Ann Ig 2019; 31: 385-391 doi:10.7416/ai.2019.2300

SHORT PAPER

Digital epidemiology: assessment of measles infection through Google Trends mechanism in Italy

O.E. Santangelo[1], S. Provenzano[1], D. Piazza[1], D. Giordano[1], G. Calamusa[1], A. Firenze[1]

Key words: Vaccine-preventable diseases, Italy, Measles Vaccine, Big Data, Internet, Measles, Medical Informatics Computing, Medical Informatics

Parole chiave: Malattie prevenibili con il vaccino, Italia, Vaccino antimorbillo, Big Data, Internet, Morbillo, Informatica medica

Abstract

Introduction. *The primary aim of this study is to evaluate the temporal correlation between Google Trends and the data on measles infection arising from the conventional surveillance system, reported by the Istituto Superiore di Sanità's (ISS) bulletin. Moreover, this study is also aimed at forecasting the trends of the reported infectious diseases cases over time.*

Materials and Methods. *The reported cases of measles were selected from January 2013 until October 2018. The data on Internet searches have been obtained from Google Trends; the research data referred to the first 48 weeks of year 2017 have been aggregated on a weekly basis. The search volume provided by Google Trends has a relative nature and is calculated as a percentage of query related to a specific term in connection with a determined place and time-frame. The statistical analyses have been performed by using the Spearman's rank correlation coefficient (rho). The statistical significance level for such analyses has been fixed in 0.05.*

Outcomes. *We have observed a strong correlation at a lag of 0 to -4 weeks (rho > 0.70) with the cases reported by ISS with the strongest correlation at a lag of -3 weeks (rho > 0.80 both for measles than for the symptoms of the measles). The database containing monthly data has shown a moderate correlation at a lag of -1 to +1 months and a strongest correlation at a lag of -1 (rho = 0.6152 for measles and rho = 0.5039 for symptoms of the measles).*

Conclusion. *The surveillance systems based on Google Trends have a potential role in public health in order to provide near real-time indicators of the spread of infectious diseases. Therefore the huge potential of this approach could be used in the immediate future as a support of the traditional surveillance systems.*

[1] *Department of Science for Health Promotion and Mother-Child Care "G. D'Alessandro", University of Palermo, Palermo, Italy*

Introduction

In the last few years the online activity based on epidemiological surveillance and forecasting is taking great attention. Prompt detection is a cornerstone in the control and prevention of infectious diseases. Google, the world's most popular search engine, currently processes 1.2 trillion searches annually worldwide (over 40,000 search queries every second) (1). Google logs the searches from its platforms and then provides a sampling on Google Trends for review and analysis by anyone, therefore provides a factual perspective on topics which currently interest and concern people.

Social media platform data and Google Trends offer an interesting tool to monitor public attention with regard to specific infectious diseases (2). Several studies have shown that this quantifiable attention is a good proxy for infectious diseases subjected to health surveillance. The data generated from queries fed into search engines are recorded and can be used for surveillance purposes. The targeted sources include Internet-search metrics, online news stories, social network data, blog and microblog data (2).

Google Trends is an online tracking system of Internet search volumes that since 2004 is used to explore web behavior related to a topic or search term.

The Google Trends database is searchable by term, geography, and time with a one-week sampling rate. Google also categories searches into topics - groups of terms that share the same concept across languages. 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. Google Trends performs a complex series of statistical operations, such as amassing data and normalizing and rescaling them.

The association between the predictive power of Google Trends and the data of official surveillance systems of various countries has been studied by various authors for different diseases, concluding that there is a statistically significant association and therefore it can offer significant information on population behavior and on disease-related phenomena (3-5).

Several Internet tools that use population-level trends in Google and other Internet search-engine queries about infectious diseases such as influenza, pertussis, and norovirus to detect and predict epidemics have been developed in recent years (4). In developing areas where traditional epidemiologic surveillance faces multiple challenges these data can help to monitor and predict infectious diseases (5). Measles is one of the most contagious diseases of humans. Measles, indeed, is a disease that can lead to serious complications, such as pneumonia (infection of the lungs), and even death. It remains an important cause of death among young children globally, despite the availability of a safe and effective vaccine. An important health public problem concerns the immunization status of the weakest groups in the population with the most common primary prevention interventions, such as vaccinations (6). In Italy, National Plan of Vaccinations (PNPV 2017-2019) plans the measles-mumps-rubella vaccine administration between the 13th and the 15th month and a booster dose at the sixth year of life and, for hepatitis B, plans three doses at the 3rd, 5th and 11th-13th month of life. The Italian PNPV, often different from those used in the children countries of origin, includes - among its objectives - the reduction of inequalities and the improvement of

the health status by promoting vaccination interventions in groups of marginalized or particularly vulnerable populations. Under the Global Vaccine Action Plan measles and rubella are targeted for elimination in five WHO Regions by 2020, in this contest surveillance of online behavior is a potential web-based disease detection system that can improve monitoring of measles. This study has been performed with the primary aim of evaluating the temporal correlation between Google Trends and conventional surveillance data generated for measles infection reported by bulletin of *Istituto Superiore di Sanità* (ISS). Thus, it is crucial to study the application of this tool for the surveillance of communicable diseases like measles.

Materials and methods

A cross-sectional study design has been used. The reported cases of measles were selected from January 2013 until October 2018. The ISS issues, on a monthly basis, a bulletin containing the cases reported in the previous months regarding measles (7). In addition, in 2017 for the first 48 weeks of such year the ISS issued a weekly bulletin that reported the number of new cases for each week of the previous weeks (7), these data have been also selected for the final analysis, weekly data are referring to 2017 only.

Data on Internet searches have been obtained from Google Trends (1). In particular, with respect to Italy, on November 30, 2018, the data have been obtained using the Italian search terms, in the "Health" category, "Morbillo" ("measles" in English) and "Sintomi del Morbillo" ("Symptoms of the Measles" in English), in the time-frame elapsing "from 1 January 2013 to 31 October 2018"; the data have been aggregated by month; research data referring to the first 48 weeks of 2017 have also been aggregated on a weekly basis.

The file in ".CSV" format has 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.

Therefore, we have created two databases: one containing monthly data (MD), and the other one providing for weekly data (WD). The data have been subdivided according to the respective databases in order to make the data coincide temporally with the monthly or weekly incidence reported in the epidemiological bulletins of the ISS; then, the data extracted from Google Trends have been moved over time (Lag), one month in the future and one month in the past as regards the database with monthly data (MD), and one, two, three and four weeks in the future and in the past regarding the weekly data (WD) database. 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 have been performed using the Spearman's rank correlation coefficient (rho). The statistical significance level for the analyses has been fixed in 0.05. The data have been analyzed using the STATA statistical software, version 14.

Results

The raw data of reported cases for weeks in 2017 year and months are shown in Figure 1. A temporal correlation has been found between the bulletin of ISS and Google search trends. Google Trends Internet search data showed a strong correlation at a lag of 0 to -4 weeks (rho > 0.70) with the reported cases of ISS (Table 1), with the strongest correlation at a lag of -3 weeks (rho = 0.8271 for measles and rho = 0.8473 for symptoms of the measles); a moderate correlation is showed at a lag of +1 to +3 weeks (Table 1). With reference to the database containing

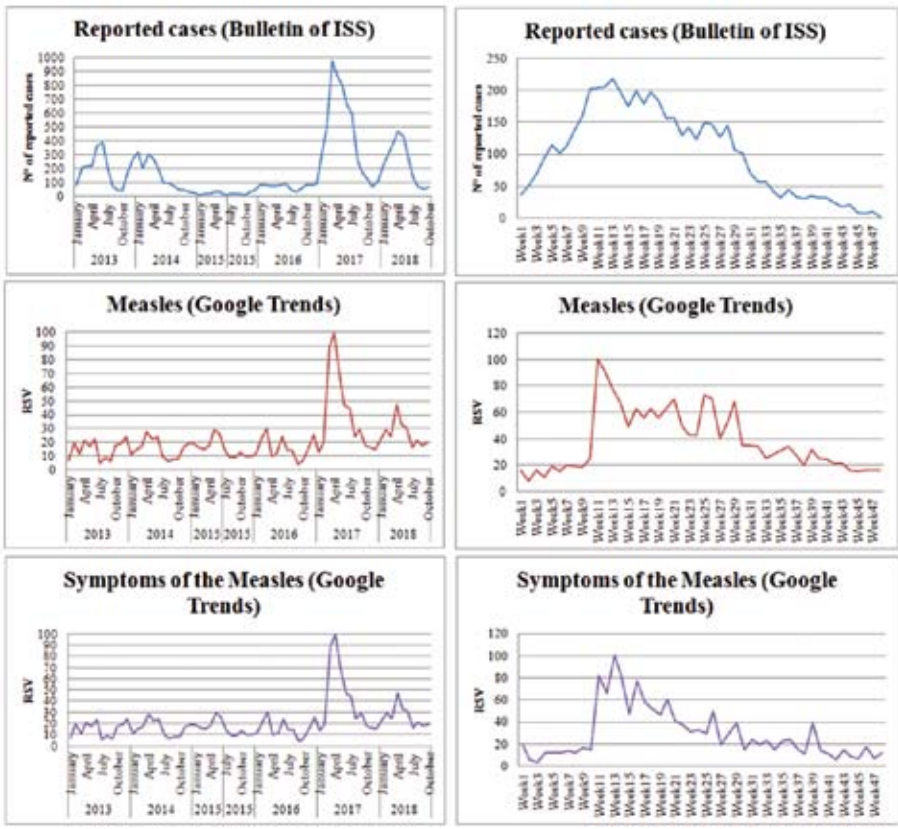

Figure 1 - Number of reported cases of measles and RSV Google Trends of search terms Measles and Symptoms of the measles. Results for months (January 2013-October 2018) and weeks (from the 1st to the 48th of the year 2017).

the monthly data (MD), at a lag of -1 to +1 months a moderate correlation is showed, but the strongest correlation is at a lag of -1 (rho = 0.6152 for measles and rho = 0.5039 for symptoms of the measles) (Table 2).

Discussion

It is very important to study the application of Google Trends for the surveillance of communicable diseases in Italy. Indeed, many studies from other parts of the world

Table 1 - Time series bi-directional cross-correlation coefficients for 4 weeks displaying relationships between Google Trends and cases reported by the ISS. In bold, the strongest correlations. Used Spearman's rank correlation coefficient.

	Lag in weeks compared to cases reported by the ISS								
Google Trends Terms	-4	-3	-2	-1	0	+1	+2	+3	+4
Measles	0.7973*	**0.8271***	0.8074*	0.7680*	0.7164*	0.5881*	0.4762*	0.3354**	0.2088
Symptoms of theMeasles	0.7989*	**0.8473***	0.8126*	0.7853*	0.7306*	0.6463*	0.5386*	0.4599**	0.3663**

*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 ISS. In bold, the strongest correlations. Used Spearman's rank correlation coefficient.

	Lag in months compared to cases reported by the ISS		
Google Trends Terms	-1	0	+1
Measles	**0.6152***	0.5803*	0.4827*
Symptoms of the Measles	**0.5039***	0.4977*	0.4554*

*p-value<0.001

suggest that this tool can be useful for disease surveillance (2).

Measles outbreaks are a common problem worldwide and surveillance is of paramount importance. Classical monitoring approaches based on conventional surveillance systems are plagued by some shortcomings, such as considerable time delay and potential underestimation of cases and/or underreporting (3). In 2009, Eysenbach defined Infodemiology and infoveillance 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*" (8). This innovative, emerging science has been used to explore public interest toward communicable diseases, such as influenza (8), as well as to detect infectious outbreaks (9).

From January 1st to October 31st 2018, no. 2368 cases of measles were reported in Italy (incidence of 47 cases per million inhabitants), no. 66 of which in October 2018. A percentage equal to 90% of the cases occurred in eight Regions; the Sicily Region reported the highest incidence. The median age of the cases is 25 years. The highest incidence has been observed in children under the age of one. In November 2018, a new death was reported in the Friuli Venezia Giulia Region involved a 23-year-old patient suffering from leukemia, bringing to eight the number of deaths reported in 2018 (10). Measles is one of the most contagious diseases that can be prevented by vaccination (R0=12-18). In addition to the above, please consider that the delays in traditional surveillance systems limit the ability of public health agencies to efficiently face epidemics. Data on Google Trends are collected and processed in near real-time and online search information produces monitoring data much faster than traditional systems (5). Firstly, we have performed correlation analyses to investigate the temporal correlations between data on measles related to Google Trends and the cases of measles reported by ISS. The results showed that a temporal correlation has been observed between the bulletin of ISS and Google search trends. In particular, a strong correlation at a lag of 0 to -4 weeks (rho > 0.70) with the reported cases of ISS with the strongest correlation at a lag of -3 weeks (rho > 0.80 both for measles than for the symptoms of the measles) was observed. Based on the stronger correlation values, the maximum correlation has been observed from 2 to 3 weeks before and this would provide sufficient time for timely action. The database with MD has shown a moderate correlation at a lag of -1 to +1 months and a strongest correlation at a lag of -1 (rho = 0.6152 for measles and rho = 0.5039 for symptoms of the measles).

The following limits can be identified in our study: the mass media (TV, radio) influence the online research of the population (11). The spike of Internet searches, for example, for "Measles" may be attributed to various factors. It may be due to the increased number of cases in the community and increased attention given by the mass media. The established correlation may not help to identify the place of an

outbreak because the Google Trends does not provide data at these levels. Moreover, temporal and geographic changes in the interface of Google Trends over time are not well documented, which may affect the search output and our study findings. Thus, the interpretation and generalization of the findings call for caution.

There is a need to demonstrate the applicability of this internet search data to be used by all states.

As shown by international studies, new integrated approaches should be explored to further improve vaccination coverage (11). In the last few years, a number of new activities have recently been planned and implemented with a progressive increase in web contacts and the involvement of several institutional bodies who contributed to the development of institutional websites dedicated to health policy with an increasingly active participation of various stakeholders. It is in this context that website VaccinarSì (12) develops with updates, sections devoted to regional problems, in-depth news analysis, and international expansion becoming an effective way to counteract vaccine hesitancy.

In the past, considering that the first reaction of people affected by influenza or flu syndrome is to search for information on the web, Google Flu Trends (GFT) aimed to predict epidemics, counting on the fact that from Big Data came an avalanche of information that would help us to do things that would be impossible with a lower volume of data, collected in a traditional way, however the use of GFT, presented in 2008 with the promise to monitor in real time the cases of influence based on the search terms associated with that disease on Google: chills, weakness, fever, headache, cough, sore throat, overestimated the prevalence of influenza in the 2012-2013 season of more than fifty percent, was inaccurate on the peak of the flu season, also due to the 'influence of the mass media that talking about it caused the frightened people to go to Google to get information, as well as errors on the search algorithm that they did to change the system (13). In the case of measles the search volume is unlikely to be determined by the patients themselves (or their relatives) since the number of cases is relatively small but the alarmism of the mass media can influence research, this process is unlikely to be more sensitive that actual disease surveillance.

Finally, caution should be used when interpreting the findings of Google Trends digital surveillance. The results of this study suggest that the surveillance systems based on Google Trends have a potential role in public health for the dynamic information and provide for near real-time indicators of the spread of infectious disease, therefore the huge potential of this approach could be used in the immediate future as a support to the traditional surveillance systems.

Funding: None.
Competing interests: None declared.
Ethical approval: Not required, the data have been provided and analyzed in anonymous and aggregated manner.
Author's contribution statement: All individuals listed as authors have substantially contributed to designing, performing or reporting the study.
Submission declaration and verification: This article has not been previously published, it is not under consideration for publication elsewhere, the relevant publication is approved by all Authors and tacitly or explicitly by the responsible authorities where the work has been carried out, and that, if accepted, it will not be published elsewhere in the same form, in English or in any other language, including electronically without the written consent of the copyright-holder. Authors declare that the data set forth under their paper are not a result of plagiarism, self-plagiarism or fraud, and that all data in the article are real, authentic and original.

Riassunto

Epidemiologia digitale: valutazione dell'infezione da morbillo con Google Trends in Italia

Introduzione. L'obiettivo principale dello studio è quello di valutare la correlazione temporale tra Google Trends e i dati del sistema di sorveglianza del morbillo

riportati dal bollettino dell'Istituto Superiore di Sanità (ISS). Inoltre, lo studio vuole prevedere le tendenze della segnalazione delle malattie infettive nel tempo.

Materiali e metodi. I casi segnalati di morbillo sono stati selezionati da gennaio 2013 a ottobre 2018. I dati sulle ricerche su Internet sono stati ottenuti da Google Trends; i dati di ricerca relativi alle prime 48 settimane del 2017 sono stati aggregati a settimana. Google Trends fornisce un volume di ricerca relativo, che viene calcolato come percentuale di query relative a un termine specifico per un determinato luogo e periodo di tempo. Le analisi statistiche sono state eseguite utilizzando il coefficiente di correlazione di Spearman (rho). Il livello di significatività statistica per le analisi era 0,05.

Risultati. Abbiamo osservato una forte correlazione con un ritardo da 0 a -4 settimane (rho> 0,70) con i casi riportati di ISS con la correlazione più forte con un ritardo di -3 settimane (rho> 0,8 sia per il morbillo che per i sintomi del morbillo). Il database con dati mensili ha mostrato una correlazione moderata con un ritardo da -1 a +1 mesi e una correlazione più forte con un ritardo di -1 (rho = 0,6152 per il morbillo e rho = 0,5039 per i sintomi del morbillo).

Conclusioni. I sistemi di sorveglianza basati su Google Trends hanno un potenziale ruolo nella salute pubblica per fornire indicatori quasi in tempo reale della diffusione delle malattie infettive, pertanto l'enorme potenziale di questo approccio potrebbe essere utilizzato nell'immediato futuro come supporto ai sistemi di sorveglianza tradizionali.

References

1. Google Trends. Available from: https://trends.google.com/trends/ [Last accessed 2019, Apr 23].
2. Milinovich GJ, Williams GM, Clements AC, Hu W. Internet-based surveillance systems for monitoring emerging infectious diseases. Lancet Infect Dis 2014; **14**(2): 160-8. doi: 10.1016/S1473-3099(13)70244-5. Epub 2013 Nov 28. Review. PubMed PMID: 24290841.
3. Pollett S, Boscardin WJ, Azziz-Baumgartner E, et al. Clin Infect Dis 2017; **64**(1): 34-41. Epub 2016 Sep 26. PubMed PMID: 27678084.
4. Ho HT, Carvajal TM, Bautista JR, et al. Using Google Trends to Examine the Spatio-Temporal Incidence and Behavioral Patterns of Dengue Disease: A Case Study in Metropolitan Manila, Philippines. Trop Med Infect Dis 2018; **3**(4). pii: E118. doi: 10.3390/tropicalmed3040118. PubMed PMID: 30423898.
5. Teng Y, Bi D, Xie G, et al. Dynamic Forecasting of Zika Epidemics Using Google Trends. PLoS One 2017; **12**(1): e0165085. doi: 10.1371/journal.pone.0165085. eCollection 2017. PubMed PMID: 28060809; PubMed Central PMCID: PMC5217860.
6. Giordano D, Provenzano S, Santangelo OE, et al. Active immunization status against measles, mumps, rubella, hepatitis B in internationally adopted children, surveyed at the university hospital of Palermo, Sicily. Ann Ig 2018; **30**(5): 431-5. doi: 10.7416/ai.2018.2243. PubMed PMID: 30062371.
7. Istituto Superiore di Sanità. Available from: https://www.epicentro.iss.it/morbillo/bollettino/Measles_WeeklyReport_N35.pdf [Last accessed 2019, Apr 23].
8. Eysenbach G. Infodemiology and infoveillance: framework for an emerging set of public health informatics methods to analyze serch, communication and publication behavior on the Internet. J Med Internet Res 2009; **11**(1): e11.
9. Gianfredi V, Bragazzi NL, Mahamid M, et al. Monitoring public interest toward pertussis outbreaks: an extensive Google Trends-based analysis. Public Health 2018; **165**: 9-15. doi: 10.1016/j.puhe.2018.09.001. Epub 2018 Oct 17. PubMed PMID: 30342281.
10. Istituto Superiore di Sanità (ISS). Morbillo & Rosolia News. Aggiornamento mensile. Rapporto N° 46 - Novembre 2018. Available from: http://www.epicentro.iss.it/morbillo/bollettino/RM_News_2018_46.pdf [Last accessed 2019, Apr 23].
11. Odone A, Signorelli C. When vaccine hesitancy makes headlines. Vaccine 2017; **35**(9): 1209-10.
12. Ferro A, Odone A, Siddu A, et al. Monitoring the web to support vaccine coverage: results of two years of the portal vaccinarSì. Epidemiol Prev 2015; **39**(4 Suppl 1): 88-93.
13. Lazer D, Kennedy R, King G, Vespignani A. Big data. The parable of Google Flu: traps in big data analysis. Science 2014; 343(6176): 1203-5. doi: 10.1126/science.1248506. PubMed PMID: 24626916.

Corresponding Author: Dr. Provenzano Sandro, Dipartimento Scienze per la Promozione della Salute e Materno Infantile "G. D'Alessandro", Via del Vespro, 129, 90127 Palermo (PA), Italy
e-mail: provenzanosandro@hotmail.it