

Original Paper

Digital Prescription–Based Influenza Activity Forecasting in Jiangxi Province, China: Comparative Modeling Analysis Using Multisource Data

Rongrong Yang^{1*}, MSc; Zilu Xu^{2,3*}, MSc; Chi Zhang^{2,3}, MSc; Liping Qiu⁴, MSc; Jie Qian², PhD; Jie Liu⁵, PhD; Yonghai Dong⁶, MSc; Zhili Zeng⁷, MPH; Rui Shen², PhD; Luzhao Feng², PhD

¹Department of HIV/AIDS and Sexually Transmitted Disease Control and Prevention, Ganzhou Center for Disease Control and Prevention, Ganzhou, Jiangxi, China

²School of Population Medicine and Public Health, Chinese Academy of Medical Sciences & Peking Union Medical College, Key Laboratory of Pathogen Infection Prevention and Control (Peking Union Medical College), Ministry of Education, State Key Laboratory of Respiratory Health and Multimorbidity; Public Health Emergency Management Innovation Center of Beijing Higher Education Innovation Center for Philosophy and Social Sciences, Beijing, China

³School of Public Health, China Medical University, Shenyang, Liaoning, China

⁴Department of HIV/AIDS and Sexually Transmitted Disease Control and Prevention, Nanchang Center for Disease Control and Prevention, Nanchang, Jiangxi, China

⁵Scientific Research Affairs Section, Jiangxi Center for Disease Control and Prevention, Nanchang, Jiangxi, China

⁶Department of Surveillance Forecasting and Emergency Response, Jiangxi Center for Disease Control and Prevention, Nanchang, Jiangxi, China

⁷Department of Infectious and Endemic Disease Control and Prevention, Jiangxi Center for Disease Control and Prevention, Nanchang, Jiangxi, China

*these authors contributed equally

Corresponding Author:

Luzhao Feng, PhD

School of Population Medicine and Public Health

Chinese Academy of Medical Sciences & Peking Union Medical College, Key Laboratory of Pathogen Infection Prevention and Control (Peking Union Medical College), Ministry of Education, State Key Laboratory of Respiratory Health and Multimorbidity; Public Health Emergency Management Innovation Center of Beijing Higher Education Innovation Center for Philosophy and Social Sciences

No. 9 Dongdang Santiao, Dongcheng District

Beijing 100730

China

Phone: 86 10 65120716

Email: fengluzhao@cams.cn

Abstract

Background: Traditional laboratory-confirmed influenza surveillance involves a 1- to 2-week reporting delay and captures only patients who have sought care and received a diagnosis, limiting early warning. Digital prescription data have been shown to signal influenza activity early, but their predictive value for forecasting remains unclear.

Objective: We aimed to evaluate the predictive value of digital prescription data for influenza forecasting by assessing lead times, contributions across forecasting horizons, and added value when combined with other real-time sources.

Methods: Using daily data for Jiangxi Province, China (2022–2024), we characterized temporal relationships for 13 multi-source indicators using prewhitened cross-correlation and performed Granger-based predictor screening restricted to the 2022 to 2023 development period. Incidence was then forecast 1 to 14 days ahead through expanding-window rolling-origin validation trained on 2022 to 2023 data and tested on 2024 data. Six models (autoregressive integrated moving average [ARIMA], seasonal ARIMA [SARIMA], Prophet, least absolute shrinkage and selection operator [LASSO], random forest, and explainable boosting machine) plus a persistence baseline were compared across 3 exogenous-input scenarios (no external input, digital prescriptions alone, and selected sources). Each indicator was also forecast from its own history and compared with observed incidence.

Results: After prewhitening, the digital prescription rate led influenza incidence by 14 days, whereas the online search index lagged incidence by 1 day, indicating near-synchronous tracking. In the development period Granger analysis, digital prescription rate, online search index, and temperature showed significant predictor-to-incidence associations after false discovery rate (FDR) correction (FDR-adjusted $P < .001$, $P = .003$, and $P = .008$, respectively), whereas nitrogen dioxide showed a marginal association (FDR-adjusted $P = .11$) and was additionally included as an environmental covariate. At 1 day, performance was similar across models and input scenarios ($R^2 = 0.925$ – 0.961 ; persistence $R^2 = 0.948$), but differences emerged at longer horizons. SARIMA performed best with the digital prescription rate as the sole exogenous input, and LASSO performed best with 4-source input, both reaching a mean R^2 of 0.766 and root mean squared error (RMSE) of 0.240 across the 14 horizons. At 14 days, single-source SARIMA achieved an R^2 of 0.510 (RMSE 0.318), and multisource LASSO achieved an R^2 of 0.551 (RMSE 0.305), a gain of 0.064 in R^2 and a 6.2% reduction in RMSE over LASSO without exogenous input. Forecasts of the digital prescription rate showed the closest temporal consistency with observed influenza incidence (mean $r = 0.897$ vs 0.866 for the online search index).

Conclusions: Digital prescriptions provide an early signal of provincial influenza activity, preceding routine reported incidence by approximately 2 weeks in this setting. Multisource input can further improve forecasting when paired with models capable of selecting informative signals, but indiscriminate integration does not necessarily help. This early signal may support earlier preparedness, including antiviral planning and health care resource allocation.

J Med Internet Res 2026;28:e96131; doi: [10.2196/96131](https://doi.org/10.2196/96131)

Keywords: influenza; digital prescriptions; early warning; forecasting; disease surveillance

Introduction

Influenza remains a major global public health challenge, causing 3 to 5 million severe cases and 290,000 to 650,000 respiratory deaths annually [1]. Accurate and timely forecasting is important for guiding public health preparedness, resource allocation, and early interventions [2]. However, traditional laboratory-confirmed sentinel surveillance involves a delay of 1 to 2 weeks between specimen collection and the availability of confirmed results in weekly reports [3]. Although notifiable disease reporting in China provides timelier daily case counts, reported incidence still reflects only cases that have completed symptom onset, facility-based care seeking, and clinical diagnosis and is weighted toward patients with more severe symptoms, whereas individuals with milder or early-stage illness may instead turn to online health services. Reported incidence may therefore lag underlying epidemic dynamics, limiting early warning.

In response to this limitation, digital data streams have emerged as complementary tools for influenza surveillance [4]. However, these sources capture different aspects of influenza-related behaviors and transmission processes and may have distinct limitations. Search trends provide timely information on health-related interests [5] but may reflect media-driven attention and other factors unrelated to disease activity [6], resulting in potential misalignment with epidemiological patterns [6]. Over-the-counter medication sales capture self-medication behavior [7], but they are not specific to influenza and may also reflect treatment for other respiratory conditions [8]. Mobility signals provide indirect information on population movement and transmission risk but require additional modeling assumptions [9]. Conventional pharmacy prescription data contain clinically authorized medication information; however, they typically depend on offline care pathways in which patients must visit health care facilities before obtaining medication, potentially

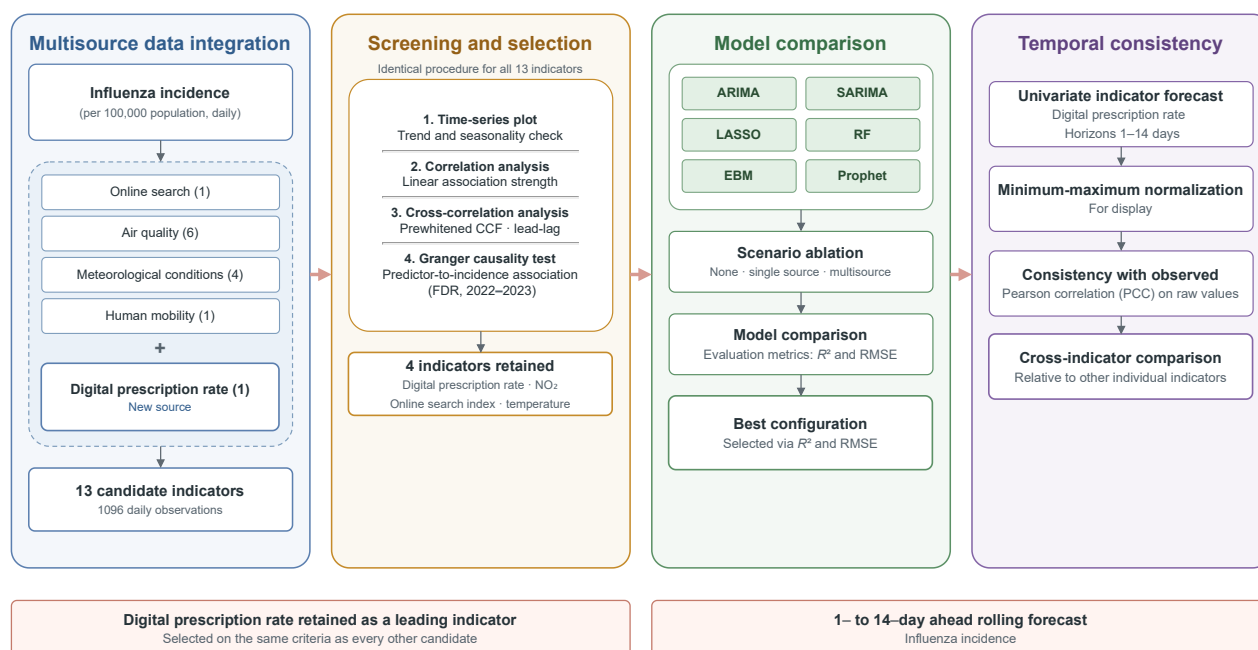
introducing delays and selecting individuals who complete the full care-seeking process [10,11]. These limitations highlight the need for timely digital health data sources that capture clinician-mediated treatment-seeking behavior.

Digital prescription data generated through online health platforms provide a potential source for addressing these limitations [12,13]. Through online consultation, physician assessment, and prescription authorization, antiviral transactions represent treatment-seeking behavior recorded close to the time of clinical decision-making [13]. Unlike search activity, which primarily reflects information-seeking behavior and may be influenced by media attention or other external factors [6], digital prescription data reflect a subsequent health care action. In China, the expansion of digital health platforms and on-demand delivery systems for prescription antivirals, including oseltamivir and baloxavir marboxil, has generated transaction-level data that are available within 24 hours [14]. Previous studies have shown that digital prescription data capture population-level influenza dynamics and provide complementary information for influenza surveillance [14]. However, their value as an exogenous predictor for forecasting influenza incidence remains insufficiently quantified, including their temporal lead and incremental forecasting contribution and the added value of combining them with other real-time sources.

In this study, we evaluated digital prescription rate as an exogenous signal for provincial influenza forecasting in Jiangxi Province, China (Figure 1). We screened candidate indicators using a uniform time-series analysis framework, assessed their temporal relationships with influenza incidence, and compared forecasting performance under different exogenous-input scenarios across multiple models. We further compared the temporal consistency of the digital prescription rate relative to other individual indicators. Specifically, we aimed to determine (1) the lead time of digital prescription rate relative to influenza incidence, (2) whether digital prescription rate improves forecasting

across different horizons, and (3) whether additional real-time sources provide incremental value under different modeling strategies.

Figure 1. Study framework. The analysis proceeded in 4 stages. Thirteen candidate indicators from 5 data categories were assessed against daily influenza incidence. Lead-lag relationships were characterized over the full study period, whereas Granger screening was restricted to the 2022 to 2023 development period; 4 indicators were retained for forecasting. Six models were compared across 3 exogenous-input scenarios at horizons of 1 to 14 days. Temporal consistency between individual indicator forecasts and influenza incidence was also assessed. ARIMA: autoregressive integrated moving average; CCF: cross-correlation function; EBM: explainable boosting machine; FDR: false discovery rate; LASSO: least absolute shrinkage and selection operator; NO₂: nitrogen dioxide; PCC: Pearson correlation coefficient; RF: random forest; RMSE: root mean squared error; SARIMA: seasonal ARIMA.



Methods

Study Design and Data Sources

We carried out a comparative predictive modeling study spanning January 1, 2022, to December 31, 2024, integrating 6 daily data streams from Jiangxi Province (digital prescriptions, influenza surveillance, online search index, meteorological conditions, air quality, and human mobility) covering 1096 consecutive calendar days. To protect privacy, all data were aggregated at the provincial level, with all individual identifiers removed.

Digital prescription data were obtained from Meituan Healthcare, an online health platform that provides digital consultation and medication delivery services in China [14]. In this study, digital prescription data comprised transaction-level records for prescription-only antiviral medications, including oseltamivir and baloxavir marboxil. Digital prescription rate was defined as the daily number of unique purchasers of these medications per 100,000 residents in Jiangxi Province. Multiple transactions made by the same purchaser on the same day were counted once. Digital prescription date therefore represent the number of individuals obtaining prescribed antiviral medications rather than the

number of prescriptions issued, filled, or delivered. Because these digital prescriptions were generated through an online health care pathway involving patient consultation, physician assessment, and prescription authorization, they reflect clinician-mediated treatment-seeking behavior but do not represent laboratory-confirmed influenza diagnoses [15].

Influenza incidence data were obtained from the China Information System for Disease Control and Prevention and expressed as the number of reported influenza cases per 100,000 population. As influenza is a statutory class C notifiable disease, cases are reported through the direct network reporting system within 24 hours of clinical diagnosis; daily counts indexed by report date are therefore available in near real time and reflect the values available at each forecast origin. Online search activity was obtained from the Baidu Index platform [16]. On the basis of our previous study [17], 8 Chinese-language search terms were predefined: “influenza,” “influenza A,” “influenza B,” “flu B,” “fever,” “sore throat,” “oseltamivir,” and “tamiflu.” Daily province-level search indexes for these terms were summed to construct a composite online search index representing overall influenza-related search activity. Human mobility data were obtained from the Baidu migration platform [18]. Daily migration rates were calculated following established

methodology as follows: migration rate = $(M \times OD \times \text{coefficient}/P) \times 10^5$, where M is the migration scale index, OD is the directional flow proportion between provinces, and P is the population of the origin province; the coefficient is as described previously [19]. Missing values in the mobility series were imputed using dynamic harmonic regression with autoregressive integrated moving average (ARIMA) errors [20]. Meteorological variables, including temperature, relative humidity, wind speed, and precipitation, were obtained from the China Meteorological Data Service Center [21]. Air quality indicators, including PM_{2.5}, PM₁₀, carbon monoxide (CO), nitrogen dioxide (NO₂), ozone (O₃), and sulfur dioxide (SO₂), were obtained from the China National Environmental Monitoring Centre [22]. All variables were aggregated at the provincial day level and aligned by calendar date.

Indicator Screening and Selection

Thirteen candidate indicators from 5 categories, including digital prescription data, online search activity, human mobility, meteorological variables, and air quality indicators, were evaluated. Pearson correlation coefficients were first calculated to characterize concurrent associations between each indicator and influenza incidence. Raw cross-correlation functions were then calculated to describe unadjusted lead-lag relationships. To reduce correlations driven by autocorrelation and shared temporal patterns, weekly and annual seasonal patterns were first removed using Fourier terms, after which prewhitened cross-correlation analysis was also performed using the classic Box-Jenkins procedure [23]. For each indicator, an ARIMA model was fitted to the seasonally adjusted indicator series, and the same filter was applied to the seasonally adjusted incidence series before cross-correlations were calculated between the resulting residuals. Both raw and prewhitened cross-correlation functions were evaluated over lags ranging from -21 to +21 days, with negative lags signaling that the indicator preceded influenza incidence.

Granger tests were performed on the seasonally adjusted series to assess the temporal predictive relationships between each indicator and influenza incidence [24]. These analyses were restricted to the 2022 to 2023 development period to preserve the independence of the held-out 2024 test period. For each indicator, the lag order was selected using the Akaike information criterion, and each test examined whether past values of an indicator improved the prediction of subsequent influenza incidence beyond the information contained in the incidence series itself. P values from the concurrent correlation analyses and Granger tests were separately adjusted for multiple comparisons using the Benjamini-Hochberg false discovery rate (FDR) procedure. Granger associations were interpreted as predictive precedence rather than biological or mechanistic causation. Forecasting scenarios were constructed by integrating significant or marginal FDR-adjusted Granger evidence with prior epidemiological evidence. For indicators with marginal Granger evidence that were included in the multisource scenario based on prior epidemiological evidence, sensitivity analyses were conducted to assess their contribution to forecasting performance.

Forecasting Framework and Analysis Methods

Influenza incidence was forecast at horizons of 1 to 14 days using an expanding-window rolling-origin validation framework. The initial models were trained using data from January 1, 2022, to December 31, 2023, and were evaluated during the held-out period from January 1, 2024, to December 31, 2024. Details for each model are provided in Table S1 in [Multimedia Appendix 1](#). Predictor lags and other model inputs were constructed exclusively from information available at or before the forecast origin to prevent future information leakage.

ARIMA and seasonal ARIMA (SARIMA) were refitted at every forecast origin, whereas Prophet, least absolute shrinkage and selection operator (LASSO), random forest (RF), and explainable boosting machine (EBM) were refitted when a Page-Hinkley test detected drift in forecast errors or after a prespecified maximum interval. Three exogenous-input scenarios were compared: an autoregressive scenario without external indicators, a prescription-only scenario incorporating digital prescription rate as the sole exogenous predictor, and a multisource scenario incorporating digital prescription rate together with the other external indicators retained during screening. Exogenous predictors were aligned using a prespecified 14-day lag, which was applied consistently across models and input scenarios. This lag corresponds to the maximum forecast horizon, ensuring that all exogenous inputs were observed at or before each forecast origin for every 1- to 14-day ahead forecast and avoiding predictor-specific lag selection based on the observed data. Six forecasting approaches representing different statistical and machine learning strategies were evaluated: ARIMA, SARIMA, LASSO regression, Prophet, RF, and EBM. ARIMA and SARIMA represented classic time-series approaches, whereas Prophet used an additive time-series structure with linear growth and Fourier-based seasonal components. LASSO evaluated the linear contributions of multiple external predictors with variable selection, EBM modeled nonlinear additive effects, and RF allowed for both nonlinear relationships and interactions. Model-specific feature construction, preprocessing procedures, model updating strategies, hyperparameter settings, tuning strategies, and software packages are described in Table S1 in [Multimedia Appendix 1](#).

To evaluate the temporal consistency between individual indicators and influenza activity across forecast horizons, each retained indicator was also modeled separately using a univariate SARIMA model under the same expanding-window rolling-origin framework. Forecast indicator trajectories were compared with observed influenza incidence at the corresponding 1- to 14-day horizons using Pearson correlation coefficients. This analysis quantified the temporal consistency between each indicator's forecast trajectory and subsequent influenza incidence.

Model Evaluation

Forecasting performance was assessed using the coefficient of determination (R^2) and root mean squared error (RMSE), calculated separately for each forecasting horizon across the target dates in the held-out period. Within each exogenous-input scenario, all models were evaluated on the same target dates and horizons. The persistence baseline carried the most recently observed influenza incidence value forward across each forecast horizon. Empirical coverage of the 95% prediction intervals was also evaluated. Analyses used Python (version 3.12; Python Software Foundation) and R (version 4.5.2; R Foundation for Statistical Computing).

Ethical Considerations

This study was approved by the institutional ethics committee of the Chinese Academy of Medical Sciences and Peking Union Medical College (approval number CAMS&PUMC-IEC-2025-141). This study used only secondary data aggregated to the provincial level, from which all individual identifiers had been removed.

Results

Data Overview

The dataset comprised 1096 consecutive daily observations from Jiangxi Province, China, covering January 1, 2022, to December 31, 2024. The dataset was structured as a province-level daily time series, with influenza incidence as the outcome. The 13 candidate indicators were digital prescription rate, online search index, human migration rate, temperature, relative humidity, wind speed, precipitation, PM_{2.5}, PM₁₀, CO, NO₂, O₃, and SO₂. The mean influenza incidence was 0.64 (SD 1.09) cases per 100,000 population. Before imputation, the mobility series was missing on 3% ($n=33$) of days; no other series had missing values. After imputation, all series were complete for subsequent analyses. Variable characteristics are reported in Table S2 in [Multimedia Appendix 1](#).

Indicator Screening and Temporal Relationships

Digital prescription rate showed a clear temporal lead while maintaining a strong concurrent association with influenza

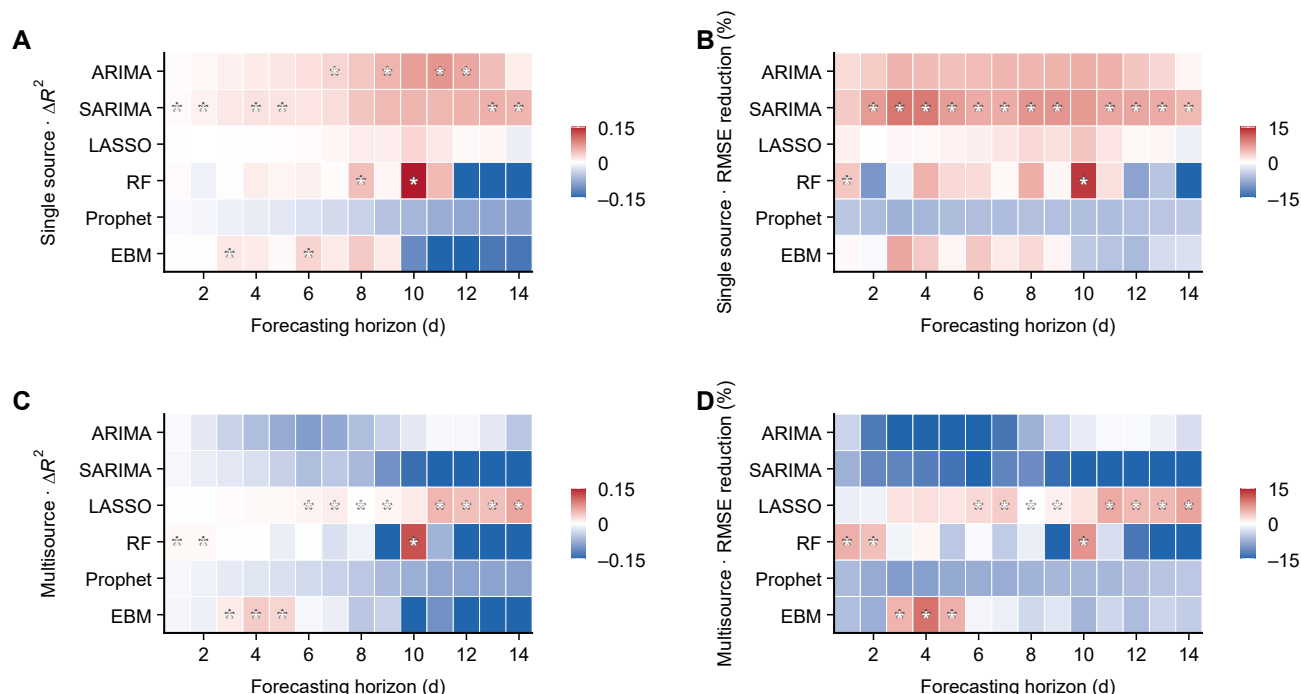
incidence. The online search index and digital prescription rate showed the strongest concurrent correlations with the incidence rate ($r=0.80$ and $r=0.75$, respectively; Table S2 and Figure S1 in [Multimedia Appendix 1](#)). The raw cross-correlation for the digital prescription rate was broad across the tested lags. After prewhitening, the digital prescription rate showed a distinct peak at a 14-day lead ($r=0.19$). The online search index had a slightly higher prewhitened peak ($r=0.21$) but lagged the reported incidence by 1 day, indicating near-synchronous tracking rather than a temporal lead (Table S3, Figure S2, and Figure S3 in [Multimedia Appendix 1](#)).

The 2022 to 2023 development period Granger analysis identified significant predictor-to-incidence associations for digital prescription rate (FDR-adjusted $P<.001$), online search index (FDR-adjusted $P=.003$), and temperature (FDR-adjusted $P=.008$; Table S3 and Figure S4 in [Multimedia Appendix 1](#)). NO₂ showed a marginal predictor-to-incidence association (FDR-adjusted $P=.11$). Given this marginal association and prior epidemiological evidence linking ambient NO₂ to influenza activity [25,26], NO₂ was additionally included as an environmental predictor in the multisource scenario.

Forecasting Performance and Added Value of Exogenous Inputs

The 2 leading exogenous-input configurations had the same rounded average performance and showed better point estimates than both the best no-exogenous model and the persistence baseline. Models were ranked within each input scenario by mean R^2 across the 14 forecasting horizons. The best single-source configuration was SARIMA with digital prescription rate, whereas the best multisource configuration was LASSO with a 4-source input including digital prescription rate, online search index, temperature, and NO₂. Both configurations had a mean R^2 of 0.766 and a mean RMSE of 0.240 cases per 100,000 population. By comparison, the best no-exogenous configuration, LASSO, had a mean R^2 of 0.748 and a mean RMSE of 0.248, whereas the persistence baseline performed less well, with corresponding values of 0.642 and 0.291 (Figure 2; Table S4 in [Multimedia Appendix 1](#)).

Figure 2. Incremental value of exogenous data over the no-exogenous baseline across forecasting horizons of 1 to 14 days. Panels A and B use a single exogenous source (digital prescription rate), whereas panels C and D use 4 sources (digital prescription rate, nitrogen dioxide, online search index, and temperature). Panels A and C show the absolute gain in R^2 relative to each model's own no-exogenous baseline, whereas panels B and D show the corresponding percentage reduction in root mean squared error (RMSE). Red indicates improved performance, blue indicates degraded performance, and white indicates no change; an asterisk marks the best-performing model at each horizon, not statistical significance. ARIMA: autoregressive integrated moving average; LASSO: least absolute shrinkage and selection operator; SARIMA: seasonal ARIMA.



The contribution of external inputs became clearer at longer forecasting horizons, but the pattern differed by model. At 1 day, R^2 ranged from 0.925 to 0.961 across models and input scenarios, and the persistence baseline achieved an R^2 of 0.948. As the horizon increased, the digital prescription rate alone was most useful for SARIMA: the mean R^2 increased from 0.733 to 0.766, and at 14 days, R^2 increased from 0.458 to 0.510, whereas RMSE decreased from 0.334 to 0.318. Multisource input mainly benefited LASSO. Although the 3 LASSO configurations performed similarly at 1 day ($R^2=0.951-0.954$), multisource LASSO achieved an R^2 of 0.551 and an RMSE of 0.305 at 14 days compared with 0.487 and 0.325 without exogenous inputs. This represented an increase in R^2 of 0.064 and a 6.2% reduction in RMSE. To assess whether this improvement depended on the inclusion of NO_2 , we performed a sensitivity analysis excluding NO_2 . Mean R^2 decreased from 0.766 to 0.752, and 14-day R^2 decreased from 0.551 to 0.499 (Table S4 in [Multimedia Appendix 1](#)), indicating that NO_2

contributed additional predictive information to LASSO in this setting. The contribution of external inputs remained model dependent. ARIMA showed a modest improvement with digital prescription rate alone, whereas external inputs generally reduced performance for Prophet, RF, and EBM; multisource input also reduced SARIMA performance (Table S4 in [Multimedia Appendix 1](#)).

Because the main gains were concentrated in single-source SARIMA and multisource LASSO, forecast trajectories and uncertainty were examined for these 2 configurations. Both tracked the timing of the 2024 influenza wave more closely at shorter horizons, whereas predicted peaks became more attenuated and the 95% prediction intervals widened as the horizon increased. Across the 14 horizons, the empirical coverage of the 95% prediction intervals was 99.2% for single-source SARIMA and 92.8% for multisource LASSO (Figures 3 and 4).

Figure 3. Predicted vs observed influenza incidence from the seasonal autoregressive integrated moving average model with a single exogenous source (digital prescription rate) at forecasting horizons of 1 to 14 days. Each panel (A-N) shows 1 horizon: the blue line represents observed influenza incidence, the red line represents the model prediction, and the shaded band represents the 95% prediction interval. The vertical dashed line marks the start of the held-out 2024 test period.

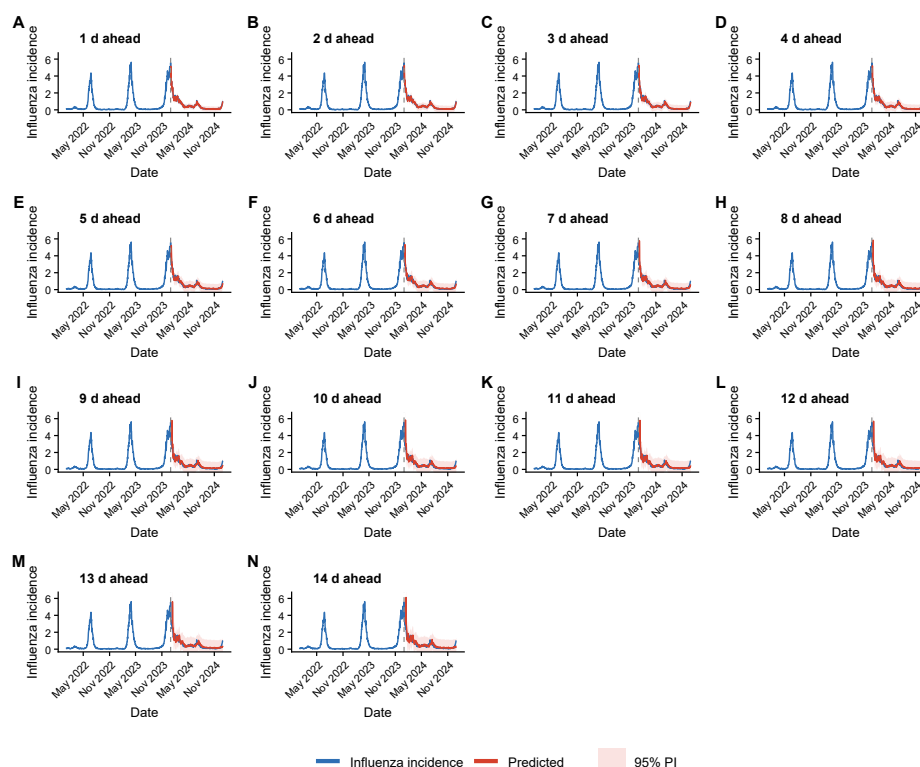
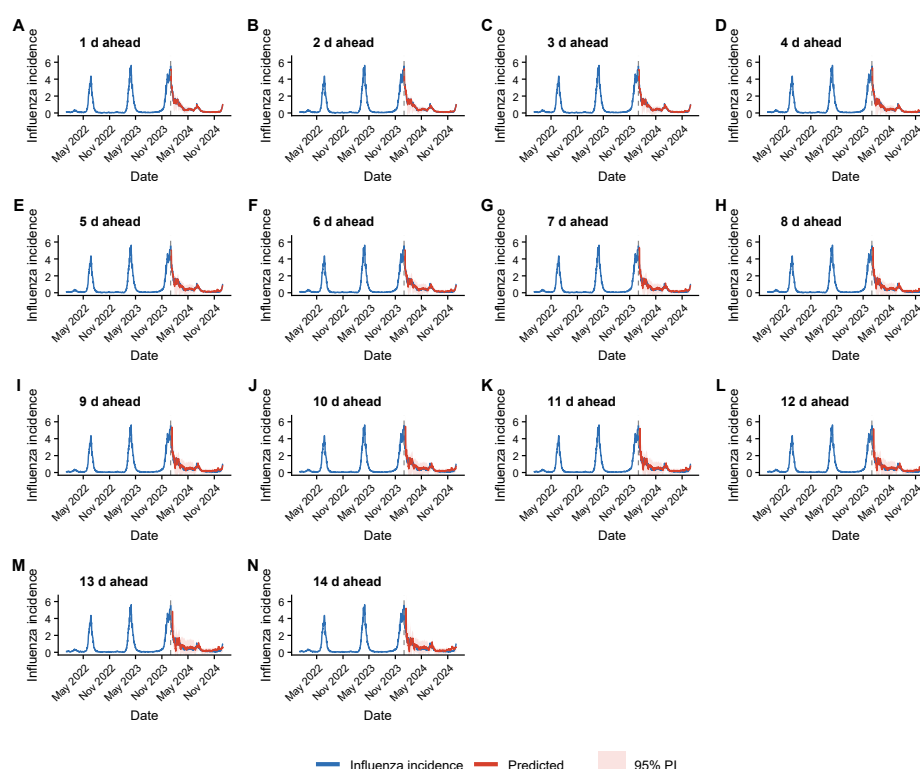


Figure 4. Predicted vs observed influenza incidence from the least absolute shrinkage and selection operator model with 4 exogenous sources (digital prescription rate, nitrogen dioxide, online search index, and temperature) at forecasting horizons of 1 to 14 days. Each panel (A-N) shows 1 horizon. The blue line represents observed influenza incidence, the red line represents the model prediction, and the shaded band represents the 95% prediction interval. The vertical dashed line marks the start of the held-out 2024 test period.

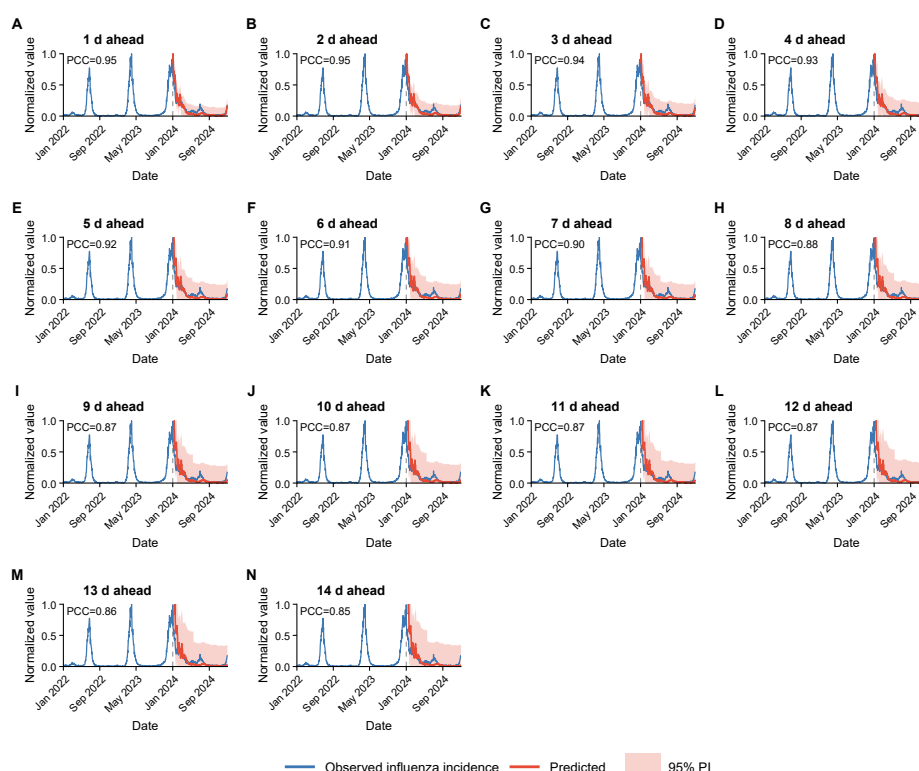


Temporal Consistency Between Indicator-Based Forecasts and Influenza Incidence

Among the 4 indicators included in forecasting, forecasts of the digital prescription rate showed the closest and most consistent temporal agreement with subsequent influenza incidence. When each retained indicator was forecast based on its own history, the digital prescription rate showed the closest agreement with influenza incidence at every horizon (Figure 5). The correlation decreased from $r=0.947$ at 1

day to $r=0.854$ at 14 days, with a mean of 0.897 across the 14 horizons. The online search index ranked second, decreasing from $r=0.897$ to $r=0.832$, with a mean of 0.866. However, its near-synchronous cross-correlation indicated concurrent tracking rather than an earlier signal. Because the indicator and incidence series share seasonal structure, these correlations should be interpreted as temporal agreement and may partly reflect shared seasonality. Temperature and NO₂ showed weaker agreement across all horizons (Figures S5-S7 in Multimedia Appendix 1).

Figure 5. Temporal consistency between the forecast digital prescription rate and observed influenza incidence across horizons of 1 to 14 days. Each panel (A-N) shows 1 horizon. The red line represents the digital prescription rate forecast from a univariate seasonal autoregressive integrated moving average model, with the 95% prediction interval shaded; the blue line represents observed influenza incidence. The vertical dashed line marks the start of the 2024 out-of-sample period; both series were minimum-to-maximum normalized to (0, 1) for display, and the Pearson correlation coefficient annotated in each panel was computed on the raw, nonnormalized values.



Discussion

Principal Findings

Digital prescription data provided an early external signal of influenza activity in Jiangxi Province. After prewhitening, it led influenza incidence by approximately 2 weeks and showed a significant predictor-to-incidence Granger association in the 2022 to 2023 development period analysis. Forecasts of the digital prescription rate also showed the closest temporal consistency with subsequent influenza incidence across the 1- to 14-day horizons. These findings support digital prescription data as a potentially useful exogenous signal for influenza forecasting in this setting.

The value of external data depended on both the model and the forecasting horizon. Differences were small at 1 day,

when recent incidence already provided a strong baseline. Incorporating digital prescription rate improved SARIMA performance, whereas the advantage of multisource input for LASSO became clearer at longer horizons. The best single-source and multisource configurations had the same rounded average performance, but adding several external sources did not consistently improve the other models. This suggests that multisource data are most useful when the model can retain informative signals and limit noise. These comparisons were conducted within a common rolling-origin framework using the same held-out dates and horizons and only information available at each forecast origin.

Digital Prescription Data as an Early Signal

Digital prescription data differ from both online behavior and routine disease surveillance. Search activity mainly captures interest or concern [6], whereas digital prescriptions follow treatment seeking, clinical assessment, prescription authorization, and medication purchase [14]. Although these prescriptions do not necessarily represent laboratory-confirmed influenza diagnoses, they are closer to treatment demand than search activity or over-the-counter medication sales [14, 27]. This may help explain why the digital prescription rate remained significantly associated with influenza incidence while also showing a clearer temporal lead.

The observed lead may reflect differences in when the two data streams are generated. Digital prescription data are generated when patients obtain prescribed antiviral medications, whereas reported influenza incidence reflects cases after health care seeking, clinical assessment, and reporting through the surveillance system. This interval should be interpreted as an empirical temporal relationship rather than a fixed biological lead or evidence that each digital prescription corresponded to a reported influenza case.

Model-Dependent Value of External Inputs

The forecasting results favored a relatively simple use of the prescription signal. Digital prescription rate alone improved SARIMA performance and produced the same rounded mean performance as the best multisource configuration. A single clinically mediated data stream may therefore provide useful predictive information without the complexity of maintaining several external sources with different update schedules and data-generating processes.

The value of multisource input became more apparent at longer horizons. At short horizons, recent incidence already contained much of the information needed for prediction. As this autoregressive signal weakened, digital prescription rate and other external data sources, including online search index, temperature, and NO₂, provided additional information related to treatment demand, epidemic conditions, and environmental context. The forecasting task included predictors with overlapping information across data sources and lag structures. By selecting informative variables and reducing the contribution of redundant predictors, LASSO may have provided a more stable and parsimonious model in this setting [28]. Although RF and EBM can capture nonlinear relationships, their additional flexibility did not result in improved performance in our dataset. These results indicate that the value of multisource input depends on how the forecasting model handles redundant or weak signals.

Operational Implications for Layered Surveillance

These findings support a layered approach to influenza monitoring. Digital prescription data may provide an early

indication of changing treatment demand, multisource models may add value at longer forecasting horizons, and routine surveillance remains necessary to confirm epidemic activity. Previous work has similarly shown that digital and traditional data streams can provide complementary information for influenza surveillance [29].

An additional warning interval of approximately 2 weeks could give health services more time to assess rising antiviral demand and plan medication supplies, staffing, and resource allocation [30]. Prediction intervals widened with the forecasting horizon, but empirical coverage remained high for the two leading configurations. These results describe forecasting performance rather than the effects of implementing the system in practice. Prospective studies are therefore needed to determine whether an earlier signal improves decisions concerning antiviral distribution, staffing, public health alerts, and other preparedness measures.

Limitations

This study has some limitations. The analysis covered one province and used aggregated data, which may mask local variation and limit generalizability. Users of digital health platforms may also differ from the wider population [31]. Differences in digital health platform penetration, health care access, and health-seeking behavior across regions may influence prescription patterns and limit the direct transfer of the observed lead to other settings. The study period included changes in health behavior during and after the COVID-19 pandemic [32], and evaluation in a single held-out year cannot establish temporal stability.

In addition, linked symptom-onset dates, laboratory results, platform penetration rates, medication supply data, and promotional information were unavailable. Changes in physician prescribing practices, antiviral availability, public awareness, or platform use could affect digital prescription rate independently of influenza activity. These factors limit interpretation of the estimated lead and make it difficult to separate epidemiological changes from changes in platform use. The lead interval may vary across influenza seasons and surveillance settings because treatment-seeking behavior and reporting delays are not constant.

Conclusions

Digital prescription data provided an early signal of provincial influenza activity, preceding routine reported incidence by approximately 2 weeks in this setting. Multisource input further improved forecasting when paired with a model capable of selecting informative signals, but adding external sources indiscriminately did not consistently improve performance. Incorporating this early signal may support earlier preparedness, including antiviral planning and health care resource allocation. Further validation across regions and influenza seasons is needed before routine implementation.

Acknowledgments

The authors would like to thank all individuals who contributed to the conduct and completion of this study. The authors declare that no generative AI tool or large language model was used in any portion of this work, including the study design, data analysis, figure generation, or writing and editing of the manuscript. All content was produced by the authors.

Funding

This study was supported by Prevention and Control of Emerging and Major Infectious Diseases–National Science and Technology Major Project (grant number 2025ZD01901900), the National Key Research and Development Program of China (grant number 2024YFC2311500), the Non-profit Central Research Institute Fund of Chinese Academy of Medical Sciences (grant number 2021-RC330-002), and the Strategic Talent Cultivation Project for Middle-Aged and Young Professionals in Public Health of Jiangxi Province. The funding sources had no role in the design and conduct of this study; collection, management, analysis, and interpretation of the data; preparation, review, and approval of the manuscript; or decision to submit it for publication.

Data Availability

The aggregated provincial-level data supporting the findings of this study are available from the corresponding author on reasonable request and subject to institutional review and the relevant data providers' permission.

Authors' Contributions

RY and ZX contributed to data analysis. RY, ZX, CZ, LQ, JQ, JL, YD, and ZZ contributed to data collection, data management, and data cleaning. RS and LF were responsible for the study design, methodology, and funding acquisition. The first draft of the manuscript was written by RY and ZX. All authors (RY, ZX, CZ, LQ, JQ, JL, YD, ZZ, RS, and LF) commented on previous versions of the manuscript. All authors have reviewed and approved the final manuscript.

Rui Shen and Luzhao Feng are co-corresponding authors. Rui Shen can be contacted at shenrui0817@163.com.

Conflicts of Interest

None declared.

Multimedia Appendix 1

Supplementary tables and figures on model implementation, candidate indicator screening and temporal relationships, forecasting performance and sensitivity analyses, and temporal consistency with influenza incidence.

[[DOCX File \(Microsoft Word File\), 1153 KB-Multimedia Appendix 1](#)]

References

1. Influenza (seasonal). World Health Organization. 2025. URL: [https://www.who.int/news-room/fact-sheets/detail/influenza-\(seasonal\)](https://www.who.int/news-room/fact-sheets/detail/influenza-(seasonal)) [Accessed 2026-01-25]
2. Reich NG, McGowan CJ, Yamana TK, et al. Accuracy of real-time multi-model ensemble forecasts for seasonal influenza in the U.S. *PLoS Comput Biol*. Nov 2019;15(11):e1007486. [doi: [10.1371/journal.pcbi.1007486](https://doi.org/10.1371/journal.pcbi.1007486)] [Medline: [31756193](https://pubmed.ncbi.nlm.nih.gov/31756193/)]
3. Mathis SM, Webber AE, León TM, et al. Evaluation of FluSight influenza forecasting in the 2021-22 and 2022-23 seasons with a new target laboratory-confirmed influenza hospitalizations. *Nat Commun*. Jul 26, 2024;15(1):6289. [doi: [10.1038/s41467-024-50601-9](https://doi.org/10.1038/s41467-024-50601-9)] [Medline: [39060259](https://pubmed.ncbi.nlm.nih.gov/39060259/)]
4. Salathé M, Bengtsson L, Bodnar TJ, et al. Digital epidemiology. *PLoS Comput Biol*. 2012;8(7):e1002616. [doi: [10.1371/journal.pcbi.1002616](https://doi.org/10.1371/journal.pcbi.1002616)] [Medline: [22844241](https://pubmed.ncbi.nlm.nih.gov/22844241/)]
5. Ginsberg J, Mohebbi MH, Patel RS, Brammer L, Smolinski MS, Brilliant L. Detecting influenza epidemics using search engine query data. *Nature*. Feb 19, 2009;457(7232):1012-1014. [doi: [10.1038/nature07634](https://doi.org/10.1038/nature07634)] [Medline: [19020500](https://pubmed.ncbi.nlm.nih.gov/19020500/)]
6. Lazer D, Kennedy R, King G, Vespignani A. Big data. The parable of Google Flu: traps in big data analysis. *Science*. Mar 14, 2014;343(6176):1203-1205. [doi: [10.1126/science.1248506](https://doi.org/10.1126/science.1248506)] [Medline: [24626916](https://pubmed.ncbi.nlm.nih.gov/24626916/)]
7. Li J, Wu C, Liu K, et al. Monitoring OTC drug sales for early detection of respiratory infectious disease outbreaks. *Front Public Health*. 2025;13:1661753. [doi: [10.3389/fpubh.2025.1661753](https://doi.org/10.3389/fpubh.2025.1661753)] [Medline: [41458427](https://pubmed.ncbi.nlm.nih.gov/41458427/)]
8. Oliveira JF, Cerqueira-Silva T, Brito PAN, et al. Anticipating influenza-like illness outbreaks via syndromic surveillance using over-the-counter drug sales and primary health care data. *NPJ Digit Public Health*. 2026;1(1):10. [doi: [10.1038/s44482-026-00014-w](https://doi.org/10.1038/s44482-026-00014-w)]
9. Chin T, Johansson MA, Chowdhury A, et al. Bias in mobility datasets drives divergence in modeled outbreak dynamics. *Commun Med (Lond)*. Jan 7, 2025;5(1):8. [doi: [10.1038/s43856-024-00714-5](https://doi.org/10.1038/s43856-024-00714-5)] [Medline: [39774250](https://pubmed.ncbi.nlm.nih.gov/39774250/)]
10. Sunamura Y, Lu XN, Hayashi Y, Saito M, Suzuki T. Prediction of influenza outbreaks using pharmacy sales data. *Total Qual Sci*. 2020;5(3):111-121. [doi: [10.17929/tqs.5.111](https://doi.org/10.17929/tqs.5.111)]

11. Adams K, Garg S, Tartof SY, et al. Patterns in prescribing and dispensing influenza antivirals among adults with influenza presenting to urgent care and emergency department settings: VISION Network, 2023-2024. *Clin Infect Dis*. Nov 6, 2025;81(4):e172-e183. [doi: [10.1093/cid/ciaf178](https://doi.org/10.1093/cid/ciaf178)] [Medline: [40184184](https://pubmed.ncbi.nlm.nih.gov/40184184/)]
12. Sun Y, Dai L, Shan Y, et al. Pathogen characteristics of respiratory infections in the season after the COVID-19 pandemic between August and December 2023: evidence from direct-to-consumer testing-based surveillance in Guangzhou and Beijing, China. *Int J Infect Dis*. Oct 2024;147:107195. [doi: [10.1016/j.ijid.2024.107195](https://doi.org/10.1016/j.ijid.2024.107195)] [Medline: [39074739](https://pubmed.ncbi.nlm.nih.gov/39074739/)]
13. Shen R, Fang C, Wang Z, et al. Digital prescription surveillance reveals seasonal patterns and environmental triggers of allergic disease activity across China. *Allergy*. May 2026;81(5):1863-1866. [doi: [10.1111/all.70203](https://doi.org/10.1111/all.70203)] [Medline: [41432218](https://pubmed.ncbi.nlm.nih.gov/41432218/)]
14. Shen R, Xu X, Yang L, et al. Real-time digital prescriptions unlock influenza dynamics: evidence from 21 million transactions. *NPJ Digit Med*. Mar 4, 2026;9(1):315. [doi: [10.1038/s41746-026-02513-9](https://doi.org/10.1038/s41746-026-02513-9)] [Medline: [41775925](https://pubmed.ncbi.nlm.nih.gov/41775925/)]
15. Measures for the supervision and administration of online drug sales [Article in Chinese]. Ministry of Justice of the People's Republic of China. 2022. URL: https://www.moj.gov.cn/pub/sfbgw/flfggz/flfggzbmzg/202305/t20230509_478415.html [Accessed 2026-07-25]
16. Baidu Index. URL: <https://index.baidu.com/v2/index.html> [Accessed 2026-07-25]
17. Shen R, Xu X, Li Y, et al. Enhancing spatiotemporal influenza prediction in China: a multi-output least absolute shrinkage and selection operator machine learning model integrating web-based search data. *Intell Med*. Nov 2025;5(4):310-317. [doi: [10.1016/j.imed.2025.02.003](https://doi.org/10.1016/j.imed.2025.02.003)]
18. Nationwide population migration details-Baidu Map migration big data platform. Baidu Maps. URL: <https://qianxi.baidu.com> [Accessed 2026-07-25]
19. Wang C, Yan J. An inversion of the constitution of the Baidu Migration Scale Index. *J Univ Electron Sci Technol China*. 2021;50:616-626. URL: <https://www.semanticscholar.org/paper/An-Inversion-of-the-Constitution-of-the-Baidu-Scale-Wang-Yan/abdfb6bcab2c4275e3280fcac3a845c0e1072ed4> [Accessed 2026-09-02]
20. Hyndman RJ, Athanasopoulos G. Dynamic harmonic regression. In: *Forecasting: Principles and Practice*. 3rd ed. OTexts; 2021.
21. China Meteorological Data Network. URL: <http://data.cma.cn> [Accessed 2026-07-25]
22. China National Environmental Monitoring Centre. URL: <https://www.cnemc.cn> [Accessed 2026-07-25]
23. Box GE, Jenkins GM, Reinsel GC, Ljung GM. *Time Series Analysis: Forecasting and Control*. 5th ed. John Wiley & Sons; 2015. ISBN: 9781118675021
24. Granger CW. Investigating causal relations by econometric models and cross-spectral methods. *Econometrica*. Aug 1969;37(3):424-438. [doi: [10.2307/1912791](https://doi.org/10.2307/1912791)]
25. Yu LJ, Li XL, Wang YH, et al. Short-term exposure to ambient air pollution and influenza: a multicity study in China. *Environ Health Perspect*. Dec 2023;131(12):127010. [doi: [10.1289/EHP12146](https://doi.org/10.1289/EHP12146)] [Medline: [38078423](https://pubmed.ncbi.nlm.nih.gov/38078423/)]
26. Yang J, Yang Z, Qi L, et al. Influence of air pollution on influenza-like illness in China: a nationwide time-series analysis. *EBioMedicine*. Jan 2023;87:104421. [doi: [10.1016/j.ebiom.2022.104421](https://doi.org/10.1016/j.ebiom.2022.104421)] [Medline: [36563486](https://pubmed.ncbi.nlm.nih.gov/36563486/)]
27. Pivette M, Mueller JE, Crépey P, Bar-Hen A. Drug sales data analysis for outbreak detection of infectious diseases: a systematic literature review. *BMC Infect Dis*. Nov 18, 2014;14:604. [doi: [10.1186/s12879-014-0604-2](https://doi.org/10.1186/s12879-014-0604-2)] [Medline: [25403237](https://pubmed.ncbi.nlm.nih.gov/25403237/)]
28. Tibshirani R. Regression shrinkage and selection via the lasso. *J R Stat Soc Series B Stat Methodol*. 1996;58(1):267-288. [doi: [10.1111/j.2517-6161.1996.tb02080.x](https://doi.org/10.1111/j.2517-6161.1996.tb02080.x)]
29. Garrido-Garcia R, Clemente L, Meyer AG, Dewey G, Yang S, Santillana M. A real-time early warning system to anticipate respiratory disease outbreaks using transfer learning. *Nat Commun*. May 19, 2026;17(1):6599. [doi: [10.1038/s41467-026-72655-7](https://doi.org/10.1038/s41467-026-72655-7)] [Medline: [42156378](https://pubmed.ncbi.nlm.nih.gov/42156378/)]
30. Mellor J, Christie R, Overton CE, et al. Forecasting influenza hospital admissions within English sub-regions using hierarchical generalised additive models. *Commun Med (Lond)*. Dec 20, 2023;3(1):190. [doi: [10.1038/s43856-023-00424-4](https://doi.org/10.1038/s43856-023-00424-4)] [Medline: [38123630](https://pubmed.ncbi.nlm.nih.gov/38123630/)]
31. Wang Z, Yu P, Yue Z, et al. Internet-based healthcare services use patterns and barriers among middle-aged and older adults in China: a cross-sectional study. *BMC Public Health*. 2025;25(1):2188. [doi: [10.1186/s12889-025-23338-3](https://doi.org/10.1186/s12889-025-23338-3)] [Medline: [40604753](https://pubmed.ncbi.nlm.nih.gov/40604753/)]
32. Tan HM, Tan MS, Chang ZY, et al. The impact of COVID-19 pandemic on the health-seeking behaviour of an Asian population with acute respiratory infections in a densely populated community. *BMC Public Health*. Jun 23, 2021;21(1):1196. [doi: [10.1186/s12889-021-11200-1](https://doi.org/10.1186/s12889-021-11200-1)] [Medline: [34158019](https://pubmed.ncbi.nlm.nih.gov/34158019/)]

Abbreviations

ARIMA: autoregressive integrated moving average
CO: carbon monoxide
EBM: explainable boosting machine
FDR: false discovery rate
LASSO: least absolute shrinkage and selection operator
NO₂: nitrogen dioxide
O₃: ozone
RF: random forest
RMSE: root mean squared error
SARIMA: seasonal autoregressive integrated moving average
SO₂: sulfur dioxide

Edited by Matthew Balcarras; peer-reviewed by Ayush Goyal, Hamed Karami, Shimpy Ralhan; submitted 25.Mar.2026; final revised version received 23.Aug.2026; accepted 24.Aug.2026; published 23.Sep.2026

Please cite as:

Yang R, Xu Z, Zhang C, Qiu L, Qian J, Liu J, Dong Y, Zeng Z, Shen R, Feng L
Digital Prescription–Based Influenza Activity Forecasting in Jiangxi Province, China: Comparative Modeling Analysis Using Multisource Data
J Med Internet Res 2026;28:e96131
URL: <https://www.jmir.org/2026/1/e96131>
doi: [10.2196/96131](https://doi.org/10.2196/96131)

© Rongrong Yang, Zilu Xu, Chi Zhang, Liping Qiu, Jie Qian, Jie Liu, Yonghai Dong, Zhili Zeng, Rui Shen, Luzhao Feng. Originally published in the Journal of Medical Internet Research (<https://www.jmir.org>), 23.Sep.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (<https://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on <https://www.jmir.org/>, as well as this copyright and license information must be included.