Acta Biomed 2020; Vol. 91, N.4: e2020190 DOI: 10.23750/abm.v91i4.8888 

Original article

# Can Google Trends and Wikipedia help traditional surveillance? A pilot study on Measles

*Omar Enzo Santangelo[1], Sandro Provenzano[1], Dimple Grigis[2], Domiziana Giordano[1], Francesco Armetta[1], Alberto Firenze[1]*

[1]Department of Health Promotion, Mother and Child Care, Internal Medicine and Medical Specialties "G. D'Alessandro", University of Palermo, Palermo, Italy; [2]University of Bergamo, Bergamo, Italy

**Summary.** *Introduction:* Cases of measles in some European countries are increasing. The aim of this study is to find the correlation between Google Trends and Wikipedia searches and the real number of cases notified. *Materials and Methods:* The data on Internet searches have been obtained from Google Trends and Wikipedia. The reported cases of measles were selected from January 2013 until December 2018 for Google Trends and July 2015 until December 2018 from for Wikipedia. We have selected data from four European Countries: Italy, France, Germany and Romania. The data extracted from Wikipedia and Google Trends have been moved over time (Lag), one month in the future and one month in the past. Cross-correlation results are obtained as product-moment correlations between the two time series. The statistical analyses have been performed by using the Spearman's rank correlation coefficient or Pearson correlation coefficient. *Results:* A temporal correlation was observed between the bulletin of ECDC and Wikipedia search trends. For Wikipedia the strongest correlation is at a lag of +1 for *rougeole* (r=0.9006) and *masern* (r=0.7023) and at lag 0 for *morbillo* (r=0.8892) and *rujeola* (r=0.5462); for Google Trends the strongest correlation at a lag 0 for *rougeole* (rho=0.7398), *symptômes rougeole* (rho=0.3399), *masern* (rho=0.6484), *sintomi morbillo* (rho=0.6029), *rujeola* (rho=0.7209), *simptome rujeola* (rho=0.5297) and at lag -1 for *masern symptom* (rho=0.4536) and *morbillo* (rho=0.5804). *Conclusions:* Google and Wikipedia could play an important role in surveillance, although these tools need to be combined with traditional surveillance systems. (www.actabiomedica.it)

**Key words:** vaccine-preventable diseases, Italy, Germany, France, Romania, Measles vaccine, Big Data, Internet, Measles, Medical Informatics Computing, Medical Informatics

## Introduction

Cases of measles in some European countries are increasing, large outbreaks with fatalities are ongoing in countries that had previously eliminated or interrupted endemic transmission (1).

Internet-based surveillance systems offer a novel and developing means of monitoring conditions of public health concern, including emerging infectious diseases (2).

The Google Trends database is searchable by term, geography and time with a one-week sampling rate. Google Trends allows a user to compare up to five terms or topics simultaneously and results are displayed as a set of time series. Google Trends normalizes the search data with the day on which more searches were made giving a reference value equal to 100, on the contrary, it assigns a reference value of 0 for the day when fewer searches were carried out. Then the data standardized are presented by Google Trends as "relative search volume" (RSV), an "Interest Index" that can take a value between 0 and 100 based on the proportion to all searches on all terms or topics (3).

The association between the predictive power of Google Trends and the data of official surveillance systems of various countries has been shown by various authors for different diseases, concluding that these data can help to monitor and predict infectious diseases (2,4).

The objective of the study is to evaluate, through two comparative studies, time correlation between Google Trends, Wikipedia Trends and the conventional surveillance data generated by the reporting of measles infection cases reported on bulletin by the European Centre for Disease Prevention and Control (ECDC).

## Materials and methods

Cross-sectional study design was used. Every month the ECDC issues a bulletin with the cases reported in European Nations in the previous months regarding measles (5).

We have selected data from four European Countries: Italy, France, Germany and Romania.

From Wikipedia Trends (6) it is possible to know how many times a specific page is viewed by users, data were extracted and aggregated on monthly basis. Then, the following data were extracted:

- a number of monthly views by users from 1 July 2015 to 31 December 2018 of the pages: *morbillo* (Italian term for measles), *rougeole* (french term for measles), *masern* (german term for measles), *rujeola* (romanian term for measles).

From Google Trends (3), on June 10, 2019, the data have been obtained using the italian, french, german and romanian search terms, in the "Health" category, *morbillo* (italian), *rougeole* (french), *masern* (german), *rujeola* (romanian) that mean "measles" in english, and *sintomi morbillo* (italian), *symptômes rougeole* (french), *masern symptome* (german), *simptome rujeola* (romanian) that mean "measles symptoms" in english, in the time-frame elapsing from 1 January 2013 to 31 December 2018; the data have been aggregated by month.

The files in ".CSV" format have been downloaded. Google Trends provides for a relative search volume (RSV), which is computed as the percentage of queries concerning a particular term for a specific location and time period, where 100 is the maximum value and 0 is the minimum value.

Then we created two databases:

- with monthly data (MDW) with the reported cases of measles in ECDC bulletin and Wikipedia Trends data from July 2015 to December 2018;
- with monthly data (MDG) with the reported cases of measles in ECDC bulletin and Google Trends data from January 2013 to December 2018.

The data extracted from Wikipedia and Google Trends have been moved over time (Lag), one month in the future and one month in the past.

Cross-correlation results are obtained as product-moment correlations between the two time series. The advantage of using cross-correlations is that it accounts for time dependence between two time-series variables.

Statistical analyses were performed using the Pearson correlation coefficient (r) for the "MDW" database and Spearman's rank correlation coefficient (rho) for the "MDG" database. The statistical significance level for the analyses was 0.05. The data were analyzed using the STATA statistical software, version 14 (7).

## Results

The raw data for Wikipedia Trends are shown in Figure 1. A temporal correlation was observed between the bulletin of ECDC and Wikipedia search trends. Regarding the database MDW, the strongest correlation is at a lag of +1 for *rougeole* (r=0.9006) and *masern* (r=0.7023) and at lag 0 for *morbillo* (r=0.8892) and *rujeola* (r=0.5462) (Table 1). Google Trends Internet search data showed the strongest correlation at a lag 0 for *rougeole* (rho=0.7398), *symptômes rougeole* (rho=0.3399), *masern* (rho=0.6484), *sintomi morbillo* (rho=0.6029), *rujeola* (rho=0.7209), *simptome rujeola* (rho=0.5297) and at lag -1 for *masern symptom* (rho=0.4536) and *morbillo* (rho=0.5804) (Table 2).

**Table 1.** Time series bi-directional cross-correlation coefficients for 1 month displaying relationships between Wikipedia Trends and cases reported by the ECDC. Used Pearson correlation coefficient

| Wikipedia Trends Terms | Lag in months compared to cases reported by the ECDC | | |
|---|---|---|---|
| | -1 (42 observations) | 0 (42 observations) | +1 (41 observations) |
| Rougeole (France) | 0.5803* | 0.8278* | **0.9006*** |
| Masern (Germany) | 0.3258** | 0.6400* | **0.7023*** |
| Morbillo (Italy) | 0.8085* | **0.8892*** | 0.6840* |
| Rujeola (Romania) | 0.5101* | **0.5462*** | 0.5390* |

*p-value<0.001 / **p-value<0.05

**Table 2.** Time series bi-directional cross-correlation coefficients for 1 month displaying relationships between Google Trends and cases reported by the ECDC. In bold, the strongest correlations. Used Spearman's rank correlation coefficient

| | Google Trends Terms | Lag in months compared to cases reported by the ECDC | | |
|---|---|---|---|---|
| | | **-1** | **0** | **+1** |
| France | rougeole | 0.6982* | **0.7398*** | 0.6726* |
| | symptômes rougeole | 0.2919** | **0.3399**** | 0.2734** |
| Germany | masern | 0.6354* | **0.6484*** | 0.5033* |
| | masern symptome | **0.4536*** | 0.4424* | 0.4441* |
| Italy | morbillo | **0.5804*** | 0.5398* | 0.4515* |
| | sintomi morbillo | 0.5787* | **0.6029*** | 0.5288* |
| Romania | rujeola | 0.6908* | **0.7209*** | 0.6869* |
| | simptome rujeola | 0.4829* | **0.5297*** | 0.4434* |

*p-value<0.001 / **p-value<0.05

## Discussion and Conclusions

The results for months at Lag 0 showed that the peaks of the curves for France and Germany anticipate by about one month the peaks of the curve deriving from the cases notified by the ECDC, while the peaks of the curve for Italy can be superimposed on the curve of the ECDC. Table 1 shows the Time series bi-directional cross-correlation coefficients for 1 month displaying relationships between Wikipedia Trends and cases reported by the ECDC. From this analysis it emerged that for France and Germany the maximum correlation between ECDC and Wikipedia data was observed at lag +1. This could mean that searches for selected terms on Wikipedia anticipate ECDC notifications by about a month. While for Italy and Romania the highest correlation occurs at Lag 0, so the search for terms on Wikipedia is about the same time as the cases notified by ECDC.

Medium or strong correlations emerge mainly at Lag 0 analyzing the data on Google trends (Table 2), probably attributable, according to the authors, to the fact that the population is currently looking for the terms present in Table 2 and therefore the number of searches is directly connected to the number of measles cases in progress. It would be possible to obtain more specific information if the ECDC bulletin were weekly, the monthly lags are still large enough to plan for a possible response to an epidemic, in other studies this type of analysis has already been carried out (8, 9).

With regard to the limits of the study, it should be noted that the media could influence the population's search for online terms. There are several reasons for the peak search for measles terms, such as the increase

of the number of cases in the community and the increased media attention (10). While for Google trends (3) it is possible to separate the data at the regional geographical level, another limitation for Wikipedia is the lack of geographical identification of a possible epidemic because Wikipedia Trends does not provide data at these levels. In addition, the temporal and geographical changes in are not well documented, which may affect the outcome of the research and the results of our study (10). Therefore, the interpretation and generalization of results require caution.

In conclusion, the results of this study suggest that Google Trends and Wikipedia-based surveillance systems have a potential role as a possible public health tool. Today, it can be a valuable tool that can flank the traditional surveillance systems ones and that in the future could be more validated and consolidated.

**Authors' contributions:** OES, SP and AF conceived, designed, coordinated and supervised the research project. OES collected samples. OES performed the data quality control, optimized the informatics database, OES performed the statistical analyses and evaluated the results. All Authors wrote the manuscript. All Authors revised the manuscript and gave their contribution to improving the paper. All authors read and approved the final manuscript.

**Conflict of interest:** Each author declares that he or she has no commercial associations (e.g. consultancies, stock ownership, equity interest, patent/licensing arrangement etc.) that might pose a conflict of interest in connection with the submitted article

## References

1. European Centre for Disease Prevention and Control. Measles; [Internet]. Available from: https://ecdc.europa.eu/en/measles , Last accessed [September 13, 2019].
2. Milinovich GJ, Williams GM, Clements AC, Hu W. Internet-based surveillance systems for monitoring emerging infectious diseases. Lancet Infect Dis 2014 14(2):1 60-8. doi: 10.1016/S1473-3099(13)70244-5. Epub 2013 Nov 28. Review.
3. Google Trends, [Internet]. Available from: https://trends.google.com/trends/, Last accessed [September 14, 2019].
4. Alicino C, Bragazzi NL, Faccio V, et al. Assessing Ebola-related web search behaviour: insights and implications from an analytical study of Google Trends-based query volumes. Infect Dis Poverty. 2015 Dec 10;4:54. doi: 10.1186/s40249-015-0090-9.
5. European Centre for Disease Prevention and Control. Publications & data. Measles; [Internet]. Available from: https://ecdc.europa.eu/en/publications-data?f%5B0%5D=diseases%3A209, Last accessed [September 15, 2019].
6. WikipediaTrends;[Internet].Availablefrom:https://tools.wmflabs.org/pageviews/?project=en.wikipedia.org&platform=all-access&agent=user&range=latest-20&pages=, Last accessed [September 15, 2019].
7. StataCorp 2015. Stata Statistical Software. Release 14. College Station, TX: StataCorp LP
8. Santangelo OE, Provenzano S, Piazza D, Giordano D, Calamusa G, Firenze A. Digital epidemiology: assessment of measles infection through Google Trends mechanism in Italy. Ann Ig. 2019 Jul-Aug;31(4):385-391. doi: 10.7416/ai.2019.2300.
9. Provenzano S, Santangelo OE, Giordano D, et al. Predicting disease outbreaks: evaluating measles infection with Wikipedia Trends. Recenti Prog Med. 2019 Jun;110(6):292-296. doi: 10.1701/3182.31610.
10. Lazer D, Kennedy R, King G, Vespignani A. Big data. The parable of Google Flu: traps in big data analysis. Science. 2014 Mar 14; 343(6176): 1203-5. doi: 10.1126/science.1248506.

Received: 25 September 2019
Accepted: 20 January 2020
Correspondence:
Dr. Sandro Provenzano,
Department of Health Promotion,
Mother and Child Care, Internal Medicine and
Medical Specialties "G. D'Alessandro", University of Palermo,
Via del Vespro, 129, 90127 Palermo (PA), Italy
Tel. +390916553641
Fax: +390916553697
E-mail: provenzanosandro@hotmail.it