Estudio de la búsqueda de información sobre la pandemia SARS-CoV-2 en Galicia

Study of the information search behaviour on the SARS-CoV-2 pandemic in Galicia

Miguel Mayo-Yáñez*[1,2,3], Lucía González-Torres[4], Christian Calvo-Henríquez[2,3,5], Carlos Chiesa-Estomba[3,6]

[1]Otorhinolaryngology – Head and Neck Surgery Department, Complexo Hospitalario Universitario A Coruña (CHUAC). [2]Clinical Research in Medicine, International Center for Doctorate and Advanced Studies (CIEDUS), Universidade de Santiago de Compostela (USC). [3]Young-Otolaryngologists of the International Federations of Oto-Rhino-Laryngological Societies (YO-IFOS) Study Group, Paris, France. [4]Pediatrics Department, Complexo Hospitalario Universitario A Coruña (CHUAC). [5]Otorhinolaryngology – Head and Neck Surgery Department, Complexo Hospitalario Universitario Santiago de Compostela (CHUS). [6]Otorhinolaryngology – Head and Neck Surgery Department, Hospital Universitario Donostia

ABSTRACT

Introduction: This manuscript analyses the use and evolution, through Google Trends as a source of information, of internet-based information-seeking behaviour related to the SARS-CoV-2 pandemic using the terms: Coronavirus, COVID-19, SARS-CoV-2 from January 1, 2020 to April 15, 2020.

Methods: A generalized linear model was used to analyse the relation between SARS-CoV-2 data epidemiology and the Search Volume Index of the terms obtained from the Google Trends query. Significant trend changes were assessed by Joinpoint methodology.

Results: A total of 7,873 SARS-CoV-2 confirmed cases were collected with an increase of 4.7% in the selected period. A relation was found between the confirmed cases (dependent variable) and the Search Volume Index of the *Coronavirus* term, with a correlation rho = 0.79 (p <0.000).

Conclusion: The analysis of search engine query data in order to create mathematical models that forecast disease spread could be useful and helpful to activate and improve strategic plan to control an outbreak.

Keywords: SARS-CoV-2; Google Trends; Correlation Data; Data Analysis; Infodemiology.

INTRODUCTION

To date, SARS-CoV-2 (COVID-19) has spread rapidly in 185 countries, with approximately 3,000,000 confirmed cases and over 200,000 deaths as of April 27, 2020.[1] Considering the global threat, the outbreak was declared a Public Health Emergency of International Concern on 30 January 2020 by the World Health Organization (WHO).[2] In the last few weeks, several scientific reports have been published on the epidemiology of the infection, the clinical course, laboratory testing and treating support.[3,4] However, there are no vaccines against SARS-CoV-2 or specific therapeutic drugs for this communicable disease.

The use of internet data has become an integral part of health informatics over the past decade. Big Data is associated with the massive computational resources needed to cope with the increasing volume and complexity of data from many sources. It has shown great potential in forecasting and better decision making, being continuously integrated into research with novel approaches and methods.[5–8] In this sense, studies have demonstrated that many people seek health information from internet sources.[9] Data from Google Trends have been shown to be valuable to monitor health information-seeking behaviour trends, epidemiology, aetiology and management of specific health conditions, predictions or detection of outbreaks.[10,11] Scrutinising such information now constitutes a new research discipline termed "Infodemiology", which is defined as "the science of distribution and determinants of information in an electronic medium, specifically the Internet, or in a population, with the ultimate aim to inform public health and public policy".[12]

It seems clear that we need urgently to develop further public health activities in order to better understand the epidemiology of the novel virus and characterize its potential impact on public health. Given the actual situation, and taking advantage of these new technologies of "Infodemiology",[12] the main objective of this study is to analyze the use and evolution, through Google Trends as a source of information, of internet-based information-seeking behaviour related to the COVID-19 pandemic in the territory of the autonomous community of Galicia (Spain). The secondary objective is to evaluate the relationship between COVID-19 searches and related mass media stories.

METHODS

Google Trends Tool

Google Trends provides access to Internet search patterns by analyzing a portion of all web queries on the Google Search website and other affiliated Google sites.[10] It analyses a sample of the billions of daily search results and provides information on geospatial and temporal patterns in search volumes for user-specified terms.

Google Trends creates a "Search Volume Index" (SVI), which represents the relative search volume for a search term indexed against the overall search volume. Search volume index values are adjusted to a normalised data scale of 0–100 based on a topic's proportion to all searches on all topics in order to reduce data redundancy and improve data integrity.[13] Different regions that show the same search interest for a term don't always have the same total search volumes.

Users are able to download the output of their searches to conduct further analyses. The value 0 does not necessarily indicate no searches, but rather indicates very low search

volumes that are not included in the results. The adjustment process also excludes queries that are made over a short time frame from the same internet protocol address and queries that contain special characters.

Data collection

Data were collected using a Windows®10 (Microsoft®, Washington, USA) laptop computer. All cookies from Google Chrome 81.0.4044.113 (Google LLC, California, USA) browser were removed before data collection was started. Based in previous studies,[14] on April 27, 2020, we queried Google Trends and downloaded the data for the following search input using the keywords individually and in combination for comparison: *Coronavirus*, *COVID-19*, *SARS-CoV-2*. Restricting the search to the geographical region of Galicia. Category filters were not included. The selected date range was from January 1, 2020 to April 15, 2020.

Once the points of each term with a greater SVI or that meant a significant change in the trend have been identified, a search on the Google Search Engine website applying the filters for the dates corresponding to those points and using the "news" query category was performed. Likewise, Galicia epidemiological data for the established date range, provided by the Ministerio de Sanidad (Gobierno de España),[15] of the SARS-CoV-2 pandemic were obtained in order to make a comparison with the search volume of the most used term.

Statistical analysis

Following the recommendations on big data analysis with Google Trends of specialized literature,[7,8,12,16] genomic and proteomic a statistical analysis was performed with the statistical package Stata® 14.2 (StataCorp LLC, Texas, USA). Statistical tests were two-tailed with a 95% confidence interval. The Spearman's correlation coefficient was used to examine the associations between search terms. A generalized linear model selecting the model according to the Akaike information criterion (AIC) and the analysis of the residuals was performed to assess the correlation between terms.[17]

The Joinpoint Regression Program, Version 4.8.0.1 (Statistical Research and Applications Branch, National Cancer Institute, USA) was used to identify significant trends changes over time for each search.[18] The Joinpoint methodology, based on Traditional Bayesian Information Criteria, is ideally suited to examine trends over time and to test whether an apparent change in trend is statistically significant.[19] The selected parameters were: A minimum of 1 week between two joinpoints were required, and a maximum of three joinpoints were allowed for describing the data. The trend was expressed by an annual percent change, a summary measure of the trend over a pre-specified fixed interval.[20]

RESULTS

Confirmed Cases (WHO data)

A total of 7,873 SARS-CoV-2 confirmed cases in Galicia were collected by the Ministerio de Sanidad between January 1 and April 15. The trend model (figure 1) indicates an increase o

4.7% ($p < 0.000$) during that period, with four well-differentiated segments found in Table 1.
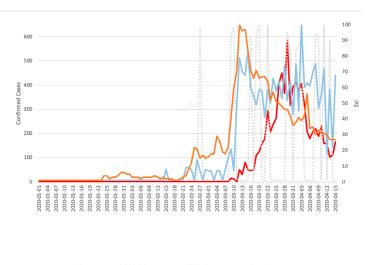
Coronavirus term

The *Coronavirus* term is the most searched, and it has suffered a progressive increase in the last 3.5 months, marking its highest peak in search volume on March 12 (figure 1). The statistical trend analysis of the *Coronavirus* term shows an increasing evolution, with a global APC of 3.6% ($p < 0.000$). Three joinpoints with their 4 corresponding segments were found in the term search volume trend (table 1). The start of the ascent begins on January 21, until February 25 (SVI = 22). Analyzing the news during that period, they are general information about the "new Wuhan coronavirus", which compared it to previous epidemics such as the flu or talked about its symptoms and the possibility that it could spread outside China. Another noteworthy news in that period is the appearance of the first case in the USA.

After a period of relative stability regarding the SVI, on March 7 (SVI = 18) there is a sharp rise until March 12 (SVI = 100). In this period, reports of the first cases and deaths in Spain, Italy, Iran, Mexico, etc. begin to appear in the news. It also matches with the WHO declaration of a pandemic and the first confinement measures in some countries. After this last peak it seems that the tendency is to decrease significantly.

COVID-19 term

The term *COVID-19* also presented an increase (Global APC = 3.7%; $p < 0.000$), but with a later onset (7 days) than the term *Coronavirus* according to trend analysis (figure 1 and table 1).

Search volume begins to climb slowly from February 21 until March 8 (from SVI = 0 to SVI = 7), at which point it undergoes a much steeper fluctuating increase to its peak on March 19 (SVI = 89) and April 3 (SVI = 100). Analyzing the news of those dates, the beginning of the increase in searches coincides with the date on which the WHO released the name of the disease. The steepest rise coincides with the beginning of the containment measures and the increase in cases in various countries.

Figure 1. Trend lines of the search volume index of the terms described and the confirmed cases.

Table 1. Trend and jointpoints of each of the search terms, as well as the number of confirmed cases of SARS-CoV-2 from January 1 to April 15, 2020.

	Segment	Lower Endpoint (Days)	Upper Endpoint (Days)	APC	Lower 95%CI	Upper 95%CI	Test Statistic (t)	Prob > |t|
Confirmed Cases	1	1	66	0.3	-0.2	0.8	1.3	0.2
	2	66	74	60.4	40.3	83.5	7.0	0.0
	3	74	88	17.9	11.9	24.3	6.2	0.0
	4	88	106	-7.7	-10.6	-4.6	-4.9	0.0
	Global	1	106	4.7	3.3	6.2	6.6	0.0
Coronavirus term SVI	1	1	33	5.5	4	7	7.6	0.0
	2	33	48	-3.3	-7.9	1.5	-1.4	0.2
	3	48	73	16	13.5	18.5	13.6	0.0
	4	73	106	-3.7	-4.9	-2.4	-5.5	0.0
	Global	1	106	3.6	2.5	4.7	6.6	0.0
COVID-19 term SVI	1	1	40	-0.0	-1.4	1.4	-0.0	1.0
	2	40	66	6.9	4.0	9.9	4.8	0.0
	3	66	74	37.1	13.4	65.7	3.3	0.0
	4	74	106	-1.3	-3.1	0.7	-1.3	0.2
	Global	1	106	3.7	1.9	5.6	4.1	0.0
SARS-CoV-2 term SVI	Global	1	106	2.4	1.5	3.3	5.1	0.0

APC, Annual Percent Change; CI95%, Confidence Interval 95%; WHO, World Health Organization; SVI, Search Volume Index.

The trend analysis showed 3 changes in trend coinciding with those previously described (table 1).

SARS-CoV-2 term

The term *SARS-CoV-2* it was the least sought after, with an APC of 2.4% (*p* <0.000) during this period, and only one segment, without jointpoints (table 1). It is the one that later begins to increase the volume of searches, doing it in a very fluctuating way (figure 1), and not clearly associating itself with news in the media.

Correlation and generalized linear model

A linear correlation was found between the confirmed cases (dependent variable) and the de SVI of the *Coronavirus* term (rho = 0.79; *p* <0.000), the rest of independent variables (*COVID-19* and *SARS-CoV-2* terms) did not show a statistically significant relationship (table 2). A linear correlation between *Coronavirus* term and *COVID-19* term (rho = 0.85; *p* <0.000), and *Coronavirus* term and *SARS-CoV-2* term (rho = 0.41; *p* <0.000) was also found.

DISCUSSION

The study of Internet search patterns has facilitated opportunities for evaluating public interest encompassing a variety of health-related topics.[12,16] Further, behavioural measures are needed in the healthcare environment and in public health planning, where national indices of progress on behaviour

measures could guide policy and conveyance planning.[12] Using internet big data analysis in healthcare research holds promise, extrapolating data from a portion of 3 billion searches daily, Google Trends represents a powerful tool for gauging public interest and has demonstrated its unique value in numerous analyses.[16]

On 31 December, at the end of 2019, some cases of pneumonia of unknown etiology were notified to the WHO Country Office in China, regarding Wuhan, a city in the province of Hubei.[21] The 3 months since the emergence of COVID-19 have demonstrated the rapid pace at which a virus can spread and at which science can develop. Every outbreak provides an opportunity to gain important information, some of which is associated with a limited window of opportunity.[22] In this case, the strategy against the pandemic seeks to detect the needs for the prevention, diagnosis and treatment of this pathology, as well as to establish work objectives and agreed-upon care recommendations and application of these to as many people as possible.

Since its launch in 2006, Google Trends has proven to be a valuable and accessible tool. It has mostly been used for monitoring and surveillance of communicable diseases and epidemics.[16,23] Despite concerns regarding the adequacy of available patient education materials,[24,25] with an increasing number of users turn to the Internet for health related background information. Our data analysis suggests that the

Table 2. Generalized linear model results.

Variable	Coefficient	Standard Error	P>|z|
Coronavirus term SVI	0.52	0.14	0.000
COVID-19 term SVI	0.02	0.03	0.523
SARS-CoV-2 term SVI	-0.01	0.02	0.703

Coronavirus term follows a linear correlation with the number of confirmed cases of SARS-CoV-2 in a statistically significant way. It can be seen that the trend lines are similar, with the *Coronavirus* term trend changes occurring a minimum of 15-30 days before the increase in cases (table 1). It could be possible that the search engine query data serves as an epidemiological model for future pandemics, or as a method of monitoring peak cases with the intention of anticipating and preparing for them. This has already been tested in other countries with this same methodology.[26]

Likewise, any news stories, new interventions or aetiology related to SARS-CoV-2 can manifest as an increase in information-seeking behaviours for any of the proposed search terms. This is consistent with the results obtained in previous studies.[14] This strengthens the idea that information-seeking behaviour is influenced by the level of awareness exposure,[23,27] and therefore, it may be possible that the mass media campaigns, in a direct and indirect way, can produce positive changes or prevent negative changes in health-related behaviours across large populations.[28]

It should be mentioned that there seems to be a correlation between *Coronavirus* and *COVID-19* or *SARS-CoV-2*. Both combinations were linked to the same influencing factors and happened during the same period of time. It seems that people use them indifferently, but possibly the term *Coronavirus* is the most used because it is more generic. Therefore, Google Trends could also be a powerful tool to understand the terminology used by patients in different regions.[14]

The Internet is unique as it serves as both a primary source of medical information and as a tool to study health information seeking behaviours. The mass media can hugely affect it. Therefore, understanding healthcare information seeking behaviour is essential in order to control and plan the quality of knowledge provided by health organisations, advocacy groups and health professionals regarding SARS-CoV-2 pandemic. A combination of the existing source containment strategy, contact investigation, infection control at health care facilities, as well as in community settings along with this new kind of search engine query data-based approach in order to create mathematical models to forecast disease spread, could be useful and helpful to activate and improve strategic plan to control an outbreak.[4] Further research is needed to understand how healthcare providers can use Google Trends to understand health information-seeking behaviours and its effect on public knowledge, awareness, disease-related anxiety and the interaction between patients and healthcare information providers. Despite this, the results are promising for future applications.

REFERENCES

1. COVID-19 Map [Internet]. Johns Hopkins Coronavirus Resource Center. [cited Apr 27, 2020]. Available in: https://coronavirus.jhu.edu/map.html

2. Coronavirus [Internet]. [cited Apr 6, 2020]. Available in: https://www.who.int/emergencies/diseases/novel-coronavirus-2019

3. Li H, Liu S-M, Yu X-H, Tang S-L, Tang C-K. Coronavirus disease 2019 (COVID-19): current status and future perspective. Int J Antimicrob Agents. 2020;105951.

4. Perrella A, Carannante N, Berretta M, Rinaldi M, Maturo N, Rinaldi L. Novel Coronavirus 2019 (Sars-CoV2): a global emergency that needs new approaches? Eur Rev Med Pharmacol Sci. 2020;24(4):2162-4.

5. Al Nuaimi E, Al Neyadi H, Mohamed N, Al-Jaroodi J. Applications of big data to smart cities. J Internet Serv Appl. 2015;6(1):25.

6. Gandomi A, Haider M. Beyond the hype: Big data concepts, methods, and analytics. Int J Inf Manag. 2015;35(2):137-44.

7. Yoo C, Ramirez L, Liuzzi J. Big data analysis using modern statistical and machine learning methods in medicine. Int Neurourol J. 2014;18(2):50-7.

8. Mavragani A, Ochoa G. Google Trends in Infodemiology and Infoveillance: Methodology Framework. JMIR Public Health Surveill. 2019;29;5(2):e13439.

9. Cacciamani GE, Bassi S, Sebben M, Marcer A, Russo GI, Cocci A, et al. Consulting «Dr. Google» for Prostate Cancer Treatment Options: A Contemporary Worldwide Trend Analysis. Eur Urol Oncol. 2019;S2588-9311(19)30105-1.

10. Google Trends [Internet]. Google Trends. [cited Sep 4, 2019]. Available in: https://trends.google.es/trends/?geo=ES

11. Ginsberg J, Mohebbi MH, Patel RS, Brammer L, Smolinski MS, Brilliant L. Detecting influenza epidemics using search engine query data. Nature. 2009;457(7232):1012-4.

12. Eysenbach G. Infodemiology and infoveillance: framework for an emerging set of public health informatics methods to analyze search, communication and publication behavior on the Internet. J Med Internet Res. 2009;11(1):e11.

13. Cómo se ajustan los datos de Trends - Ayuda de Tendencias de búsqueda [Internet]. [cited Sep 6, 2019]. Available in: https://support.google.com/trends/answer/4365533?hl=es&ref_topic=6248052

14. Faoury M, Upile T, Patel N. Using Google Trends to understand information-seeking behaviour about throat cancer. J Laryngol Otol. 2019;133(7):610-4.

15. Situación de COVID-19 en España [Internet]. [cited Apr 27, 2020]. Available in: https://covid19.isciii.es/

16. Nuti SV, Wayda B, Ranasinghe I, Wang S, Dreyer RP, Chen SI, et al. The Use of Google Trends in Health Care Research: A Systematic Review. PLoS ONE. 2014;22;9(10):e109583

17. Akaike H. A new look at the statistical model identification. IEEE Trans Autom Control. 1974;19(6):716-23.

18. Joinpoint Regression Program, Version 4.8.0.1. April, 2020; Statistical Research and Applications Branch, National Cancer Institute.

19. Kim HJ, Fay MP, Feuer EJ, Midthune DN. Permutation tests for joinpoint regression with applications to cancer rates. Stat Med. 2000;19(3):335-51.

20. Average Annual Percent Change (AAPC) and Confidence Interval [Internet]. Joinpoint Help System. [cited Apr 27, 2020]. Available in: https://surveillance.cancer.gov/help/joinpoint/setting-parameters/method-and-parameters-tab/apc-aapc-tau-confidence-intervals/average-annual-percent-change-aapc

21. Promed Post – ProMED-mail [Internet]. [cited Apr 6, 2020]. Available in: https://promedmail.org/promed-post/

22. Fauci AS, Lane HC, Redfield RR. Covid-19 - Navigating the Uncharted. N Engl J Med. 2020;382(13):1268-9.

23. 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;7(3):185-9.

24. Svider PF, Agarwal N, Choudhry OJ, Hajart AF, Baredes S, Liu JK, et al. Readability assessment of online patient education materials from academic otolaryngology-head and neck surgery departments. Am J Otolaryngol. 2013;34(1):31-5.

25. Tulsieram KL, Arocha JF, Lee J. Readability and Coherence of Department/Ministry of Health HPV Information. J Cancer Educ. 2018;33(1):147-53.

26. Ayyoubzadeh SM, Ayyoubzadeh SM, Zahedi H, Ahmadi M, R Niakan Kalhori S. Predicting COVID-19 Incidence Through Analysis of Google Trends Data in Iran: Data Mining and Deep Learning Pilot Study. JMIR Public Health Surveill. 2020; 6(2): e18828

27. Schootman M, Toor A, Cavazos-Rehg P, Jeffe DB, McQueen A, Eberth J, et al. The utility of Google Trends data to examine interest in cancer screening. BMJ Open. 2015;5(6):e006678

28. Wakefield MA, Loken B, Hornik RC. Use of mass media campaigns to change health behaviour. Lancet. 2010;376(9748):1261-71.