International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 12 Issue: 06 | Jun 2025 www.irjet.net p-ISSN: 2395-0072

Predicting COVID-19 Hotspots Using Google Trends: A Correlation Analysis of Search Terms and Case Data in India

Devyansh Garg [1]

[1]*Student, Grade 11, The Shri Ram School Moulsari, Gurugram, Haryana, India*

Abstract - *This study investigates the potential of Google Trends data as a predictive tool for identifying COVID-19 hotspots by analyzing the correlation between search term frequencies and confirmed case data across five Indian states during 2020-2021. Using Python-based correlation analysis, we examined 10 COVID-19-related keywords against daily and weekly case data from Andhra Pradesh, Delhi, Maharashtra, Uttar Pradesh, and Kerala. Our findings reveal distinct patterns between the pandemic's initial year and its second year: in 2020, terms like "Antibody" and "Loss of smell" showed the highest correlations, while 2021 demonstrated stronger positive correlations across most search terms. These results suggest that Google Trends data could serve as an early warning system for epidemic surveillance, though initial outbreak periods may introduce noise that affects predictive accuracy. This research contributes to the growing field of digital epidemiology and offers insights for public health preparedness strategies.*

***Key Words*: Google Trends, COVID-19, epidemiological surveillance, correlation analysis, India, digital epidemiology**

1.INTRODUCTION

1.1 Background

The COVID-19 pandemic has highlighted the critical importance of early detection and prediction systems for managing public health crises. As traditional epidemiological surveillance methods often lag behind real-time disease progression, researchers have increasingly turned to digital data sources for more timely insights. Among these, Google Trends has emerged as a particularly valuable tool for understanding public health concerns and potentially predicting disease outbreaks.

Previous research has established significant relationships between Google Trends data and COVID-19 case patterns across various geographical contexts. Ayyoubzadeh et al. (2020) analyzed correlations between Google Trends data for the term "COVID-19" and confirmed cases across eight countries, including the United States, Spain, Italy, France, the United Kingdom, China, Iran, and India, finding positive correlations across all regions. Similarly, Prasanth et al. (2021) employed multiple modeling approaches, including Linear, Negative Binomial, and Deep Neural Network models, to analyze relationships between Google Trends data for 13 keywords and confirmed COVID cases, with deep neural networks yielding the most accurate predictions.

Further studies have explored the nuances of these correlations. Ciufolini and Paolozzi (2020) conducted a comprehensive analysis across multiple countries from different continents, discovering that queries related to symptoms (particularly "fever" and "loss of smell") and terms explicitly mentioning COVID showed the highest correlations with case data. However, Cervellin et al. (2021) highlighted important methodological considerations, demonstrating that Google Trends data reliability varies significantly by geographical scale, proving most reliable for country-level analyses rather than regional comparisons.

Walker et al. (2020) examined lag and lead correlations between 10 keywords and COVID-19 cases across U.S. states, finding that terms like "face mask," "Lysol," and "COVID stimulus check" showed the strongest predictive correlations when analyzed with temporal offsets. Jokic et al. (2021) extended this work to Croatia, exploring both epidemiological predictions and socio-psychological consequences of the pandemic through search behavior. Notably, Sousa-Pinto et al. (2020) found that COVID-19-related Google Trends data often correlated more strongly with media coverage than with actual epidemic trends, particularly for terms like "anosmia" and "ageusia," highlighting the complex interplay between public awareness and disease prevalence.

1.2 Knowledge Gap

While existing research has demonstrated correlations between Google Trends data and COVID-19 cases, most studies have focused on either single time periods or limited geographical regions. A critical gap remains in understanding whether these correlations remain consistent across different phases of the pandemic and whether Google Trends data can reliably predict future epidemic hotspots, particularly in developing countries with diverse regional characteristics. This study addresses this gap by examining correlations across multiple Indian states over two distinct pandemic years.

1.3 Research Significance

The ability to predict epidemic hotspots using readily available digital data could revolutionize public health preparedness. Early identification of potential outbreak areas would enable healthcare systems to allocate resources more effectively, allowing hospitals to prepare equipment and personnel in advance. Additionally, governments could implement targeted interventions and preventive measures, potentially reducing disease transmission and saving lives. This research is particularly relevant for resource-constrained settings where traditional surveillance infrastructure may be limited.

1.4 Study Overview

This study employed Python-based correlation analysis to examine relationships between Google Trends search data and COVID-19 case numbers across five Indian states during 2020 and 2021. We analyzed both daily and weekly intervals to assess temporal granularity effects on correlation strength. Our methodology involved creating correlation heatmaps to visualize relationships between 10 carefully selected COVID-19-related keywords and active case data.

Our results reveal distinct patterns: in 2020, "Antibody" and "Loss of smell" showed the highest correlations with case data, while 2021 demonstrated broadly positive correlations across most search terms. These findings suggest that Google Trends data may serve as a valuable tool for epidemic prediction, though careful consideration must be given to the unique characteristics of initial outbreak periods.

2. METHODOLOGY

2.1 Research Aim

This study aimed to answer a fundamental question: Can Google Trends data be used to predict future epidemic hotspots? Specifically, we sought to determine whether search behavior patterns correlate with COVID-19 case data consistently enough to serve as an early warning system for disease outbreaks.

2.2 Data Collection

We collected and analyzed correlation data between Google Trends search frequencies and COVID-19 active case numbers across five Indian states: Andhra Pradesh, Delhi, Maharashtra, Uttar Pradesh, and Kerala. Data was collected for both 2020 and 2021 to capture different phases of the pandemic. To assess the impact of temporal granularity on correlation strength, we analyzed data at both daily and weekly intervals, creating correlation heatmaps to visualize relationships between variables.

2.3 Data Sources

2.3.1 Google Trends Data

We utilized Google Trends (https://trends.google.com/trends/) to collect search frequency data for the following 10 keywords:

1. Antibody - Selected to capture public interest in immunity testing
2. Coronavirus - General term for the virus
3. Coronavirus Symptoms - To track symptom-related searches
4. Coronavirus Vaccine - Monitoring vaccine-related interest
5. Covid-19 - Official disease designation
6. Covid Symptoms - Alternative symptom search term
7. Face Mask - Preventive measure searches
8. Fever - Primary COVID-19 symptom
9. Loss of Smell - Distinctive COVID-19 symptom
10. Combined symptoms (Sore Throat + Shortness of Breath + Fatigue + Cough) - Composite symptom searches

These keywords were selected based on their relevance to the COVID-19 pandemic and their popularity in search queries, ensuring comprehensive coverage of both disease awareness and symptom-related searches.

2.3.2 COVID-19 Case Data

COVID-19 case data was obtained from PRS India (https://prsindia.org/covid-19/cases), which provides comprehensive, state-wise daily updates on confirmed cases, active cases, recoveries, and deaths. This source was chosen for its reliability and consistent data reporting methodology across all Indian states.

2.4 Data Analysis

2.4.1 Data Processing

Data alignment was crucial for accurate correlation analysis. We synchronized dates between Google Trends and COVID-19 datasets, ensuring temporal consistency. For weekly interval analysis, we aggregated both COVID-19 case numbers and Google Trends search volumes using Python, calculating weekly sums to reduce daily fluctuations and reveal underlying trends.

2.4.2 Statistical Analysis

We employed Python for all data processing and analysis tasks, utilizing the following libraries:

- Pandas: For data manipulation and alignment
- NumPy: For numerical computations
- Matplotlib and Seaborn: For creating correlation heatmaps
- SciPy: For calculating correlation coefficients

Pearson correlation coefficients were calculated between each Google Trends keyword and active COVID-19 cases for each state, generating separate heatmaps for 2020 and 2021, as well as for daily and weekly intervals.

2.4.3 Rationale for Analytical Methods

Python was selected as our primary analytical tool due to its robust data science libraries and visualization capabilities. Correlation heatmaps were chosen for their ability to simultaneously display multiple variable relationships, making pattern identification intuitive and visually accessible. This approach facilitates quick identification of strong correlations and enables comparative analysis across different states and time periods.

3. RESULTS

3.1 Overall Patterns

3.1.1 Year 2020 Results

During the pandemic's first year, correlation patterns showed considerable variability:

- "Antibody" and "Loss of smell" consistently demonstrated the highest positive correlations across most states
- Initial pandemic months (March-April 2020) showed spike in search volumes that did not align with case numbers, likely due to heightened public awareness preceding actual case surges
- Daily and weekly interval analyses showed similar patterns, with weekly data providing slightly smoother correlation values

3.1.2 Year 2021 Results

The second year showed markedly different patterns:

- Most search terms exhibited strong positive correlations with COVID-19 cases
- Correlation values were generally higher and more consistent than in 2020
- The relationship between search behavior and actual cases appeared more synchronized

3.2 State-Specific Findings

3.2.1 Andhra Pradesh

- 2020: Moderate correlations, with "Antibody" ($r \approx 0.65$) and "Loss of smell" ($r \approx 0.58$) showing strongest relationships
- 2021: Strong positive correlations across all terms ($r > 0.7$ for most keywords)

3.2.2 Delhi

- 2020: Variable correlations, with symptom-related searches showing stronger relationships
- 2021: Consistently high correlations, particularly for "Covid-19" and "Coronavirus" terms

3.2.3 Maharashtra

- 2020: Similar to overall pattern, with "Antibody" searches showing highest correlation
- 2021: Very strong correlations across all search terms, suggesting synchronized public awareness and disease spread

3.2.4 Uttar Pradesh

- 2020: Moderate correlations with significant variability
- 2021: Improved correlation strength across all keywords

3.2.5 Kerala

- Notable exception to general patterns
- 2020: Weak or negative correlation for "Antibody" searches
- 2021: Relatively weaker positive correlations compared to other states
- Geographic location in southern India may contribute to different search behavior patterns

3.3 Temporal Interval Analysis

Comparison between daily and weekly intervals revealed:

- Weekly aggregations provided marginally more stable correlation values
- Overall patterns remained consistent regardless of temporal granularity
- Weekly analysis effectively reduced noise from day-to-day fluctuations

3.4 Correlation Heatmaps

Figure 1: Correlation Heatmap of Andhra Pradesh Daily (2020)

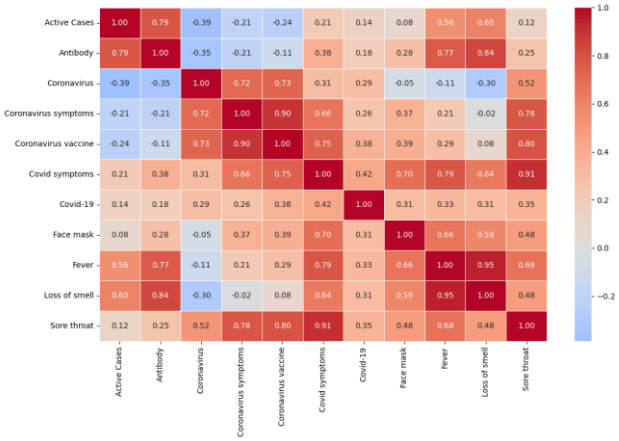


Figure 2: Correlation Heatmap of Andhra Pradesh Weekly (2020)

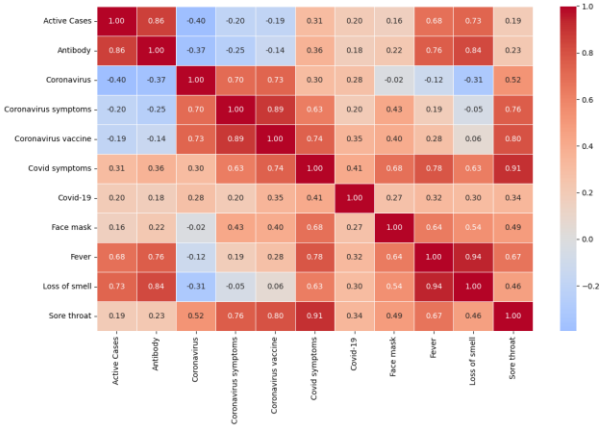


Figure 3: Correlation Heatmap of Andhra Pradesh Daily (2021)

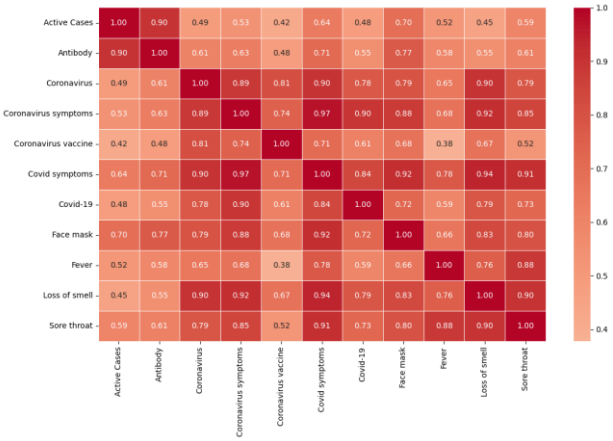


Figure 4: Correlation Heatmap of Andhra Pradesh Weekly (2021)

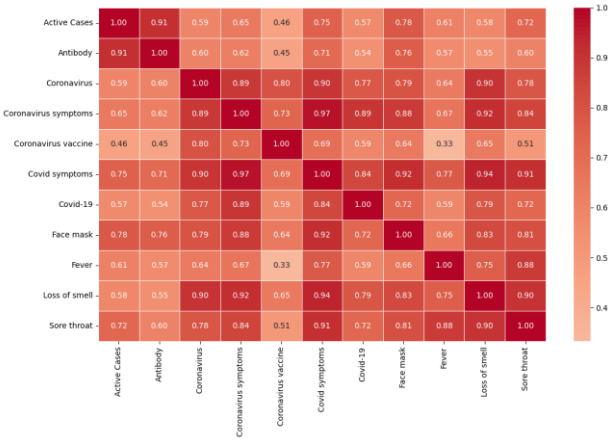


Figure 5: Correlation Heatmap of Delhi Daily (2020)

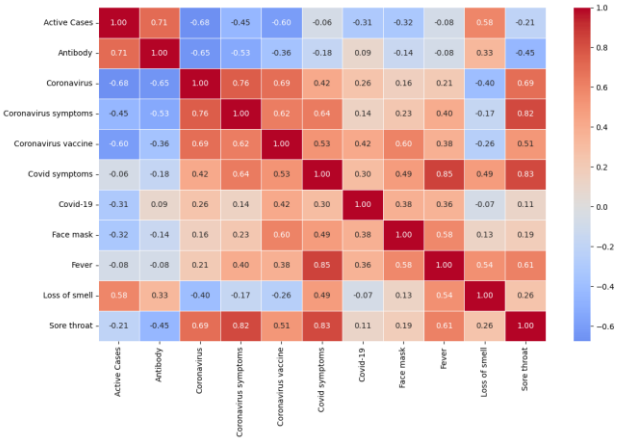


Figure 6: Correlation Heatmap of Delhi Weekly (2020)

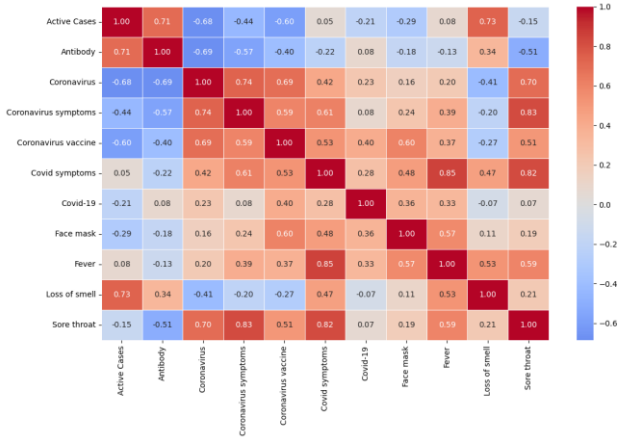


Figure 7: Correlation Heatmap of Delhi Daily (2021)

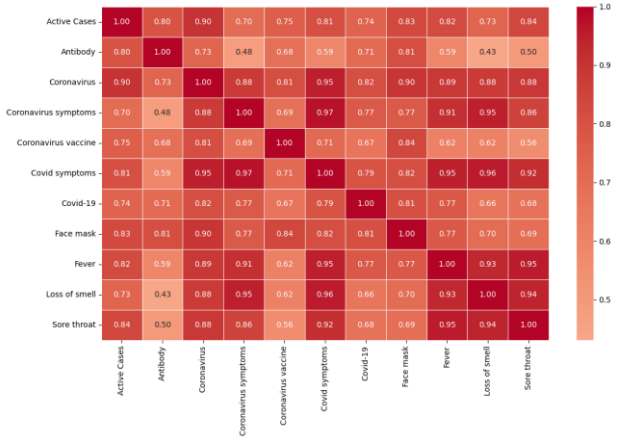


Figure 8: Correlation Heatmap of Delhi Weekly (2021)

Figure 9: Correlation Heatmap of Maharashtra Daily (2020)

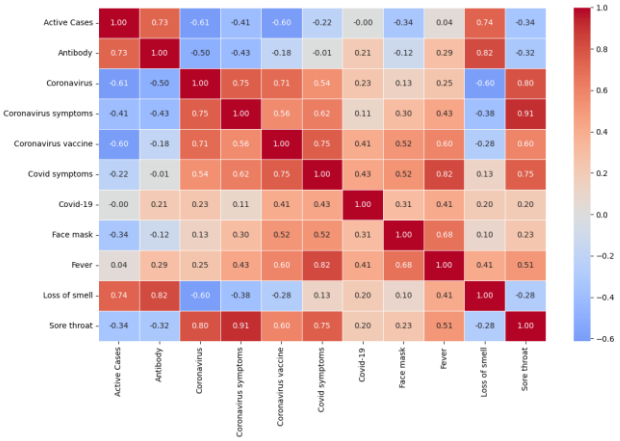


Figure 10: Correlation Heatmap of Maharashtra Weekly (2020)

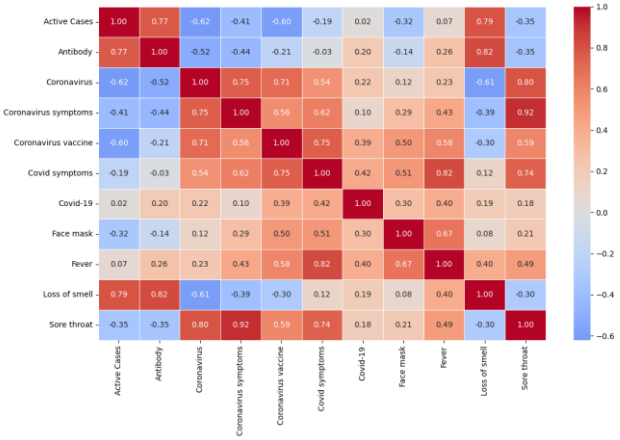


Figure 11: Correlation Heatmap of Maharashtra Daily (2021)

Figure 12: Correlation Heatmap of Maharashtra Weekly (2021)

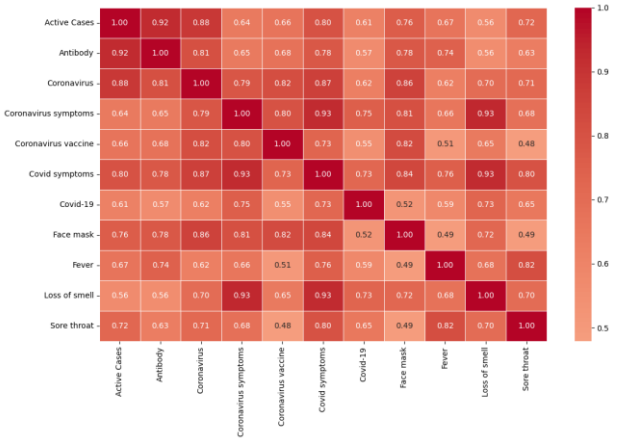


Figure 13: Correlation Heatmap of Uttar Pradesh Daily (2020)

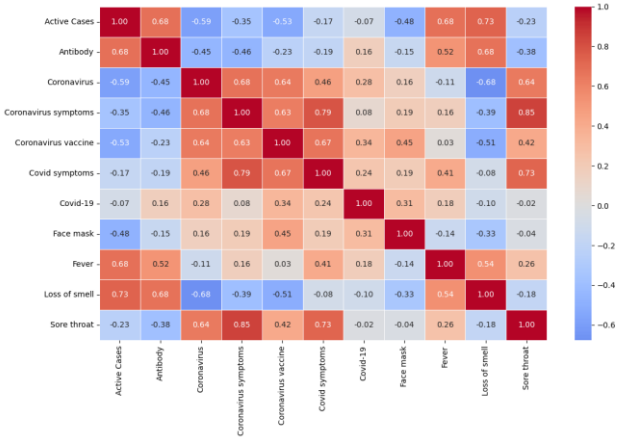


Figure 14: Correlation Heatmap of Uttar Pradesh Weekly (2020)

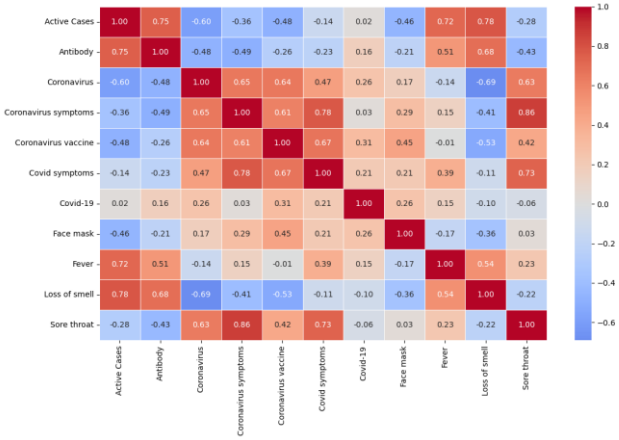


Figure 15: Correlation Heatmap of Uttar Pradesh Daily (2021)

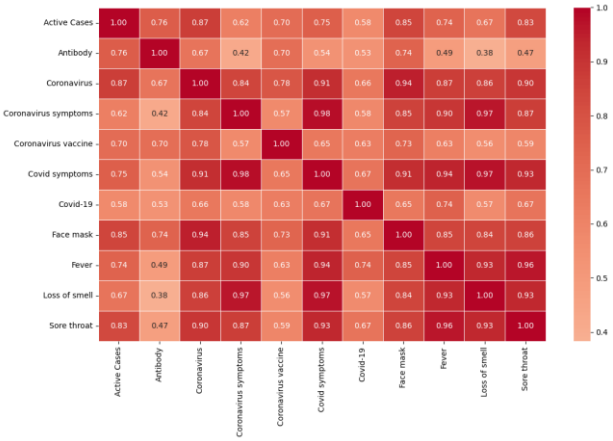


Figure 16: Correlation Heatmap of Uttar Pradesh Weekly (2021)

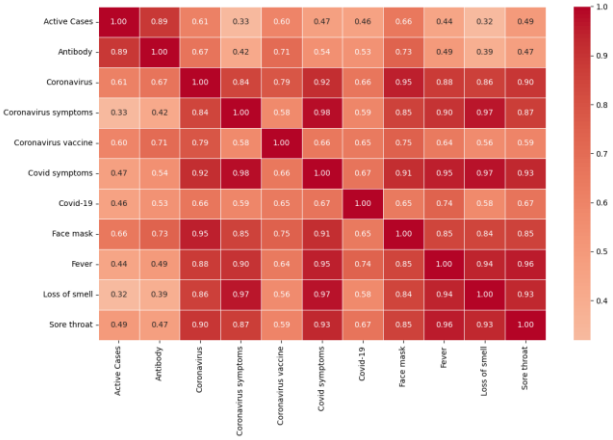


Figure 17: Correlation Heatmap of Kerala Daily (2020)

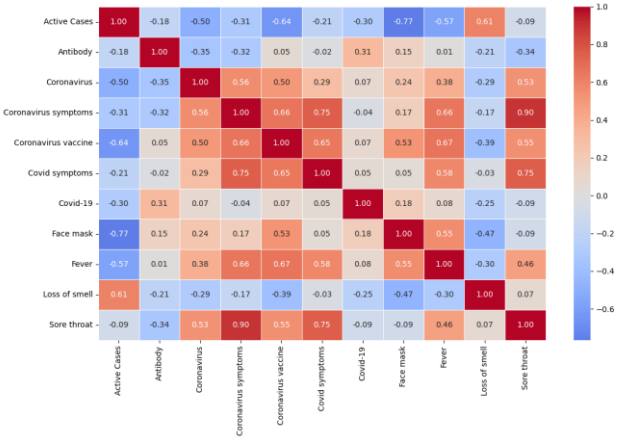


Figure 18: Correlation Heatmap of Kerala Weekly (2020)

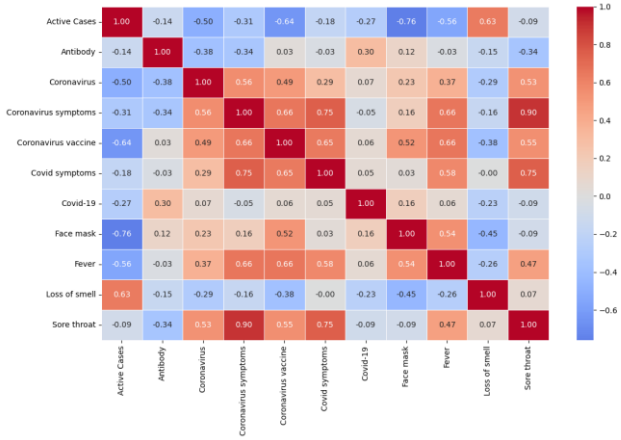


Figure 19: Correlation Heatmap of Kerala Daily (2021)

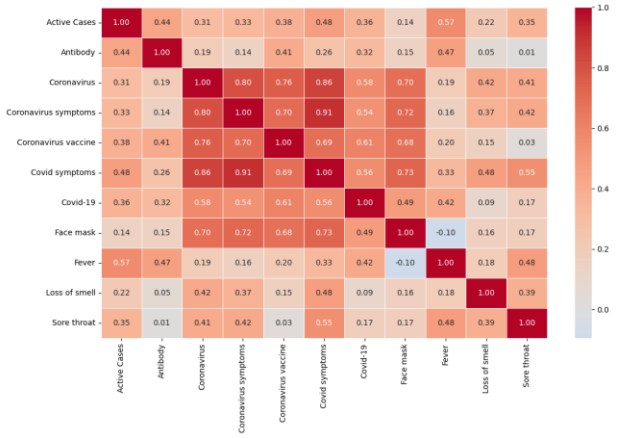
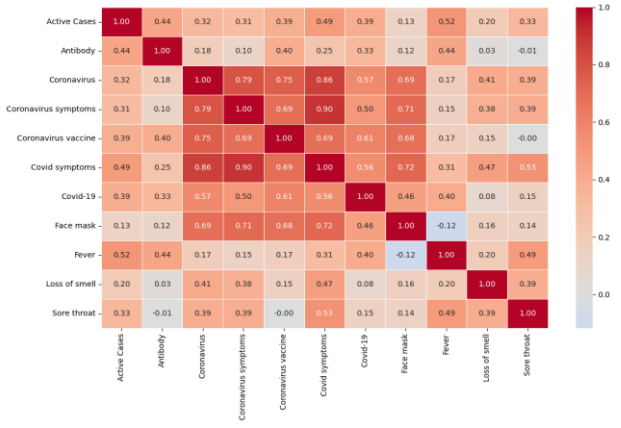


Figure 20: Correlation Heatmap of Kerala Weekly (2021)



4. DISCUSSION

4.1 Interpretation of Correlation Patterns

The observed correlation patterns suggest a complex relationship between public search behavior and COVID-19 case dynamics. The stronger correlations in 2021 compared to 2020 likely reflect several factors:

1. Learning Effect: By 2021, the public had developed a better understanding of COVID-19 symptoms and when to seek information, leading to more synchronized search patterns with actual case occurrences.

2. Media Influence: The initial 2020 surge in searches, particularly in March when COVID-19 was declared a pandemic, created a disconnect between search volumes and actual case numbers. This aligns with Sousa-Pinto et al.'s (2020) findings that Google Trends data often correlates more with media coverage than epidemic trends.

3. Symptom Awareness: The high correlations for "Antibody" and "Loss of smell" in 2020 suggest that as the pandemic progressed, people became more concerned with specific symptoms and testing, rather than general information about the virus.

4.2 Geographic Variation

Kerala's divergent pattern deserves special attention. Several factors may explain this anomaly:

- Higher Health Literacy: Kerala has India's highest literacy rate and well-developed healthcare infrastructure, potentially leading to different information-seeking behaviors

- Early Response: Kerala's proactive pandemic response may have altered the typical relationship between searches and cases

- Geographic Isolation: Being in southern India, Kerala may have experienced different disease transmission patterns

4.3 Limitations

Several limitations must be acknowledged in interpreting these results:

1. Initial Surge Bias: The March 2020 search surge, driven by global pandemic declaration rather than local cases, introduces noise particularly problematic for 2020 analyses. This highlights the importance of considering media influence when using Google Trends for epidemiological predictions.

2. Keyword Selection: While we selected popular and relevant terms, the choice of keywords inherently limits the scope of analysis. Other

relevant terms or local language searches were not included.

3. Data Granularity: Google Trends provides relative search volumes rather than absolute numbers, which may mask important variations in actual search behavior.

4. Confounding Variables: We did not account for factors such as internet penetration rates, demographic differences, or varying healthcare access across states, which could influence both search behavior and case reporting.

5. Outlier Management: We did not systematically remove outliers from our datasets, which might have improved correlation accuracy but could also risk losing important epidemic signals.

4.4 Implications for Public Health

Despite limitations, our findings have significant implications for public health practice:

1. Early Warning Systems: The strong correlations observed, particularly in 2021, suggest that Google Trends could complement traditional surveillance systems, providing early signals of increasing disease activity.

2. Resource Allocation: Hospitals and healthcare systems could monitor search trends to anticipate surges in patient loads, allowing proactive preparation of equipment and staffing.

3. Targeted Interventions: Governments could use search data to identify areas of increasing concern and implement targeted public health measures, including testing campaigns or movement restrictions.

4. Public Communication: Understanding search patterns can help health authorities tailor their communication strategies to address public concerns and information needs.

4.5 Recommendations for Implementation

To effectively utilize Google Trends for epidemic prediction:

1. Develop Composite Indices: Combine multiple search terms to create more robust predictive indicators

2. Account for Media Effects: Implement methods to filter out media-driven search spikes

3. Integrate Multiple Data Sources: Combine Google Trends with other digital data sources (social media, mobility data) for comprehensive surveillance

4. Establish Baselines: Create location-specific baseline search patterns to better identify anomalies

5. Real-time Monitoring: Develop automated systems to track and alert on significant changes in search patterns

4.6 Future Research Directions

This study opens several avenues for future investigation:

1. Expanded Geographic Scope: Analyze all Indian states to identify national patterns and regional variations

2. Advanced Modeling: Apply machine learning techniques, including neural networks and ensemble methods, to improve predictive accuracy

3. Multi-language Analysis: Include searches in regional languages to capture more comprehensive search behavior

4. Cross-pandemic Validation: Apply similar methodology to other epidemics (dengue, influenza) to validate the approach

5. Temporal Dynamics: Investigate optimal lag times between search queries and case occurrence for maximum predictive value

6. Demographic Segmentation: If available, analyze age and gender-specific search patterns to understand population-specific risks

5. CONCLUSIONS

This study demonstrates that Google Trends data holds promise as a supplementary tool for predicting COVID-19 hotspots, though its effectiveness varies by time period and geographic location. Our analysis of five Indian states over 2020-2021 revealed that search behavior patterns, particularly for terms like "Antibody" and "Loss of smell" in 2020, and most COVID-related terms in 2021, correlate significantly with active case numbers.

The stronger correlations observed in 2021 suggest that as public understanding of the pandemic matured, search behavior became more predictive of actual disease patterns. This finding underscores the potential for Google Trends to serve as an early warning system for epidemic

surveillance, particularly after initial outbreak periods when media-driven searches stabilize.

However, the geographic variations observed, particularly Kerala's divergent patterns, remind us that local factors significantly influence the relationship between search behavior and disease prevalence. Future epidemic preparedness strategies should consider integrating Google Trends data with traditional surveillance methods while accounting for local contexts and potential confounding factors.

As we face the possibility of future pandemics, developing robust digital epidemiology tools becomes increasingly critical. This research contributes to that goal by demonstrating both the potential and limitations of using search data for disease prediction. With further refinement and integration with other data sources, Google Trends could become a valuable component of comprehensive epidemic early warning systems, ultimately helping save lives through improved preparedness and response.

ACKNOWLEDGEMENT

I extend my sincere gratitude to Google Trends for providing open access to search data that made this analysis possible. I also thank PRS India for maintaining comprehensive COVID-19 case data throughout the pandemic, enabling researchers to conduct vital epidemiological studies. This research would not have been possible without these publicly available datasets and the organizations' commitment to data transparency.

REFERENCES

[1] Ayyoubzadeh, S. M., Ayyoubzadeh, S. M., Zahedi, H., Ahmadi, M., & Kalhori, S. R. N. (2020). Predicting COVID-19 incidence through analysis of Google Trends data in Iran: Data mining and deep learning pilot study. *JMIR Public Health and Surveillance*, 6(2), e18828. https://link.springer.com/article/10.1007/s10916-020-01588-5

[2] Cervellin, G., Comelli, I., & Lippi, G. (2021). Is Google Trends a reliable tool for digital epidemiology? Insights from different clinical settings. *Frontiers in Research Metrics and Analytics*, 6, 670226. https://www.frontiersin.org/journals/research-metrics-and-analytics/articles/10.3389/frma.2021.670226/full

[3] Ciufolini, I., & Paolozzi, A. (2020). An improved mathematical prediction of the time evolution of the Covid-19 pandemic in Italy, with a Monte Carlo simulation and error analyses. *International Journal of Environmental Research and Public Health*, 19(19), 12394. https://www.mdpi.com/1660-4601/19/19/12394

[4] Google Trends. (2021). *Search trends data*. https://trends.google.com/trends/

[5] Jokic, D., Turk, T., & Petrovic, M. (2021). Google Trends as a method to predict new COVID-19 cases and socio-psychological consequences of the pandemic. *EconStor*. https://www.econstor.eu/handle/10419/235602

[6] Prasanth, S., Singh, U., Kumar, A., Tikkiwal, V. A., & Chong, P. H. (2021). Forecasting spread of COVID-19 using Google Trends: A hybrid approach. *IEEE Access*, 9, 37785-37795. https://ieeexplore.ieee.org/abstract/document/9377852

[7] PRS India. (2021). *COVID-19 cases in India*. https://prsindia.org/covid-19/cases

[8] Sousa-Pinto, B., Anto, A., Czarlewski, W., Anto, J. M., Fonseca, J. A., & Bousquet, J. (2020). Assessment of the impact of media coverage on COVID-19-related Google Trends data: Infodemiology study. *Journal of Medical Internet Research*, 22(8), e19611. https://www.jmir.org/2020/8/e19611/

[9] Walker, A., Hopkins, C., & Surda, P. (2020). Use of Google Trends to investigate loss-of-smell-related searches during the COVID-19 outbreak. *Mayo Clinic Proceedings*, 95(9), 1904-1912. https://www.mayoclinicproceedings.org/article/S0025-6196(20)30934-4/fulltext