Acta Biomed 2021; Vol. 92, N. 1: e2021056 DOI: 10.23750/abm.v92i1.9790

Original article

Correlation between flu and Wikipedia's pages visualization

Vincenza Gianfredi[1], *Omar Enzo Santangelo*[2], *Sandro Provenzano*[3]

[1]School of Medicine, Vita-Salute San Raffaele University, Milan, Italy
[2]Azienda Socio Sanitaria Territoriale di Lodi, Lodi, Italy
[3]Azienda Ospedaliera Universitaria Policlinico "P. Giaccone", Palermo, Italia.

Abstract. *Introduction:* This study aimed to assess if the frequency of the Italian general public searches for influenza, using the Wikipedia web-page, are aligned with Istituto Superiore di Sanità (ISS) influenza cases. *Materials and Methods:* The reported cases of flu were selected from October 2015 to May 2019. Wikipedia Trends was used to assess how many times a specific page was read by users; data were extracted as daily data and aggregated on a weekly basis. The following data were extracted: number of weekly views by users from the October 2015 to May 2019 of the pages: Influenza, Febbre and Tosse (Flu, Fever and Cough, in English). Cross-correlation results are obtained as product-moment correlations between the two times series. *Results:* Regarding the database with weekly data, temporal correlation was observed between the bulletin of ISS and Wikipedia search trends. The strongest correlation was at a lag of 0 for number of cases and Flu (r=0.7571), Fever and Cough (r=0.7501). The strongest correlation was at a lag of -1 for Fever and Cough (r=0.7501). The strongest correlation was at a lag of 1 for number of cases and Flu (r=0.7559), Fever and Cough (r=0.7501). *Conclusions:* A possible future application for programming and management interventions of Public Health is proposed.

Key words: Flu; Italy; Internet; Medical Informatics Computing; Vaccine-preventable diseases; Medical Informatics.

Introduction

Influenza is a single strand-RNA viral vaccine-preventable disease that affect millions of people each year and causing thousands of deaths (1). According to the World Health Organization (WHO), approximately 20% of the global population is yearly infected by influenza, and approximately 70.000 people died only considering the European Region (1). Same trend is also registered in Italy, where, according to the National Institute of Health (Istituto Superiore di Sanità, ISS), a range between 4-15% of population is infected yearly (2).

Influenza is still a public health issue not only because of its high incidence rate, but also because its high burden in terms of health-care costs, lost working hours, and premature death (mortality rate ≈ 13 x 100.000, in Italy) (3). However, even if a large proportion of flu burden could be avoided, thanks to a safe vaccine, flu vaccination coverage is still largely below the threshold (4-8), causing periodic epidemic worldwide. In this context, prevention of flu spread is fundamental in order to control disease outbreak. Nevertheless, identification of new cases, through classical surveillance systems, is a critical point because largely affected by under-diagnosis and under-reporting (9). Moreover, traditional surveillance systems are expensive since health-care workers manually enter data, that aggregately collected, are sent to national Health Ministry to be further analyzed. This approach results in a time-lag that can range from a minimum of weeks to several months, that might affect the prompt preventive reaction. At the contrary, novel surveillance systems based on disease-related internet activity traces, using for instance web-page views (most frequently Wikipedia web-pages) (10, 11), social media

posts (12), or search queries (most frequently Google Trends) (13-15) are become even more attractive because faster and cheaper. The hypothesis behind this approach is that an increase in flu cases is followed by a higher number of people who experienced flu symptoms, which in turn corresponds to higher flu related web search (or flu related posts) from the public. These novel surveillance systems statistically associate data from the traditional surveillance to the internet activities, in order to explore public interest and to inform mathematical models that can predict the outbreak going (16). This promising and flourishing science is known as infodemiology or infoveillance and could overcome some of the traditional systems' issues because based on real-time monitoring.

Therefore, the aim of the current study was to assess if the frequency of the Italian general public searches for influenza, using the Wikipedia web-page, are aligned with ISS influenza cases. Even if influenza syndrome might range between few and mild respiratory symptoms to complicated pneumonia requiring hospitalization; typical manifestations are characterized by fever, generally higher than 38°C, accompanied by cough, usually dry, persistent, and lasting 2 weeks or more. For this reason, we mainly focused our analysis using the keywords flu, fever and cough.

Materials and methods

A cross-sectional study design was used. The reported cases of flu were selected from October 2015 to May 2019. Every week from the 42nd week of the current year to the 17th week of the following year the Istituto Superiore di Sanità (ISS) issues a bulletin with the flu cases reported in the previous week (17).

From Wikipedia (18) it is possible to know how many times a specific page is viewed by users, data were extracted as daily data and aggregated on a weekly basis, corresponding to the weeks reported in the ISS bulletins. The following data were extracted:

- number of weekly views by users from the October 2015 to May 2019 of the pages: Influenza, Febbre and Tosse (Flu, Fever and Cough, in English).

The data extracted from Wikipedia have been moved over time (Lag), one week in the future and one week in the past as regards the database with weekly data.

Cross-correlation results are obtained as product-moment correlations between the two times 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). According to a rule of thumb there is a strong correlation if $r > 0.7$, moderate correlation if the value of r is between 0.3 and 0.7 and weak correlation if $r< 0.3$ (19). The statistical significance level for the analyses was 0.05. The data were analyzed using the STATA statistical software, version 14 (20).

Results

Based on results, a temporal correlation was observed between the bulletin of ISS and Wikipedia search trends (Fever, Cough and Flu). Table 1 shows correlation for number of flu reported cases and the search terms of Wikipedia for weeks at Lag 0. A strong correlation was found for: "Number of cases" and "Flu" (r=0.7571), "Fever" and "Cough" (r=0.7501). A moderate correlation was found for: "Number of cases" and "Fever" (r=0.5320), "Fever" and "Flu" (r=0.6822). Weak correlation was found for "Cough" and "Flu" (r=0.3345).

Table 2 shows the number of flu reported cases and the search terms of Wikipedia for weeks at Lag -1. A strong correlation was found for: "Fever" and "Cough" (r=0.7501). A moderate correlation was found for: "Number of cases" and "Flu" (r=0.6956), "Fever" and "Flu" (r=0.6822), "Cough" and "Flu" (r=0.3345).

Table 3 shows correlation for number of flu reported cases and the search terms of Wikipedia for weeks at Lag +1. A strong correlation was found for: "Number of cases" and "Flu" (r=0.7559), "Fever" and "Cough" (r=0.7501). A moderate correlation was found for: "Number of cases" and "Fever" (r=0.5108), "Fever" and "Flu" (r=0.6822), "Cough" and "Flu" (r=0.3345).

Table 1. Number of reported cases of flu and of search terms of Wikipedia, results for weeks at Lag 0. Used Pearson correlation coefficient

		Number of cases	Fever	Cough	Flu
Number of cases	r	1.0000			
	observations	112			
Fever	r	0.5320	1.000		
	p-value	<0.001			
	observations	112	112		
Cough	r	0.1288	0.7501	1.0000	
	p-value	0.1760	<0.001		
	observations	112	112	112	
Flu	r	0.7571	0.6822	0.3345	1.0000
	p-value	<0.001	<0.001	<0.001	
	observations	112	112	112	112

Table 2. Number of reported cases of flu and of search terms of Wikipedia, results for weeks at Lag -1. Used Pearson correlation coefficient

		Number of cases	Fever	Cough	Flu
Number of cases	r	1.0000			
	observations	111			
Fever	r	0.4823	1.000		
	p-value	<0.001			
	observations	111	112		
Cough	r	0.1009	0.7501	1.0000	
	p-value	0.2920	<0.001		
	observations	111	112	112	
Flu	r	0.6956	0.6822	0.3345	1.0000
	p-value	<0.001	<0.001	<0.001	
	observations	111	112	112	112

Table 3. Number of reported cases of flu and of search terms of Wikipedia, results for weeks at Lag +1. Used Pearson correlation coefficient

		Number of cases	Fever	Cough	Flu
Number of cases	r	1.0000			
	observations	111			
Fever	r	0.5108	1.000		
	p-value	<0.001			
	observations	111	112		
Cough	r	0.1294	0.7501	1.0000	
	p-value	0.1757	<0.001		
	observations	111	112	112	
Flu	r	0.7559	0.6822	0.3345	1.0000
	p-value	<0.001	<0.001	<0.001	
	observations	111	112	112	112

Discussion

In this study we found a large correlation between flu cases and Wikipedia search volume for flu, and fever ($p < 0.001$ for both) but not for cough. In particular, the correlation is stronger for flu, and medium when fever is considered. Moreover, strong correlations are also found between keywords. This means that people who are interested in one of these keywords also read the others Wikipedia web pages. In other words, there is a significant correlation between keywords and flu reported cases, and among keywords. This result remains consistent even using different time lag, becoming more stronger when a time lag of 0 was adopted. This result confirms the hypothesis that the spreading of the infection is followed by the increasing public interest on symptoms and disease, rising the internet search volume.

Even if, internet search volume, using different web pages or social media, was largely assessed in other countries (mainly Americas) (16, 21), no previous study assesses the phenomenon in Italy. This is important because one limit of this approach is that results might be biased by the characteristics of the population, cultural aspects, and availability of electronic devices. As for instance, McIver and Brownstein in their study showed as a high media attention on influenza raised the search volume and predate the epidemic up to 2 weeks before the reported cases (22). Sex and age are other important aspects that should be considered. Actually, older people might be less prone to use smartphones or computers or might be less expert in search information on internet. According to the National Institute of Statistics (Istituto Nazionale di Statistica, ISTAT) approximately 10% of people aged 60-70 years regularly use internet(23), however, people over 65 represent a quarter of the total Italian population(24), making Italy one of the most elderly European country. Women more frequently than men search on internet for health-related information, at the same time different level of financial deprivation, education and health literacy might also affect the results (25). However, the search volume of information-seeking through Wikipedia can be considered a good proxy of the general information-seeking behavior (26). This is due to the high accessibility, usability and perceived reliability of Wikipedia, proved by the fact that Wikipedia often rank highly in Google search results (27). Moreover, Wikipedia represents a prominent health information resource being one of the most frequently consulted web-pages for seeking health information (27). Furthermore, Wikipedia is used not only by the public, but even for educational (both by students and health professionals) and research purpose (27). Even if this wide use of Wikipedia might help to deeply understand human behavior, it might also generate some noise signals that may increase difficulties in data interpretation (28). Moreover, Wikipedia has also some limitations. Firstly, the quality and accuracy of the contents, that mainly affect the spread of correct information among general public and students. Secondly, the lower flexibility of Wikipedia that only allowed for counts of page views, reduces the regional or local degree of data interpretation if compared to other online resources as Google trends or Twitter (29). However, since Wikipedia has language-tailored pages and considering that Italian is only spoken in Italy, this makes the analysis of the language-specific pages (as Italian) more accurate than analyses of more international languages (as for instance English or Spanish).

The analysis presented in this paper might also be conducted using the web-page of (inter)national health institutions or local health units (30). This might be relevant for public health workers in order to better understand the public interests and to measure how frequently the public interact with informative and educational materials published (31). These data provide a real-time feed-back extremely useful to plan future health communication campaigns (32). Actually, in our historical context, public health workforce should even more use the internet-based skills (both for communication and for research) in order to produce evidence-based data and practices (33).

Limitations

The study has some limits: the spike of Internet searches 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 Wikipedia does not provide data at these levels. Moreover, temporal and geographic changes in the interface of Wikipedia 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.

Conclusion

To conclude, our results showed the association between number of weekly views of flu, caught and fever Wikipedia web pages and the spread of influenza in Italy, in the period October 2015 - May 2019. These results confirm the important role and usefulness of the infosurveillance systems in public health. Particularly because, providing data on internet research volume they can promptly inform about the spread of infectious diseases. Moreover, inforsurveillance systems offer data in a timeliness and cheapest manner.

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

Author's contribution statement: All individuals listed as authors have substantially contributed to designing, performing or reporting the study.

References

1. World Health Organization. Influenza – estimating burden of disease 2020 [Available from: http://www.euro.who.int/en/health-topics/communicable-diseases/influenza/seasonal-influenza/burden-of-influenza.
2. Epicentro. Influenza, aspetti epidemiologici in Italia 2020 [Available from: https://www.epicentro.iss.it/influenza/epidemiologia-italia.
3. Rizzo C, Bella A, Viboud C, Simonsen L, Miller MA, Rota MC, et al. Trends for influenza-related deaths during pandemic and epidemic seasons, Italy, 1969–2001. Emerg Infect Dis. 2007;13(5):694–9.
4. Alagna E, Santangelo OE, Raia DD, Gianfredi V, Provenzano S, Firenze A. Health status, diseases and vaccinations of the homeless in the city of Palermo, Italy. Ann Ig. 2019;31(1):21–34.
5. Gianfredi V, Nucci D, Salvatori T, Orlacchio F, Villarini M, Moretti M, et al. "PErCEIVE in Umbria": evaluation of anti-influenza vaccination's perception among Umbrian pharmacists. J Prev Med Hyg. 2018;59(1):E14-E9.
6. Rossi D, Croci R, Affanni P, Odone A, Signorelli C. Influenza vaccination coverage in Lombardy Region: a twenty-year trend analysis (1999–2019). Acta Biomed. 2020;91(3-S):141–5.
7. Odone A, Chiesa V, Ciorba V, Cella P, Pasquarella C, Signorelli C. Influenza and immunization: a quantitative study of media coverage in the season of the <<Fluad case>>. Epidemiol Prev. 2015;39(4 Suppl 1):139–45.
8. Signorelli C, Odone A, Miduri A, Cella P, Pasquarella C, Gozzini A, et al. Flu vaccination in elite athletes: A survey among Serie A soccer teams. Acta Biomed. 2016;87(2):117–20.
9. Gianfredi V, Moretti M, Fusco Moffa I. Burden of measles using disability-adjusted life years, Umbria 2013–2018. Acta Biomed. 2020;91(3-S):48–54.
10. Mahroum N, Bragazzi NL, Sharif K, Gianfredi V, Nucci D, Rosselli R, et al. Leveraging Google Trends, Twitter, and Wikipedia to Investigate the Impact of a Celebrity's Death From Rheumatoid Arthritis. J Clin Rheumatol. 2018;24(4):188–92.
11. Provenzano S, Santangelo OE, Giordano D, Alagna E, Piazza D, Genovese D, et al. Predicting disease outbreaks: evaluating measles infection with Wikipedia Trends. Recenti Prog Med. 2019;110(6):292–6.
12. Gianfredi V, Bragazzi NL, Nucci D, Martini M, Rosselli R, Minelli L, et al. Harnessing Big Data for Communicable Tropical and Sub-Tropical Disorders: Implications From a Systematic Review of the Literature. Front Public Health. 2018;6:90.
13. Bragazzi NL, Barberis I, Rosselli R, Gianfredi V, Nucci D, Moretti M, et al. How often people google for vaccination: Qualitative and quantitative insights from a systematic search of the web-based activities using Google Trends. Hum Vaccin Immunother. 2017;13(2):464–9.
14. Gianfredi V, Bragazzi NL, Mahamid M, Bisharat B, Mahroum N, Amital H, et al. Monitoring public interest toward pertussis outbreaks: an extensive Google Trends-based analysis. Public Health. 2018;165:9–15.
15. 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;31(4):385–91.
16. Priedhorsky R, Daughton AR, Barnard M, O'Connell F, Osthus D. Estimating influenza incidence using search query deceptiveness and generalized ridge regression. PLoS Comput Biol. 2019;15(10):e1007165.
17. Istituto Superiore di Sanità. Sistema di Sorveglianza Integrata dell'Influenza 2020 [Available from: https://old.iss.it/site/RMI/influnet/pagine/rapportoInflunet.aspx.
18. Wikipedia. Analisi di visualizzazioni delle pagine 2020 [Available from: https://tools.wmflabs.org/pageviews.
19. Mukaka MM. Statistics corner: A guide to appropriate use of correlation coefficient in medical research. Malawi Med J. 2012;24(3):69–71.

20. StataCorp. Stata Statistical Software. In: Station C, editor.: StataCorp LP; 2015.
21. Hickmann KS, Fairchild G, Priedhorsky R, Generous N, Hyman JM, Deshpande A, et al. Forecasting the 2013–2014 influenza season using Wikipedia. PLoS Comput Biol. 2015;11(5):e1004239.
22. McIver DJ, Brownstein JS. Wikipedia usage estimates prevalence of influenza-like illness in the United States in near real-time. PLoS Comput Biol. 2014;10(4):e1003581.
23. Istituto Nazionale di Statistica. Internet@Italia. Domanda e offerta di servizi online e scenari di digitalizzazione. Rome; 2018.
24. Istituto Nazionale di Statistica. Popolazione e famiglie 2020 [Available from: http://www4.istat.it/it/anziani/popolazione--e-famiglie.
25. Bidmon S, Terlutter R. Gender Differences in Searching for Health Information on the Internet and the Virtual Patient-Physician Relationship in Germany: Exploratory Results on How Men and Women Differ and Why. J Med Internet Res. 2015;17(6):e156.
26. Gabarron E, Lau AY, Wynn R. Is There a Weekly Pattern for Health Searches on Wikipedia and Is the Pattern Unique to Health Topics? J Med Internet Res. 2015;17(12):e286.
27. Smith DA. Situating Wikipedia as a health information resource in various contexts: A scoping review. PLoS One. 2020;15(2):e0228786.
28. Generous N, Fairchild G, Deshpande A, Del Valle SY, Priedhorsky R. Global disease monitoring and forecasting with Wikipedia. PLoS Comput Biol. 2014;10(11):e1003892.
29. Sharpe JD, Hopkins RS, Cook RL, Striley CW. Evaluating Google, Twitter, and Wikipedia as Tools for Influenza Surveillance Using Bayesian Change Point Analysis: A Comparative Analysis. JMIR Public Health Surveill. 2016;2(2):e161.
30. Bragazzi NL, Gianfredi V, Villarini M, Rosselli R, Nasr A, Hussein A, et al. Vaccines Meet Big Data: State-of-the-Art and Future Prospects. From the Classical 3Is ("Isolate-Inactivate-Inject") Vaccinology 1.0 to Vaccinology 3.0, Vaccinomics, and Beyond: A Historical Overview. Front Public Health. 2018;6:62.
31. Gianfredi V, Odone A, Fiacchini D, Rosselli R, Battista T, Signorelli C. Trust and reputation management, branding, social media management nelle organizzazioni sanitarie: sfide e opportunità per la comunità igienistica italiana. J Prev Med Hyg. 2019;60(3):E108-E9.
32. Gianfredi V, Grisci C, Nucci D, Parisi V, Moretti M. [Communication in health.]. Recenti Prog Med. 2018;109(7):374–83.
33. Gianfredi V, Balzarini F, Gola M, Mangano S, Carpagnano LF, Colucci ME, et al. Leadership in Public Health: Opportunities for Young Generations Within Scientific Associations and the Experience of the "Academy of Young Leaders". Front Public Health. 2019;7:378.

Correspondence:
Received: 13 May 2020
Accepted: 10 December 2020
Dr. Omar Enzo Santangelo
Azienda Socio Sanitaria Territoriale di Lodi
piazza Ospitale, 10
Lodi, Italy.
E-mail: omarenzosantangelo@hotmail.it