BMC Medical Research
Methodology

RESEARCH Open Access

A data-driven approach to study temporal characteristics of COVID-19 infection and death Time Series for twelve countries across six continents

Sabyasachi Guharay[1]*

Abstract

Background In this work, we implement a data-driven approach using an aggregation of several analytical methods to study the characteristics of COVID-19 daily infection and death time series and identify correlations and characteristic trends that can be corroborated to the time evolution of this disease. The datasets cover twelve distinct countries across six continents, from January 22, 2020 till March 1, 2022. This time span is partitioned into three windows: (1) pre-vaccine, (2) post-vaccine and pre-omicron (BA.1 variant), and (3) post-vaccine including post-omicron variant. This study enables deriving insights into intriguing questions related to the science of system dynamics pertaining to COVID-19 evolution.

Methods We implement a set of several distinct analytical methods for: (a) statistical studies to estimate the skewness and kurtosis of the data distributions; (b) analyzing the stationarity properties of these time series using the Augmented Dickey-Fuller (ADF) tests; (c) examining co-integration properties for the non-stationary time series using the Phillips-Ouliaris (PO) tests; (d) calculating the Hurst exponent using the rescaled-range (R/S) analysis, along with the Detrended Fluctuation Analysis (DFA), for self-affinity studies of the evolving dynamical datasets.

Results We notably observe a significant asymmetry of distributions shows from skewness and the presence of heavy tails is noted from kurtosis. The daily infection and death data are, by and large, nonstationary, while their corresponding log return values render stationarity. The self-affinity studies through the Hurst exponents and DFA exhibit intriguing local changes over time. These changes can be attributed to the underlying dynamics of state transitions, especially from a random state to either mean-reversion or long-range memory/persistence states.

Conclusions We conduct systematic studies covering a widely diverse time series datasets of the daily infections and deaths during the evolution of the COVID-19 pandemic. We demonstrate the merit of a multiple analytics frameworks through systematically laying down a methodological structure for analyses and quantitatively examining the evolution of the daily COVID-19 infection and death cases. This methodology builds a capability for tracking dynamically evolving states pertaining to critical problems.

Keywords COVID-19 time series, Heavy-Tailed distribution, Stationarity, Co-integration, Self-Affinity studies, Hurst exponents, Detrended Fluctuation Analysis, State transitions

*Correspondence:
Sabyasachi Guharay
sguhara2@gmu.edu
Full list of author information is available at the end of the article

Background

The coronavirus disease, COVID-19, struck the globe, with apparently no preparedness for tracking and mitigation. On March 11, 2020, the World Health Organization (WHO) declared the COVID-19 as a "global pandemic" [1]. The RNA virus, labelled as SARS-CoV-2 (Severe Acute Respiratory Syndrome Coronavirus 2), infected over half a billion people worldwide and caused approximately seven million deaths [2]. This global pandemic has generated enormous research activities drawing in knowledge from diverse disciplines in sciences and engineering. Researchers have been pursuing activities with the goal to learn about the properties of this new evolving virus, including its temporal dynamics.

A vast literature exists addressing diverse problems on COVID-19 ranging from fundamental virology properties to mathematical and statistical analyses, modeling and simulation, data inspection, evidence-informed decision making, and many others. A few recent articles and references therein provide a glimpse of the diversity in research activities, especially from the viewpoint of quantitative methodology for COVID-19 studies [3–10]. The research work broadly focuses in three key areas: (1) genomic characteristics of the SARS-CoV-2; (2) forecasting of the COVID-19 pandemic using machine learning methods; (3) nowcasting of the COVID-19 pandemic, in several discrete areas of the world. In this pursuit, virologists have made rapid progress on understanding the genomic characteristics of the SARS-CoV-2 through gene sequencing. Scientists and engineers have studied questions related to the mathematical composition of the gene sequences using fractal techniques [11–14] – these studies report unique signals that exhibit special attributes for the gene sequences for SARS-CoV-2. In the context of forecasting and nowcasting, model-based data science techniques, have been implemented for a few specific geographic regions [15–21]; these studies include time-series analyses and machine-learning methods, such as Random Forest, Ridge Regression, Support Vector Machines, and ensemble regression models based on deep learning architectures. From the epidemiological viewpoints, models are developed that expand upon the Susceptible-Exposed-Infectious-Recovered (SEIR), time dependent susceptible-infection-recovered (TSIR) and other differential equations-based frameworks. These models mimic the way the COVID-19 disease spread and thus simulate future transmission scenarios under various assumptions [22–28]. Recent literature points to empirical evidence suggesting that there is no universal method that can accurately forecast pandemic data [29]. Work on nowcasting for the COVID-19 diseases reports policy effects [30], along with the temporal dynamics [31]. Some basic studies report multifractal characteristics of the US daily COVID-19 cases to predict short-range behaviors [32]. The epidemiological dynamics based on human mobility have been reported [33]. Several articles discuss general modeling landscapes in this area [34, 35]; and these literatures and references therein identify key unsolved problems in this domain. A recent article reports a case study using seven quantitative indicators ranging from entropy to higher moments and discusses forecasting the transition from endemic to epidemic phases for contagious diseases, specifically for COVID-19 [36].

This article analyzes the dynamics of the daily number of new infection and death cases by sampling countries across the six inhabited continents of the globe. Unique to this work is the implementation of multiple independent quantitative techniques to study the characteristics of infection and death time series datasets. This work attempts to answer several key questions relevant to the COVID-19 pandemic evolution.

(a) What are the key fundamental statistical attributes of infection and death time series pertaining to this global pandemic?

(b) Can self-affinity paradigms, commonly used in non-linear dynamics, be efficiently implemented for analyzing the COVID-19 time series data?

Does this approach enable deriving new information, especially, characteristic metrics for tracking the state of dynamical systems?

(c) Can the answers in (a) and (b) be extended to gain insights into questions related to universality and scalability, i.e., information invariance?

We investigate data characteristics and trends over diverse geolocations to systematically probe into the above questions and gain insights into the evolution of the COVID-19 pandemic. Using time series data from the beginning of the pandemic, namely January 22, 2020 till March 1, 2022 (in many instances, coinciding with the fall of the number of infection cases in the omicron wave BA.1 variant), we break down the analysis into three distinct time windows, namely, Time Window 1 (TW1), Time Window 2 (TW2), and Time Window 3 (TW3). The window TW1 covers time from the beginning till March 1, 2021 – this is up to the time when most of the world's population did not have access to vaccines; all three vaccines, namely Johnson & Johnson, Pfizer and Moderna were available to the public around the end of TW1 [1]. The time window TW2 includes the time till October 10, 2021 when the vaccines became available to the general public; the delta variant of the SARS-CoV-2 virus then prevailed across the globe. Finally, the time window TW3 extends TW2 to cover prevalence of the omicron variant, and this includes time till March 1, 2022. We denote the entire time period as Full Window (FW). Uniqueness to these three time windows can be noted from the standpoints of natural immunity (primarily without vaccines), impact of vaccines in the presence of emergence of different variants, and other factors (that may be primarily dealing with socio-economic issues, human behaviors/practices of life, and policies as well).

This work establishes a path for applying four specific orthogonal analytical techniques in a toolbox, each with unique merits:

(1) Statistical studies pertaining to the higher moments in the large time series data sets;
(2) Examine the stationarity properties in the time evolving infection and death data over different time domains;
(3) Co-integration studies to identify relationships in in non-stationary time series data sets;
(4) Self-affinity studies (widely used in nonlinear dynamics) to gain insights into the dynamical behavior of infection and death data sets. Weaving through the knowledge gained from each technique enables addressing several key questions that are asked in this work. This aspect of using multiple analytics strengthens the uniqueness and merit of this work on methodology. The specific contributions of this article are highlighted below.
We examine the higher moments, namely, the skewness and kurtosis, to understand the symmetry and tail behavior in the data. Through data stationarity and co-integration studies, we investigate whether the infection and death time series structurally move in the same manner.
We study the self-affinity characteristics of both the infection and daily death time series by calculating both the Hurst exponent from the rescaled range (R/S) and the Detrended Fluctuation Analysis (DFA) scaling exponent. This investigates any shift of trends, of the characteristic exponent metrics with the evolution of the pandemic.

The methodologies, their implementations and the results reported here build a framework that can be applied to other geographical regions and also to different time window specifications.

In Sect. "Materials & Methods" on materials and methodology, we first discuss the datasets for the COVID-19 pandemic and the basic mathematical transformations for analyses. This is followed by descriptions of the statistical analysis methodologies and the methodologies for self-affinity studies. These disparate quantitative methodologies form the crux of the analytics toolbox. Section 3 provides results using the analytics toolbox, and it also simultaneously includes related discussions. Conclusions, including the key limitations of the analytical techniques and the merit of the analytics approach taken here, are narrated in Sect. 4; the areas for future research, especially in the context of analytics for dynamically evolving diseases are highlighted as well.

Materials & Methods

We begin by describing the materials used in this study, namely, the COVID-19 data and relevant analytical transformations. We proceed next by describing the methodologies for determing statistical characteristics of the time series data. The final subsection involves describing the methodology for studying the self-affinity characteristics of the time series data.

Materials: COVID-19 Data

Numerous publicly available sources have reported data on the COVID-19 pandemic. Johns Hopkins University's (JHU) COVID-19 Data Repository [37] by the Center for Systems Science & Engineering is an enterprise data aggregator for this pandemic – it covers data from various sources, such as, the WHO, Center for Disease Control (CDC), and WorldoMeter [2]. These data repositories featured daily updates as well as corrections due to changes in data aggregations. It is worth noting that several researchers have studied the reliability of COVID-19 data sources, including the JHU data repository [38–40]; the aggregation and anomaly detection features in the JHU repository have been remarked to be up to the standards [39]. Detailed studies on comparisons of different data repositories and addressing questions related to missing data/data imputation and data augmentation have merits; however, this is beyond the scope of the current work. We use the R software package, namely, *covid19.analytics* [41] to access the JHU COVID-19 Data Repository for confirmed new cases of infections and deaths**.** Sufficient diversity in data is considered to demonstrate the merit of the multiple analytics approach taken here – the data span over all six inhabited continents of the world, namely, North & South America, Europe (including Scandinavia), Asia, Africa, Middle East and Oceania. We study the daily infection and death counts for United States of America (USA), Brazil, the United Kingdom (UK), Germany, Sweden, Israel, Japan, South Korea, New Zealand, South Africa, Nigeria, and India. In this study, the sample datasets are carefully selected through a set of rationales that point to broadness and diversities; this foundational work builds ground with sufficient illustrations so it can be adapted by researchers for further enhancements; either for studies of COVID-19 or for examining other time series data. The selection rationales and corresponding examples of the countries are the following:

(1) Choosing at least one country from each of the six inhabitable continents of Earth, including Asia, Europe, Africa, North America, South American and Oceania;
(2) Including countries where the number of infections and deaths are high with respect to their population such as US, Brazil, UK, Nigeria, Germany;
(3) Including countries with unique mitigation, such as Sweden (herd immunity), New Zealand (complete border lockdown), Japan and South Korea (adapting special social behavior such as universal mask usage);
(4) Including countries where onset of the COVID-19 variants began such as South Africa (Omicron) and India (Delta);
(5) Including countries with strong vaccination policies, such as Israel.

We have thus selected twelve countries across the six continents that regularly reported the daily infection and death data. It is emphasized that the application of the methodologies discussed in this paper are not limited to these twelve countries and can be adapted to other cases as well. We scrutinized the data and did not observe any missing data during the time periods used in this study, and thus have used all the available data from the repository [37] for the countries and time periods of interest. It is noteworthy that several authors have studied outlier detection with the COVID-19 time series space [42, 43]. These studies show that even using non-parametric methods, such as Local Outlier Factor (LoF) [42] or Compositional Functional Data Analysis (CFDA) [43], the results are not always clear. For countries like Iceland, Luxembourg, and Belarus, studies based on CFDA reported potential outliers [43]. These studies suggest that based on the underlying assumptions, different outlier techniques may yield different results. A systematic rigorous study of outliers in all of the COVID-19 data repositories can be intriguing and very valuable. It is important to note that many challenges can occur due to questions related to masking and swamping, especially in high dimensional data [44]. In any outlier identification and removal procedure, one key concern is that of the robustness with respect to possible misclassification errors, i.e., masking (Type I) and swamping (Type II) [45]. The problem of outlier detection is well acknowledged here, but it is beyond the scope of this work.

It is emphasized that the application of the methodologies discussed in this paper are not limited to these twelve countries and can be adapted to other cases as well. Table 1 shows a summary of the characteristics of the countries and the variables of interest along with the details of the time periods of data. Figure 1(a) - (d) show illustrations of the daily infection and death time series data across the full window (FW) time period: with infection and death data for USA in Fig. 1(a) and (b) and for Brazil in Fig. 1(c) and (d). Both USA and Brazil show several spikes throughout the two-year pandemic time period for the infection data. For the daily death counts, USA shows several peaks, while Brazil data exhibit a couple of peaks. Since many analytical methods rely on stationarity of the data, we scrutinize this point and transform datasets appropriately, to satisfy the fundamental requirements for respective analytical methods.

Differencing a non-stationary dataset generally renders stationarity in the transformed data. Economists often use log returns instead of raw price data to assess and exploit the data stationarity [46]. The log returns have a natural interpretation, in that it models the percent change. We calculate the daily log return defined as: $\mathrm{r_t} = \log\left(\frac{\mathrm{C_t}}{\mathrm{C_{t-1}}}\right)$, where $\mathrm{C_t}$ is the count of the daily number of cases (either infections or deaths) at time t and the time interval unit is one day. In Figs. 2(a) through 2(d), we show sample time series plots of the calculated daily log return of the time series data that are shown earlier in Figs. 1(a) through 1(d). From a visual inspection of these time series, the data appear to be stationary. However, we formally test this proposition and discuss it in the next section.

Methodologies for Examining Statistical Characteristics of Time Series Data

To understand the complexity in the time series data, we first study the higher moments [47–50], namely, the skewness and the kurtosis. The excess kurtosis (> 3) provides evidence of heavy or thin tails in the data, while the skewness addresses questions related to the symmetry of the data distribution. We address questions on stationarity of the datasets. Note that this test is particularly critical for co-integration studies as well as for self-affinity analysis, especially by Hurst exponent estimate methodologies. Although well-known in the statistics community, it is important to briefly discuss the basic premises of the methodologies to, study the time series data, here for COVID-19 infection and death counts data.

Fundamental statistical property studies: Skewness & Kurtosis

The commonly used metrics for assessing the basic statistical characteristics involve examining the moments of data distributions. The first and second moments, namely, the arithmetic mean and the variance are the measures indicating the centrality and dispersion in the data, respectively. Higher-order moments, namely, the third and fourth moments, provide further significant geometric interpretations of the data distributions.

Table 1 Summary of the characteristics of the countries (along with respective regions) and the variables of interest. Note: FW is not included here as this is a natural aggregation of all other time periods. TW1 ranges from January 22, 2020 till March 1, 2021. TW2 spans from March 2, 2021 till October 10, 2021. TW3 covers from March 2, 2021 till March 1, 2022. Note that the total number of observations for each time period of interest are the same for each country

Region	Country	Variable of Interest	Time Period of Interest		
North America	United States of America (USA)	Daily Infection Counts	TW1	TW2	TW3
North America	United States of America (USA)	Daily Death Counts	TW1	TW2	TW3
South America	Brazil	Daily Infection Counts	TW1	TW2	TW3
South America	Brazil	Daily Death Counts	TW1	TW2	TW3
Europe	United Kingdom (UK)	Daily Infection Counts	TW1	TW2	TW3
Europe	United Kingdom (UK)	Daily Death Counts	TW1	TW2	TW3
Europe	Germany	Daily Infection Counts	TW1	TW2	TW3
Europe	Germany	Daily Death Counts	TW1	TW2	TW3
Scandinavia	Sweden	Daily Infection Counts	TW1	TW2	TW3
Scandinavia	Sweden	Daily Death Counts	TW1	TW2	TW3
Middle East	Israel	Daily Infection Counts	TW1	TW2	TW3
Middle East	Israel	Daily Death Counts	TW1	TW2	TW3
East Asia	Japan	Daily Infection Counts	TW1	TW2	TW3
East Asia	Japan	Daily Death Counts	TW1	TW2	TW3
East Asia	South Korea	Daily Infection Counts	TW1	TW2	TW3
East Asia	South Korea	Daily Death Counts	TW1	TW2	TW3
South Asia	India	Daily Infection Counts	TW1	TW2	TW3
South Asia	India	Daily Death Counts	TW1	TW2	TW3
Oceania	New Zealand	Daily Infection Counts	TW1	TW2	TW3
Oceania	New Zealand	Daily Death Counts	TW1	TW2	TW3
Africa	South Africa	Daily Infection Counts	TW1	TW2	TW3
Africa	South Africa	Daily Death Counts	TW1	TW2	TW3
Africa	Nigeria	Daily Infection Counts	TW1	TW2	TW3
Africa	Nigeria	Daily Death Counts	TW1	TW2	TW3

Skewness, the third-order moment, provides the degree of asymmetry of the distribution. The sample skewness, *skew*, is defined as the following for a univariate dataset $x_1, x_2, \ldots, x_N$ with a mean of $\langle x \rangle$ and sample standard deviation s:

$$skew = \sum_{i=1}^{N} \frac{[x_i - \langle x \rangle]^3/N}{s^3}$$

If $skew < 0$, it indicates left-skew (left shifted) in the data while $skew > 0$ indicates right skew (right shifted), and $skew = 0$ indicates a full symmetry in the data. The normal distribution has a skewness of zero.

The sample excess kurtosis, k, is defined as the following for univariate dataset $x_1, x_2, \ldots, x_N$ with a mean of $\langle x \rangle$ and sample standard deviation s:

$$k = \sum_{i=1}^{N} \frac{[x_i - \langle x \rangle]^4/N}{s^4} - 3$$

This definition is used to compare with the standard normal distribution that has a kurtosis of three. Note that $k > 0$ indicates a heavy-tailed or leptokurtic distribution. If $k < 0$, indicates a light tailed or platykurtic distribution; finally, $k = 0$ indicates a mesokurtic distribution, such as the normal distribution.

Stationarity property studies

There are analytical tools which rely on the fundamental assumption of stationarity in the dataset [46]. Strong stationarity requires that the data generation process for the time series have an unconditional joint probability distribution which does not change when shifted in time. Mathematically speaking, if $\{X_t\}$ is a stochastic process and $F_X\left(x_{t_{1+\tau}}, \cdots, x_{t_{n+\tau}}\right)$ represent the cumulative distribution function (CDF) of the unconditional joint distribution of $\{X_t\}$ at times $t_{1+\tau}, \ldots, t_{n+\tau}$. then, $\{X_t\}$ is strongly stationary if the following holds true [46]:

$$F_X\left(x_{t_{1+\tau}}, \cdots, x_{t_{n+\tau}}\right) = F_X\left(x_{t_1}, \cdots, x_{t_n}\right)$$

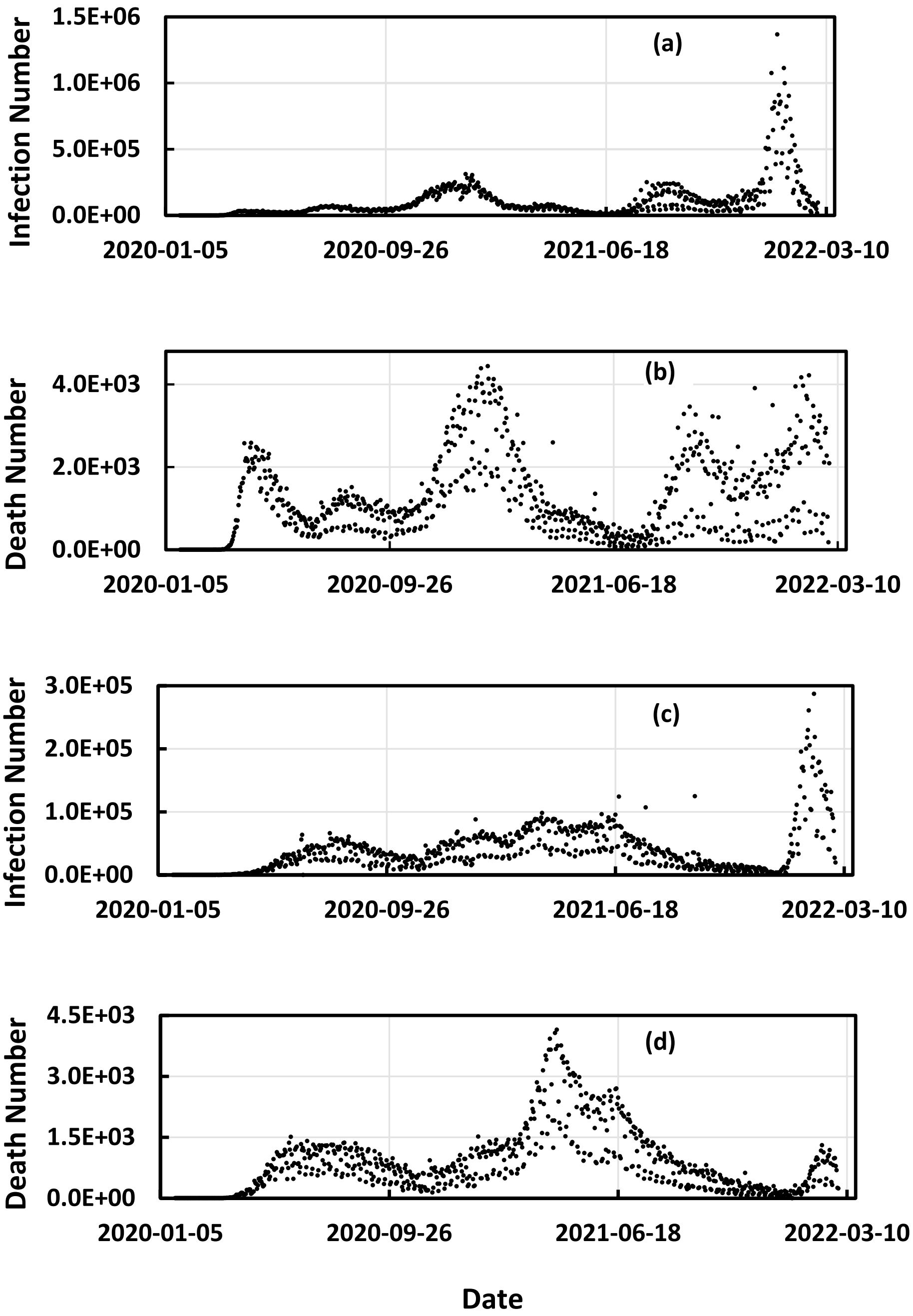

Fig. 1 Time series plots of the confirmed daily infection and death counts for USA and Brazil: (**a**) The top panel shows the daily infection counts for USA across FW; (**b**) The second panel shows the USA daily death counts across the FW; **(c)** the third panel shows the daily infection counts in Brazil across FW; **(d)** the bottom panel shows the daily death counts in Brazil across the FW. Note, the two countries of USA and Brazil are selected as illustrative examples

A more practical definition of stationarity is that of the weak stationarity or covariance stationarity as testing the strong stationarity condition is impractical for real-world datasets [51]. Weak stationarity requires that mean and the auto-covariance function do not vary with respect to time and that the variance is finite.

Unit root and stationary tests determine if a time series is non-stationary. These methodologies include Phillips Perron (PP), Augmented Dickey Fuller (ADF), Kwiatkowski-Phillips-Schmidt-Shin (KPSS) tests [52, 53]. Each of these statistical hypothesis tests has its own strengths and weaknesses. For smaller data sets,

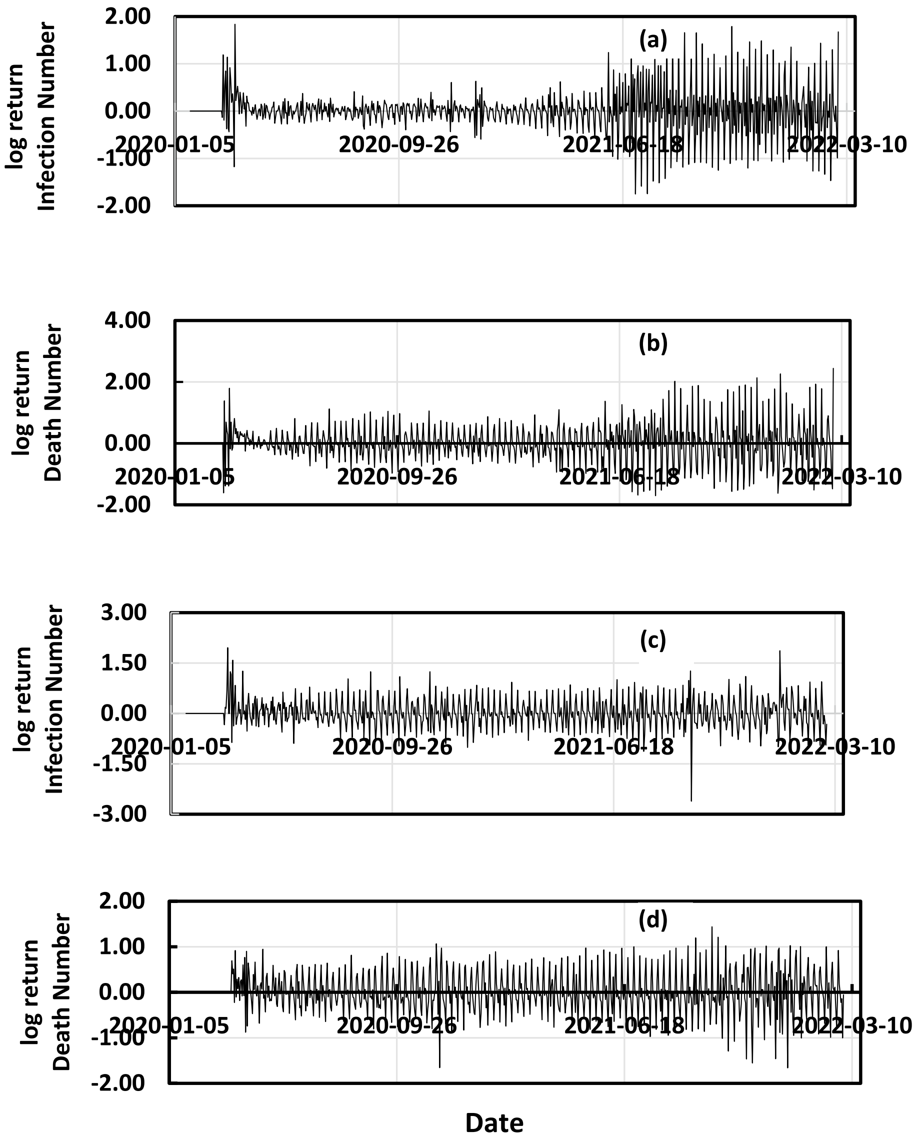

Fig. 2 Time series plots of the daily infection and death log return data for USA and Brazil – the raw time series data shown in Fig. 1 to calculate corresponding log return values: (**a**) The top panel shows the daily infection log return for USA across FW; (**b**) The second panel shows the USA daily death log return across the FW; **(c)** the third panel shows the daily infection log return in Brazil across FW; **(d)** the bottom panel shows the daily death log return in Brazil across the FW. Note, the two countries of USA and Brazil are selected as illustrative examples

the ADF test is simple and most useful [54]. In this work, we employ the ADF test where the null hypothesis indicates the presence of a unit root and the alternate hypothesis indicates stationarity in the data. For the lag length, we use automatic lag length selection optimization based on the modified Akaike information criterion (MAIC) [55]. We implement the *CADFtest* package in R to perform this calculation [56].

Co-Integration relationship studies

Co-integration was first introduced to study the long-run relationships among variables, primarily in economics [57]. The main objective is to check if two or more non-stationary time series are statistically correlated. To eliminate any “spurious” correlation, a rigorous statistical hypothesis test is conducted to evaluate if a vector of coefficients exists to form a stationary linear combination of both non-stationary time series. If this exists, then the two individually non-stationary time series are said to be co-integrated [58]. If two non-stationary time series are co-integrated, they have a long-run equilibrium and share an underlying common stochastic trend. In epidemiology, joint mortality models have been used to imply a long-run relationship between mortality rates across different demographics using co-integration [59]. There are several different methods for statistical testing for co-integration, for example the Engle-Granger two-step method, Johansen test and Phillips-Ouliaris (PO) test [46]. While the Johansen test requires large sample sizes, this is not required for PO [60]. We, therefore, use the PO test. In this context, we study if the non-stationary daily infection counts data are co-integrated with the non-stationary daily death data for a particular lag. Specifically, we are testing if the death counts move with the same trends or are correlated with the infection counts. Epidemiologists have observed a seven-day lag between the daily infection counts and the daily death counts [61, 62]. Higher positive lags between deaths and infections are plausible, and its impact can be evaluated following the implementation steps of the methodologies laid down in this work.

Methodologies for examining self-affinity characteristics of time series data

Understanding the self-similarity and long-range correlation structures in time series datasets provides insights into the inherent structural dynamics of the epidemiological problem. In this context, several questions may occur:

(a) If the viral dynamics follow a path with a mean-reversion trend; or
(b) If the viral dynamics follow a random path; or
(c) if the viral dynamics approach a long-term memory state.

Self-affinity characteristics can serve as a guide to obtain answer to the above questions. Two methods are used in this work: (1) R/S analysis to compute the Hurst exponent for stationary time series; (2) Detrended Fluctuation Analysis (DFA) scaling exponent α to measure the long-range correlation structure in any type of time series.

Hurst exponent calculated from the R/S analysis

The Hurst exponent is a technique for detecting the underlying temporal dependency. Hurst [63] proposed the rescaled range (R/S) analysis to study the scaling behaviors of complex systems. Given a stationary time series $X \equiv \{X_t : t = 1, 2, \ldots, n\}$ with mean $\langle x \rangle_n$ and variance $S^2(n)$, one can proceed forward to calculate the R/S ratio.

Defining *R(n)* and *S(n)* as

$$R(n) = \underset{1 \le i \le n}{max}\, X(i, n) - \underset{1 \le i \le n}{min}\, X(i, n)$$

$$S(n) = \sqrt{\left[\frac{1}{n}\sum_{i=1}^{n}\left(x_i - \langle x \rangle_n\right)^2\right]}$$

The ratio of the *R(n)/S(n)* follows [63, 64]:

$$\frac{R(n)}{S(n)} \propto \left[\frac{n}{2}\right]^H$$

where H is the Hurst exponent for the stationary time series *X*.

Theoretically, when H > 0.5, the time series is expected to have long-range correlation and show persistence and long-memory. If H < 0.5, then the data in the time series are anti-correlated and mean-reverting. H = 0.5 indicates the time series behaves like white noise. There are other ways to estimate the Hurst exponent, including multifractal detrended fluctuation analysis, detrending moving average, and the generalized Hurst exponent approach. For heavy-tailed distributions with limited data size, the rescaled range analysis (R/S) is shown to be a robust approach [64] for estimating the Hurst Exponent.

Detrended Fluctuation Analysis (DFA) and Scaling Exponent

DFA is a relatively newer method. This technique was first used while examining correlations in DNA sequences [65]. Unlike the R/S analysis (which uses the range function), the DFA uses the squared fluctuations around the trend of the signal as a measure of dispersion. This method involves detrending of the sub-periods, and thus can be used for both stationary and non-stationary time series.

Given that a time series Y_t of length N is scaled by its arithmetic mean $\langle Y \rangle$, its cumulative sum is calculated as the following:

$$Z(j) = \sum_{i=1}^{j}[Y(i) - \langle Y \rangle]$$

The cumulative sum series, *Z(j)* is next segmented into time windows of different lengths *ΔT*, which yields a set of random walks of varying sizes. Then, these random walks are detrended within the windows by locally estimating their trends as best fit polynomials of order p, $Z^{(p)}_{\Delta T}(j)$. Typically, a fit of $p = 1$, namely, a linear fit, is done, which is referred to as DFA-1 [66]. These trends are removed using mean squared deviations which results in the squared fluctuation function as shown below:

$$F^{(p)}(\Delta T) = \sqrt{\left[\frac{1}{N}\sum_{j=1}^{N}\left(Z(j) - Z_{\Delta T}^{(p)}(j)\right)^2\right]}$$

Then similar to the R/S analysis, if the time series has long range correlations, it would be reflected in the power-law relationship of the fluctuation function, $F^{(p)}(\Delta T)$, as the following:

$$F^{(p)}(\Delta T) \propto [\Delta T]^{\alpha}$$

The scaling exponent, α, is estimated by the best fit of a log–log plot using linear regression [65]. The exponent with $0 < \alpha < 0.5$, indicates anti-correlation or mean reversion property of the time series [67]. For white noise, $\alpha = 0.5$. If $1/2 < \alpha < 1$, it indicates persistent long-range correlations. If $\alpha = 1$, then the time series behaves like 1/f noise. If $\alpha > 1$, the time series is non-stationary. For $\alpha = 1.5$, the non-stationary time series corresponds to Brownian noise [68]. There has been a recent study on the effectiveness of the DFA scaling exponent on short scales [69], and how that can impact the results.

Results and Discussions

Statistical Characterisitics of COVID-19 Infection and Death Time Series

Following the methodologies discussed in Sec. "Materials & Methods", the key results are presented and discussed in the subsections below.

COVID-19 data distribution characteristics

Figure 3 (a) – (c) show three different types of distributions for the purpose of illustrations. All three panels in Fig. 3 exhibit a positive skew. Note the presence of a heavy-tail in the USA data (Fig. 3(a)), while the Brazil

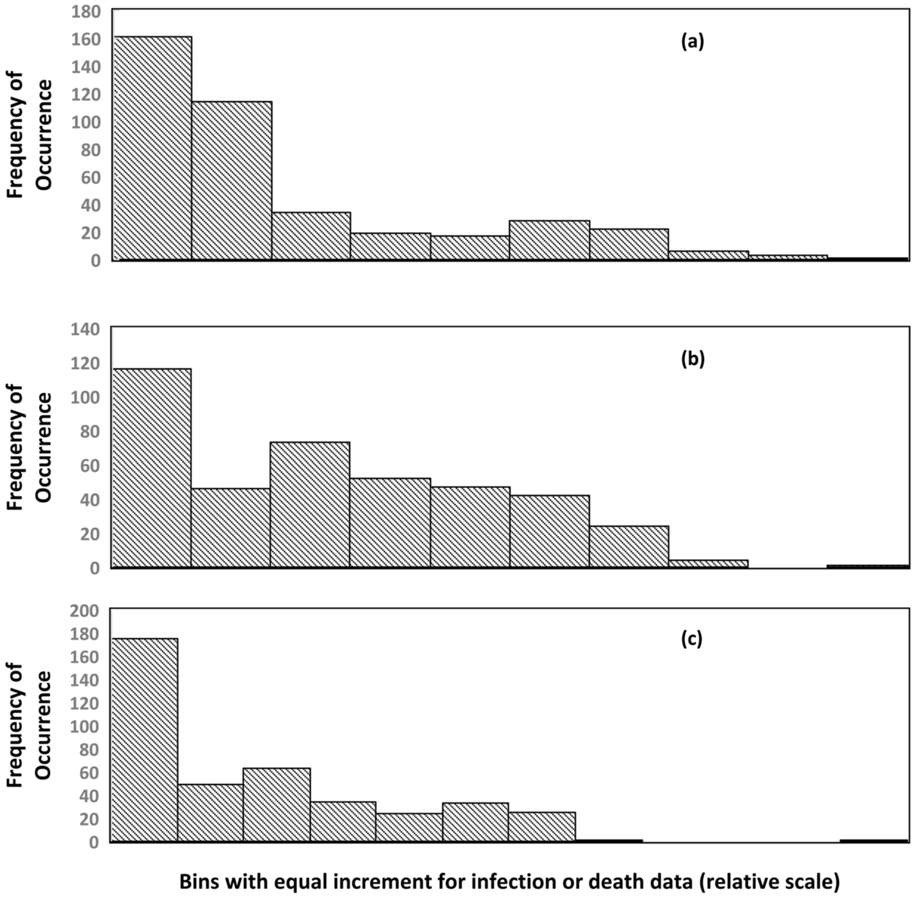

Fig. 3 Notional illustrations for histogram plots showing different distributional characteristics of the data. (**a**) the top panel shows a leptokurtic distribution – here, the daily infection counts for USA across TW1;(**b**) the middle panel shows a platykurtic distribution – here, the daily infection counts for Brazil across TW1; **(c)** the lower panel shows a mesokurtic distribution – here, the daily death counts for India across TW1. Note that the distribution is showing a heavy tail for US infection data, no significant tail for India and a light-tail for Brazil

data in Fig. 3(b) shows a lighter tail. For Fig. 3(c) (for daily death in India) we do not observe the presence of either a heavy or a thin tail.

The skewness and kurtosis values for the three characteristic time windows are shown in Figs. 4 and 5. A summary of these results is included in the Tables 1 and 2 in the Appendix, for infection and death time series data, respectively. A right-sided skew in both the daily confirmed infection and death count datasets is universally observed across all time windows for all twelve countries. This observation implies that the arithmetic mean exceeds the median for all these datasets. Thus, there are a higher number of large values (daily infection/death counts) than low values. Regarding the fourth moment, none of the countries exhibits a mesokurtic type of distribution for the confirmed daily infection counts, i.e., a zero-excess kurtosis. Approximately 10% of the daily infection counts data show platykurtic distributions, while the remaining datasets exhibit heavy-tailed data, namely leptokurtic distributions. For daily new death counts, India uniquely shows close to a mesokurtic distribution, in that excess kurtosis $| k | \leq 0.1$ in the TW1 period. For the remaining countries, over 90% show heavy-tail distribution properties. Brazil and UK are unique in that they both exhibit a platykurtic distribution for both infection and death counts data. Brazil shows this for TW1 and TW2, while UK shows it for the TW2 period only. South Korea showed a platykurtic distribution for daily infection counts data for TW2. Platykurtic distributions indicate that there is less likelihood of observing extreme low/high events. Most noticeably, extremely high kurtosis values ($\gtrsim 60$) are noted in infection data for UK, Israel, New Zealand, and Nigeria during TW3 – this indicates thin peaks with very heavy

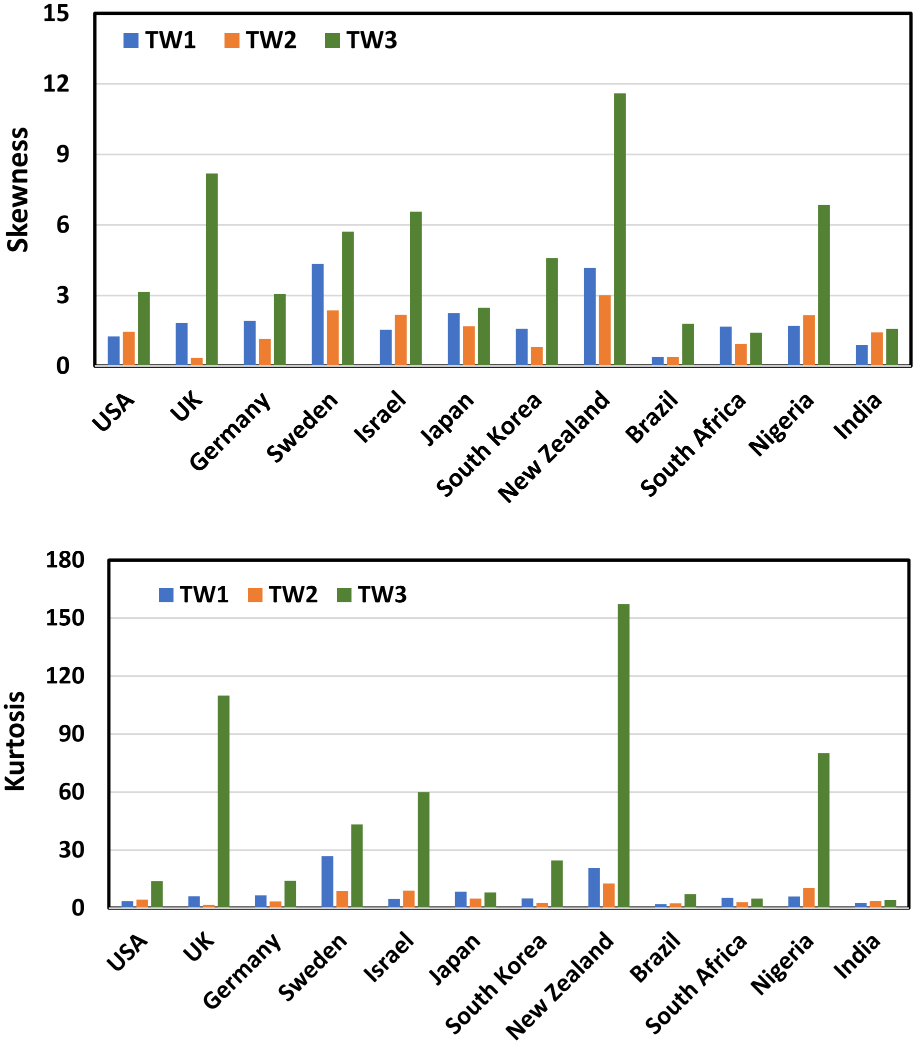

Fig. 4 Bar graphs for skewness (top figure) and kurtosis (bottom figure) for raw infection time series. Results are shown for three distinct time windows, namely, TW1, TW2 and TW3 – the countries are labelled along the horizontal axis

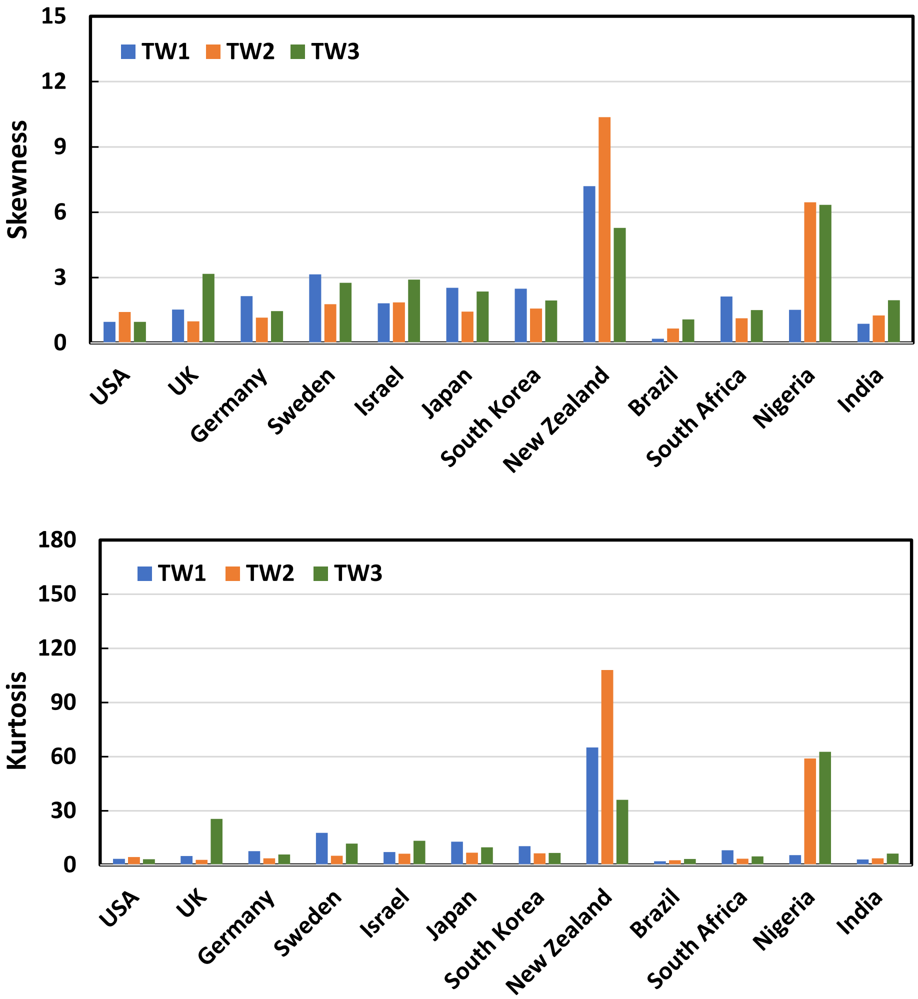

Fig. 5 Bar graphs for skewness (top figure) and kurtosis (bottom figure) for raw death time series. Results are shown for three distinct time windows, namely, TW1, TW2 and TW3 – the countries are labelled along the horizontal axis

tails– see Fig. 4. We note from Fig. 5 that the death data indicate similar high kurtosis values for New Zealand during TW1 and TW2 and for Nigeria during TW2 and TW3. Note we observe the presence of heavy right-side shifted distributions, with skewness ($\gtrsim 6$) for: (a) UK and Israel in infection data (TW3), and (b) New Zealand and Nigeria in both infection (TW3) and death data (TW1 and TW2).

Overall, we observe the overwhelming presence of heavy tails in both the infection and death data sets. While examining skewness and kurtosis for the log return data, we note kurtosis values close to or exceeding 60, i.e., signature of extreme heavy tails in the infection data for UK (TW1 and TW3) and in the death data for New Zealand (TW1 and TW2). For all other cases, we note much smaller kurtosis values. The distributions, in general, are much less shifted from the center (symmetry) in all cases (compared to corresponding cases for the raw data). In terms of distinguishing the COVID-19 data for the twelve countries from the study of the higher moments, we find that both skewness and kurtosis, in general, have limited capabilities to find distinct signatures (see Figs. 4 and 5). An intriguing common theme noted in these statistical metrics of the data is that over 90% of the countries are leptokurtic and have positive skew values. Note notionally similar understandings in the context of forecasting [36], where the authors report the higher moments, namely, skewness and kurtosis, bearing a low predictive power.

Stationarity of COVID-19 time series data

For the ADF stationarity tests, we use a standard 95% confidence level to test the null hypothesis for the presence of a unit root. The key results are summarized below – see additional detailed results with judgement of stationarities for each case in Tables 3 through 6 in

Table 2 Co-integration results using the Phillips-Ouliaris (PO) test between daily infection data and daily death data with a positive seven-day lag in the FW time period. The null hypothesis indicates that there is no co-integration, and we use a 95% confidence level. DD* represents Death Data with a positive lag of seven days. Note: a – represents statistically significant co-integration in FW; b – represents statistically significant co-integration in TW1; c – represents statistically significant co-integration in TW2; d – represents statistically significant co-integration in TW3

Time Windows		**DD***	**DD***	**DD***	**DD***	**DD***	**DD***	**DD***	**DD***	**DD***	**DD***	**DD***	**DD***
		USA	**UK**	**Germany**	**Sweden**	**Israel**	**Japan**	**South Korea**	**New Zealand**	**Brazil**	**South Africa**	**Nigeria**	**India**
Infection Data	USA	(a,b,c,d)											
	UK		(a,b,d)										
	Germany			(a,b,c,d)									
	Sweden				(a,b,c,d)								
	Israel					(a,b,c,d)							
	Japan						(b)						
	South Korea							(c)					
	New Zealand								(b)				
	Brazil									(a,b,c,d)			
	South Africa										(a,b,c,d)		
	Nigeria											(a,b,c,d)	
	India												(a,b)

the Appendix. Note that both raw time series data as well as corresponding log return data are systematically investigated.

(i) For raw infection daily counts data, South Korea showed a stationary behavior in only TW1; the data for all other countries showed non-stationary behavior during all time windows, namely, FW, TW1, TW2, and TW3.

For infection log return data, all countries showed a stationary signature during all time windows, namely, FW, TW1, TW2, and TW3.

(ii) For death daily counts data, only New Zealand showed stationary behavior in FW, and TW2. In all other time windows, New Zealand daily infection counts did not show stationary behavior. All other countries showed non-stationary behavior for the daily death data during all time windows, namely, FW, TW1, TW2, and TW3.

For death log return data, all countries showed stationary signature during all time windows, namely, FW, TW1, TW2, and TW3.

The stationary behavior for infection in South Korea during TW1 can be possibly due to strict mitigation, and this policy may have affected its low infection counts prior to the vaccine [70]. For confirmed daily death counts data, the observation of stationarity across the FW and TW2 for New Zealand can be attributed to death counts of zero during the post-vaccine pre-omicron period. New Zealand was probably able to avoid the brunt negative effects of the pandemic due to its geographic isolation and strict lockdown [71]. With the start of omicron variant, New Zealand's daily death count behavior changes, and this is reflected in our observation that corresponding time series during TW3 is non-stationary (see Table 4 in the Appendix).

Co-integration between nonstationary infection and death time series for structural similarities

As discussed in the methodologies section, the basics for co-integration stand on evaluating the linear combination of two or more nonstationary time series and examine if this results in a stationary time series, i.e., the classic I(1) (integrated of order 1) to I(0) (integrated of order 0) transformation in time series theory [46]. We have examined co-integration for different lag values between the infection and death time series, namely, 7, 14 and 21 days. Table 2 shows the results across the FW, TW1, TW2 and TW3 time windows for all countries, where both the infection and death counts are non-stationary. Note that New Zealand is not included for FW, as the New Zealand death counts are stationary. Similarly, for TW1, South Korea is omitted, as the infection data is stationary in this time period. During the TW2 time period, New Zealand daily death time series are observed to be stationary and are omitted. Note that Japan and South Korea do not show any statistically significant co-integration relationship. Epidemiologically speaking, this indicates that the seven-day lagging between death and infection cases does not reveal statistically significant similarities in the FW time period. Except New Zealand, all other countries show statistically significant co-integration between the daily infection and daily death data with a lag of seven days for the FW time period. Next, during TW1 when the vaccines were not available (except for limited cases), we observe that except South Korea (not eligible due to data stationarity), all other countries show co-integration between the infection and the death data. During TW2, the countries, namely, UK, Japan and India do not show co-integration between their own infection data and death counts data. Except New Zealand (not eligible due to having stationary data), all other countries show co-integration. Finally, during TW3, the countries, namely, Japan, South Korea, India, and New Zealand, do not show co-integration between the daily infection counts and their respective daily death counts. These results demonstrate that the time window is highly sensitive in determining the long-range correlation between the non-stationary infection counts and the death counts. The seven-day lag period is not universally observed across all countries in the selected samples. For large countries like USA and Brazil, we do observe a statistically significant co-integration relationship between infection counts and death counts with a positive seven-day lag across all time windows. However, India does not show this same pattern as the delta-wave affected this country post vaccination [72, 73]. Further detailed investigations can be done for this area. During the TW3 time period (which includes the omicron variant), 1/3 of the countries in our sample did not exhibit statistically significant co-integration relationships between the infection counts and death cases with a positive seven-day lag. This underscores the importance of further detailed investigation on the direct causal relationship between infection and death cases for the omicron variant; however, this investigation is beyond the scope of this article. Further sensitivity analyses with varying lags, namely, 14 days and 21 days, are included in Tables 7 and 8 in the Appendix.

Self-affinity characteristics for dynamical system evaluation

Our goal in the self-affinity studies is to get insights into the temporal dynamics of the SARS-CoV-2 virus. Since the Hurst exponent estimated through the R/S analysis can only be calculated for stationary time series, we have studied the log return time series for both infection and death cases. To establish an analytical platform for a comparison of the Hurst exponent with alpha from DFA, the same log return datasets are analyzed for both cases.

Hurst exponent analysis for COVID-19 time series

In Fig. 6, we show the Hurst exponent results for daily infection counts and daily death counts across the FW time window. The exponent values for infection time series is on average higher than the corresponding death time series data. The countries, namely, Sweden, South Korea, and India show Hurst exponents greater than 0.5. This indicates that the percentage change for infection cases for these three countries have a long-range memory. The countries, namely, New Zealand, Nigeria and Brazil show a Hurst exponent of around 0.5, and thus indicates that the percent change in infection cases is Gaussian in nature. For the remaining six countries, namely, USA, UK, Germany, Japan, South Africa and Israel, the Hurst exponent is below 0.5, and thus it indicates trends towards mean-reversion or anti-correlation in the percent change of infection cases. For the Hurst exponents for log return death data, only Sweden shows a value around 0.5, and this indicates white noise behavior in the percent change in death cases. All other countries show Hurst exponent below 0.5. New Zealand shows the lowest value (<0.25); this supports trends towards a mean-reversion state. With this broad, macroscopic view of the Hurst exponent results across the FW time window, we now study the behavior in segmented time windows, namely, during TW1, TW2 and TW3.

Hurst exponents during characteristic segmented time windows and state transitions Since the overwhelming majority of the daily infection and daily death count time series data are non-stationary (see Sec. "Stationarity of COVID-19 Time Series Data" and Tables 3 and 4 in the Appendix), we do not calculate the Hurst exponent for those cases. Figure 7 shows Hurst exponent results from the R/S analysis.

First, Fig. 7(a) shows the Hurst exponent time evolution for the infection log return data for all twelve countries. Figure 7(b) shows the Hurst exponent time evolution for the death log return data. A line for H = 0.5 is drawn to indicate a baseline for white noise.

Note several dashed lines are drawn to illustrate the key points. For the USA daily infection log return data, a 40% decline in the Hurst exponent is noted from TW1 to TW2, and then a 14% rise again in TW3. For Brazil, a 30% decline is noted in the Hurst exponent from TW1 to TW2, and then a 16% rise in TW3. This can perhaps be explained by competing issues associated with the prevailing circumstances, namely, vaccines and emergence of a highly transmissible omicron variant—the omicron BA.1 variant caused a sharp rise in infection cases in the early part of 2022. For India, the Hurst exponent increased by 14% from TW1 to TW2. This can perhaps be due to the delta variant originated from India and caused severe problems with India's healthcare system in early 2021. However, the Hurst exponent decreased by 7% from TW2 to TW3 for India, as the delta wave gradually receded there. South Korea and Israel show similar trends. The Hurst exponent shows an increase of 14% from TW1 to TW2 and then another 3% increase from TW2 to TW3 for Israel. For most of the other countries, there has been a consistent trend where a steady decline in the Hurst exponent occurs between TW1 and TW2 and then attains almost a similar value between TW2 to TW3. The average Hurst exponent across the twelve countries shows about 6% decline from TW1 to TW2

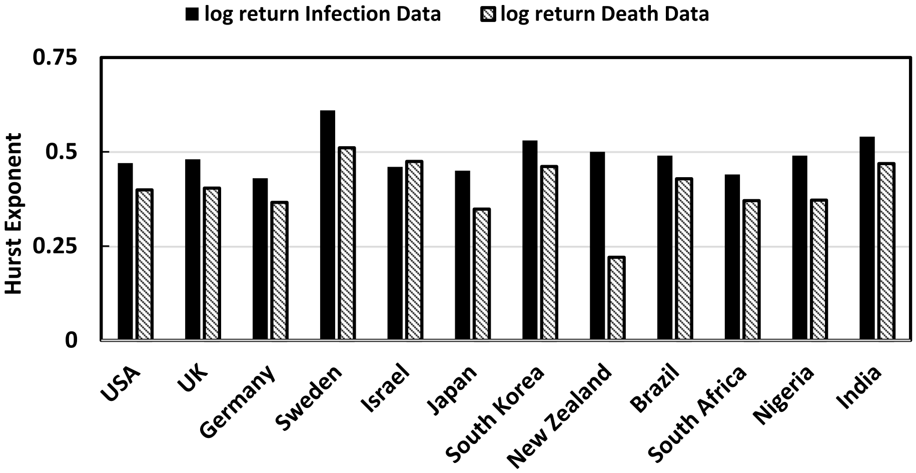

Fig. 6 Hurst exponent from R/S analysis for daily infection and death log return data for all countries – here the entire time series data covering January 22, 2020 to March 1, 2022 are used. Note distinct differences in the Hurst exponent values between log return infection and death datasets

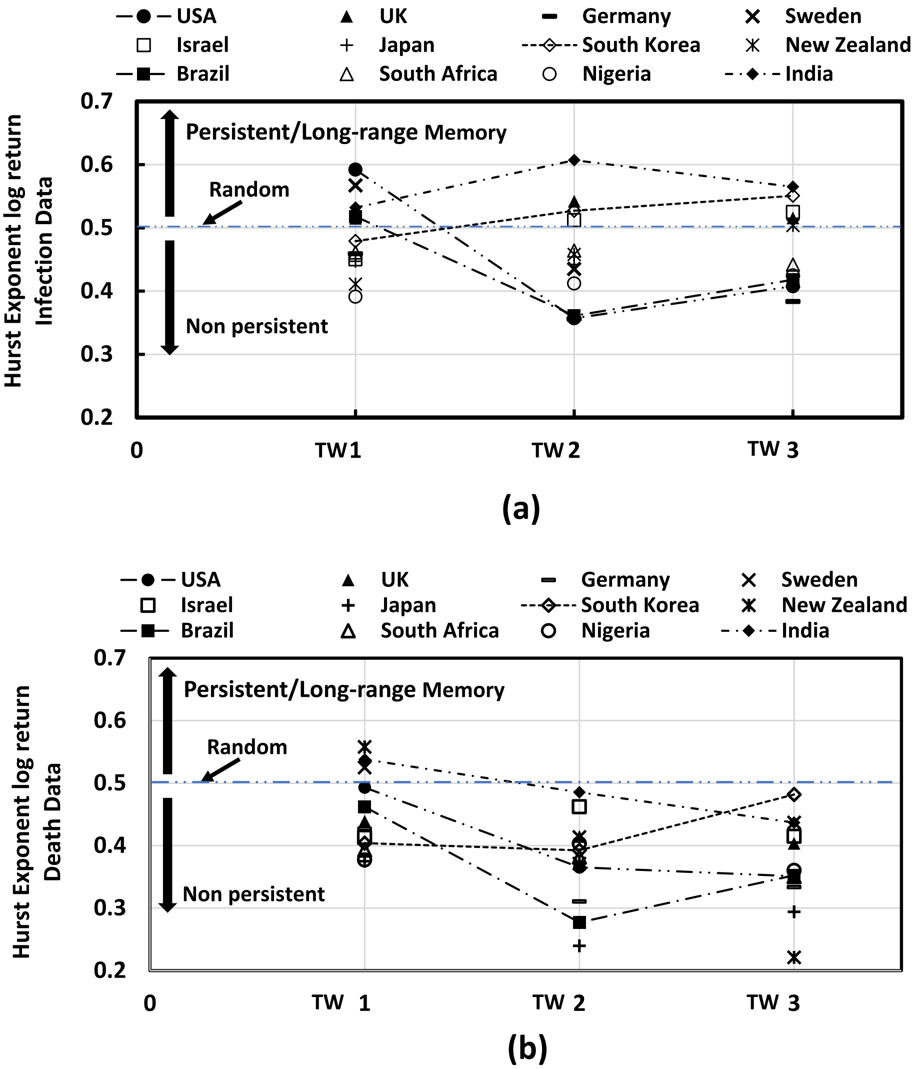

Fig. 7 Time evolution of the Hurst exponent results for all twelve countries' daily infection and death log return time series: (**a**) The top panel shows the time evolution from TW1-TW2-TW3 for the log return infection data; **(b)** the bottom panel shows the time evolution from TW1-TW2-TW3 for the log return death count data. A dashed horizontal line is drawn along H = 0.5 to indicate a white noise state. H < 0.5 represents data corresponding to an anti-correlated or mean-reversing state, while H > 0.5 shows persistence or long-range memory in the data. We have shown examples of the state transitions from long-range memory/persistence to random to mean-reversion using the dashed lines

and then about 3% decline from TW2 to TW3. This indicates that, in general, the percent changes in the infection cases show a tendency to move away from persistence/long-range memory and move in the direction towards mean-reversion and anti-correlated states. This can perhaps be attributed to the fact that as time progressed a higher population percentage has been able to receive vaccines and also acquire natural immunity. However, these actions also are tugging against an opposite reaction as time progressed forward, namely, the opening up of society and relaxing measures, such as social-distancing, wearing masks, etc. This force is competing with the fact that the SARS-CoV-2 is mutating from the alpha variant to the delta variant to the highly contagious omicron variant. Due to these competing forces, there is not a clear downward trending, i.e., a negative slope for the average Hurst exponent values across the twelve countries between the time periods considered here. The Hurst exponent results point towards local changes, as illustrated by transitions for the USA and Brazil. A universal behavior can be expected with a homogeneous overall situation, namely, global vaccines and/or herd immunity.

Now we discuss the results in Fig. 7(b) that correspond to the Hurst exponent time evolution of the death time series (log return data). The line for H = 0.5 indicates a baseline for white noise. We observe that a few countries in TW1 show Hurst exponent values larger than 0.5; specifically, note the results for Sweden, India, and New Zealand. Starting in TW2, all countries present Hurst exponent values below 0.5, with the exception for India and Israel showing the exponent value near 0.5. In TW3, the Hurst exponent for South Kore is near 0.5, while the remaining countries show Hurst exponent values below 0.5. The Hurst exponent value for New Zealand is close to zero – this trend indicates mean-reversion or anti-correlation in the data.

Several illustrative state transitions are indicated by dashed lines for the death log return data. For the USA, a 26% drop in the Hurst exponent is noted from TW1 to TW2 and then another 4% drop from TW2 to TW3. For Brazil, there is a 33% drop in the Hurst exponent from TW1 to TW2 and then a 27% increase from TW2 to TW3. For India, there is a 10% drop in the Hurst exponent from TW1 to TW2 and then another 10% drop from TW2 to TW3. Overall, the average Hurst exponent across the twelve countries drops 16% from TW1 to TW2 and then another 2% from TW2 to TW3.

The general trend indicates that as time evolved in the pandemic and vaccines became available, the death percent changes tend towards being anti-correlated and mean-reverting. This provides some quantitative evidence that the vaccines and natural immunity may have had a positive or favorable effect on the percent change in daily death counts. During TW2 and TW3, the vaccines were by and large available to the general public, while the only defense that the public, in general, had was from lockdowns and social distancing during TW1.

DFA Scaling Exponent Studies

To further establish the trends through self-affinity studies, we examine the DFA scaling exponent, α – the methodology is discussed in Sec. "Detrended Fluctuation Analysis (DFA) and Scaling Exponent". Figure 8 shows the Detrended Fluctuation Analysis scaling exponent results.

Figure 8(a) shows the estimated DFA α for the infection time series data (log return data). Here, we plot the evolution of this scaling exponent for all twelve countries across TW1 through TW3 time segments. We have indicated the line $\alpha = 0.5$ to indicate white noise, and we show several dashed lines to illustrate state transitions.

An analysis of the USA log return infection data, for example, shows an 81% decline in the DFA α from TW1 to TW2. The scaling exponent is far above 0.5 during TW1 for USA, while it falls to approximately 0.17—0.18 in TW2 and TW3 time segments. The DFA α for India, on the other hand, shows less than 0.5 in TW1, but it increases 80% in TW2 and continues to remain above 0.5 in TW3. For Israel, on the other hand, a 7% decline is noted from TW1 to TW2, and it remains at the same level (value of 0.52) for TW3. In general, the DFA scaling exponent for Israel remains around 0.5 across the three time periods.

Looking at the average scaling exponent across the twelve countries for the log return of the infection data, we note an approximately 29% drop in the scaling exponent from TW1 to TW2. Another 5% drop in the average DFA scaling exponent occurs going from TW2 to TW3.

For the death log return data, Fig. 8(b) shows a downward trend (on average) from TW1 to TW2. From TW2 to TW3, the death log return time series exhibits on average a flat trend.

A few specific state transition cases are highlighted – the dashed lines can be followed in these illustrations; all others can be similarly analyzed in Fig. 8(b). During TW1, the DFA scaling exponent for the USA is above 0.5. Then, during TW2 an approximately 93% drop in the DFA scaling exponent occurs, and it increases by about 25% between TW2 to TW3. For India, the DFA scaling exponent is above 0.5 in TW1 and increases by 6% in TW2. It then drops below 0.5 in TW3 with a decrease of 26% between TW2 and TW3. The average DFA scaling exponent value across the twelve countries decreases 49% from TW1 to TW2 and then another 4% decrease from TW2 to TW3.

Overall, the average DFA scaling exponent shows significantly larger decline for the death time series than for the infection time series between TW1 and TW3. The death time series have moved away from a long-range memory and approached towards an anti-correlation or mean-reversal state from TW1 through TW3 on average. This corroborates the results from Hurst exponent analysis discussed earlier.

Key Comparisons of Hurst Exponent and DFA scaling for Assigning Characteristics to COVID-19 Evolution

Overall, we observe the following key behaviors and trends that shed important light on questions related to state transitions in the COVID-19 data.

(i)The Hurst exponent analysis shows that during TW1 when vaccines were unavailable (except limited cases), several countries show long-range memory and persistence in both infection and death log return data. This includes India, US, UK, Sweden and Brazil for infection data, and Sweden, New Zealand, and India for death data.

During TW2 and TW3, about one-third of the countries show Hurst exponent values above or around 0.5 for log return infection, while all countries show Hurst exponent values less than 0.5 for log return death.

(ii)The DFA scaling exponent analysis shows the values for US, UK, Israel and Brazil to be above 0.5 for the

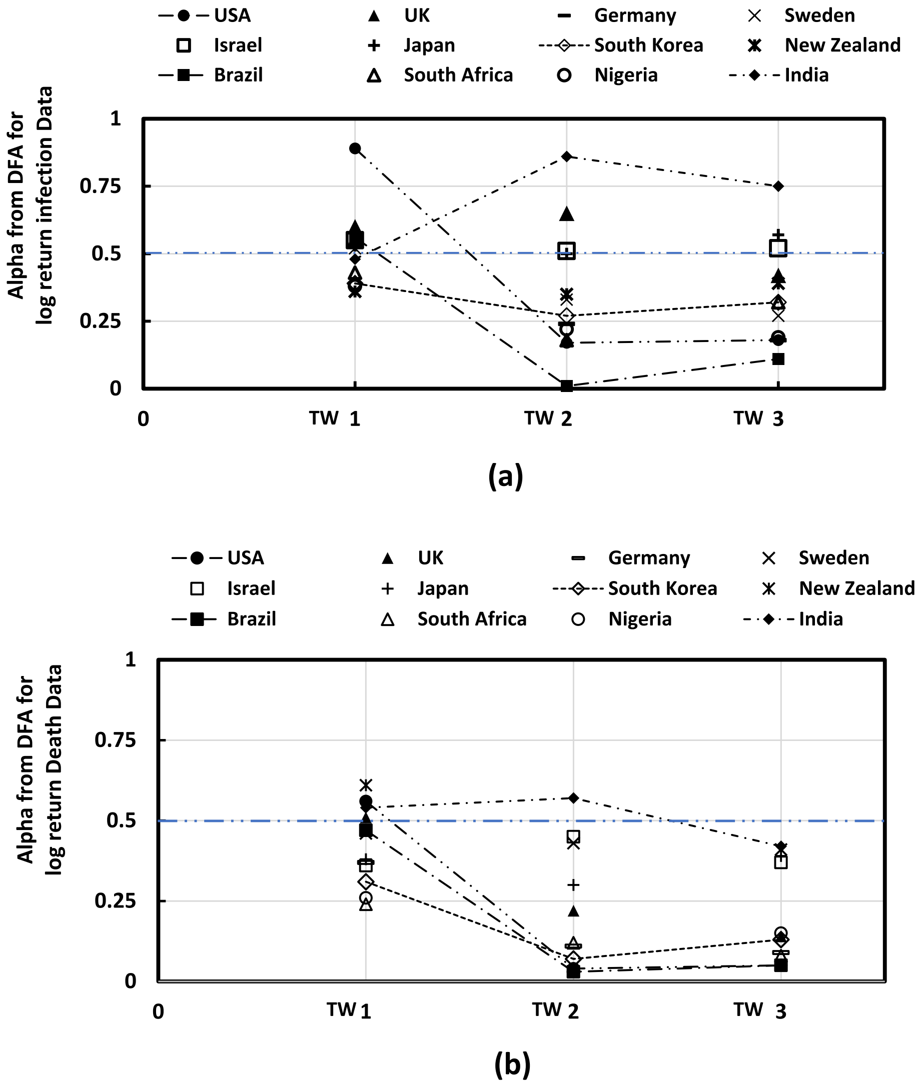

Fig. 8 Time evolution of the scaling exponent, α, calculated from the Detrended Fluctuation Analysis (DFA) for all twelve countries' daily infection and death log return time series: (**a**) The top panel shows the time evolution from TW1-TW2-TW3 for the log return infection data; **(b)** the bottom panel shows the time evolution from TW1-TW2-TW3 for the log return death data. A horizontal dashed line shows α = 0.5 to indicate a white noise state. We have shown examples of the state transitions from long-range memory/persistence to random to mean-reversion using the dashed lines

infection time series during TW1—indicating long-range memory and persistence. For the log return death time series in TW1, the countries, namely, USA, New Zealand, and India show DFA scaling exponents above 0.5.

All countries, except India, show DFA scaling exponents below 0.5 for death time series during TW2. For infection data, approximately one-sixth of the countries show DFA scaling exponents above 0.5 during TW2 and TW3 time period, indicating long-range memory. The rest are all below 0.5 which indicates trends towards non-persistence or anti-correlations.

(iii) From both the average Hurst exponent calculated from R/S analysis and the average DFA scaling exponent, we consistently observe a negative trend in the exponent values corresponding to the log return death data going from TW1 to TW3. This behavior is also present for the log return infection data, but the percent change is relatively less prominent. This trend correlates with several factors: (1) Several countries show a positive uptick, in the Hurst exponent and/or DFA scaling exponent, between TW2 and TW3 for the infection log return data. (2) A large global jump in the number of infection cases due to the omicron variant of the SARS-CoV-2 spread throughout the globe is reported. (3) The omicron variant likely affected the infection daily percent changes as the vaccines and natural immunity may not have fully mitigated against contracting the COVID-19 disease. However, the vaccines and the natural immunity may have shown resilience in preventing deaths from the COVID-19 disease. The observations from the self-affinity analyses that both the Hurst exponent and the DFA scaling exponent show a clear downward trend for the average values as the time window progresses forward can be used as quantitative metrics to characterize the underlying dynamical system.

Conclusions

This work builds a structured methodological approach using an analytics toolbox to study the temporal dynamics of the COVID-19 pandemic, specifically the daily infection and death counts. Through a systematic study of diverse datasets across a comprehensive time window, this article demonstrates the utility of an ensemble of multiple data analytics methodology. This study shows the design of a toolbox, in examining statistical and self-affinity characteristics of a dynamically evolving system, here, COVID-19 daily infections and deaths. With segmentation of the time into three characteristic windows from the beginning of the pandemic (January 22, 2020) till March 1, 2022, we have analyzed daily new infection and death time series across diverse geographical regions spanning twelve countries over six different continents.

We summarize the key findings below.

(i) All of the infection and death time series show a positive skewness in the data for the twelve countries across all time windows.

(ii) The vast majority, i.e., approximately 90% of the cases, for both infection and death time series show a leptokurtic distribution, with varying characteristics, especially from the standpoint of heavy tail distribution attributes.

(iii) From co-integration studies, we find no universal rule that the infection count data is co-integrated with the death count data for a lag of seven days. Approximately two-thirds of the countries in our data sample show co-integration between the non-stationary infection cases and its corresponding non-stationary death time series counts. Sensitivity studies with respect to lag days, from seven days to twenty-one days, point to questions on universality, with an indication towards local attributes instead of diffusion through the entire geosystem.

(iv) From the self-affinity studies across all twelve countries, we observe a trend of moving from long-memory to mean reversion state for the log return death data.

The infection time series shows a relatively weaker negative trend (with respect to death time series data) for the average Hurst and DFA exponent value across TW1 to TW3. Several upticks in the Hurst and DFA exponent values going from TW2 to TW3 can be contributing here. This can be possibly attributed to the fact that the omicron variant may have caused general spikes in most places around the globe in the daily infection number of cases.

This quantitative analysis builds a foundation for assessment of data trends. Tracking both Hurst exponent and DFA scaling exponent enables us to follow the system state. Note that low values approaching zero are desirable from an epidemiological viewpoint – this indicates anti-persistence in the time series and thus mean-reversion. Usage of these analytical tools in the context of new data for new situations can be continuously explored using the foundation laid down in this work.

While demonstrating the merit of an ensemble of multiple analytics for examining large dynamical datasets, several topics for future explorations are conceived. With the ongoing complex mutations of the SARS-CoV-2, such as various sub-omicron variants, many questions emerge on the competing dynamics of additional population segments having vaccines. Continuing the statistical analyses and the Hurst and DFA scaling exponent estimation enables building valuable knowledge, especially to determine the merits and limitations of the individual versus multiple analytics with newer data.

One of the key strengths of this approach is that we have analyzed the temporal dynamics of the COVID-19 pandemic using three independent analytical methodologies: (a) statistical characteristics through examining higher moments of the distribution; (b) stationarity tests of all time series datasets and co-integration between non-stationary time series datasets, (c) and finally, self-affinity properties using Hurst exponent and DFA scaling exponent estimates. While each of these methodologies may individually have certain assumptions and limitations (see the respective sections on methodologies in Sec. "Methodologies for Examining Statistical Characteristics of Time Series Data" and "Methodologies for Examining Self-Affinity Characteristics of Time Series Data"), their usage in an analytics toolbox provides an aggregation of lessons learned from all and reduces individual bias in any one particular method. Although the merits of synergistic aggregation are not quantitatively assessed (beyond the scope of this work), the links in the analytics chain comprising higher moments analysis-stationarity & co-integration studies-self-affinity studies and value added by the elements can be qualitatively recognized. A formal mathematical construct, based on for example, information theory centric analyses, can shed light on quantitative assessments of relative merits of diverse analytics and their aggregations, including model-based approaches as well. Such an endeavor in the future can add important values and guide users to dynamically perform scenario-based analysis of alternatives and assess relative merits of aggregation of results from different analytical methods.

Since this is a data driven approach and it does not incorporate any specific model or parametric assumptions, any corruptions or imperfections in the source data can yield questionable conclusions. Hence, the credibility of the original data source [35] is very critical. As noted earlier, several independent studies [38–40] have shown that the JHU repository is a reasonable data source to use. With a broader perspective it can be noted that a systematic comparison of different data repository for all data has merits. An important point related to the use of higher moments, namely, skewness and kurtosis, is that although these statistical metrics of data distributions give insights into general macroscopic trends for the COVID-19 daily infection and death data, its ability to determine clear distinctions in the characteristic signatures of the data between the countries is limited. Limitations of low predictive power of the skewness and kurtosis are reported in a recent work with respect to forecasting contexts [36].

As outlook for the future, the analytics framework reported in this work can be tested with datasets pertaining to new dynamical problems. In addition, other quantitative approaches, for example, spectral based methods (such as Wavelet transform [74] and/or Fast Fourier Transform (FFT)) and entropy-based methods (such as, Renyi entropy and Mutual information function (MIF) [75]) may also be added to the analytics toolbox to yield additional information and enhance capabilities. Further study may be warranted in this area.

Additionally, quantitative evidence of regime changes can be explored using Markov regime switching models [76]. Genetic based algorithms and minimum description length (MDL) principles can be used for non-stationary time series [77] to identify the structural breaks, and it can quantitatively indicate when these breaks lead to potential regime changes. Finally, linking this work to epidemiological model-based approaches, such as SEIR/TSIR/others, have merits to build a grand ensemble methodology and answer various critical questions, especially to build policy related strategies.

Abbreviations

SARS-CoV-2	Severe Acute Respiratory Syndrome Coronavirus 2
COVID	COronaVIrus Disease
SEIR	Susceptible-Exposed-Infectious-Recovered model
TSIR	Time dependent Susceptible-Infection-Recovered model
TW1	Time Window 1 ranges from January 22, 2020 through March 1, 2021
TW2	Time Window 2 ranges from March 2, 2021 through October 10, 2021
TW3	Time Window 3 ranges from March 2, 2021 through March 1, 2022
FW	Full Window ranges from January 22, 2020 – March 1, 2022
JHU COVID-19 Data Repository	Johns Hopkins University's COVID-19 Data Repository by the Center for Systems Science & Engineering
PP test	Phillips Perron test for unit root
ADF test	Augmented Dickey Fuller test for unit root
KPSS test	Kwiatkowski-Phillips-Schmidt-Shin test for unit root
MAIC	Modified Akaike Information Criterion
PO test	Phillips-Ouliaris test for co-integration
R/S analysis	Rescaled range analysis for Hurst exponent
DFA	Detrended Fluctuation Analysis
I(1)	Integrated of order 1 for time series
I(0)	Integrated of order 0 for time series

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12874-024-02423-y.

Supplementary Material 1.

Supplementary Material 2.

Acknowledgements

The author acknowledges valuable comments and suggestions of the reviewers. The author did not receive any financial support, in terms of research funds or otherwise, from any organization for the submitted work. The author acknowledges conducting the computing analysis using the open-source R software platform. In addition to the R packages mentioned in the text, several additional R packages were used to perform calculations: (a) *timeDate* package (for computing skewness and kurtosis); (b) *nonlinearTseries* package (for computing DFA scaling exponent); *pracma* package (for computing Hurst exponent); *tseries* package (for calculating the PO test).

Authors' contributions

SG wrote the paper.

Funding

None.

Availability of data and materials

All of the data is accessed from the JHU COVID-19 Repository as shown in reference [37]. Specific data used in this study can be downloaded using the covid19.analytics R package as shown in reference [41]. In addition, the datasets used and/or analyzed during the current study can be available from the corresponding author on request.

Declarations

Ethics approval and consent to participate

N/A.

Consent for publication

N/A.

Competing interests

The authors declare no competing interests.

Author details

[1]Systems Engineering & Operations Research, George Mason University, Fairfax, VA 22030, USA.

Received: 9 November 2023 Accepted: 26 November 2024
Published online: 03 January 2025

References

1. Staff, A. J. M. C. "A timeline of COVID-19 vaccine developments in 2021." American Journal of Managed Care. https://www.ajmc.com/view/a-timeline-of-covid-19-vaccine-devel opments-in-2021. Published June 3 (2021).
2. Worldometer, Dadax. COVID-19 coronavirus pandemic. World Health Organization; 2020. www.worldometers.info.
3. Faddy MJ, Pettitt AN. Comparisons of statistical distributions for cluster sizes in a developing pandemic. BMC Med Res Methodol. 2022;22(1):1–7.
4. Aslam M. Design of a new Z-test for the uncertainty of COVID-19 events under Neutrosophic statistics. BMC Med Res Methodol. 2022;22(1):1–6.
5. Neil-Sztramko SE, Belita E, Traynor RL, Clark E, Hagerman L, Dobbins M. Methods to support evidence-informed decision-making in the midst of COVID-19: creation and evolution of a rapid review service from the National Collaborating Centre for Methods and Tools. BMC Med Res Methodol. 2021;21(1):1–10.
6. Xin Y, Nevill CR, Nevill J, Gray E, Cooper NJ, Bradbury N, Sutton AJ. Feasibility study for interactive reporting of network meta-analysis: experiences from the development of the MetaInsight COVID-19 app for stakeholder exploration, re-analysis and sensitivity analysis from living systematic reviews. BMC Med Res Methodol. 2022;22(1):1–8.
7. Srinivasa RG, Aslam M. Inspection plan for COVID-19 patients for Weibull distribution using repetitive sampling under indeterminacy. BMC Medical Research Methodology. 2021;21(1):1–15.
8. Zhou, Qi, Qinyuan Li, Janne Estill, Qi Wang, Zijun Wang, Qianling Shi, Jingyi Zhang et al. "Methodology and experiences of rapid advice guideline development for children with COVID-19: responding to the COVID-19 outbreak quickly and efficiently." BMC Medical Research Methodology. 2022;22(1)1-24.
9. Ziemann S, Paetzolt I, Grüßer L, Coburn M, Rossaint R, Kowark A. Poor reporting quality of observational clinical studies comparing treatments of COVID-19–a retrospective cross-sectional study. BMC Med Res Methodol. 2022;22(1):1–11.
10. Haddad C, Sacre H, Zeenny RM, Hajj A, Akel M, Iskandar K, Salameh P. Should samples be weighted to decrease selection bias in online surveys during the COVID-19 pandemic? Data from seven datasets. BMC Med Res Methodol. 2022;22(1):1–11.
11. Namazi H, Selamat A, Krejcar O. Complexity-based analysis of the alterations in the structure of coronaviruses. Fractals. 2021;29(02):2150123.
12. Meraz M, Vernon-Carter EJ, Rodriguez E, Alvarez-Ramirez J. A fractal scaling analysis of the SARS-CoV-2 genome sequence. Biomed Signal Process Control. 2022;73: 103433.
13. Dehipawala S, Cheung E, Tremberger G, Cheung T. Entropy and Fractal Dimension Study of the TDP-43 Protein Low Complexity Domain Sequence in ALS Disease Severity and SARS-CoV-2 Gene Sequences in Virulence Variability. Entropy. 2021;23(8):1038.
14. Perez JC Lounnas V, Tan M, Azalbert X, Perronne C. May the SARS-CoV-2 OMICRON variant signal the end of the pandemic–a Fibonacci fractal analysis. Arch Microbiol Immunol. 2022;6(1):1–6. https://fortunepublish.com/articles/may-the-sarscov2-omicron-variant-signal-the-end-of-the-pandemic-ndash-afibonacci-fractal-analysis.html.
15. Chakraborty T, Ghosh I. Real-time forecasts and risk assessment of novel coronavirus (COVID-19) cases: A data-driven analysis. Chaos, Solitons Fractals. 2020;135: 109850.
16. Kırbaş, İsmail, Adnan Sözen, Azim Doğuş Tuncer, and Fikret Şinasi Kazancıoğlu. "Comparative analysis and forecasting of COVID-19 cases in various European countries with ARIMA, NARNN and LSTM approaches." Chaos, Solitons & Fractals. 2020;138:110015.
17. Ribeiro, Matheus Henrique Dal Molin, Ramon Gomes da Silva, Viviana Cocco Mariani, and Leandro dos Santos Coelho. "Short-term forecasting COVID-19 cumulative confirmed cases: Perspectives for Brazil." Chaos, Solitons & Fractals. 2020;135: 109853.
18. Hu Z, Ge Q, Li S, Xiongd M. Artificial Intelligence Forecasting of COVID-19 in China. Int J. 2020;6(1):71–94.
19. Sujath, R., Jyotir Moy Chatterjee, and Aboul Ella Hassanien. "A machine learning forecasting model for COVID-19 pandemic in India." Stochastic Environmental Research and Risk Assessment 2020;34(7):959–972.
20. Refisch L, Lorenz F, Riedlinger T, Taubenböck H, Fischer M, Grabenhenrich L, Wolkewitz M, Binder H, Kreutz C. Data-driven prediction of COVID-19 cases in Germany for decision making. BMC Med Res Methodol. 2022;22(1):1–13.
21. Lucas B, Vahedi B, Karimzadeh M. A spatiotemporal machine learning approach to forecasting COVID-19 incidence at the county level in the USA. Int J Sci Analytics. 2023;15(3):247–66. https://link.springer.com/content/pdf/10.1007/s41060-021-00295-9.pdf.
22. Ray, Debashree, Maxwell Salvatore, Rupam Bhattacharyya, Lili Wang, Jiacong Du, Shariq Mohammed, Soumik Purkayastha et al. "Predictions, role of interventions and effects of a historic national lockdown in India's response to the COVID-19 pandemic: data science call to arms." Harvard Data Science Review. 2020;Suppl 1):2020.
23. Wu JT, Leung K, Leung GM. Nowcasting and forecasting the potential domestic and international spread of the 2019-nCoV outbreak originating in Wuhan, China: a modelling study. The Lancet. 2020;395(10225):689–97.
24. Fanelli D, Piazza F. Analysis and forecast of COVID-19 spreading in China, Italy and France. Chaos, Solitons Fractals. 2020;134: 109761.
25. Ghosh I, Chakraborty T. An integrated deterministic–stochastic approach for forecasting the long-term trajectories of COVID-19. International Journal of Modeling, Simulation, and Scientific Computing. 2021;12(03):2141001.
26. Hachtel GD, Stack JD, Hachtel JA. Forecasting and modeling of the COVID-19 pandemic in the USA with a timed intervention model. Sci Rep. 2022;12(1):1–14.
27. Ambrosio, Benjamin, and M. A. Aziz-Alaoui. "On a coupled time-dependent SIR models fitting with New York and New-Jersey states COVID-19 data." Biology. 2020;9(6):135.
28. Chen Y-C, Ping-En Lu, Chang C-S, Liu T-H. A time-dependent SIR model for COVID-19 with undetectable infected persons. IEEE Transactions on Network Science and Engineering. 2020;7(4):3279–94.
29. Chakraborty, Tanujit, Indrajit Ghosh, Tirna Mahajan, and Tejasvi Arora. "Nowcasting of COVID-19 confirmed cases: Foundations, trends, and challenges." In Modeling, Control and Drug Development for COVID-19 Outbreak Prevention, pp. 1023–1064. Springer, Cham, 2022.
30. Hamilton MA, Hamilton D, Soneye O, Ayeyemi O, Jaradat R. An analysis of the impact of policies and political affiliation on racial disparities in COVID-19 infections and deaths in the USA. International Journal of Data Science and Analytics. 2022;13(1):63–76.
31. Sivakumar B, Deepthi B. Complexity of COVID-19 Dynamics. Entropy. 2021;24(1):50.
32. Mariani, M. C., W. Kubin, P. K. Asante, O. K. Tweneboah, and M. P. Beccar-Varela. "Multifractal Analysis of Daily US COVID-19 Cases." In Proceedings of the 10th Annual AHSE, STEM/STEAM and Education Conference, Honolulu, HI, USA, pp. 9–11. 2021.
33. Du B, Zhao Z, Zhao J, Le Yu, Sun L, Lv W. Modelling the epidemic dynamics of COVID-19 with consideration of human mobility. International Journal of Data Science and Analytics. 2021;12(4):369–82.
34. Cao, Longbing, and Qing Liu. "COVID-19 Modeling: A Review." arXiv preprint arXiv:2104.12556 (2021).
35. Kamalov F, Cherukuri KA, Thabtah F. Machine learning applications to Covid-19: a state-of-the-art survey. In: 2022 Advances in Science and Engineering Technology International Conferences (ASET). IEEE; 2022. p. 1–6. https://ieeexplore.ieee.org/abstract/document/9734959.
36. Waku, Jules, Kayode Oshinubi, Umar Muhammad Adam, and Jacques Demongeot. "Forecasting the endemic/epidemic transition in COVID-19 in some countries: Influence of the vaccination." Diseases. 2023;11(4):135.
37. Miller M. 2019 Novel Coronavirus COVID-19 (2019-nCoV) Data Repository: Johns Hopkins University Center for Systems Science and Engineering. Bulletin-Association of Canadian Map Libraries and Archives (ACMLA). 2020;164:47–51.
38. Cramer, Estee Y., Yuxin Huang, Yijin Wang, Evan L. Ray, Matthew Cornell, Johannes Bracher, Andrea Brennen, A.J.C. Rivadeneira, A. Gerding, K. House, and D. Jayawardena, "The United States COVID-19 forecast hub dataset." Scientific Data 9, no. 1 (2022): 462.
39. Dong, Ensheng, Jeremy Ratcliff, Tamara D. Goyea, Aaron Katz, Ryan Lau, Timothy K. Ng, Beatrice Garcia, E. Bolt, S. Prata, D. Zhang, and R.C. Murray, "The Johns Hopkins University Center for Systems Science and

Engineering COVID-19 Dashboard: data collection process, challenges faced, and lessons learned." The Lancet Infectious Diseases 22, no. 12 (2022): e370-e376.

40. Miller, April R., Samin Charepoo, Erik Yan, Ryan W. Frost, Zachary J. Sturgeon, Grace Gibbon, Patrick N. Balius, C.S. Thomas, M.A. Schmitt, D.A. Sass, and J.B. Walters, "Reliability of COVID-19 data: An evaluation and reflection." PLOS One 17, no. 11 (2022): e0251470.
41. Ponce, Marcelo, and Amit Sandhel. "covid19. analytics: An R Package to Obtain, Analyze and Visualize Data from the 2019 Coronavirus Disease Pandemic." Journal of Open Source Software 6, no. 60 (2021): 2995.
42. Brzezińska, Agnieszka Nowak, and Czesław Horyń. "Outliers in COVID-19 data based on Rule representation-the analysis of LOF algorithm." Procedia Computer Science 192 (2021): 3010–3019.
43. Boado-Penas MD, Eisenberg J. Pandemics: Insurance and Social Protection. Springer Nature; 2022. https://library.oapen.org/bitstream/handle/20.500.12657/51459/1/9783030783341.pdf.
44. Serfling R, Wang S. General foundations for studying masking and swamping robustness of outlier identifiers. Statistical Methodology. 2014;20:79–90.
45. Wang S, Serfling R. On masking and swamping robustness of leading nonparametric outlier identifiers for multivariate data. J Multivar Anal. 2018;166:32–49.
46. Zivot E, Wang J. Modeling financial time series with S-PLUS, vol. 2. New York: Springer; 2006.
47. Kim H-Y. Statistical notes for clinical researchers: assessing normal distribution (2) using skewness and kurtosis. Restorative Dentistry & Endodontics. 2013;38(1):52–4.
48. Joanes DN, Gill CA. Comparing measures of sample skewness and kurtosis. Journal of the Royal Statistical Society: Series D (The Statistician). 1998;47(1):183–9.
49. Bono R, Arnau J, Alarcón R, Blanca MJ. Bias, precision, and accuracy of skewness and kurtosis estimators for frequently used continuous distributions. Symmetry. 2019;12(1):19.
50. Wyszomirski T. Detecting and displaying size bimodality: kurtosis, skewness and bimodalizable distributions. J Theor Biol. 1992;158(1):109–28.
51. Park K Il, Park M. Fundamentals of probability and stochastic processes with applications to communications. Cham: Springer International Publishing; 2018. https://link.springer.com/book/10.1007/978-3-319-68075-0.
52. Florescu I. Probability and stochastic processes. John Wiley & Sons; 2014.
53. Fedorová D. Selection of unit root test on the basis of length of the time series and value of ar (1) parameter. Statistika. 2016;96(3):3.
54. McLeod AI, Yu H, Mahdi E. Time series analysis with R. In: Handbook of statistics, vol. 30. Elsevier; 2012. p. 661–712. https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=c47618aa589815bc77b904d8510e852b59957fec#page=680.
55. Ng S, Perron P. Lag length selection and the construction of unit root tests with good size and power. Econometrica. 2001;69(6):1519–54.
56. Lupi C. Unit root CADF testing with R. J Stat Softw. 2010;32:1–19.
57. Granger, Clive WJ. "Some properties of time series data and their use in econometric model specification." Journal of Econometrics. 1981;16(!):121–130.
58. Murray MP. A drunk and her dog: an illustration of cointegration and error correction. Am Stat. 1994;48(1):37–9.
59. Jarner, Søren F., and Snorre Jallbjørn. "Pitfalls and merits of cointegration-based mortality models." Insurance: Mathematics and Economics. 2002;90: 80–93.
60. Phillips, Peter CB, and Sam Ouliaris. "Asymptotic properties of residual based tests for cointegration." Econometrica: Journal of the Econometric Society (1990): 165–193.
61. Chruściel Piotr T, Szybka SJ. On the lag between deaths and infections in the first phase of the Covid-19 pandemic. medRxiv. 2021:2021–01. https://www.medrxiv.org/content/medrxiv/early/2021/01/04/2021.01.01.21249115.full.pdf.
62. Jin, Raymond. "The lag between daily reported COVID-19 cases and deaths and its relationship to age." Journal of Public Health Research 10, no. 3 (2021).
63. Hurst HE. Long-term storage capacity of reservoirs. Trans Am Soc Civ Eng. 1951;116(1):770–99.
64. Barunik J, Kristoufek L. On Hurst exponent estimation under heavy-tailed distributions. Physica A. 2010;389(18):3844–55.
65. Peng, C-K., Sergey V. Buldyrev, Shlomo Havlin, Michael Simons, H. Eugene Stanley, and Ary L. Goldberger. "Mosaic organization of DNA nucleotides." Physical Review E 49, no. 2 (1994): 1685.
66. Setty, Venkat Anurag, and A. Surjalal Sharma. "Characterizing detrended fluctuation analysis of multifractional Brownian motion." Physica A: Statistical Mechanics and its Applications 419 (2015): 698–706.
67. Peng, C-K., Shlomo Havlin, H. Eugene Stanley, and Ary L. Goldberger. "Quantification of scaling exponents and crossover phenomena in nonstationary heartbeat time series." Chaos: an interdisciplinary journal of nonlinear science. 1995;5(1):82–87.
68. Hardstone R, Poil S-S, Schiavone G, Jansen R, Nikulin VV, Mansvelder HD, Linkenkaer-Hansen K. Detrended fluctuation analysis: a scale-free view on neuronal oscillations. Front Physiol. 2012;3:450.
69. Carpena P, Gómez-Extremera M, Bernaola-Galván PA. On the Validity of Detrended Fluctuation Analysis at Short Scales. Entropy. 2022;24(1):61.
70. Chen H, Shi L, Zhang Y, Wang X, Sun G. A cross-country core strategy comparison in China, Japan, Singapore and South Korea during the early COVID-19 pandemic. Glob Health. 2021;17(1):1–10.
71. Kung S, Doppen M, Black M, Hills T, Kearns N. Reduced mortality in New Zealand during the COVID-19 pandemic. The Lancet. 2021;397(10268):25.
72. Thangaraj, Jeromie, Wesley Vivian, Pragya Yadav, CP Girish Kumar, Anita Shete, Dimpal A. Nyayanit, D. Sudha Rani, Abhinendra Kumar et al. "Predominance of delta variant among the COVID-19 vaccinated and unvaccinated individuals, India, May 2021." Journal of Infection 84, no. 1 (2022): 94–118.
73. Yang W, Shaman J. COVID-19 pandemic dynamics in India and impact of the SARS-CoV-2 Delta (B.1.617. 2) variant. PREPRINT-medRxiv. 2021. https://pesquisa.bvsalud.org/portal/resource/midias/ppmedrxiv-21259268.
74. Omata K, Shimazaki A. Wavelet analysis of COVID-19 pandemic. Journal of Advanced Simulation in Science and Engineering. 2023;10(2):214–20.
75. Rana, Subas, Nasid Habib Barna, and John A. Miller. "Exploring the Predictive Power of Correlation and Mutual Information in Attention Temporal Graph Convolutional Network for COVID-19 Forecasting." In International Conference on Big Data, pp. 18–33. Cham: Springer Nature Switzerland, 2023.
76. Kim C-J, Piger J, Startz R. Estimation of Markov regime-switching regression models with endogenous switching. Journal of Econometrics. 2008;143(2):263–73.
77. Diks, Cees, and Valentyn Panchenko. "A new statistic and practical guidelines for nonparametric Granger causality testing." Journal of Economic Dynamics and Control. 2006;30(9-10):1647–1669.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.