Since January 2020 Elsevier has created a COVID-19 resource centre with free information in English and Mandarin on the novel coronavirus COVID-19. The COVID-19 resource centre is hosted on Elsevier Connect, the company's public news and information website.

Journal of Infection 84 (2022) e1–e3

Contents lists available at ScienceDirect

Journal of Infection

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

Letter to the Editor

A model study on predicting new COVID-19 cases in China based on social and news media

An article in this Journal reported that the WeChat keyword index had a strong correlation with the trend of COVID-19 in China.[1] Compare with traditional monitoring method, forecasting and early warning of infectious diseases based on the social and news media data with the advantages of timeliness and low cost.[2] Therefore, there have been some reports on the forecasting and early warning of COVID-19 based on Twitter, Weibo, and other social media indexes or posts.[3–7] However, data for the above social media-based predictions of the COVID-19 pandemic were sourced only from single social media platforms, such as Twitter and Weibo. As users aged under 30 years account for more than half of Twitter and Weibo users, the user demographics of Twitter and Weibo are too one-sided in age composition, which is not preferable in statistical analysis; thus, Twitter and Weibo cannot represent all social media.[5] In addition, few reports use news media data, which will lead to low comprehensiveness and objectivity of prediction.

In this study, we collected the daily new confirmed cases in China released by the National Health Commission from January 1, 2020, to March 18, 2020, totaling 78 days, as the data for analysis and prediction (See Supplementary Appendix).[8] We chose these dates because the health commissions at all levels in China officially began to count newly confirmed cases every day starting from January 1, 2020, and newly confirmed cases in Wuhan, China, fell to 0 on March 18, 2020. Correspondingly, a web crawler technology was used to capture public information from major news websites, electronic newspapers, Weibo, WeChat, and other APPs. The meta-search crawler obtained data from search engine webpages using 32 keywords, such as “fever”, “pyrexia”, and “cough” et al. (See Supplementary Appendix). We included more than 1000 mainstream news outlets and electronic newspapers in China, such as China News, China Daily, and People’s Daily. By analyzing data from major news media websites, electronic newspapers, Weibo, WeChat, and other APPs related to the COVID-19 pandemic, we obtained the daily total relative index of each keyword sourced from different platforms.

We calculated the daily total relative indexes of the 32 keywords and their correlation coefficients with daily new confirmed COVID-19 cases in China (Fig. 1A). The keywords showing strong correlation (Pearson correlation coefficients $>$ 0•8) with new confirmed cases were identified to be “fever”, “cough”, “fatigue”, “coronavirus”, and “novel coronavirus”. We plotted the daily relative indexes of these five keywords and the trend curves of daily new confirmed cases in China for visual analysis (Fig. 1B). The trend curves showed that “coronavirus” and “novel coronavirus” had the best correlation with new confirmed cases in China, which is consistent with the histogram results.

The SARS-CoV-2 infection has an incubation period of 1–14 days, there may be a certain lag period before the relative indexes of keywords can show a correlation with new confirmed cases. By shifting the data to calculate the correlation coefficient (See Supplementary Appendix), we found that the Pearson and Spearman correlation coefficients of these five keywords with new confirmed cases in China both reached their peak at the time lag of 3 days (Fig. 1C and D), indicating that the relative indexes of keywords on that day had the greatest correlation with the number of new confirmed cases in China 3 days later. The Pearson correlation coefficient of the five keywords reached 0.84, 0.82, 0.86, 0.92 and 0.92 respectively at the time lag of 3 days (See Supplementary Appendix).

Our study involves more than 1000 social and news media sources, we calculated the correlation coefficients between the relative indexes of keywords with strong correlation in various media and the number of new confirmed cases in China with a 3-day lag (See Supplementary Appendix). Our results clearly showed that Weibo and electronic newspapers had relatively low average correlation coefficients with new confirmed cases in China (Pearson correlation coefficients $<$ 0•8). The average Pearson correlation coefficients of other media (Wechat, news, websites and APPs) reached 0.83, 0.87, 0.83 and 0.88 respectively. Therefore, in the subsequent analyses, we excluded the relative indexes of Weibo and electronic newspapers and kept only the sum of the relative indexes of WeChat, news, websites, and APPs.

Ultimately, for daily new confirmed cases and keywords index, we used principal component analysis, best subset selection, partial least squares regression, stepwise regression, and elastic net regression as prediction models respectively to eliminate collinearity and over-fitting. Moreover, we identified the optimal prediction model by comparing the residuals of each model and used the 10-fold cross-validation method and 0•632 bootstrapping to verify each model (See Supplementary Appendix). The number of predictive variables of each model as well as the adjusted R^2, cross-validation R^2, cross-validation standard deviation S, adjusted residual sum of squares (RSS), and adjusted mean square error (MSE) were selected from the results and compared (Table 1).

Table 1 shows that best subset selection, partial least squares regression, stepwise regression, and elastic net regression all achieved good performance and prediction accuracy. Partial least squares regression was the best model we identified according to the parameters. It had the best performance and the lowest error among the five models (Adjusted R^2=93•75%, Cross-validation R^2=88•26%, RSS=7•143, MSE=0•101). The effects of 0.632 bootstrapping training set and prediction set verify the results of Table 1 (See Supplementary Appendix). In the results of predicting future cases by using the date before February 19, 2020 as the training set, the R^2 of principal component analysis, best subset

https://doi.org/10.1016/j.jinf.2022.01.009

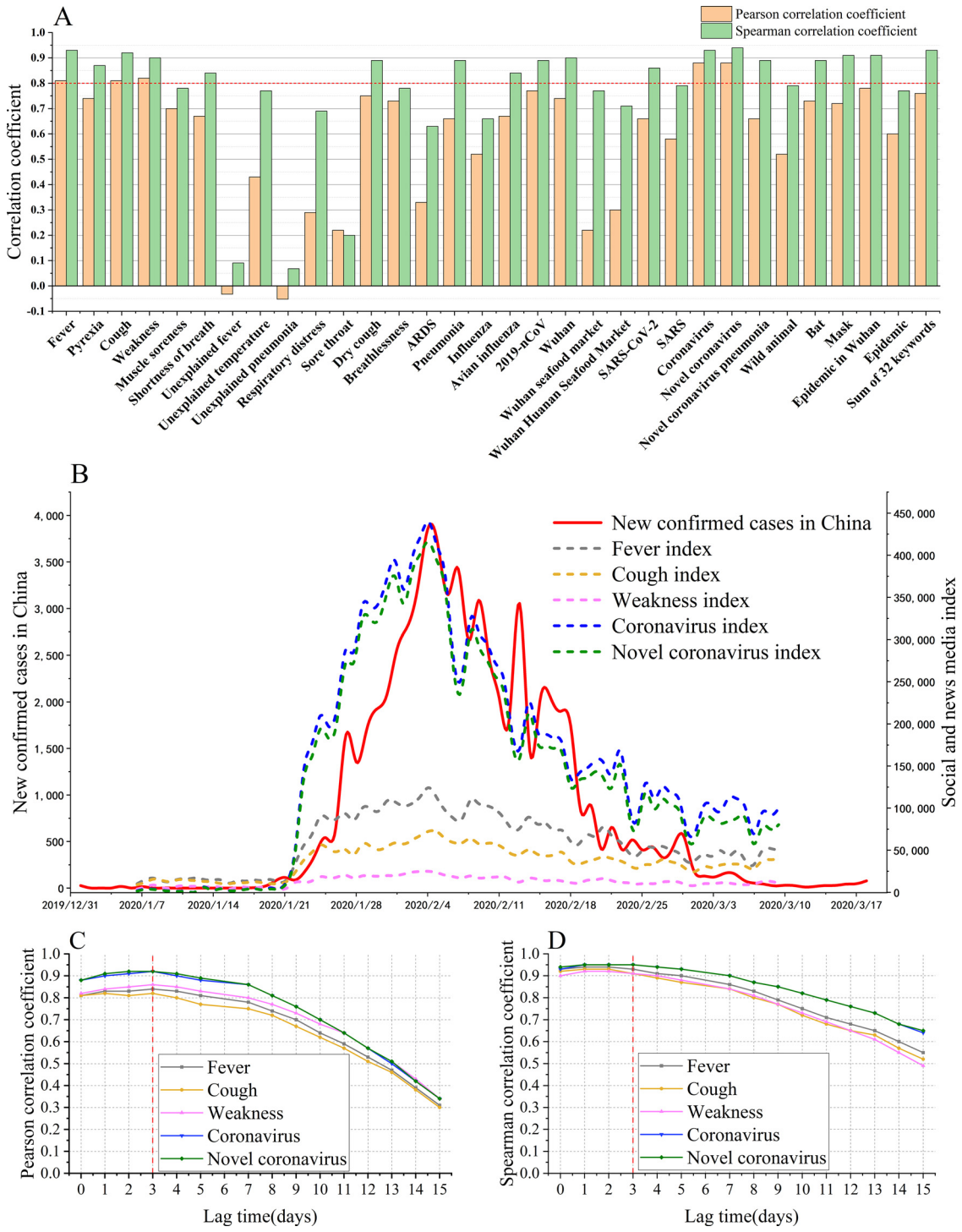

Fig. 1. Correlation and hysteresis between the relative indexes of 32 keywords and daily new confirmed cases in China. (A) Correlation coefficients between relative indexes of keywords and daily new confirmed cases in China. (B) Trend curves of relative indexes of keywords and daily new confirmed cases in China. (C) Time lag based on Pearson correlation coefficient. (D) Time lag based on Spearman correlation coefficient.

selection, partial least squares regression, stepwise regression, and elastic net regression were 68•54%, 79•32%, 89•19%, 79•32%, and 77•60%, respectively. Partial least squares regression has the best goodness-of-fit.

In this article, the correlation and hysteresis between more than 1000 social and news media and COVID-19 cases were analyzed and calculated. The results showed that compared with social media, news media had stronger average correlation, played a more important role in COVID-19 prediction, and was a data source that cannot be ignored. Using social and news media data, we proposed five different prediction models to predict the daily new confirmed cases in China, compared the five models, and selected partial least squares regression as the optimal model. This comprehensive model had excellent accuracy and low error and can effectively predict the daily new confirmed cases in China 3 days in advance based on social and news media data. In the future, our proposed model could be a powerful supplement to traditional methods of infectious disease surveillance.

Table 1
Performance parameters of five prediction models.

Model	Adjusted R^2	Cross-validation R^2	S	RSS	MSE	Number of predictive variables
Principal component analysis	73•42%	71•54%	0•646	29•986	0•395	1
Best subset selection	80•98%	80•21%	0•538	21•169	0•282	2
Partial least squares regression	93•75%	88•26%	NA	7•143	0•101	6
Stepwise regression	80•98%	80•21%	0•538	21•169	0•282	2
Elastic net regression	85•15%	81•75%	0•517	15•870	0•220	9

Note: 1. The final regression equation and various evaluation indexes of best subset selection and stepwise regression were the same.
2. NA represents the default value, and for model verification, partial least squares regression does not require calculation of the response degree *S*.

Declaration of Competing Interest

The authors declare that they have no competing interests.

Funding

This work was financially supported by grants from the China Mega-Project on Infectious Disease Prevention (No. 2017ZX10303401).

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.jinf.2022.01.009.

References

1. Lu Y, Zhang L. Social media WeChat infers the development trend of COVID-19. *J Infect* 2020;**81**(1):e82–3.
2. Larsen D, Dinero RE, Asiago-Reddy E, et al. A review of infectious disease surveillance to inform public health action against the novel coronavirus SARS-CoV-2. *SocArXiv* 2020. doi:10.31235/osf.io/uwdr6.
3. Ray EL, Wattanachit N, Niemi J, et al. Ensemble predictions of coronavirus disease 2019 (COVID-19) in the US. *MedRXiv* 2020. doi:10.1101/2020.08.19.20177493.
4. Zou D, Wang L, Xu P, Chen J, Zhang W, Gu Q. Epidemic model guided machine learning for COVID-19 predictions in the United States. *medRxiv* 2020. doi:10.1101/2020.05.24.20111989.
5. Yousefinaghani S, Dara R, Mubareka S, Sharif S. Prediction of COVID-19 waves using social media and Google search: a case study of the US and Canada. *Front Public Health* 2021;**9**:656635.
6. Shen C, Chen A, Luo C, Zhang J, Feng B, Liao W. Using reports of symptoms and diagnoses on social media to predict COVID-19 case counts in Mainland China: observational infoveillance study. *J Med Internet Res* 2020;**22**(5):e19421.
7. Panuganti BA, Jafari A, MacDonald B, DeConde AS. Predicting COVID-19 incidence using anosmia and other COVID-19 symptomatology: preliminary analysis using Google and Twitter. *Otolaryngol Head Neck Surg* 2020;**163**(3):491–7.
8. National Health Commission of the People's Republic of China, Epidemic situation notification. Available from: http://www.nhc.gov.cn/xcs/yqtb/list_gzbd.shtml (Accessed at 17 February 2020).

Mengxuan Lin[1]
Academy of Military Medical Sciences, Academy of Military Science of Chinese PLA, Beijing, China

Hui Chen[1]
Department of Infectious Disease Prevention and Control, Center for Disease Control and Prevention of Chinese People's Liberation Army, Beijing, China

Yuqi Wang[1]
Department of Information Management, Peking University, Beijing, China

Shaofu Qiu, Mingjuan Yang, Xinying Du
Department of Infectious Disease Prevention and Control, Center for Disease Control and Prevention of Chinese People's Liberation Army, Beijing, China

Tao Zheng*
Academy of Military Medical Sciences, Academy of Military Science of Chinese PLA, Beijing, China

Hongbin Song*, Ligui Wang*
Department of Infectious Disease Prevention and Control, Center for Disease Control and Prevention of Chinese People's Liberation Army, Beijing, China

*Corresponding authors.
E-mail addresses: zhengtao_66@163.com (T. Zheng), hongbinsong@263.net (H. Song), wangligui1983@126.com (L. Wang)
[1] These authors contributed equally to this work.