Forecasting influenza epidemics by integrating internet search queries and traditional surveillance data with the support vector machine regression model in Liaoning, from 2013 to 2015

Feng Liang [1] , **Peng Guan** [1] , **Wei Wu** [1] , **Desheng Huang** [Corresp. 1, 2]

[1] Deartment of Epidemiology, School of Public Health, China Medica University, Shenyang, Liaoning, China

[2] Department Mathematics, School of Fundamental Sciences, China Medical University, Shenyang, China

Corresponding Author: Desheng Huang
Email address: dshuang@cmu.edu.cn

Background. Influenza epidemic posed significant social and economic challenges in China. Internet search query data have been identified as a valuable source for the detection of emerging influenza epidemics. However, the selection of the search queries and the adoption of prediction method are the crucial challenges when it comes to improving predictions. The purpose of this study was to explore the application of SVM regression model in merging search engine query data and traditional influenza data.

Methods. Official monthly reported number of influenza cases in Liaoning province, China was acquired from China National Scientific Data Center for Public Health from January 2011 to December 2015. Based on Baidu Index, a publicly available search engine database, search queries potentially related to influenza over the corresponding period were identified. The Support Vector Machine (SVM) regression model was built to be used for predictions , the choice of three parameters (C, γ, ε) in the SVM regression model was determined by leave-one-out cross-validation (LOOCV) during the model construction process. The model's performance was evaluated by the evaluation metrics including Root Mean Square Error (RMSE), Root Mean Squared Percent Error (RMSPE) and Mean Absolute Percentage Error (MAPE). **Results.** In total, 17 search queries related to influenza were generated through the initial query selection approach and were adopted to construct the SVM regression model, including 9 queries in the same month, 3 queries at a lag of one month, 1 query at a lag of two months and 4 queries at a lag of three months. The SVM model performed well when with the parameters (C=2, γ=0.005 , ε=0.0001), based on the ensemble data integrating the influenza surveillance data and Baidu search query data .

Conclusions. The results demonstrated the feasibility of using internet search engine query data as the complementary data source for influenza surveillance and the efficiency of SVM regression model in tracking the influenza epidemics in Liaoning.

1 **Forecasting influenza epidemics by integrating internet search**

2 **queries and traditional surveillance data with the support vector**

3 **machine regression model in Liaoning, from 2013 to 2015**

4 **Feng Liang[1], Peng Guan[1], Wei Wu[1], Desheng Huang[1,2]**

5 [1] Department of Epidemiology, School of Public Health, China Medical University,

6 Shenyang 110122, China;

7 [2] Department of Mathematics, School of Fundamental Sciences, China Medical

8 University, Shenyang 110122, China

10 Corresponding Author:

11 Desheng Huang[1,2]

12 No. 77 Puhe Road, Shenyang North New Area, Shenyang 110122, China

13 Email address: dshuang@cmu.edu.cn.

14   **ABSTRACT**

15   **Background.** Influenza epidemic posed significant social and economic challenges in

16   China. Internet search query data have been identified as a valuable source for the

17   detection of emerging influenza epidemics. However, the selection of the search queries

18   and the adoption of prediction method are the crucial challenges when it comes to

19   improving predictions. The purpose of this study was to explore the application of SVM

20   regression model in merging search engine query data and traditional influenza data.

21   **Methods.** Official monthly reported number of influenza cases in Liaoning province,

22   China was acquired from China National Scientific Data Center for Public Health from

23   January 2011 to December 2015. Based on Baidu Index, a publicly available search

24   engine database, search queries potentially related to influenza over the corresponding

25   period were identified. The Support Vector Machine (SVM) regression model was built to

26   be used for predictions, the choice of three parameters (C, γ, ε) in the SVM regression

27   model was determined by leave-one-out cross-validation (LOOCV) during the model

28   construction process. The model's performance was evaluated by the evaluation metrics

29   including Root Mean Square Error (RMSE), Root Mean Squared Percent Error (RMSPE)

30   and Mean Absolute Percentage Error (MAPE). **Results.** In total, 17 search queries

31   related to influenza were generated through the initial query selection approach and were

32   adopted to construct the SVM regression model, including 9 queries in the same month,

33   3 queries at a lag of one month, 1 query at a lag of two months and 4 queries at a lag of

34   three months. The SVM model performed well when with the parameters (C=2, γ=0.005,

35   ε=0.0001), based on the ensemble data integrating the influenza surveillance data and

36   Baidu search query data. **Conclusions.** The results demonstrated the feasibility of using

37   internet search engine query data as the complementary data source for influenza

38   surveillance and the efficiency of SVM regression model in tracking the influenza

39   epidemics in Liaoning.

INTRODUCTION

Seasonal influenza is a serious public health problem and remains rampant across the world. According to the latest estimates from the United States Centers for Disease Control and Prevention (US-CDC), there are about 3 to 5 million cases of severe illnesses, and about 2.9 to 6.5 million deaths each year caused by influenza epidemics (*World Health Organization, 2018; World Health Organization, 2017*). National Health and Family Planning Commission of the People's Republic of China reported that China has 456,718 influenza cases with the incidence rate of 33.0994 per 100 thousand in 2017. *(National Health and Family Planning Commission of the People's Republic of China, 2018)*. Influenza epidemics posed significant social and economic Challenges in China (*Yang et al., 2015; Wang et al., 2015*). It is necessary to establish a real-time flu surveillance system for rapid and effective responses in China.

A national noticeable infectious disease reporting system has been established to continuously report the influenza cases in China, while the system reports the flu activity one month before, putting the flu data lagged for a month. Traditional flu surveillance methods for prediction were mainly based on hospital or laboratory data (*Wang et al., 2017*). The idea of applying internet search query data for the infectious diseases prediction was from Ginsberg in 2009 (*Ginsberg et al., 2009*), who presented a brand-new method providing nearly real-time surveillance of influenza-like illness and

60   overcoming the limitations of lag-time in the traditional flu surveillance systems of the

61   United States. Online search query data have a stronger tendency and immediacy and

62   can maintain full synchronization with the flu epidemic. In addition, internet search query

63   data can be measured in real time. In order to monitor the infectious diseases activity in

64   time, numerous studies have been emerging recently based on online search query data

65   or social media data, including Google (*Seo & Shin, 2017*; *Yang et al., 2017*; *Xu et al.*,

66   *2017*; *Pollett et al., 2017*), Yahoo (*Polgreen et al., 2008*),Naver (*Shin et al., 2016*),

67   Daum (*Woo et al., 2016*; *Seo et al., 2014*), Baidu search engine (*Guo et al., 2017*); Twitter

68   (*Wagner et al., 2017*; *Kagashe et al., 2017*; *Allen et al., 2016*; *Yun et al., 2016*) and Weibo

69   (*Fung et al., 2013*; *Zhang et al., 2015*) social media; Wikipedia (*Hickmann et al., 2015*;

70   *McIver & Brownstein, 2014*), hospital or clinicians' database (*Bouzille et al., 2018*;

71   *Santillana et al., 2014*), etc. As Google has been pulled out of mainland China in 2010,

72   Google search query data and Google Flu Trends cannot be accessible in mainland

73   China. This article will construct a forecasting model for influenza based on the ensemble

74   data integrating traditional influenza cases data and Baidu search data, which is the most

75   popular search engine in China.

76   Support vector machines (SVM) are supervised learning models with associated

77   learning algorithms, the application of SVM to classification and regression has been a

78   hot topic recently. For solving the regression problem, support vector machines have

79 been applied to many fields, air quality forecasting (*Liu et al.*, *2017*), water demanding

80 and water quality prediction (*Ghalehkhondabi et al.*, *2017; Zhang et al.*, *2017*), and

81 biomedicine (*Nickerson et al.*, *2016*), etc. SVM can efficiently perform a non-linear

82 classification which is based on the kernel trick, the inputs can be implicitly mapped into

83 high-dimensional feature spaces. Lampos, V et al. indicated that a nonlinear query

84 modeling approach presented the lowest cumulative nowcasting error (*Lampos et al.*,

85 *2015*). Woo, H et al. found that the SVM regression model based on weekly influenza

86 incidence data and query data from the Korean website Daum performed well (*Woo et*

87 *al.*, *2016*). Guo et al. comprehensively assessed six machine learning algorithms based

88 on Baidu search engine data and Dengue case data in Guangdong proposed that SVM

89 regression model had a better performance than other forecasting techniques (*Guo et al.*,

90 *2017*). Thus, this article attempted to build a SVM regression model to predict the flu

91 activity.

92   The influenza epidemic situation varies greatly among different regions. China has a

93 vast territory that spans tropical, subtropical and temperate regions, and it is a big

94 challenge to establish an influenza prediction mechanism in the whole country. Liaoning

95 province is located in both coastal and bordered region in the northeastern part of China,

96 where the feasibility of influenza prediction models based on internet search query data

97 is still unknown. Thus, the purpose of this study was to investigate whether an early

98   warning model utilizing both online influenza query data and traditional surveillance data

99   could improve influenza prediction.

100 **MATERIALS AND METHODS**

101 **Study Setting and Data collection**

102 Liaoning is a coastal province in the northeast of China with a population of approximately

103 43.77 million in 2016 and a temperate continental monsoon climate. Official monthly

104 reported number of influenza cases in Liaoning province, China was acquired from China

105 National Scientific Data Center for Public Health (http://www.phsciencedata.cn) from

106 January 2011 to December 2015. China National Scientific Data Center for Public Health

107 is open for those registered users in mainland China and the latest influenza incidence

108 data was the data in December 2015. Based on Baidu Index (http://index.baidu.com), an

109 online keyword research tool which is publicly open for the public across the globe, search

110 queries potentially related to influenza over the corresponding period were identified.

111 'Influenza' was first adopted as a primary indicator term to find more related queries about

112 influenza on the Chinese website (http://tool.chinaz.com/baidu/words.aspx). The website

113 is a free online platform that providing internet keyword mining of the Baidu search engine

114 in mainland China. Monthly average volume of those related search queries from Liaoning

115 was extracted from Baidu index website, from January 2011 to December 2015.

116 **Statistical Analysis**

117 The related influenza search terms were ranked first and those terms that having no

118 data within one calendar year during the study period were excluded. Pearson

119 correlation analysis was performed to explore the correlation between influenza-related

120 search queries and the reported number of influenza cases in Liaoning. Those search

121 terms that with the statistically significant correlation coefficient above 0.4 were sent to

122 the construction of SVM regression model. The selection of maximum cross-correlation

123 coefficient has been proposed in previous studies. *(Guo et al., 2017; Yuan et al., 2013)*

124 The influenza case surveillance data was divided into two parts, the fitting dataset

125 and the validation dataset in SVM regression model. Forty-five months' data from January

126 2011 to September 2014 was used for model training, and the rest 15 months' data from

127 October 2014 to December 2015 was used as the test set for model prediction. The

128 choice of three parameters (C, γ, ε) in the SVM regression model was determined by

129 leave-one-out cross-validation (LOOCV) during the model construction process. Three

130 metrics were adopted to measure the performance of the Support Vector Machine

131 regression model, including Root Mean Square Error (RMSE), Root Mean Squared

132 Percent Error (RMSPE), and Mean Absolute Percentage Error (MAPE). These three

133 metrics are measures of prediction accuracy of a forecasting method in statistics. RMSE

134   is very sensitive to the extreme errors or very small errors in a set of measurements,

135   therefore RMSE can well reflect the precision of the forecasting. RMSPE is a percent

136   difference between predicted and true values. MAPE is the most common measure of

137   forecast error and it functions best when there are no extremes to the data (including

138   zeros). The definitions of these three metrics are provided below. The notation in the study

139   is as follows: y_i denotes the observed value of the influenza cases at time t_i, $\hat{y}_i$ denotes

140   the predicted value by SVM regression model at time t_i.

141   Root Mean Squared Error (RMSE), a measure of the difference between predicted

142   and true values, is defined as:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \qquad (1)$$

143   Root Mean Squared Percent Error (RMSPE), a measure of the percent difference

144   between predicted and true values, is defined as:

$$RMSPE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\frac{y_i - \hat{y}_i}{y_i})^2} \times 100\% \qquad (2)$$

145   Mean Absolute Percentage Error (MAPE), is the mean or average of the absolute

146   percentage errors of forecasts and is defined as:

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}|\ \frac{y_i - \hat{y}_i}{y_i}\ |\times 100\% \tag{3}$$

147   The statistical analysis and the construction of Support Vector Machine regression

148   model were performed using R statistical software version 3.4.2 with package e1071.

149   **RESULTS**

150   **Baidu search terms filtering.**

151   According to the filtering criteria, there were 46 search terms left due to the available

152   sequential data within one calendar year during the study period (Table S1). The Pearson

153   correlation analysis was made between search terms and influenza cases across different

154   lag periods (in the same month, at a lag of one month, at a lag of two months and three

155   months). The correlation value of search terms with the influenza cases in the non-flu

156   season is provided as a basal level of their relationship (Table 1). Twenty-nine of the

157   remaining 46 terms were excluded, because their Pearson correlation coefficients

158   between the search terms and influenza cases were less than 0.4 across every lag period.

159   A total of 17 search terms which were strongly correlated with influenza cases across

160   different lag periods were retained for the construction of SVM regression model,

161   including 9 queries in the same month, 3 queries at a lag of one month, 1 query at a lag

162   of two months and 4 queries at a lag of three months (Table 2). Meanwhile, the amount

163   of influenza cases might have an impact on the amount of incident cases of the following

164   months. The Pearson correlation analysis was performed to compare the relationship

165   between the reported number of influenza cases of the month and historically reported

166   number of influenza cases. The correlation coefficients were 0.672, 0.498 and 0.151 at

167   the lag time of one month, two months and three months, respectively. The reported

168   number of influenza cases at the lag time of one month has shown the strongest

169   correlation, thus it was submitted to SVM regression model.

170   **Parameter selection of SVM regression model**

171   The mathematical formula of SVM regression model is provided below: $\alpha,\ \alpha^*$ are

172   Lagrangian operator. C is the upper bound of all variables, Q is a k by k positive

173   semidefinite matrix, $Q_{ij} = y_i y_j K\ (x_i,\ x_j)$, and $K(x_i,\ x_j)$ is the kernel.

$$\min_{\alpha,\alpha^*} \frac{1}{2}\left(\alpha - \alpha^*\right)^{\mathrm{T}} Q\left(\alpha - \alpha^*\right) + \varepsilon \sum_{i=1}^{k}\left(\alpha_i + \alpha_i^*\right) + \sum_{i=1}^{k} y_i\left(\alpha_i - \alpha_i^*\right)$$

$$\mathrm{s.t.}\quad 0 \le \alpha_i, \alpha_i^* \le C, i = 1,\dots,k,\qquad\qquad (4)$$

$$\sum_{i=1}^{k}\left(\alpha_i - \alpha_i^*\right) = 0.$$

174   The expression of radial basis function is provided below:

$$K\left(x_i, x_j\right) = \exp\left\{-\gamma \mid x_i - x_j \mid^2\right\} \tag{5}$$

175 During the process of leave-one-out cross-validation, we started from the default

176 value (γ = 0.0556, ε = 0.1), then we adjusted the C value to observe the model fitting

177 results (Table 3). The same method was applied to the selection of the other two

178 parameters, γ and ε (Table 4, Table 5). The values of these three parameters were

179 evaluated according to the lowest test error, then the optimal parameters of the model

180 were determined (C=2, γ=0.005, ε=0.0001).

181 **Comparison and prediction of SVM regression models from different data**

182 **sources.**

183 Compared with the model based on influenza cases data at the lag time of one month

184 and the source of Baidu search data, the SVM regression model based on ensemble data

185 integrating historical influenza surveillance data and Baidu search data showed the best

186 accuracy with lowest RMSE (42.654) and best robustness with lowest MAPE (26.197%),

187 as seen in Table 6.

188 The predicted values of the above three models and the actual number of influenza

189 cases from October 2014 to December 2015 have been presented in Figure 1. It was

190 easily to find that the SVM model prediction's curve was almost identical when comparing

191 the model based on internet search query data with the model based on ensemble data,

192   and trend of the curve were consistent with the overall development trend of the actual

193   influenza cases curve. The SVM regression model based on ensemble data was capable

194   of predicting the timing and magnitude of most periods, whereas it failed to predict the

195   influenza outbreak peak in March 2015. The predictions of the model based on flu data

196   at the lag time of one month were significantly lower than the actual value in the previous

197   six-month forecasting, but the overall trend was consistent in the following 9 months. The

198   residual of each predictor is displayed in Figure 2.

199   **DISCUSSION**

200   This article presented an efficient SVM regression model to predict flu activity and track

201   the epidemic orbit in Liaoning province of China. The entire analysis demonstrated that

202   the SVM regression model based on ensemble data was better than the model based on

203   conventional data. It could be a complement to the traditional surveillance for influenza

204   dynamics.

205   With the rapid development and popularity of the Internet, this new method of

206   infectious diseases surveillance system based on online search query data is more

207   convenient and accurate. According to the 41st Statistical Report on Internet Development

208   released by China Internet Network Information Center (CINIC), the internet users in

209   China have steadily increased and up to 772 million, and the internet popularity rate

210 reached 55.8%, exceeding the global average level until Dec 2017 (*China Internet*

211 *Network Information Center, 2018*). Baidu search engine is the most widely accepted

212 search engine in China, making it the most representative and available data source for

213 the studies targeting tracking the online seeking behavior of Chinese people. Based on

214 Baidu search query data, Chinese scholars have made great efforts in the field of disease

215 monitoring, such as Norovirus (*Liu et al.*, *2017*), Dengue (*Li et al.*, *2017*), Hand, foot, and

216 mouth disease (HFMD) (*Du Z et al.*, *2017*), and epidemic erythromelalgia (EM) (*Gu et al.*,

217 *2015*). These forecasting models got great performances in the field of early warning.

218 However, most of the researches focused on southeastern coastal regions of China, such

219 as Guangdong and Zhejiang, and few disease prediction models was constructed and

220 applied in the coastal areas in the northeast China. It is a significant attempt to predict the

221 influenza activity in Liaoning province located in the northeast of China. This article could

222 provide some hints and lessons for the flu forecasting and alerting in the Northeast of

223 China.

224 Strong correlation between influenza cases and search terms of Baidu was found in

225 the present study. Influenza is characterized by a short incubation period and a sudden

226 onset of symptoms such as fever, cough (usually dry), etc., it is reasonable that most of

227 search terms about flu virus, symptoms and therapy closely correlated with influenza

228 cases at the same month. The most effectively vaccine injection timing is about one to

229   two months prior to the flu season. Winter and spring are the peak flu seasons in Liaoning,

230   China, thus September and October are the best months for flu vaccination in the study

231   area. The search behavior about flu vaccine is often earlier than the vaccination timing,

232   so we found that the search terms of Baidu at a lag of 3 months had a strong correlation

233   with the occurrence of influenza cases.

234   The present study showed that the forecasting model based on internet search query

235   was better than the model based on traditional data in terms of accuracy and stability.

236   The results were consistent with the results of other studies (*Guo et al.*, *2017*; *Yuan et*

237   *al.*, *2013*). However, Olson et al. (*Olson et al.*, *2013*) investigated the reliability of Google

238   Flu Trends (GFT) of 2003 to 2013 and compared the flu timing and intensity between

239   forecasting data and actual influenza incidence at the national, regional and local levels.

240   They concluded that GFT data could not serve as the reliable surveillance for seasonal

241   or pandemic influenza and traditional surveillance are still irreplaceable. The main reason

242   was that GFT was based on internet data without considering the epidemiological factors

243   such as the age distribution of patients, geographical location, illness complaints or

244   clinical manifestations. Our study proved the advantages of ensemble source data

245   integrating traditional influenza incidence data and search engine data in the field of

246   forecasting. Meanwhile, there may be some space to improve the SVM model presented

247   in the present study. Although most of the forecasting values were fitted well with actual

248   influenza cases in the SVM regression model, they failed to identify the influenza's peak

249   in March 2015. The climatic factors have great impact on Influenza incidence *(Gomez-*

250   *Barroso et al., 2017),* thus the possible reason of their missing might be that March are

251   the cold month and the flu peak seasons in northeastern China while internet search

252   query data could not distinguish the situation.

253   Several limitations in influenza forecasting model based on ensemble data integrating

254   traditional influenza cases data and Baidu search engine data need to be mentioned.

255   Firstly, media report may influence the internet searching behavior, which will have an

256   impact on the performance of forecasting model directly. In addition, without considering

257   the impact factors of influenza, such as seasonal and meteorological factors, the

258   forecasting results may have bias to some degree. Furthermore, correlation analysis of

259   the search keywords mainly was based on previous vocabularies data. However, in pace

260   with the rapid changes of the internet environment, many fresh online search vocabularies

261   produced at every moment. The fresh vocabularies were hard to be tracked and usually

262   have been overlooked.

263   **CONCLUSIONS**

264   The present study built a forecasting model based on ensemble data integrating Baidu

265   search query data and traditional flu data in Liaoning province. The model based on

266 ensemble data showed the best accuracy and best robustness in SVM regression model,

267 rather than the models based on other single data sources. It could be a complement of

268 the traditional surveillance for influenza dynamics in Liaoning.

269 **ACKNOWLEDGMENTS**

270 This study was partly based on the information from the data provided by China National

271 Scientific Data Center for Public Health and Baidu index website.

272 **REFERENCES:**

273 **Allen C, Tsou M H, Aslam A, Nagel A, Gawron J M**. **2016**. Applying GIS and Machine

274 Learning Methods to Twitter Data for Multiscale Surveillance of Influenza. *PLoS*

275 *One.* **11**:e157734. DOI: 10.1371/journal.pone.0157734

276 **Bouzille G, Poirier C, Campillo-Gimenez B, Aubert M L, Chabot M, Chazard E,**

277 **Lavenu A, Cuggia M**. **2018**. Leveraging hospital big data to monitor flu epidemics.

278 *Comput Methods Programs Biomed.* **154**:153-160. DOI:

279 10.1016/j.cmpb.2017.11.012

280 **China Internet Network Information Center. 2018.** The 41st Statistical Report on

281 Internet Development. Available at:

282 http://cnnic.cn/gywm/xwzx/rdxw/201801/t20180131_70188.htm (accessed 12

283 March 2018)

284 **Du Z, Xu L, Zhang W, Zhang D, Yu S, Hao Y**. **2017**. Predicting the hand, foot, and

285   mouth disease incidence using search engine query data and climate variables: an

286   ecological study in Guangdong, China. *BMJ Open.* **7**:e16263. DOI:

287   10.1136/bmjopen-2017-016263

288   **Fung I C, Fu K W, Ying Y, Schaible B, Hao Y, Chan C H, Tse Z T**. **2013**. Chinese

289   social media reaction to the MERS-CoV and avian influenza A(H7N9) outbreaks.

290   *Infectious Diseases of Poverty.* **2**:31. DOI: 10.1186/2049-9957-2-31

291   **Ghalehkhondabi I, Ardjmand E, Young W N, Weckman G R**. **2017**. Water demand

292   forecasting: review of soft computing methods. *Environmental Monitoring and*

293   *Assessment.* **189**:313. DOI: 10.1007/s10661-017-6030-3

294   **Ginsberg J, Mohebbi M H, Patel R S, Brammer L, Smolinski M S, Brilliant L**. **2009**.

295   Detecting influenza epidemics using search engine query data. *Nature.* **457**:1012-

296   1014. DOI: 10.1038/nature07634

297   **Gomez-Barroso D, Leon-Gomez I, Delgado-Sanz C, Larrauri A**. **2017**. Climatic

298   Factors and Influenza Transmission, Spain, 2010-2015. *Int J Environ Res Public*

299   *Health.* **14**. DOI: 10.3390/ijerph14121469

300   **Gu Y, Chen F, Liu T, Lv X, Shao Z, Lin H, Liang C, Zeng W, Xiao J, Zhang Y, Huang**

301   **C, Rutherford S, Ma W**. **2015**. Early detection of an epidemic erythromelalgia

302   outbreak using Baidu search data. *Sci Rep.* **5**:12649. DOI: 10.1038/srep12649

303   **Guo P, Liu T, Zhang Q, Wang L, Xiao J, Zhang Q, Luo G, Li Z, He J, Zhang Y, Ma**

304   **W**. **2017**. Developing a dengue forecast model using machine learning: A case

305   study in China. *PLoS Negl Trop Dis.* **11**:e5973. DOI: 10.1371/journal.pntd.0005973

306   **Guo P, Zhang J, Wang L, Yang S, Luo G, Deng C, Wen Y, Zhang Q**. **2017**.

307   Monitoring seasonal influenza epidemics by using internet search data with an

308   ensemble penalized regression model. *Sci Rep* **7**:46469. DOI: 10.1038/srep46469

309   **Hickmann K S, Fairchild G, Priedhorsky R, Generous N, Hyman J M, Deshpande**

310   **A, Del V S**. **2015**. Forecasting the 2013-2014 influenza season using Wikipedia.

311   *PLoS Computational Biology.* **11**:e1004239. DOI: 10.1371/journal.pcbi.1004239

312   **Kagashe I, Yan Z, Suheryani I**. **2017**. Enhancing Seasonal Influenza Surveillance:

313   Topic Analysis of Widely Used Medicinal Drugs Using Twitter Data. *Journal of*

314   *Medical Internet Research.* **19**:e315. DOI: 10.2196/jmir.7393

315   **Lampos V, Miller A C, Crossan S, Stefansen C**. **2015**. Advances in nowcasting

316   influenza-like illness rates using search query logs. *Sci Rep.* **5**:12760. DOI:

317   10.1038/srep12760

318   **Li Z, Liu T, Zhu G, Lin H, Zhang Y, He J, Deng A, Peng Z, Xiao J, Rutherford S, Xie**

319   **R, Zeng W, Li X, Ma W**. **2017**. Dengue Baidu Search Index data can improve the

320   prediction of local dengue epidemic: A case study in Guangzhou, China. *PLoS Negl*

321   *Trop Dis.* **11**:e5354. DOI: 10.1371/journal.pntd.0005354

322   **Liu B C, Binaykia A, Chang P C, Tiwari M K, Tsao C C**. **2017**. Urban air quality

323   forecasting based on multi-dimensional collaborative Support Vector Regression

324   (SVR): A case study of Beijing-Tianjin-Shijiazhuang. *PLoS One.* **12**:e179763. DOI:

325   10.1371/journal.pone.0179763

326   **Liu K, Huang S, Miao Z P, Chen B, Jiang T, Cai G, Jiang Z, Chen Y, Wang Z, Gu H,**

327   **Chai C, Jiang J**. **2017**. Identifying Potential Norovirus Epidemics in China via

328   Internet Surveillance. *Journal Of Medical Internet Research.* **19**:e282. DOI:

329   10.2196/jmir.7855

330   **McIver D J, Brownstein J S**. **2014**. Wikipedia usage estimates prevalence of influenza-

331   like illness in the United States in near real-time. *PLoS Computational Biology.*

332   **10**:e1003581. DOI: 10.1371/journal.pcbi.1003581

333   **National Health and Family Planning Commission of the People's Republic of**

334   **China. 2018.** National Statutory Epidemic Situation in 2017. Available at

335   http://www.nhfpc.gov.cn/jkj/s3578/201802/de926bdb046749abb7b0a8e23d929104.

336   shtml (accessed 11 March 2018)

337   **Nickerson P, Tighe P, Shickel B, Rashidi P**. **2016**. Deep neural network architectures

338   for forecasting analgesic response. *Conf Proc IEEE Eng Med Biol Soc.* **2016**:2966-

339   2969. DOI: 10.1109/EMBC.2016.7591352

340   **Olson D R, Konty K J, Paladini M, Viboud C, Simonsen L**. **2013**. Reassessing

341   Google Flu Trends data for detection of seasonal and pandemic influenza: a

342   comparative epidemiological study at three geographic scales. *PLoS Computational*

343   *Biology.* **9**:e1003256. DOI: 10.1371/journal.pcbi.1003256

344   **Polgreen P M, Chen Y, Pennock D M, Nelson F D**. **2008**. Using internet searches for

345   influenza surveillance. *Clinical Infectious Diseases.* **47**:1443-1448. DOI:

346   10.1086/593098

347   **Pollett S, Boscardin W J, Azziz-Baumgartner E, Tinoco Y O, Soto G, Romero C,**

348   **Kok J, Biggerstaff M, Viboud C, Rutherford G W**. **2017**. Evaluating Google Flu

349   Trends in Latin America: Important Lessons for the Next Phase of Digital Disease

350   Detection. *Clinical Infectious Diseases.* **64**:34-41. DOI: 10.1093/cid/ciw657

351   **Santillana M, Nsoesie E O, Mekaru S R, Scales D, Brownstein J S**. **2014**. Using

352   clinicians' search query data to monitor influenza epidemics. *Clinical Infectious*

353   *Diseases.* **59**:1446-1450. DOI: 10.1093/cid/ciu647

354   **Seo D W, Jo M W, Sohn C H, Shin S Y, Lee J, Yu M, Kim W Y, Lim K S, Lee S I**.

355   **2014**. Cumulative query method for influenza surveillance using search engine data.

356   *Journal Of Medical Internet Research.* **16**:e289. DOI: 10.2196/jmir.3680

357   **Seo D W, Shin S Y**. **2017**. Methods Using Social Media and Search Queries to Predict

358   Infectious Disease Outbreaks. *Healthc Inform Res.* **23**:343-348. DOI:

359   10.4258/hir.2017.23.4.343

360   **Shin S Y, Kim T, Seo D W, Sohn C H, Kim S H, Ryoo S M, Lee Y S, Lee J H, Kim W**

361   **Y, Lim K S**. **2016**. Correlation between National Influenza Surveillance Data and

362   Search Queries from Mobile Devices and Desktops in South Korea. *PLoS One.*

363   **11**:e158539. DOI: 10.1371/journal.pone.0158539

364   **Wagner M, Lampos V, Yom-Tov E, Pebody R, Cox I J**. **2017**. Estimating the

365   Population Impact of a New Pediatric Influenza Vaccination Program in England

366   Using Social Media Content. *Journal of Medical Internet Research.* **19**:e416. DOI:

367   10.2196/jmir.8184

368   **Wang C, Li Y, Feng W, Liu K, Zhang S, Hu F, Jiao S, Lao X, Ni H, Xu G**. **2017**.

369   Epidemiological Features and Forecast Model Analysis for the Morbidity of Influenza

370   in Ningbo, China, 2006-2014. *Int J Environ Res Public Health.* **14**. DOI:

371   10.3390/ijerph14060559

372   **Wang X, Cai J, Yao W, Zhu Q, Zeng M**. **2015**. Socio-economic impact of influenza in

373   children:a single-centered hospital study in Shanghai. *Zhonghua Liu Xing Bing Xue*

374   *Za Zhi.* **36**:27-30.

375   **World   Health   Organization.   2018.**   Influenza   (Seasonal).   Available   at:

376   http://www.who.int/mediacentre/factsheets/fs211/en/ (accessed   11 March 2018).

377   **World Health Organization. 2017.** Up to 650 000 people die of respiratory diseases

378   linked to seasonal flu each year. Available at:

379   http://www.who.int/mediacentre/news/releases/2017/seasonal-flu/en/ (accessed 11

380   March 2018)

381   **Woo H, Cho Y, Shim E, Lee J K, Lee C G, Kim S H**. **2016**. Estimating Influenza

382   Outbreaks Using Both Search Engine Query Data and Social Media Data in South

383   Korea. *Journal Of Medical Internet Research.* **18**:e177. DOI: 10.2196/jmir.4955

384   **Xu Q, Gel Y R, Ramirez R L, Nezafati K, Zhang Q, Tsui K L**. **2017**. Forecasting

385   influenza in Hong Kong with Google search queries and statistical model fusion.

386   *PLoS One.* **12**:e176690. DOI: 10.1371/journal.pone.0176690

387   **Yang J, Jit M, Leung K S, Zheng Y M, Feng L Z, Wang L P, Lau E H, Wu J T, Yu H**

388   **J**. **2015**. The economic burden of influenza-associated outpatient visits and

389   hospitalizations in China: a retrospective survey. *Infectious Diseases of Poverty.*

390   **4**:44. DOI: 10.1186/s40249-015-0077-6

391   **Yang S, Santillana M, Brownstein J S, Gray J, Richardson S, Kou S C**. **2017**. Using

392   electronic health records and Internet search information for accurate influenza

393   forecasting. *BMC Infectious Diseases* **17**:332. DOI: 10.1186/s12879-017-2424-7

394   **Yuan Q, Nsoesie E O, Lv B, Peng G, Chunara R, Brownstein J S**. **2013**. Monitoring

395   influenza epidemics in china with search query from baidu. *PLoS One* **8**:e64323.

396   DOI: 10.1371/journal.pone.0064323

397   **Yun G W, Morin D, Park S, Joa C Y, Labbe B, Lim J, Lee S, Hyun D**. **2016**. Social

398   media and flu: Media Twitter accounts as agenda setters. *International Journal Of*

399   *Medical Informatics* **91**:67-73. DOI: 10.1016/j.ijmedinf.2016.04.009

400   **Zhang E X, Yang Y, Di Shang R, Simons J J, Quek B K, Yin X F, See W, Oh O S,**

401   **Nandar K S, Ling V R, Chan P P, Wang Z, Goh R S, James L, Tey J S**. **2015**.

402   Leveraging social networking sites for disease surveillance and public sensing: the

403   case of the 2013 avian influenza A(H7N9) outbreak in China. *Western Pac Surveill*

404   *Response J.* **6**:66-72. DOI: 10.5365/WPSAR.2015.6.1.013

405   **Zhang L, Zou Z, Shan W**. **2017**. Development of a method for comprehensive water

406   quality forecasting and its application in Miyun reservoir of Beijing, China. *J Environ*

407   *Sci (China).* **56**:240-246. DOI: 10.1016/j.jes.2016.07.017

Pearson correlation coefficients between search terms from Baidu search engine and the number of influenza cases in Liaoning, for the time period Jan 2013-Dec 2015

Search terms	The same month	Lag one month	Lag two months	Lag three months	Non-flu season
Flu	**0.608****	0.536**	0.500**	0.395**	0.672**
flu symptoms	**0.618****	0.374**	0.273*	0.04	0.430*
influenza type A	**0.489****	0.134	0.005	-0.287*	0.048
influenza vaccine	0.259*	0.436**	0.645**	**0.764****	0.328
Is it necessary to get vaccinated against the flu?	0.103	0.362**	0.626**	**0.814****	0.255
flu virus	**0.621****	0.435**	0.282*	0.175	0.711**
the symptom of flu	**0.656****	0.438**	0.209	-0.054	0.235
influenza drugs	**0.639****	0.431**	0.218	-0.116	0.337
the symptoms of type A flu	0.157	0.029	-0.061	-0.225	0.320
prevent flu	0.623**	0.644**	**0.663****	0.511**	0.374*
swine flu	0.371**	0.290*	0.172	0.075	0.459*
H1N1 flu	0.021	-0.157	-0.231	**-0.409****	-0.302
Beijing flu	0.249	0.055	0.037	-0.148	0.313
swine flu symptoms	0.032	0.142	-0.104	-0.101	-0.039
how to prevent flu	0.023	-0.025	0.056	0.007	-0.006
viral flu	**0.484****	0.339**	0.234	-0.023	0.587**
how to prevent flu	0.129	0.087	0.201	0.125	-0.014
Spanish flu	0.459**	**0.479****	0.405**	0.409**	0.491**

flu prevention	0.178	0.012	-0.092	-0.133	0.380*
side effects of flu vaccine	0.084	0.379**	0.657**	**0.804****	0.254
the prevention measures of flu	-0.079	-0.056	0.043	0.061	-0.108
type A H1N1 flu	0.089	-0.024	-0.281*	-0.362**	-0.379*
flu therapy	**0.438****	0.183	0.168	-0.075	0.291
the prevention of flu	-.335**	-0.357**	-0.251	-0.285*	-0.394*
influenza epidemic	0.09	-0.075	-0.207	-0.375**	0.108
influenza vaccine price	-0.273*	-0.185	-0.002	0.126	-0.124
type A flu	0.383**	0.326*	0.025	-0.193	0.508**
type A flu virus	**0.586****	0.395**	0.276*	0.002	0.369*
new type of flu	0.352**	**0.430****	0.037	-0.076	0.387*
type A influenza	0.016	-0.222	-0.138	-0.345**	-0.197
love flu strain	0.054	-0.165	-0.166	-0.296*	-0.425*
flu concept stock	0.266*	0.226	0.011	-0.072	0.349
seasonal influenza	-0.046	-0.122	-0.12	-0.172	0.515**
love flu	-0.300*	-0.310*	-0.350**	-0.368**	-0.114
type A H1N1 flu virus	-0.109	-0.22	-0.24	-0.131	-0.160
new flu	-0.048	-0.064	-0.151	-0.25	-0.016
how to treat swine flu	0.223	**0.547****	0.476**	0.381**	0.200

Influenza transmission route	0.228	0.1	-0.09	-0.076	0.346
The route of transmission of flu	0.216	-0.011	-0.12	-0.164	0.207
treatment program of A type H1N1 flu	-0.145	-0.19	-0.24	-.304*	0.102
flu (space) symptom	0.147	0.042	-0.068	-0.157	0.358
H1N1 flu symptom	0.346**	0.196	0.141	0.064	0.260
Sheep flu	0.011	-0.038	-0.053	-0.107	0.211
super flu	0.314*	0.248	0.133	0.16	0.172
the symptom of swine flu	0.134	0.071	-0.025	-0.124	0.341
Taiwan flu	0.124	0.066	0.104	-0.188	0.374*

2 Note: * indicates the P value with statistically significance at 0.05 level.

3 ** indicates the P value with statistically significance at 0.01 level.

Strongly correlated search terms with the number of influenza cases in different lag periods

Lag time	Search keywords
The same month	Flu, flu symptoms, influenza type A, flu virus, the symptoms of flu, influenza drugs, viral flu, flu therapy, type A flu virus
Lag one month	Spanish flu, new type of flu, how to treat swine flu
Lag two months	prevent flu
Lag three months	influenza vaccine, is it necessary to get vaccinated against the flu, H1N1 flu, side effects of flu vaccine

The SVM model precision of different C values.(γ=0.05556, ε=0.1)

C	Train Error	Test Error
0.0001	8812.675	8834.564
0.001	8768.452	8806.329
0.01	8363.176	8532.467
0.1	5831.012	6661.826
1	14647.06	4052.645
2	498.3551	**3900.983**
3	215.0484	4003.317
4	175.4402	4116.998
5	147.7603	4215.681
10	76.7374	4756.99
100	**71.55703**	4792.467

The SVM model precision of different γ values. (C=1,ε=0.1)

γ	Train Error	Test Error
0.0001	8130.444	8199.893
0.001	4533.864	4933.536
0.005	1932.273	**3239.05**
0.01	1604.143	3502.44
0.02	1493.212	3655.345
0.03	1465.351	3773.955
0.04	1466.39	3873.576
0.05	**1459.852**	3982.198
0.1	1470.783	4796.493
1	2441.952	7936.262

The SVM model precision of different ε values. (C=1, γ=0.05556)

ε	Train Error	Test Error
0.0001	**1388.189**	**3985.145**
0.001	1388.479	3986.231
0.01	1392.287	3993.167
0.05	1412.391	4027.2
0.08	1440.697	4051.417
0.09	1452.264	4052.576
0.1	1464.717	4052.645
0.2	1539.904	4060.86
0.3	1709.969	4153.16
0.4	1965.82	4346.04
0.5	2361.972	4643.752
1	5522.041	7408.229

Similarity metrics between 3 data sources: the number of influenza cases at a lag of one month, Baidu keywords, ensemble data, for the time period Oct 2014-Dec 2015

	RMSE	RMSPE (%)	MAPE (%)
Influenza cases at a lag of one month	82.874	40.658	35.150
Baidu keywords	43.472	30.438	26.806
Ensemble model	**42.654**	**29.687**	**26.197**

Figure 1

The performance of the three available predictors.
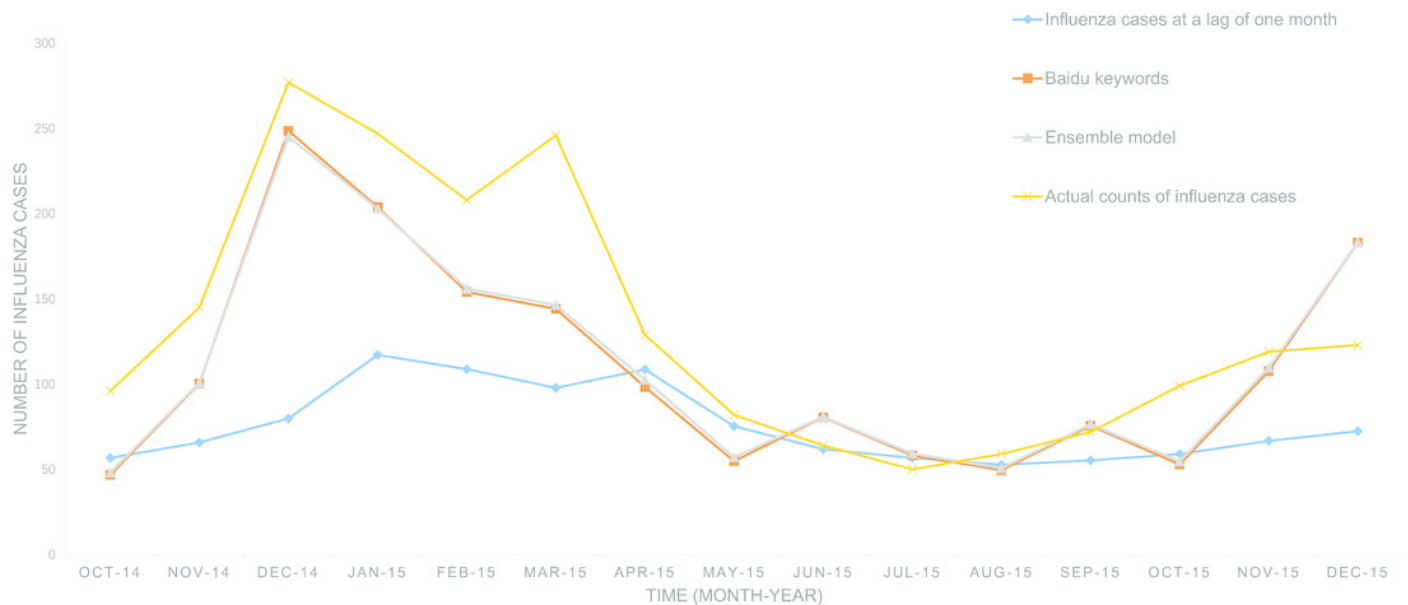
Figure 2

The residual of the three predictors.