Environment International 117 (2018) 284–291

Contents lists available at ScienceDirect

Environment International

journal homepage: www.elsevier.com/locate/envint

Using Google Trends and ambient temperature to predict seasonal influenza outbreaks

Yuzhou Zhang[a], Hilary Bambrick[a], Kerrie Mengersen[b], Shilu Tong[a,c,d], Wenbiao Hu[a,*]

[a] School of Public Health and Social Work, Institute of Health and Biomedical Innovation, Queensland University of Technology, Brisbane, Queensland, Australia
[b] Science and Engineering Faculty, Mathematical and Statistical Science, Queensland University of Technology, Brisbane, Queensland, Australia
[c] School of Public Health and Institute of Environment and Human Health, Anhui Medical University, Hefei, Anhui, China
[d] Shanghai Children's Medical Centre, Shanghai Jiao-Tong University, Shanghai, China

ARTICLE INFO

Keywords:
Early warning
Prediction
Search terms
Seasonal influenza
Temperature

ABSTRACT

Background: The discovery of the dynamics of seasonal and non-seasonal influenza outbreaks remains a great challenge. Previous internet-based surveillance studies built purely on internet or climate data do have potential error.

Methods: We collected influenza notifications, temperature and Google Trends (GT) data between January 1st, 2011 and December 31st, 2016. We performed time-series cross correlation analysis and temporal risk analysis to discover the characteristics of influenza epidemics in the period. Then, the seasonal autoregressive integrated moving average (SARIMA) model and regression tree model were developed to track influenza epidemics using GT and climate data.

Results: Influenza infection was significantly corrected with GT at lag of 1–7 weeks in Brisbane and Gold Coast, and temperature at lag of 1–10 weeks for the two study settings. SARIMA models with GT and temperature data had better predictive performance. We identified autoregression (AR) for influenza was the most important determinant for influenza occurrence in both Brisbane and Gold Coast.

Conclusions: Our results suggested internet search metrics in conjunction with temperature can be used to predict influenza outbreaks, which can be considered as a pre-requisite for constructing early warning systems using search and temperature data.

Handling Editor: Olga-Ioanna Kalantzi

1. Introduction

The discovery of the dynamics of seasonal and non-seasonal influenza outbreaks remains a great challenge (Lipsitch et al., 2011). While vaccination is effective in preventing infection, seasonal influenza remains epidemics and results in an estimated three to five million cases of severe illness and about 250,000 to 500,000 deaths each year worldwide (World Health Organization, n.d.).

There is a delay of up to 2 weeks between the onset of disease and when notification data is compiled into traditional surveillance reports (Chan et al., 2010). This lag in reporting limits the ability of such conventional surveillance systems to provide timely epidemiologic intelligence and delays the response of health officers to possible outbreaks (Project, 2011).

In order to prepare for the next severe influenza epidemics and provide a timely, effective response, researchers have proposed several new approaches to achieve near real-time detection and even prediction of emerging and spread of influenza outbreaks (Simonsen et al., 2016). Over the past decade the increasing number of internet users around world has provided new sources of data potentially valuable for identifying influenza outbreaks (Kang et al., 2013; Cho et al., 2013; Shin et al., 2016; Seo et al., 2014; Polgreen et al., 2008).

We recognised that previous models built purely on internet-based or climate factors do have potential error (Lazer et al., 2014; Urashima et al., 2003; Pollett et al., 2016). Previous studies reported that media bias can adversely impact internet-based surveillance systems (Althouse et al., 2011). For instance, Google Flu Trends (GFT) predicted more than double the peak of influenza-like illness (ILI) cases than the Centers for Disease Control and Prevention (CDC) in 2013 (Lazer et al., 2014). A major reason for the overestimation may be the widespread media coverage of the severe flu season, which may result in many searches by people who were not ill (Butler, 2013). A previous study

* Corresponding author.
E-mail addresses: yuzhou.zhang@hdr.qut.edu.au (Y. Zhang), h.bambrick@qut.edu.au (H. Bambrick), k.mengersen@qut.edu.au (K. Mengersen), s.tong@qut.edu.au, s.tong@ahmu.edu.cn, tongshilu@scmc.com.cn (S. Tong), w2.hu@qut.edu.au (W. Hu).

https://doi.org/10.1016/j.envint.2018.05.016
Received 12 January 2018; Received in revised form 4 April 2018; Accepted 7 May 2018
Available online 16 May 2018

used Autoregression model with Google search data to capture changes in people's online search behaviour over time. The findings suggested that this approach could reduce the predictive errors (Yang et al., 2015). This study aims to assess whether the development of an empirical time series model combining internet-based influenza search metrics and temperature can predict influenza outbreaks and reduce the potential errors introduced from factors such as fear based searching.

2. Methods

2.1. Data collection

Weekly influenza notifications in Brisbane and the Gold Coast (Queensland Hospital and Health Services areas) during the period from January 1st, 2011 to December 31st, 2016 were retrieved from Queensland Health Influenza Surveillance Annual Reporting. The reports provide a profile of influenza from a number of laboratory confirmed notifications.

A climate dataset of two study settings were obtained from Australian Climate Data Online System. The daily maximum temperature (°C) data of Brisbane and the Gold Coast for the study period were collect from Brisbane Basic Climatological Station and Gold Coast Seaway Basic Climatological Station respectively.

The search term “influenza” was chosen for analysis in the study. For the purpose of the study, a search query is a complete, exact sequence of terms issued by internet users (Ginsberg et al., 2009); we did not combine several search terms, although we hope to explore these options in future work. A .CSV file for search term during the study period was downloaded from Google Trends (GT) website to collect weekly influenza Internet search trend data. As GT cannot provide search metrics data at city level in Australia, the search query data at Queensland state level was collected in this study. We hypothesized GT data for Queensland can represent that for Brisbane and the Gold Coast since the total population of these two cities account for nearly 67% of Queensland population (Australian Government, n.d.). Additionally, the two cities have more access to the internet (Brisbane: 82.4%, Gold Coast: 80.1%) comparing with other Queensland regions (Queensland Government, n.d.-a).

2.2. Time-series cross correlation analysis

To assess the correlations between influenza notifications with GT and climate variables, time-series cross correlation between weekly influenza occurrence, GT and mean maximum temperature was carried out in the study. Because the variables are strongly associated with each other with different time lags, only those with maximal correlation coefficient were performed to construct the models in the study (Sang et al., 2015).

2.3. Temporal risk analysis

The seasonal parameters of interest were used to estimate influenza outbreak timing and duration (Bloom-Feshbach et al., 2013). With the current knowledge of influenza epidemic, it is still hard to identify whether an outbreak appears suddenly for a certain period of time (Wen et al., 2006). We discover influenza outbreaks between May and October which is the influenza season in Queensland (Queensland Government, n.d.-b). The first week of an influenza outbreak in Brisbane was defined in this paper when case numbers continued to increase for 6 weeks within a calendar year (Neuzil et al., 2000). However, it was not a reliable definition of outbreak for the Gold Coast as influenza activity was lower in the Gold Coast than in Brisbane. If the influenza activity was very low during a season, it was difficult to identify the peak week and no peak was selected (Paget et al., 2007). Thus, the definitions of the first outbreak week in the Gold Coast was when the influenza notification exceeded 1% of mean annual influenza cases number (Bloom-Feshbach et al., 2013). The increasing duration of an epidemic was defined as the number of weeks between first and peaking week. Thus, the increasing duration index of an outbreak can be described as the proportion of the increasing duration in a calendar year. This index (α) is defined as:

$$\alpha = IW/TW$$

where IW is the total number of increasing weeks of an outbreak during each calendar year and TW is the total number of weeks in a calendar year (52 weeks).

Increasing intensity refers to the likely increasing magnitude within an outbreak. Incidence rate, as an index to measuring the magnitude of new cases occurring during a specified period, it cannot reflect the spread speed during the period. Increasing intensity index can assess the severity of an epidemic by focusing on successive weeks when cases have occurred (Wen et al., 2006). This index (β) is formulated as:

$$\beta = (y - b)/x$$

where y is the observed influenza notifications; b is the base level of the formula, which is defined as the starting value of the series data and x is the number of weeks for increasing duration. These parameters' values are based on the linear regression equation of the notifications in the increasing duration of an outbreak. The index evaluates the spread speed of an outbreak by focusing on successive weeks when cases increase rapidly. The β value will become bigger if most cases are temporally concentrated throughout the outbreak. A low value of β describes an outbreak that is more temporally dispersed.

To discover whether GT is a valuable data source to detect the likely rising magnitude within an outbreak, this index was also performed for GT data. We used the first week for GT as performed in influenza analysis, but used GT's own peaking week in the analysis. Thus, the increasing duration for GT is defined as the total number of weeks between the first week of influenza outbreak and GT peaking week.

2.4. Seasonal autoregressive integrated moving average (SARIMA) model with GT and temperature

As influenza has a strong seasonal characteristics in time series (Dushoff et al., 2004), SARIMA models were developed to control the effects of seasonality in the forecast of influenza epidemics. We used influenza notification as the dependent variable, GT and mean maximum temperature as the independent variable. GT and mean maximum temperature with maximal cross correlation coefficient were performed to construct the models of Brisbane and the Gold Coast. Generally, there are three significant components of a SARIMA model, including autoregressive (AR), differencing and moving average (MA). Three parameters are typically selected when fitting this model: (p, d, q); where p is the order of the AR, d is the order of the differencing, and q is the order of the MA (Box et al., 2015). To test the goodness-of-fit of the model, autocorrelation and partial autocorrelation of residuals were checked. In addition, Bayesian information criterion (BIC), the stationary R square (R^2), the Root Mean Squared Error (RMSE) and the Maximum Absolute Percent Error (MAPE) were also used to examine the goodness-of-fit of the model. We used the same data file in the temporal risk analysis to construct and validate SARIMA models. The data file was divided into two data sets: we aimed to construct the SARIMA models using the least amount of weekly data (Brisbane: week 19–30 (2011), week 18–29 (2012–2016); Gold Coast: week 19–31 (2011), week 18–30 (2012–2016)); and the rest data was used as a test data set to validate the model. This method could assist us to predict more weeks' notifications during an outbreak period using limited weekly data. Moreover, a comparison of performance of SARIMA models that either included or excluded GT and temperature data was undertaken. A SARIMA model can be considered as a good model if it has a large R^2 value and a small BIC value. The better model was select as the predictive model.

We reported 3 metrics to evaluate the predictive performance of SARIMA model: Pearson correlation, RMSE and MAPE. The definitions of all evaluation metrics were given in Supplementary Information file.

2.5. Regression tree analysis

Regression tree models were performed in the study to segment the weekly GT and temperature data into subsets that were most likely to be associated with weekly influenza outbreaks. GT and mean maximum temperature with maximal cross correlation coefficient were performed to construct the models. We also used influenza data with maximal autoregression (AR) coefficient in model construction, as AR is a significant component in influenza forecast model (Dugas et al., 2013). Cross-validation was used to select the tree size using estimated prediction errors. The best model is defined as having the smallest tree size and an estimated error rate within one standard error of the minimum (Breiman et al., 1984).

All data analyses were performed through using SPSS Statistics software, version 23 (SPSS Inc.; Chicago, IL, USA) and R version 3.4.0. Statistical significance was set at $P < 0.05$. All data was checked for completeness and accuracy before analysis.

3. Results

3.1. Descriptive analysis

During the study period, the mean weekly influenza cases were 183.0 (range: 0 to 1894) in Brisbane and 35.2 (range: 0 to 454) at the Gold Coast. Both the highest weekly counts of influenza cases of Brisbane and the Gold Coast over the study period were observed in week 34, 2015 (Fig. 1). Moreover, both the highest mean influenza occurrence of Brisbane and the Gold Coast during the study period was observed in week 34 with the value of 816.5 and 146.5 respectively (Fig. S1).

Table 1
The proportion of main circulating influenza type, subtypes and vaccine composition of Queensland from 2011 to 2016. (Proportion = The annual notifications of a certain type or subtype/The annual total notifications; N.A.: Unavailable).

Year	Type	Proportion	Subtype	Proportion	Vaccine composition
2011	Influenza A	80.1%	H1N1	29.4%	Influenza A: H1N1 and H3N2; Influenza B
2012	Influenza A	74.6%	H3N2	11.5%	Same as 2011
2013	Influenza A	65.2%	H1N1	12.8%	Same as 2011
2014	Influenza A	86.4%	H1N1	12.0%	Same as 2011
2015	Influenza B	66.5%	N.A.	N.A.	Same as 2011
2016	Influenza A	92.1%	H3N2	9.5%	Same as 2011

The weekly trends of influenza occurrence, GT metrics and mean maximum temperature were shown in Fig. 1. There were similar trends between the weekly counts of influenza cases and the weekly GT for influenza in the cities during the study period. However, there were opposite tendencies between the weekly influenza notifications and the weekly mean maximum temperature over the study period. The peaks of influenza notifications occurred during the temperature troughs. Moreover, there were different main influenza types and subtypes circulated each year with similar vaccine composition at the study period (Table 1) (Queensland Government, n.d.-b).

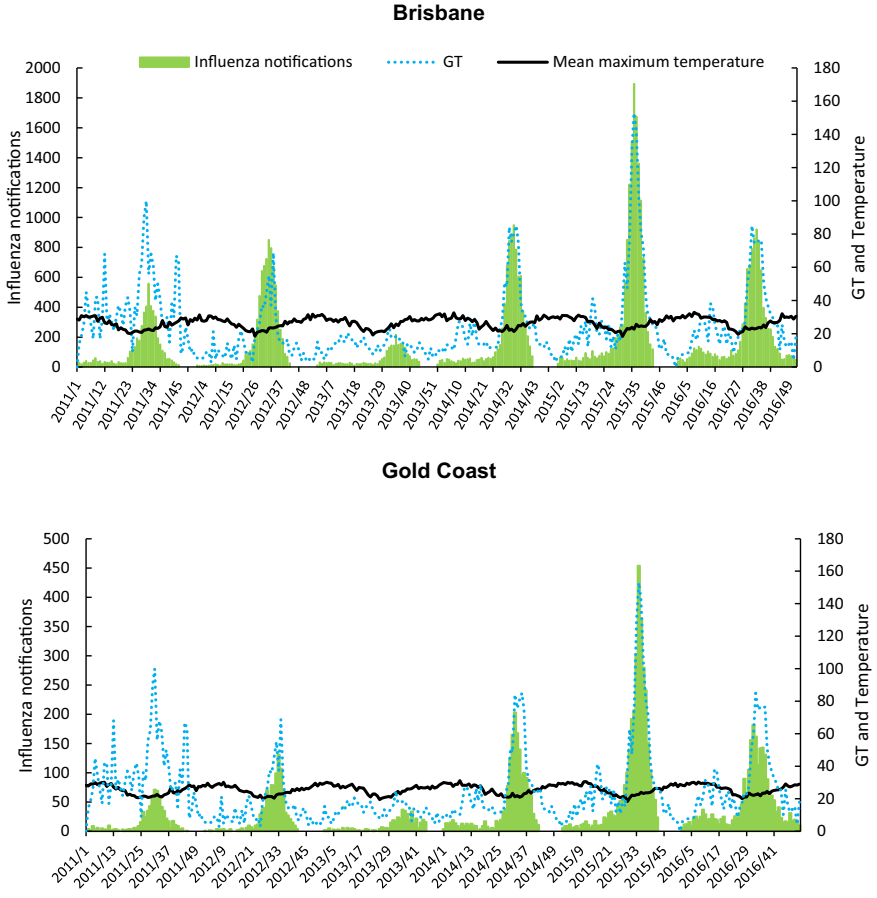

Fig. 1. Weekly distribution of influenza cases, GT data and mean maximum temperature in Brisbane and Gold Coast, 2011–2016.

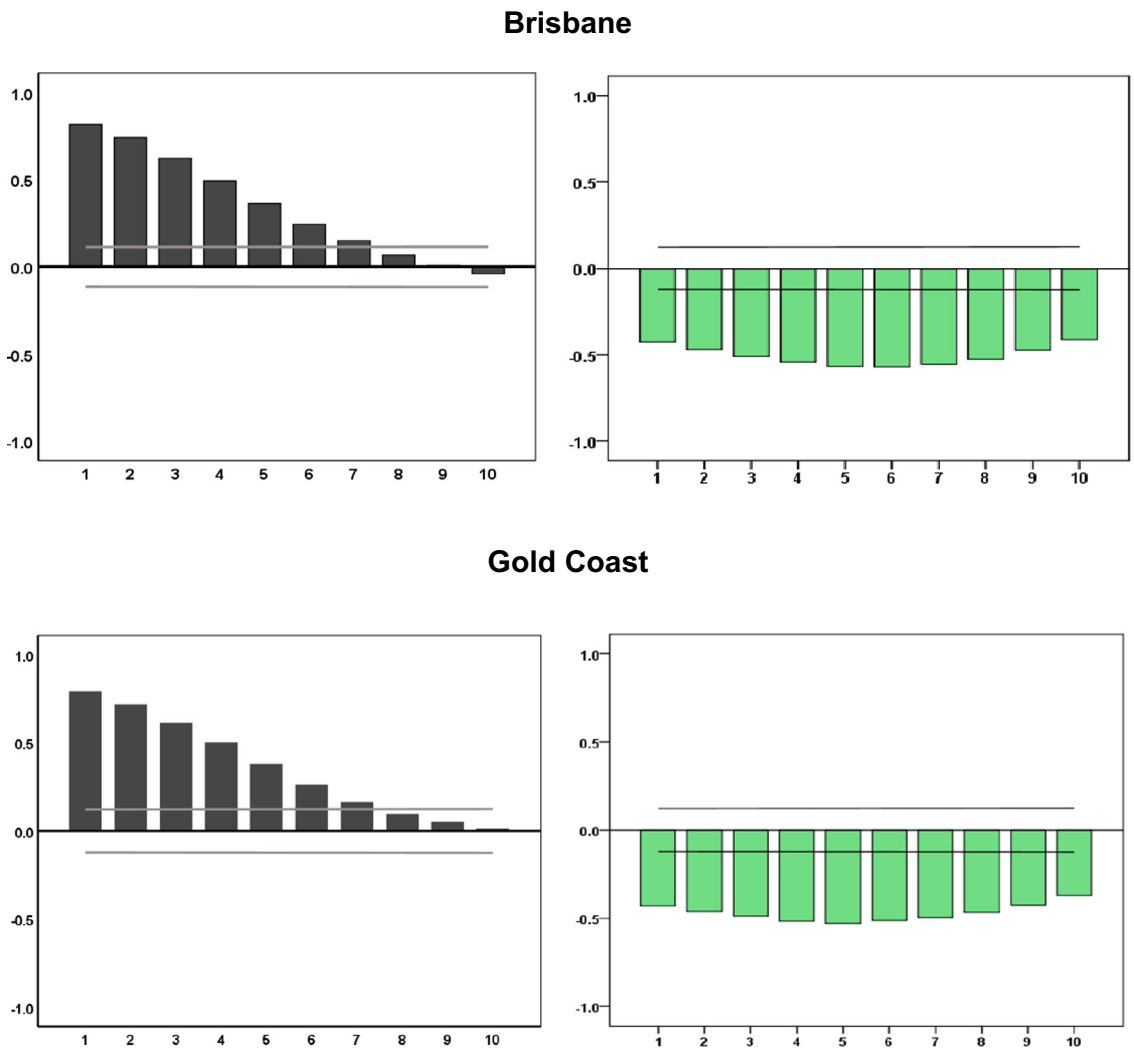

Fig. 2. Cross-correlation between influenza notifications with GT search query data and temperature data. Yellow bars indicate the value of GT data from 2011 to 2016. The values of mean maximum temperature are indicated by green bars. Confidence intervals (95%) are indicated by the dashed lines (X axis: lag value, Y axis: CCF value). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Table 2
Statistics of temporal risk indices of Brisbane and Gold Coast from 2011 to 2016.

		Peaking influenza weekly notifications	IW (weeks)	Increasing duration index (α) for influenza	Increasing intensity index (β) for influenza	Peaking GT weekly metrics	IW (weeks)	Increasing duration index (α) for GT	Increasing intensity index (β) for GT
Brisbane	2011	557	12	0.23	43.64	100	12	0.23	6.50
	2012	850	15	0.29	62.48	69	17	0.33	3.07
	2013	216	9	0.17	22.35	25	6	0.12	2.28
	2014	949	11	0.21	103.67	85	11	0.23	8.13
	2015	1894	12	0.23	155.03	153	12	0.23	11.18
	2016	921	11	0.21	97.61	85	9	0.17	8.53
Gold Coast	2011	71	8	0.15	8.21	100	12	0.23	11.63
	2012	131	11	0.21	10.95	69	17	0.33	5.22
	2013	37	10	0.19	3.12	25	6	0.12	2.14
	2014	203	7	0.13	32.04	85	11	0.23	9.47
	2015	454	10	0.19	39.90	153	12	0.23	14.86
	2016	182	7	0.13	24.57	85	9	0.17	9.40

3.2. Time-series cross correlation analysis

Time-series cross correlation analysis demonstrated that weekly influenza occurrence to be positively correlated with weekly GT metrics for search terms “influenza” at the time lags of 1–7 weeks in Brisbane and the Gold Coast (Fig. 2) (Table S1). Additionally, the correlations between weekly influenza notifications and weekly mean maximum temperature were negative at the lags of 1–10 weeks in the two study settings. The number of searches for influenza term in Brisbane and the Gold Coast had the correlations of 0.82 and 0.80 ($P < 0.05$) respectively with the number of influenza notifications in the following week, and that searches further apart in time had lower correlations. The

Table 3
Spearman's Correlation of annual temporal risk indices of influenza outbreaks in Brisbane and Gold Coast, 2011–2016.

		Increasing duration index (α)	Increasing intensity index (β) for influenza	Increasing intensity index (β) for GT
Brisbane	Peaking influenza weekly notifications	0.24	1.00**	0.89*
	Increasing duration index (α)		0.24	0.15
	Increasing intensity index (β) for influenza			0.89*
Gold Coast	Peaking influenza weekly notifications	−0.27	1.00**	0.66
	Increasing duration index (α)		−0.27	−0.27
	Increasing intensity index (β) for influenza			0.66

** Correlation is significant at the 0.01 level (2-tailed).
* Correlation is significant at the 0.05 level (2-tailed).

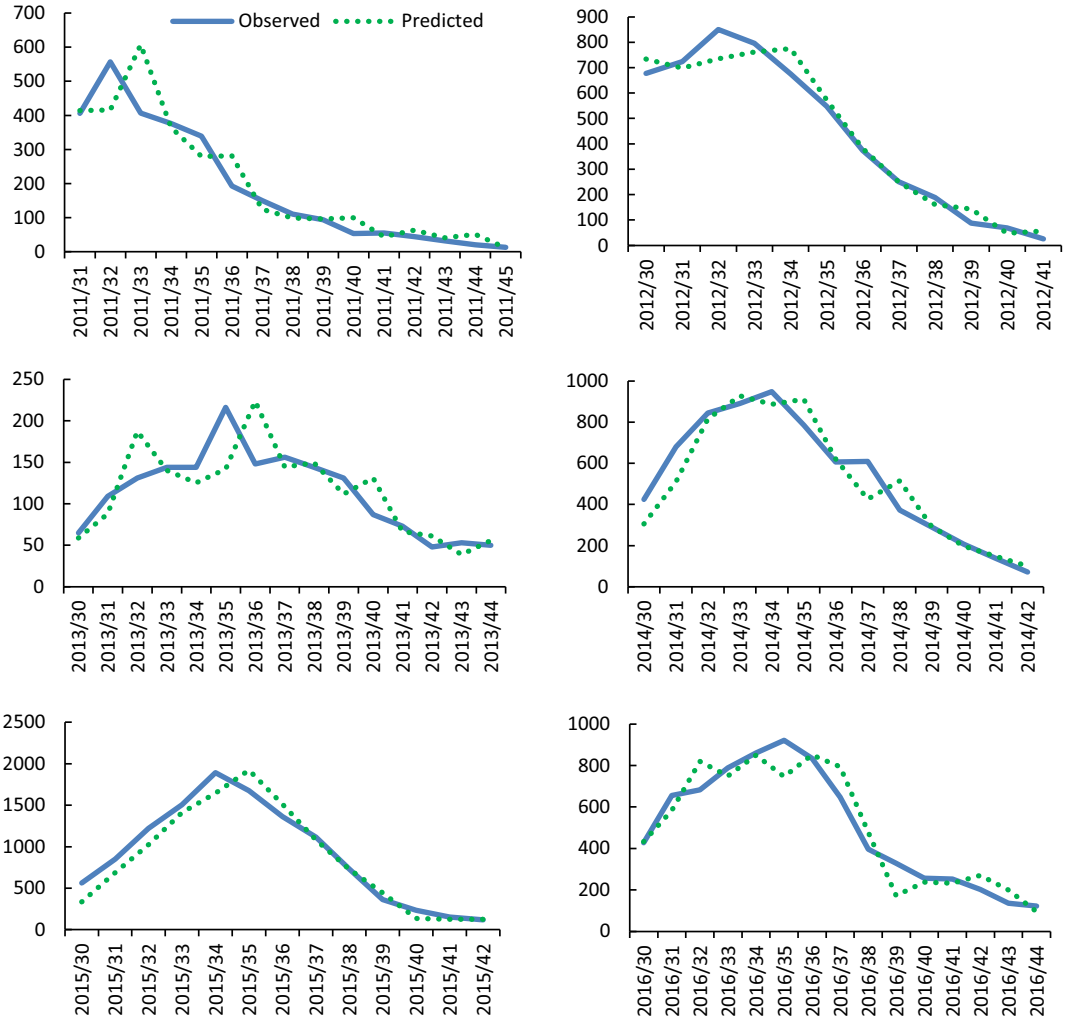

Fig. 3. Observed and predicted values during 2011–2016 outbreaks based on the SARIMA model in Brisbane (X axis: date (week), Y axis: influenza notifications).

strongest correlation between influenza infections and mean maximum temperature in Brisbane was found at lag of 6 weeks, and at the Gold Coast it was at a lag of 5 weeks. The variables with time lags that had the maximal correlation coefficient were performed to construct the models.

3.3. Temporal risk of influenza outbreaks analysis

2012 experienced the highest Increasing duration index (α) in Brisbane and the Gold Coast with the value of 0.29 and 0.21 respectively (Table 2) (Fig. S2). The largest Increasing intensity index (β) for influenza and GT were observed in 2015 in the study settings. The index for influenza and GT are 155.03 and 11.18 in Brisbane, and the values of the Gold Coast are 39.90 and 14.86 separately.

We then calculated the correlation among these three indices as shown in Table 3. It could be found that both the value of Increasing intensity index (β) for influenza and GT is highly correlated with peaking weekly influenza notifications. The scatterplot matrix with the regression lines in Fig. S3 displays the relationships between the peaking weekly influenza notifications and Increasing intensity index (β) for influenza and GT.

3.4. SARIMA model

GT and mean maximum temperature with maximal cross correlation coefficient were used to construct the models of Brisbane and the Gold Coast. We used GT at 1-week lag and mean maximum temperature at 6-week lag to construct the models of Brisbane with maximal

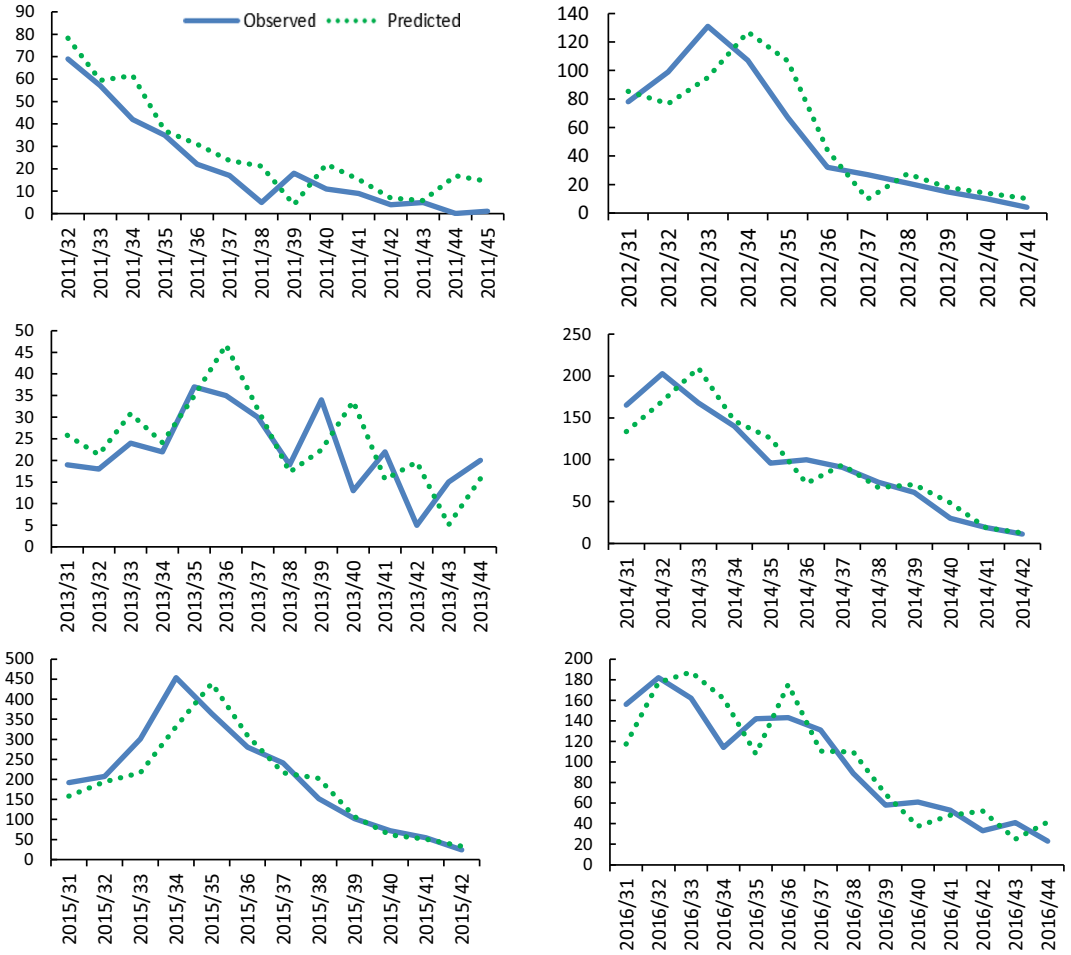

Fig. 4. Observed and predicted values during 2011–2016 outbreaks based on the SARIMA model in Gold Coast (X axis: date (week), Y axis: influenza notifications). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

correlation coefficient of 0.81 and 0.57 respectively. GT at 1-week lag and mean maximum temperature at 5-week lag were used in the model of Gold Coast with maximal correlation coefficient of 0.80 and 0.53 separately. Influenza surveillance data from week 19 to 30, 2011 and week 18 to 29, 2012 to 2016 was used to construct a SARIMA model of Brisbane; and that from week 19 to 31, 2011 and week 18 to 30, 2012 to 2016 of the Gold Coast was used as the dataset of model training. The SARIMA model (3,0,2) (1,0,0) and (3,0,1) (1,0,0) with GT metrics and mean maximum temperature had the best fit to the data in Brisbane and the Gold Coast separately. The analysis of goodness-of-fit revealed that the SARIMA models fitted the data well, as the ACF and PACF fluctuate randomly near zero in the residuals of this model (Fig. S4). The model with GT and temperature had better performance with the larger R^2 and smaller BIC, RMSE and MAPE values, compared to the variables excluded models (Table S2). Thus, the models that incorporated GT and temperature data were selected as the predictive models for validation. Results of SARIMA models are presented in Table S3. Models were used to predict the weekly influenza notifications for the rest outbreak weeks, and were validated by Queensland Health annual influenza surveillance reports (Fig. S5) (Fig. 3) (Fig. 4).

The evaluation of predictive performance of the SARIMA models was presented in Table S4. This table indicated that the predictive capacity of the models is performed well with high overall Pearson correlations (Brisbane: 0.97, Gold Coast: 0.94, $P < 0.01$). The model is also robust as indicated by the size of the overall MAPE (Brisbane: 0.16, Gold Coast: 0.29), which measures that the model is low off-target with respect to the observed influenza notifications.

3.5. Regression tree analysis

We used the same lagged GT and temperature data in SARIMA model to construct regression tree models. AR for influenza at 1-week lag was also included in the model construction. Fig. 5 demonstrated that the 1-week lagged AR for influenza was the first classifying factor in the models of Brisbane and Gold Coast. Thus, the 1-week lagged AR was the most important factor influencing the variation in the weekly influenza cases number of the two study settings. However, the lagged GT played different roles in models of the two cities. The lagged GT is the third classifying factor in the model of Gold Coast, and the fourth one in the model of Brisbane. The mean influenza notifications increased by over 5.3-fold (1510/284) in the model of Brisbane when AR for influenza at 1-week lag was ≥1168. Moreover, when AR for influenza at 1-week lag was ≥187, the mean influenza notifications increased by over 5-fold (271/54) in the Gold Coast.

4. Discussion

Descriptive analysis showed similar trends between the weekly counts of influenza cases and the weekly GT for influenza in both Brisbane and the Gold Coast during the study period. Additionally, our results have clearly shown a significant positive relationship between the occurrence of influenza outbreaks and GT data. Although several previous work indicated that many cases of ILI were not caused by influenza and other respiratory viruses infections may also trigger an internet search (Thursky et al., 2003; Stiver, 2003), the findings of this study support the hypothesis that internet search metrics have utility in

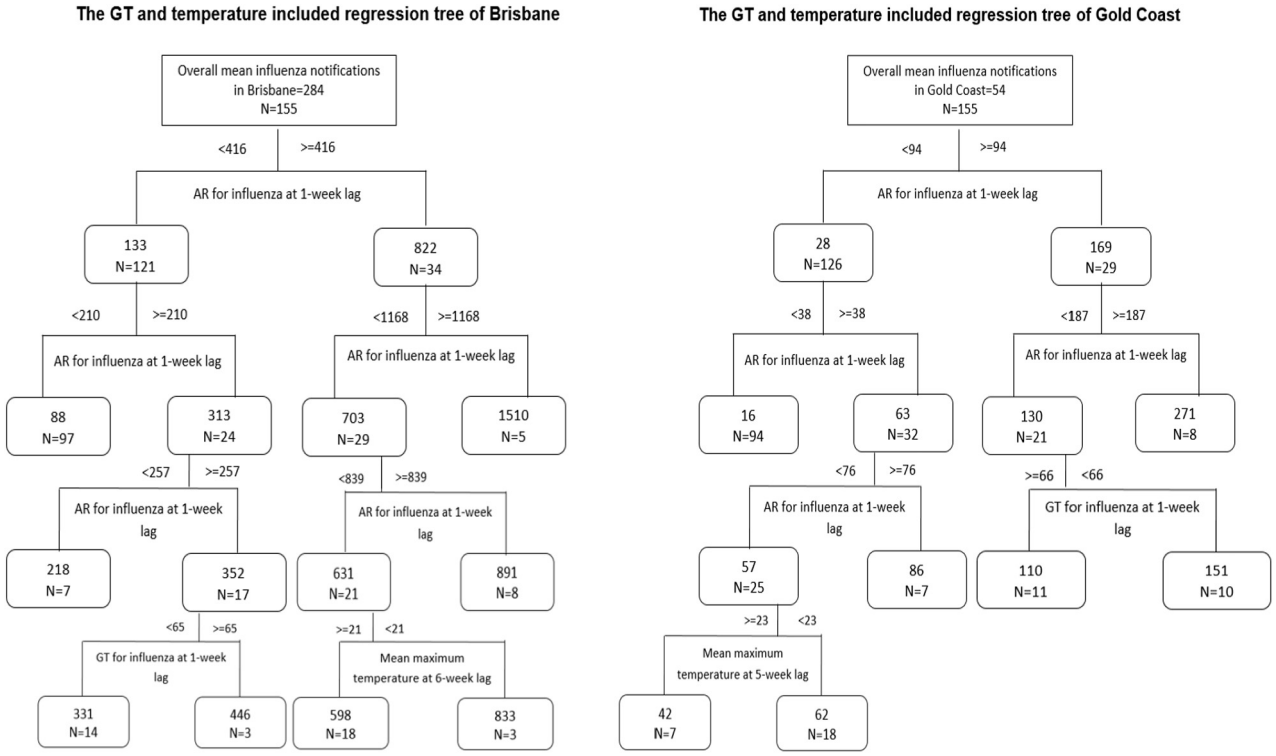

Fig. 5. The regression tree modeling the hierarchical relationship between weekly influenza notifications with GT and temperature data in Brisbane and the Gold Coast between 2011 and 2016. (The regression trees showed the threshold values, mean weekly influenza notifications; N is the total week count of occurrence of influenza outbreaks; AR: autoregression.)

monitoring and even predicting influenza outbreaks. However, a significant negative correlation between influenza notifications and mean maximum temperature was observed in the study. The results are in agreement with previous studies, which reported that lower temperature could contribute to an increasing activity of influenza (Jaakkola et al., 2014; Shoji et al., 2011).

Results of cross correlation indicated that predicting of influenza events may be possible using internet search metrics and temperature data. Significant positive correlations were exhibited at 1–7 lag weeks for GT data of Brisbane and Gold Coast. This result can be considered as a pre-requisite for constructing early warning systems for influenza using GT, as this finding indicates how much faster internet metrics may collect the data. The signalling of variation in GT metrics can provide adequate time to government and health authorities to implement influenza preventive measures.

It is very important to perform the increasing duration index (α) for public health officers and environment agencies as it reflects the effectiveness of the prevention or control measures used over the epidemic period (Wen et al., 2006). A larger value of α indicates that the cases are less likely to disappear once they occur and more chance to result in virus mutation (Wen et al., 2006). We hope to explore the reasons for the variations in the index values in future work.

It should be noted that the correlation coefficient between peaking influenza weekly notifications and Increasing intensity index (β) for influenza is 1.00 in the two study settings. It indicated that a bigger size of outbreak of influenza could be found when case numbers increased more rapidly during the outbreak period. Furthermore, a strong positive association between peaking weekly influenza cases number and Increasing intensity index (β) for GT was observed in Brisbane with the value of 0.89. This result demonstrated GT can be seen as a valuable data source to reflect the magnitude of influenza outbreaks of Brisbane as well. A bigger size of outbreak can be found when GT metrics rapidly rise during an influenza epidemic. The indicators will be valuable for public health personnel to identify spatial risk areas where the influenza event with higher transmission intensity and larger outbreak size.

The result of this study suggests the SARIMA models that included internet search metrics and temperature data provided a better fit to influenza monitoring data than the models that excluded either variable, and the results showed good predictive capacity of the models as well. The incorporation of internet search metrics and temperature based models into conventional influenza surveillance systems has the potential to bolster capacity in the monitoring and predicting of influenza outbreaks by reducing the effect of this lag on the system.

The regression tree models identified that the 1-week lagged AR for influenza was a key determinant of occurrence of influenza outbreaks in both Brisbane and the Gold Coast. However, GT played different roles in Brisbane and Gold Coast. The appearances in the two areas may be due to the differences in the social demographic characteristics, the availability, and popularity of the internet, human migration and meteorological factors (Liu et al., 2016). The results demonstrated that the models, in general, built on AR for influenza, GT and temperature data provide the threshold value for the potential risk factors of influenza activity linking with official influenza data reported by health authorities.

There are four limitations in the study. First, we only collected the GT data for search term “influenza” in the study, the accuracy of GT may be improved if more search terms, such as the term “flu” were explored in the future study (Nagel et al., 2013). Second, the GT data for Queensland only being available at state rather than city level may affect the accuracy of this study. Third, the accuracy of GT may be influenced by different levels of access to internet (Milinovich et al., 2014). Fourth, it is acknowledged that different internet-seeking behaviours; and self-reporting and media-driven bias between different sectors of community (Milinovich et al., 2014). Previous study reported that media bias can adversely influence internet-based surveillance systems (Althouse et al., 2011; Zhang et al., 2017).

5. Conclusion

To conclude, our results suggested internet search metrics in conjunction with temperature can be used to predict influenza outbreaks, which can be considered as a pre-requisite for constructing early warning systems for influenza surveillance using internet search metrics and temperature. The findings will facilitate government, health authorities and the public to predict and respond to influenza outbreaks.

Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Conflict of interest

The authors declare no conflict of interest.

Acknowledgments

Y. Z. was supported by the China Scholarship Council Postgraduate Scholarship and the Queensland University of Technology Higher Degree Research Tuition Fee Sponsorship. W. H. was supported by an Australian Research Council Future Fellowship (award number FT140101216). We also thank the Queensland Health and Australian Bureau of Meteorology for providing the data on influenza cases and climate variables, respectively. W. H. designed this study. Y. Z. collected and analysed the data and drafted this manuscript. W. H., H. B., K. M. and S. T. interpreted the results and revised the manuscript.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.envint.2018.05.016.

References

Althouse, B.M., Ng, Y.Y., Cummings, D.A., 2011. Prediction of dengue incidence using search query surveillance. PLoS Negl. Trop. Dis. 5 (8), e1258.
Australian Government, 2017. Regional Population Growth, Australia, 2015–16. Available from: http://www.abs.gov.au/ausstats/abs@.nsf/mf/3218.0.
Bloom-Feshbach, K., Alonso, W.J., Charu, V., Tamerius, J., Simonsen, L., Miller, M.A., et al., 2013. Latitudinal variations in seasonal activity of influenza and respiratory syncytial virus (RSV): a global comparative review. PLoS One 8 (2), e54445.
Box, G.E., Jenkins, G.M., Reinsel, G.C., Ljung, G.M., 2015. Time Series Analysis: Forecasting and Control. John Wiley & Sons.
Breiman, L., Friedman, J.H., Olshen, R.A., Stone, C.J., 1984. Classification and Regression Trees. Wadsworth & Brooks, Monterey, CA.
Butler, D., 2013. When Google got flu wrong. Nature 494 (7436), 155.
Chan, E.H., Brewer, T.F., Madoff, L.C., Pollack, M.P., Sonricker, A.L., Keller, M., et al., 2010. Global capacity for emerging infectious disease detection. Proc. Natl. Acad. Sci. 107 (50), 21701–21706.
Cho, S., Sohn, C.H., Jo, M.W., Shin, S.-Y., Lee, J.H., Ryoo, S.M., et al., 2013. Correlation between national influenza surveillance data and Google Trends in South Korea. PLoS One 8 (12), e81422.
Dugas, A.F., Jalalpour, M., Gel, Y., Levin, S., Torcaso, F., Igusa, T., et al., 2013. Influenza forecasting with Google flu trends. PLoS One 8 (2), e56176.
Dushoff, J., Plotkin, J.B., Levin, S.A., Earn, D.J., 2004. Dynamical resonance can account for seasonality of influenza epidemics. Proc. Natl. Acad. Sci. U. S. A. 101 (48), 16915–16916.
Ginsberg, J., Mohebbi, M.H., Patel, R.S., Brammer, L., Smolinski, M.S., Brilliant, L., 2009. Detecting influenza epidemics using search engine query data. Nature 457 (7232), 1012–1014.
Jaakkola, K., Saukkoriipi, A., Jokelainen, J., Juvonen, R., Kauppila, J., Vainio, O., et al., 2014. Decline in temperature and humidity increases the occurrence of influenza in cold climate. Environ. Health 13 (1), 22.
Kang, M., Zhong, H., He, J., Rutherford, S., Yang, F., 2013. Using google trends for influenza surveillance in South China. PLoS One 8 (1), e55205.
Lazer, D., Kennedy, R., King, G., Vespignani, A., 2014. The parable of Google flu: traps in big data analysis. Science 343 (6176), 1203–1205.
Lipsitch, M., Finelli, L., Heffernan, R.T., Leung, G.M., 2011. Redd, for the H1N1 Surveillance Group SC. Improving the evidence base for decision making during a pandemic: the example of 2009 influenza A/H1N1. Biosecur Bioterror. 9 (2), 89–115.
Liu, K., Wang, T., Yang, Z., Huang, X., Milinovich, G.J., Lu, Y., et al., 2016. Using Baidu search index to predict dengue outbreak in China. Sci. Rep. 6.
Milinovich, G.J., Williams, G.M., Clements, A.C., Hu, W., 2014. Internet-based surveillance systems for monitoring emerging infectious diseases. Lancet Infect. Dis. 14 (2), 160–168.
Nagel, A.C., Tsou, M.-H., Spitzberg, B.H., An, L., Gawron, J.M., Gupta, D.K., et al., 2013. The complex relationship of realspace events and messages in cyberspace: case study of influenza and pertussis using tweets. J. Med. Internet Res. 15 (10).
Neuzil, K.M., Wright, P.F., Mitchel, E.F., Griffin, M.R., 2000. The burden of influenza illness in children with asthma and other chronic medical conditions. J. Pediatr. 137 (6), 856–864.
Paget, J., Marquet, R., Meijer, A., van der Velden, K., 2007. Influenza activity in Europe during eight seasons (1999–2007): an evaluation of the indicators used to measure activity and an assessment of the timing, length and course of peak activity (spread) across Europe. BMC Infect. Dis. 7 (1), 141.
Polgreen, P.M., Chen, Y., Pennock, D.M., Nelson, F.D., Weinstein, R.A., 2008. Using internet searches for influenza surveillance. Clin. Infect. Dis. 47 (11), 1443–1448.
Pollett, S., Boscardin, W.J., Azziz-Baumgartner, E., Tinoco, Y.O., Soto, G., Romero, C., et al., 2016. Evaluating Google Flu Trends in Latin America: important lessons for the next phase of digital disease detection. Clin. Infect. Dis. 64 (1), 34–41.
Project, T.S., 2011. Assessment of syndromic surveillance in Europe. Lancet 378 (9806), 1833–1834.
Queensland Government Internet access in Queensland 2008. Available from: http://www.qgso.qld.gov.au/products/reports/internet-access-qld-c06/internet-access-qld-c06.pdf.
Queensland Government Influenza surveillance reporting 2017. Available from: https://www.health.qld.gov.au/clinical-practice/guidelines-procedures/diseases-infection/surveillance/reports/flu.
Sang, S., Gu, S., Bi, P., Yang, W., Yang, Z., Xu, L., et al., 2015. Predicting unprecedented dengue outbreak using imported cases and climatic factors in Guangzhou, 2014. PLoS Negl. Trop. Dis. 9 (5), e0003808.
Seo, D.-W., Jo, M.-W., Sohn, C.H., Shin, S.-Y., Lee, J., Yu, M., et al., 2014. Cumulative query method for influenza surveillance using search engine data. J. Med. Internet Res. 16 (12), e289.
Shin, S.-Y., Kim, T., Seo, D.-W., Sohn, C.H., Kim, S.-H., Ryoo, S.M., et al., 2016. Correlation between national influenza surveillance data and search queries from mobile devices and desktops in South Korea. PLoS One 11 (7), e0158539.
Shoji, M., Katayama, K., Sano, K., 2011. Absolute humidity as a deterministic factor affecting seasonal influenza epidemics in Japan. Tohoku J. Exp. Med. 224 (4), 251–256.
Simonsen, L., Gog, J.R., Olson, D., Viboud, C., 2016. Infectious disease surveillance in the big data era: towards faster and locally relevant systems. J. Infect. Dis. 214 (suppl_4) (S380–S5).
Stiver, G., 2003. The treatment of influenza with antiviral drugs. Can. Med. Assoc. J. 168 (1), 49–57.
Thursky, K., Cordova, S.P., Smith, D., Kelly, H., 2003. Working towards a simple case definition for influenza surveillance. J. Clin. Virol. 27 (2), 170–179.
Urashima, M., Shindo, N., Okabe, N., 2003. A seasonal model to simulate influenza oscillation in Tokyo. Jpn. J. Infect. Dis. 56 (2), 43–47.
Wen, T.-H., Lin, N.H., Lin, C.-H., King, C.-C., Su, M.-D., 2006. Spatial mapping of temporal risk characteristics to improve environmental health risk identification: a case study of a dengue epidemic in Taiwan. Sci. Total Environ. 367 (2), 631–640.
World Health Organization Influenza (Seasonal) 2016 [cited 2017 2017 Jun 11]. Available from: http://www.who.int/mediacentre/factsheets/fs211/en/.
Yang, S., Santillana, M., Kou, S.C., 2015. Accurate estimation of influenza epidemics using Google search data via ARGO. Proc. Natl. Acad. Sci. 112 (47), 14473–14478.
Zhang, Y., Milinovich, G., Xu, Z., Bambrick, H., Mengersen, K., Tong, S., et al., 2017. Monitoring pertussis infections using internet search queries. Sci. Rep. 7 (1), 10437.