Analysing the Dynamics of the COVID-19 Pandemic

Prakarsh Jhajharia

Received April 25, 2024
Accepted June 24, 2024
Electronic access July 15, 2024

This research aims to better understand infection dynamics and epidemiological factors associated with the COVID-19 pandemic. By employing advanced statistical methodologies such as linear, polynomial, ridge and LASSO regression, the study provides valuable information for future pandemic preparedness. We examined the COVID-19 pandemic by analysing global data collected from March 1, 2020, to March 1, 2022. Our findings challenge the conventional quarantine period of 14 days as recommended by the World Health Organisation, rather suggesting an optimal quarantine period of 17 days. To our surprise, no clear correlations emerge between testing rates, vaccine administration, and reported cases. Demographic analyses show no direct links between population density and infection rate. These insights contribute to pandemic preparedness, offering a comprehensive understanding for future global health strategies.

Keywords: Computational Biology and Bioinformatics; Computational Epidemiology; COVID - 19; Pandemic

Introduction

This research aims to unravel essential insights into the dynamics of the COVID-19 pandemic, with the goal of advancing our comprehension of infection patterns, factors influencing mortality, and the global impact of demographics. This involved analysing a variety of features, including infection rates, mortality rates and demographic statistics. Additionally, the project aimed to analyse common hypotheses about the spread and severity of COVID-19, ranging from the effectiveness of quarantine measures to the influence of demographic features[1].

Our research encompasses a comprehensive analysis of data collected from diverse, reliable sources including the Word Bank and World Health Organisation, spanning from March 1, 2020, to March 1, 2022. We employ a range of statistical methods, to analyse the data and uncover details that could significantly contribute towards our understanding of the COVID-19 pandemic.

By uncovering key patterns and correlations in the spread and impact of COVID-19, our research shows how demographic features, healthcare facilities, and social practices influence the pandemic[2]. For instance, understanding the relationship between ICU admissions and mortality rate can help healthcare systems better prepare for such pandemics. Additionally, the analysis of the effectiveness of health measures, like quarantine periods, helps determine the viability of such measures. These findings further help policymakers and health officials in coming up with effective strategies to reduce the impact of future pandemics.

The existing literature on the COVID-19 pandemic reflects a diverse range of studies that predominantly focus on various aspects of the virus's spread, impacts, and mitigation strategies. Noteworthy contributions include studies by He J, Chen G, Jiang Y, et al. (2020) which investigate regional infection patterns and the effectiveness of control measures[3]. Additionally, the work of Liang, C. K., and Chen, L. K. (2022) as well as the work of Tchicaya A, Lorentz N, Leduc K and de Lanchy G. (2021) delve into mortality rates and healthcare system responses[4,5]. However, the current research distinguishes itself by adopting a comprehensive global perspective, utilising an extensive dataset spanning 193 countries. The application of advanced statistical methods, such as Linear Regression and Polynomial Regression, sets it apart, allowing for a nuanced exploration of infection dynamics and demographic influences. Unlike previous studies, this research challenges conventional beliefs, presenting unexpected findings on quarantine periods, critical admission timeframes, and the intricate relationships between testing, vaccination, and reported cases. These unique contributions enrich the existing body of literature and provide valuable insights for future pandemic preparedness and global health strategies.

Data Collection

After extensive research spanning several days, we identified a comprehensive dataset on ourworldindata.org, which offered day-wise statistics on a wide array of features related to COVID-19 (Figure 1)[6]. These ranged from daily new cases to daily new vaccinations. Additionally, we sourced demographic data from data.worldbank.org, which proved instrumental in uncovering some previously unseen correlations between demography and the spread of COVID-19[7]. To establish a connection between COVID-19 and other health conditions, such as cardiac arrest and diabetes, we utilised data from who.int.[8]

Due to recurring issues with data reporting, the dataset underwent cross-validation with alternative sources to ensure accuracy. To minimise the risk of inaccurate findings, the study only included data from countries with comprehensive records during the specified timeframe. Additionally, a thorough review of the data sources provided by ourworldindata.org revealed that the primary data originated from UN sources and official government records.

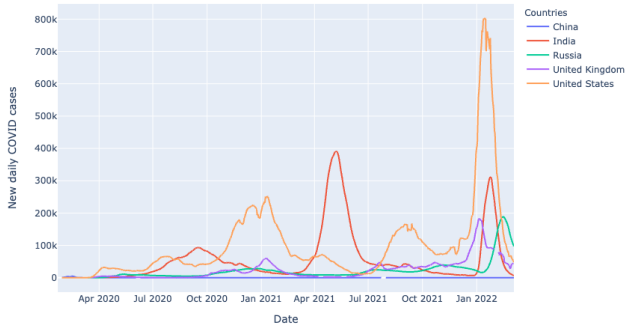

Fig. 1 Graph depicting daily new COVID-19 cases in China, India, Russia, the United Kingdom, and the United States. [The graph shows a sharp increase in daily new cases indicating the emergence of a wave during that period, as evidenced in India and the United States. Conversely, relatively flat lines indicate the stabilisation of COVID-19 cases, as observed in China.]

Features

In terms of the features analysed, our study encompassed 20 major aspects of COVID-19 across 193 countries (Table 1). These included daily new cases, deaths, tests, vaccinations, hospital and ICU admissions, and various demographic features like population, population density, median age, and the proportion of the population aged 65 or older and 70 or older. Economic indicators like GDP per capita were also considered, along with health-related features such as reproduction rate, availability of handwashing facilities, life expectancy, human development index, mortality rate, cardiac arrest death rate, and the prevalence of diabetes.

Limitations

In our study, it is of utmost importance to acknowledge the limitations of the data that might impact our findings. The model relies exclusively on data collected from March 1, 2020, to March 1, 2022. This limited time frame takes into account only the initial and intermediate stages of the COVID-19 pandemic, thereby being unable to capture the long-term effects and subsequent waves of the virus. Another major limitation is the potential inaccuracy in the data. This inaccuracy arises from various features, including but not limited to, false reporting by government officials, discrepancies in testing and reporting of cases across different countries, and delays in data collection (Figure 2). Such inaccuracies can introduce biases and errors in the model. Additionally, the model may not account for unmeasured features, such as socio-political features and public health interventions (like lockdowns and social distancing measures) that can significantly impact the spread of the pandemic. Despite these limitations, the study has mitigated inaccuracies by cross-referencing across multiple data sources. Additionally, data normalisation has been employed, and only countries with robust datasets have been included to reduce the risk of false reporting by government officials. Furthermore, the results have been compared with findings from other research in the field to ensure the integrity and accuracy of the data presented.

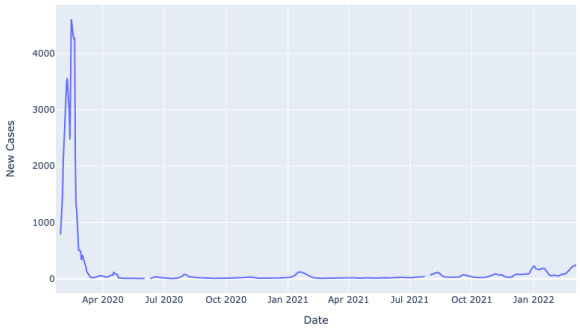

Fig. 2 Graph illustrating daily COVID-19 cases in China from March 1, 2020, to March 1, 2022. [This graph highlights the underreporting of cases in China. Post-April 2020, the reported cases drop nearly to zero, with multiple data gaps further indicating a misrepresentation of the actual situation.]

Feature Scaling

An integral part of our data preparation process was feature scaling, a technique employed to ensure uniformity in the values of various features within our dataset. Feature scaling prevents any single feature with larger values from disproportionately influencing the model. We considered two primary methods of feature scaling, Normalisation and Standardisation[9]. We chose normalisation for our model primarily because it allowed for a fair and equitable comparison across all countries. By scaling all data points to a consistent range, normalisation ensured that no single country's data could dominate due to scale differences alone. The sensitivity to outliers in normalisation is also particularly important in analysing the COVID-19 pandemic across various countries. It highlights anomalies or sudden spikes in

Features	Purpose
New Cases	The number of new COVID-19 cases reported daily.
New Deaths	The number of new deaths due to COVID-19 daily.
New Hospital Patients	The number of patients hospitalised due to COVID-19 daily.
New ICU Patients	The number of COVID-19 patients admitted into Intensive Care Units daily.
New Tests	The number of new tests conducted for COVID-19 daily.
Positive Rate	The percentage of COVID-19 tests that return positive.
New Vaccinations	The number of COVID-19 vaccine doses administered daily.
Population	The total number of people living in a specific region or country.
Population Density	The number of people living per unit of area (e.g., per square kilometre).
Median Age	The age that divides a population into two numerically equal groups: half are younger, half older.
Population aged 65 or older	The percentage of the population that is 65 years old or older.
Population aged 70 or older	The percentage of the population that is 70 years old or older.
GDP Per Capita	The gross domestic product divided by the population, indicating average economic output per person.
Human Development Index	A composite index measuring average achievement in key dimensions of human development.
Life Expectancy	The average number of years a person is expected to live.
Reproduction Rate	The average number of children a woman would have in her lifetime.
Handwashing Facilities	The availability of handwashing facilities.
Diabetes	Prevalence of diabetes in the population.
Mortality Rate	The rate of deaths in a population.
Cardiac Arrest Death Rate	The rate of deaths due to cardiac arrests.

Table 1 Features in our dataset along with their purposes.

case numbers, which could indicate critical changes in the pandemic's spread. Furthermore, by maintaining the integrity of these outliers, normalisation helps in preserving crucial information that could be pivotal in understanding the dynamics of the virus transmission and the effectiveness of control measures across different regions. Standardisation, by transforming data to have zero mean and unit variance, could dilute the impact of outliers by pulling their values closer to the mean, potentially under-representing critical variations such as extremely high number of cases or sudden spikes in deaths. This can lead to a misleading analysis where significant anomalies might have been overlooked.

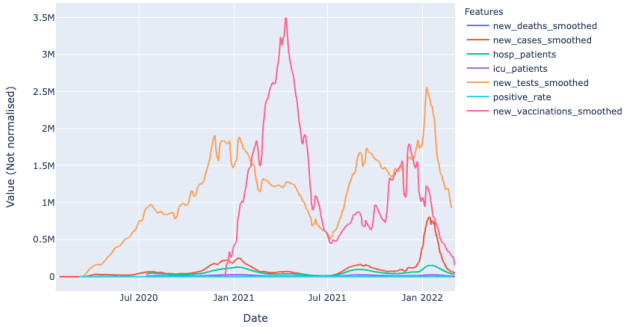

Fig. 3 Graph depicting the non-normalized values of various features for the United States of America. [The figure illustrates that certain features, such as new cases and new tests, have significantly higher overall values compared to others.]

Methods

Linear Regression

Linear regression is a machine learning method that is used to find the relationship between a fixed or dependent variable and other independent variables. Its main goal is to find the line of best fit (Figure 5).

To calculate the line of best fit, linear regression uses a slope-intercept form,

$$y_i = \beta_0 + \beta_1 x_i \tag{1}$$

In our approach, we used Linear Regression to create a model that utilises a range of features to accurately predict a target feature (Figure 6). This method utilised the fact that various independent variables, when analysed together, could provide a c precise prediction of another dependent variable. We developed this model by carefully selecting relevant features that were likely to influence the target variable. This selection was based on a combination of statistical analysis and common knowledge. Once the features were chosen, we applied the Linear Regression model to plot the line of best fit.

To analyse the dynamics between COVID-19 diagnosis and subsequent health outcomes, we utilised a lag-based analysis using linear regression. We began by examining a range of

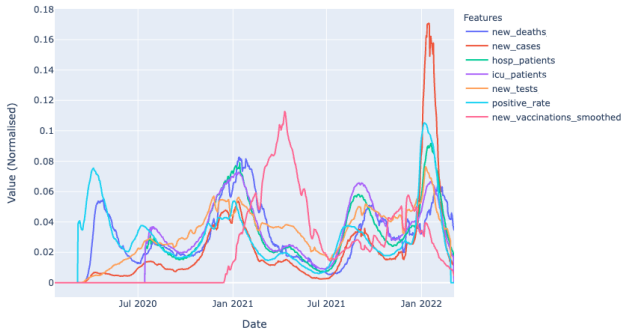

Fig. 4 Graph depicting the normalised values of various features for the United States of America. [The figure demonstrates that all features are now equally weighted, ensuring no single feature dominates the analysis.]

time lags from 0 to 30 days to explore potential delays between initial diagnosis and ensuing outcomes. For each lag, we constructed linear regression models where daily counts of new cases, hospitalizations, ICU admissions, etc. served as independent variables, and outcomes up to 30 days later were the dependent variables. We calculated correlation coefficients from these models to measure the strength of correlations at different lags, identifying the time intervals where these relationships were most pronounced. This methodology allowed us to determine the critical periods for intervention and monitoring, providing a quantitative basis for our understanding of the disease's progression as can be seen in the first three hypotheses.

Fig. 5 Graph illustrating a Linear Regression Model for Daily New COVID-19 Cases versus Daily New Deaths in the United States of America. [The figure reveals a moderate correlation between these two factors, suggesting that an increase in daily cases tends to correspond with an increase in daily deaths.]

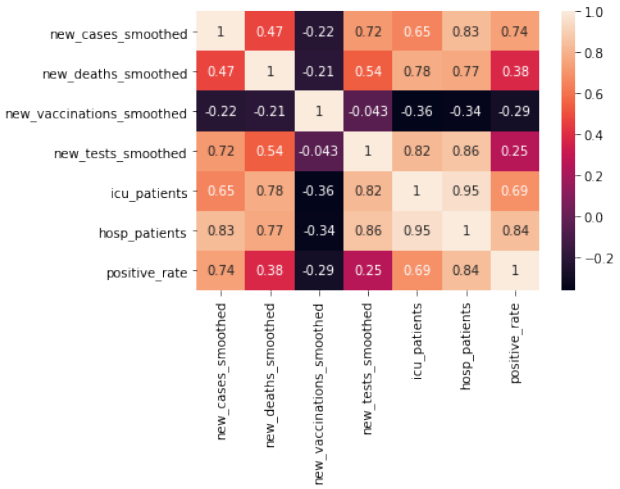

Fig. 6 Heatmap depicting correlations among the top seven features of the dataset.

Polynomial Regression

In cases where the data points are linear, a linear regression algorithm is effective, but in situations where the data shows a nonlinear relationship, linear regression falls short. Unlike linear regression, which models the relationship as a straight line, polynomial regression can fit data with a nonlinear relationship. The loss function remains the Mean Squared Error (MSE).

The general form of a polynomial regression model is:

$$y_i = \beta_0 + \beta_1 x + \beta_2 x^2 + \ldots + \beta_n x^n + \varepsilon \tag{2}$$

In our dataset, we employed Polynomial Regression to explore potential non-linear relationships among various features (Figure 7). This approach was important because some relationships in the data may not be apparent through linear analysis. Polynomial Regression allows us to model and understand complex relationships where the impact of one variable on another is not simply linear but could be quadratic, cubic, or of even higher order.

Ridge and LASSO Regression

To explore other forms of relationships among various features in the dataset we employed Ridge and LASSO Regression (Figure 8 and 9). This was essential as it allowed us to model and notice key correlations that were not apparent using linear and polynomial regression models. Ridge and LASSO regression allows us to understand complex relationships that may involve multicollinearity and redundant data.

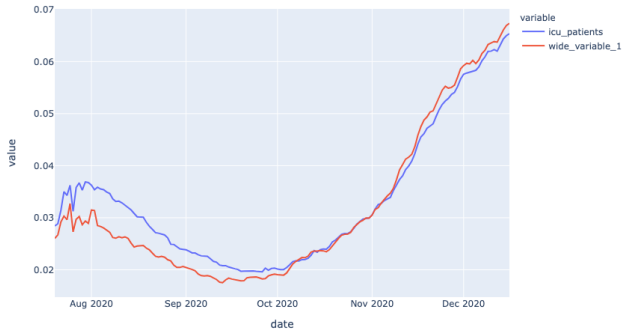

Fig. 7 Graph illustrating a Polynomial Regression model used for predicting the number of New ICU Patients. [The model, trained on three features: New Tests, New Cases, and New Hospital Admissions, displays actual values in blue and predicted values in red. This visualisation indicates that Polynomial Regression may be a suitable model for this dataset.]

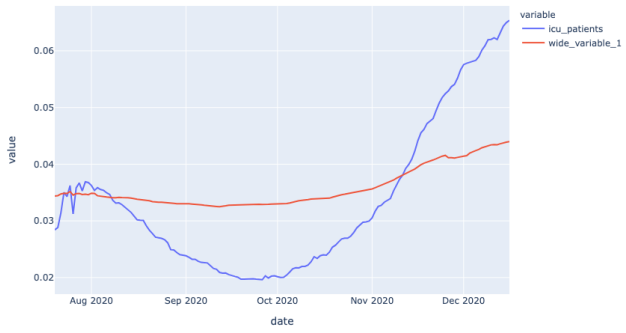

Fig. 8 Graph illustrating a Ridge Regression Model used for predicting the number of New ICU Patients. [The model, trained on three features: New Tests, New Cases, and New Hospital Admissions, displays actual values in blue and predicted values in red. This visualisation indicates that Ridge Regression may not be the most suitable model for this dataset.]

Results

Introduction

In this study, we delve into various aspects of the COVID-19 pandemic, exploring critical hypotheses under three main groups. Each hypothesis aims to unravel different aspects of the pandemic, from health outcomes and healthcare dynamics to societal impacts.

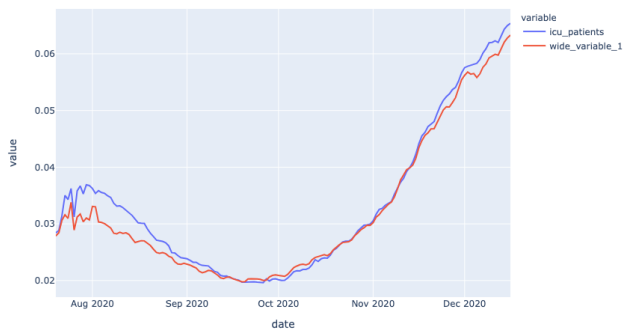

Fig. 9 Graph illustrating a Lasso Regression Model used for predicting the number of New ICU Patients. [The model, trained on three features: New Tests, New Cases, and New Hospital Admissions, displays actual values in blue and predicted values in red. This visualisation indicates that Lasso Regression may be a suitable model for this dataset.]

Group 1 - Pandemic Progression and Health Outcomes:

1. **Hypothesis 1** investigates the adequacy of the 14-day quarantine period, finding a need to potentially extend it.

2. **Hypothesis 2** examines the timing of hospital or ICU admissions post-COVID-19 diagnosis, revealing an earlier than expected timeframe.

3. **Hypothesis 3** explores the correlation between ICU admissions and subsequent deaths, indicating a close temporal relationship.

Group 2 - Testing and Vaccination Dynamics:

1. **Hypothesis 4** assesses the correlation between the number of COVID-19 tests and new cases, finding an unclear link.

2. **Hypothesis 5** analyses the relationship between case numbers and vaccine administration, discovering no evident correlation.

Group 3 - Demographic and Societal Factors:

1. **Hypothesis 6** looks into the correlation between total cases per million and population density, concluding no direct relationship.

2. **Hypothesis 7** probes the link between total case numbers and overall population size, also finding no correlation.

Fig. 10 A compilation of four graphs representing different regression models: Linear Regression (Top Left), Polynomial Regression (Top Right), Ridge Regression (Bottom Left), and Lasso Regression (Bottom Right). [These models, aimed at predicting Daily New Cases, are trained using seven features: New Tests, New Deaths, Positive Rate, New Vaccinations, New ICU Admissions, and New Hospital Admissions. Each graph displays the actual values in blue and the predicted values in red. This visualisation suggests that the Linear, Lasso, and Polynomial Regression models may be suitable for this dataset, whereas the Ridge Regression model appears less fitting.]

Pandemic Progression and Health Outcomes

Hypothesis 1: The optimal quarantine period for COVID-19 is 14 days[10]

Explanation: This hypothesis investigates the likely time frame for COVID-19 fatalities post-diagnosis. We examined the correlation between new COVID-19 cases and subsequent deaths to predict the ideal quarantine duration. The Hypothesis enabled us to statistically predict the adequate amount of days one should remain in quarantine and whether the widespread assumption of the adequate quarantine period being 14 days was true or false.

Findings: On analysing the data for a month after a person is diagnosed with COVID-19 we observed the strongest correlation between the cases and deaths to be at approximately 17 days (Figure 11). The hypothesis examines the number of days after which a person diagnosed with COVID-19 is most likely to succumb to the virus. It is predicated on the assumption that following this peak period, the severity of the virus will naturally diminish, suggesting that 17 days is a suitable quarantine duration. Considering these findings, it is recommended to extend the COVID-19 quarantine period to about 17 days for optimal effectiveness.

Hypothesis 2: A person diagnosed with COVID-19 is most likely to be admitted to the hospital or ICU after 14 days.

Explanation: This hypothesis examines the timeframe within which COVID-19 patients are typically admitted to the hospital or intensive care unit (ICU) post-diagnosis. The goal is to determine the critical period for hospitalisation or intensive care following a positive COVID-19 test, with an initial assumption that this period is around 14 days[11].

Findings: Our data analysis over the course of a month following a COVID-19 diagnosis revealed that the highest correlation between the diagnosis and hospital or ICU admissions occurs at approximately 9 days with quite high correlation from

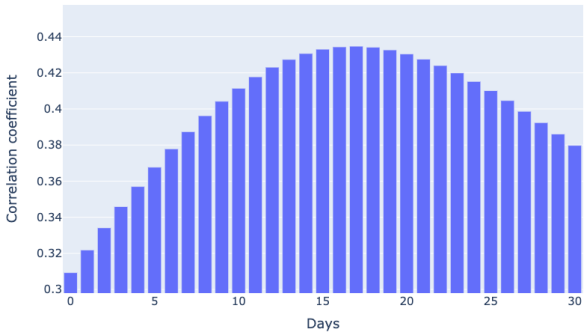

Fig. 11 This graph illustrates the correlation coefficient between daily new COVID-19 cases and deaths over a 0 to 30-day period following the report of a new case. [The correlation sharply increases, reaching its peak at 17 days, and then shows a gradual decline.]

5 to 10 days. This peak correlation suggests a shorter timeframe than the initially hypothesised 14 days for critical care admission (Figure 12). Based on these findings, the critical period for hospitalisation or ICU admission for COVID-19 patients is sooner than previously thought, occurring around 5 to 10 days after diagnosis rather than the initially assumed 14 days.

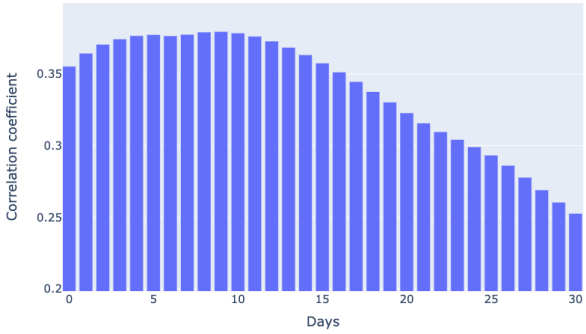

Fig. 12 This graph illustrates the correlation coefficient between the date of COVID-19 diagnosis and subsequent hospital or ICU admissions over a 0 to 30-day period. [The correlation reaches its maximum at 9 days post-diagnosis, and then shows a gradual decline.]

Hypothesis 3: There is a correlation between the daily number of new ICU admissions and the number of new COVID-19 related deaths.

Explanation: This hypothesis aims to explore the relationship between the number of new patients admitted to Intensive Care Units (ICU) each day and the number of new deaths due to COVID-19. The focus is on understanding how changes in ICU admissions may be predictive of fatality rates.

Findings: Upon analysing the data, we found that the strongest correlation between the daily increase in ICU admissions and new deaths occurs at a lag of 2 days. This indicates that the number of new ICU admissions today can be a significant predictor of the number of deaths two days later (Figure 13). The peak correlation at 2 days highlights the critical importance of ICU admissions as a potential early indicator of mortality trends in the pandemic.

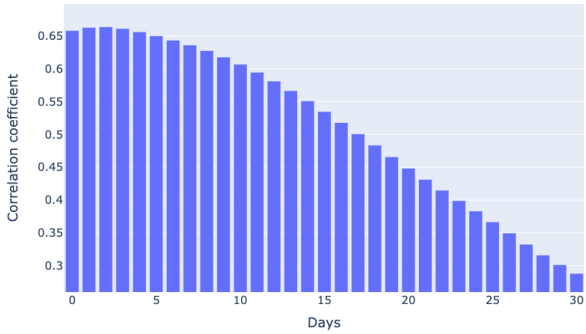

Fig. 13 This graph represents the correlation coefficient between the daily number of new ICU admissions and the number of new COVID-19 related deaths over a 0 to 30-day period. [The correlation reaches its peak at a 2-day interval, suggesting a close temporal relationship between ICU admissions and fatalities.]

Testing and Vaccination Dynamics

Hypothesis 4: There is a significant correlation between the number of new COVID-19 tests conducted and the number of new cases detected.

Explanation: This hypothesis investigates the relationship between the frequency of COVID-19 testing and the detection of new cases. The objective is to understand how the volume of testing impacts the reported case numbers, potentially indicating the extent of undiagnosed cases in the population [12].

Findings: Our analysis reveals a correlation between the number of new tests conducted and the number of new COVID-19 cases reported. However, the correlation coefficient (0.4706) obtained is not adequate to assert a definitive correlation. This suggests that while there may be a relationship between testing rates and case detection, other factors could also be influencing the number of reported cases (Figure 14).

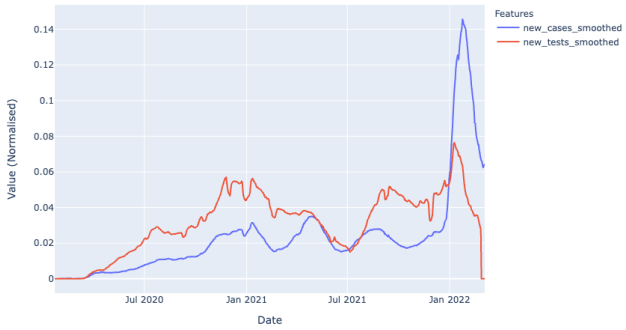

Fig. 14 This graph shows the daily number of new COVID-19 tests conducted and the daily number of new cases reported. [It indicates a potential link between the volume of tests and case numbers, but the correlation is not strong enough to confirm a direct relationship.]

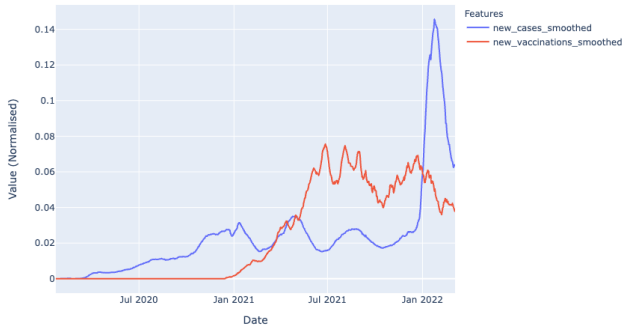

Fig. 15 This graph compares the daily number of new COVID-19 cases with the daily number of vaccines administered. [It demonstrates no visible pattern or correlation between these two variables, indicating that the number of cases does not directly relate to the number of vaccines given.]

Hypothesis 5: The number of COVID-19 cases is correlated with the number of vaccines administered.

Explanation: This hypothesis seeks to explore the potential relationship between the number of reported COVID-19 cases and the volume of COVID-19 vaccines administered. The aim is to determine if an increase in vaccinations correlates with a change in the number of reported cases, which could indicate the effectiveness of vaccination campaigns in controlling the spread of the virus [13].

Findings: Upon analysing the data, we found no significant correlation between the number of COVID-19 cases and the number of vaccines administered. This lack of correlation suggests that the relationship between these two variables is not straightforward and may be influenced by a variety of other factors such as vaccine efficacy, population immunity levels, virus variants, and public health practices (Figure 15). The analysis reveals that there is no correlation between the number of COVID-19 cases and the number of vaccines administered, leading to the rejection of the hypothesis. Past studies in this field have found the vaccines most effective after 14 or 21 days [14]. This hypothesis, however, aimed at finding the immediate impact of administering the vaccine. Another factor that might cause minor discrepancies is the occurrence of spikes in new test cases during a wave, while the number of vaccines administered stays relatively constant as evident in Figure 15 from Jan 2022.

Demographic and Societal Factors

Hypothesis 6: The number of total COVID-19 cases per million has a direct correlation with population density.

Explanation: This hypothesis examines whether a direct correlation exists between the population density of an area and the total number of COVID-19 cases reported per million people. The objective is to assess if densely populated areas have higher rates of COVID-19 cases, as might be expected due to closer contact among residents [15].

Findings: After analysing the data, we found no significant correlation between population density and the number of COVID-19 cases per million. This absence of correlation suggests that the spread of COVID-19 is not dependent on population density (Figure 16). The findings lead to the rejection of the hypothesis. The influence of population density on the number of COVID-19 cases can be mitigated by a variety of factors. Effective public health measures such as testing, contact tracing, and lockdowns, along with the population's adherence to health guidelines like social distancing and mask-wearing, play crucial roles. Additionally, the strength of healthcare infrastructure may influence a number of cases. The dataset showed that countries with comparable population densities experienced widely varying rates of COVID-19 cases per million, explaining the absence of a clear correlation between population density and the number of cases.

Hypothesis 7: The total number of COVID-19 cases is directly correlated with the total population of a country.
Explanation: This hypothesis explores the possibility of a direct correlation between the total population of a region and the cumulative number of COVID-19 cases. The aim is to assess whether larger populations inherently experience higher absolute numbers of COVID-19 cases.

Findings: Our analysis indicates no significant correlation between the total population of an area and the total number of COVID-19 cases. This lack of correlation suggests that the overall number of cases in a region is not simply a function of its population size (Figure 17). The findings lead to the hy-

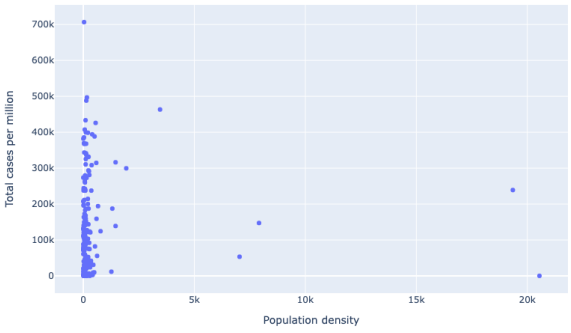

Fig. 16 This graph illustrates the relationship between population density and the total number of COVID-19 cases per million. [It shows no apparent correlation, indicating that high or low population density does not necessarily predict the rate of COVID-19 cases.]

pothesis not being accepted. A key reason for this outcome is the observation that most of the countries have populations under 50 million but exhibit a wide variance in case numbers, demonstrating a lack of consistent trend across different population sizes. Outliers such as China and India, the only countries with populations exceeding 400 million, also display significant discrepancies in their total case counts, further reinforcing our findings.

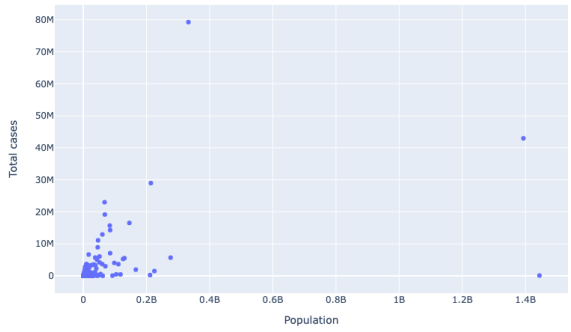

Fig. 17 This graph compares the total population of various regions with their respective total numbers of COVID-19 cases. [The absence of a clear pattern or correlation suggests that total population size does not directly influence the total case count.]

Discussion

In this research, we have extensively analysed various dynamics of the COVID-19 pandemic, employing a robust statistical methodology to challenge existing theories and uncover new insights. Our findings bring forth some unexpected results, particularly regarding the relationships between quarantine periods, population demographics, and the spread of the virus. These results are compared with similar studies to provide a broader context and validate the robustness of our conclusions.

Pandemic Progression and Health Outcomes

Our findings suggest an optimal quarantine period of 17 days, contrary to the widely adopted 14-day period. This is supported by Khadem Charvadeh, Yasin et al (2022), who also reported a longer quarantine period for COVID-19 cases based on empirical data[16]. A study conducted in India found similar results, suggesting upto a 28 day quarantine period in certain cases[17]. The initial two studies conducted in China, which focused on determining the incubation period of COVD-19, primarily included samples of 100 and 10 hospitalised patients, respectively[18,19]. These studies initially recommended a 14-day quarantine period based on their findings. However, the subsequent availability of more advanced data and enhanced statistical methodologies has allowed for a more accurate assessment of the optimal quarantine duration. Additionally, our finding that a person diagnosed with COVID-19 is most likely to be admitted to the hospital or ICU between 5-10 days complements another study conducted across 114 Belgian hospitals, where the observed time frame for hospitalisation was between 3 and 10.4 days20.

Testing and Vaccination Dynamics

A study conducted by Gaid, Michael & Salloum, Said in 2021 corroborates our findings, showing that testing rates do not significantly influence the total number of reported COVID-19 cases[20]. This aligns with our analysis which also found no substantial correlation between the extent of testing and the prevalence of the virus, suggesting that other factors may play more critical roles in the dynamics of the pandemic's spread. It is also important to note that the majority of research pertaining to COVID-19 vaccines focuses on the severity of the disease rather than the number of cases. This distinction is crucial as it highlights that while vaccines are highly effective at reducing hospitalizations and deaths, their immediate impact on the transmission rates and total case numbers might be less pronounced. This perspective supports our findings, suggesting that vaccination rates alone may not be a reliable indicator of changes in case numbers in the short term.

Demographic and Societal Factors

In line with our results, the study by Hamidi, S., Sabouri, S., & Ewing, R. (2020) found no direct link between population density and infection rates of COVID-19, proposing that COVID-19 dynamics is more significantly influenced by connectivity than population density[21]. Conversely, studies conducted in Algeria

and France documented a strong correlation between population density and infection rates in densely populated areas [22,23]. Additionally, the discrepancies between this study and those conducted in France and Algeria can primarily be attributed to differences in methodology. While the studies in France and Algeria focused on detailed, local data within each country at a microscopic level, our study encompassed a broader, global analysis, considering data from all countries worldwide.

Conclusion

In conclusion, our search, backed up by a robust dataset from Our World in Data and the World Bank, has provided a comprehensive analysis of the COVID - 19 pandemic using advanced statistical methodologies. By meticulously collecting and scrutinising data spanning key features such as infection rates, mortality and demographics, we developed a detailed understanding of the pandemic's dynamics. Our methodology which includes Linear Regression, Polynomial Regression, Ridge Regression, and Lasso Regression, allowed for a nuanced exploration of the data, revealing complex relationships and patterns.

In our study, we examined several additional hypotheses with intriguing outcomes. We discovered that an extended quarantine period, potentially up to 20 days, could be more effective than the standard 14 days. Additionally, the critical time frame for hospital or ICU admissions post-diagnosis is shorter than previously thought, at around 5-10 days. We found an unclear link between the number of COVID-19 tests conducted and new cases detected, suggesting that case numbers may be influenced by factors beyond just testing rates. Similarly, our analysis revealed no significant correlation between the volume of vaccine administration and the number of reported cases, pointing to more complex interactions in the efficacy of vaccination campaigns. Regarding demographic influences, we observed no direct relationship between population density or total population size and COVID-19 case numbers.

The findings from our study are pivotal in several ways. They offer a deeper understanding of the effectiveness of various public health measures, highlight critical periods in the progression of the disease, and underline the impact of demographic factors on the spread and severity of the virus. These insights are not only crucial for enhancing preparedness and response strategies for future pandemics but also serve as a valuable contribution to the global knowledge base in epidemiology and public health. Our research highlights the importance of data-driven analysis in tackling global health challenges and sets a precedent for future studies in the field.

Acknowledgements

I express my sincere gratitude to Lucien Werner for providing guidance at every step of this research paper. I am also thankful to my teachers and parents for nurturing my passion for pursuing data science.

References

1 S. Talic, S. Shah, H. Wild, D. Gasevic, A. Maharaj, Z. Ademi, X. Li, W. Xu, I. Mesa-Eguiagaray, J. Rostron, E. Theodoratou, X. Zhang, A. Motee, D. Liew and D. Ilic, *Effectiveness of public health measures in reducing the incidence of covid-19, SARS-CoV-2 transmission, and covid-19 mortality: systematic review and meta-analysis*, BMJ (Clinical research ed.), 375, e068302.

2 H. Zawbaa, A. El-Gendy, H. Saeed, H. Osama, A. Ali, D. Gomaa, M. Abdelrahman, H. Harb, Y. Madney and M. Abdelrahim, *International journal of clinical practice*, **75**, 14116.

3 J. He, G. Chen, Y. Jiang, R. Jin, A. Shortridge, S. Agusti, M. He, J. Wu, C. Duarte and G. Christakos, *The Science of the total environment*, **747**, 141447.

4 C. Liang and L. Chen, *Archives of gerontology and geriatrics*, **98**, 104587.

5 A. Tchicaya, N. Lorentz, K. Leduc and G. Lanchy, *PloS one*, **16**, 0256857.

6 *Our World in Data*, `https://ourworldindata.org/coronavirus`.

7 *The World Bank*, `https://data.worldbank.org/indicator/SP.POP.TOTL`.

8 *World Health Organisation*, `https://covid19.who.int/data`.

9 M. Ali, Peshawa and R. Faraj, *Data Normalization and Standardization: A Technical Report*.

10 B. Nussbaumer-Streit, V. Mayr, A. Dobrescu, A. Chapman, E. Persad, I. Klerings, G. Wagner, U. Siebert, C. Christof, C. Zachariah and G. Gartlehner, *The Cochrane database of systematic reviews*, **4**, 013574.

11 M. Cevik, J. Marcus, C. Buckee and T. Smith, *Clinical infectious diseases : an official publication of the Infectious Diseases Society of America*, **73**, 170– 176.

12 L. Lopes-Júnior, E. Bomfim, D. Silveira, R. Pessanha, S. Schuab and R. Lima, *BMJ open*, **10**, 040413.

13 F. Zhou, T. Hu, X. Zhang, K. Lai, J. Chen and X. Zhou, *Journal of infection and public health*, **15**, 499–507.

14 I. Mohammed, *Human vaccines immunotherapeutics*, **18,1**, 2027160.

15 D. Shukla, S. Bhadoria, M. Bansal and R. Changulani, *Journal of family medicine and primary care*, **11**, 1314–1321.

16 K. Charvadeh and Yasin, *Statistics in biosciences*, **14,1**, 175–190.

17 M. K. Das, *Cureus*, **14,5 e24999**, year.

18 C. Yeo, *The lancet. Gastroenterology hepatology*, **5,4**, 335–337.

19 Z.-D. Guo, *Emerging infectious diseases*, **26,7**, 1583–1591.

20 M. Gaid and S. Salloum, *Explore the Relationship Between COVID-19 Testing Rates with the Number of Cases*, 10.1007/978-3-030-76346-6_4.

21 S. Hamidi, S. Sabouri and R. Ewing, *Journal of the American Planning Association*, **86**, 495–509.

22 N. Kadi and M. Khelfaoui, *Bull Natl Res Cent*, **44**, 138.

23 R. Pascoal and H. Rocha, *Physica A: Statistical Mechanics and its Applications*, **593**, year.

Authors

Prakarsh Jhajharia is a high school science student at La Martiniere for Boys, Kolkata, India. Prakarsh is a dedicated data scientist with a goal to find solutions to various societal problems and misconceptions. His hobbies include participating in debates and Model United Nations.