Title: Predicting social distancing index during COVID-19 outbreak through online search engines trends

P. C. Lins-Filho¹, M. M. S. Araújo¹, T. S. Macêdo¹, A. K. A. Ferreira¹, M. C. F. Melo¹, E. L. M. S. Silva¹, J. L. M. Freitas¹, A. F. Caldas Jr¹,²

¹ Universidade Federal de Pernambuco, Recife, Pernambuco, Brazil
² Universidade de Pernambuco, Camaragibe, Pernambuco, Brazil

Corresponding Author:

PhD Arnaldo de França Caldas Jr

Address for correspondence: Estrada de Aldeia, Km 13, Prive Portal de Aldeia, Aldeia, Camaragibe, Pernambuco, Brasil. Tel: +55 (81) 999713652 E-mail: caldasjr@alldeia.com.br

Running Title: Predicting SDI with online search volume
Summary

Online-available information has been considered an accessory tool to estimate epidemiology and collect data on diseases and population behavior patterns. This study aimed to explore the potential use of Google and YouTube relative search volume to predict social distancing index in Brazil during COVID-19 outbreak and verify the correlation between social distancing measures with the course of the epidemic. Data concerning the social distancing index, epidemiological data on COVID-19 in Brazil and the search engines trends for “Coronavirus” were retrieved from online databases. Multiple linear regression was performed and resulted in a statistically significant model evidencing that Google and YouTube relative search volumes are predictors of the social distancing index. The Spearman correlation test revealed a weak correlation between social distancing measures and the course of the COVID-19 epidemic. Health authorities can apply these data to define the proper timing and location for practicing appropriate risk communication strategies.

Keywords: COVID-19; Social Isolation; Internet; Consumer Health Information.
Introduction

The World Health Organization have recently declared South America as the new coronavirus epicenter, mainly because of the situation in Brazil that registers the most cases and deaths in Latin America [1]. Given the COVID-19 pandemic, robust risk communication is urgently needed particularly in the most affected countries [2]. Internet query platforms, which allows to interact with internet-based data, have been considered a source of potentially useful and accessible resources, especially aimed to identify outbreaks and implement intervention strategies [3, 4]. Online-available information has been considered as a surrogate tool for estimating epidemiology and gathering data about patterns of disease and population behavior [5-7].

As the online queries on COVID-19 increases globally reflecting the interest of people to be aware about this emerging infectious disease, mining online data and search patterns on electronic resources might provide a better support to manage this worldwide health crisis [8]. Internet searches and social media data have been reported to correlate with traditional surveillance data and can even predict the outbreak of disease epidemics several days or weeks earlier [9]. Evidence points that Google Trends could potentially define the proper timing and location for practicing appropriate risk communication strategies for affected populations and be employed to predict outbreak trends of the novel coronavirus [2, 5, 10].

Previous investigations reported the use of Internet search engines as source of data for public health surveillance and diseases incidence prediction worldwide, as zika, in Brazil and Colombia [11]; influenza, in the United States [12]; malaria, dengue fever and chikungunya, in India [13]; and Middle East respiratory syndrome, in Korea [14]. However, there are no reports on assessing the relative search volume (RSV) of search engines to predict social distancing behavior during infectious diseases outbreaks.
Predictions might support in health resource management and planning for prevention purposes [5]. As COVID-19 treatment protocol is still uncertain it is especially important to prevent the virus dissemination in society [15]. Currently, the virus spread prevention approaches focus on hand hygiene, social distancing and quarantine [16]. Social distancing is designed to reduce interactions between people in a broader community, in which individuals may be infectious but have not yet been identified hence not yet isolated. This measure is particularly useful in settings where community transmission is believed to have occurred, but where the linkages between cases is unclear, and where restrictions placed only on persons known to have been exposed is considered insufficient to prevent further transmission [15, 17, 18]. Collective infection control measures can reduce the disease incidence, though at the price of a prolongation of the epidemic period [19]. Therefore, it is important to raise information on these measures.

A recent study that assessed the impact of online information on the individual-level intention to voluntarily self-isolate during the pandemic concluded that in order to enhance individuals' motivation to adopt preventive measures such as social distancing, actions should focus on raising consciousness on the severity of the situation, in addition, information overload had a significant impact on individuals' threat and coping perceptions, and through them on self-isolation intention [20]. Thus, the aim of the present investigation was to predict social distancing index (SDI) through Google and YouTube search trends and investigate the correlation between the SDI with epidemiological data on COVID-19 outbreak in Brazil.

Methods

In Brazil, the current market share of Google among the search engines is over 97% [21]. Google Trends (https://trends.google.com/trends/) data is a randomly collected
sample of Google search queries, each piece of data is categorized and tagged with a topic. Each data point is divided by the total searches in a specific location over a time period to compare relative popularity. Google Trends portrays search frequency output as a normalized data series and the resulting numbers are scaled on a range of 0 to 100 based on a topic proportion to all searches on all topics. Scores represents search interest relative to the highest point on the graph for that time period and geographic location. A value of 100 is the peak popularity of a term. A value of 50 indicates that the term is half as popular as it was at its peak of popularity [22]. This same methodology was applied to YouTube search trends. To the present investigation the Brazilian Portuguese correspondent topic for “coronavirus”, which held most popularity, was used.

The Social Distancing Index was created to help combat the spread of COVID-19, since its launch it has been improved with the sole objective of providing an increasingly accurate data for public authorities and research institutes. In order to achieve the index, highly accurate geolocation data was treated with a distance algorithm. Polygons from all regions of the Brazilian Institute of Geography and Statistics were adopted in order to ensure a more accurate categorization and more reliable data [23]. Data is available on Inloc website, displayed as map and chart.

Epidemiological data concerning COVID-19 outbreak in Brazil were collected from the Brazilian government Health Ministry database, available online [24]. Statistical data on daily new cases, cumulative number of cases, cumulative number of deaths and recovered cases were retrieved.

All databases were assessed for data collection on 23 May 2020, and the information corresponds to the period from 23 February to 20 May 2020.
Data were submitted to statistical analysis, all tests were applied considering an error of 5% and the confidence interval of 95%, and the analyzes were carried out using SPSS software version 23.0 (SPSS Inc. Chicago, IL, USA). Although the hypothesis of normal distribution of data was not confirmed by the Kolmogorov-Smirnov test, the statistical analysis was performed by the application of nonparametric tests. The strength of the association between distinct measures was tested with Spearman rank correlation. Multiple linear regression was performed to verify whether Google and YouTube relative search volumes are predictors of the social distancing index in Brazil.

**Results**

The multiple linear regression analysis resulted in a statistically significant model \[F (2,85) = 32.045; p<0.001; R^2=0.430\]. Therefore, Google RSV \((\beta = 1.226; t = 7.887; p<0.001)\) and YouTube RSV \((\beta = -0.930; t = -5.980; p<0.001)\) are predictors of the social distancing index in Brazil. The equation that describes this relationship is \((SDI) = 34.347 + 0.422 \text{(Google RSV)} + (-0.359) \text{(YouTube RSV)}.\)

In Brazil for the time span analyzed the mean SDI score was approximately 43%, the maximum of social distancing observed during this period was 62.2%. In mean scores over 3312 new cases of COVID-19 were confirmed daily. In the moment of data collection, the reported total number of cumulative deaths, confirmed cases and recovered cases were 18859, 291579 and 116683, respectively. The mean values of search engines RSV are shown in table 1.

Correlation between SDI and the other studied variables was found to be varying from weak to moderate, statistically significant correlation was found with all measures tested except for cumulative recovered cases, as shown in table 2.

**Discussion**
Evidence suggests that collective isolation measures have been highly effective in controlling the spread of the COVID-19 [16, 25]. However, maintaining isolation for many months may have even worse consequences than an epidemic wave that runs an acute course, the isolation measures should be thoughtfully planned and executed based on current stage of pandemic [26]. As observed in table 1 the mean score for SDI was 43.126 with a discrete standard deviation when compared with the standard deviation of the daily new cases. A weak correlation was found between the isolation measure with epidemiological data from COVID-19, this may represent that in Brazil social isolation measures are poorly associated with the course of the disease in the country. In addition, these findings may be correlated with failure to control the increase in cases that were added daily, on average, by 3312.26 new cases, during the period covered by this investigation. The absence of concise social distancing policies and, furthermore, the political instability at the center of the Brazilian government poses a deadly distraction in the middle of a public health emergency [1].

The weak correlation between social distancing and the course of the disease can also be observed in the time series pattern seen in figure 1 (section A), where while the number of cases per day and cumulative deaths show an ascending pattern, the social distancing index remains with slight changes, with the exception of a slight increase at the end of March, when cases start to show an ascending profile. At the end of the time series, when the daily new cases are at peak, SDI was under de mean observed in the evaluated period.

A correlation was found between search engines RSV and SDI (table 2). To further investigate this correlation a multiple linear regression was performed, data extracted from this analysis showed that Google RSV and YouTube RSV are predictors to social distancing. In figure 1 (section B) it is possible to observe de matching behavior of this
measures along the time series, with a similar pattern of peaks and decrease starting at the end of April.

The positive correlation found with Google RSV may be associated with the access to information, since raising awareness on the severity of the situation and the importance on following the advice from health organizations is a key point on achieving self-isolation intention [20].

The negative association observed between YouTube RSV and SDI through de multiple linear regression may correlate with the low quality of YouTube content on COVID-19 reported in previous studies [27-29], since misinformation can hinder the communication of health professionals and organizations with general public and even reduce compliance with treatments or medical advices [30, 31].

The findings of the present study support the evidence that online-available information can potentially assist conventional epidemiologic tools for estimating data about patterns of disease and population behavior [5, 6]. Relative search volume of Google and YouTube could define the proper timing and location for practicing appropriate risk communication strategies. Health authorities might apply these data to measure the effect of the transmission of information on the population and to obtain feedback from research statistics.

Acknowledgments

P.C., T.M., M.A. and E.L. were supported by a PhD scholarship from Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES).

Declaration of interest

None
Funding

No funding was received for this work.

References


(4) Bhattacharya S. Predicting emerging and re-emerging disease outbreaks through internet search trends: An analysis from India. AIMS Public Health 2019; 6: 1-3.


(17) Wilder-Smith A, Chiew CJ, Lee VJ. Can we contain the COVID-19 outbreak with the same measures as for SARS?. *The Lancet Infectious Diseases* 2020; 20: e102-e107.


Table 1. Frequency measures of SDI, search engines RSV and COVID-19 daily new cases for the evaluated period

<table>
<thead>
<tr>
<th></th>
<th>Mean (SD)</th>
<th>Median</th>
<th>Minimum</th>
<th>Maximum</th>
</tr>
</thead>
<tbody>
<tr>
<td>Social distancing index</td>
<td>43.126 (8.130)</td>
<td>43.30</td>
<td>26.60</td>
<td>62.20</td>
</tr>
<tr>
<td>Google RSV</td>
<td>37.470 (23.649)</td>
<td>34.50</td>
<td>2</td>
<td>100</td>
</tr>
<tr>
<td>YouTube RSV</td>
<td>19.520 (21.035)</td>
<td>11.50</td>
<td>2</td>
<td>100</td>
</tr>
<tr>
<td>Daily new cases</td>
<td>3312.26 (4564.24)</td>
<td>1180</td>
<td>0</td>
<td>19951</td>
</tr>
</tbody>
</table>

*RSV – Relative Search Volume

Table 2. Correlation between SDI measures, COVID-19 epidemiological data and search engines RSV

<table>
<thead>
<tr>
<th></th>
<th>Social distancing index</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Google RSV</td>
<td>p0.567</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>YouTube RSV</td>
<td>p0.367</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Daily new cases</td>
<td>p0.370</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Cumulative cases</td>
<td>p0.391</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Cumulative deaths</td>
<td>p0.396</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Recovered cases</td>
<td>p0.650</td>
<td>0.547</td>
</tr>
</tbody>
</table>

*RSV – Relative Search Volume

Mann-Whitney U test
Figure 1. Time series of social distancing index related to Google and YouTube RSVs and COVID-19 new cases and cumulative deaths in Brazil.
A. SDI related to COVID-19 daily new cases and cumulative deaths

B. SDI related to search engines RSVs