Methods and Applications

An Open-Source Data Driven Hybrid Modeling System for Infectious Disease Surveillance and Early Warning

Jianyi Zhang[1]; Haoliang Cui[1]; Yiwen Xing[1]; Zekun Wang[2]; Wenkai Luo[1]; Chaozhuo Wei[3]; Zhongwei Jia[1,#]

ABSTRACT

Introduction: The increasing trend of globalization has led to a heightened risk of imported epidemics; however, existing surveillance systems remain fragmented and reliant on laboratory confirmation. We developed an open-source data-driven hybrid modeling system to provide earlier and more reliable alerts, designed to complement China's multipoint trigger early-warning framework.

Methods: This system integrates heterogeneous signals, including official epidemiology, digital traces, mobility, meteorology, and pathogen genomics, using semantic harmonization and a hybrid analytic stack. Seasonality-adjusted baselines with anomaly detection, mobility- and climate-aware SEIR models, and short-horizon learners generated calibrated early-warning scores. Thresholds were constrained by positive predictive value. Pilot studies were conducted for coronavirus disease 2019 (COVID-19) in Yantai and severe fever with thrombocytopenia syndrome virus (SFTSV) in Shandong and Henan, with tuberculosis indicators embedded for programmatic use.

Results: Across deployments, the system achieved 83.3% sensitivity and 76.9% positive predictive value, providing a median lead time of 9.3 days before official confirmation. Forecasting accuracy reached 92.1% for COVID-19 in Yantai, 90.3% for SFTSV in Shandong, and 89.8% for SFTSV in Henan. Early warnings were aligned with subsequent confirmations and supported targeted screening and resource allocation.

Conclusion: An open-source data-driven hybrid modeling system can deliver calibrated and timely alerts across diverse pathogens. By broadening inputs, enabling cross-agency linkage, and offering operator-oriented dashboards, it serves as a practical complement to China's national early-warning system and has the potential for scaling out with One Health inputs.

Globalization and increased human mobility have raised the risk of infectious diseases. International tourist arrivals and global traffic have roughly doubled since 2000 (*1*). During the coronavirus disease 2019 (COVID-19) pandemic, imported cases repeatedly seeded local outbreaks in China, while the expanding distribution of severe fever with thrombocytopenia syndrome across East Asia illustrates cross border spread of vector-borne diseases (*2*–*3*). Tuberculosis (TB) remains a persistent global threat; with 10.6 million new cases and 1.3 million deaths in 2022; and rebounds in China underscore the need for improved prevention along travel corridors (*4*–*5*).

China has developed a nationwide surveillance backbone, including the National Notifiable Infectious Disease Reporting System (NIDRIS) and the China Infectious Disease Automated-alert and Response System (CIDARS) that provide direct case reporting and rule-based signal generation from statutory notifiable diseases (*6*–*7*). More recently, national guidance emphasizes multi-point trigger early-warning architecture aimed at integrating multiple data sources, enhancing interoperability, and supporting multi-agency collaboration (*8*–*9*). However, most current pilot studies and applications rely primarily on report-based analytics, such as space-time scan statistics, which identify spatiotemporal clusters but remain constrained by delayed confirmation, limited data inputs, and weak predictive power (*10*–*11*). These limitations reduce actionable lead time and restrict applicability to pathogens with long incubation periods or non-specific clinical presentations.

Epidemic intelligence research has explored statistical, mechanistic models, and machine learning approaches separately; however, few studies combine them in hybrid frameworks balancing interpretability and accuracy (*12*–*13*). Existing studies often lack interoperability standards and operator-facing dashboards, limiting their scalability and usability in real-world decision-making environments.

To address these gaps, we introduce an open-source data-driven hybrid modeling system designed to complement China's national multi-point-trigger early-warning architecture. The system integrates heterogeneous open and partner-shared signals — including epidemiological reports, digital traces, mobility, meteorology, and pathogen genomics — through semantic harmonization and hybrid analytics, including seasonality-adjusted baselines, anomaly detection, mobility- and climate-aware SEIR models, and short-horizon sequence learners. Interoperable HL7 FHIR-aligned data contracts enable scalable integration with health, customs, and laboratory systems (*14*). While operator-oriented dashboards follow established design principles for interpretability and oversight (*15*). We present the system's architecture and pilot evidence across COVID-19 and severe fever with thrombocytopenia syndrome virus (SFTSV) and show how the same framework embeds TB indicators for programmatic use, bridging open-source data intelligence with the national early-warning workflow.

METHODS

We selected three pathogen targets to test the system's One Health versatility across distinct transmission modes and timescales: COVID-19 (acute respiratory disease requiring rapid community forecasting), SFTSV (vector-borne disease requiring ecological integration), and tuberculosis (chronic disease requiring long-term strategic planning). The pilot sites were chosen based on disease burden and data feasibility; for SFTSV, Shandong and Henan provinces were selected as high-endemicity regions in China, providing sufficient case volume to validate vector-driven models. For COVID-19, Yantai was selected as a representative coastal port city that experienced distinct waves of local transmission triggered by importation. This setting offered clear onset-to-suppression dynamics essential for validating the community forecasting model's sensitivity to intervention measures. The system continuously ingests heterogeneous data, performs semantic harmonization, runs hybrid analytics (statistical baselines, mechanistic models, and deep-sequence learners), and emits a calibrated early warning score (EWS) for operations. Personally identifiable information was not collected or processed (Figure 1).

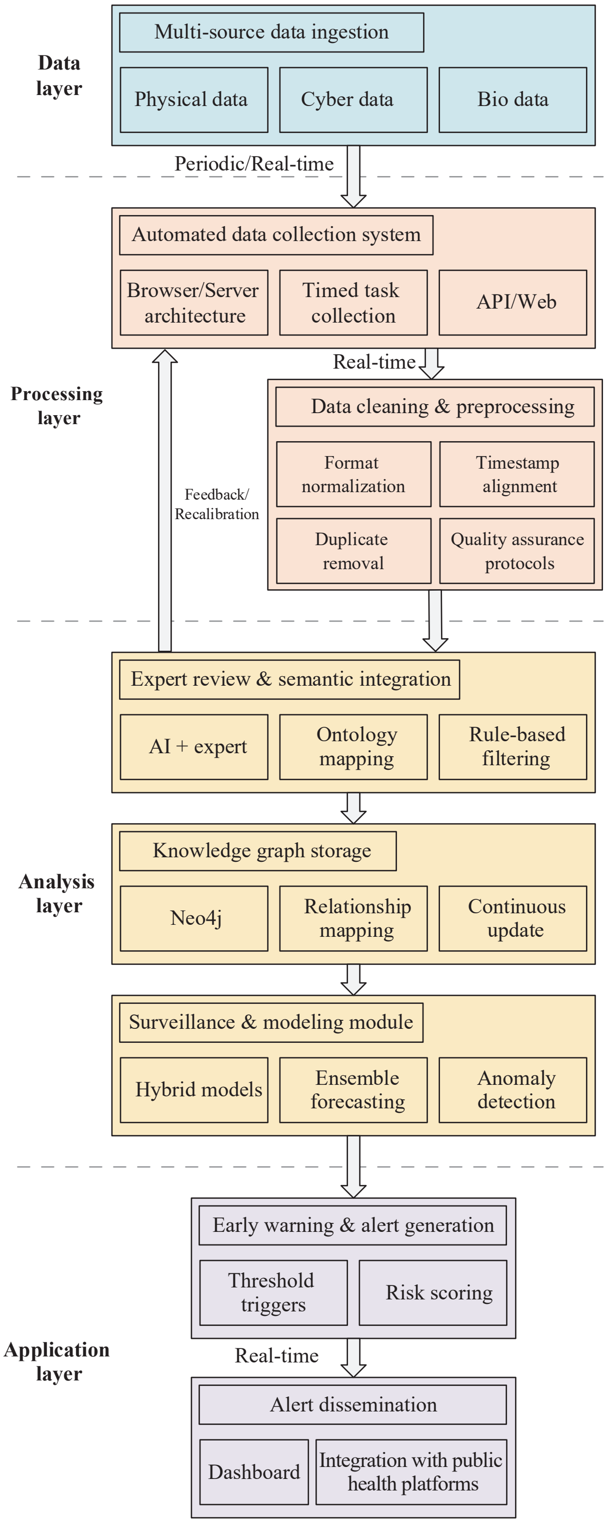


FIGURE 1. Workflow from data ingestion to alert dissemination.

Data Sources and Preprocessing

In this study, the term "open-source data" refers to open-source intelligence (OSINT) and publicly available datasets that are accessible without proprietary restrictions. These include official bulletins, digital signals, meteorological records, and anonymized mobility data, distinct from internal hospital records or confidential line-list data.

Official epidemiology relies on national/provincial bulletins and WHO/ECDC situation updates; digital epidemiology integrates multichannel digital traces such as search engine queries (Baidu Index, Wikipedia Pageviews), social media discussions (Weibo), and content from aggregators (Douyin/TikTok China, Toutiao), all equipped with geotags and temporal stamps; genomics provides sequence metadata for pathogen context; context & covariates encompass human mobility, meteorology, and holiday markers; and for the vector signal SFTSV, national tick index surveillance data is digitized from official graphs using scale-mean abstraction to create daily or weekly exogenous drivers, with monthly series derived through calendar aggregation.

Model Development Methodology

In this study, we developed three distinct model components integrated through a hybrid framework, with detailed methodologies provided in the supplementary materials. The SFTSV model utilizes a network transmission approach where the vector driver is approximated by a Fourier series fitted to the 2018–2019 national tick index, assuming stationary seasonal phenology. For the COVID-19 model, an agent-based SEIR model was implemented on a dynamic contact graph; biological parameters were fixed to literature values to ensure identifiability, focusing calibration solely on the effective contact probability. Rifampicin-resistant tuberculosis (RR-TB) incidence was estimated following the WHO-recommended mathematical procedure (Supplementary Table S1, available at https://weekly.chinacdc.cn). Finally, these outputs were integrated via hybrid fusion, employing logistic stacking as a meta-learner to weigh mechanistic and deep learning signals based on their historical performance.

RESULTS

System-level Alert Performance

Across pilots, the system operated at a pre-specified threshold tuned for decision utility (PPV constraint ≥ 0.70). Against officially confirmed events, the system achieved 83.30% sensitivity and 76.90% positive predictive value (PPV), with a median lead time of 9.30 days before first confirmation. Alerts and confirmatory timelines are illustrated in the dashboard traces (Figure 2E–F); adjudication logs indicate that most false positives arose from short sub-threshold anomalies that did not consolidate into confirmed events (Table 1).

Site-level Forecasting Performance

For SFTSV monthly incidence forecasting in Shandong and Henan, the model's predictions closely tracked observed trends in both provinces, as illustrated in Figure 3A–B. Using the pre-specified accuracy metric with bootstrap 95% confidence intervals (*CI*s), Shandong (2013–2015) achieved 90.29% accuracy (95% *CI*: 85.79%, 93.84%). Henan (2009–2014; including Xinyang) achieved 89.81% (95% *CI*: 86.24%, 93.08%).

Peak months and troughs aligned with the seasonality captured by the mechanistic (tick- and human-driven) transmission terms, and the model reproduced the interannual amplitude differences without overfitting (Figure 2A–B).

In the COVID-19 community forecasting conducted for Yantai, community-scale forecasts achieved 92.15% accuracy (95% *CI*: 86.99%, 93.96%) under the same definition. In peak-focused validation with 10,000 simulations (Poisson-drawn initial seeds within the 95% interval), the model achieved a peak timing accuracy of 88.43% (95% *CI*: 88.26%, 88.59%) and a peak magnitude accuracy of 91.16% (95% *CI*: 91.04%, 91.30%). The forecast trajectories and observed counts are shown in Figure 2C–D.

At the PPV-constrained threshold, the median lead time was 9.3 days (overall). Most detected events had ≥ 7 days' advance notice; short-lead alerts (<7 days) clustered in late-season periods with compressed confirmation cycles (timeline examples in Figure 3E).

Our TB model, adapted from the recurrent framework of Li (5), closely reproduced historical trends (R^2=0.95 for total incidence, 0.99 for RR-TB incidence, and 0.82 for TB deaths), with a posterior mean force of infection of 2.35 per year (95% *CI*: 1.16, 3.58). Projections to 2030 indicated an incidence rate of 33.7 per 100,000 (95% *CI*: 30.80, 38.30), below Li's estimate of 44.9 but above the End TB target of 13, suggesting China's 2024–2030 goal (43) is attainable. The model was implemented as an interactive Shiny application to support visualization and policy use.

DISCUSSION

This study demonstrates that the system can combine diverse open signals with hybrid models to

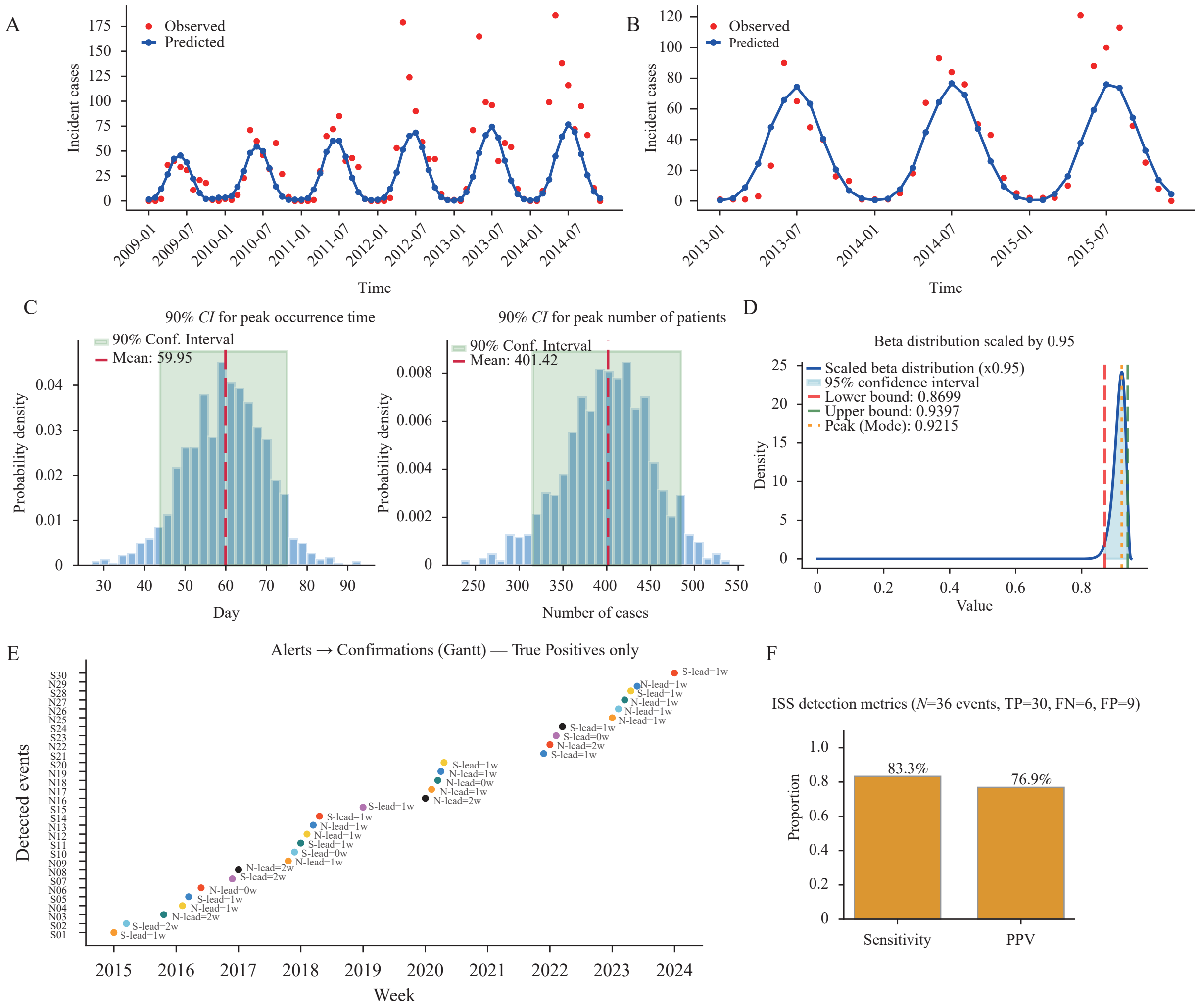


FIGURE 2. Model performance across pilots. (A) Monthly SFTSV cases in Henan (2009–2014); (B) Monthly SFTSV cases in Shandong (2011–2015). (C) Posterior distributions of COVID-19 predicted outcome intervals; (D) Beta-distributed probability calibration for COVID-19 with 95% confidence intervals; (E) System alerts and official confirmations for true-positive events; (F) Aggregate detection metrics including sensitivity and positive predictive value.

Abbreviation: PPV=positive predictive value; SFTSV=severe fever with thrombocytopenia syndrome virus; *CI*=confidence interval; TP=true positive; FN=false negative; FP=false positive; COVID-19=coronavirus disease 2019.

produce calibrated early-warning scores constrained by positive predictive values, reducing false alerts while preserving sensitivity. We tested three pathogen contexts — COVID-19, SFTSV, and TB — and observed their practical utility in both acute and chronic use cases.

Semantic harmonization organized multi-source evidence into consistent geotemporal units, reducing ambiguity in sparse or fast-moving events. Hybrid modeling integrated statistical baselines, mobility- and climate-aware SEIR models (including a human-tick-human pathway for SFTSV), and short-horizon learners to preserve epidemiologic interpretability while capturing nonlinearity. PPV-constrained probability calibration translated model outputs into actionable alerts, improving resource allocation and limiting alert fatigue. Together, these choices enabled earlier, more precise alerts that aligned well with observed trends without overfitting to site-specific conditions.

Relationship to Prior Work and Added Value

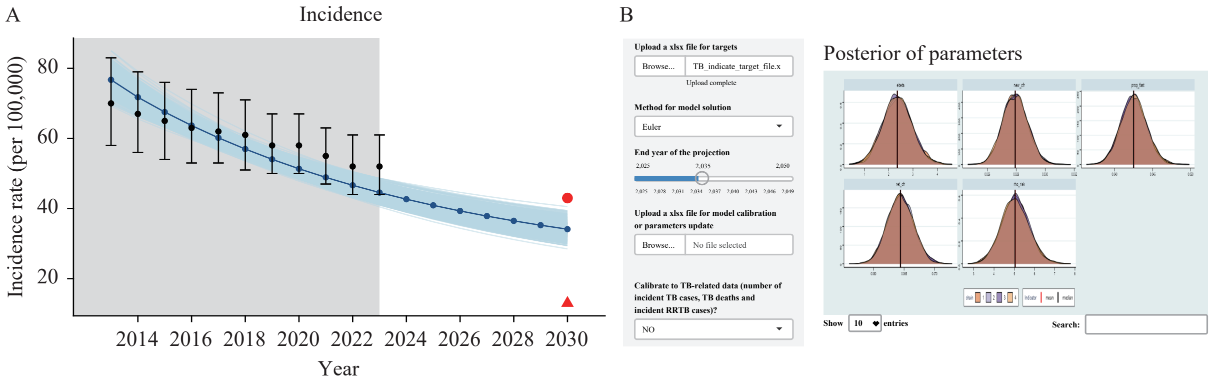
While frameworks such as EWARS support outbreak management (8–9), their reliance on statutory reports limits their timeliness (6). Previous studies have often traded interpretability (statistical baselines) for short-term accuracy (machine learning), frequently lacking multisource integration. Our system advances the field by 1) hybridizing statistical, mechanistic, and sequence-based learners to balance

TABLE 1. System- and site-level performance summary.

Setting	Pathogen	Outcome granularity	Detection sensitivity, %	PPV, %	Median lead time, days	Forecast accuracy, % (95% *CI*)	Peak timing accuracy, % (95% *CI*)	Peak magnitude accuracy, % (95% *CI*)	$R^2_{totalincidence}$	$R^2_{RR-TBincidence}$	$R^2_{TBdeaths}$
Overall (all pilots)	Mixed	Event-level alerts	83.3	76.9	9.3						
Yantai, Shandong	COVID-19	Community time series				92.15 (86.99, 93.96)	88.43 (88.26, 88.59)	91.16 (91.04, 91.30)			
Shandong	SFTSV	Monthly incidence				90.29 (85.79, 93.84)	–	–			
Henan	SFTSV	Monthly incidence				89.81 (86.24, 93.08)	–	–			
China (National)	TB	National annual incidence							0.95	0.99	0.82

Note: "–" means no data. The indicators of TB are presented in the form of proportions.

Abbreviation: COVID-19=coronavirus disease 2019; SFTSV=severe fever with thrombocytopenia syndrome virus; *CI*=confidence interval; TB=tuberculosis.



Upload a xlsx file for targets

TB_indicate_target_file.x

Upload complete

Method for model solution

Euler

End year of the projection

2,025 2,035 2,050

2,025 2,028 2,031 2,034 2,037 2,040 2,043 2,046 2,049

Upload a xlsx file for model calibration or parameters update

No file selected

Calibrate to TB-related data (number of incident TB cases, TB deaths and incident RRTB cases)?

NO

Show 10 entries

FIGURE 3. Dashboard (TB). (A) Incidence Trends and Projections; (B) Posterior Parameters and Model Application
Note: The system showed high stability in robustness checks under parameter perturbations.
Abbreviation: TB=tuberculosis.

interpretability with adaptability; 2) integrating open signals beyond statutory notifications; and 3) achieving high predictive accuracy and meaningful lead times relative to uncalibrated systems.

Qualitatively, the hybrid framework offers distinct advantages over the single-method baselines. Mechanistic SEIR models capture long-term seasonal trends but lag during stochastic onsets, whereas deepsequence learners offer high sensitivity but lack epidemiological transparency. By fusing these approaches, our system stabilizes forecasts during peaks while improving sensitivity during early onsets. Quantitatively, the system demonstrated a median lead time of 9.3 days relative to official confirmation. Given the inherent reporting lags in traditional passive surveillance (7), this represents a substantial window for pre-emptive intervention.

Pathogen Landscape Perspective

A modular design allows flexible application across pathogen contexts. The same framework generates outbreak alerts for COVID-19 and SFTSV while embedding TB analytics to strengthen screening and continuity of care. This adaptability enables emergency response and long-term control through a unified operational surface.

Operational utility differs across pathogen types. For acute outbreaks, such as COVID-19 and SFTSV, the system functions as a tactical early-warning tool, issuing short-horizon alerts (lead time <14 days) to trigger immediate containment measures such as targeted screening or vector control. For chronic diseases such as TB, the system serves a strategic forecasting function, projecting long-term trends (e.g., to 2030) to guide resource allocation and policy target setting. This multimodal capability aligns with the tiered surveillance architecture advocated in recent national guidance on intelligent multi-point trigger systems (8–9).

Implications for Public Health Practice

Embedding such a system into an operational setting can accelerate detection, improve the allocation of quarantine and laboratory resources, and better align vector control with clinical responses during high-risk periods. These functions align with national guidance on building a multipoint-trigger early-warning architecture (*8*–*9*) and with international calls to strengthen public health forecasting (*12*). The system also benefits from interoperable data contracts, such as HL7 FHIR, which facilitate scalable integration across health, customs, and laboratory agencies (*14*), as well as operator-oriented dashboards designed for real-time decision support (*15*). In the short term, priorities include regular recalibration, expanded data exchange with partner agencies, and the incorporation of operator feedback loops. In the medium term, multisite evaluations are required to provide robust evidence of improved timeliness and efficiency.

First, data-related issues exist in multiple aspects. Digital traces are susceptible to “media noise”, and smartphone-derived mobility data may underrepresent the elderly. Meanwhile, meteorological data face spatiotemporal alignment challenges. Second, there are ecological and modeling-related constraints. The national tick index has limited local granularity in terms of ecological constraints. Structural simplifications, assuming uniform mixing for COVID-19 or simplified vector-host cycles for SFTSV, may overlook microenvironmental heterogeneity. Finally, there are coverage and parameter-related problems. Pilot coverage was geographically limited, and parameter uncertainty persists as PPV thresholds need recalibration and TB models depend on uncertain latent progression parameters.

Conflict of Interest: No conflicts of interest.

Funding: Supported by the National Natural Science Foundation of China (72274210, 72174004).

doi: 10.46234/ccdcw2026.036

# Corresponding author: Zhongwei Jia, jiazw@bjmu.edu.cn.

[1] Department of Global Health, School of Public Health, Peking University, Beijing, China; [2] Beijing Municipal Health Big Data and Policy Research Center, Beijing, China; [3] Division of Surveillance, Early Warning and Emergency Response, Heilongjiang Provincial of Disease Control and Prevention, Harbin City, Heilongjiang Province, China.

Submitted: October 11, 2025
Accepted: February 09, 2026
Issued: February 20, 2026

REFERENCES

1. Yu XJ, Liang MF, Zhang SY, Liu Y, Li JD, Sun YL, et al. Fever with thrombocytopenia associated with a novel bunyavirus in China. N Engl J Med 2011;364(16):1523 – 32. https://doi.org/10.1056/NEJMoa1010095.
2. Cui HL, Shen SJ, Chen L, Fan ZY, Wen Q, Xing YW, et al. Global epidemiology of severe fever with thrombocytopenia syndrome virus in human and animals: a systematic review and meta-analysis. Lancet Reg Health West Pac 2024;48:101133. https://doi.org/10.1016/j.lanwpc.2024.101133.
3. Miao D, Dai K, Zhao GP, Li XL, Shi WQ, Zhang JS, et al. Mapping the global potential transmission hotspots for severe fever with thrombocytopenia syndrome by machine learning methods. Emerg Microbes Infect 2020;9(1):817 – 26. https://doi.org/10.1080/22221751.2020.1748521.
4. World Health Organization (WHO). Global tuberculosis report 2023. Geneva: WHO; 2023. https://www.who.int/publications/i/item/9789240083851.
5. Li T, Zhang B, Du X, Pei SJ, Jia ZW, Zhao YL. Recurrent pulmonary tuberculosis in China, 2005-2021. JAMA Netw Open 2024;7(8):e2427266. https://doi.org/10.1001/jamanetworkopen.2024.27266.
6. Sun HM, Hu WH, Wei YY, Hao YT. Drawing on the development experiences of infectious disease surveillance systems around the world. China CDC Wkly 2024;6(41):1065 – 74. https://doi.org/10.46234/ccdcw2024.220.
7. Ren X, Wang LP, Cowling BJ, Zeng LJ, Geng MJ, Wu P, et al. Systematic review: national notifiable infectious disease surveillance system in China. Online J Public Health Inform 2019;11(1):e62534. https://doi.org/10.5210/ojphi.v11i1.9897.
8. National Bureau of Disease Control and Prevention, National Health Commission of the People’s Republic of China, National Development and Reform Commission, Ministry of Education of the People’s Republic of China, Ministry of Civil Affairs of the People’s Republic of China, Ministry of Finance of the People’s Republic of China, et al. Guiding opinions on the establishment of a sound and intelligent multi-point trigger infectious disease surveillance and early warning system. 2024. https://www.gov.cn/zhengce/zhengceku/202408/content_6971481.htm. (In Chinese).
9. Yang WZ, Lan YJ, Lyu W, Leng ZW, Feng LZ, Lai SJ, et al. Establishment of multi-point trigger and multi-channel surveillance mechanism for intelligent early warning of infectious diseases in China. Chin J Epidemiol 2020;41(11):1753 – 57. https://doi.org/10.3760/cma.j.cn112338-20200722-00972.
10. Liu QQ, Li JH, Liu SY, Tang L, Wang XQ, Huang AD, et al. The Epidemiological Characteristics and Spatiotemporal Clustering of Measles—China, 2005-2022. China CDC Wkly 2024;6(27):665 – 9. https://doi.org/10.46234/ccdcw2024.123.
11. Levin-Rector A, Kulldorff M, Peterson ER, Hostovich S, Greene SK. Prospective spatiotemporal cluster detection using SaTScan: tutorial for designing and fine-tuning a system to detect reportable communicable disease outbreaks. JMIR Public Health Surveill 2024;10:e50653. https://doi.org/10.2196/50653.
12. Lutz CS, Huynh MP, Schroeder M, Anyatonwu S, Dahlgren FS, Danyluk G, et al. Applying infectious disease forecasting to public health: a path forward using influenza forecasting examples. BMC Public Health 2019;19(1):1659. https://doi.org/10.1186/s12889-019-7966-8.
13. Reich NG, Brooks LC, Fox SJ, Kandula S, McGowan CJ, Moore E, et al. A collaborative multiyear, multimodel assessment of seasonal influenza forecasting in the United States. Proc Natl Acad Sci USA 2019;116(8):3146 – 54. https://doi.org/10.1073/pnas.1812594116.
14. Vorisek CN, Lehne M, Klopfenstein SAI, Mayer PJ, Bartschke A, Haese T, et al. Fast Healthcare Interoperability Resources (FHIR) for interoperability in health research: systematic review. JMIR Med Inform 2022;10(7):e35724. https://doi.org/10.2196/35724.
15. Rabiei R, Bastani P, Ahmadi H, Dehghan S, Almasi S. Developing public health surveillance dashboards: a scoping review on the design principles. BMC Public Health 2024;24(1):392. https://doi.org/10.1186/s12889-024-17841-2.

SUPPLEMENTARY MATERIAL

DATA PROCESSING LAYER CONSTRUCTION

As illustrated in the Processing Layer of Figure 1, the system follows a two-stage construction protocol. 1) Automated Data Collection: Built on a Browser/Server (B/S) architecture, utilizing timed task collection modules to continuously ingest data via public APIs and web-scraping interfaces. 2) Data Cleaning & Preprocessing: Raw inputs undergo a standardized pipeline: 1) Format Normalization parses heterogeneous text/JSON into structured schemas; 2) Timestamp Alignment synchronizes all records to UTC+8; 3) Duplicate Removal eliminates redundant entries via hash matching; and 4) Quality Assurance Protocols (e.g., winsorization) filter outliers to ensure data integrity. Source-region normalization was performed using a Z-score with a small ridge.

$$\tilde{x}_{s,t,r} = \frac{x_{s,t,r} - \mu_{s,r}}{\sigma_{s,r} + \varepsilon} \tag{1}$$

MODEL DEVELOPMENT METHODOLOGY

SFTSV Network Transmission Model (Human-Tick-Human)

Tick Index as a Seasonal Driver: A Fourier series approximates the tick indexTickIndex$(t)$to capture the annual and biannual cycles:

$$\text{TickIndex}(t) = A_0 + A_1 \sin(\omega t) + B_1 \cos(\omega t) + A_2 \sin(2\omega t) + B_2 \cos(2\omega t) \tag{2}$$

estimated using the nonlinear least-squares method. This driver transmits the force of infection from the tick into a human channel.

The monthly observed incidence $\{Y_k\}_{k=1}^{n}$ was modeled as a Poisson variable, with the mean $\lambda_k(\theta)$ implied by the state equations:

$$\begin{aligned} &Y_k \sim \text{Pois}(\lambda_k(\theta)), \\ &\log \mathcal{L}(\theta) = \sum_{k=1}^{n} \log(\text{Pois}(Y_k \mid \lambda_k) + 10^{-7}). \end{aligned} \tag{3}$$

where the small constant prevents numerical underflow. Posterior computation uses a grid/maximum A Posteriori (MAP) strategy with log-posterior stabilization:

$$P(\theta \mid D) \propto \exp\left(\log P(D \mid \theta) + \log P(\theta) - \max_{\theta'} \log P(D \mid \theta')\right) \tag{4}$$

followed by normalization and MAP selection. Sensitivity checks use a random-start local search over the prior box.

We utilized 2018–2019 national tick surveillance data to derive a normalized seasonal function of vector activity. The use of this later-period index to drive the historical model (2009–2015) is consistent with long-term epidemiological observations, indicating that vector phenology follows relatively stable annual cycles driven by climatic factors. Based on this, we assumed that the seasonal shape remained stationary, while allowing the transmission amplitude to fit dynamically to the observed incidence in each year.

COVID-19 Community Network Model (agent SEIR on Dynamic Contact Graph)

Resident Classes: Residents were grouped according to mobility patterns: 1) fixed workplace/school (regular, high-frequency travel), 2) no fixed workplace/retired/home-based (variable patterns), and 3) low mobility (elderly/children).

Graph construction. Each household is a node with attributes (demography, mobility propensity). Edges represent residential proximity (within unit/building/compound). During simulation, edges are dynamically augmented by co-presence in venues (workplace/school, transit, shops/parks/clinics), forming a time-varying multilayer graph.

A semi-mechanistic approach was adopted to ensure parameter identifiability. Biological parameters were fixed based on established literature: the mean latent period was set to 5.2 days and the infectious period to 5.8 days.

Agents transitioned between compartments (Susceptible-Exposed-Infectious-Recovered) based on these fixed intervals and a time-varying force of infection derived from the dynamic contact graph. Consequently, the model calibration focused solely on estimating the effective contact probability (β_t) and the initial seed size, thereby reducing dimensionality and preventing overfitting.

Estimation of Incidence of RR-TB (Rifampicin-Resistant Tuberculosis)

Total tuberculosis incidence $I(t)$ was obtained from the compartmental model as

$$I(t) = \omega_{fast} \cdot E_{fast}(t) + \omega_{slow} \cdot E_{slow}(t) + \rho \cdot R(t) \tag{5}$$

We then applied the mathematical procedure recommended by the World Health Organization to derive the incidence of rifampicin-resistant (RR) TB, expressed by the following equation:

$$I_{rr}(t) = I(t)[(1-f) \cdot p_{new}(t) \cdot ((1-r(t)) + r(t) \cdot \rho_{RRTB}) + f \cdot p_{ret}(t)] \tag{6}$$

Deep sequence Learners and Fusion

Parallel sequence models (LSTM/temporal CNN) use multistream inputs$X_{t,i}$(normalized signals, recent cases, mobility, and meteorology) to predict h-step-ahead counts:

$$\hat{y}_{t+h,i} = g_\omega([X_{t-w+1,i}, \ldots, X_{t,i}]) \tag{7}$$

trained with a Huber loss and correlation-promoting regularizer:

$$\mathcal{J} = \sum_{t,i} \mathcal{L}\left(y_{t,i}, \hat{y}_{t,i}\right) - \lambda \operatorname{corr}\left(y_{\cdot,i}, \hat{y}_{\cdot,i}\right) \tag{8}$$

Outputs from statistical baselines, mechanistic models, and learners were Platt-scaled to calibrated probabilities $p_{t,i}^{(m)}$ and fused by logistic stacking.

$$\Pr(\text{event}_{t,i} = 1) = \sigma\left(\alpha_0 + \sum_m \alpha_m \operatorname{logit} p_{t,i}^{(m)}\right) \tag{9}$$

An EWS combines stacked risk with anomaly flags:

$$\text{EWS}_{t,i} = \sum_m w_m \tilde{p}_{t,i}^{(m)} + \eta_1 \{S_{t,i} > h\} + \eta_2 \max(Z_{t,i}, 0), \tag{10}$$

and triggers when $\text{EWS}_{t,i} \geq \tau$. Threshold τ is tuned to maximize F_1 under a minimum PPV constraint (PPV≥ 0.70) using rolling-origin evaluation.

The logistic stacking layer functions as a meta-learner that assigns dynamic weights to the component models based on their historical validation performance. This mechanism is designed to reconcile conflicting signals; for instance, the system tends to assign higher weights to mechanistic SEIR outputs during stable seasonal periods (capturing regular trend lines), while upweighting digital signals and deep learners during irregular onset phases, where statistical anomalies precede official reporting. This dynamic weighting strategy aligns with the principles of ensemble model averaging, in which component models are weighted according to their predictive performance in specific contexts.

Semantic Harmonization and Knowledge Graph

An NLP pipeline performs named-entity and relation extraction over unstructured text and bulletins and mapping to controlled vocabulary (disease/syndrome, host/vector, place, and intervention). Entities and relations populate a knowledge graph $G = (V, E)$ that 1) resolves aliases and administrative hierarchies (district→city→province), 2) stores provenance, and 3) serves efficient retrieval for modeling features and operator dashboards.

Statistical Baselines and Anomaly Detection

Syndromic/aggregate counts $Y_{t,r}$ are modeled using a seasonality-adjusted quasi-Poisson Generalized Linear Model (GLM) with trend and holiday effects; standardized residuals feed a one-sided Cumulative Sum (CUSUM) to generate anomaly evidence:

$$Y_{t,r} \sim \mathrm{QP}(\lambda_{t,r}), \log \lambda_{t,r} = \beta_0 + \beta_1 t + f_{\mathrm{seas}}(t) + \gamma^{\top} H_t + \delta_r \tag{11}$$

$$Z_{t,r} = \frac{Y_{t,r} - \hat{\lambda}_{t,r}}{\sqrt{\widehat{\mathrm{Var}}(Y_{t,r})}}, S_{t,r} = \max\{0, S_{t-1,r} + Z_{\mathrm{t,r}} - k\}, alert\ if S_{t,r} > h \tag{12}$$

SFTSV Network Transmission Model (Human-Tick-Human) Calculation Details

Human–tick–human force of infection. Let S_t, E_t, I_t, R_t, D_t denote human compartments, $N_t = S_t + E_t + I_t + R_t$ The total infection pressure is the sum of tick to human and human to human components:

$$\lambda_t = \beta_{th}(t) \frac{S_t}{N_t} \mathrm{TickIndex}(t) + \beta_{hh}(t) \frac{S_t I_t}{N_t} \tag{13}$$

Seasonal/behavioral modulation of transmission:

$$\begin{aligned} \beta_{hh}(t) &= 10^{b_h}, \\ \beta'_{th}(t) &= 10^{b_t} \frac{\sin(\alpha_t t/365 + 2\pi\phi/365) + 1}{2}. \end{aligned} \tag{14}$$

where $\beta'_{th}(t)$ absorbs the proportionality constant K between the observed tick index and effective contacts (only $\beta'_{th}(t)$ is estimated).

State evolution (discrete-time):

$$\begin{aligned} S_{t+1} &= S_t - \lambda_t, \\ E_{t+1} &= E_t + \lambda_t - \sigma E_t, \\ I_{t+1} &= I_t + \sigma E_t - \mu I_t - \delta I_t, \\ R_{t+1} &= R_t + \delta I_t, \\ D_{t+1} &= D_t + \mu I_t. \end{aligned} \tag{15}$$

Here σ is progression from exposed to infectious (latent-to-infectious rate), δ is recovery, μ is disease-induced mortality (fixed from guidelines/meta-analysis as literature-based constants for the baseline run).

Bayesian parameter estimation. We estimate b_h, b_t, α_t, ϕ under weakly informative uniform priors (reflecting seasonality and scale).

$$\begin{aligned} b_h &\sim \mathrm{Unif}(-10, -6), \\ b_t &\sim \mathrm{Unif}(-10, -6), \\ \alpha_t &\sim \mathrm{Unif}(5.98, 6.61), \\ \phi &\sim \mathrm{Unif}(10, 100). \end{aligned} \tag{16}$$

COVID-19 Community Network Model (agent SEIR on Dynamic Contact Graph) Calculation Details

Two transmission phases. Phase-1 (lockdown/home isolation), household droplets and environment/aerosol exposures dominate; Phase-2 (reopening) includes venue-specific contacts (work/school, transit, leisure). Let $\mathcal{N}_i^{\mathrm{home}}$ represent household neighbors, $\mathcal{M}$ venue types, $c_{im}(t)$ expected contacts, and $\pi_{im}(t)$ time fraction in the venue $m; \eta$ vaccine effectiveness (susceptibility reduction). The probability of infection for node i on day t is modeled as:

$$p_i^{(1)}(t) = 1 - (1-\eta) \underbrace{\exp(-\rho_{\mathrm{env}} \tau_i^{\mathrm{home}}(t))}_{\text{no env. infection}} \underbrace{\prod_{j \in \mathcal{N}_i^{\mathrm{home}}} (1 - \rho_{\mathrm{home}} \{I_j(t) = 1\}),}_{\text{no household infection}} \tag{17}$$

$$p_i^{(2)}(t) = 1 - (1-\eta) \exp(-\rho_{\mathrm{env}} \tau_i(t)) \prod_{j \in \mathcal{N}_i^{\mathrm{home}}} (1 - \rho_{\mathrm{home}} \{I_j(t) = 1\}) \prod_{m \in \mathcal{M}} (1 - \rho_m)^{c_{im}(t) \pi_{im}(t)}. \tag{18}$$

Here, $\rho.$ is the per-contact transmission probability, and τ denotes the time in the environment. The state evolution follows the agent-level SEIR with daily updates.

$$\begin{aligned} &X_i(t) \in \{S, E, I, R\}, \\ &\Pr\{X_i(t+1) = E \mid X_i(t) = S\} = p_i(t), \\ &\Pr\{X_i(t+1) = I \mid X_i(t) = E\} = \sigma, \\ &\Pr\{X_i(t+1) = R \mid X_i(t) = I\} = \gamma. \end{aligned} \tag{19}$$

and vaccination acts as a multiplicative susceptibility reduction $(1-\eta)$. For network-level inference, a cavity/percolation recursion approximates the probability that infection does not traverse edge $j \to i$ up to time t:

$$\begin{aligned} \psi_{j\to i}(t) &= \prod_{k\in\mathcal{N}(j)i} [1-\beta_{jk}(1-\psi_{k\to j}(t-1))], \\ \pi_i(t) &= 1-\prod_{j\in\mathcal{N}(i)} [1-\beta_{ij}(1-\psi_{j\to i}(t))]. \end{aligned} \tag{20}$$

which improves stability on large graphs.

Estimation of Incidence of RR-TB (Rifampicin-Resistant Tuberculosis) Calculation Details

Tuberculosis natural history. The compartmental model employed to represent the natural history of TB largely replicates the structure proposed by Li et al., with only minor modifications. We reimplemented the model in Stan, a probabilistic programming language written in C++, to situate the analysis of tuberculosis and other infectious diseases within a coherent Bayesian framework. This model distinguishes between two latency pathways: fast (E_{fast}) and slow ($\mathrm{E}_{\mathrm{slow}}$). As no consensus on the definition of fast-progressing latent infection exists in the literature, we adopted the progression parameters reported by Menzies et al., which are also used by the WHO. The annual rate of progression from the fast-progression latent compartment to active tuberculosis (ω_{fast}) was set at 0.0826 per person-year, whereas the corresponding rate for the slow-progression latent compartment (ω_{slow}) was fixed at 0.0006 per person-year. Disease status was classified into two groups based on the WHO guidance issued in 2013: newly diagnosed TB (I_{new}) and retreated TB (I_{ret}). Although the 2024 WHO revision refers to the latter category as being re-registered for treatment, the retreated designation was retained in the present analysis. Future iterations of the model will adopt the updated terminology. Individuals who were cured or completed their treatment were transferred to the compartment, represented as recovery (R). Individuals in compartment R remain at risk of recurrence and can return to an infectious state (I_{ret}). The model equations are as follows:

$$\begin{aligned} \frac{dS}{dt} &= B\cdot N-\beta\frac{S(I_{new}+I_{ret})}{N}-M\cdot S, \\ \frac{dE_{fast}}{dt} &= (1-g)\cdot\beta\frac{S(I_{new}+I_{ret})}{N}-\omega_{fast}\cdot E_{fast}-M\cdot E_{fast}, \\ \frac{dE_{slow}}{dt} &= g\cdot\beta\frac{S(I_{new}+I_{ret})}{N}-\omega_{slow}\cdot E_{slow}-M\cdot E_{slow}, \\ \frac{dI_{new}}{dt} &= \omega_{fast}\cdot E_{fast}+\omega_{slow}\cdot E_{slow}-(M+CFR_{new})\cdot I_{new}-\eta_{new}\cdot I_{new}-f_{new}\cdot I_{new}, \\ \frac{dI_{ret}}{dt} &= f_{new}\cdot I_{new}+f_{new}\cdot I_{new}+\rho\cdot R-(M+CFR_{ret})\cdot I_{ret}-\eta_{ret}\cdot I_{ret}, \\ \frac{dR}{dt} &= \eta_{new}\cdot I_{new}+\eta_{ret}\cdot I_{ret}-M\cdot R-\rho\cdot R. \end{aligned} \tag{21}$$

Data flux and model calibration. Gao et al. estimated that the prevalence of LTBI in China in 2013 was 18.08%, corresponding to approximately 247 million infections. Houben and Dodd estimated that there were 350 million infections in China in 2014. We rely on the estimation conducted by Gao et al. because it is derived from large-scale empirical data collected in China and, therefore, aligns more closely with the study population and methodological framework. Therefore, our model-fitting starting point was set to 2013.

Robustness Checks

The performance estimates were stable in the rolling-origin evaluation under small perturbations of priors for the SFTSV transmission modifiers and venue contact parameters for COVID-19. Bootstrap confidence intervals

SUPPLEMENTARY TABLE S1. RR-TB calculation parameters and explanations.

Description	Definition	Value
B	Birth rate	Birth rates were modelled using logistic regression. $\text{logit}(B) = \ln\frac{B}{1-B} = b_{0,B} + \mathrm{b}_B\mathrm{t}$ Differentiating the implied logistic curve with respect to time gives the corresponding ordinary differential equation $\frac{dB}{dt} = \mathrm{b}_B\mathrm{B}(1-B)$
M	Background mortality rate	Background mortality rates were modelled using logistic regression. $\text{logit}(M) = \ln\frac{M}{1-M} = b_{0,M} + \mathrm{b}_M\mathrm{t}$ Differentiating the implied logistic curve with respect to time gives the corresponding ordinary differential equation $\frac{dM}{dt} = \mathrm{b}_M\mathrm{M}(1-M)$
β	Transmission rate	Calibrated to match epidemiological data, with uniform priors [0, 30].
g	Proportion of infected individuals that transition into the "slow" progression LTBI compartment	Fixed, 0.91
ω_{fast}	The annual rate of progression from the "fast" progression LTBI to active TB.	Fixed, 0.0826
ω_{slow}	The annual rate of progression from the "slow" progression LTBI to active TB.	Fixed, 0.0006
CFR_{new}	Case fatality rate for newly diagnosed patients	Calibrated to match epidemiological data, with uniform priors [0, 0.2].
CFR_{ret}	Case fatality rate for retreated patients	Calibrated to match epidemiological data, with uniform priors [0, 0.2].
η_{new}	Treatment success rate for newly diagnosed patients	Region varying. For China, the overall value was 0.94.
η_{ret}	Treatment success rate for retreated patients	Region varying. For China, the overall value was 0.85.
f_{new}	Treatment failure rate for newly diagnosed patients	Region varying. For China, the overall value was 0.0227.
f_{ret}	Treatment failure rate for retreated patients	Region varying. For China, the overall value was 0.0527.
ρ	Overall recurrence rate	Region varying. For China, the overall value was 0.0047.
$\mathrm{r}(\mathrm{t})$	The proportion of recurrent cases out of the sum of new and recurrent cases at time t.	Computed from the model as $\mathrm{r}(\mathrm{t}) = \frac{\rho\cdot\mathrm{R}(\mathrm{t})}{\mathrm{I}(\mathrm{t})}$
f	The cumulative risk that an incident case undergoes a non-relapse retreatment, defined as retreatment following treatment failure or return after default.	f_{new} was used as a proxy for f.
$p_{new}(t)$	The estimated proportions of RR-TB among newly diagnosed patients at time t.	Different from the methods used by WHO, we constructed logistic regression to estimate the proportions. $\text{logit}(p_{new}(t)) = \ln\frac{p_{new}(t)}{1-p_{new}(t)} = \mathrm{b}_{0,new} + \mathrm{b}_{new}\mathrm{t}$
$p_{ret}(t)$	The estimated proportions of RR-TB among retreated patients at time t.	Different from the methods used by WHO, we constructed logistic regression to estimate the proportions. $\text{logit}(p_{ret}(t)) = \ln\frac{p_{ret}(t)}{1-p_{ret}(t)} = \mathrm{b}_{0,ret} + \mathrm{b}_{ret}\mathrm{t}$
ρ_{RRTB}	The risk of RR-TB in recurrent cases relative to previously untreated cases.	Calibrated to match epidemiological data, with uniform priors [1, 10].

Abbreviation: LTBI=latent tuberculosis infection; TB=tuberculosis.

overlapped across sites, and alert metrics were consistent when restricted to weeks with complete multisource coverage.

Uncertainty Quantification

We addressed the parameters and predictive uncertainty using distinct approaches tailored to each model component. For the mechanistic SFTSV and TB models, we employed a Bayesian framework and reported parameter estimates with 95% credible intervals (CrIs) derived from the posterior distributions. For the COVID-19 community model and deepsequence learners, where analytical posteriors were intractable, we utilized bootstrap resampling (N=2,000 iterations) to generate 95% confidence intervals (*CIs*) for all predictive horizons.

Evaluation Plan

Event Matching and Definitions: An ‘event match’ was defined as a system alert (EWS > threshold) occurring within a 14-day window preceding the official confirmation date of a case cluster. To account for confirmation delays, the official date was adjusted by subtracting the median reporting lag (2 days for COVID-19 and 5 days for SFTSV) derived from the historical NIDRIS data. Multiple alerts triggered within a single window were aggregated into a single ‘detected event’ to prevent double-counting. Forecast agreement (SFTSV monthly): Accuracy was defined as $(1 - \text{NMAE}) \times 100\%$ with 95% bootstrap CIs. Although standard statistical evaluations typically use RMSE or MAE, we adopted this percentage-based metric to facilitate intuitive interpretation by public health operators. This choice prioritizes communicative utility in operational dashboards over strict statistical conventions, which is consistent with the user-centered design principles in public health surveillance.

$$\begin{aligned}\text{Sensitivity} &= \frac{\text{TP}}{\text{TP} + \text{FN}},\\ \text{PPV} &= \frac{\text{TP}}{\text{TP} + \text{FP}},\\ \text{Specificity} &= \frac{\text{TN}}{\text{TN} + \text{FP}}.\end{aligned} \tag{22}$$

$$\text{Brier} = \frac{1}{n}\sum_{t,i}(p_{t,i} - y_{t,i})^2 \tag{23}$$

$$\begin{aligned}\text{NMAE} &= \frac{\frac{1}{n}\sum_{k=1}^{n}|y_k - \hat{y}_k|}{\max(y) - \min(y)},\\ \text{Accuracy} &= (1 - \text{NMAE}) \times 100\%.\end{aligned} \tag{24}$$