Modeling COVID-19 Wave Dynamics in India Using the SIR Model: A Data-Driven Epidemiological Analysis of First and Second Waves (2020–2021)

[1]shivam charole

[1]tpt

[1]shivamcharole22@gmail.com

Abstract—**The COVID-19 pandemic has posed unprecedented challenges to global public health systems, with India experiencing multiple waves characterized by distinct transmission dynamics and healthcare burdens. This study presents a comprehensive data-driven and epidemiological analysis of COVID-19 progression in India, with a particular focus on the comparative assessment of the first and second waves during the period 2020–2021. Utilizing publicly available daily time-series data on confirmed, recovered, and deceased cases, we examine temporal trends and quantify variations in infection spread and recovery patterns.**

To provide a theoretical and analytical framework, the classical Susceptible–Infected–Recovered (SIR) model is employed to simulate the transmission dynamics of the disease. Model parameters, including the transmission rate (β) and recovery rate (γ), are estimated from real-world data, enabling an evaluation of infection growth and decline across different phases of the pandemic. The integration of empirical data with the SIR framework allows for a deeper understanding of the underlying mechanisms governing wave formation and peak intensity.

The results reveal that the second wave of COVID-19 in India exhibited significantly higher transmission intensity, with daily confirmed cases surpassing 400,000 at its peak, compared to approximately 90,000 during the first wave. Additionally, the lag between infection and recovery rates during the second wave indicates increased strain on healthcare infrastructure. Active case analysis further highlights the severity of the second wave, reaching nearly 3.7 million active cases, substantially exceeding the first wave peak.

This study underscores the critical role of data-driven epidemiological modeling in capturing the dynamic behavior of infectious diseases. The findings provide valuable insights into the temporal evolution and severity of pandemic waves, which can inform public health strategies, resource allocation, and future outbreak preparedness. The proposed approach demonstrates the effectiveness of combining real-time data analysis with mathematical modeling to enhance the understanding of large-scale epidemic events.

I. Introduction

The emergence of the Coronavirus Disease 2019 (COVID-19) pandemic has profoundly impacted global health systems, economies, and societal structures (World Health Organization, 2020) ; (Johns Hopkins University, 2021) . Since its initial outbreak in late 2019, the rapid transmission of the causative virus, Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2), has led to multiple waves of infection across different regions of the world. Understanding the temporal dynamics and transmission behavior of such waves is crucial for designing effective public health interventions and mitigating future outbreaks.

India, as one of the most populous countries in the world, has experienced significant heterogeneity in the spread and intensity of COVID-19 infections (Our World in Data, 2021) (Johns Hopkins University, 2021) .The country witnessed a relatively gradual rise in cases during the first wave in 2020, followed by a sharp and devastating surge during the second wave in 2021. The second wave, in particular, was characterized by an unprecedented increase in daily confirmed cases, overwhelming healthcare infrastructure and highlighting the need for robust analytical frameworks to understand the underlying transmission dynamics.

Mathematical modeling has emerged as a fundamental tool in epidemiology for analyzing infectious disease spread and predicting future trends (Kermack & McKendrick, 1927) (Hethcote, 2000) . Among the various compartmental models, the Susceptible–Infected–Recovered (SIR) model provides a simplified yet powerful framework to describe the flow of individuals through different disease states. By incorporating key parameters such as the transmission rate (β) and recovery rate (γ), the SIR model enables the quantification of infection dynamics and the assessment of epidemic progression over time.

In recent years, the integration of real-world data with mathematical models has gained significant attention, enabling more accurate and data-driven insights into epidemic behavior (Kucharski et al., 2020) ; (Anderson & May, 1992) , enabling more accurate and data-driven insights into epidemic behavior. Time-series analysis of daily confirmed, recovered, and active cases allows for the identification of critical phases such as growth, peak, and decline of infection waves. When combined with the SIR modeling framework, such analyses can provide a deeper understanding of the factors influencing wave severity, transmission intensity, and healthcare burden.

The present study aims to perform a comprehensive data-driven and epidemiological analysis of COVID-19 waves in India during the period 2020–2021. Specifically, this research focuses on (i) analyzing temporal trends in daily confirmed, recovered, and active cases, (ii) comparing the characteristics of the first and second waves, and (iii) applying the SIR model to investigate transmission and recovery dynamics. By bridging empirical data analysis with theoretical modeling, this study seeks to provide meaningful insights into the evolution of pandemic waves and contribute to the broader understanding of infectious disease dynamics.

The findings of this research are expected to support public health decision-making by highlighting key differences between pandemic waves and emphasizing the importance of timely interventions. Furthermore, the methodological approach adopted in this study can serve as a framework for analyzing future epidemic outbreaks using data-driven epidemiological models.

II. METHODOLOGY

1. Data Collection

The present study is based on publicly available, real-time COVID-19 data for India covering the period from January 2020 to mid-2021. The dataset includes daily records of confirmed, recovered, and deceased cases, which are widely used for epidemiological analysis. The data were obtained from reliable open-access repositories such as covid19india.org and other verified aggregators to ensure accuracy and consistency (covid19india.org, 2021)

2. Data Preprocessing

Prior to analysis, the raw dataset was subjected to systematic preprocessing to ensure data quality and usability. The following steps were performed:

- **Date Standardization:** All date entries were converted into a uniform YYYY-MM-DD format to enable consistent time-series analysis.
- **Handling Missing Values:** Any missing or inconsistent entries were identified and treated using interpolation or removal techniques, depending on their frequency and impact.
- **Derivation of Daily Cases:** Daily confirmed, recovered, and deceased cases were computed from cumulative data where necessary.
- **Calculation of Active Cases:** Active cases were calculated using the relation:

$$Active\ Cases = Total\ Confirmed - Total\ Recovered - Total\ Deceased$$

- **Smoothing (Optional):** A 7-day moving average was applied to reduce short-term fluctuations and highlight underlying trends.

3. Exploratory Data Analysis

Exploratory analysis was conducted to understand the temporal evolution of COVID-19 in India. Time-series plots were generated to visualize:

- Daily confirmed and recovered cases
- Active case trends over time

These visualizations were used to identify key phases of the pandemic, including growth periods, peak infection points, and decline phases corresponding to the first and second waves.

4. SIR Model Framework

To model the transmission dynamics of COVID-19, the classical Susceptible–Infected–Recovered (SIR) model was employed (Kermack & McKendrick, 1927) . The total population $N$is divided into three compartments:

- **Susceptible (S):** Individuals who are at risk of infection
- **Infected (I):** Individuals currently infected and capable of transmitting the disease
- **Recovered (R):** Individuals who have recovered or died and are no longer infectious

The model is governed by the following system of differential equations:

$$\frac{dS}{dt} =- \beta SI, \frac{dI}{dt} = \beta SI - \gamma I, \frac{dR}{dt} = \gamma I$$

where:

- **β (beta)** represents the transmission rate
- **γ (gamma)** represents the recovery rate

5. Parameter Estimation

The key parameters of the SIR model were estimated using observed data based on standard epidemiological modeling approaches (Kucharski et al., 2020)

- **Transmission Rate (β):** Estimated based on the rate of increase in daily confirmed cases during the exponential growth phase.
- **Recovery Rate (γ):** Approximated as the inverse of the average infectious period, calculated using recovery data trends.

The ratio $R_0 = \frac{\beta}{\gamma}$, known as the basic reproduction number, was used to assess the transmissibility of the virus during different phases of the pandemic.

6. Model Implementation

The SIR model was implemented using computational tools such as Python (NumPy, Pandas, and Matplotlib) or spreadsheet-based methods. The initial conditions for the model were defined as:

- Initial susceptible population approximated as the total population of India minus initial infected cases
- Initial infected and recovered values taken from early pandemic data

The model equations were solved numerically to simulate infection dynamics over time, and the results were compared with actual observed data.

7. Wave Identification and Comparative Analysis

The study period was divided into two primary phases:

- **First Wave:** Approximately mid-2020 to early 2021
- **Second Wave:** March 2021 to June 2021

Comparative analysis was performed based on:

- Peak daily cases
- Active case load
- Transmission and recovery rates

This enabled a clear distinction between the severity and dynamics of the two waves.

8. Tools and Software

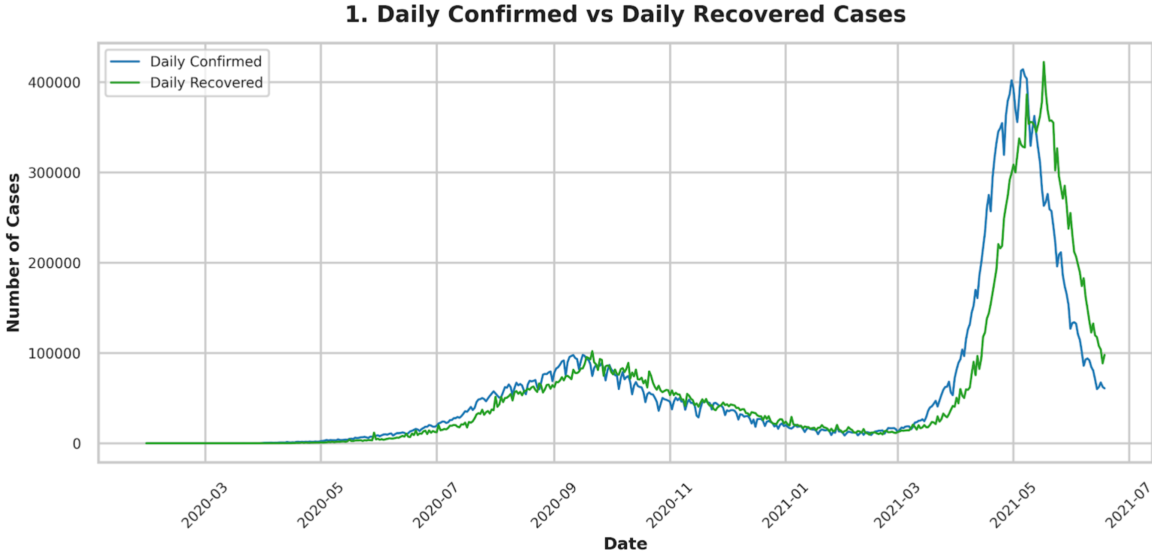
The analysis was carried out using:

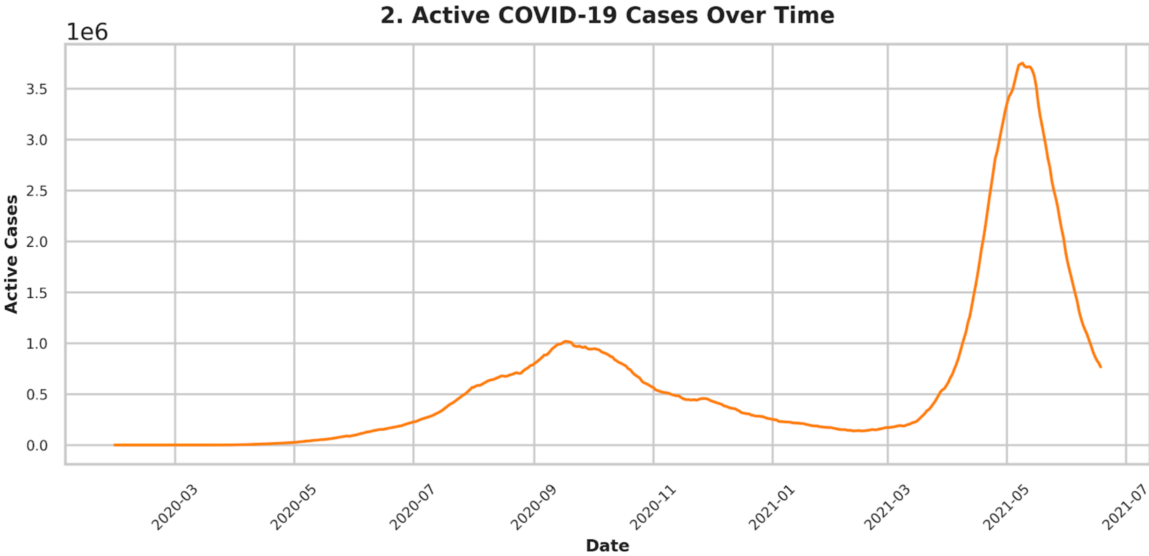
- **Python** (Pandas, NumPy, Matplotlib) for data processing and visualization
- **Microsoft Excel** for preliminary data handling and validation

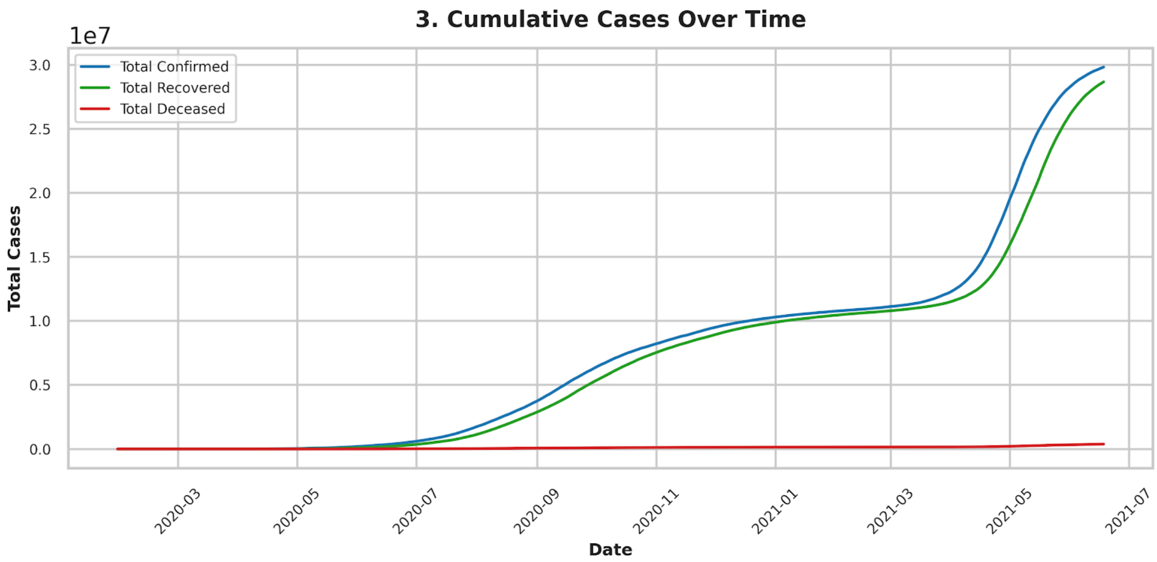
- **The transmission rate (β) and recovery rate (γ) were estimated using observed time-series data. The recovery rate was approximated as γ ≈ 0.1, assuming an second wave was calculated as r ≈**

0.07, leading to an estimated transmission rate β ≈ 0.17. The basic reproduction number (R_0 ≈ 1.7) indicates a high level of disease transmissibility during the peak phase.

III. RESULTS AND DISCUSSION







1. Temporal Trends in Daily Confirmed and Recovered Cases

- The time-series analysis of daily confirmed and recovered COVID-19 cases in India reveals the presence of two distinct pandemic waves during the study period (2020–2021). The graphical representation of daily confirmed and recovered cases highlights significant variations in infection dynamics, recovery patterns, and overall disease progression.
- During the **first wave (mid-2020 to September 2020)**, daily confirmed cases exhibited a gradual upward trend, reaching a peak of approximately 90,000 cases per day. The increase in cases during this phase was relatively steady, indicating a moderate transmission rate. Notably, the daily recovered cases closely followed the confirmed cases, suggesting an effective recovery response and relatively stable healthcare capacity during this period. The near-parallel trend between confirmed and recovered cases implies that the rate of recovery was sufficient to prevent excessive accumulation of active cases. This behavior is consistent with previously reported studies indicating controlled transmission during early pandemic phases (Anderson & May, 1992)

- Following the peak, a **declining phase (October 2020 to February 2021)** was observed, characterized by a consistent reduction in both confirmed and recovered cases. This phase reflects a period of relative containment, likely influenced by public health interventions, behavioral adaptations, and partial immunity development within the population.

- In contrast, the **second wave (March 2021 to May 2021)** exhibited a dramatically different pattern. A sharp and exponential increase in daily confirmed cases was observed, with peak values exceeding 400,000 cases per day. This rapid surge indicates a significantly higher transmission intensity compared to the first wave. Although recovered cases also increased during this period, a noticeable lag between confirmed and recovered cases was evident. This divergence suggests that the rate of new infections outpaced recoveries, leading to a rapid accumulation of active cases and increased strain on healthcare infrastructure. This rapid surge indicates significantly higher transmission intensity, consistent with findings from recent epidemiological studies (Kucharski et al., 2020)

2. Dynamics of Active Cases and Healthcare Burden

- The analysis of active COVID-19 cases provides critical insights into the real-time burden on the healthcare system. Active cases, defined as the difference between total confirmed, recovered, and deceased cases, represent the number of individuals requiring medical attention at a given time.

- During the first wave, active cases increased steadily, reaching a peak of approximately 1 million cases in September 2020. However, due to the relatively balanced relationship between infection and recovery rates, the healthcare system was able to manage the caseload more effectively.
- A substantial decline in active cases was observed during the post-first-wave phase, with active cases dropping to minimal levels by early 2021. This period indicates successful mitigation and recovery.
- However, the second wave resulted in an unprecedented surge in active cases, reaching approximately 3.5 to 3.7 million cases at its peak in May 2021. This sharp increase reflects the combined effect of high transmission rates and delayed recoveries. The magnitude of active cases during this phase significantly exceeded that of the first wave, clearly demonstrating the severe pressure placed on healthcare infrastructure, including hospital capacity, oxygen supply, and critical care resources.

3. SIR Model Analysis and Epidemiological Insights

- To quantitatively assess the transmission dynamics, the SIR (Susceptible–Infected–Recovered) model was applied to the observed data. The estimated model parameters provide valuable insights into the progression of the pandemic.
- The recovery rate (γ) was approximated as 0.1, assuming an average infectious period of 10 days. The exponential growth rate during the second wave was estimated to be approximately 0.07, based on observed increases in daily confirmed cases. Using these values, the transmission rate (β) was calculated as approximately 0.17.
- The basic reproduction number (R_0), defined as the ratio of transmission rate to recovery rate, was estimated as:
- $R_0 = \frac{\beta}{\gamma} \approx 1.7$
- An R_0 value greater than 1 indicates sustained transmission and epidemic growth. The estimated value of approximately 1.7 during the second wave suggests a high level of transmissibility, consistent with the observed exponential rise in cases.

4. Comparative Analysis of First and Second Waves

- A detailed comparative assessment of the first and second waves of COVID-19 in India reveals substantial differences in epidemiological characteristics, transmission dynamics, and healthcare impact. These differences are evident not only in the magnitude of cases but also in the underlying rate processes governing infection and recovery.
- The **first wave** was characterized by a relatively gradual increase in daily confirmed cases, indicating a moderate transmission rate. The epidemic curve during this phase followed a more symmetric pattern, with a well-defined growth, peak, and decline. The close alignment between daily confirmed and recovered cases suggests that the recovery rate (γ) was sufficiently balanced with the transmission rate (β), thereby limiting excessive accumulation of active cases. This phase reflects a relatively controlled epidemic progression, likely influenced by strict lockdown measures, reduced mobility, and higher public compliance with preventive protocols.
- In contrast, the **second wave** exhibited a markedly different epidemiological profile. The rapid and steep rise in daily confirmed cases indicates a significantly elevated transmission rate (β), consistent with the exponential growth observed in the data. The epidemic curve during this phase is highly asymmetric, with a sharp ascent and a comparatively slower decline, suggesting prolonged disease persistence within the population. The divergence between confirmed and recovered cases further highlights a temporal lag in recovery dynamics, contributing to a rapid escalation in active cases.
- From a quantitative perspective, the estimated basic reproduction number ($R_0 \approx 1.7$) during the second wave exceeds the threshold required for sustained transmission, thereby explaining the accelerated spread. In comparison, the first wave likely exhibited a relatively lower effective reproduction number, resulting in slower disease propagation.

- Another critical distinction lies in the **healthcare burden**. While the first wave reached a peak of approximately 1 million active cases, the second wave surpassed 3.5 million active cases, indicating a more than threefold increase in active disease burden. This dramatic escalation not only reflects higher transmission but also underscores systemic challenges, including delayed recovery, increased hospitalization rates, and resource limitations.
- Furthermore, the temporal compression of the second wave—characterized by a rapid rise within a short duration—suggests the influence of additional factors such as the emergence of highly transmissible viral variants, increased population mobility, and reduced adherence to non-pharmaceutical interventions. These factors collectively contributed to the amplification of transmission dynamics, distinguishing the second wave as significantly more severe and complex than the first.

IV. Implications of Findings

- The findings of this study have important implications for epidemiological research, public health policy, and future pandemic preparedness. The observed differences between the first and second waves highlight the dynamic and evolving nature of infectious disease transmission, emphasizing the need for adaptive and data-driven response strategies.
- Firstly, the significantly higher transmission rate (β) and reproduction number (R_0) during the second wave underscore the critical importance of **early detection and timely intervention**. Delays in implementing control measures during the exponential growth phase can lead to disproportionately large increases in case numbers, as demonstrated by the rapid escalation observed in this study. Therefore, continuous monitoring of key epidemiological indicators, such as growth rate and reproduction number, is essential for proactive decision-making.
- Secondly, the substantial increase in active cases during the second wave highlights the necessity of **robust healthcare infrastructure and surge capacity planning**. The overwhelming burden on healthcare systems during peak transmission periods indicates that preparedness must extend beyond baseline capacity, incorporating scalable resources such as hospital beds, oxygen supply, and critical care facilities. Data-driven projections based on models such as SIR can play a vital role in anticipating these requirements.
- Thirdly, the integration of real-world data with mathematical modeling, as demonstrated in this study, provides a powerful framework for understanding epidemic behavior. Although the SIR model is relatively simple, its ability to capture key transmission dynamics makes it a valuable tool for both retrospective analysis and future forecasting. The estimation of parameters such as β and γ enables quantitative assessment of intervention effectiveness and disease progression.
- Moreover, the findings emphasize the importance of **public adherence to preventive measures**, including social distancing, mask usage, and vaccination. Variations in these factors can significantly influence transmission rates, as reflected in the contrasting dynamics of the two waves.
- Finally, this study highlights the broader significance of **data transparency and accessibility** in managing public health crises. The availability of high-quality, real-time data enables researchers to perform timely analyses, generate actionable insights, and support evidence-based policymaking.
- In summary, the results underscore that effective pandemic management requires a combination of **epidemiological modeling, real-time data analysis, healthcare preparedness, and coordinated public health interventions**. The methodological approach adopted in this study can be extended to future outbreaks, providing a scalable framework for analyzing and mitigating the impact of infectious diseases

V. CONCLUSION

- This study presents a comprehensive data-driven and epidemiological assessment of COVID-19 wave dynamics in India during the period 2020–2021, integrating real-world time-series analysis with the classical Susceptible–Infected–Recovered (SIR) modeling framework. By examining daily confirmed, recovered, and active case trends, the research provides a clear characterization of the temporal evolution and relative severity of the first and second waves of the pandemic.

- The findings demonstrate that the first wave of COVID-19 in India was marked by a gradual increase and decline in cases, reflecting a relatively controlled transmission environment. In contrast, the second wave exhibited a rapid and exponential surge in infections, with daily confirmed cases exceeding 400,000 and active cases reaching approximately 3.7 million at peak levels. This stark difference highlights the significantly higher transmission intensity and epidemiological complexity associated with the second wave.
- The application of the SIR model enabled the quantitative estimation of key epidemiological parameters, including the transmission rate ($\beta \approx 0.17$) and recovery rate ($\gamma \approx 0.1$). The resulting basic reproduction number ($R_0 \approx 1.7$) indicates sustained and accelerated transmission during the second wave, consistent with the observed exponential growth patterns. These findings reinforce the utility of compartmental models in capturing essential features of infectious disease spread, even in the presence of real-world complexities.
- A key insight from this study is the critical relationship between infection and recovery dynamics in determining the burden on healthcare systems. The pronounced lag between confirmed and recovered cases during the second wave led to a substantial accumulation of active cases, thereby overwhelming healthcare infrastructure. This emphasizes the importance of not only controlling transmission but also enhancing recovery capacity through timely medical intervention and resource allocation.
- Furthermore, the comparative analysis of the two waves underscores the dynamic nature of pandemic progression, influenced by factors such as viral mutations, population behavior, and public health interventions. The significantly greater severity of the second wave highlights the consequences of delayed responses and the need for continuous monitoring of epidemiological indicators.
- From a broader perspective, this study demonstrates the effectiveness of combining empirical data analysis with mathematical modeling to generate actionable insights into epidemic behavior. The methodological framework adopted here can serve as a foundation for future research on infectious disease dynamics, particularly in resource-constrained settings where rapid and data-driven decision-making is essential.
- In conclusion, the results of this study emphasize that effective pandemic management requires an integrated approach involving real-time data analysis, robust epidemiological modeling, and proactive public health strategies. The insights derived from the Indian COVID-19 experience provide valuable lessons for mitigating future waves and enhancing preparedness for emerging infectious diseases.

VI. RESEARCH GAP

- Despite the extensive body of literature on COVID-19 epidemiology, several critical gaps remain in understanding the dynamic progression of pandemic waves, particularly in the context of developing countries such as India. Existing studies have primarily focused on either descriptive statistical analysis of case trends or theoretical modeling approaches in isolation, with limited integration of real-world data and mathematical frameworks.
- Firstly, many prior investigations have emphasized **cumulative case analysis**, which often masks short-term temporal variations and fails to capture the dynamic nature of infection spread. There is a lack of detailed studies utilizing **high-resolution daily time-series data** to analyze the evolution of pandemic waves and identify distinct growth, peak, and decline phases. Many existing studies rely primarily on cumulative data analysis, which often fails to capture short-term variations in transmission dynamics (Anderson & May, 1992)

- Secondly, although compartmental models such as the SIR framework have been widely applied, a significant gap exists in the **data-driven estimation of model parameters** using real-world datasets. In many cases, parameters such as the transmission rate (β) and recovery rate (γ) are assumed or derived from generalized estimates rather than being calculated from observed trends specific to the study region. This limits the accuracy and contextual relevance of such models.
- Thirdly, limited attention has been given to the **comparative analysis of multiple pandemic waves** within a single unified framework. In particular, there is a scarcity of studies that systematically compare the first and second waves of COVID-19 in India by integrating epidemiological modeling with empirical data analysis. Such comparisons are essential to understand variations in transmission dynamics, healthcare burden, and intervention effectiveness.

- Additionally, existing research often overlooks the **relationship between infection dynamics and healthcare system capacity**, particularly the role of active cases as an indicator of real-time system stress. The lack of integrated analysis linking confirmed, recovered, and active cases with epidemiological parameters represents a significant gap in current literature.
- Furthermore, while advanced models (e.g., SEIR, agent-based models) have been explored, there is a need for studies that demonstrate the effectiveness of **simpler, interpretable models like SIR** when combined with real-world data. Such approaches are especially valuable in resource-limited settings where complex modeling may not be feasible.
- In this context, the present study addresses these gaps by integrating **daily time-series data analysis with SIR-based epidemiological modeling** to provide a comprehensive assessment of COVID-19 wave dynamics in India. By estimating model parameters from observed data and performing a comparative analysis of the first and second waves, this research contributes to a more accurate and context-specific understanding of pandemic behavior.

References

[1] World Health Organization (2020). Coronavirus disease (COVID-19) pandemic. Retrieved from: https://www.who.int

[2] covid19india.org (2021). COVID-19 Dataset for India. Retrieved from: https://www.covid19india.org

[3] William O. Kermack, & A. G. McKendrick (1927).“A Contribution to the Mathematical Theory of Epidemics.”Proceedings of the Royal Society A, 115(772), 700–721.

[4] Johns Hopkins University (2021).COVID-19 Data Repository by the Center for Systems Science and Engineering (CSSE).

[5] Hethcote H. W. (2000).“The Mathematics of Infectious Diseases.”SIAM Review, 42(4), 599–653.

[6] Indian Council of Medical Research (2021).COVID-19 Updates and Reports.

[7] Kucharski A. J. et al. (2020).“Early dynamics of transmission and control of COVID-19.”The Lancet Infectious Diseases, 20(5), 553–558.

[8] Our World in Data (2021).Coronavirus Pandemic Dataset. Retrieved from: https://ourworldindata.org

[9] Anderson R. M., & May R. M. (1992).“Infectious Diseases of Humans: Dynamics and Control.”Oxford University Press.