BMC Infectious Diseases

RESEARCH

Open Access

Investigating Google Trends to forecast acute febrile illness outbreaks in North India reported through the Integrated Disease Surveillance Program

Madhur Verma[1*], Kamal Kishore[2*], Pragyan Paramita Parija[3], Soumya Swaroop Sahoo[1], Dolly Gambhir[4], Usha Gupta[4] and Rakesh Kakkar[1]

Abstract

Background Acute Febrile Illness (AFI) like Malaria, Dengue, Chikungunya, and Enteric fever still remain the most common cause of seeking healthcare in low-middle-income countries and need to be constantly monitored for any impending outbreak. Digital epidemiology promises to assist traditional health surveillance. The health data (including AFI) collated by Google using specialised platforms like Google Trends (GT) is known to correlate with actual disease trends. The present study thus aims to assess the potential of GT to support routine surveillance system and forecast AFI outbreaks reported through the Indian Integrated Disease Surveillance Programme (IDSP).

Methods We utilised Haryana's IDSP portal to retrieve the weekly data of the most commonly reported infectious diseases causing AFI between 2011 and 2020. Internet search trends were downloaded using GT. Descriptive statistics estimated the burden of the AFI and Bland–Altman's plot depicted statistical agreement between the two. We adopted the Box-Jenkins approach to attain the final SARIMA model and explain the time-dependent weekly incidence of AFI.

Results The time series plot of the reported AFI displayed trends. Martin- Bland plots depicted acceptable agreement between two datasets for all Chikungunya and Dengue. Among the models evaluated, the Malaria model [SARIMA(1,1,1)(1,1,1)] demonstrated the best performance with a balanced fit and reasonable accuracy, while the Enteric Fever model [SARIMA(0,1,0)(1,1,1)] exhibited low prediction error but weak seasonal significance. In contrast, the Dengue [SARIMA(1,1,0)(1,1,0)] and Chikungunya [ARIMA(1,0,0)(0,0,0)] models had high forecast errors, limiting their predictive reliability. Overall, GT supplemented the prediction performance of the SARIMA models with adjusted R^2 of 46%, 50%, 50%, and 52% compared to the original 43%, 49%, 20%, and 48%.

Conclusions Our study observed modest improvements in GT-based SARIMA forecasting models compared to routine IDSP mechanisms for predicting AFI outbreaks in Haryana, highlighting the potential for further enhancement. As more granular GT data becomes available, its integration with traditional surveillance systems could significantly enhance forecasting accuracy for AFI and other infectious disease outbreaks. At no additional cost to the health

*Correspondence:
Madhur Verma
drmadhurverma@gmail.com
Kamal Kishore
kkishore.pgi@gmail.com
Full list of author information is available at the end of the article

system, GT can serve as a valuable, real-time digital epidemiology tool, strengthening public health preparedness and enabling timely interventions for the early containment of emerging diseases.

Keywords Digital health, Surveillance, Public health, Vector-Borne Diseases

Background

Acute Febrile Illness (AFI) still remains the most common cause of seeking healthcare in lower-middle income countries such as India and significantly contributes to premature morbidity and mortality [1]. The most common pathogens majorly contributing to the AFI burden include Malaria, Dengue, Enteric Fever, and Chikungunya [2]. In a tropical country, including India, factors like rapid changes in the environment, globalisation, frequent travel, migration, urbanisation, and encroachment of forest areas increase human-vector interaction, leading to sporadic outbreaks and increased incidence of AFI in the community [3]. The rising burden and frequency calls for concerted actions and necessitates early detection of outbreaks through robust surveillance [4, 5]. In India, conventional infectious disease surveillance is supported by the Integrated Disease Surveillance Program (IDSP)—a health program launched by the Ministry of Health and Family Welfare, New Delhi (MoHFW) in 2004 [6]. The IDSP collects and reports 22 notifiable epidemic-prone diseases at the district level for planning and executing pre-emptive and active outbreak threats. The program, however, had a few inadequacies, such as delayed reporting and data validity. To mitigate these deficiencies, IDSP was revised, and revamped as the Integrated Health Information Platform (IHIP) in 2021. The IHIP is a real-time, web-based data platform with Geographic Information system-enhanced data representation and analytical tools [7].

Despite addressing many inherent weaknesses of IDSP by IHIP, there is still a significant time lag between the first trigger of an outbreak and the response due to the existing workflow pattern [8, 9]. Therefore, there is a constant search for alternate outbreak monitoring systems for data triangulation that can supplement the IDSP in aggregating timely intelligence to proactively contain the disease outbreak. The rapid expansion of mobile phones and similar smart-device networks, along with increasing penetration of reliable internet connectivity, has revolutionised different facets of health, such as service delivery, health beneficiary tracking, data storage and dissemination [10]. This has empowered healthcare providers and the beneficiaries [11]. Within this context of digital health, digital epidemiology is an emerging field that uses technology to explore a population's disease patterns and health dynamics and supplement conventional field-based surveillance systems [12, 13]. Digitalisation is supported through internet-based platforms akin to Google, the most commonly used internet search engine globally. Besides providing solutions to user-initiated queries, Google also monitors health-related data, including infectious diseases [14]. This monitoring is more comprehensive through specialised services exemplified by Google Mobility Reports and Google Search Trends (GT) [15, 16].

Prior research has demonstrated that objectively measured attention by the GT approximates the actual infectious disease behaviour [17]. We previously ascertained a temporal correlation between the timings of infectious disease outbreaks as per IDSP reporting and GT in Haryana and Chandigarh, located in the northern region of India [18]. Subsequently, we hypothesise that such temporal correlations can help forecast the outbreaks in the community earlier than the existing monitoring mechanism, and GT can supplement IDSP by decreasing the lag between disease onset and disease reporting, assuming the stability of the study population and vector distribution within the study period. Therefore, with this background, we conducted the present study in Haryana with the primary aim of assessing the potential of GT to forecast AFI outbreaks due to Malaria, Dengue, Chikungunya, and Enteric fever that are reported through the IDSP.

Methods

Study design

Ecological study using secondary data analysis.

Data source

Disease surveillance data

The weekly frequency of cases collected between January 2011 and March 2020 under IDSP was obtained from Haryana's State Surveillance Units (SSU) functional under the Health Department, which was compiled to generate monthly estimates and match with the monthly format of the GT data. The IDSP integrates an IT-enabled disease surveillance system for epidemic-prone diseases supplemented by laboratory-based disease surveillance. It helps in real-time monitoring, detection and response to outbreaks. The IDSP data at the national level is managed by the Central Surveillance Unit (CSU), National Centre for Disease Control (NCDC), New Delhi (India). CSU receives its data from the country's different states and union territories through the SSU, which is supported by the District Surveillance Unit (DSU) functional at the district headquarters (Fig. 1).

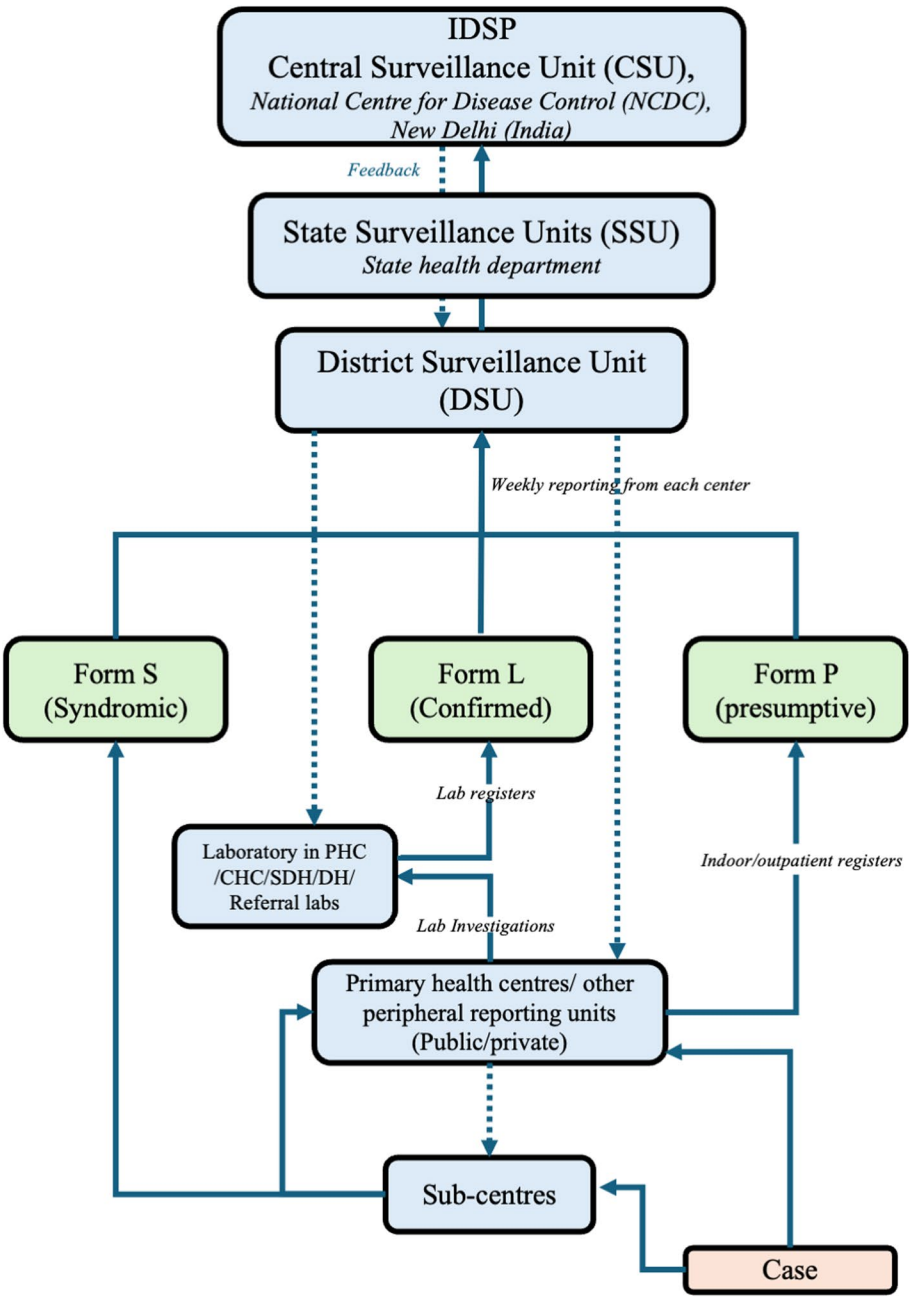

Fig. 1 Flow Chart depicting the flow of information in the Integrated Disease Surveillance Program in Haryana, India

- The front-line health Worker at the subcentre (which is the lowest level of the public health system) reports data using the 'S' (syndromic surveillance) forms every week to their medical officers stationed at the next higher level i.e. at the Primary Health (PHC) care centre, Community Health Centers, District Hospitals, Medical colleges, and private medical practitioners.
- The medical officers also report 22 diseases to the DSU using the 'P'(presumptive form) form based on clinical examination and making a presumptive diagnosis.
- Out of 22 diseases, the laboratories (public and private) report 12 laboratory-confirmed cases using "L" (Laboratory-based surveillance) forms. The cases identified between Monday to Sunday are compiled and forwarded to the reporting officer by the start of next week.

After verification and compilation at the DSU, the data are transmitted to CSU via SSU the following week. Of the 22 diseases reported through P form, we

used data concerning four major AFI (Malaria, Chikungunya, Dengue, and Enteric fever) for our study as they currently contribute the maximum to reported AFI cases leading to hospitalisation in our country [19]. The data were retrieved, processed, and analysed after obtaining ethical approval.

Google search trends data

Monthly Google search data were retrieved using Google Trends (GT) as per previously published methodology [18]. Google maintains a database of all searches stratified by duration and region on its GT platform, which can be downloaded for free [20]. The GT data are a sample of real-time (from the last seven days) and historical (from 2004 to 36 h before to search) Google search queries that have been appropriately anonymised, classified, and topic-tagged. For examining the relative popularity, each piece of data are divided by the total number of searches conducted in a particular region for a particular time period. Depending on the topic's share of searches across all topics, the GT represents search frequency output as a normalised data series at a scale of 0 to 100. Numbers correspond to the Google search query interest at a particular time and location.

First, the Google Trends tool was utilised to identify the most popular search term related to the concerned AFI to extract the search interest. The search parameters were set to "Haryana," "All Categories," and "Web search," and the time range was defined as "01st January 2011-31st March 2020." After comparative searches using search terms commonly used by the inhabitants of the state to identify a disease, the key terms that were the most relevant and widely used were "Malaria," "Dengue," "Typhoid", and "Chikungunya" (Supplementary Material 1), and were used as the final search terms for our study. Further, a list of the ten most correlated terms was also extracted for the 4 AFI to recheck if any other commonly used term was more popular on Google (Supplementary Material 2). Subsequently, using the selected search terms, Google monthly search hits were extracted and downloaded for the diseases as '.csv' files for additional analysis and interpretation (Supplementary Material 3).

Statistical analysis

Data collection and cleaning

The IDSP and GT weekly data points were merged using the month-year column. The merged file was cleaned, coded, and used for analysis using SPSS version 22. Descriptive statistics were used to estimate the burden of AFI and were depicted using absolute numbers, period prevalence of AFI per 100,000 population *([Period Prevalence = Total number of cases during the period/ projected population]*100,000).* Yearly changes in GT were depicted using median (minimum and maximum values) values. We used Haryana's 2011 census population as the base to calculate the projected population till 2020 using the decadal growth rate to have the yearly denominators as per standard guidelines [21]. The time series plot was used to assess the initial trend in the data. Subsequently, we reported Bland–Altman's plot to assess the agreement between GT and IDSP for each AFI [22]. The X-axis represents the mean of both methods for each time point calculated as *(Google Trends Value + Case Frequency)/2*. The Y-axis represents the difference between the two methods for each time point. The plots' dashed horizontal lines represent the 95% Limits of Agreement (LoA) calculated as (mean difference ± 1.96 × Standard Deviation of Differences). If most data points fall within the LoA, it suggests good agreement. However, we also calculated the Intraclass Correlation Coefficient (ICC) values to objectively measure agreement between GT and IDSP data that range from 0 to 1. The values between 0.0–0.39, 0.40–0.59, 0.60–0.74 and 0.75–1.00 depict poor, moderate, good and excellent agreement between the two variables.

We applied time series analysis to account for correlation in the data collected on equally spaced time order. First, we calculated cross-correlation to explore the possibilities of using GT lag to build our model, but it was insignificant and hence excluded. Subsequently, we investigated the possibility of building a time series model without using GT lag. We applied Seasonal Autoregressive Integrated Moving Average (SARIMA) and Autoregressive Integrated Moving Average (ARIMA) models to estimate the relationship between GT (predictor variable) and IDSP data (criterion variable). Differencing of order one was used to make stationary time series [23–25]. The parameter p is the order of autoregression, d is the difference, and q is the order of moving average, and P, D, Q, and S are the order of seasonal autoregression, seasonal difference, seasonal moving average, and seasonal length, respectively, of SARIMA (p,d,q) (P, D, Q) model. We adopted the Box-Jenkins approach for attaining the final SARIMA model in the following steps:

a. *Identification of the Model*: we made time series plots to detect non-stationarity in the data and used 1st order differencing to get stationary time series. Subsequently, we plotted PACF (Partial Autocorrelation Function) and ACF (Autocorrelation Function) to identify the MA (Moving Average) & AR(Autoregressive) components of the time series.

b. E*stimating Parameters:* we used indices such as AIC (Akaike information criterion), BIC (Bayes information criterion), and Ljung-Box to build and select the final model.
c. *Diagnostic Evaluation*: Finally, we evaluated the adequacy of the final models by making residual plots of ACF and PACF [26].

The models were evaluated using the BIC, Adjusted R square, and Mean Absolute Percentage Error (MAPE) indices. BIC is a likelihood-based criterion favouring a parsimonious model over a complex model by incorporating a penalty for adding a parameter in the model—the lower value of BIC is preferred. The utility of the model integrated with GT was compared with the original model, only including IDSP data using an adjusted R-square, and a higher value indicated a better model. MAPE values assessed the average magnitude of error the model produces and its ability to forecast. Normally, MAPE around 20% is considered a reasonably good value for the final model. We used a 2-tailed $p<0.05$ to declare the statistical significance.

Ethical statement

All experimental protocols were approved by the institutional ethics committee of the Post Graduate Institute of Medical Education and Research Chandigarh (India), through the letter number INT/IEC/2020/SPL-927, dated 24th July 2020. The waived informed consent to participate was approved by the ethics committee of the Post Graduate Institute of Medical Education and Research Chandigarh. The consent to use IDSP data were formally obtained from the Director General of Health Services, Department of Health, Government of Haryana wide letter number 32l3-IDSP/020–01, Dated 03rd Jan, 2020. All methods were carried out in accordance with relevant guidelines and regulations.

Results

In the study area, we first depicted the year-wise reported cases and period prevalence of four AFI between January 2011 and March 2020 (Table 1, Fig. 2). Malaria and Enteric fever contributed maximally of the four AFI, followed by Dengue and Chikungunya. Malaria prevalence was almost consistent across the years, except in 2019, with more than a 65% increase in reported cases. Dengue cases did not show any systematic increase or decrease till 2016; however, the cases increased meteorically by more than four times from 2017 onwards. Chikungunya prevalence in our study is significantly less than other AFI. The prevalence of Chikungunya was significantly higher in 2016 as compared to the preceding year; the same, however, was on a decreasing trend afterward. Enteric fever prevalence displayed an increasing trend from 2011 onwards.

We individually described the GT over the study period for all four AFI (Table 2). The time trends for Dengue, Malaria, Chikungunya, and Enteric fever demonstrated spikes in Google searches corresponding to the reported cases (Fig. 3a-d). Further, Fig. 4 displays Martin Bland plots to assess the agreement between the frequency of reported cases and GT. Most of the differences between IDSP and GT data were observed to be within the LoA for Chikungunya and Dengue. The ICC agreement between AFI frequencies and GT for Dengue, Malaria, Chikungunya, and Enteric fever were 0.51, 0.12, 0.65, and 0.25, respectively, implying that Chikungunya IDSP data depicted strongest agreement with GT data, followed by

Table 1 Year-wise trends of reported cases and period prevalence of diseases causing AFI in the state of Haryana (India) between 2011–20

Study year	Projected Population (Per census 2011)	Malaria		Dengue		Chikungunya		Enteric fever	
		Reported Number of cases	Period Prevalence/ 100,000	Reported Number of cases	Period Prevalence/ 100,000	Reported Number of cases	Period Prevalence/ 100,000	Reported Number of cases	Period Prevalence/ 100,000
2011	2,53,51,000	63,262	250	847	3.34	36	0.14	20,882	82.3
2012	2,57,72,000	128,271	498	2,818	10.93	21	0.08	31,914	123.8
2013	2,61,93,000	127,077	485	3,831	14.63	2	0.007	52,165	199.2
2014	2,66,75,000	114,795	430	1,928	7.23	10	0.03	81,928	307.2
2015	2,66,14,000	134,895	507	4,431	16.65	5	0.02	89,247	335.4
2016	2,74,55,000	119,083	434	2,964	10.80	2199	8.1	84,196	306.8
2017	2,78,61,000	118,130	424	12,455	44.70	106	0.38	125,304	449.7
2018	2,82,66,000	120,129	425	15,709	55.58	89	0.31	127,116	0.0
2019	2,86,72,000	203,404	709	15,082	52.60	21	0.07	128,259	447.3
2020	2,90,77,000	121,273	417	1,821	6.26	4	0.02	29,083	100.1

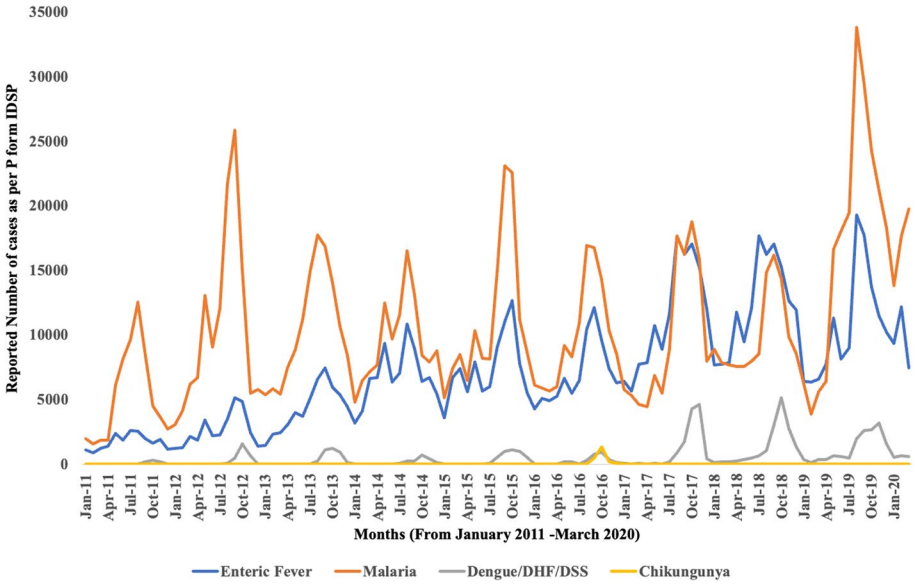

Fig. 2 Epi-curves depicting the frequency of cases of febrile illness in Haryana, India, from 2011 to 2020

Table 2 Year-wise description of the google trends for diseases causing acute febrile illnesses in Haryana (India) between 2011–20

Study Year	Malaria GT median, (min- max)	Dengue GT median, (min- max)	Chikungunya GT median, (min- max)	Enteric fever GT median, (min- max)
2011	17.5(0–55)	2(0–19)	0(0–4)	25.5(0–81)
2012	20.5(0–100)	2(0–38)	0(0–3)	17(0–71)
2013	20(11–54)	4.5(0–41)	0(0–3)	26.5(0–65)
2014	16(0–44)	3(0–18)	0(0–4)	43.5(0–88)
2015	23.5(0–71)	3.5(0–100)	0(0–2)	34.5(0–80)
2016	21(10–55)	4(0–36)	0.5(0–100)	36(16–100)
2017	18.5(10–34)	3.5(0–49)	1(0–6)	39(24–58)
2018	17.5(9–27)	3(0–29)	1(0–2)	43(25–89)
2019	19.5(11–34)	3(0–15)	1(0–1)	43(26–81)
2020Q1	16(12–20)	2(0–2)	1(0–1)	36(29–67)

Dengue, and Enteric fever, while Malaria depicted weakest agreement.

After initial descriptive analysis, we constructed a model for AFI prevalence data of IDSP and GT search volumes. The data depicted non-stationarity for Malaria, Dengue, and Enteric fever data, and we addressed the same with differencing of order 1(Fig. 5a-d). Models created using the Box-Jenkins method depicted that Malaria was best fitted using SARIMA $(1,1,1)$ $(1,1,1)_{12}$ model, Dengue was best fitted with SARIMA $(1,1,0)$ $(1,1,1)_{12}$ model, Enteric fever with SARIMA $(0,1,0)$ $(1,1,1)_{12}$, and Chikungunya with ARIMA $(0,0,1)$ $(0,0,0)_{12}$model compared to other models. The Malaria model [SARIMA(1,1,1)(1,1,1)] demonstrated a moderate fit (Adj-$R^2=0.46$), well-behaved residuals (Ljung-Box $p=0.39$), and a reasonable prediction error (MAPE=20.74%), indicating stable performance. For Dengue, the SARIMA(1,1,0)(1,1,0) model showed a slightly better fit (Adj-$R^2=0.51$) and a lower BIC (12.95), but its high MAPE (429.39%) and insignificant AR(1) coefficient ($p=0.80$) suggest weak predictive power. The Chikungunya model [ARIMA(1,0,0)(0,0,0)] exhibited the highest fit (Adj-$R^2=0.53$) and the lowest BIC (9.50), but its extremely high MAPE (749.79%) due to its sensitivity to outliers indicates poor forecasting accuracy. The Enteric Fever model [SARIMA(0,1,0)(1,1,1)] had a moderate fit (Adj-$R^2=0.52$) and good residual independence (Ljung-Box $p=0.78$), with a lower MAPE (19.11%), but SAR(1) was not statistically significant ($p=0.24$).

Fig. 3 Dual-axis plots displaying monthly febrile illness and Google trends for Haryana, India, from 2011 till 2020

The addition of the GT improved the prediction performance of the SARIMA models from adjusted R^2 of 46%, 50%, 50%, and 52% from the original 43%, 49%, 20%, and 48% (Table 3). The model performances were evaluated using ACF and PACF of residuals (plotting model values against actual values). Figure 5 displays that there are random variations from the origin zero (0), and hence, the fitted models are adequate. Overall, the Malaria model performed the best, followed by enteric fever. Dengue and Chikungunya were weak in enhancing the predictive accuracy.

Discussion

All four AFIs assessed in our study contribute significantly to the burden of communicable diseases, as reported by the IDSP in Haryana and neighbouring states

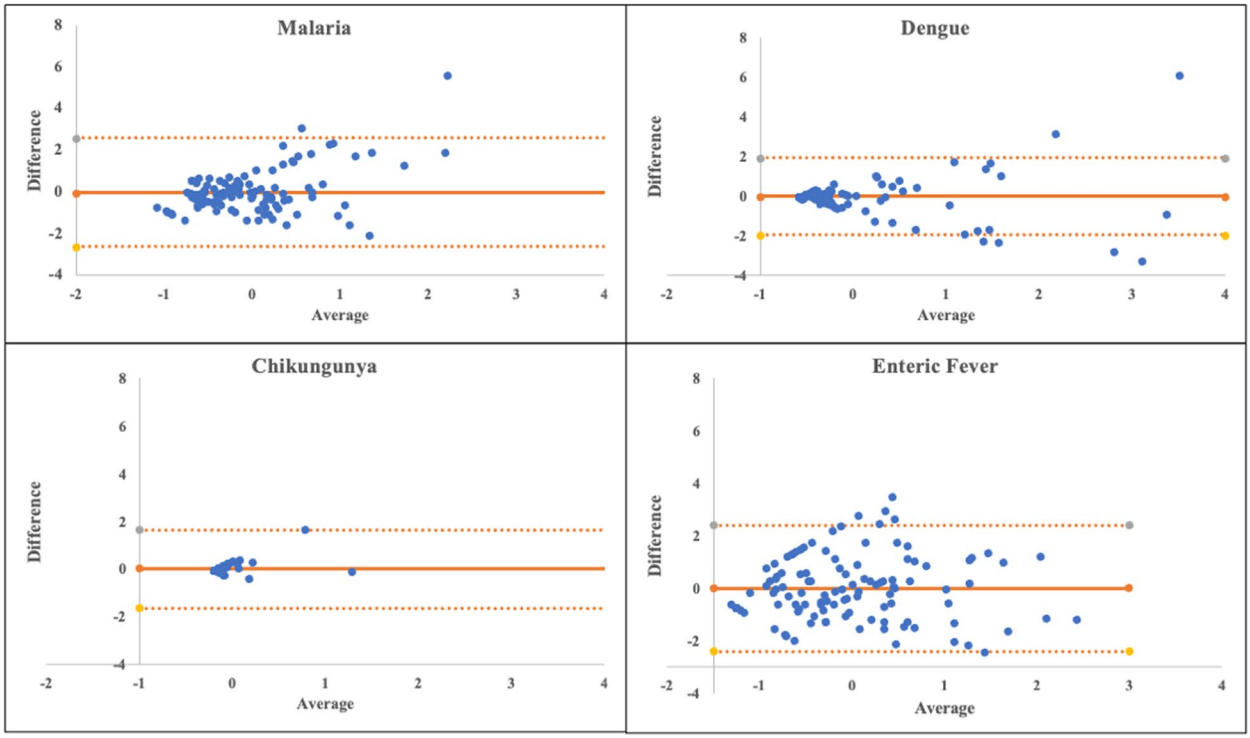

Fig. 4 Martin Bland plots depicting agreement between IDSP and Google trends data for acute febrile illnesses reported between 2011–20

[1, 18]. Since the burden of AFI in tropical regions is unlikely to decrease in the near future, an alternative surveillance method can augment a timely response [27, 28]. We used GT to enhance classical methods (rather than replace them) and help address their limitations. Thus, we assessed the utility of concatenating the data created through internet searches in improving routine epidemiological appraisals. We observed certain key findings. First, Malaria and Enteric fever depicted the highest burden in our study area. Second, Google trends corroborated well with the disease burden. Third, the GT-based forecasting model offered modest improvements over the IDSP-based model, and was a significant predictor for all the AFI except Dengue. Fourth, Enteric fever depicted only seasonal trends in contrast to non-seasonal autoregressive and moving average trends depicted by Chikungunya, which was also the simplest model and was dependent only on its previous values, while Malaria and Dengue depicted well using SARIMA models. Lastly, our GT integrated models demonstrated statistical robustness, with well-behaved residuals (Ljung-Box $p > 0.05$), moderate-to-good adjusted R^2 values higher compared to only IDSP-based models, and stable parameter estimates, indicating reliable fit within expected predictive boundaries, though further refinements are needed to enhance forecast accuracy (higher MAPE values), particularly for Dengue and Chikungunya. The trends in seasonality suggest the influence of seasonal and environmental factors. Our results demonstrate an inferential relationship between the GT search volumes and IDSP's AFI reporting in Haryana. Internet searches as per GT and the actual disease burden in the population can be coherent and concurrently point to a foreseeable outbreak. The media coverage also leads to inquisitiveness and awareness regarding the disease and can contribute to a spurt in internet search activity. Nevertheless, the GT can serve as a quasi-real-time infoveillance tool that quantifies the disease of interest per the user's web interest and can thus supplement the infectious disease outbreak prediction [29, 30]. Furthermore, ARIMA models help predict the temporal dependence between observations of IDSP data on GT using a time series.

Malaria has been a major public health problem in the state. There was a high burden throughout the study period, with frequent peaks coinciding with the rainy season in the state. The Best Model was for Malaria (Good model fit, reasonable MAPE, independent residuals, and significant parameters) depicted using SARIMA(1,1,1) $(1,1,1)_{12}$ model, that captures both long-term trends and seasonal patterns while keeping prediction errors under control. A previous study from South Africa depicted adjusted-R^2 from routine monitoring data-based SARIMA models to be around 0.41, which is very similar to our study [31]. Another study from Pakistan observed

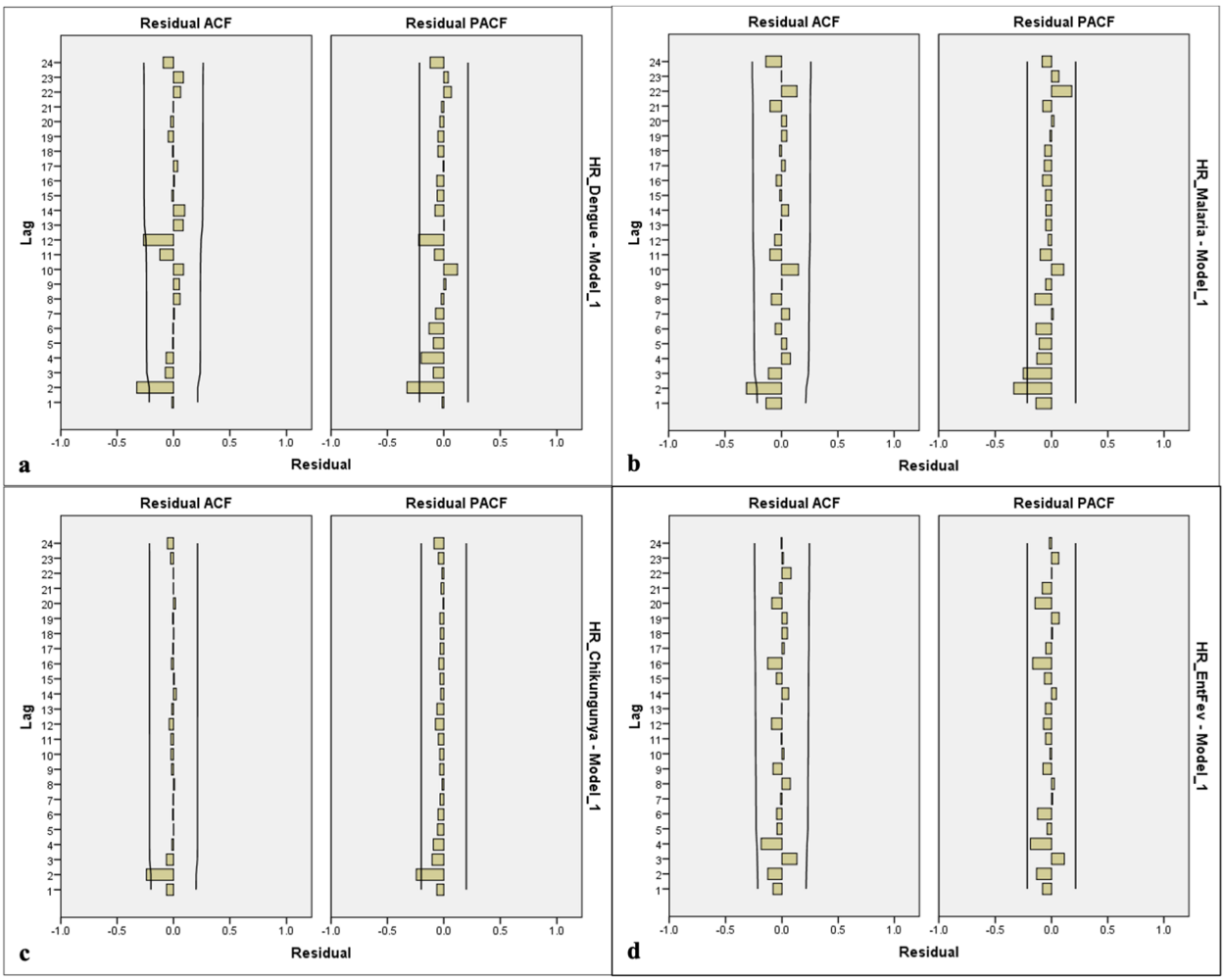

Fig. 5 Autocorrelation (ACF) and Partial autocorrelation function (PACF) of Acute febrile illnesses time series data

that SARIMA (2,1,2)(1,1,0)12 showed the best results [32]. In contrast, Wangdi et al. and Kumar et al. preferred assigning ARIMA(2,1,1)$(0,1,1)_{12}$ and ARIMA(0,1,1)$(0,1,0)_{12}$ to forecast Malaria in Bhutan and New Delhi [33, 34]. Further, the calculated BIC in the Bhutan study was 1812.4, with an error of −8.12%compared to BIC of 16.32 and 20.6% error in our study. A previous study from Indonesia offered SARIMA (0,1,1)$(1,1,1)_{12}$ as their final model. It achieved a comparable MAPE (21.6%) that echoed our findings to conclude that such models can make predictions based on the historical dataset [35].

We observed that Dengue depicted spikes similar to Malaria but were less intense. It can probably be attributed to the fact that the vector for Dengue (*Aedes aegypti* mosquito) can be controlled by the residents of the community as it dwells in standing water with a short range of flight, compared to anopheles mosquito (vector for Malaria) that dwells in ponds, has a longer range of flight, and its control is strongly influenced by public health measures implemented during the season. There is also evidence that herd immunity may limit the transmission of Dengue serotypes in some areas. We observed that Dengue GT positively correlated with the IDSP data. However, the strength of this relationship can vary based on factors such as public awareness, internet accessibility, and regional differences in information-seeking behavior [36]. Similar to our results, another longitudinal analysis from Indonesia depicted that the Dengue incidence and GT results are nonstationary processes ($P=0.01$). Still, their particular linear combination is a stationary process and allows us to construct error-corrected models. Hence, GT can be an initial indicator of upcoming Dengue outbreaks [35]. The Dengue model [SARIMA(1,1,0)(1,1,0)] captures both short-term and seasonal effects but demonstrated a moderate fit (Adj-$R^2=0.51$), a relatively low BIC (12.95), high prediction error (MAPE = 429.39%) and the insignificance of the autoregressive parameter ($p=0.80$) indicate poor predictive reliability,

Table 3 Time series models to forecast febrile illness outbreaks in Haryana using data (2011–20) from the integrated disease surveillance program and google trends

Disease	Model	Indices	Parameters	Significance
Malaria	SARIMA(1,1,1)(1,1,1)	Adj-R^2: 0.46 [0.43][a] BIC: 16.32 [16.31] MAPE: 20.74 [22.35] Ljung-Box: $p=0.39$	AR(1): −0.74 MA(1): −0.90 SAR(1): −0.26 MAR(1): 0.80 GRM: 23.49	< 0.001 < 0.001 = 0.11 = 0.008 = 0.007
Dengue	SARIMA(1,1,0)(1,1,0)	Adj-R^2: 0.51 [0.49] BIC: 12.95 [12.92] MAPE: 429.39 [205.69] Ljung-Box: $p=0.15$	AR(1): −0.03 SAR(1): −0.79 GRM: 2.86	< 0.80 < 0.001 = 0.12
Chikungunya	ARIMA(1,0,0)(0,0,0)	Adj-R^2: 0.53 [0.20] BIC: 9.50 [9.50] MAPE: 749.79 [749.79] Ljung-Box: $p=0.99$	AR(1): −0.25 GRM: 10.14	= 0.02 < 0.001
Enteric fever	SARIMA(0,1,0)(1,1,1)	Adj-R^2: 0.52 [0.48] BIC: 15.49 [15.50] MAPE: 19.11 [16.11] Ljung-Box: $p=0.78$	SAR(1): −0.29 MAR(1): 0.64 GRM: 19.67	= 0.24 = 0.01 = 0.02

Abbreviations: SARIMA Seasonal Auto-Regressive Integrated Moving Average Model, *ARIMA* Auto-Regressive Integrated Moving Average Model,*Adj-*R^2 Adjusted R-square value, *BIC* Bayes information criterion, *MAPE* Mean Absolute Percentage Error, *AR* Autoregressive, *MA* Moving Average, *SAR* Seasonal autoregressive, *MAR* Moving Autoregressive, *GRM* General Regression Models

[a] The figures in square parentheses depict the values of the model without including the data from Google trends

necessitating further refinement for improved forecasting accuracy. In another study from Brazil, the SARIMA $(2,1,2)(1,1,1)_{12}$ model was coherent with the Dengue incidence trends, supporting our findings [35]. However, studies from Venezuela, Singapore, and Bangladesh also depict that the GT-based estimation methods performed poorly in low transmission areas with unfavourable climates for Dengue transmission but performed satisfactorily during times of higher incidence, similar to our estimates [37, 38].

Chikungunya depicted a spike in the later stages of the study period. This is because itre-emerged in Haryana during the study period, and the population was particularly susceptible to infection. There was a strong correlation between Chikungunya and GT, similar to another study from Venezuela [37]. The best-fitted model for Chikungunya was the simplest [ARIMA(1,0,0)(0,0,0)]. It demonstrated the highest adjusted R^2 (0.53) among all models, suggesting a relatively good fit to the observed data. The model had the lowest BIC (9.50), indicating better parsimony. The Ljung-Box test ($p=0.99$) confirmed that the residuals were independent, implying that the model did not suffer from autocorrelation issues. The autoregressive terms were statistically significant, confirming their relevance in the model. While the model effectively captured past trends, the high MAPE suggests extremely poor forecast accuracy, and such a model may not be suitable for practical forecasting and requires further refinement, possibly by incorporating seasonal components or additional exogenous predictors to improve predictive accuracy. Our results reiterate that Chikungunya transmission in Haryana has transitioned similarly to other naïve geographical areas, making it more foreseeable and thus offering a targeted approach while designating our study area for further field trials to control outbreaks [35].

Enteric fever depicted a high burden across the study period, which is a cause of concern. The disease, however, depicted poor agreement with GT on Bland Altman's plot. Low agreement in regions where Enteric fever is endemic can be attributed to a general familiarity with the disease, leading to reduced online searches. Conversely, it can also be attributed to a lack of awareness, which could result in misinterpretation of symptoms, causing individuals to search for other conditions. The model [SARIMA(0,1,0)(1,1,1)], however, exhibited a moderate fit (Adjusted $R^2=0.52$), and a balanced trade-off between model complexity and goodness-of-fit (BIC value: 15.49). Still, the MAPE was among the lowest across all models, signifying better predictive accuracy. The non-significant Ljung-Box test confirms the model's reliability in terms of residual behaviour. However, the MAR, GRM (p-value < 0.05) and SAR (p-value > 0.05) coefficients suggest that seasonal patterns may not have a role in trends. Despite this limitation, the model remains a reliable forecasting tool due to its low error rate and well-behaved residuals, though further refinement of seasonal components can enhance predictive powers. Gao J et al. in China selected the SARIMA $(0,1,7)(1,0,1)_{12}$ model, which considers a large number of past error terms to make predictions and relies more on

past seasonal values and error correction, whereas our model was much simpler [39]. The seasonal component is vital for Enteric fever and played a significant role in our and Gao J et al. model. The disease incidence trend points towards seasonality factors (temperature, precipitation, humidity, etc.). There are other dynamic factors such as water hygiene which, however, could not be evaluated in our assessment. Our results suggest an increasing incidence of Enteric fever. Several factors contribute to the rising disease burden that can be controlled through improvements in WASH (Water, Sanitation, and Hygiene) measures to break the chain of transmission. GT data has also shown a significant correlation with the disease burden and can be used to supplement surveillance activities [40].

There are a few policy implications of this study. Entirely banking upon these data might be deceptive and lead to overestimation, similar to reasons that lead to discontinuing the Google flu trends [41]. So, instead of replacing the conventional model with "Big Data," we aimed to strengthen the original model. The substantial correlation between the reported monthly case numbers and the volume of Google searches does not imply that it makes the prediction more accurate. The SARIMA model has previously considered both time series' strong seasonality, which is mostly to blame. The SARIMA model could be helpful for modelling and forecasting monthly febrile illnesses. The SARIMA time series model was able to forecast AFI incidence with acceptable accuracy in our study area.GT data can help detect emerging AFI from vector-borne diseases in similar geographical terrains with logistic constraints.

The major strength of our study is the use of disease burden data collected through established protocols with decent acceptability in our country. The major limitation of this study is that we could not validate the model due to the COVID-19 pandemic disrupting the routine surveillance program due to restrictions imposed, and thus, we restrict our conclusions based on an assessment of the utility of GT. As we expect the availability of a more granular level of GT data in the near future, validation of the model can be taken up in future studies. The routine surveillance data were available only in public health sectors and excluded diagnoses made in private health facilities that constitute the bulk of medical services in the country. Further, Google search is influenced by various factors, including mobile phone availability, demography, and media activities, which can act as potential confounders. Media coverage can correlate to GT, as seen with different Public Health Emergencies of International concern. This can be seen as a limitation to using GT as there may be periods of inconsistent detection similar to rumour-surveillance**.** Another limitation is the diversity of search terms and languages within a region, posing inherent challenges in retrieving actual volume searches. All at once, no prediction can be sure as the past seldom reiterates itself in the future. Besides, forecasts are influenced by several factors, namely the data's reliability and psychological factors, such as people's perceptions and reactions to the hazards arising from the epidemic. Despite the perceived limitations of online search data, its usage for shaping public health policy and monitoring outbreaks and epidemics in recent times has gained momentum.

In conclusion, we report a high burden of acute febrile illnesses in Haryana (India). Most AFI studies from India are limited to trends or implementation research concerning preventive interventions and their effectiveness. Analysis conducted in the present study further such efforts by endeavouring to construct a predictive mechanism that can be deployed for AFI forecasting at the national and sub-national levels based on data retrieved through IDSP that is likely to improve with the introduction of the IHIP in the state of Haryana. GT is a surrogate indicator of health inquisitiveness and literacy, providing rich information to the scientific community. We observed modest improvements in the GT-based SARIMA forecasting model over routine surveillance mechanisms for the selected diseases causing AFI outbreaks in Haryana. In resource-constrained countries like India, GT can help improve public health preparedness for early containment of infectious diseases through timely response at no extra cost to the health system. Such a mechanism that can support forecasting mechanisms should be seen as an excellent forte for developing appropriate mitigation plans through adequate resource allocation. However, we expect that with the availability of more granular GT data in the future, it can serve as a valuable addition to the IHIP portal in IDSP for forecasting AFI and other infectious disease outbreaks, and the results are likely to be more robust and reliable. Future research should establish quasi-real-time infoveillance linkage by integrating GT-derived data corresponding to field data for purposive action in mitigating the effects of febrile illness outbreaks.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12879-025-10801-0.

Supplementary Material 1.

Acknowledgements
We thank the Department of Health and Family Welfare, Government of Haryana, for their constant support throughout this project.

Clinical Trial
Not applicable.

Authors' contributions
MV and KK conceptualised the study, conducted data analysis, interpreted the results, and prepared the first draft of the manuscript. PPP and SSS did the data curation and contributed to the draft of the manuscript. DG, UG, and RK helped in getting approvals from the state government and data acquisition and gave comments on the flow of the manuscript and the implications of the emerging recommendations.

Funding
The authors have not received any specific funding for this study.

Data availability
The disease-specific data were made available to the authors from the state IDSP cell of the health department following necessary approvals and can be shared by the corresponding author only upon reasonable request and receiving formal approval from the state government. The Google Trends data is available in the public domain and can be accessed from https://trends.google.com/trends/.

Declarations

Ethics approval and consent to participate
All experimental protocols were approved study by the institutional ethics committee of the Post Graduate Institute of Medical Education and Research Chandigarh (India), wide letter number INT/IEC/2020/SPL-927, dated 24th July, 2020. The consent to use IDSP data were formally obtained from the Director General of Health Services, Department of Health, Government of Haryana wide letter number 32l3-IDSP/020–01, Dated: 03rd Jan, 2020. All methods were carried out in accordance with relevant guidelines and regulations.

Consent for publication
Not applicable.

Competing interests
The authors declare no competing interests.

Author details
[1]Department of Community & Family Medicine, All India Institute of Medical Sciences Bathinda, Punjab 151001, India. [2]Department of Biostatistics, Post Graduate Institute of Medical Education and Research, Chandigarh 160012, India. [3]Department of Community Medicine, All India Institute of Medical Sciences Vijaypur, Jammu 184120, India. [4]Department of Health and Family Welfare, Government of Haryana, Sector-6, Panchkula, Haryana 134109, India.

Received: 3 October 2024 Accepted: 13 March 2025
Published online: 28 March 2025

References

1. Verma M, Panwar S, Sahoo SS, Grover GS, Aggarwal S, Tripathy JP, et al. Mapping the stability of febrile illness hotspots in Punjab from 2012 to 2019- a spatial clustering and regression analysis. BMC Public Health. 2023;23(1):2014.
2. Grundy BS, Houpt ER. Opportunities and challenges to accurate diagnosis and management of acute febrile illness in adults and adolescents: A review. Acta Trop. 2022;227:106286.
3. Carlson CJ, Bannon E, Mendenhall E, Newfield T, Bansal S. Rapid range shifts in African Anopheles mosquitoes over the last century. Biol Lett. 2023;19(2):20220365.
4. Morens DM, Daszak P, Markel H, Taubenberger JK. Pandemic COVID-19 Joins History's Pandemic Legion. MBio. 2020;11(3):e00812-20.
5. Martin R, Fall IS. Field Epidemiology Training Programs to accelerate public health workforce development and global health security. Int J Infect Dis. 2021;110:S3-5.
6. Ministry of Health and Family Welfare; Government of India. Integrated Disease Surveillance Programme(IDSP). [Cited 2022 Dec 26]. Available from: https://www.idsp.mohfw.gov.in/
7. Tanu T, Sagar V, Kumar D. IHIP – A leap into India's dream of digitalizing healthcare. Indian J Community Med. 2023;48(1):201.
8. Goel K, Chaudhuri S, Saxena A. India's strategy on surveillance system-A paradigm shift from an Integrated Disease Surveillance Programme (IDSP) to an Integrated Health Information Platform (IHIP). Clin Epidemiol Glob Heal. 2022;15:101030.
9. Kumar A, Goel M, Jain R, Khanna P. Tracking the implementation to identify gaps in integrated disease surveillance program in a block of district Jhajjar (Haryana). J Fam Med Prim Care. 2014;3(3):213.
10. DeSouza SI, Rashmi MR, Vasanthi AP, Joseph SM, Rodrigues R. Mobile phones: The next step towards healthcare delivery in rural India? PLoS One. 2014;9(8):e104895.
11. Thimbleby H. Technology and the future of healthcare. J Public health Res. 2013;2(e28):160–7.
12. Salathé M. Digital epidemiology: what is it, and where is it going? Life Sci Soc policy. 2018;14(1):1.
13. Park HA, Jung H, On J, Park SK, Kang H. Digital epidemiology: Use of digital data collected for non-epidemiological purposes in epidemiological studies. Healthc Inform Res. 2018;24(4):253–62.
14. Verma M, Goel S, Sinha P, Singh M, Upadhyay K. Interest in online tobacco cessation services during the COVID-19 pandemic in India-insights from google trends. Indian J Commun Med. 2024. https://doi.org/10.4103/ijcm.ijcm_265_23.
15. Kishore K, Jaswal V, Verma M, Koushal V. Exploring the Utility of Google Mobility Data During the COVID-19 Pandemic in India: Digital Epidemiological Analysis. JMIR Public Heal Surveill. 2021;7(8):e29957.
16. Broniatowski DA, Paul MJ, Dredze M. Twitter: Big data opportunities. Science 2014;345(6193):148–148.
17. Nuti S V., Wayda B, Ranasinghe I, Wang S, Dreyer RP, Chen SI, et al. The Use of Google Trends in Health Care Research: A Systematic Review. Voracek M, editor. PLoS One. 2014;9(10):e109583.
18. Verma M, Kishore K, Kumar M, Sondh AR, Aggarwal G, Kathirvel S. Google Search Trends Predicting Disease Outbreaks: An Analysis from India. Healthc Inform Res. 2018;24(4):300.
19. Abhilash KP, Jeevan J, Mitra S, Paul N, Murugan T, Rangaraj A, et al. Acute undifferentiated febrile illness in patients presenting to a Tertiary Care Hospital in South India: clinical spectrum and outcome. J Glob Infect Dis. 2016;8(4):147.
20. Google. Google trends [Internet]. 2022 [cited 2022 Dec 26]. Available from: https://trends.google.com/trends/?geo=IN
21. National Commission on Population, Ministry of Health and Family Welfare Directorate General of Health Services. Population projections for India and states 2011–2036 [Internet]. Report of the technical group on population projections. 2019 [cited 2022 Dec 20]. Available from: https://main.mohfw.gov.in/sites/default/files/Population Projection Report 2011–2036 - upload_compressed_0.pdf
22. Bland JM, Altman DG. Measuring agreement in method comparison studies. Stat Methods Med Res. 1999;8(2):135–60.
23. Midekisa A, Senay G, Henebry GM, Semuniguse P, Wimberly MC. Remote sensing-based time series models for malaria early warning in the highlands of Ethiopia. Malar J. 2012;11(1):1–10.
24. Amusa LB, Twinomurinzi H, Okonkwo CW. Modeling COVID-19 incidence with Google Trends. Front Res metrics Anal. 2022;15:7.
25. Wongkoon S, Jaroensutasinee M, Jaroensutasinee K. Assessing the temporal modelling for prediction of dengue infection in northern and northeastern. Thailand Trop Biomed. 2012;29(3):339–48.
26. Box GE, Jenkins G, Reinsel G, Ljung G. Time Series Analysis: Forecasting and Control. 5th ed. Balding DJ, Cressie NAC, Fitzmaurice GM, Givens GH, Goldstein H, Molenberghs G, et al., editors. New Jersey: Wiley; 2016.
27. Caminade C, McIntyre KM, Jones AE. Impact of recent and future climate change on vector-borne diseases. Ann N Y Acad Sci. 2019;1436(1):157.
28. Semenza JC, Suk JE. Vector-borne diseases and climate change: a European perspective. FEMS Microbiol Lett. 2018;365(2):fnx244.
29. Mavragani A, Ochoa G. Google Trends in Infodemiology and Infoveillance: Methodology Framework. JMIR public Heal Surveill. 2019;5(2):e13439.
30. Rovetta A. Reliability of Google Trends: Analysis of the Limits and Potential of Web Infoveillance During COVID-19 Pandemic and for Future Research. Front Res Metrics Anal. 2021;25(6):28.

31. Adeola AM, Botai JO, Mukarugwiza Olwoch J, De W. Rautenbach HCJ, Adisa OM, De Jager C, et al. Predicting malaria cases using remotely sensed environmental variables in Nkomazi, South Africa. Geospat Health. 2019;14(676):81–91.
32. Riaz M, Sial MH, Sharif S, Mehmood Q. Epidemiological Forecasting Models Using ARIMA, SARIMA, and Holt – Winter Multiplicative Approach for Pakistan. J Environment Public Health. 2023;2023(1):8907610.
33. Wangdi K, Singhasivanon P, Silawan T, Lawpoolsri S, White NJ, Kaewkungwal J. Development of temporal modelling for forecasting and prediction of malaria infections using time-series and ARIMAX analyses: A case study in endemic districts of Bhutan. Malar J. 2010;9(1):1–9.
34. Kumar V, Mangal A, Panesar S, Yadav G, Talwar R, Raut D, et al. Forecasting malaria cases using climatic factors in delhi, India: a time series analysis. Malar Res Treat. 2014;2014:482851.
35. Permanasari AE, Hidayah I, Bustoni IA. SARIMA (Seasonal ARIMA) implementation on time series to forecast the number of Malaria incidence. In: 2013 International Conference on Information Technology and Electrical Engineering (ICITEE). IEEE; 2013. p. 203–7.
36. Gluskin RT, Johansson MA, Santillana M, Brownstein JS. Evaluation of Internet-Based Dengue Query Data: Google Dengue Trends. PLoS Negl Trop Dis. 2014;8(2): e2713.
37. Strauss RA, Castro JS, Reintjes R, Torres JR. Google dengue trends: An indicator of epidemic behavior. The Venezuelan Case. Int J Med Inform. 2017;104:26–30.
38. Althouse BM, Ng YY, Cummings DAT. Prediction of Dengue Incidence Using Search Query Surveillance. Crockett RJK, editor. PLoS Negl Trop Dis. 2011;5(8):e1258.
39. Gao J, Li J, Wang M. Time series analysis of cumulative incidences of typhoid and paratyphoid fevers in China using both Grey and SARIMA models. PLoS One. 2020;15(10):e0241217.
40. Wang MY, Tang NJ. The correlation between Google trends and salmonellosis. BMC Public Health. 2021;21(1):1–11.
41. Kandula S, Shaman J. Reappraising the utility of Google Flu Trends. PLoS Comput Biol. 2019;15(8):1–16.

Publisher's Note