



Available online at <http://scik.org>

Commun. Math. Biol. Neurosci. 2026, 2026:97

<https://doi.org/10.28919/cmbn/10199>

ISSN: 2052-2541

PREDICTIVE MODELS FOR ACUTE RESPIRATORY INFECTION IN INDONESIA USING OFFICIAL SURVEILLANCE DATA AND GOOGLE TRENDS

ILFA STEPHANE^{1,*}, MAHMUD ISNAN², HERU SAPUTRA³, DODI DEVIANTO⁴, ILHAM TRI MAULANA³,

BUDI RAHMADYA⁵

¹Department of Digital Business, Universitas Metamedia, Padang, 25133, Indonesia

²Computer Science Department, School of Computer Science, Bina Nusantara University, Jakarta, 11480, Indonesia

³Department of Information Systems, Universitas Metamedia, Padang, 25133, Indonesia

⁴Department of Mathematics and Data Sciences, Universitas Andalas, Padang, 25175, Indonesia

⁵Department of Computer System, Universitas Andalas, Padang, 25175, Indonesia

Copyright © 2026 the author(s). This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Abstract: Acute respiratory infections (ARI) remain a major public health concern, with substantial regional variation in prevalence. Google Trends may provide digital search signals that complement official health survey data, but its utility for describing provincial variation in ARI prevalence in Indonesia remains insufficiently evaluated. This study evaluated whether Google Trends search indicators could serve as complementary ecological signals for describing variation in ARI prevalence across Indonesian provinces. This cross-sectional ecological study integrated symptomatic ARI prevalence reported in the 2023 Indonesian Health Survey (SKI) with Google Trends Interest by Region data for 34 provinces. The analyses included descriptive statistics, Pearson and Spearman correlations, Benjamini-Hochberg

*Corresponding author

E-mail address: ilfastephane@gmail.com

Received August 3, 2026

false discovery rate (FDR) correction, and multicollinearity assessment. Candidate statistical-learning and machine-learning models, including OLS, LASSO, Ridge, Elastic Net, Support Vector Regression, Random Forest, and XGBoost, were evaluated using nested cross-validation. In the main analysis of 34 provinces, correlations between search indicators and ARI prevalence were generally weak and were not statistically significant after FDR correction. All evaluated models showed limited out-of-fold generalizability. In the sensitivity analysis excluding East Nusa Tenggara and Papua, LASSO using flu, cough and cold, and shortness of breath achieved the best relative performance (RMSE 4.233, MAE 3.292, and global out-of-fold R^2 0.290). Google Trends provided only limited evidence of utility as a complementary and exploratory ecological signal for describing variation in ARI prevalence across provinces. These findings do not support its use as a substitute for official health surveys, forecasting models, or early warning systems.

Keywords: acute respiratory infection; Google trends; digital epidemiology; complementary surveillance; machine learning.

2020 AMS Subject Classification: 62P10, 62J05, 68T09.

1. INTRODUCTION

Acute respiratory infections remain a major contributor to the global burden of infectious diseases. The Global Burden of Disease 2021 analysis identified upper respiratory infections among the most prevalent acute illnesses worldwide, while lower respiratory infections continued to account for substantial infectious-disease mortality. This burden is reflected in morbidity, mortality, health-care visits, hospitalizations, treatment use, and economic costs to health systems [1–3]. In Indonesia, the 2023 Indonesian Health Survey (SKI) reported an ARI prevalence of 23.50% across all age groups, with marked variation among provinces [4]. This heterogeneity highlights the need for representative epidemiological information to support the planning and evaluation of respiratory disease-control policies.

Respiratory disease surveillance systems are a crucial component of public health because they can support the detection of changes in disease patterns, outbreak preparedness, and prioritization of interventions [5, 6]. In the context of severe respiratory disease, Severe Acute Respiratory

Infection (SARI) surveillance has been used in various countries to monitor the circulation of respiratory pathogens, including influenza, SARS-CoV-2, and other epidemiologically relevant pathogens [7–10]. However, conventional surveillance continues to face challenges, including reporting delays, limited coverage of sentinel facilities, reliance on laboratory infrastructure, and the need for significant human resources [8–10]. Therefore, modern surveillance approaches increasingly emphasize a complementary framework, or mosaic surveillance, namely the use of multiple data sources rather than replacing one data source with another [5].

The development of digital epidemiology and infodemiology opens up opportunities to use digital footprints to complement public health surveillance systems [11, 12]. One widely used source is Google Trends, which provides an index of search interest relative to specific terms at specific locations and time periods [13–15]. These data have been used in various health studies, including studies of influenza, influenza-like illness, respiratory syncytial virus, COVID-19, and other respiratory conditions [16–20]. However, Google Trends does not directly measure disease incidence. Its values more closely reflect information-seeking behavior, which can be influenced by media attention, internet access, digital literacy, changes in community behavior, and Google Trends normalization mechanisms [13–15, 21]. Therefore, the use of Google Trends in epidemiology requires validation against official health data and careful interpretation.

Several previous studies have utilized Google Trends or online search data for case forecasting, early outbreak detection, time-series monitoring of respiratory diseases, and the development of artificial intelligence-based approaches for infectious disease surveillance [16, 19, 20, 22–24]. However, evidence on the use of Google Trends to explain variations in survey-based disease prevalence at the national level in developing countries is still limited. In the Indonesian context, this limitation is significant because geographic, social, internet access, and health care access variations can influence the relationship between digital search signals and official epidemiological indicators. Furthermore, the small size of provincial aggregate data requires a transparent, regularized, and rigorously evaluated modeling approach to avoid overly predictive claims.

This study aims to evaluate whether Google Trends search interest indicators can serve as complementary ecological signals to describe variations in ARI prevalence across provinces in Indonesia. This study does not aim to build a forecasting system, outbreak detection system, or final national model. Instead, it assesses the relationship between digital search indicators and survey-based ARI prevalence. It compares the performance of several statistical and machine learning approaches in out-of-fold estimation at the provincial level.

2. MATERIALS AND METHODS

2.1. Study Design. This study used a cross-sectional ecological design with province as the unit of analysis. The reference period was 2023. The study outcome was symptomatic ARI prevalence at the provincial level as reported in the 2023 SKI, while predictors were Google Trends relative search interest indices at the provincial level. Because the unit of analysis was the province, all findings were interpreted as ecological associations and were not used to infer individual-level relationships [25, 26].

The study workflow consisted of extracting ARI prevalence from the 2023 SKI report, collecting and preprocessing Google Trends data, harmonizing provincial units to ensure compatibility with Google Trends, integrating the dataset, conducting descriptive and exploratory analyses, evaluating statistical learning and machine learning models, performing internal validation using nested cross-validation, conducting sensitivity analysis, and interpreting the models with appropriate inferential restrictions. The workflow is summarized in Figure 1.

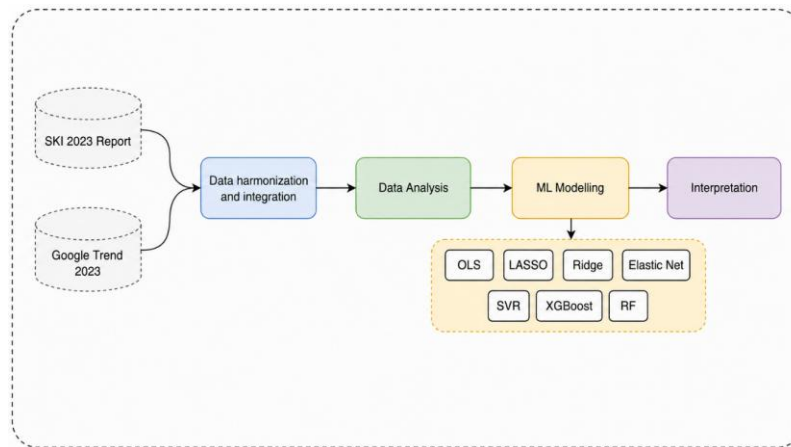


FIGURE 1. Workflow for integrating 2023 SKI and Google Trends data in the provincial ARI prevalence analysis.

2.2. Data Sources and ARI Outcome Definition. This study used two data sources. The outcome source was provincial-level symptomatic ARI prevalence reported in the 2023 Indonesian Health Survey (SKI) by the Ministry of Health of the Republic of Indonesia. The 2023 SKI is a population-based national health survey with a cross-sectional design used to generate health indicators at national and subnational levels [4]. The predictor source was Google Trends Interest by Region for Indonesia in 2023. Thus, ARI prevalence was treated as official provincial health survey data, whereas Google Trends was used as an indicator of relative search interest in terms related to respiratory symptoms and disease.

The primary outcome was symptomatic ARI prevalence at the provincial level. This prevalence was used directly as an official health survey outcome. The analysis focused on evaluating the ecological relationship between ARI prevalence reported in the 2023 SKI and harmonized Google Trends search indicators at the provincial level.

ARI prevalence was treated as a continuous outcome expressed as a percentage. Before integration with Google Trends, prevalence figures were checked to ensure consistency of province names, completeness of observations, and comparability of analytical units. This process was conducted as a data audit, not as a re-estimation procedure.

2.3. Google Trends Acquisition and Preprocessing. Google Trends data were collected through the Interest by Region function using the pytrends 4.9.2 library in Python. Retrieval parameters were Indonesia as the region, January 1 to December 31, 2023 as the time frame, Web Search, All Categories, Indonesian locale (hl = id-ID), and time zone tz = 420. Seven Indonesian search terms were used: batuk (cough), pilek (runny nose), sakit tenggorokan (sore throat), ISPA (ARI), flu (flu), batuk pilek (cough and cold), and sesak napas (shortness of breath). The first four terms were classified as theory-driven because they represented diagnoses or symptoms relevant to the SKI definition, whereas flu, batuk pilek, and sesak napas were classified as exploratory/data-informed based on the literature, clinical relevance, and Google Trends Related Queries.

Each search term was retrieved separately. Therefore, the 0-100 score obtained from Google Trends was interpreted as a relative search interest index for the same term and time period, not as

an absolute search volume. Scores across keywords were not directly compared because each term is normalized independently by Google Trends.

2.4. Regional Harmonization and Data Integration. The 2023 SKI reports ARI prevalence according to the provincial structure applicable to the survey, whereas Google Trends Interest by Region followed the regional structure available on the Google Trends platform at the time of retrieval. To obtain consistent analytical units, regional data were harmonized into 34 provincial units compatible with Google Trends coverage. After province names were standardized, SKI and Google Trends data were merged using the same provincial units. The integration audit included checking for name discrepancies, duplicate observations, and missing values.

2.5. Exploratory Statistical Analysis. Descriptive statistics were used to summarize the distribution of outcomes and search indicators. Extreme observations in ARI prevalence were identified using the interquartile range (IQR) criterion. Extreme observations were not excluded from the main analysis unless there was evidence of measurement, processing, or data entry error. Sensitivity analyses were conducted separately, excluding provinces identified as extreme observations.

The bivariate relationship between each search indicator and ARI prevalence was evaluated using Pearson correlation with 95% confidence intervals based on Fisher's transformation. The Pearson correlation coefficient was calculated as shown in Equation (1). Spearman correlation was used as a supporting analysis to evaluate the consistency of the direction of the rank-based relationship. To reduce the risk of type I error due to seven simultaneous tests, p-values were corrected using the Benjamini-Hochberg False Discovery Rate (FDR) procedure [27]. Inter-predictor correlations and the Variance Inflation Factor (VIF), calculated using Equation (2), were used to assess information redundancy and potential multicollinearity.

$$r = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^n (y_i - \bar{y})^2}} \quad (1)$$

where x denotes the Google Trends search index, y denotes the corresponding provincial ARI prevalence, and n is the number of provinces.

$$VIF_j = \frac{1}{(1-R_j^2)} \quad (2)$$

Here, R^2 for predictor j is obtained by regressing that predictor on the remaining predictors; larger VIF values indicate greater linear redundancy among predictors.

2.6. Candidate Predictor Sets and Models Evaluated. Six candidate predictor sets were evaluated using Ordinary Least Squares: theory-driven predictors (batuk [cough], pilek [runny nose], sakit tenggorokan [sore throat], and ISPA [ARI]); data-informed/exploratory predictors (flu, batuk pilek [cough and cold], and sesak napas [shortness of breath]); top bivariate predictors (batuk pilek, pilek, and batuk); parsimonious predictors (batuk pilek and pilek); a single best predictor (batuk pilek); and all-keywords exploratory predictors (all seven indicators). Predictor sets derived from outcome exploration were interpreted as exploratory rather than confirmatory analyses.

Statistical learning and machine learning benchmarks were conducted using three data-informed/exploratory predictors: flu, batuk pilek (cough and cold), and sesak napas (shortness of breath). The models compared were OLS, LASSO [28], Ridge [29], Elastic Net [30], Support Vector Regression (SVR) with a linear kernel, SVR with a radial basis function (RBF) kernel [31], Random Forest [32], and XGBoost [33]. LASSO was treated as one regularization model evaluated in the benchmark, rather than as a stand-alone feature pre-selection stage before model comparison. Models were implemented in Python using the scikit-learn library [34]. LASSO coefficients were estimated by minimizing the L1-penalized least-squares objective in Equation (3).

$$\hat{\beta} = \arg \min_{\beta_0, \beta} \left\{ \frac{1}{2n} \sum_{i=1}^n \left(y_i - \beta_0 - \sum_{j=1}^p x_{ij} \beta_j \right)^2 + \lambda \sum_{j=1}^p |\beta_j| \right\} \quad (3)$$

In Equation (3), λ controls the strength of L1 regularization; increasing λ produces stronger coefficient shrinkage and can reduce selected coefficients toward zero.

2.7. Internal Validation and Performance Metrics. Model performance was evaluated using nested cross-validation to reduce bias in performance estimates when model selection and model evaluation are performed on the same data [35]. The outer loop used leave-one-out cross-validation (LOOCV) to generate out-of-fold predictions for each province. The inner loop used 5-fold cross-

validation for hyperparameter selection, as required by the models. The entire predictor standardization and hyperparameter selection process was performed within the pipeline on the training data for each outer fold, so information from the test provinces was not used during model training.

The main performance metrics were RMSE and MAE, calculated using Equations (4) and (5). The global out-of-fold R^2 was calculated once using all observed values and the out-of-fold predictions generated by the outer LOOCV, as shown in Equation (6). R^2 values were not calculated for each fold because each test fold contained only one observation. The main analysis was conducted on all 34 provinces. Sensitivity analyses were conducted by excluding East Nusa Tenggara and Papua, bringing the total number of analytical units to 32 provinces. The results of the main and sensitivity analyses are reported separately.

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2} \quad (4)$$

$$MAE = \frac{1}{n} \sum_i^n |y_i - \hat{y}_i| \quad (5)$$

$$R_{\text{OOF}}^2 = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i^{\text{OOF}})^2}{\sum_{i=1}^n (y_i - \bar{y})^2} \quad (6)$$

In these equations, y is the observed ARI prevalence, $\hat{y}$ is the corresponding model prediction, and $\hat{y}^{\text{OOF}}$ is the out-of-fold prediction obtained for a province when that province was excluded from the outer-loop training set.

2.8. Model Interpretation. The model with the best relative performance in the sensitivity analysis was refitted to the entire sensitivity dataset for descriptive interpretation purposes. Standardized coefficients were used to describe the direction of the conditional association between the search indicator and ARI prevalence in the model. The interpretation of the coefficients was not used as causal evidence, for individual-level inference, or to directly generalize to provinces excluded from the sensitivity analysis.

3. RESULTS

3.1. Data Standarization. The ARI prevalence data used in this study were obtained from provincial prevalence figures reported in the 2023 SKI. Before being combined with Google Trends, the data were checked to ensure completeness of observations, consistency of province names, suitability of prevalence units, absence of regional duplication, and absence of missing values in the analytical units.

To form an analytical dataset consistent with Google Trends output, ARI prevalence data from the 2023 SKI were harmonized into 34 provincial units. Harmonization was performed to ensure that the geographic units in the 2023 SKI and Google Trends followed the same regional structure. After standardizing region names and merging the data, all provinces in the SKI dataset had corresponding observations in the Google Trends dataset.

3.2. Integration and Characteristics of the Analytical Dataset. The final analytical dataset consisted of 34 provincial observations, one outcome variable representing ARI prevalence from the 2023 SKI, and seven Google Trends search indicators: batuk (cough), pilek (runny nose), sakit tenggorokan (sore throat), ISPA (ARI), flu, batuk pilek (cough and cold), and sesak napas (shortness of breath). No unmatched provinces, duplicate observations, or missing values were identified after merging.

The simple average ARI prevalence across provinces was 20.57%, with a median of 18.81% and a standard deviation of 6.29 percentage points (Table 1). The lowest prevalence was observed in the Riau Islands (11.45%), while the highest was observed in East Nusa Tenggara (36.28%). The value of 20.57% is a simple average of provincial estimates and does not represent the survey-weighted national prevalence.

TABLE 1. Descriptive statistics of ARI prevalence and Google Trends search indicators (n = 34)

Variable	Mean	Median	SD	Minimum	Maximum
Prevalence of ARI (%)	20.57	18.81	6.29	11.45	36.28
Cough	80.62	84.00	11.06	61	100
Runny nose	60.79	60.50	19.91	29	100

Variable	Mean	Median	SD	Minimum	Maximum
Sore throat	72.50	72.50	14.14	51	100
ARI	63.62	60.50	17.17	38	100
Flu	50.26	46.00	16.63	25	100
Cough and cold	56.76	57.50	17.77	30	100
Shortness of breath	65.29	64.50	13.71	38	100

Examination of the prevalence distribution using the IQR criterion identified East Nusa Tenggara and Papua as extreme observations, with prevalences of 36.28% and 35.52%, respectively. Data examination did not indicate that these values were due to measurement, processing, or data entry errors. Both were retained in the main analysis, while sensitivity analyses were conducted on the 32 provinces after excluding these two regions. The geographic distribution of ARI prevalence is shown in Figure 2.



FIGURE 2. Distribution of ARI prevalence by province in Indonesia in 2023. The map shows the prevalence of ARI across 34 provincial units harmonized with Google Trends coverage.

3.3. Bivariate Relationships Between Search Indicators and ARI Prevalence. Pearson correlation coefficients between the seven Google Trends search indicators and ARI prevalence ranged from -0.048 to 0.328 (Table 2). The largest positive correlation was observed for batuk pilek (cough and cold; $r = 0.328$; 95% CI -0.011 to 0.600; $p = 0.058$), followed by pilek (runny nose; $r = 0.283$), batuk (cough; $r = 0.229$), and ISPA (ARI; $r = 0.227$). Sesak napas (shortness of breath) and sakit tenggorokan (sore throat) showed very small negative coefficients, at -0.048 and -0.031, respectively. All 95% confidence intervals included zero.

PREDICTIVE MODELS FOR ACUTE RESPIRATORY INFECTION IN INDONESIA

TABLE 2. Pearson correlation between Google Trends search indicators and ARI prevalence (n = 34)

Indicators	Pearson r	95% CI	p	FDR-adjusted p
Cough and cold	0.328	-0.011 to 0.600	0.058	0.344
Runny nose	0.283	-0.061 to 0.567	0.105	0.344
Cough	0.229	-0.118 to 0.527	0.192	0.344
ISPA (ARI)	0.227	-0.120 to 0.525	0.197	0.344
Flu	0.074	-0.271 to 0.402	0.676	0.861
Shortness of breath	-0.048	-0.380 to 0.295	0.788	0.861
Sore throat	-0.031	-0.365 to 0.310	0.861	0.861

After Benjamini-Hochberg correction, no search indicator reached statistical significance. Spearman correlation analysis showed a direction and ordering of associations that were generally similar to the Pearson results. For batuk pilek (cough and cold), the Spearman correlation was 0.327 ($p = 0.059$). Thus, there was insufficient statistical evidence to support the use of any single search indicator as an independent explanation for provincial variation in ARI prevalence. The correlation pattern and 95% confidence intervals are visualized in Figure 3.

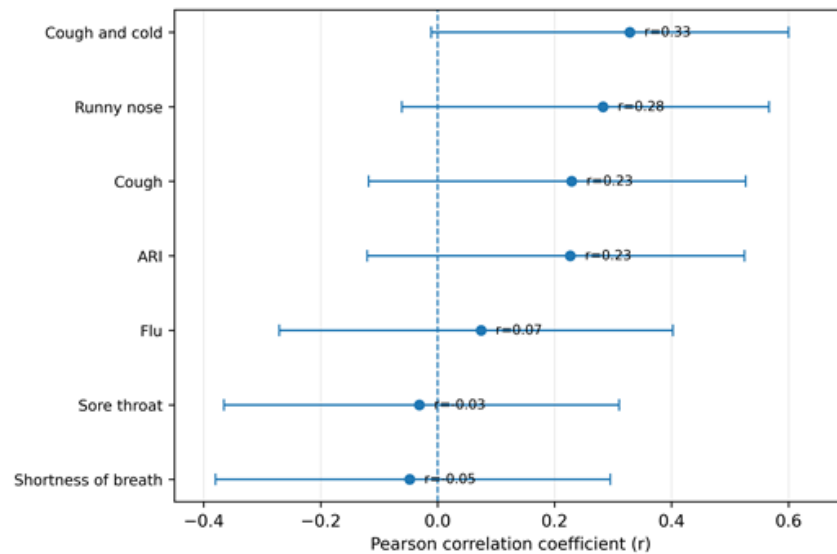


FIGURE 3. Pearson correlations between Google Trends search indicators and provincial ARI prevalence. Dots indicate Pearson correlation coefficients, and horizontal lines indicate 95% confidence intervals.

3.4. Interpredictor Correlations and Multicollinearity. Inter-predictor correlations showed very high overlap between pilek (runny nose) and batuk pilek (cough and cold), with a correlation coefficient of approximately 0.94. In the model including all seven indicators, the VIF values for pilek and batuk pilek were 12.27 and 11.64, respectively, in the main analysis, and 17.46 and 12.04, respectively, in the sensitivity analysis. In contrast, the combination of flu, batuk pilek, and sesak napas showed lower multicollinearity; in the sensitivity analysis, the VIFs of the three predictors ranged from 1.22 to 1.60. These findings suggest that the model including all indicators was susceptible to multicollinearity, whereas the three-indicator combination could be used without problematic multicollinearity. VIF results were not used to infer causal relationships. The interpredictor correlation structure is visualized in Figure 4.

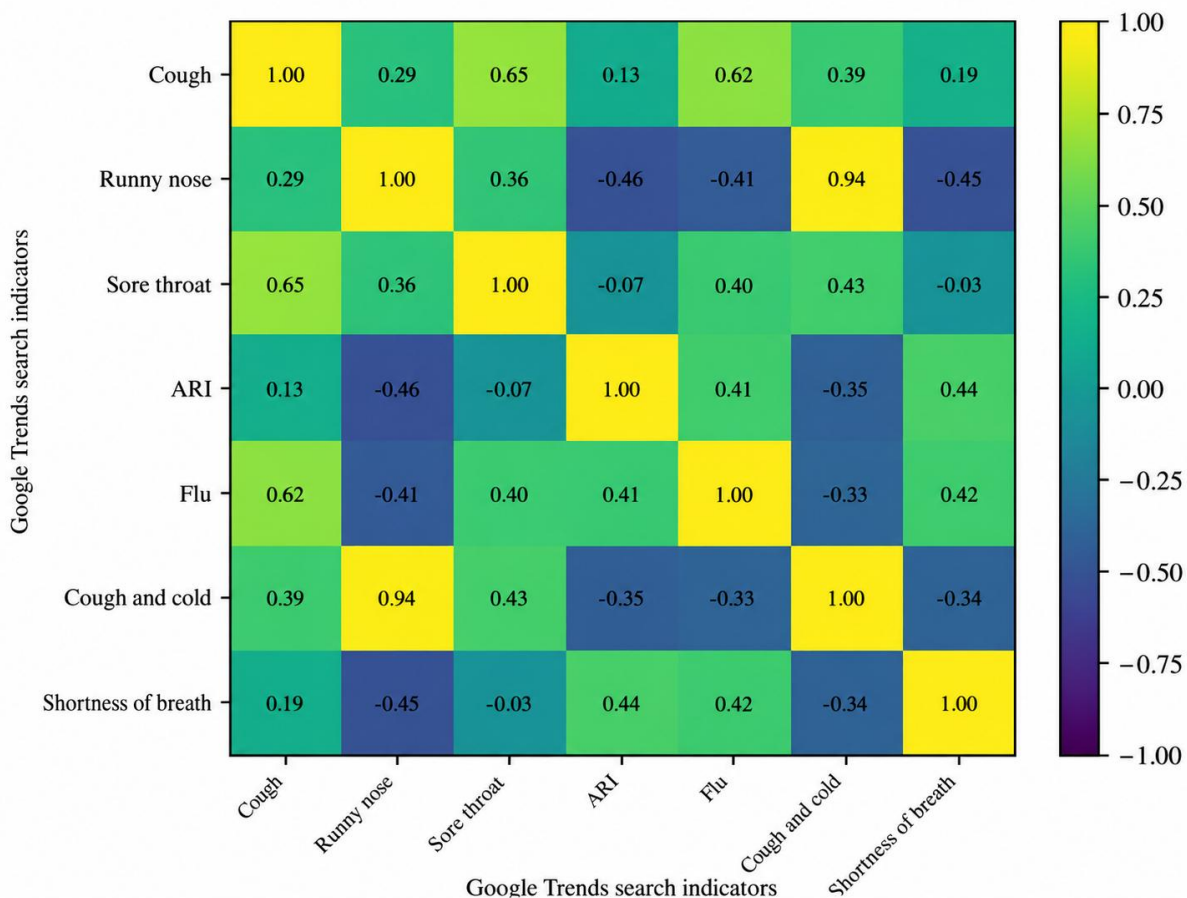


FIGURE 4. Pearson correlation matrix among the seven Google Trends search indicators in the main analysis ($n = 34$). The matrix highlights substantial redundancy between runny nose and cough and cold ($r \approx 0.94$).

3.5. Candidate Multivariate Models in the Main Analysis. Six candidate OLS models were compared in the main analysis of 34 provinces. The single-best-predictor model using only batuk pilek (cough and cold) produced the lowest LOOCV RMSE of 6.149, with a LOOCV MAE of 4.980 and a global out-of-fold R^2 of 0.017. The theory-driven model had an adjusted R^2 of 0.212, but its cross-validation error was larger. The all-keywords model produced the highest adjusted R^2 of 0.234, but had the largest LOOCV RMSE (7.502) and a global out-of-fold R^2 of -0.463. Overall, adding predictors did not consistently improve out-of-fold performance in the main analysis.

3.6. Benchmarking Statistical-Learning and Machine-Learning Models in the Main Analysis. Benchmarking across 34 provinces using three data-informed/exploratory predictors showed that none of the models produced a positive global out-of-fold R^2 . XGBoost produced the lowest RMSE among the compared models (6.326), with an MAE of 4.866 and a global out-of-fold R^2 of -0.040. Random Forest produced an RMSE of 6.491 and a global out-of-fold R^2 of -0.095. Linear regression, regularization models, and linear SVR produced RMSE values ranging from 6.964 to 7.103, with global out-of-fold R^2 values ranging from -0.261 to -0.312. Thus, in the main analysis of 34 provinces, there was no evidence of adequate generalization ability among the evaluated models.

3.7. Sensitivity Analysis Excluding East Nusa Tenggara and Papua. Sensitivity analysis was conducted by excluding East Nusa Tenggara and Papua, reducing the number of analytical units to 32 provinces. After these two provinces were excluded, the performance of several candidate models improved. The data-informed/exploratory model using flu, batuk pilek (cough and cold), and sesak napas (shortness of breath) produced an adjusted R^2 of 0.331, LOOCV RMSE of 4.288, LOOCV MAE of 3.455, and a global out-of-fold R^2 of 0.272. The all-keywords model had an adjusted R^2 of 0.316, but its cross-validation performance was lower (LOOCV RMSE 4.845; global out-of-fold R^2 0.070). These results indicate that the improved performance of the three-predictor model emerged only in the sensitivity analysis and cannot be assumed to apply unconditionally to all 34 provinces.

3.8. Model Benchmarking in the Sensitivity Analysis. Model benchmarking on the 32-province dataset was performed using the same three predictors. LASSO produced the best relative performance among the compared models, with a nested-CV RMSE of 4.233, a nested-CV MAE of 3.292, and a global out-of-fold R^2 of 0.290. Conventional linear regression produced an RMSE of 4.288 and a global out-of-fold R^2 of 0.272. Ridge and Elastic Net performed very closely, while the nonlinear models did not provide consistent performance improvements. Although LASSO had the best metrics in the sensitivity analysis, its improvement over OLS and other regularization models was modest. A comparison of model performance between the main and sensitivity analyses is presented in Figure 5.

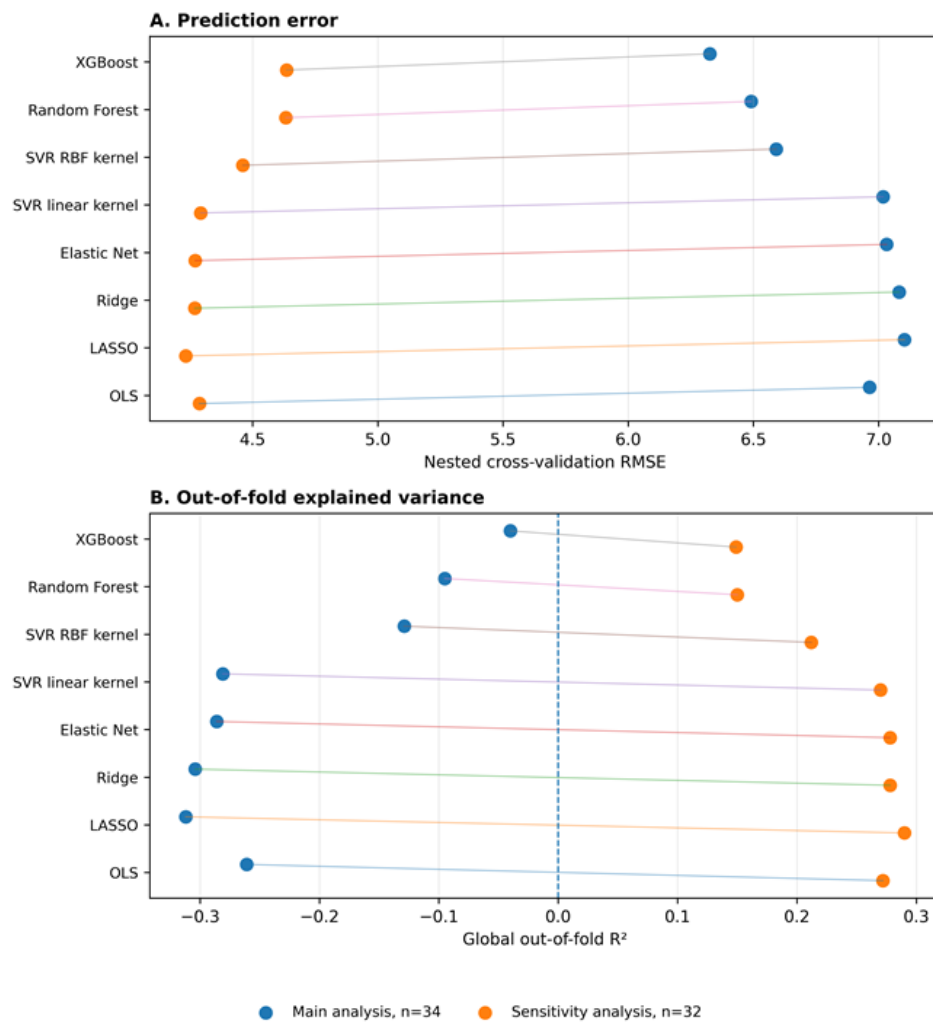


FIGURE 5. Comparison of model performance in the main and sensitivity analyses. Panel A shows nested cross-validation RMSE, while Panel B shows global out-of-fold R^2 .

3.9. Interpretation of the LASSO Model in the Sensitivity Analysis. LASSO was selected as the model with the best relative performance in the sensitivity analysis for 32 provinces, not as the final national model. When the model was refitted to the entire sensitivity dataset, the alpha value selected through 5-fold cross-validation was 0.3039. The resulting standardized coefficients were +2.372 for flu, +1.267 for cough and cold, and -1.987 for shortness of breath.

The interpretive model in the sensitivity analysis can be written as follows:

$$\hat{Y} = 19.614 + 2.372[(\text{flu} - 50.938)/16.485] + 1.267[(\text{batuk pilek} - 56.250)/17.922] - 1.987[(\text{sesak napas} - 64.406)/12.857].$$

Positive coefficients for flu and batuk pilek (cough and cold) indicated positive conditional associations with ARI prevalence after other predictors in the model were controlled. Conversely, the negative coefficient for sesak napas (shortness of breath) indicated a negative conditional association after flu and batuk pilek (cough and cold) were controlled. The direction of the coefficients should not be interpreted causally. Because East Nusa Tenggara and Papua were not included in the sensitivity model, the LASSO results cannot be directly generalized to these two regions or to all of Indonesia without additional validation. The direction and magnitude of the standardized coefficients are presented in Figure 6. Agreement between observed and out-of-fold predicted prevalence for the selected sensitivity model is shown in Figure 7.

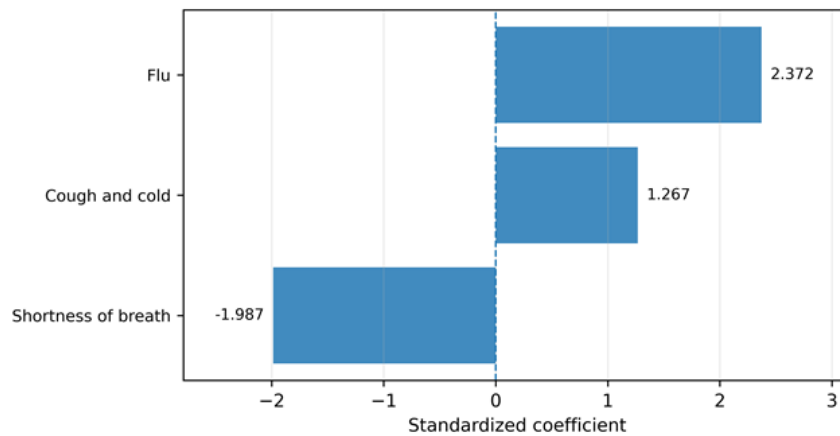


FIGURE 6. Standardized coefficients of the LASSO model in the sensitivity analysis. Coefficients were obtained from the LASSO model fitted to the 32-province sensitivity dataset and should not be interpreted causally.

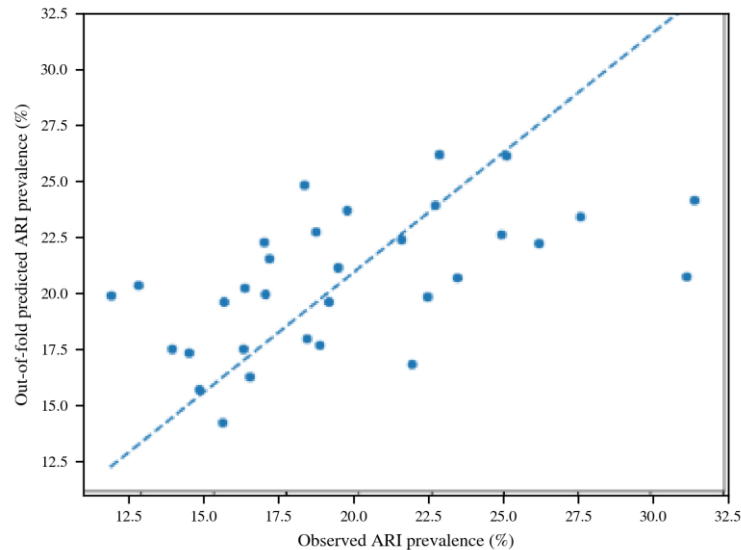


FIGURE 7. Observed versus out-of-fold predicted ARI prevalence for the LASSO model in the sensitivity analysis ($n = 32$). The diagonal line represents perfect agreement between observed and predicted prevalence; predictions were generated by the outer leave-one-out cross-validation procedure.

4. DISCUSSION

The main findings of this study indicate that Google Trends search signals do not yet show a strong, consistent relationship with ARI prevalence across provinces in Indonesia. In the primary analysis of 34 provinces, all bivariate correlations were weak, all confidence intervals encompassed zero, and no indicators remained significant after FDR correction. Modeling results reinforce these findings, as most models did not demonstrate adequate out-of-fold generalization ability. Therefore, the primary significance of these results is not that Google Trends can be used directly to estimate ARI prevalence, but rather that its utility at the provincial level is limited, context-dependent, and requires validation against official health data.

These findings expand the digital epidemiology literature on respiratory diseases, which has previously yielded mixed results. Several studies have reported that online search data can correlate with respiratory surveillance indicators, particularly when analyzed as a time series. Santillana et al. demonstrated that combining search data, social media, and traditional surveillance data can improve estimates of influenza activity [20]. Uda et al. also reported a strong correlation between the relative search volume of Google Trends and national respiratory syncytial virus

surveillance data in Japan [36]. However, other studies have shown that high performance on historical data is not always consistent. Olson et al. emphasized the need for caution in using Google Flu Trends for seasonal and pandemic influenza surveillance [19], while Lazer et al. pointed out the risks of overfitting, algorithmic changes, and overestimation in online search-based systems [37]. In this context, the results of this study align with the literature, which positions Google Trends as a supplementary signal rather than a direct measure of disease incidence.

The differences in the results of this study compared to some previous studies can be explained by differences in objectives and units of analysis. Many Google Trends studies on respiratory diseases utilize weekly or monthly variations for nowcasting, forecasting, or detecting trend changes [16, 17, 20, 36]. This study used a cross-sectional ecological design and assessed variations in survey-based prevalence over a single year. This design does not capture the temporal dynamics that are typically the key strength of Google Trends. Furthermore, Google Trends scores represent relative search interest, not the number of cases. These scores can be influenced by internet access, digital literacy, media attention, language preferences, healthcare-seeking behavior, and Google Trends' normalization mechanisms [13–15, 21]. These factors can weaken the relationship between digital searches and survey prevalence, particularly in countries with high geographic and social heterogeneity.

Model benchmark results also indicate that using more complex algorithms does not automatically improve predictive performance. In the main analysis, nonlinear models such as Random Forest and XGBoost did not produce a positive global out-of-fold R^2 . In the sensitivity analysis, LASSO performed relatively well, but its improvement over OLS, Ridge, and Elastic Net was small. This pattern suggests that the primary limitation is not the algorithm type per se, but rather the small aggregate sample size, regional heterogeneity, and the limited information conveyed by the search indicators. Therefore, the LASSO model in the sensitivity analysis is best understood as an exploratory finding that illustrates a possible signal in a subset of provinces, rather than as a final national model or evidence of causal relationships.

The primary contribution of this study is to demonstrate the limits of Google Trends for describing variations in survey-based ARI prevalence at the subnational level in developing

countries. This article does not prove that Google Trends is highly predictive, but rather demonstrates that the relationship between digital search behavior and official epidemiological indicators is not stable across regions. Consequently, Google Trends can be used to generate hypotheses, enrich analytical dashboards, or complement mosaic surveillance frameworks. Still, it is not sufficient as a sole basis for prevalence estimation, resource allocation, forecasting, or early warning systems.

Several limitations should be considered. The cross-sectional ecological design does not support individual-level, causal, or temporal inference. The sample of 34 provinces limits model capacity and increases uncertainty in performance estimates. The outcome was derived from prevalence reported in the 2023 SKI, but survey sampling uncertainty was not incorporated into the modeling framework. Google Trends provides a relative index that may vary across retrievals and is normalized separately for each keyword. Regional factors, including internet access, digital literacy, demographic composition, air quality, meteorology, population density, and access to health care, were not modeled. Future research should use temporal data, repeat Google Trends retrievals to assess index stability, and conduct spatial, temporal, and external validation before this approach is considered for operational use.

5. CONCLUSION

Google Trends search indicators did not demonstrate a strong or consistent relationship with acute respiratory infection (ARI) prevalence across Indonesian provinces. In the main analysis of 34 provinces, bivariate correlations were generally weak and were not statistically significant after FDR correction, while neither statistical-learning nor machine-learning models showed adequate out-of-fold generalization. A more promising signal emerged only in the 32-province sensitivity analysis, in which LASSO using flu, cough and cold, and shortness of breath achieved the best relative performance. Google Trends may therefore be considered a complementary and exploratory ecological signal, but it is not sufficiently robust to replace official health survey data, conventional surveillance systems, epidemiological forecasting, or early warning systems. Its use in public health decision-making requires stronger temporal, spatial, and external validation.

AUTHOR CONTRIBUTIONS

Ifa Stephane: Conceptualization, Methodology, Project administration, Formal analysis, Writing – original draft. Mahmud Isnani: Data curation, Methodology, Software, Formal analysis, Validation, Visualization, Writing – review & editing. Heru Saputra: Investigation, Data curation, Software, Visualization, Writing – review & editing. Dodi Devianto: Methodology, Supervision, Formal analysis, Validation. Ilham Tri Maulana: Formal analysis, Validation. Budi Rahmadya: Formal analysis, Validation.

DATA AVAILABILITY

The 2023 SKI data used in this study were obtained from the Ministry of Health of the Republic of Indonesia through an official data access request mechanism. The authors are unable to share these data freely because their use is subject to access approval and a Confidentiality Agreement. Requests for access to the 2023 SKI data can be submitted through the official data service platform of the Ministry of Health of the Republic of Indonesia (<https://layanandata.kemkes.go.id>). Data use was conducted in accordance with the approved data access procedure and Confidentiality Agreement; permission to use the data in this study is stated under number FRM/SMKI-PUSDATIN/70. Google Trends data used in this study are aggregate, region-based search interest data that can be retrieved using the search parameters described in the Methods section.

ACKNOWLEDGEMENTS

The authors gratefully acknowledge funding support from the Ministry of Higher Education, Science, and Technology of the Republic Indonesia (Kementerian Pendidikan Tinggi, Sains, dan Teknologi Republik Indonesia, Kemdiktisaintek) through the Basic Research/Inter-University Collaborative Research (PKPT) scheme under research contract Number: 289/C3/DT.05.00/PL_Baru/2026. The authors thank the Ministry of Health of the Republic of Indonesia, particularly the Center for Data and Information Technology, for facilitating access to the 2023 Indonesian Health Survey (SKI) data through the official data service platform. The authors also acknowledge Google Trends as the source of the aggregate search-interest data used in this study. All analyses, interpretations, and conclusions presented in this article are solely those of the authors and do not necessarily reflect the official views of the funding body or data providers.

CONFLICT OF INTERESTS

The authors declare that there is no conflict of interests.

REFERENCES

- [1] S.B. Sirota, M.C. Doxey, R.-M.V. Dominguez, R.G. Bender, A. Vongpradith, et al., Global, Regional, and National Burden of Upper Respiratory Infections and Otitis Media, 1990–2021: a Systematic Analysis from the Global Burden of Disease Study 2021, *Lancet Infect. Dis.* 25 (2025), 36-51. [https://doi.org/10.1016/S1473-3099\(24\)00430-4](https://doi.org/10.1016/S1473-3099(24)00430-4).
- [2] C. Chen, Y. You, Y. Du, W. Zhou, D. Jiang, et al., Global Epidemiological Trends in the Incidence and Deaths of Acute Respiratory Infections from 1990 to 2021, *Heliyon* 10 (2024), e35841. <https://doi.org/10.1016/j.heliyon.2024.e35841>.
- [3] C. Cui, T.T. Timbrook, C. Polacek, Z. Heins, N.A. Rosenthal, Disease Burden and High-Risk Populations for Complications in Patients with Acute Respiratory Infections: A Scoping Review, *Front. Med.* 11 (2024), 1325236. <https://doi.org/10.3389/fmed.2024.1325236>.
- [4] Badan Kebijakan Pembangunan Kesehatan, Survei Kesehatan Indonesia (SKI) dalam Angka, Kementerian Kesehatan Republik Indonesia, Jakarta, 2024. <https://repository.badankebijakan.kemkes.go.id/id/eprint/5539/>.
- [5] J.A. Mott, I. Bergeri, H.C. Lewis, A.W. Mounts, S.C. Briand, Facing the Future of Respiratory Virus Surveillance: “The Mosaic Surveillance Framework”, *Influenza Other Respir. Viruses* 17 (2023), e13122. <https://doi.org/10.1111/irv.13122>.
- [6] J. Fitzner, S. Qasmieh, A.W. Mounts, B. Alexander, T. Besselaar, et al., Revision of Clinical Case Definitions: Influenza-Like Illness and Severe Acute Respiratory Infection, *Bull. World Health Organ.* 96 (2018), 122-128. <https://doi.org/10.2471/BLT.17.194514>.
- [7] A.T. Aman, T. Wibawa, H. Kosasih, R.H. Asdie, I. Safitri, et al., Etiologies of Severe Acute Respiratory Infection (SARI) and Misdiagnosis of Influenza in Indonesia, 2013-2016, *Influenza Other Respir. Viruses* 15 (2021), 34-44. <https://doi.org/10.1111/irv.12781>.
- [8] M. Brady, R. Duffy, L. Domegan, A. Salmon, B. Maharjan, et al., Establishing Severe Acute Respiratory Infection (SARI) Surveillance in a Sentinel Hospital, Ireland, 2021 to 2022, *Eurosurveillance* 28 (2023), 2200740. <https://doi.org/10.2807/1560-7917.ES.2023.28.23.2200740>.

- [9] L. Marron, R. Duffy, J. O'Donnell, L. Domegan, An Evaluation of the Severe Acute Respiratory Infection Surveillance System in Ireland, *BMC Public Health* 25 (2025), 492. <https://doi.org/10.1186/s12889-025-21645-3>.
- [10] F.K. Lomholt, K.L. Møller, J. Nielsen, P. Valentiner-Branth, L.S. Vestergaard, Surveillance of Severe Acute Respiratory Infections Using ICD-10 Diagnosis Codes and National Electronic Health Records, Denmark, 2022 to 2024, *Eurosurveillance* 30 (2025), 2400801. <https://doi.org/10.2807/1560-7917.ES.2025.30.33.2400801>.
- [11] G. Eysenbach, Infodemiology and Infoveillance: Framework for an Emerging Set of Public Health Informatics Methods to Analyze Search, Communication and Publication Behavior on the Internet, *J. Med. Internet Res.* 11 (2009), e11. <https://doi.org/10.2196/jmir.1157>.
- [12] M. Salathé, L. Bengtsson, T.J. Bodnar, D.D. Brewer, J.S. Brownstein, et al., Digital Epidemiology, *PLoS Comput. Biol.* 8 (2012), e1002616. <https://doi.org/10.1371/journal.pcbi.1002616>.
- [13] A. Mavragani, G. Ochoa, Google Trends in Infodemiology and Infoveillance: Methodology Framework, *JMIR Public Health Surveill.* 5 (2019), e13439. <https://doi.org/10.2196/13439>.
- [14] S.V. Nuti, B. Wayda, I. Ranasinghe, S. Wang, R.P. Dreyer, et al., The Use of Google Trends in Health Care Research: A Systematic Review, *PLoS One* 9 (2014), e109583. <https://doi.org/10.1371/journal.pone.0109583>.
- [15] R. Alibudbud, Google Trends for Health Research: Its Advantages, Application, Methodological Considerations, and Limitations in Psychiatric and Mental Health Infodemiology, *Front. Big Data* 6 (2023), 1132764. <https://doi.org/10.3389/fdata.2023.1132764>.
- [16] O. Cai, B. Sousa-Pinto, United States Influenza Search Patterns Since the Emergence of COVID-19: Infodemiology Study, *JMIR Public Health Surveill.* 8 (2022), e32364. <https://doi.org/10.2196/32364>.
- [17] H.D. Ravkin, E. Yom-Tov, L. Nesher, The Effect of Nonpharmaceutical Interventions Implemented in Response to the COVID-19 Pandemic on Seasonal Respiratory Syncytial Virus: Analysis of Google Trends Data, *J. Med. Internet Res.* 24 (2022), e42781. <https://doi.org/10.2196/42781>.
- [18] J. Choi, Y. Cho, E. Shim, H. Woo, Web-Based Infectious Disease Surveillance Systems and Public Health Perspectives: A Systematic Review, *BMC Public Health* 16 (2016), 1238. <https://doi.org/10.1186/s12889-016-3893-0>.
- [19] D.R. Olson, K.J. Konty, M. Paladini, C. Viboud, L. Simonsen, Reassessing Google Flu Trends Data for Detection of Seasonal and Pandemic Influenza: A Comparative Epidemiological Study at Three Geographic Scales, *PLoS Comput. Biol.* 9 (2013), e1003256. <https://doi.org/10.1371/journal.pcbi.1003256>.

- [20] M. Santillana, A.T. Nguyen, M. Dredze, M.J. Paul, E.O. Nsoesie, et al., Combining Search, Social Media, and Traditional Data Sources to Improve Influenza Surveillance, *PLoS Comput. Biol.* 11 (2015), e1004513. <https://doi.org/10.1371/journal.pcbi.1004513>.
- [21] A. Rovetta, Reliability of Google Trends: Analysis of the Limits and Potential of Web Infoveillance During COVID-19 Pandemic and for Future Research, *Front. Res. Metr. Anal.* 6 (2021), 670226. <https://doi.org/10.3389/frma.2021.670226>.
- [22] H. McClymont, S.B. Lambert, I. Barr, S. Vardoulakis, H. Bambrick, et al., Internet-Based Surveillance Systems and Infectious Diseases Prediction: An Updated Review of the Last 10 Years and Lessons from the COVID-19 Pandemic, *J. Epidemiol. Glob. Health* 14 (2024), 645-657. <https://doi.org/10.1007/s44197-024-00272-y>.
- [23] I. Villanueva-Miranda, G. Xiao, Y. Xie, Artificial Intelligence in Early Warning Systems for Infectious Disease Surveillance: A Systematic Review, *Front. Public Health* 13 (2025), 1609615. <https://doi.org/10.3389/fpubh.2025.1609615>.
- [24] R. Qasrawi, G. Issa, S. Thwib, R. AbuGhoush, M. Amro, et al., The Role of Machine Learning in Infectious Disease Early Detection and Prediction in the MENA Region: A Systematic Review, *Inform. Med. Unlocked* 56 (2025), 101651. <https://doi.org/10.1016/j.imu.2025.101651>.
- [25] S. Schwartz, The Fallacy of the Ecological Fallacy: the Potential Misuse of a Concept and the Consequences, *Am. J. Public Health* 84 (1994), 819-824. <https://doi.org/10.2105/AJPH.84.5.819>.
- [26] T. Loney, N.J. Nagelkerke, The Individualistic Fallacy, Ecological Studies and Instrumental Variables: A Causal Interpretation, *Emerg. Themes Epidemiol.* 11 (2014), 18. <https://doi.org/10.1186/1742-7622-11-18>.
- [27] Y. Benjamini, Y. Hochberg, Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing, *J. R. Stat. Soc. Ser. B Stat. Methodol.* 57 (1995), 289-300. <https://doi.org/10.1111/j.2517-6161.1995.tb02031.x>.
- [28] R. Tibshirani, Regression Shrinkage and Selection via the Lasso, *J. R. Stat. Soc. Ser. B Stat. Methodol.* 58 (1996), 267-288. <https://doi.org/10.1111/j.2517-6161.1996.tb02080.x>.
- [29] A.E. Hoerl, R.W. Kennard, Ridge Regression: Biased Estimation for Nonorthogonal Problems, *Technometrics* 12 (1970), 55-67. <https://doi.org/10.1080/00401706.1970.10488634>.
- [30] H. Zou, T. Hastie, Regularization and Variable Selection via the Elastic Net, *J. R. Stat. Soc. Ser. B Stat. Methodol.* 67 (2005), 301-320. <https://doi.org/10.1111/j.1467-9868.2005.00503.x>.

- [31] A.J. Smola, B. Schölkopf, A Tutorial on Support Vector Regression, *Stat. Comput.* 14 (2004), 199-222.
<https://doi.org/10.1023/B:STCO.0000035301.49549.88>.
- [32] L. Breiman, Random Forests, *Mach. Learn.* 45 (2001), 5-32. <https://doi.org/10.1023/A:1010933404324>.
- [33] T. Chen, C. Guestrin, XGBoost: A Scalable Tree Boosting System, in: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ACM, New York, NY, USA, 2016, pp. 785-794. <https://doi.org/10.1145/2939672.2939785>.
- [34] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, et al., Scikit-learn: Machine Learning in Python, *J. Mach. Learn. Res.* 12 (2011), 2825-2830.
- [35] S. Varma, R. Simon, Bias in Error Estimation when Using Cross-Validation for Model Selection, *BMC Bioinformatics* 7 (2006), 91. <https://doi.org/10.1186/1471-2105-7-91>.
- [36] K. Uda, H. Hagiya, T. Yorifuji, T. Koyama, M. Tsuge, et al., Correlation between National Surveillance and Search Engine Query Data on Respiratory Syncytial Virus Infections in Japan, *BMC Public Health* 22 (2022), 1517. <https://doi.org/10.1186/s12889-022-13899-y>.
- [37] D. Lazer, R. Kennedy, G. King, A. Vespignani, The Parable of Google Flu: Traps in Big Data Analysis, *Science* 343 (2014), 1203-1205. <https://doi.org/10.1126/science.1248506>.