

Methods and Applications

Developing Machine Learning Prediction Model for Daily Influenza Reported Cases Using Multichannel Surveillance Data — A City, Hubei Province, China, 2023–2025

Xinyue Zhang¹; Xinyi Sang¹; Beibei Liu¹; Quanyu Wang¹; Xiuran Zuo²; Sheng Wei^{1,3,#}; Qi Wang^{1,#}

ABSTRACT

Introduction: Public health surveillance is crucial for decision-making. Given the significant threat of influenza to public health, developing predictive models using multichannel surveillance systems is imperative.

Methods: Data were collected from multichannel surveillance systems, including hospitals, search engines, and climatological and air pollutant surveillance systems, in a southern Chinese city from January 2023 to January 2025. Spearman's correlation analysis assessed the relationships between variables and reported influenza cases. Several machine learning models were used to predict trends in reported cases.

Results: Correlation analysis showed that all four surveillance systems were related to influenza, with 27 variables correlated with daily reported cases. The Long Short-Term Memory model, established based on variables with the highest lagged correlations (5-day to 7-day lag) through combined surveillance systems, outperformed other models for 5-day forecasts ($R^2=0.92$; mean absolute error=156.92; mean absolute percentage error=79.95%; root Mean Squared Error=292.33).

Conclusions: Data from various surveillance systems effectively track influenza epidemics. The model shows potential for infectious disease surveillance and epidemic preparedness.

Influenza, an acute respiratory infectious disease caused by the influenza virus, threatens global public health due to its high incidence, transmissibility, and severe complications (1). Effective surveillance is crucial for timely public health interventions.

Influenza spread is influenced by climatic, human migration, social media, and socioeconomic status (2). Integrating multisource data outbreak prediction

remains challenging.

Artificial intelligence (AI) has introduced new scope in disease prediction. Deep learning (DL), a machine learning (ML) subset, enables model optimization through self-supervised learning, showing superior prediction performance (3). These technologies show potential for disease prediction by capturing complex patterns (4).

This study explored the relationship between influenza cases and surveillance systems in a southern Chinese city, using AI techniques to establish prediction models for influenza epidemics, to refine monitoring strategies and inform public health responses.

METHODS

This study was conducted in a major southern Chinese city with 13 million residents in late 2023. Located at 114° 30' E and 30° 58' N, it has a subtropical monsoon humid climate with hot summers and cold winters. Influenza occurs more in winter and spring. Several hospitals serve as influenza surveillance sentinel to collect representative data systematically. To build the prediction model, data were collected from multichannel surveillance systems, including influenza cases, hospitals, meteorological and air pollutant surveillance systems, and search engine data (Supplementary Table S1, available at <https://weekly.chinacdc.cn/>).

Multi-Surveillance systems

Daily influenza reported cases were obtained from China Information System for Disease Control and Prevention (5). All case details were identified. After tallying daily case counts, missing data for 70 days (70/731, 9%) was identified. Linear interpolation was chosen for its ability to preserve temporal structure and provide reliable estimates for short-term missing values, making it preferable to mean imputation or deletion.

Daily fever clinic attendance and nine related symptoms, including runny nose, cough, sore throat, dyspnea, fever, headache, joint pain, myalgia, and fatigue, were collected from Electronic Medical Records (EMRs) of outpatient visits.

Baidu, a Chinese search engine, provides a search index for tracking keyword trends and user behavior. Baidu Search Index (BSI) data covering desktop and mobile queries were collected using influenza symptoms keywords (<https://index.baidu.com/> [accessed 2025-01-05]).

Air pollutant surveillance data, including concentrations of particulate matter with aerodynamic diameter $\leq 2.5 \mu\text{m}$ ($\text{PM}_{2.5}$), particulate matter with aerodynamic diameter $\leq 10.0 \mu\text{m}$ (PM_{10}), nitrogen dioxides (NO_2), sulfur dioxide (SO_2), carbon monoxide (CO), and ozone (O_3), were obtained from the city's Department of Ecology and Environment website.

Meteorological surveillance data were obtained from China Meteorological Data Sharing Service Center (<http://data.cma.cn/en> [accessed 2025-01-05]), including daily mean temperature (Tmean), mean relative humidity (RHmean), mean air pressure, mean wind speed, mean precipitation, and mean visibility range. The mean absolute humidity (AHmean) was calculated based on Tmean and RHmean as following formula (6):

$$AHmean = \{6.112 \times \exp[(17.67 \times Tmean) / (Tmean \pm 243.5)] \times RHmean \times 2.1674\} / (273.15 + Tmean)$$

Model Construction

Spearman's correlation coefficients were calculated for all variables for daily reported cases, as data showed strong non-parametric distributions. Each variable was normalized to 0–1 to facilitate model compatibility.

This study used ML methods to build prediction models: ensemble models (Random Forest and eXtreme Gradient Boosting), linear regression, and instance-based models (K-Nearest Neighbor). Two DL models, the Gated Recurrent Unit and Long Short-Term Memory (LSTM), were used.

In the model construction process, variables with $P < 0.05$ in correlation analysis were selected as model inputs. Lagged correlations evaluated relationships between each variable and reported cases from -7 to 0 days. For prediction on day t , variables with highest correlation coefficients from $t-7$ to t were selected.

For the DL models, grid search determined optimal

hyperparameter combinations. The selected hidden sizes were 32, 64, 128, and 256. The number of layers ranged from one to three in steps by one. Dropout values ranged from 0.3 to 0.5 step by 0.05. The learning rate values selected were 0.01, 0.05, 0.001, and 0.0005. Adam was selected as the optimizer.

The data from January 2023 to January 2025 showed two peaks in influenza cases and an upward trend in late 2024 (Supplementary Figure S1, available at <https://weekly.chinacdc.cn/>). The dataset was split in an 8:2 ratio into a training set (January 2023 to August 2024) and testing set (August 2024 to January 2025) for model training and validation. This ensured adequate training data while maintaining a predictive baseline for validation. The model performance was evaluated by comparing predicted and actual data using the coefficient of determination (R^2) (values close to 1 indicate better prediction), mean absolute error (MAE), mean absolute percentage error (MAPE), and root mean square error (RMSE) (values close to 0 indicate better prediction).

Model Comparison

To evaluate model validity and robustness of the selected dataset and constructed model, comparative analysis of the performances of the various models was conducted. Prediction models were developed using four surveillance systems: hospital (H), search engine (B), meteorological (M), and air pollutant (P). These models involved individual systems and their combinations: *Model H*, *Model B*, *Model M*, *Model P*, *Model H+B*, *Model H+M*, *Model H+P*, *Model B+M*, *Model B+P*, *Model M+P*, *Model H+B+M*, *Model H+B+P*, *Model H+M+P*, *Model B+M+P*, *Model All*, totaling 15 models.

Statistical Analysis

All the time-series data were smoothed. All statistical analyses in this study were performed using Python (version 3.12.0; Python Software Foundation, Wilmington, Delaware, USA) and TensorFlow (version 2.0.0; Google LLC, Mountain View, California, USA).

RESULTS

Description of Data from Surveillance Systems

From January 2023 to January 2025, the reported

influenza cases showed three distinct peaks (Supplementary Figure S1): the first in early 2023, second from December 2023 to January 2024, and third in late 2024. These peaks highlight the seasonal and fluctuating nature of influenza infections.

During monitoring, the number of fever clinic visits and nine symptoms showed trends similar to those of influenza cases but with discrepancies. Fever clinic visits increased in early 2023 and from October 2023 to January 2024, matching the first two peaks, but not the third in late 2024 (Supplementary Figure S2, available at <https://weekly.chinacdc.cn/>). The daily frequencies of nine influenza-related symptoms showed similar patterns (Supplementary Figure S3, available at <https://weekly.chinacdc.cn/>). Daily BSI trends, climatic factors, and air pollutants are detailed in Supplementary Figures S4–S6 (available at <https://weekly.chinacdc.cn/>).

Correlation of Surveillance Systems

Figure 1 shows the lag correlations between all surveillance data and the reported cases over the monitoring period. Most correlations were statistically significant ($P<0.001$). Symptoms like sore throat, cough, and myalgia showed strong positive correlations at shorter lags (0–3 days). Myalgia reached a correlation of 0.694 [95% confidence interval (CI): 0.654, 0.730] at Coef-0, sore throat was 0.634 (95% CI: 0.588, 0.675) at Coef-0, and cough was 0.566 (95% CI: 0.514, 0.613) at Coef-1. The meteorological factors and air pollutants were weakly correlated. Tmean had a correlation of -0.456 at Coef-7 and -0.425 at Coef-3, while O_3 had -0.337 at Coef-7 and -0.321 at Coef-5.

Prediction Model of the Influenza Cases

Prediction models for reported influenza cases were

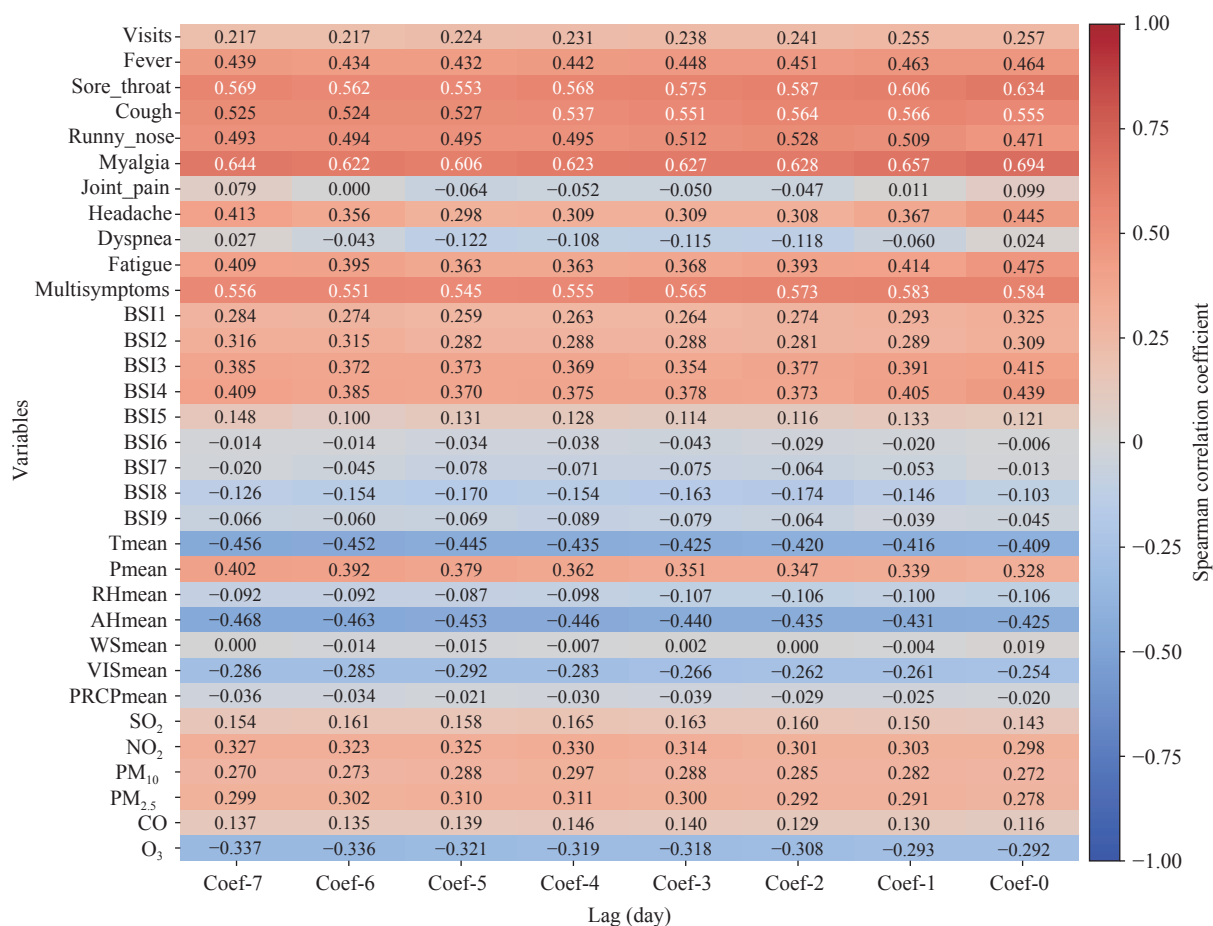


FIGURE 1. Spearman correlation analysis between different surveillance systems and reported cases with a 7-day lag before.

Abbreviation: Tmean=daily mean temperature; Pmean=daily mean air pressure; RHmean=daily mean relative humidity; AHmean=daily mean absolute humidity; WSmean=daily mean wind speed; VISmean=daily mean visibility; PRCPmean=daily mean precipitation.

developed using variables from four surveillance systems. For DL models, a temporal window was incorporated, and the LSTM model performed best with a 14-day window ($R^2=0.81$, MAE=170.59, MAPE=60.26%, and RMSE=453.23) (Supplementary Table S2, available at <https://weekly.chinacdc.cn/>). Prediction models were then built using various ML algorithms for predictions from 0 to 7 days ahead. The LSTM model, using variables with highest lagged correlations (5- to 7-day lag), showed strong prediction performance for 5 days forecasts ($R^2=0.92$; MAE=156.92; MAPE=79.95%; RMSE=292.33)

(Table 1).

The prediction performances of different surveillance systems combinations were compared. The model using all variables performed the best ($R^2=0.92$, MAE=156.92, MAPE=79.95%, and RMSE=292.33) (Figure 2). The models based on a single surveillance system performed poorly, and the best-performing model was *Model H* ($R^2=0.25$, MAE=428.57, MAPE=221.61%, and RMSE=896.41). Among the two system combinations, *Model H+M* was best ($R^2=0.73$, MAE=248.17, MAPE=125.64%, RMSE=533.21). For the three-system combinations, *Model*

TABLE 1. Comparison of different lag days of data between different models.

Model performance	Lag (day)							
	-7	-6	-5	-4	-3	-2	-1	0
MAE								
RF	337.87	338.28	341.30	334.08	312.13	300.46	288.00	242.29
XGBoost	287.21	278.68	306.39	296.40	256.80	270.32	262.13	194.60
LR	331.20	345.66	363.05	459.53	463.15	453.79	518.47	520.87
KNN	318.65	322.00	335.00	315.57	310.96	292.79	272.76	249.22
GRU	213.40	201.27	245.67	248.84	222.84	209.87	245.61	235.47
LSTM	200.40	229.21	156.92	238.00	244.77	258.08	214.40	170.59
MAPE (%)								
RF	123.67	126.36	133.84	134.04	129.53	127.40	124.04	118.06
XGBoost	123.83	113.94	126.43	129.99	106.45	117.50	100.04	91.91
LR	290.16	299.56	290.85	405.42	409.27	396.86	466.83	553.25
KNN	101.26	110.71	112.25	101.42	104.63	95.35	85.36	89.93
GRU	92.66	80.78	99.19	70.50	74.83	73.84	104.70	87.31
LSTM	82.45	114.17	79.95	124.37	132.32	110.60	61.55	60.26
RMSE								
RF	850.54	847.87	839.32	801.37	734.83	702.84	662.17	522.37
XGBoost	676.90	666.64	740.27	670.44	622.87	619.18	627.01	436.74
LR	563.26	575.52	607.19	665.98	659.59	653.02	699.16	601.48
KNN	835.86	824.83	840.50	789.46	792.10	753.85	721.15	572.74
GRU	416.80	505.86	611.00	646.34	553.33	509.26	574.51	581.57
LSTM	467.92	407.89	292.33	405.50	490.79	574.22	570.83	453.23
R^2								
RF	0.25	0.26	0.27	0.34	0.44	0.49	0.55	0.72
XGBoost	0.53	0.54	0.44	0.54	0.60	0.61	0.60	0.80
LR	0.67	0.66	0.62	0.54	0.55	0.56	0.50	0.63
KNN	0.28	0.30	0.27	0.36	0.35	0.41	0.46	0.66
GRU	0.84	0.76	0.65	0.61	0.71	0.76	0.69	0.68
LSTM	0.79	0.83	0.92	0.84	0.77	0.69	0.69	0.81

Abbreviation: MAE=mean absolute error; RF=Random Forest; XGBoost=eXtreme Gradient Boosting; SVM=Support Vector Machine; LR=Linear Regression; KNN=K-Nearest Neighbors; GRU=Gated Recurrent Unit; LSTM=Long Short-Term Memory; MAPE=mean absolute percentage error; RMSE=root mean square error; R^2 =coefficient of determination.

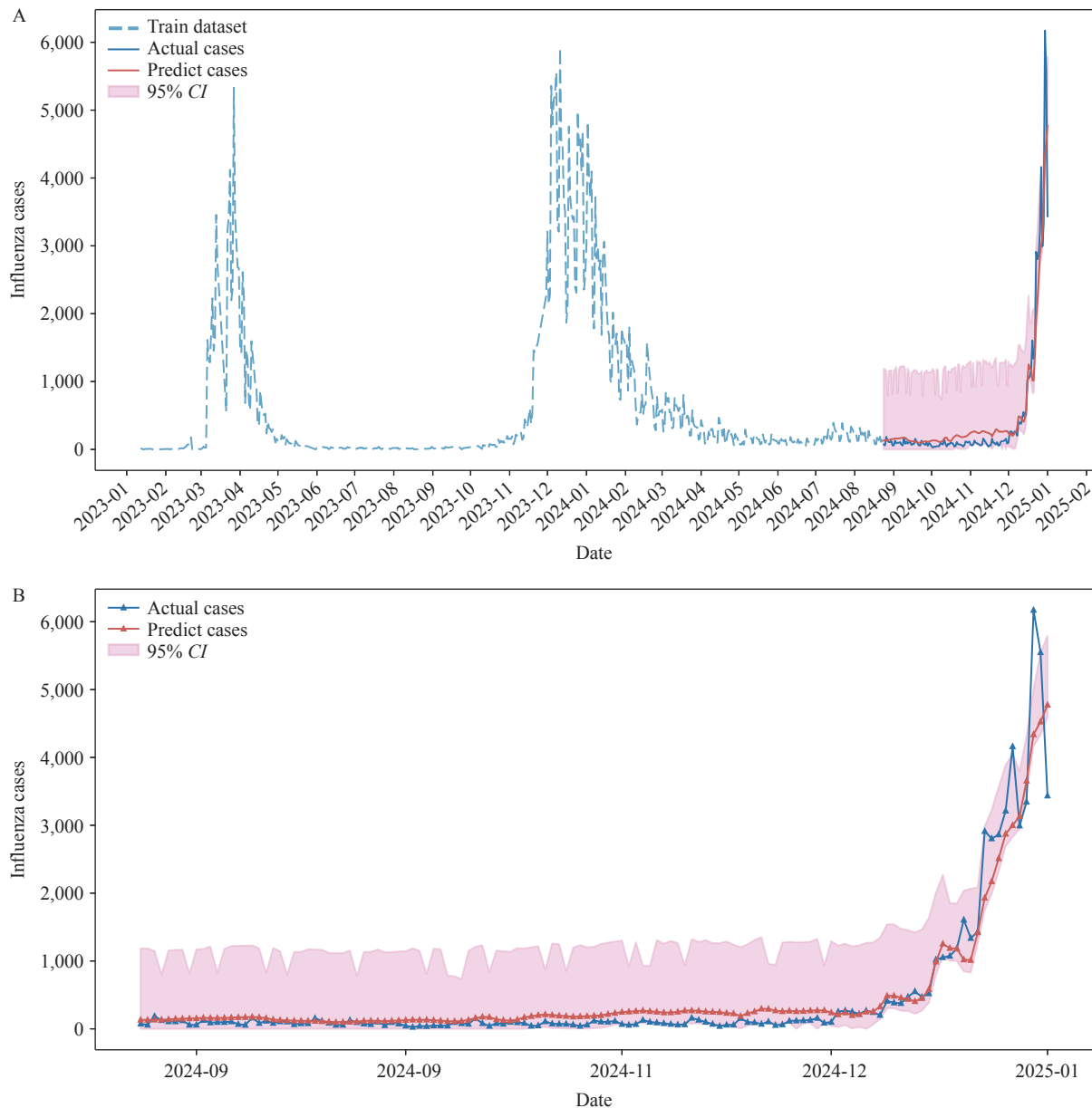


FIGURE 2. Prediction diagram for the LSTM model. (A) All monitoring period; (B) Test set period.

Abbreviation: LSTM=Long Short-Term Memory; CI=confidence interval.

$H+M+P$ ($R^2=0.84$, $MAE=201.06$, $MAPE=63.64\%$, $RMSE=418.51$) showed the highest prediction efficiency (Table 2).

DISCUSSION

Prediction models were developed to track the influenza epidemic in a southern Chinese city by integrating data from multichannel surveillance systems. The LSTM model combining all surveillance data demonstrated high prediction accuracy, with $R^2=0.92$, $MAE=156.92$, $MAPE=79.95\%$, and $RMSE=$

292.33, outperforming other models.

Compared to existing models, our approach shows improved prediction accuracy by integrating multiple data sources. Previous studies relied mainly on single-source data, such as clinical report (7) or Google Trends (8), limited by reporting delays, definition changes, and data errors (9–10). Our models used EMRs, social media, meteorological, and air pollutant data were designed to mitigate forecasting errors. The combined model performed best, highlighting the value of diverse data in assessing influenza trends. The early phase of the influenza outbreak highlighted the

TABLE 2. Comparison of different combinations of data.

Model performance of different combinations	MAE	MAPE (%)	RMSE	R ²
Model H	428.57	221.61	896.41	0.25
Model B	414.27	141.49	971.23	0.11
Model M	456.34	159.50	946.81	0.16
Model P	435.46	131.86	1000.15	0.06
Model H+B	370.35	168.43	793.79	0.41
Model H+M	248.17	125.64	533.21	0.73
Model H+P	286.82	105.25	670.63	0.58
Model B+M	275.89	102.85	651.88	0.60
Model B+P	317.25	109.17	703.38	0.54
Model M+P	254.46	102.41	548.00	0.72
Model H+B+M	215.49	108.40	480.24	0.78
Model H+B+P	197.02	52.24	522.32	0.74
Model H+M+P	201.06	63.64	418.51	0.84
Model B+M+P	267.35	118.07	550.24	0.72
Model All	156.92	79.95	292.33	0.92

Abbreviation: MAE=mean absolute error; MAPE=mean absolute percentage error; RMSE=root mean square error; R²=coefficient of determination.

inadequacy of confirmed case data for traditional surveillance (11). A multifaceted monitoring approach is essential to improve epidemic predictions.

Data quality and diversity significantly affect model performance. EMRs provide detailed clinical information. Fever clinic visits exhibited a weaker correlation ($r=0.257$), as fever clinic patients may have other diseases and some patient with influenza may not visit fever clinics. Symptoms like myalgia, cough, and sore throat showed stronger positive correlations with influenza cases at shorter lags, which underscore the importance of symptom-based surveillance (12). Social media data provide real-time insights into potential outbreaks, with search volumes for symptom keywords correlating with influenza cases (13). Studies indicate that combining Internet-based queries and climate data improves the accuracy and timeliness of infectious disease warning systems (14). Our findings show positive correlations between air pollutants (SO₂, NO₂, PM₁₀, PM_{2.5}, and CO) and negative correlation between O₃ and influenza cases, consistent with previous studies (15–16). This highlights the importance of integrating air pollutant data for accurate influenza forecasting.

The LSTM model demonstrated improved accuracy through multisource data. Hospital surveillance enhanced prediction performance, consistent with correlation results. This approach provides understanding of environmental factors, public health

interventions, and disease dynamics. Although a 5–7-day lag generally performed well, some combinations weakened due to flu's complex spread mechanisms involving virus survival and human behavior. Environmental variables can extend virus survival time, potentially causing delays between changes and observable influenza case increases. Our multisource surveillance data integrated clinical, laboratory, and syndromic monitoring systems, with reporting delays contributing to extended lag period. Despite similarities in respiratory disease spread factors, significant heterogeneity existed. Early prediction improves response strategies, resource allocation, and outbreak management.

This study has certain limitations. The absence of population mobility and vaccination rates may restrict the capacity of the model to capture influenza transmission dynamics. Data quality from less reliable sources may affect performance, and the lack of external validation limits generalizability. Seasonal variations may lead to dispersed data patterns and higher noise levels during non-epidemic seasons. Future research should incorporate vaccination data and explore additional data sources like behavioral patterns and environmental factors. External validation and detailed data preprocessing, such as smoothing or cross-validation, could enhance generalization.

This study demonstrates that multichannel data integration improves respiratory infectious disease

prediction accuracy and timeliness, with implications for public health responses. Ongoing research will refine these models for other health threats.

Conflicts of interest: No conflicts of interest.

Acknowledgments: The staff members at the Center for Disease Prevention and Control and Health Information Center for data verification.

Funding: Supported by the National Key Research and Development Program of China (grant no. 2022YFC2305103).

doi: 10.46234/ccdcw2025.234

Corresponding authors: Qi Wang, wangqi_tj@hust.edu.cn; Sheng Wei, weis@sustech.edu.cn.

¹ Department of Epidemiology and Biostatistics, School of Public Health, Tongji Medical College, Huazhong University of Science and Technology, Wuhan City, Hubei Province, China; ² Health Information Center of Wuhan, Wuhan City, Hubei Province, China; ³ School of Public Health and Emergency Management, Southern University of Science and Technology, Shenzhen City, Guangdong Province, China.

Copyright © 2025 by Chinese Center for Disease Control and Prevention. All content is distributed under a Creative Commons Attribution Non Commercial License 4.0 (CC BY-NC).

Submitted: September 19, 2025

Accepted: October 25, 2025

Issued: October 31, 2025

REFERENCES

- Iuliano AD, Roguski KM, Chang HH, Muscatello DJ, Palekar R, Tempia S, et al. Estimates of global seasonal influenza-associated respiratory mortality: a modelling study. *Lancet* 2018;391(10127):1285 – 300. [https://doi.org/10.1016/S0140-6736\(17\)33293-2](https://doi.org/10.1016/S0140-6736(17)33293-2).
- Zhang YZ, Bambrick H, Mengersen K, Tong SL, Hu WB. Using internet-based query and climate data to predict climate-sensitive infectious disease risks: a systematic review of epidemiological evidence. *Int J Biometeorol* 2021;65(12):2203 – 14. <https://doi.org/10.1007/s00484-021-02155-4>.
- Brownstein JS, Rader B, Astley CM, Tian HY. Advances in artificial intelligence for infectious-disease surveillance. *N Engl J Med* 2023;388(17):1597 – 607. <https://doi.org/10.1056/NEJMr2119215>.
- Kraemer MUG, Tsui JLH, Chang SY, Lytras S, Khurana MP, Vanderslott S, et al. Artificial intelligence for modelling infectious disease epidemics. *Nature* 2025;638(8051):623 – 35. <https://doi.org/10.1038/s41586-024-08564-w>.
- Dong YH, Wang LP, Burgner DP, Miller JE, Song Y, Ren X, et al. Infectious diseases in children and adolescents in China: analysis of national surveillance data from 2008 to 2017. *BMJ* 2020;369:m1043. <https://doi.org/10.1136/bmj.m1043>.
- Nottmeyer LN, Sera F. Influence of temperature, and of relative and absolute humidity on COVID-19 incidence in England - A multi-city time-series study. *Environ Res* 2021;196:110977. <https://doi.org/10.1016/j.envres.2021.110977>.
- Krymova E, Béjar B, Thanou D, Sun T, Manetti E, Lee G, et al. Trend estimation and short-term forecasting of COVID-19 cases and deaths worldwide. *Proc Natl Acad Sci USA* 2022;119(32):e2112656119. <https://doi.org/10.1073/pnas.2112656119>.
- Prasanth S, Singh U, Kumar A, Tikkiwal VA, Chong PHJ. Forecasting spread of COVID-19 using google trends: a hybrid GWO-deep learning approach. *Chaos Solitons Fractals* 2021;142:110336. <https://doi.org/10.1016/j.chaos.2020.110336>.
- Holmdahl I, Buckee C. Wrong but useful - what covid-19 epidemiologic models can and cannot tell US. *N Engl J Med* 2020;383(4):303 – 5. <https://doi.org/10.1056/NEJMp2016822>.
- Chin V, Samia NI, Marchant R, Rosen O, Ioannidis JPA, Tanner MA, et al. A case study in model failure? COVID-19 daily deaths and ICU bed utilisation predictions in New York state. *Eur J Epidemiol* 2020;35(8):733 – 42. <https://doi.org/10.1007/s10654-020-00669-6>.
- Morgan OW, Abdelmalik P, Perez-Gutierrez E, Fall IS, Kato M, Hamblion E, et al. How better pandemic and epidemic intelligence will prepare the world for future threats. *Nat Med* 2022;28(8):1526 – 8. <https://doi.org/10.1038/s41591-022-01900-5>.
- Uyeki TM, Hui DS, Zambon M, Wentworth DE, Monto AS. Influenza. *Lancet* 2022;400(10353):693 – 706. [https://doi.org/10.1016/S0140-6736\(22\)00982-5](https://doi.org/10.1016/S0140-6736(22)00982-5).
- Milunovich GJ, Williams GM, Clements AC, Hu WB. Internet-based surveillance systems for monitoring emerging infectious diseases. *Lancet Infect Dis* 2014;14(2):160 – 8. [https://doi.org/10.1016/S1473-3099\(13\)70244-5](https://doi.org/10.1016/S1473-3099(13)70244-5).
- Xu L, Zhou C, Luo ST, Chan DK, McLaws ML, Liang WN. Modernising infectious disease surveillance and an early-warning system: the need for China's action. *Lancet Reg Health West Pac* 2022;23:100485. <https://doi.org/10.1016/j.lanwpc.2022.100485>.
- Zhang R, Lai KY, Liu WH, Liu YH, Ma XW, Webster C, et al. Associations between short-term exposure to ambient air pollution and influenza: an individual-level case-crossover study in Guangzhou, China. *Environ Health Perspect* 2023;131(12):127009. <https://doi.org/10.1289/EHP12145>.
- Guo F, Zhang P, Do V, Runge J, Zhang K, Han ZS, et al. Ozone as an environmental driver of influenza. *Nat Commun* 2024;15(1):3763. <https://doi.org/10.1038/s41467-024-48199-z>.

SUPPLEMENTARY MATERIAL

Model Introduction

This study constructed a prediction model for daily incidence of reported influenza cases using four channels surveillance data sources: hospital surveillance (H), Baidu Search Index (B), meteorological surveillance (M), and air pollutant surveillance (P). Six machine learning (ML) models were applied, covering common single and ensemble learning methods. The single models included linear regression (LR) and K-nearest neighbors (KNN). The ensemble learning models comprised random forest (RF) and extreme gradient boosting (XGBoost). Deep learning models included gated recurrent unit (GRU) and long short-term memory (LSTM). Model inputs (x) were different data combinations from the four sources, with outputs (y) as daily reported case numbers. Data from January 2023 to January 2025, were used, with 80% for training and 20% for testing. The study aimed to identify the optimal model for predicting influenza reported cases by comparing the predictive performance of different models.

Linear regression (LR) is a supervised ML algorithm that learns from labeled datasets to map data points to an optimal linear function. LR performs well in capturing linear relationships in time series and can handle small-scale data effectively. For formulas, see Equation [1], where β represents the partial regression coefficients and represents the residuals.

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_n x_n + \quad (1)$$

K-Nearest Neighbors (KNN) is a versatile ML algorithm used for both classification and regression tasks. For regression, it predicts a value based on the average of the values of its k-nearest neighbors. The choice of k is critical as it balances the bias-variance trade-off. KNN is non-parametric, making no assumptions about the underlying data distribution, and is straightforward to implement.

Random Forest (RF) is an ensemble of tree-structured classifiers. Each tree predicts based on an independent and identically distributed random vector. In regression, it outputs numbers, trained on datasets from random vector distributions. By integrating multiple decision trees, RF reduces overfitting risk, handles nonlinear data and complex feature relationships, and is somewhat robust to missing data. For formulas, see Equation [2]. $l_{\text{new}}^{\text{RF}}$ denote the RF's prediction for a new sample, which is the aggregated output from all decision trees in the model. m represent the number of decision trees in the RF. The prediction from the j -th decision tree for the new sample is denoted as l_j^{tree}

$$l_{\text{new}}^{\text{RF}} = \frac{1}{m} \sum_{j=1}^m l_j^{\text{tree}} \quad (2)$$

XGBoost is an optimized distributed gradient-boosting library that implements ML algorithms in the Gradient Boosting framework. It's designed for efficiency, flexibility, and portability. XGBoost excels in performance, especially in nonlinear time-series problems. It can handle heterogeneous features and their complex interactions. The formula is shown in Equation [3]. $i^{(t)}$ represents the model's prediction for the t -th sample, reflecting the latest prediction after cumulative optimization up to iteration t . $\sum_{k=1}^{t-1} f_k(x_i)$ is the cumulative prediction from previous iterations up to $t-1$. $f_t(x_i)$ denotes the prediction from the newly added tree at the t -th iteration, which aims to correct the model's prediction errors and bring the results closer to the true targets.

$$i^{(t)} = \sum_{k=1}^{t-1} f_k(x_i) + f_t(x_i) \quad (3)$$

LSTM networks were developed to capture long-term dependencies in time-series data. An LSTM unit comprises three gated units and two states. The three gates are the forget gate (f_t), input gate (i_t), and output gate (o_t), which regulate the information flow from the previous timestep. The two states are the hidden state (h_t) and cell state (c_t). The hidden state h_t acts as the short-term memory, passing information to the next timestep and the output layer. The cell state c_t serves as the long-term memory, updated by the forget and input gates. Here's the structure in detail:

- Forget Gate Layer: Determines what information to discard from the cell state using a sigmoid function.
- Input Gate Layer: Updates the cell state with new information through a sigmoid function and a tanh function.
- Output Gate Layer: Decides the next hidden state using a sigmoid function and a tanh function.

For the mathematical representation, the LSTM equations are as follows:

$$f_t = \sigma(W_f[C_{t-1}, h_{t-1}, x_t] + b_f) \quad (4)$$

$$i_t = \sigma(W_i[C_{t-1}, h_{t-1}, x_t] + b_i) \quad (5)$$

$$o_t = \sigma(W_o[C_{t-1}, h_{t-1}, x_t] + b_o) \quad (6)$$

GRU is an LSTM variant. It combines the LSTM's forget and input gates into a single update gate and merges the cell state (c_t) with the hidden state (h_t), resulting in a simpler structure compared to the standard LSTM. This design reduces the number of parameters, enabling faster training or requiring less data for effective generalization.

Dataset Splitting

As shown in [Supplementary Figure S1](#), the total cases of daily incidence of reported influenza cases in the city from January 2023, to January 2025, had two full peaks in 2023–2025 and an upward trend at the end of 2024. For each model, the dataset was sequentially split into training and testing sets at an 8:2 ratio. This gave the training set enough data for model fitting and the testing set a starting trend for prediction.

Model Evaluation

Model assessment used regression metrics: R^2 , MAPE, MAE, and RMSE. R^2 indicates the percentage of variation in the target variable explained by the features. R^2 ranges from 0 to 1, with higher values indicating a better fit. MAPE measures the average percentage difference between predicted and actual values, reflecting the average deviation of predictions from true values. MAE is the average of absolute errors, providing a clear measure of prediction accuracy. RMSE, the square root of mean squared error, is widely used as it penalizes larger errors more than MAE. For MAPE, MAE, and RMSE, values closer to 0 indicate better model performance. The calculation equations for these metrics are presented below:

$$R^2 = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2} \quad (7)$$

$$MAPE = \frac{1}{n} \sum_{i=1}^n \left| \frac{y_i - \hat{y}_i}{y_i} \right| \times 100\% \quad (8)$$

$$MAE = \frac{1}{n} \sum_{i=1}^n |\hat{y}_i - y_i| \quad (9)$$

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2} \quad (10)$$

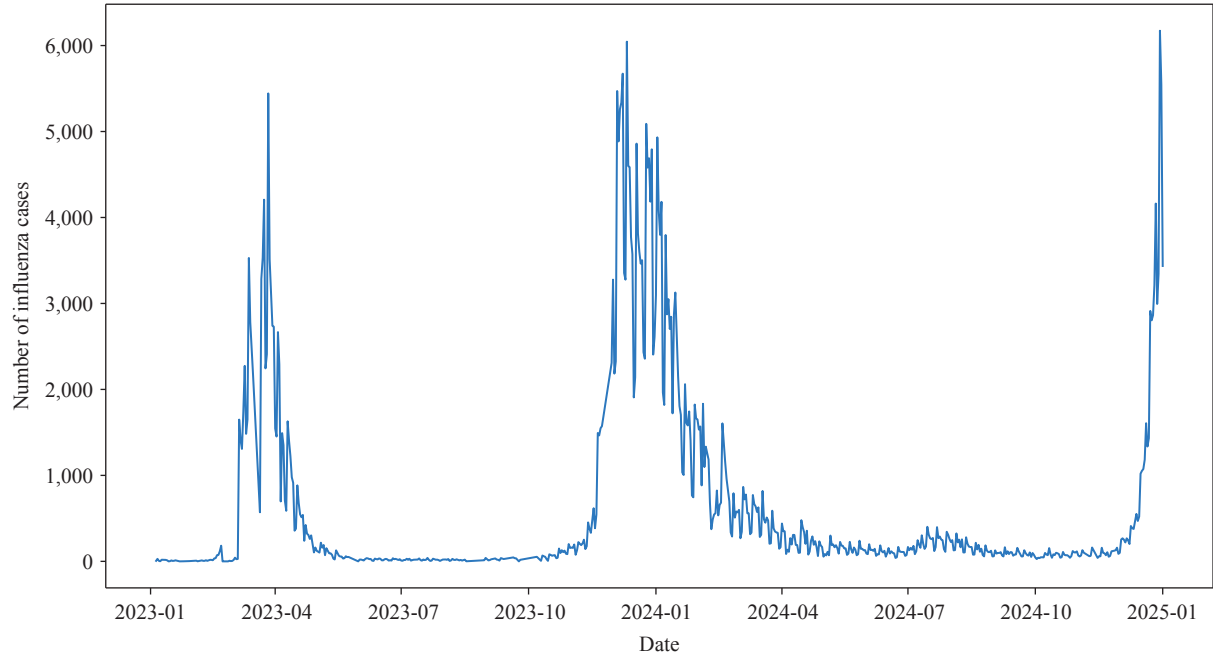
SUPPLEMENTARY TABLE S1. Different surveillance systems and dataset descriptions.

Multichannel surveillance data	Description	Timeline	Date range
Cases Report data			
Influenza_cases	Daily Reported Influenza Cases	1 day	Jan 5, 2023–Jan 1, 2025
Hospital Surveillance data			
visits	Fever Clinic Visits	1 day	Jan 5, 2023–Jan 1, 2025
Outpatients with respiratory symptoms	Daily Outpatients with respiratory symptoms		Jan 5, 2023–Jan 1, 2025
Fever	Outpatient chief complaint of fever	1 day	
Sore_throat	Outpatient chief complaint of sore throat	1 day	
Cough	Outpatient chief complaint of cough	1 day	
Runny_nose	Outpatient chief complaint of runny nose	1 day	
Myalgia	Outpatient chief complaint of myalgia	1 day	
Joint_pain	Outpatient chief complaint of joint pain	1 day	
Headache	Outpatient chief complaint of headache	1 day	
Dyspnea	Outpatient chief complaint of dyspnea	1 day	
Fatigue	Outpatient chief complaint of fatigue	1 day	
Multisymptoms	Outpatient chief complaint of two or more respiratory symptoms	1 day	
Search Engine Surveillance data			Jan 1, 2023–Jan 1, 2025
BSI1	Baidu search rank related with fever	1 day	
BSI2	Baidu search rank related with sore throat	1 day	
BSI3	Baidu search rank related with cough	1 day	
BSI4	Baidu search rank related with runny nose	1 day	
BSI5	Baidu search rank related with myalgia	1 day	
BSI6	Baidu search rank related with joint pain	1 day	
BSI7	Baidu search rank related with headache	1 day	
BSI8	Baidu search rank related with dyspnea	1 day	
BSI9	Baidu search rank related with fatigue	1 day	
Meteorological Surveillance data			Jan 1, 2023–Jan 1, 2025
Tmean	Daily average temperature	Every 3h per day	
Pmean	Daily average air pressure	Every 3h per day	
RHmean	Daily average relative humidity	Every 3h per day	
AHmean	Daily average absolute humidity	Every 3h per day	
WSmean	Daily average wind speed	Every 3h per day	
VISmean	Daily average visibility	Every 3h per day	
PRCPmean	Daily average precipitation	Every 3h per day	
Air pollutant Surveillance data			Jan 1, 2023–Jan 1, 2025
SO2	Daily average concentration of SO ₂	1 day	
NO2	Daily average concentration of NO ₂	1 day	
PM10	Daily average concentration of PM ₁₀	1 day	
PM2.5	Daily average concentration of PM _{2.5}	1 day	
CO	Daily average concentration of CO	1 day	
O3	Daily average concentration of O ₃	1 day	

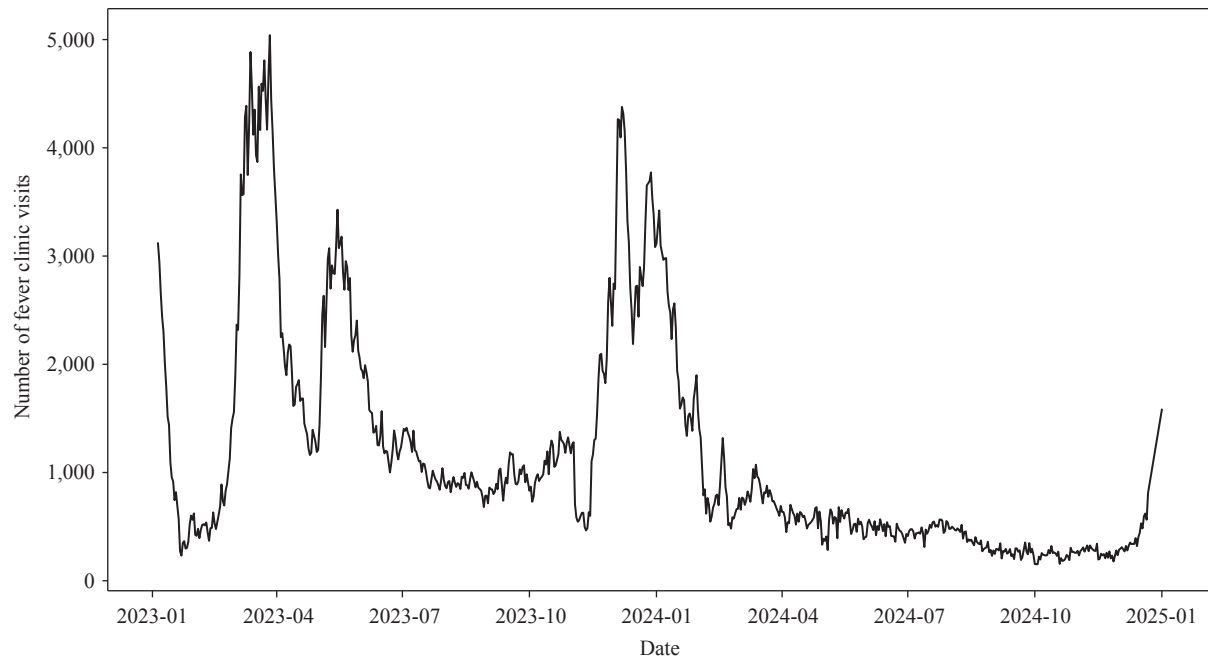
SUPPLEMENTARY TABLE S2. Comparison of different window sizes of data between deep learning models.

Model performance	Window size				
	1 day	3 days	5 days	7 days	14 days
MAE					
GRU	277.10	274.69	283.20	238.26	235.47
LSTM	227.66	248.37	240.24	230.47	170.59
MAPE (%)					
GRU	49.73	92.24	95.22	102.65	87.31
LSTM	84.55	67.29	59.42	60.49	60.26
RMSE					
GRU	767.65	688.28	707.67	561.88	581.57
LSTM	651.43	653.26	638.03	607.43	453.23
R ²					
GRU	0.40	0.52	0.50	0.68	0.68
LSTM	0.57	0.67	0.59	0.64	0.81

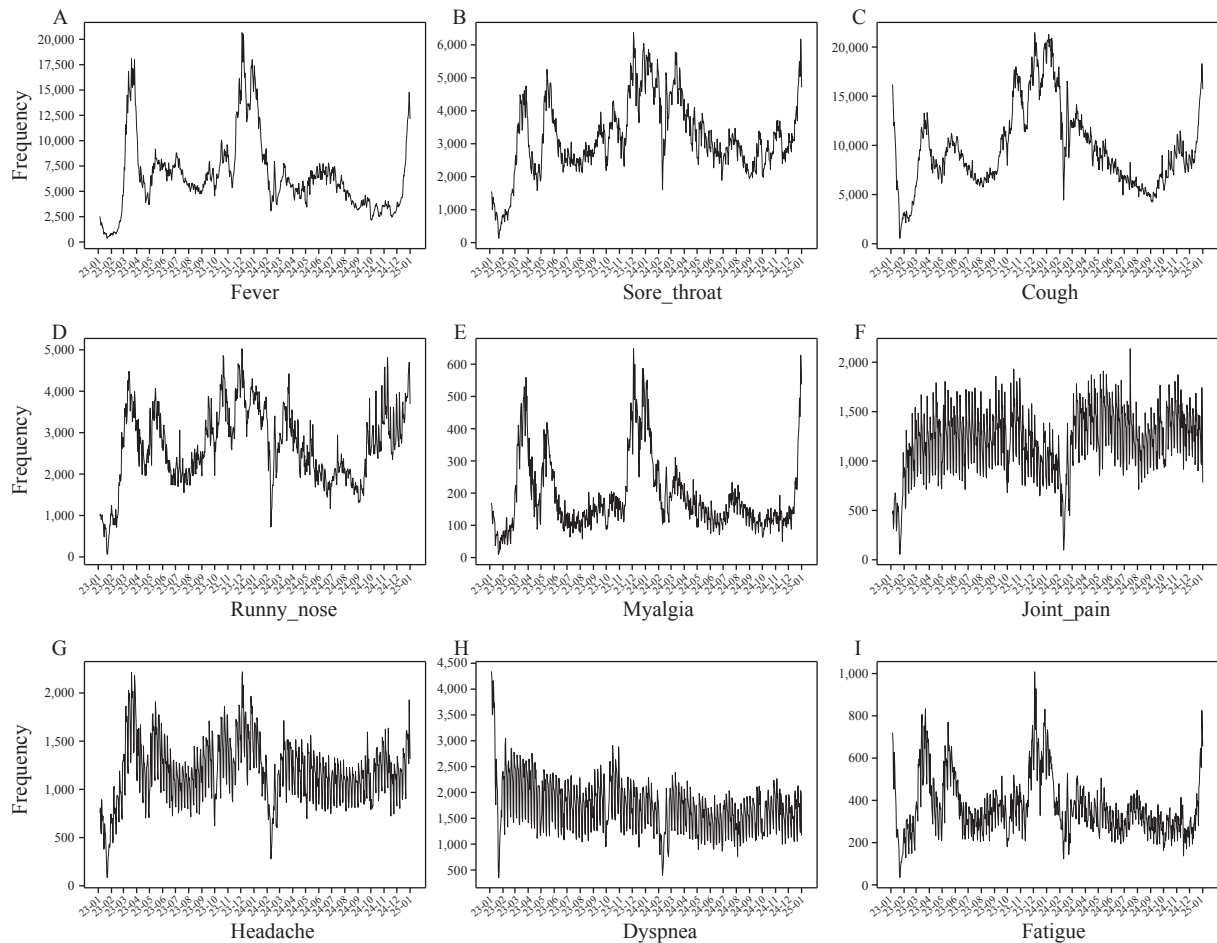
Abbreviation: MAE=mean absolute error; GRU=Gated Recurrent Unit; LSTM=Long Short-Term Memory; MAPE=mean absolute percentage error; RMSE=root mean square error; R²=coefficient of determination.



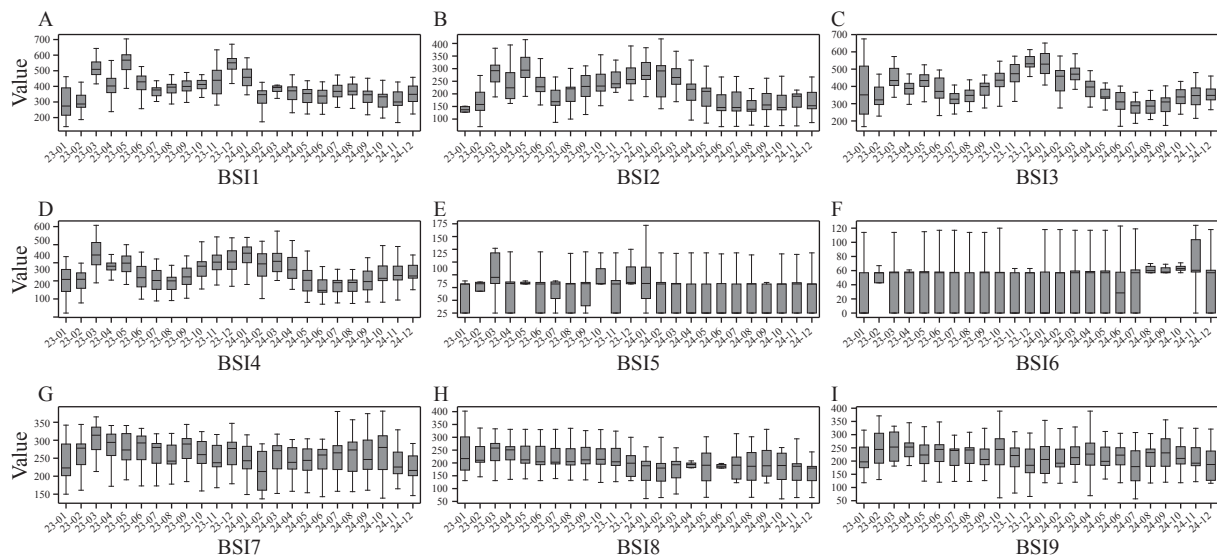
SUPPLEMENTARY FIGURE S1. The Reported Cases of Influenza every day between January 2023 and January 2025.



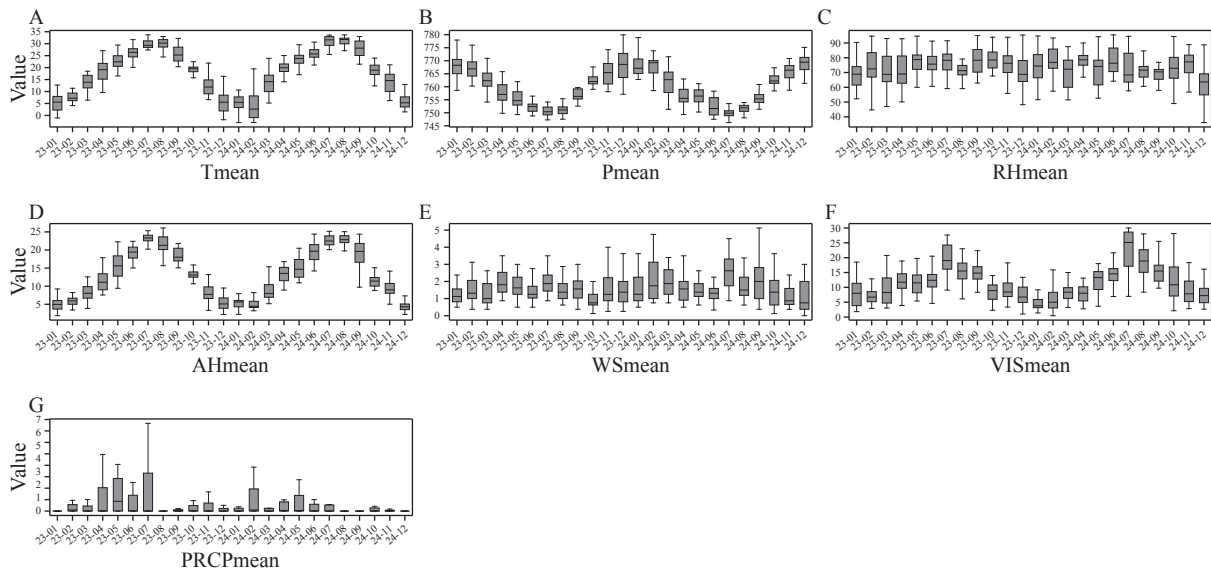
SUPPLEMENTARY FIGURE S2. The Fever Clinic Visits every day between January 2023 and January 2025.



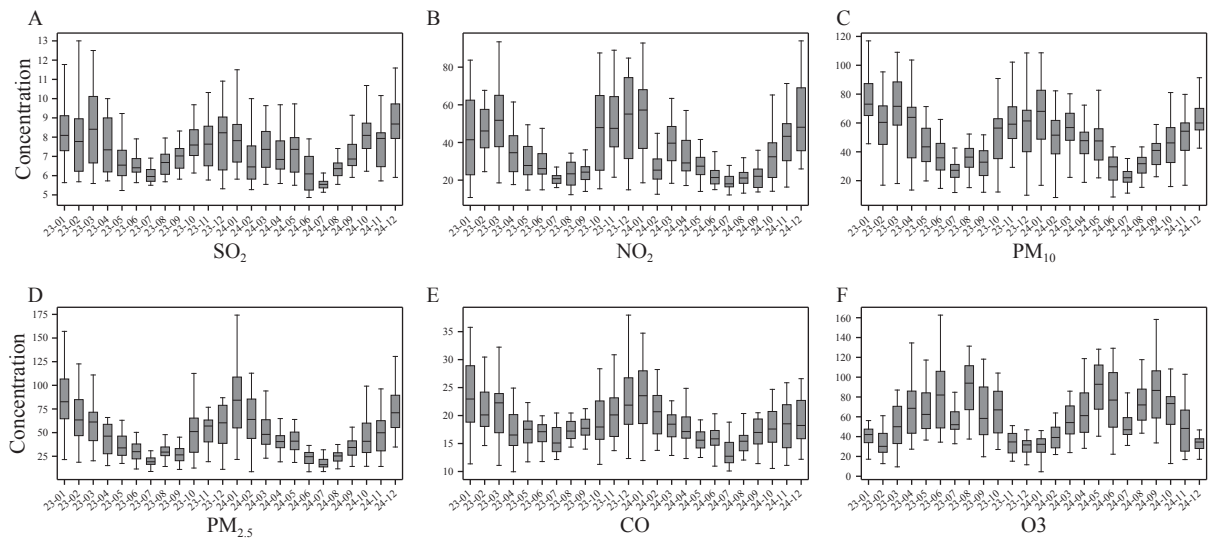
SUPPLEMENTARY FIGURE S3. The frequency of nine related symptoms every day between January 2023 and January 2025. (A) Daily frequency of fever; (B) Daily frequency of sore throat; (C) Daily frequency of cough; (D) Daily frequency of runny nose; (E) Daily frequency of myalgia; (F) Daily frequency of joint pain; (G) Daily frequency of headache; (H) Daily frequency of dyspnea; (I) Daily frequency of fatigue.



SUPPLEMENTARY FIGURE S4. The counts of nine related BSIs between January 2023 and January 2025. (A) Daily counts of BS1; (B) Daily counts of BS12; (C) Daily counts of BS13; (D) Daily counts of BS14; (E) Daily counts of BS15; (F) Daily counts of BS16; (G) Daily counts of BS17; (H) Daily counts of BS18; (I) Daily counts of BS19.



SUPPLEMENTARY FIGURE S5. The description of daily climate factors between January 2023 and January 2025. (A) Daily mean temperature; (B) Daily mean air pressure; (C) Daily mean relative humidity; (D) Daily mean absolute humidity; (E) Daily mean wind speed; (F) Daily mean visibility; (G) Daily mean precipitation.



SUPPLEMENTARY FIGURE S6. The description of daily air pollutants between January 2023 and January 2025. (A) Daily concentration of SO₂; (B) Daily concentration of NO₂; (C) Daily concentration of PM₁₀; (D) Daily concentration of PM_{2.5}; (E) Daily concentration of CO; (F) Daily concentration of O₃.