Epidemiology and Infection

cambridge.org/hyg

Prediction of Zika-confirmed cases in Brazil and Colombia using Google Trends

S. Morsy[1,2,*], T.N. Dang[2,3,*], M.G. Kamel[2,4], A.H. Zayan[2,5], O.M. Makram[2,6], M. Elhady[2,7], K. Hirayama[8] and N.T. Huy[9,10]

[1]Medical Biochemistry and Molecular Biology Department, Faculty of Medicine, Tanta University, Tanta, Egypt; [2]Online Research Club (http://www.onlineresearchclub.org); [3]University of Medicine and Pharmacy, Ho Chi Minh City, Vietnam; [4]Faculty of Medicine, Minia University, Minia 61519, Egypt; [5]Faculty of Medicine, Menoufia University, Menoufia, Egypt; [6]Faculty of Medicine, October 6 University, 6th October City, Egypt; [7]Department of Pediatrics, Faculty of Medicine (for girls), Al-Azhar University, Cairo 11651, Egypt; [8]Department of Immunogenetics, Institute of Tropical Medicine (NEKKEN), Leading Graduate School Program, and Graduate School of Biomedical Sciences, Nagasaki University, 1-12-4 Sakamoto, Nagasaki 852-8523, Japan; [9]Evidence Based Medicine Research Group & Faculty of Applied Sciences, Ton Duc Thang University, Ho Chi Minh City 700000-760000, Vietnam and [10]Department of Clinical Product Development, Institute of Tropical Medicine (NEKKEN), Leading Graduate School Program, and Graduate School of Biomedical Sciences, Nagasaki University, 1-12-4 Sakamoto, Nagasaki 852-8523, Japan

Short Paper

*Authors equally contributed to the manuscript.

Cite this article: Morsy S, Dang TN, Kamel MG, Zayan AH, Makram OM, Elhady M, Hirayama K, Huy NT (2018). Prediction of Zika-confirmed cases in Brazil and Colombia using Google Trends. *Epidemiology and Infection* **146**, 1625–1627. https://doi.org/10.1017/S0950268818002078

Received: 23 June 2017
Revised: 30 May 2018
Accepted: 27 June 2018
First published online: 30 July 2018

Key words:
Brazil; Colombia; Google Trends; prediction; Zika

Author for correspondence:
Nguyen Tien Huy, E-mail: nguyentienhuy@tdt.edu.vn

Abstract

Zika virus infection in humans has been linked to severe neurological sequels and foetal malformations. The rapidly evolving epidemics and serious complications made the frequent updates of Zika virus mandatory. Web search query has emerged as a low-cost real-time surveillance system to anticipate infectious diseases' outbreaks. Hence, we developed a prediction model that could predict Zika-confirmed cases based on Zika search volume in Google Trends. We extracted weekly confirmed Zika cases of two epidemic countries, Brazil and Colombia. We got the weekly Zika search volume in the two countries from Google Trends. We used standard time-series regression (TSR) to predict the weekly confirmed Zika cases based on the Zika search volume (Zika query). The basis TSR model – using 1-week lag of Zika query and using 1-week lag of Zika cases as a control for autocorrelation – was the best for predicting Zika cases in Brazil and Colombia because it balanced the performance of the model and the advance time in the prediction. Our results showed that we could use Google search queries to predict Zika cases 1 week earlier before the outbreak. These findings are important to help healthcare authorities evaluate the outbreak and take necessary precautions.

Zika virus infection in humans has changed in character from an endemic self-limited mild illness to an epidemic disease [1]. Developing accurate tools to predict Zika infection spread is required for early prevention of the disease [2]. The purpose of this analysis is to explore whether web-based query could effectively predict Zika virus spread.

On 2 October 2016, Pan American Health Organisation (PAHO) released an epidemiological report of Zika virus in different countries [3]. Each report contains the number of confirmed and suspected cases in each country as reported by Ministry of Health in these countries. For our analysis, we selected Colombia and Brazil because there was continuous monitoring for both confirmed and suspected cases. In addition, both countries were considered as most epidemic countries in South America. PAHO report for Brazil included both suspected and confirmed cases from January 2016 to 9 July 2016, which corresponds to the first epidemiologic week of 2016 till the 27th epidemiologic week of 2016. For Colombia, the report had data from 9 August 2015 to 21 May 2016, which corresponds to the 32nd epidemiologic week of 2015 to the 20th epidemiologic week of 2016. We used Webplotdigitiser software to extract the weekly confirmed Zika cases of Brazil and Colombia [4]. We only extracted confirmed cases not suspected nor reported cases to avoid overestimation of the epidemic. That is because the case definitions for Zika suspected included rashes with one of the following symptoms: fever, usually <38.5 °C, conjunctivitis (non-purulent/hyperaemic), arthralgia, myalgia and peri-articular oedema with the history of travelling to one of the epidemic areas [5]. These criteria are similar to many infectious diseases that caused reporting of a huge number of Zika cases, mainly suspected cases, while confirmed Zika cases represented only a minimum of these numbers. This can be proved by epidemiological reports released by PAHO in October 2016 in which we noticed a big difference between reported and confirmed cases. We used only confirmed cases to avoid overestimation of the epidemic because we have noticed a big difference between confirmed and reported cases that will affect our results.

To get the web search volume for the word Zika in this specific time period, we used Google Trends (https://trends.google.com/trends/) to get the weekly search volume for word 'Zika',

CAMBRIDGE UNIVERSITY PRESS

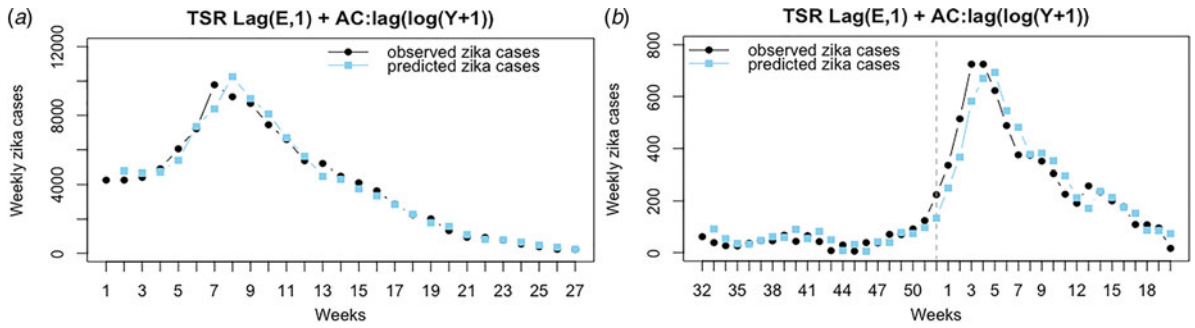

Fig. 1. The figure shows the pattern of observed Zika cases and predicted Zika cases using the model TSR lag (E, 1) + AC: lag (log ($Y+1$), 1) in Brazil (a) and Colombia (b). (a) Brazil, basis TSR model with lag one of Zika query as a predictor and the lag one of log value of Zika case as controlling for the auto-correlation. (b) Colombia, basis TSR model with lag one of Zika query as a predictor and the lag one of log value of Zika case as controlling for the auto-correlation. The vertical line defines years 2015 and 2016.

termed Zika query. We did not use other words for signs and symptoms of Zika because it was similar to other diseases that can cause misjudgement of search volume. The steps of searching the Google Trends and processing the query data for the analysis are explained in the Supplementary video 1.

We used a standard time-series regression (TSR), particularly the Poisson distributed lag model (PDLM) to examine the association between weekly Zika cases (i.e. the outcome) and weekly Zika query (i.e. the predictor). A quasi-Poisson distribution of the outcome is assumed to account for the overdispersion (the presence of expected increasing variance among the data). We also considered important features of the application of TSR to infectious diseases, such as the lag association (e.g. the last week Zika query could be associated with this week Zika cases), the strong auto-correlations and the controlling for the long-term trend. These features are discussed in detail in Imai *et al.* [6]. This model has been considered the best in the prediction of dengue cases when compared with other models including standard multiple regression model (SMR) and seasonal autoregressive-integrated moving average model (SARIMA) [7].

The general model is specified as follow:

$$Y_t \sim \text{quasi} - \text{Poisson}(\mu_t)$$

$$\begin{aligned} \log \mu_t &= \alpha + \beta_1 \text{ Lag } E_{t-k} + \beta_2 \text{time} + \beta_{\text{AC}} \text{AC} \\ &= \text{Basis TSR} + \beta_{\text{AC}} \text{AC}, \end{aligned} \quad (1)$$

where Y_t is the weekly Zika count on week t, μ_t is the mean parameter of the Poisson distribution, alpha (α) is the intercept, and Lag E_{t-k} is the Zika query in week t minus lag k ($k = 0, 1, 2, 3$).

Time is a variable that takes consecutive numbers ranging from 1 on the day on which observations began to 27 on the final day of the observation period in Brazil data, and to 41 in Columbia data. The time variable was used to control the long-term trend in Zika cases (assumed an increase linear trend) following Bhaskaran *et al.*'s method [8]. AC stands for the auto-correlation term. We invite the reader to refer Imai *et al.* for the nature of the technical details of this model [6].

We used R software version 3.4.3 for all the described analyses [9]; we used Epi [10], tsModel [11] and bbmle [12] packages.

In total, seven different models were constructed, and the performance of them was validated based on the dispersion value, which was used for the evaluation of the model as reported by Imai *et al.*[6] (i.e. the smaller the dispersion value, the better the model in predicting Zika cases). The seven constructed models for each country with their dispersion values are described in Supplementary Tables S1 and S2.

In addition, we also conducted a sensitivity analysis to determine whether the results were dependent on modelling choices. We replaced the time variable by the peak indicator variable (i.e. two values: 1 indicates high-peak weeks, 0: otherwise). The high-peak weeks were defined as the weeks containing Zika case counts greater than the median value of Zika case counts of the whole study period.

The best model in predicting Zika cases in Brazil was the model with basis TSR, including lag zero of Zika query plus lag one of Zika cases as controlling for auto-correlation (i.e. TSR lag (Zika, 0) + AC: lag (log ($Y+1$), 1)) (Supplementary Table S1). Whereas the model with basis TSR, including lag one of Zika query plus lag one of Zika cases as controlling for auto-correlation came into second (i.e. TSR lag (Zika, 1) + AC: lag (log ($Y+1$), 1)). Similarly, the best model in predicting Zika cases in Colombia is TSR lag (Zika, 0) + AC: lag (log ($Y+1$), 1), and the model TSR lag (Zika, 1) + AC: lag (log ($Y+1$), 1) took second place (Supplementary Table S2).

For the real application, the model that can predict Zika cases in future would be preferable. Therefore, in this study, we would recommend using the model TSR lag (Zika, 1) + AC: lag (log (Y + 1), 1) in predicting Zika cases in Brazil and Colombia because it balanced the performance of the model and the advance time of prediction. The pattern of observed Zika cases and predicted Zika cases using the model TSR lag (Zika, 1) + AC: lag (log (Y + 1), 1) in Brazil and Colombia is shown in the (Fig. 1). The correlation coefficients are 0.986 and 0.918 in Brazil and Colombia, respectively, indicating a good predictive capacity of the models. The results of sensitivity analysis were consistent with the results of the original models, suggesting that our results are robust and not likely affected by modelling choices.

Our study explored the possibility to use Google Trends as a low-cost available Zika bio-surveillance system in developing countries. Our model was robust for the prediction of Zika in the two countries 1 week in advance, which can help to activate timely vector control by local authorities, and community-based preventive measures. It has been shown that Zika followed the same time period and geographic distribution of dengue and Chikungunya viruses in Brazil [13, 14, 15]. This is because of the concurrent transmission of these viruses by the same vector. In addition, the model can be used for monitoring other arboviral

diseases. After current tropical urbanisation, increasing global transportation and global warming, there is a spread of *Aedes spp.* to other regions in the world [16]. With the presence of these vectors plus the circulating arboviruses in human blood, this will be adequate for another arboviral-emerging disease [16]. More arboviral diseases are expected in the literature to be the next global outbreak including Venezuelan equine encephalitis virus, Mayaro and Oropouche [17]. Venezuelan equine encephalitis virus had the same symptoms of Zika including rash, fever, headache, myalgia and arthralgia. The similarity between the symptoms of Zika, Chikungunya and Mayaro virus can lead to misdiagnosis of these diseases as Zika.[18] Theoretically, the similarity between viruses can result in an abnormal increase in search volume or at least change in the trend which will give an initial overview of the state of arboviral circulation. Hence, the model can reflect the status of arboviruses in these two countries. Yet, more research is needed to confirm this theory. With no research tool to discover the epidemic potential of these arboviruses, monitoring Zika can help predicting the status of arboviruses.

Prediction of Zika cases using Google Trends was investigated in previous papers [19, 20]. They used the suspected cases of Zika, and correlated the Zika-related Google searches, Twitter microblogs and HealthMap news reports with the suspected cases of Zika in Colombia, El Salvador, Honduras, Venezuela and Martinique. In our study, however, we used the confirmed Zika cases for correlation and prediction which will give more reliable and consistent results. Another point of our study is the source of data. Our data were directly extracted from PAHO reports, which is considered far more reliable than Twitter microblogs and HealthMap. We tried these data before and we found an overlapped and duplicate data that were immediately discarded, and we decided to depend only on official reports provided in PAHO. For the statistical model, McGough *et al.* used LASSO regression model for prediction whereas, we used the PDLM. Phung *et al.* [7] validated the three different models comprising: SMR, SARIMA and PDLM for the prediction of dengue cases and they found that PDLM was the most accurate for prediction.

In conclusion, we could use Zika query to predict Zika cases 1 week in advance, which provides a useful tool for monitoring and controlling Zika outbreaks.

Supplementary material. The supplementary material for this article can be found at https://doi.org/10.1017/S0950268818002078

Acknowledgements. The authors would like to express our gratitude to the reviewer for his/her time and efforts in reviewing and enriching this paper's quality.

Funding. No funding was received for this work.

Conflict of interest. None.

References

1. **Kindhauser MK *et al.*** (2016) Zika: the origin and spread of a mosquito-borne virus. *Bulletin of the World Health Organization* **94**, 675–686C.
2. **Huff A *et al.*** (2016) FLIRT, a web application to predict the movement of infected travelers validated against the current Zika virus epidemic. *International Journal of Infectious Diseases* **53**, 97–98.
3. **Mitchell C and Mitchell C**. (2017) *PAHO WHO | Countries and Territories with Autochthonous Transmission in the Americas Reported in 2015-2017. Pan American Health Organization/World Health Organization*, Washington, DC, USA (Accessed October 2016).
4. **Rohatgi A. WebPlotDigitizer, 4.1 ed, 2018 (Accessed October 2016).**
5. **Sanchez JD and Sanchez JD**. (2016) PAHO WHO | Zika resources: case definitions. *Pan American Health Organization/World Health Organization*, Washington, DC, USA (Accessed October, 2016).
6. **Imai C *et al.*** (2015) Time series regression model for infectious disease and weather. *Environmental Research* **142**(Suppl. C), 319–327.
7. **Phung D *et al.*** (2015) Identification of the prediction model for dengue incidence in Can Tho city, a Mekong Delta area in Vietnam. *Acta Tropica* **141**, 88–896.
8. **Bhaskaran K *et al.*** (2013) Time series regression studies in environmental epidemiology. *International Journal of Epidemiology* **42**, 1187–1195.
9. **Team RC. R** (2017) A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. In: R Core Team.
10. **Bendix Carstensen MPELMH** (2018) Epi: A Package for Statistical Analysis in Epidemiology. R package version 2.24.
11. **McDermott RDP, with contributions from A. tsModel: Time Series Modeling for Air Pollution and Health. R package version 0.6. In (2013).**
12. **Team BB and Core RD** (2017) bbmle: Tools for General Maximum Likelihood Estimation. R package version 1.0.20. In.
13. **Benelli G and Mehlhorn H** (2016) Declining malaria, rising of dengue and Zika virus: insights for mosquito vector control. *Parasitology Research* **115**, 1747–1754.
14. **Cardoso CW *et al.*** (2015) Outbreak of exanthematous illness associated with Zika, chikungunya, and dengue viruses, Salvador, Brazil. *Emerging Infectious Diseases* **21**, 2274–2276.
15. **Roth A *et al.*** (2014) Concurrent outbreaks of dengue, chikungunya and Zika virus infections – an unprecedented epidemic wave of mosquito-borne viruses in the Pacific 2012–2014. *Eurosurveillance* **19**, 20929.
16. **Weaver SC and Reisen WK**. (2010) Present and future arboviral threats. *Antiviral Research* **85**, 328.
17. **Rodríguez-Morales AJ *et al.*** (2017) Mayaro, Oropouche and Venezuelan equine encephalitis viruses: following in the footsteps of Zika? *Travel Medicine and Infectious Disease* **15**, 72–73.
18. **Paniz-Mondolfi AE *et al.*** (2016) Chikdenmazika syndrome: the challenge of diagnosing arboviral infections in the midst of concurrent epidemics. *Annals of Clinical Microbiology and Antimicrobials* **15**, 42.
19. **McGough SF *et al.*** (2017) Forecasting Zika incidence in the 2016 Latin America outbreak combining traditional disease surveillance with search, social media, and news report data. *PLoS Neglected Tropical Diseases* **11**, e0005295.
20. **Teng Y *et al.*** (2017) Dynamic forecasting of Zika epidemics using google trends. *PLoS ONE* **12**, e0165085.