Column in Laboratory Medicine

Is digital epidemiology reliable?—insight from updated cancer statistics

Giuseppe Lippi[1], **Gianfranco Cervellin**[2]

[1]Section of Clinical Biochemistry, University of Verona, Verona, Italy; [2]Emergency Department, University Hospital of Parma, Parma, Italy
Correspondence to: Prof. Giuseppe Lippi. Section of Clinical Biochemistry, University Hospital of Verona, Piazzale LA Scuro, 37134 Verona, Italy. Email: giuseppe.lippi@univr.it.

Submitted Nov 13, 2018. Accepted for publication Nov 21, 2018.
doi: 10.21037/atm.2018.11.55
View this article at: http://dx.doi.org/10.21037/atm.2018.11.55

Digital epidemiology is an innovative and constantly expanding scientific discipline, whose popularity has been catalyzed by enhanced access to scientific information and digital tools. The most obvious definition of digital epidemiology is that suggested by Marcel Salathé, according to whom "*digital epidemiology is epidemiology based on digital sources data engendered outside public healthcare systems*" (1). The worldwide popularity of this new discipline has constantly increased during the past decades, thanks to better accessibility to Internet-generated information, encompassing data attainable from social media and/or Web-based Search Engines.

Google Trends (Google Inc. Mountain View, CA, United States) is indeed one of the most popular tools currently used in digital epidemiology. This free Web resource has been developed for garnering information on the number of Google searches throughout different periods of time and across different geographical locations (2), so that the final results of a Google Trends analysis for one or more search keywords mirror the overall number of Google searches for those terms. Results are typically expressed in a scale comprised between <1 and 100, where the highest value (i.e., 100) reflects the maximum peak of Google searches (and hence the highest popularity) for the search term, a value <1 reflects a numerically insignificant volume of Google Searches compared to the peak value, whilst the other values comprised between 100 and <1 mirror a proportionally lower popularity of search terms compared to the highest peak.

In the last decade growing evidence has been made available that Google Trends analyses may be a reliable tool for providing estimates of popularity of many diseases and treatments, which globally parallel real-world epidemiology of disease and therapeutics usage (3,4). Notably, Domnich *et al.* recently showed that query-based model in Google trends were capable to accurately predict the peak time of influenza-like illness (5). Similarly, accurate results were obtained by Teng *et al.* for predicting Zika Virus epidemics (6), by Wang *et al.* for forecasting vesicular stomatitis (7), and by Marques-Toledo *et al.* for predicting Dengue outbreaks (8). Beside infectious diseases, the accuracy of Google Trends analysis for gathering insights into real world epidemiology of human diseases remains mostly unexplored or unproven.

According to the recent statistics of the World Health Organization (WHO), cancer is one of the leading causes of death around the world, accounting for an estimated 9.6 million deaths in year 2018. In the specific field of cancer, several articles have been published on public interest on malignancies, but no definitive evidence has been brought that Google Trends analyses may be really helpful to assist establishing the worldwide epidemiology of malignancies. Notably, a recent article published by Phillips *et al.* (9) showed that online Google search volumes were significantly correlated with cancer incidence at the state level in the US, but no information was provided outside that country.

Therefore, to explore whether or not digital epidemiology would actually mirror the real disease epidemiology, updated data on cancer statistics (i.e., for the year 2018) were retrieved from GLOBOCAN (Global Cancer Incidence, Mortality and Prevalence), a project developed by the International Agency for Research on Cancer (IARC), which provides estimates of incidence, prevalence and mortality from 36 different forms of cancer in 185 worldwide countries (10). The data on the number of new cases of the ten most frequent types of cancers (*Table 1*)

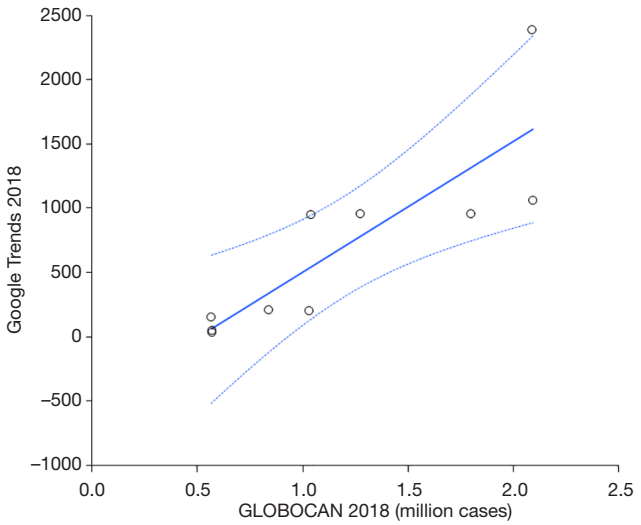

Figure 1 Correlation between GLOBOCAN (Global Cancer Incidence, Mortality and Prevalence) and Google Trends statistics for the ten most frequent forms of cancer.

Table 1 GLOBOCAN (Global Cancer Incidence, Mortality and Prevalence) and Google Trends data for the ten most frequent forms of cancer

Cancer	GLOBOCAN 2018 (million cases)
Lung	2.094
Breast	2.089
Colorectal	1.801
Prostate	1.276
Skin	1.042
Stomach	1.033
Liver	0.841
Esophagus	0.572
Cervix	0.570
Thyroid	0.567

were then compared with those obtained with a Google Trends search limited to the past 12 months (i.e., between 12 November, 2017 and 12 November, 2018), using the search terms "Lung cancer" AND "Breast cancer", AND "Colorectal cancer" AND "Prostate cancer", AND "Skin cancer" AND "Stomach cancer", AND "Liver cancer" AND "Esophagus cancer" AND "Cervix cancer" AND "Thyroid cancer", with no geographical restriction.

The results of this analysis are shown in *Figure 1*. A highly significant correlation [Pearson's correlation, 0.85; 95% confidence interval (CI), 0.46–0.96; P=0.002] was found between GLOBOCAN and Google Trends data. This would actually mean that Google Trends analysis displayed an overall 85% efficiency for predicting the current worldwide cancer incidence. Notably, the major difference concerned the statistics for lung and breast cancer incidence, with GLOBOCAN data being virtually identical for these two forms of cancer, whilst the volume of Google searches was almost double for breast than for lung cancer. This is not really surprising, since the highest peak of Google searches for breast cancer was recorded in October 2018, corresponding to the "WHO Breast Cancer Awareness Month", and cumulating approximately 20% of all Google searches throughout the past 12 months.

Public health epidemiology is conventionally based on information garnered from health care systems, which can only collect data from diagnosed or treated patients, thus generating a virtually incomplete picture. Another important drawback of conventional health epidemiology is that published data are frequently outdated, since it takes quite a long time to collect, pool and analyze information, especially when statistics are based on a large number of worldwide healthcare resources or surveys. On the other hand, online tools have either inherent drawbacks. These typically include local on-line resources availability, which may be still limited in many low-income countries, as well as the fact that on-line resources such as Google Trends might occasionally generate epidemiologic pictures of human diseases that are at least partially different from those originating from more conventional approaches, because Google information may be searched for many other reasons than for personal diseases. Finally, some biases may be observed during episodic peaks of popularity for certain diseases, in concomitance with specific (social) media coverage and health campaigns, such as that earlier illustrated for the "WHO Breast Cancer Awareness Month".

Irrespective of some limitations, the advantages and weaknesses of "conventional" and "digital" epidemiology would lead us to conclude that digital epidemiology not only offers remarkable clue on how a certain disease is perceived by the general public, but it shall also be seen as a suitable and timely instrument for filling some gaps in traditional healthcare epidemiology, allowing expedited responses to public healthcare issues. It is now virtually unquestionable that the role of digital epidemiology could only increase in the foreseeable future, as broadband accessibility will further

widen around the globe, thus substantially improving its accuracy and efficiency. In the era of "big data", and with its inherent limitations, digital epidemiology shall hence be regarded as a valuable complement for more traditional epidemiologic approaches.

Acknowledgements

None.

Footnote

Conflicts of Interest: The authors have no conflicts of interest to declare.

References

1. Salathé M. Digital epidemiology: what is it, and where is it going?. Life Sci Soc Policy 2018;14:1.
2. Kristoufek L. Can Google Trends search queries contribute to risk diversification? Sci Rep 2013;3:2713.
3. 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;7:185-9.
4. Lippi G, Mattiuzzi C, Cervellin G, et al. Direct oral anticoagulants: analysis of worldwide use and popularity using Google Trends. Ann Transl Med 2017;5:322.
5. Domnich A, Panatto D, Signori A, et al. Age-related differences in the accuracy of web query-based predictions of influenza-like illness. PLoS One 2015;10:e0127754.
6. Teng Y, Bi D, Xie G, et al. Dynamic Forecasting of Zika Epidemics Using Google Trends. PLoS One 2017;12:e0165085.
7. Wang J, Zhang T, Lu Y, et al. Vesicular stomatitis forecasting based on Google Trends. PLoS One 2018;13:e0192141.
8. Marques-Toledo CA, Degener CM, Vinhal L, et al. Dengue prediction by the web: Tweets are a useful tool for estimating and forecasting Dengue at country and city level. PLoS Negl Trop Dis 2017;11:e0005729.
9. Phillips CA, Barz Leahy A, Li Y, et al. Relationship Between State-Level Google Online Search Volume and Cancer Incidence in the United States: Retrospective Study. J Med Internet Res 2018;20:e6.
10. Bray F, Ferlay J, Soerjomataram I, et al. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J Clin 2018;68:394-424.

Cite this article as: Lippi G, Cervellin G. Is digital epidemiology reliable?—insight from updated cancer statistics. Ann Transl Med 2019;7(1):15. doi: 10.21037/atm.2018.11.55

