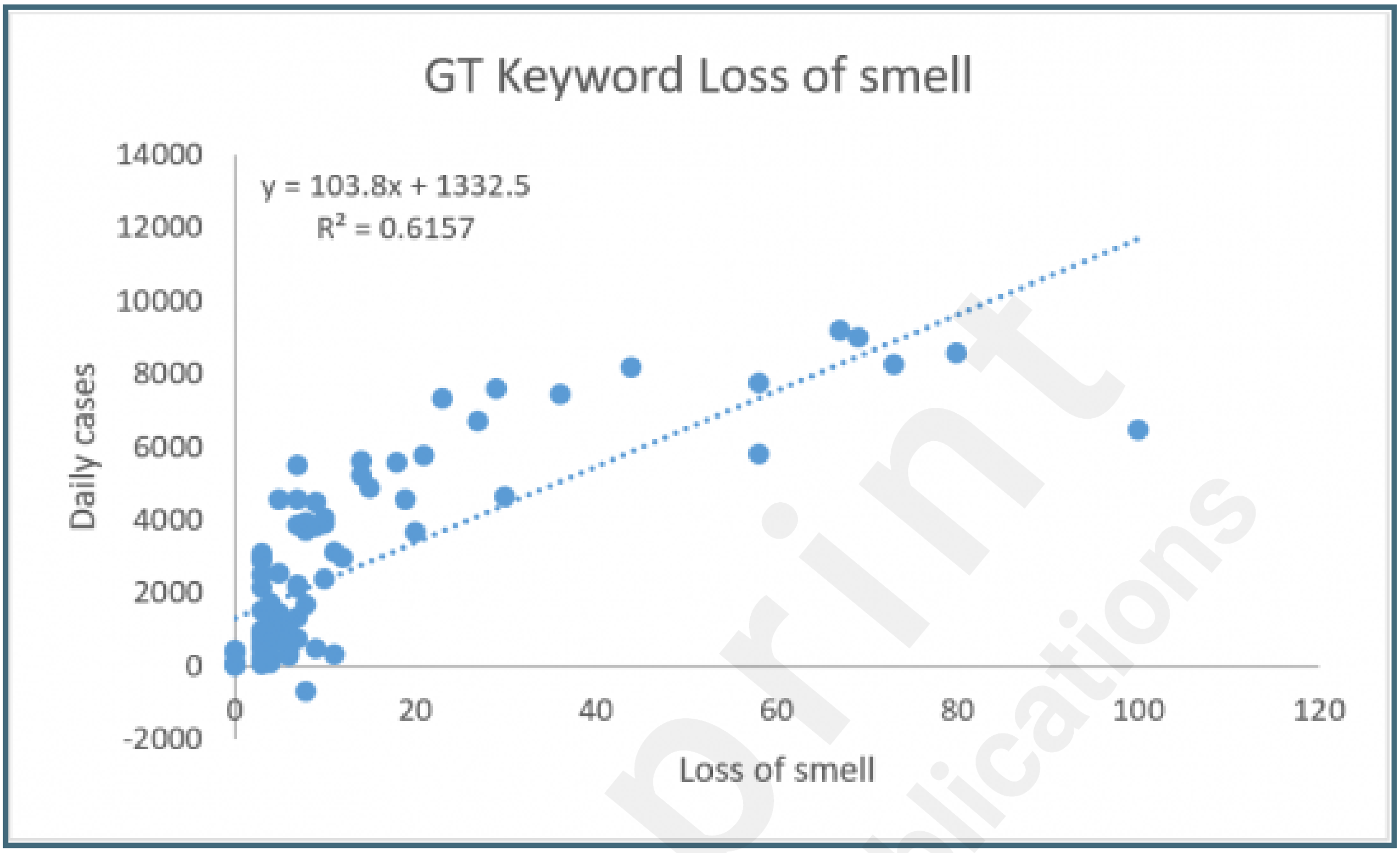
Preprint JMIR Publications

Linear trend in each of the relationships 4/8.



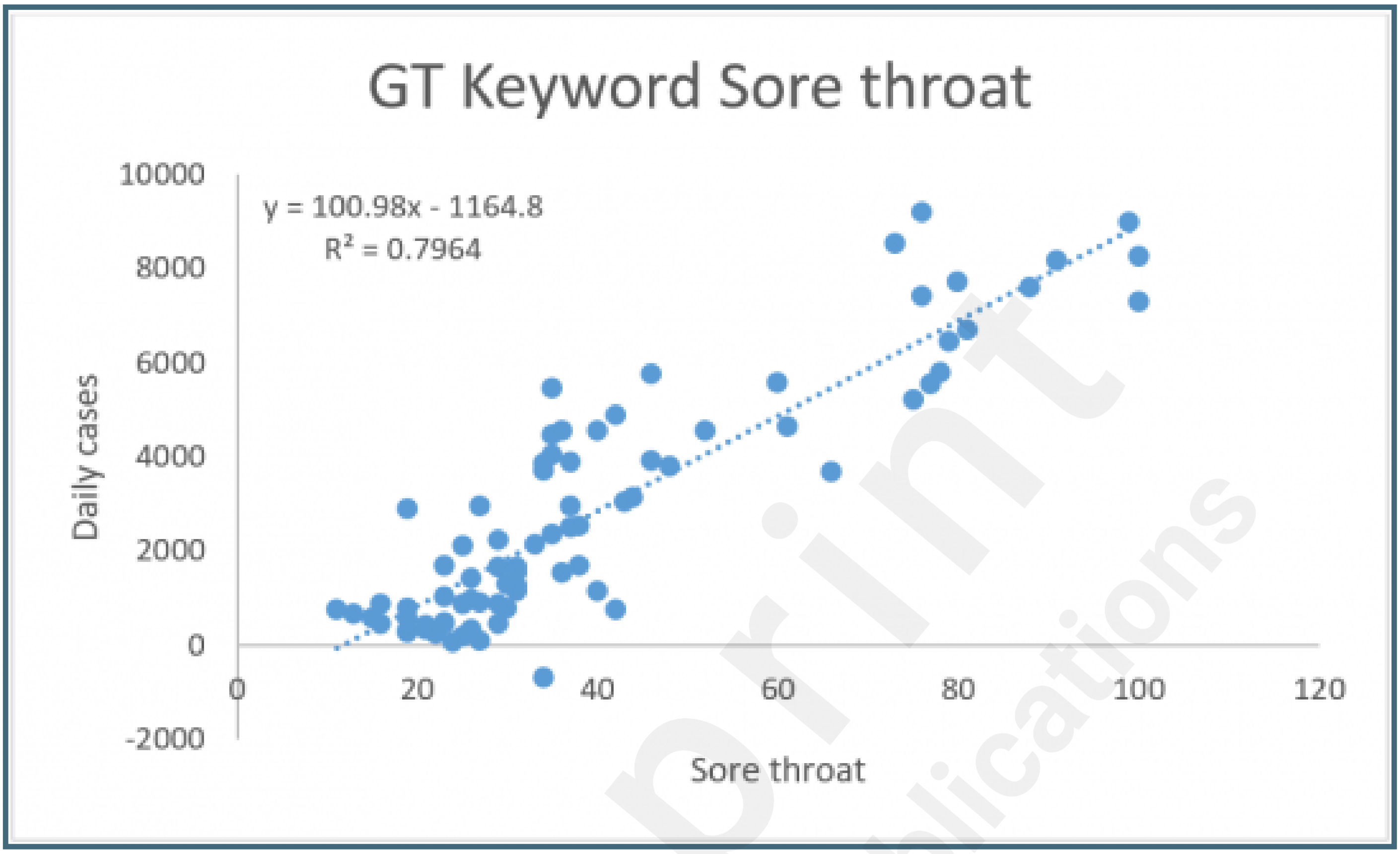
Preprint JMIR Publications

Linear trend in each of the relationships 5/8.



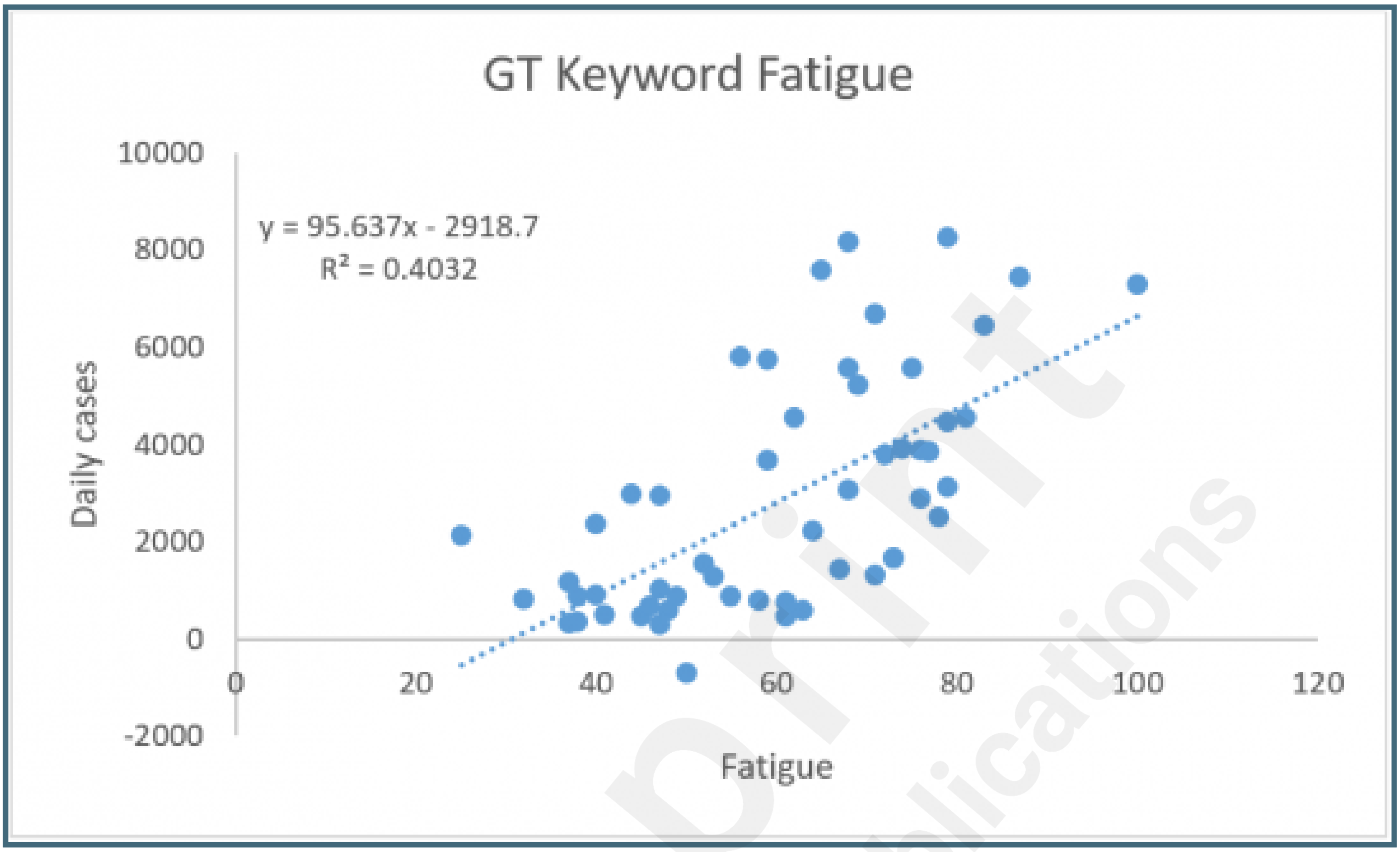
Preprint JMIR Publications

Linear trend in each of the relationships 6/8.



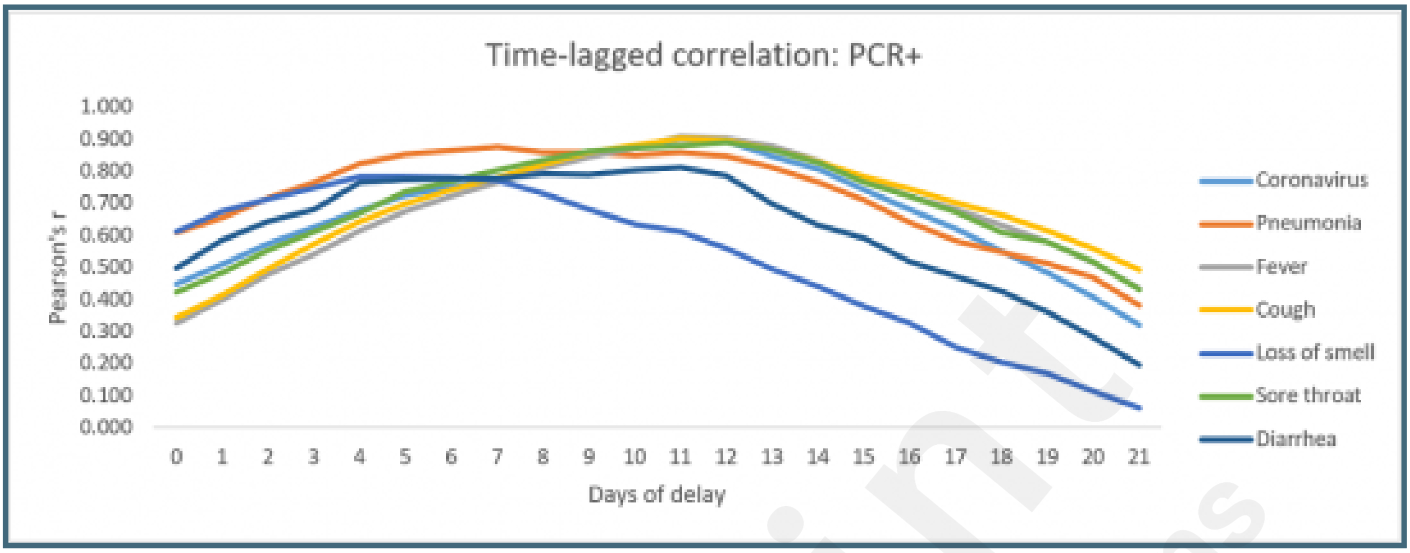
Preprint JMIR Publications

Linear trend in each of the relationships 8/8.



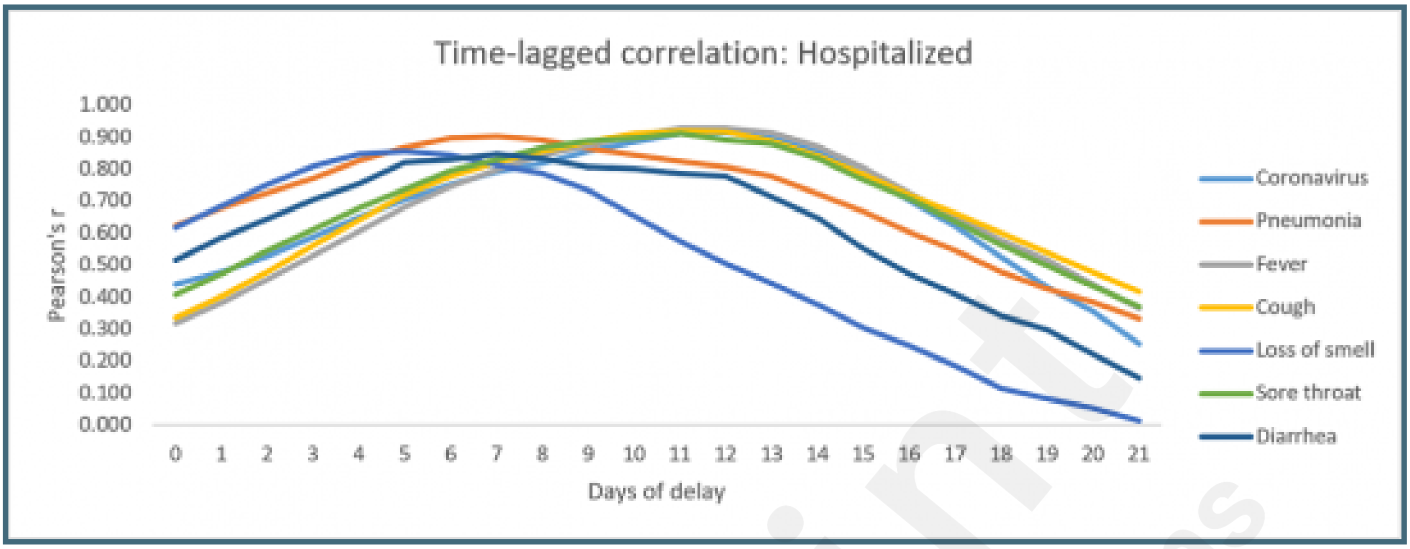
Preprint JMIR Publications

Day with the best correlation between symptom keyword-positive case 1/4.



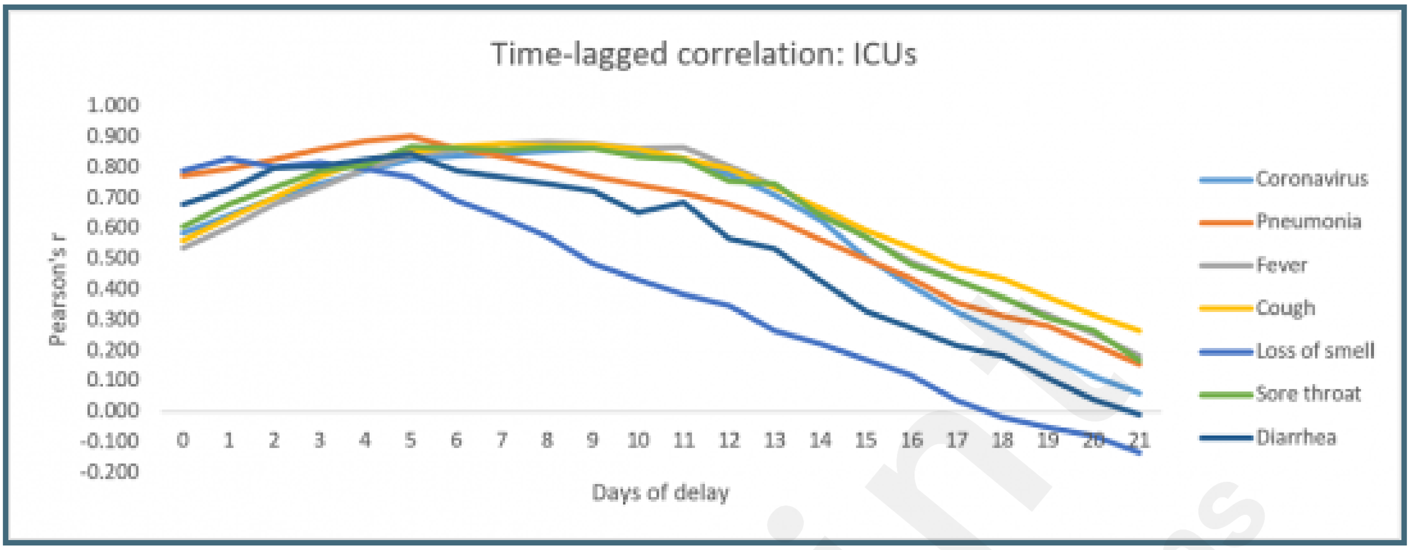
Preprint JMIR Publications

Day with the best correlation between symptom keyword-positive case 2/4.



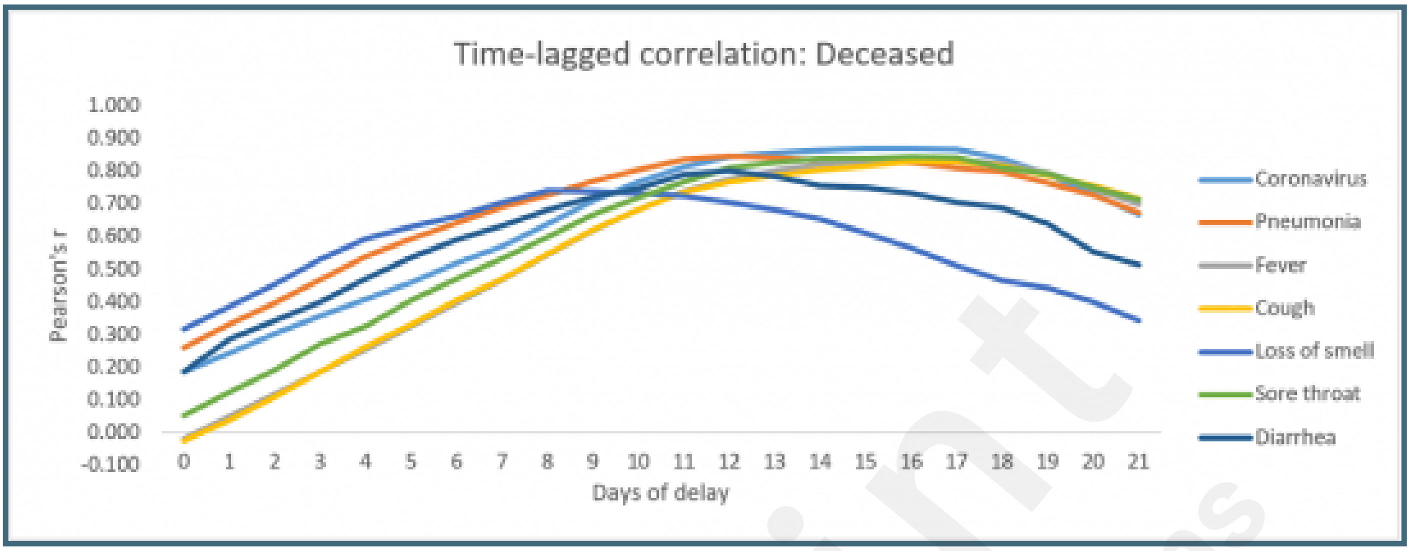
Preprint JMIR Publications

Day with the best correlation between symptom keyword-positive case 3/4.



Preprint JMIR Publications

Day with the best correlation between symptom keyword-positive case 4/4.



Preprint JMIR Publications

Multimedia Appendixes

Preprint JMIR Publications

Screenshot for study "Individuals' concerns, predict the spread of the coronavirus (COVID 19): the case of Spain"; alternative title "Association of Google search data with the incidence of the COVID 19: evidence from Spain". Alberto Jimenez, MA ; Rosa M. Estevez-Reboredo, PhD ; Miguel A. Santed, PhD ; Victoria Ramos, PhD.
URL: https://asset.jmir.pub/assets/8311658f4bc5b60aad2d2e597092af64.png

Data sets for study "Individuals' concerns, predict the spread of the coronavirus (COVID 19): the case of Spain"; alternative title "Association of Google search data with the incidence of the COVID 19: evidence from Spain". Alberto Jimenez, MA ; Rosa M. Estevez-Reboredo, PhD ; Miguel A. Santed, PhD ; Victoria Ramos, PhD.
URL: https://asset.jmir.pub/assets/36db2fd6c35fe8dceb0313c220ee0f84.xlsx

Preprint JMIR Publications

TOC/Feature image for homepages

Preprint JMIR Publications

Girl with mask and smartphone.

Preprint
JMIR Publications