# Dynamic Forecasting of Zika Epidemics Using Google Trends

Yue Teng[1,2☯]*, Dehua Bi[2,3☯], Guigang Xie[2☯], Yuan Jin[2,4], Yong Huang[1,2], Baihan Lin[5], Xiaoping An[1,2], Dan Feng[6]*, Yigang Tong[1,2]*

1 Beijing Institute of Microbiology and Epidemiology, Beijing, China, 2 State Key Laboratory of Pathogen and Biosecurity, Beijing, China, 3 Department of Mechanical and Mechatronics Engineering, University of Waterloo, Waterloo, Ontario, Canada, 4 Beijing Institute of Biotechnology, Beijing, China, 5 Computational Neuroscience Program, Department of Psychology, Physics, and Computer Science and Engineering; Institute for Protein Design, University of Washington, Seattle, United States of America, 6 Division of Standard Operational Management, Institute of Hospital Management, Chinese PLA General Hospital, Beijing, China

☯ These authors contributed equally to this work.
* yueteng@sklpb.org (YT); danfeng@sklpb.org (DF); tong.yigang@gmail.com (YT)
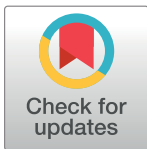
Abstract

We developed a dynamic forecasting model for Zika virus (ZIKV), based on real-time online search data from Google Trends (GTs). It was designed to provide Zika virus disease (ZVD) surveillance and detection for Health Departments, and predictive numbers of infection cases, which would allow them sufficient time to implement interventions. In this study, we found a strong correlation between Zika-related GTs and the cumulative numbers of reported cases (confirmed, suspected and total cases; *p<0.001*). Then, we used the correlation data from Zika-related online search in GTs and ZIKV epidemics between 12 February and 20 October 2016 to construct an autoregressive integrated moving average (ARIMA) model (0, 1, 3) for the dynamic estimation of ZIKV outbreaks. The forecasting results indicated that the predicted data by ARIMA model, which used the online search data as the external regressor to enhance the forecasting model and assist the historical epidemic data in improving the quality of the predictions, are quite similar to the actual data during ZIKV epidemic early November 2016. Integer-valued autoregression provides a useful base predictive model for ZVD cases. This is enhanced by the incorporation of GTs data, confirming the prognostic utility of search query based surveillance. This accessible and flexible dynamic forecast model could be used in the monitoring of ZVD to provide advanced warning of future ZIKV outbreaks.

Introduction

Zika virus (ZIKV) is transmitted to people primarily by mosquitoes [1]. Prior to 2015, outbreaks had occurred in Africa, Southeast Asia, and the Pacific Islands [2, 3, 4]. In May 2015, the presence of Zika virus disease (ZVD) was confirmed in Brazil. ZIKV has subsequently reportedly been spreading throughout the Americas, with epidemics occurring in many countries [5, 6]. The World Health Organization declared ZIKV, and its suspected link to birth defects, an international public health emergency in February 2016 [7, 8]. Traditional,

healthcare-based and government-implemented, ZVD monitoring is resource intensive and slow. Early surveillance of infectious disease prevalence, when followed by an urgent response, can reduce the effects of disease outbreaks [9]. Surveillance of online behavior, such as queries in search engines, is a potential web-based disease detection system that can improve monitoring [10]. Google Trends has been shown to have the potential to go beyond early detection and forecast future influenza and Dengue outbreaks [11–14]. Several studies have used autoregressive integrated moving average (ARIMA) models for the forecasting of influenza prevalence from Google Flu Trends [15, 16]. The real-time nature of GTs surveillance and the demonstrated strong correlation of GTs with infectious disease mean GTs offers a potential tool for timely epidemic detection and prevention [17]. However, the forecasting capabilities of GTs for ZIKV outbreaks remain unknown. In this study, we examined the ZIKV-related GTs temporally correlated with ZVD epidemics, and developed an improved dynamic forecasting method for ZVD activity in the worldwide using an ARIMA model to predict future patterns of ZIKV transmission.

Materials and Methods

Data collection and statistical analysis

Google Trends, an online tracking system of Internet hit-search volumes (Google Inc.), was used to explore web behavior related to the ZIKV outbreaks. GTs data for ZIKV in worldwide was mined of the key word "Zika" from 12 February 2016 to 9 November 2016 (Yearly EPI Week 6 to 45) to cover the 2016 period of the ZIKV epidemic, and was downloaded directly from https://www.google.com/trends/explore?date=all&q=zika on 9 November 2016. Although Google Trends normalizes the search data with the day having the most searches set equal to 100, we obtained and analyzed the relative search volumes for each day based on the data (100) in 12 February 2016 (Yearly EPI Week 6; data shown in S1 Table). The number of ZIKV confirmed, suspected and total cases in the worldwide were retrieved from the PAHO (Pan American Health Organization), available at http://www.paho.org/hq/ and WHO (World Health Organization), available at http://www.who.int/emergencies/zika-virus/situation-report/en/ (last accessed on 9 November 2016, S2 Table). To detect the cumulative GTs volumes relative to reported ZIKV cases, we used the Pearson Product-Moment Correlation to assess linear correlation. All calculations were performed in Python 2.7 with the Scipy library.

Linear regression model

As a baseline model for comparison, the data is fitted with a linear regression model to establish the relationship of the cases to the GTs data. The linear model is constructed with R version 2.14 (http://www.r-project.org/) and the parameters are obtained automatically. Prediction results are plotted together with the proposed ARIMA model to show the comparison outcome.

Reconstructed ARIMA model

For the time series analysis, we fitted an autoregressive integrated moving average (ARIMA) (0, 1, 3) model by using R version 2.14 (http://www.r-project.org/). The autoregressive integrated moving average (ARIMA) forecasting model in this study was developed from the training and testing sets that were extracted from the data sets. We used the data set including the 9 months from 12 February to 9 November 2016 (Yearly EPI Week 6 to 45) in the analysis. The data from 12 February to 25 August 2016 (Yearly EPI Week 6 to 34) were used for training the forecasting model, and the validation was performed on the 8 weeks' data from 1 September to 20 October 2016 (Yearly EPI Week 35 to 42). We applied the Box-Jenkins approach to

ARIMA (p, d, q) modeling of time series. This model-building process is designed to take advantage of associations in the sequentially lagged relationship that usually exists in periodically collected data [18, 19]. The parameter for the model includes p, the order of autoregression; d, the degree of difference and q, the order of moving average. Several models were initially considered in our study (S3 Table). The residual analysis and the Akaike Information Criterion (AIC) was used to compare the goodness-of-fit among the ARIMA models. The Ljung-Box test was used to measure the ACF of the residuals (S3 Table). Using this model selection procedure, we selected one non-seasonal difference term for stationary (d) and three lags of moving average terms (q), resulting in a model of ARIMA (0, 1, 3). To predict the future values, the developed ARIMA model was fitted to the entire data from 12 February to 20 October 2016 (Yearly EPI Week 6 to 42) and used to forecast over a time span of 3 weeks, covering 27 October and 9 November 2016 (Yearly EPI Week 43 and 45).

Results

Correlations between data on ZIKV outbreaks and ZIKV-related GTs

Our analyses used the data from 12 February to 9 November 2016 (Yearly EPI Week 6 to 45), covering 9 months of the GTs data (Fig 1 and S1 Table) and the reported ZIKV epidemic data (confirmed, suspected and total cases, S2 Table). Alongside the "stepped" increases in the
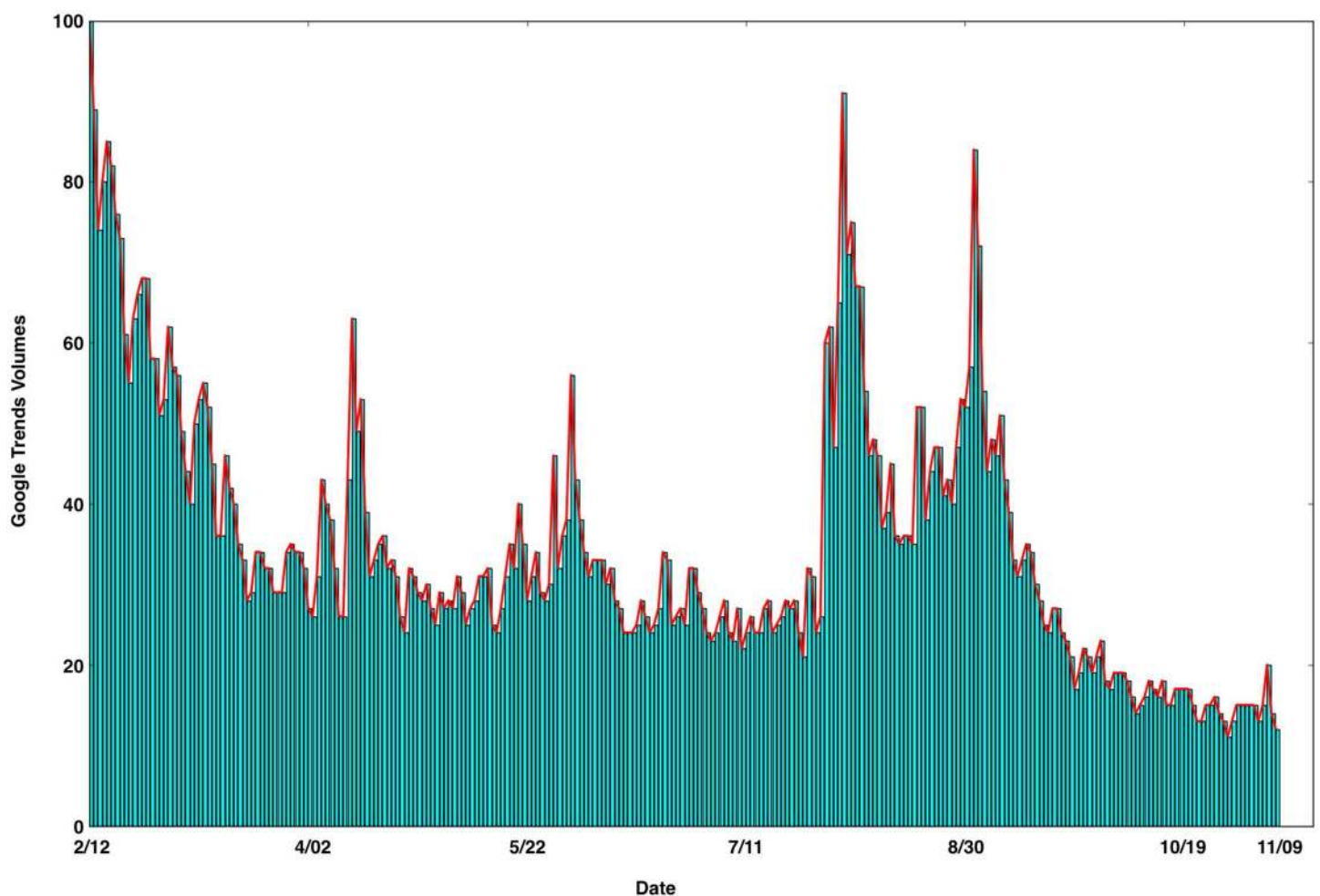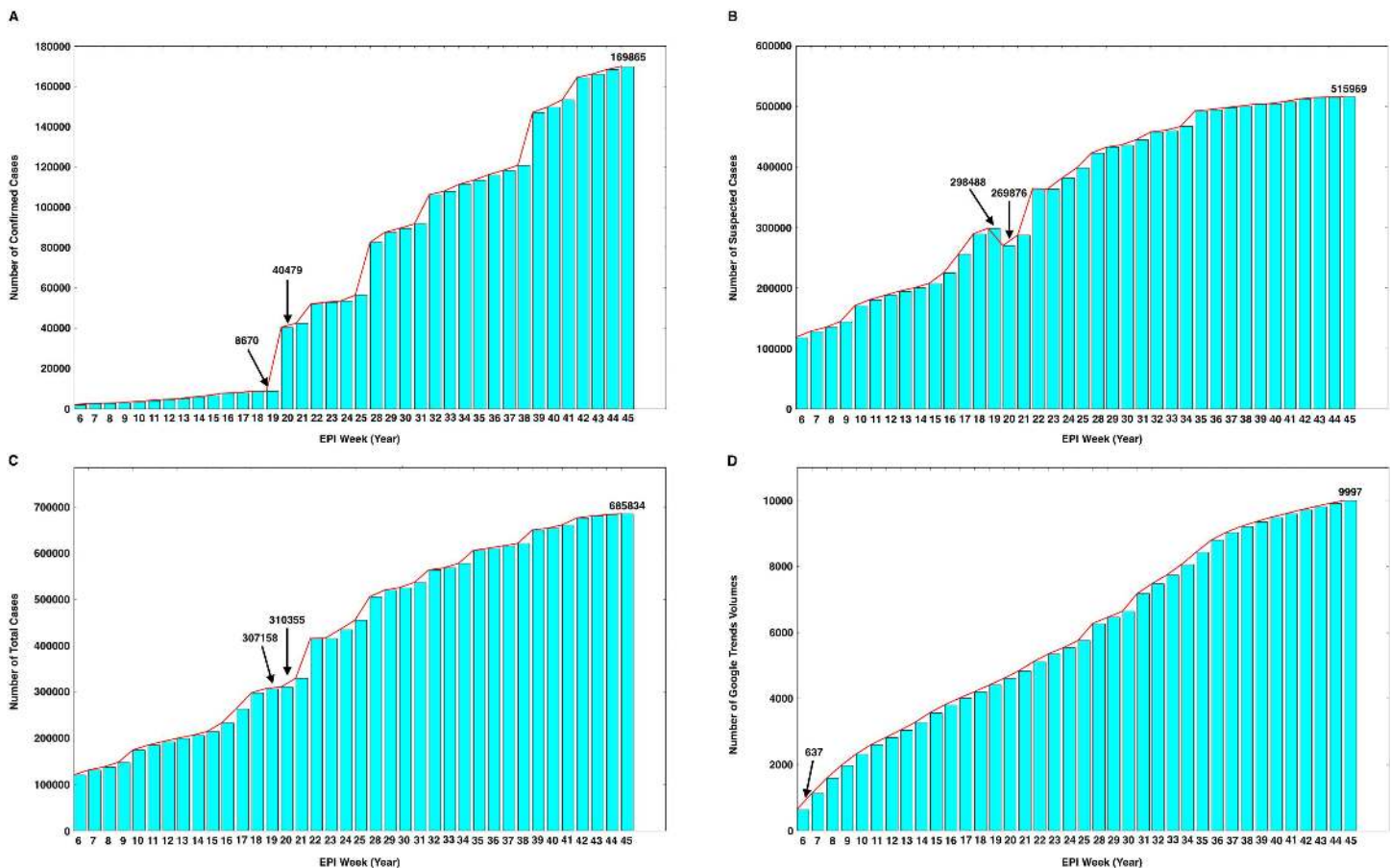

Fig 1. Time series plot of the daily GTs volumes from 12 February to 9 November 2016 (Yearly EPI Week 6 to 45).

doi:10.1371/journal.pone.0165085.g001

reported ZIKV cases (confirmed, suspected and total cases) in this epidemic, there were dramatically increased numbers of Zika-related online searches in Google (Fig 2). From 12 May to 19 May 2016 (Yearly EPI Week 19 to 20), the ZIKA epidemics entered a period of rapid growth with a large cumulative number of confirmed cases (increasing from 8,670 to 40,479; Fig 2A). Meanwhile, the cumulative number of suspected cases fell from 298,488 to 269,876 (Fig 2B), because the suspected cases were diagnosed during this period. Then, the number of total reported cases increased slightly from 307,158 to 310,355 (Fig 2C). Until 20 October 2016 (Yearly EPI Week 42), a total of 164,352 confirmed cases and 512,345 suspected cases were reported in the S2 Table. The GTs data showed that the volume of dynamic Zika-related online searches continuously grew from February to November 2016 (Fig 2D). We performed Pearson Product-Moment Correlation analyses to examine temporal correlations between the accumulative volumes of Zika-related search queries and the cumulative numbers of reported cases. The result indicated that the data on Zika-related GTs had statistically significant and positive correlations with the cumulative numbers of confirmed cases (R = 0.968, $p<0.001$), suspected cases (R = 0.980, $p<0.001$) and total cases (R = 0.988, $p<0.001$) of ZIKV.

Statistical machine learning and reconstructed ARIMA model

On the basis of the revealed correlation between the data of GTs and the number of reported cases, we split the data into training (75%, from 12 February to 25 August 2016) and testing

Fig 2. Time series plots of the number of reported ZIKV confirmed cases (A), suspected cases (B), total cases (C) and the accumulated Google search volumes (D) during the ZVD epidemic from 12 February to 9 November 2016 (Yearly EPI Week 6 to 45).

doi:10.1371/journal.pone.0165085.g002
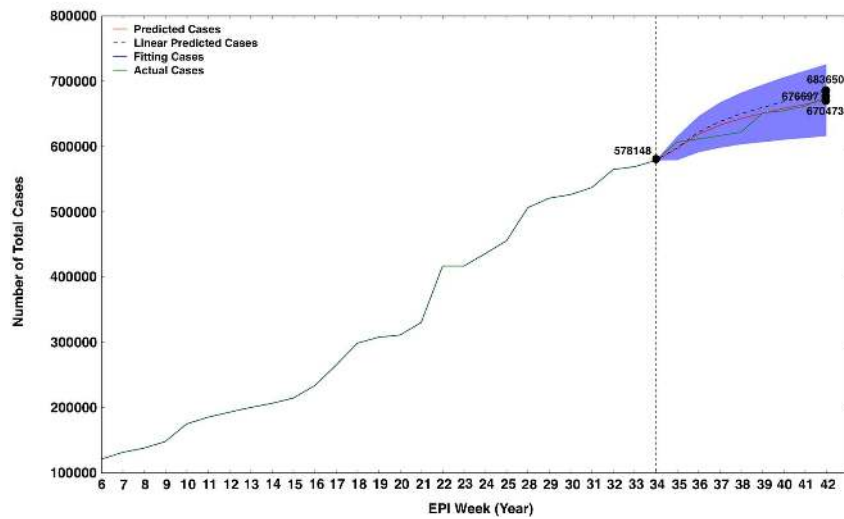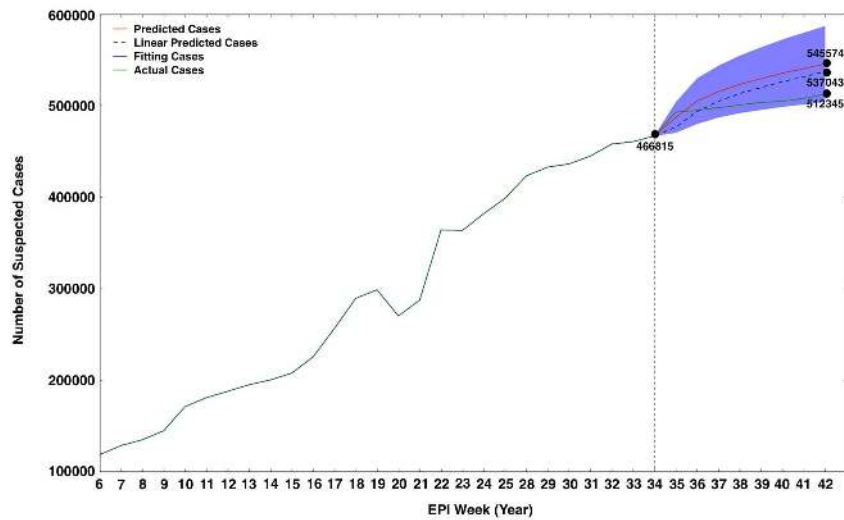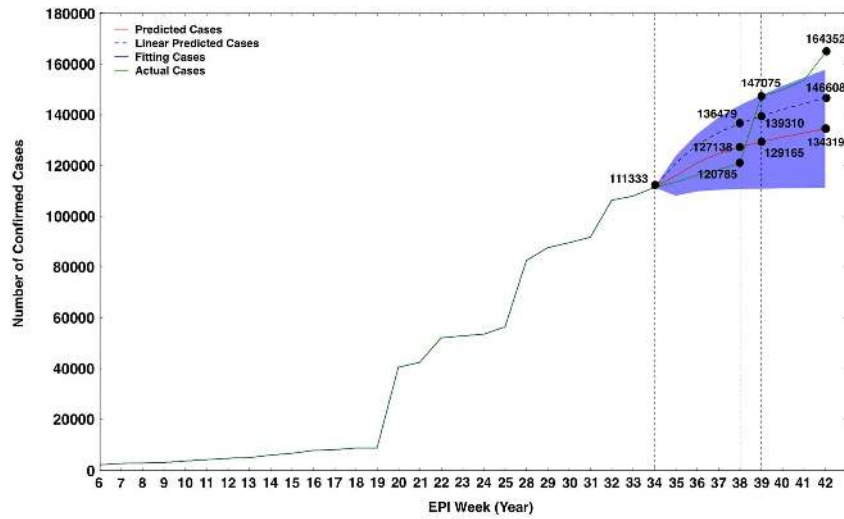
Fig 3. Numbers of reported confirmed cases (A), suspected cases (B) and total cases (C) in the testing set compared with the simulation data by the advanced ARIMA (0, 1, 3) model for training set using the data of Google Trends as the external regressor. The red, blue and green solid lines represent the predicted, training and actual number of cases, respectively. The black dash lines represent the prediction of the linear baseline model. The blue region represents the 95% confidence interval predicted by the ARIMA (0, 1, 3) model.

(25%, from 26 August to 20 October 2016) sets. We analyzed the training set to observe whether the simulation data is similar to the reported data in testing set using the advanced autoregressive integrated moving average (ARIMA (0, 1, 3)) model. In this reconstructed model, we used the collected online search data as the external regressor in a prediction model to assist the historical ZVD epidemic data in improving the quality of prediction. Using this model with the training data of GTs as a predictor, we found that the cumulative number of reported confirmed case (164,352) in 20 October 2016 (Yearly EPI Week 42) in the testing set was above the 95% confidence interval of the simulated data (111,119 to 157,663; Fig 3A). However, Fig 3B and 3C showed that the number of suspected cases (512,345) and the number of the total cases (676,697) in 20 October 2016 (Yearly EPI Week 42) in the testing set fell within the 95% confidence interval of the simulated data (504,034–587,115 for suspected cases and 615,347–725,599 for total cases, respectively). We also used a simple baseline model to do a sanity test as a reference, which could reproduce features in the data and predicted the number of confirmed, suspected and total cases in 20 October 2016 (Yearly EPI Week 42; 146,608 for confirmed cases, 537,043 for suspected cases and 683,650 for total cases in Fig 3A, 3B and 3C). However, the complicated ARIMA model showed a better performance for the prediction with a lower value of AIC than the simple linear predictor (S3 Table). These observations implied that the Google Trends information would improve the prediction of the size of ZIKV outbreaks.

Forecasting the ZIKV outbreaks in worldwide during early november 2016

Based on the data on Zika-related online searches and ZIKV epidemics in GTs between 12 February and 20 October 2016, we used the reconstructed ARIMA (0, 1, 3) model to forecast ZVD outbreaks during the current 3 weeks. In this model, we used the online search data as the external regressor to enhance the forecasting model and assist the historical ZVD epidemic data in improving the quality of the predictions while responding to disease outbreaks. We forecasted the cumulative number of ZIKV cases as a simulation of the continuation of the time series. This model with GTs data as the predictor estimated that the forecast cumulative number of confirmed cases would increase to reach 169,722.4 (95% CI: 154,746.8–184,698.0) by 9 November (Yearly EPI Week 45), which was closest to the actual reported infection cases (169,865; Fig 4A). The predicted number of suspected cases would grow up to 524,231.6 (95% CI: 498,863.8–549,599.4) on 9 November 2016 (Yearly EPI Week 45; Fig 4B), which was slightly more than the reported suspected cases (515,969). Finally, the actual number of total reported cases was 685,834 on 9 November 2016 (Yearly EPI Week 45) and the predicted number of total cases was 693,106 (95% CI: 662,396.1–723,815.9) (Fig 4C), which was above the reported cases (S4 Table). These forecasts of ZVD outbreaks suggested that ZIKV disease transmission in the worldwide remains intense during November 2016.

Conclusions

ZVD outbreaks are now a common and growing problem worldwide [20, 21 22]. Delays in traditional surveillance systems limit the ability of public health agencies to respond efficiently to ZIKV epidemics [23]. Because data on GTs are collected and processed in near real-time,
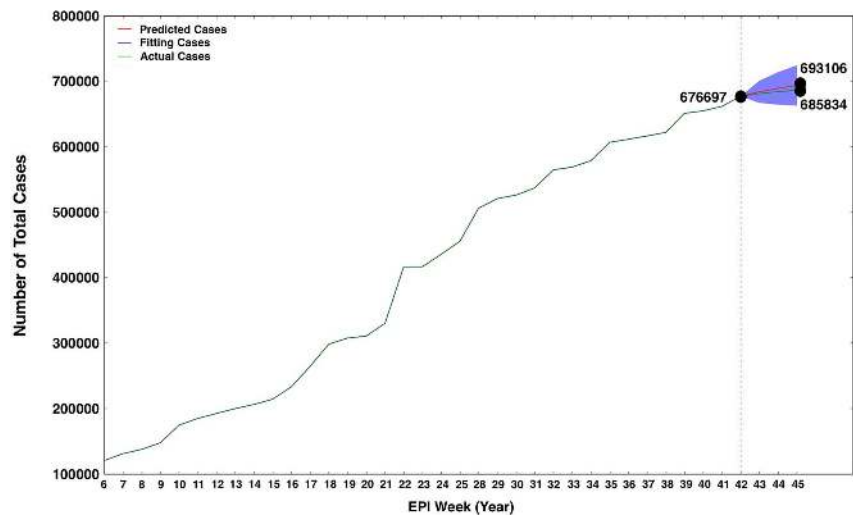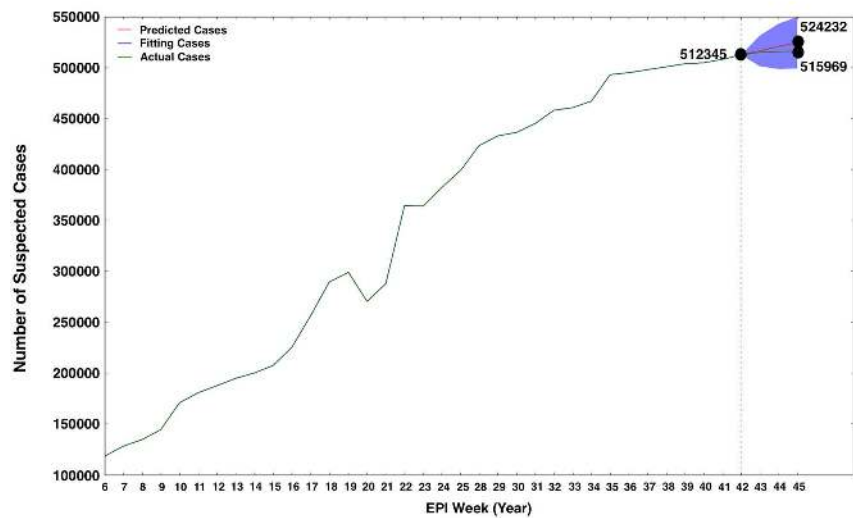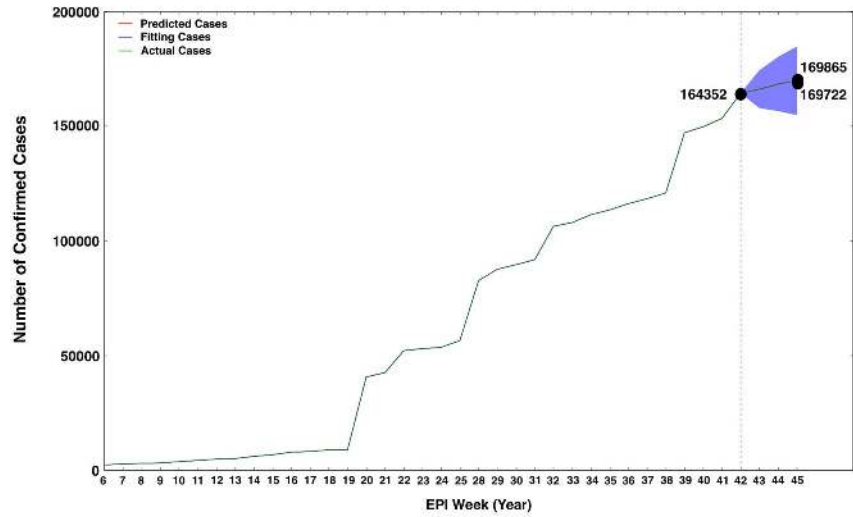
Fig 4. Forecasts of the number of ZIKV confirmed cases (A), suspected cases (B) and total cases (C) in worldwide between 27 October and 9 November 2016 (Yearly EPI Week 43 and 45) by the advanced ARIMA (0, 1, 3) model, which was improved by aggregating historical logs with the most-current 3 weeks' data of Zika-related Google Trends as a estimating predictor to estimate ZVD cases. The red, blue and green solid lines represent the predicted, training and actual number of cases. The blue region represents the 95% confidence interval predicted by the ARIMA (0, 1, 3) model.

doi:10.1371/journal.pone.0165085.g004

online search information produces monitoring data much faster than traditional systems [24, 25]. We first performed correlation analyses to investigate the temporal correlations between data on ZIKV-related GTs and reported cases of ZVD. The result showed that the Zika-related GTs had a strong correlation with confirmed, suspected and total cases of ZIKV. Based on the correlation data, the advanced model ARIMA (0, 1, 3) was improved by aggregating historical logs and estimated data of online search queries associated with Zika as a predictor to forecast ZVD cases. And we found that the complicated ARIMA model showed a good performance for the prediction with the lower value of AIC than the simple linear predictor (Fig 3). The results also indicated that the predicted data by ARIMA model are quite similar to the actual data during ZIKV epidemic early November 2016 (Fig 4). The results in this study showed that the novel surveillance tool of GTs can also provide dynamic timely information to public health agencies and provide near real-time indicators of the spread of infectious disease [26]. However, to be effective for monitoring disease activity on local geographic areas, it must be considered within the local context of ZIKV transmissibility.

Supporting Information

S1 Table. Daily Google Trends Volumes from 12 February to 9 November, 2016 (Yearly EPI Week 6 to 45). The search region is set to 'Worldwide' and the key wrod is the 'Zika'.
(XLSX)

S2 Table. Number of reported confirmed cases, reported suspected cases, reported total cases and cumulative Google Trends volumes about ZIKV in the worldwide during the ZVD epidemic of 12 February to 9 November 2016 (Yearly EPI Week 6 to 45; 39 weeks).
(XLSX)

S3 Table. Comparison of tested models.
(XLSX)

S4 Table. Predictions of the numbers of confirmed, suspected and total cases from 27 October to 9 November 2016 (Yearly EPI Week 6 to 45) by the advanced ARIMA (0, 1, 3) model, which was improved by aggregating historical logs with data of Zika-related Google Trends as an estimating predictor to forecast ZVD cases.
(XLSX)

Acknowledgments

Author Contributions

Conceptualization: YT DHB YJ DF YGT.

Data curation: DHB GGX BHL YH XPA YT.

Formal analysis: DHB GGX BHL XPA YT.

Funding acquisition: YT DF.

Investigation: YT DHB YJ DF YGT.

Methodology: DHB GGX BHL YH YT.

Project administration: YT DHB YJ DF YGT.

Resources: DHB GGX BHL YGT YT.

Software: DHB GGX BHL YH YT.

Supervision: YT DF YGT.

Validation: DHB GGX BHL YGT YT.

Visualization: YT DHB.

Writing – original draft: YT DHB.

Writing – review & editing: YT DHB.

References

1. Chouin-Carneiro T., Bega-Rua A., Bazeille M., Yebakima A., Girod R., Goindin D., et al. Differential Susceptibilities of Aedes aegypti and Aedes albopictus from the Americas to Zika Virus. PLoS Negl. Trop. Dis. 2016 10, e0004543. doi: 10.1371/journal.pntd.0004543 PMID: 26938868

2. Hayes EB. Zika virus outside Africa. Emerg. Infect. Dis. 2009 15, 1347–1350. doi: 10.3201/eid1509.090442 PMID: 19788800

3. Grard G., Caron M., Mombo I.M., Nkoghe D., Ondo S.M., Jiolle D., et al. Zika virus in Gabon (Central Africa)–2007: a new threat from Aedes albopictus? PLoS Negl. Trop. Dis. 2014 8, e2681. doi: 10.1371/journal.pntd.0002681 PMID: 24516683

4. Duffy M.R., Chen T.H., Hancock W.T., Powers A.M., Kool J.L., Lanciotti R.S., et al. Zika virus outbreak on Yap Island, federated states of Micronesia. N. Engl. J. Med. 2009 360, 2536–2543. doi: 10.1056/NEJMoa0805715 PMID: 19516034

5. Mlakar J., Korva M., Tul N., Popovic M., Poljsak-Prijatelj M., Mraz J., et al. Zika virus associated with microcephaly. N. Engl. J. Med. 2009 374, 951–958.

6. Johansson M.A., Mier-y-Teran-Romero L., Reefhuis J., Gilboa S.M., Hills S.L. Zika and the Risk of Microcephaly. N. Engl. J. Med. 2016 375, 1–4. doi: 10.1056/NEJMp1605367 PMID: 27222919

7. Rodrigues L.C. Microcephaly and Zika virus infection. The Lancet. 2016 387, 2070–2072.

8. Cauchemez S., Besnard M., Bompard P., Dub T., Guillemette-Artur P., Eyrolle-Guignot D., et al. Association between Zika virus and microcephaly in French Polynesia 2013–15: a retrospective study. The Lancet. 2016 387, 2125–2132.

9. Tizzoni M., Bajardi P., Poletto C., Ramasco J.J., Balcan D., Goncalves B., et al. Real-time numerical forecast of global epidemic spreading: case study of 2009 A/H1N1pdm. BMC Medicine. 2012 10, 165. doi: 10.1186/1741-7015-10-165 PMID: 23237460

10. Ginsberg J., Mohebbi M.H., Patel R.S., Brammer L., Smolinski M.S., Brilliant L. Detecting influenza epidemics using search engine query data. Nature. 2009 457, 1012–4. doi: 10.1038/nature07634 PMID: 19020500

11. Gluskin R.T., Johansson M.A., Santillana M., Brownstein J.S. Evaluation of Internet-based dengue query data: Google Dengue Trends. PLoS Negl. Trop. Dis. 2014 8, e2713. doi: 10.1371/journal.pntd.0002713 PMID: 24587465

12. Yang S., Santillana M. & Kou S.C. Accurate estimation of influenza epidemics using Google search data via ARGO. Proc. Natl. Acad. Sci. USA 2015 112, 14473–8. doi: 10.1073/pnas.1515373112 PMID: 26553980

13. Majumder M.S., Santillana M., Mekaru S.R., McGinnis D.P., Khan K., Brownstein J.S. Utilizing Nontraditional Data Sources for Near Real-Time Estimation of Transmission Dynamics During the 2015–2016 Colombian Zika Virus Disease Outbreak. JMIR Public Health Surveill 2016 2, e30.

14. Bakker K.M., Martinez-Bakker M.E., Helm B., Stevenson T. et al. Digital epidemiology reveals global childhood disease seasonality and the effects of immunization. Proc. Natl. Acad. Sci. USA 2016 113, 6689–94. doi: 10.1073/pnas.1523941113 PMID: 27247405

15. Soebiyanto R.P., Adimi F. & Kiang R.K. Modeling and predicting seasonal influenza transmission in warm regions using climatological parameters. PLoS One 2015 5: e9450.

16. Dugas A.F., Jalalpour M., Gel Y., Levin S., Torcaso F., Igusa T., et al. Influenza forecasting with Google Flu Trends. PLoS One. 2015 8, e56176.

17. Southwell B.G., Dolina S., Jimenez-Magdaleno K., Squiers L.B., Kelly B.J. Zika Virus-Related News Coverage and Online Behavior, United States, Guatemala, and Brazil. Emerg. Infect. Dis. 2016 22, 1320–1. doi: 10.3201/eid2207.160415 PMID: 27100826

18. Stock J.H. & Watson M. Introduction to econometrics 3rd ed. Harlow, UK: Pearson Education 2011.

19. Akhtar S. & Rozi S. An autoregressive integrated moving average model for short-term prediction of hepatitis C virus seropositivity among male volunteer blood donors in Karachi, Pakistan. World J Gastroenterol 2009 15, 1607–12. doi: 10.3748/wjg.15.1607 PMID: 19340903

20. Faria N.R., Azevedo R.S.S., Kraemer M.U.G., Souza R., Cunha M.S., Hill S.C., et al. Zika virus in the Americas: Early epidemiological and genetic findings. Science 2016 352, 345–349. doi: 10.1126/science.aaf5036 PMID: 27013429

21. Vogel G. Evidence grows for Zika virus as pregnancy danger. Science 2016 351, 1123–4. doi: 10.1126/science.351.6278.1123 PMID: 26965596

22. Malone R.W., Homan J., Callahan M.V., Glasspool-Malone J., Damodaran L., Schneider A.D.B., et al. Zika Virus: Medical Countermeasure Development Challenges. PLoS Negl. Trop. Dis. 2016 10, e0004530. doi: 10.1371/journal.pntd.0004530 PMID: 26934531

23. Lucey D.R. & Gostin L.O. The Emerging Zika Pandemic: Enhancing Preparedness. JAMA. 2016 315, 865–6. doi: 10.1001/jama.2016.0904 PMID: 26818622

24. Check Hayden E. Spectre of Ebola haunts Zika response. Nature 2016 531, 19. doi: 10.1038/531019a PMID: 26935676

25. Kelly B., Squiers L., Bann C., Stine A., Hansen H., Lynch M. Perceptions and plans for prevention of Ebola: results from a national survey. BMC Public Health 2015 15,1136. doi: 10.1186/s12889-015-2441-7 PMID: 26572610

26. Towers S., Afzal S., Bernal G., Bliss N., Brown S., Espinoza B., et al. Mass media and the contagion of fear: the case of Ebola in America. PLoS One. 2015 10, e0129179. doi: 10.1371/journal.pone.0129179 PMID: 26067433