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

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.