medRxiv preprint doi: https://doi.org/10.1101/2021.06.09.21258612; this version posted June 12, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

Second wave of COVID-19 in India could be predicted with genomic surveillance of SARS-CoV-2 variants coupled with epidemiological data: A tool for future

*Ashutosh Kumar[1], Prakhar Dwivedi[1], Gopichand Kumar[1], Ravi K. Narayan[1], RakeshK. Jha[1], Rakesh Parashar[2], Chetan Sahni[3], Sada N. Pandey[4]

1- Department of Anatomy, All India Institute of Medical Sciences (AIIMS), Patna, India
2- India Health Lead at Oxford Policy Management Limited, Oxford, UK
3- Department of Zoology, Banaras Hindu University (BHU), Varanasi, India
4- Department of Anatomy, Institute of Medical Sciences, Banaras Hindu University (BHU), Varanasi, India

***Corresponding author**

Department of Anatomy,

All India Institute of Medical Sciences (AIIMS),

Patna, India, Postal code-801507

Email: drashutoshkumar at aiimspatna.org

Abstract

India has witnessed a devastating second wave of COVID-19, which peaked during the last week of April and the second week of May, 2021. We aimed to understand whether the arrival of second wave was predictable and whether it was driven by the existing SARS-CoV-2 strains or any of the emerging variants. We analyzed the monthly distribution of the genomic sequence data for SARS-CoV-2from India and correlated that with the epidemiological data for new cases and deaths, for the corresponding period of the second wave. Our analysis shows that the first indications of arrival of the second wave were observable by January, 2021, and by March, 2021 it was clearly predictable. B.1.617 lineage variants drove the wave, particularly B.1.617.2 (a.k.a. delta variant). We propose that genomic surveillance of the SARS-CoV-2 variants augmented with epidemiological data can be a promising tool for predicting future COVID-19 waves.

Main text:

India witnessed a devastating second wave of COVID-19 that started towards the end of February 2021 and peaked towards the end of April and during the first half of May 2021. An unwarned arrival of the second COVID-19 wave and the exponential surge in the infections crumpled the country's epidemic response system and health infrastructure. This resulted in

NOTE: This preprint reports new research that has not been certified by peer review and should not be used to guide clinical practice.

medRxiv preprint doi: https://doi.org/10.1101/2021.06.09.21258612; this version posted June 12, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

massive suffering and loss of lives, which could have been significantly minimized if timely forecasts were available. A lot of policy discourse has, since then, revolved around whether this wave was predictable and preventable. However, there are no definitive answers backed by the evidence available to this question. This article seeks to ascertain whether a specific variant of the SARS-CoV-2 drove the second wave and whether it was possible to predict this wave.

The second wave in India was arguably triggered by an emerging lineage of SARS-CoV-2 variants B.1.617, particularly its sub-lineage B.1.617.2 (a.k.a. delta variant) (1). B.1.617 lineage was recently recognized as a global variant of concern (VOC) by World Health Organization. B.1.617 variant was first reported in India in October 2020 (9), and the strain evolved to three more sub-lineages B.1.617.1-3. As per the most recent update from WHO, currently B.1.617.1 (a.k.a. kappa variant) is a variant of interest (VOI) and B.1.617.2 gained status of VOC replacing B.1.617 (2). B.1.617 contained mutations in the key spike protein regions involved in interactions with the host and induction of neutralizing antibodies (S: L452R, E484Q, D614G, del681, del1072) (3). The sub-lineages contained lineage defining spike mutations (L452R, D614G) as well as developed new mutations: B.1.617.1 (S: T95I, G142D, E154K, L452R, E484Q, D614G, P681R, Q1071H), B.1.617.2 (S: T19R, G142D, 156del, 157del, R158G, L452R, T478K, D614G, P681R, D950N), and B.1.617.3 (S: T19R, L452R, E484Q, D614G, P681R) (4).

Accumulating evidence suggests that B.1.617 lineage variants are more transmissible (5-10) and perhaps more lethal (8) than B.1.1.7 (a.k.a. alpha variant), which had been a dominant strain in Indian population before the arrival of second wave (1). The studies have shown a significant reduction in the neutralization against B.1.617 lineage variants by antibodies received from natural infections and manycurrently used COVID-19 vaccines and multiple monoclonal antibodies (5-8). Higher transmissibility and immunoescape is being reported for B.1.617.2, which is currently the fastest growing SARS-CoV-2 strain in the Indian population (1, 9-10). Until December 2020, B.1.1.7 was a dominant strain in the Indian population, although the first cases of B.1.617 lineage had started appearing by then. By April 2021 India entered a full blown second COVID-19 wave (1, 2).

To ascertain whether the second wave was predictable and whether it was driven by any existing SARS-CoV-2 lineage or by a single or group of emerging variants, we analyzed two types of data set retrieved from secondary sources. First, we analyzed monthly distribution of the genomic sequence data for SARS-CoV-2 from India available in EpiCoV^{TM} database of Global Initiative on Sharing All Influenza Data (GISAID). We also analyzed the epidemiological data for new cases and deaths from COVID-19 (downloaded from Worldometer: https://www.worldometers.info/coronavirus/coronavirus/country/india/) for the period from 1stDecember 2020 to 30thApril 2021. We analyzed a total of 10115 SARS-CoV-2 genomic sequences, which were uploaded in GISAID database (last reporting date 7.06.2021), showing dates of collection for the samples falling in the period of study. We then plotted the genomic sequence data and epidemiological data together and drew graphical correlations. Data for the individual variants were categorized, and graphs were plotted to visualize the trends.

medRxiv preprint doi: https://doi.org/10.1101/2021.06.09.21258612; this version posted June 12, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

Observing the data trends gives a glimpse of the formation of the second COVID-19 wave in India with clear indications that SARS-CoV-2 strains may have driven it (Figs. 1-2). By December 2020, 8 SARS-CoV-2 Pango lineages and their multiple sub-lineages were circulating in Indian population including four VOCs (B.1.1.7, B.1.351, P1, and B.1.617.2) and three variants of interest (VOIs) (B.1.617.1, B.1.127/B.1.429 and B.1.525). However, B.1.1.7 was most dominant variant being found in majority of the genomic sequences uploaded in GISAID database from India. B.1.617 lineage variants collectively (B.1.617+) showed an increasingly upward trend since day of origin and surpassed other VOCs including B.1.1.7 by January, 2021, and kept rising up to end of April. In contrast, B.1.1.7 showed a downward trend by March, 2021. This indicated B.1.617 lineage variants were the dominant variant hereafter and by April 2021 were detected in 67% of SARS-CoV-2 sequences uploaded on GISAID. The rise of B.1.617 lineage showed a closely matched pattern with epidemiological data (new cases and deaths) in the time period of December, 2020 to April, 2021(Fig. 1), which was not observed with any other variant, indicating that this lineage was the prime driver of the rising second wave, which peaked between last week of April and second week of May, 2021 as being reflected in the trend of daily new cases and deaths (Fig. S1).

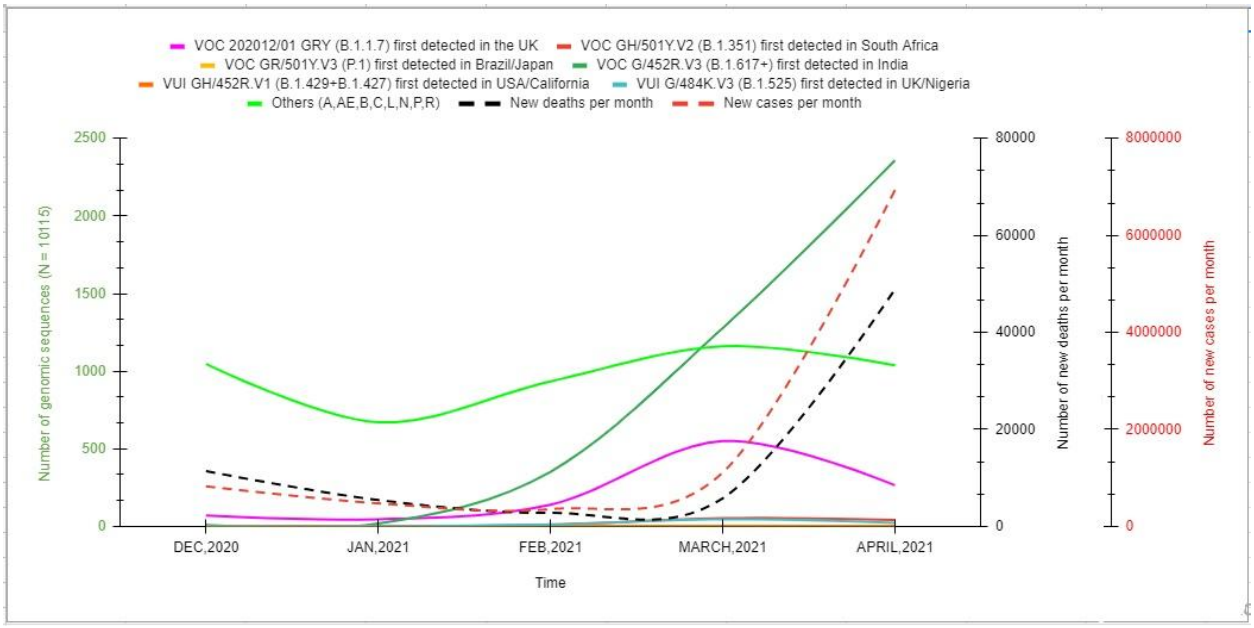
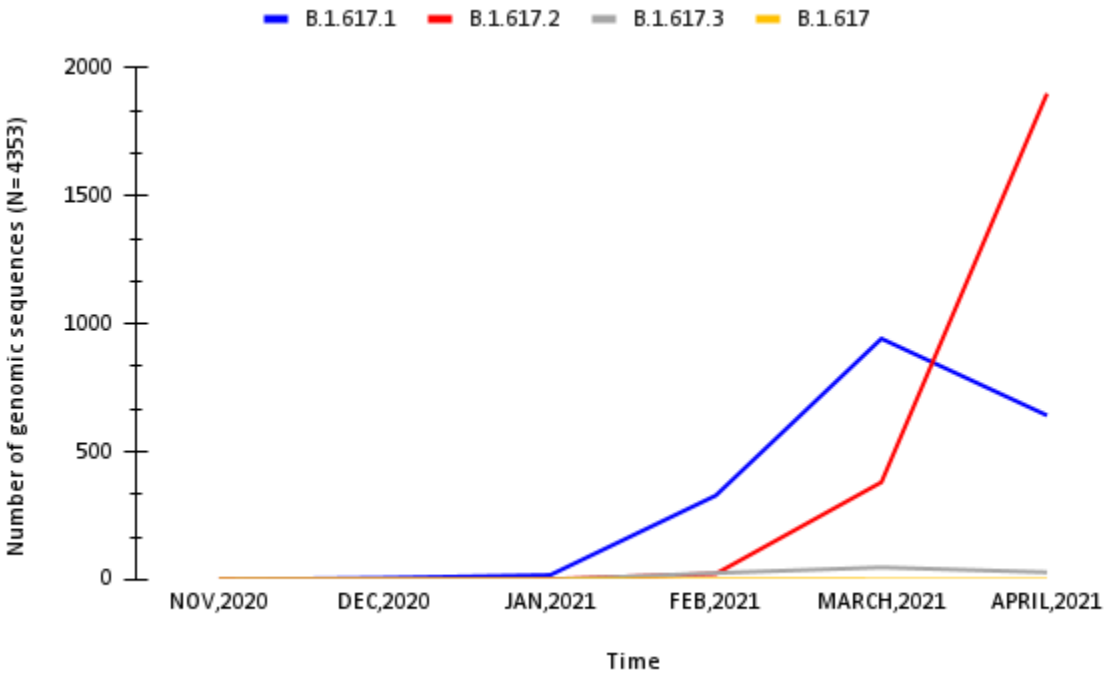


Figure 1 Monthly distribution of SARS-CoV-2 variants in genomic sequence data from India and their correlation with daily new COVID-19 cases and deaths for the period of 1st December, 2020 to 30th April, 2021. Data source: SARS-CoV-2 genomic sequence—GISAID database: https://www.gisaid.org/; Epidemiological data—Worldometer: https://www.worldometers.info/coronavirus/coronavirus/country/india)

medRxiv preprint doi: https://doi.org/10.1101/2021.06.09.21258612; this version posted June 12, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

To know whether the rise in B.1.617 lineage variants were localized to certain geographical regions which may have influenced the collective data trends, we further analyzed the genomic sequence data from the states and union territories individually. A similar increase in the detection of B.1.617 lineage variants was observable in most of the states and union territories of India (for which genomic data was available), with fewer exceptions (Fig. S2). However, totally different patterns were visible in Kerala and Punjab, where B.1.1.7 was still a dominant variant (Fig. S2), and remotely placed states/union territories, such as, Ladakh and Andaman-Nicobar group of islands (Fig. S2), where a previously dominant variant (B.1 lineage) was still prevalent by December, and very few genomic sequences have been uploaded from these places in the period of study.

Further, we wanted to know which particular B.1.617 lineage variants were dominating in the studied time period. Interestingly, an intra-lineage competition was distinctly visible between the sub-lineages of B.1.617 (Fig. 2). The single sample of B.1.617 was reported on 25. 02. 2021 (date of collection) (EPI-ISL_1544002) and thereafter it has not been detected. First case of B.1.617.2 was detected as early as 21st November, 2020 (EPI_ISL_2373501) followed by B.1.617.1 on 1st December, 2020 (EPI_ISL_1372093), and B.1.617.3 on 14th December, 2021 (EPI_ISL_2099648). B.1.617.2 followed by B.1.617.1 was detected in much greater number of samples than B.1.617.3 since first case. B.1.617.1 took a downturn by March, 2021, but B.1.617.2 showed a steep rise and by April, 2021 B.1.617.2 was singly detected in 50% of the SARS-CoV-2 samples from India uploaded in GISAID database.



medRxiv preprint doi: https://doi.org/10.1101/2021.06.09.21258612; this version posted June 12, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

Figure 2 Origin and spread of B.1.617 lineage SARS-CoV-2 variants in Indian population. Data source: SARS-CoV-2 genomic sequence—GISAID database: https://www.gisaid.org.(Data analyzed for the period of 1st November, 2020 to 30th April, 2021. First genomic sequence of **B.1.617 lineage** was uploaded on GISAID database on 21stNovember, 2020.)

We observe that that the first indications for arrival of the second wave were clearly observable by January, 2021 when rise of B.1.617 lineage surpassed all VOCs and VOIs (Fig.1). Moreover, by March, 2021, the wave was clearly predictable, when B.1.617.2 showed a steep rise in parallel with a similar rise in monthly new cases and deaths. Analyses taken together distinctly delineate that the formation of the second COVID-19 wave in India was closely associated with the rise of B.1.617 lineage variants, particularly its sub-lineage B.1.617.2 (Figs 1-2). Our findings get corroborated by a recent preprint article published by a group of scientists affiliated with INSACOG—Indian SARS-CoV-2 Genomic Consortia, who observed a similar pattern in rise of B.1.617 lineage, primarily B.1.617.2 variant in Delhi before second wave (1). The group also showed that this variant is as much as 50 % more transmissible than B.1.1.7 and dominated in the samples (76%) analyzed from vaccine breakthrough infections in Delhi (1).

There have been multiple limitations in our study which could have impacted the interpretation of the findings. Firstly, the samples used in our analyses were not representative for populations asfor many geographical regions that has been greatly disproportioned. Hence, the genomic sequence data presented in this study doesn't reflect the accurate epidemiological scale of spread of the variants in the reported geographical regions, but only shows their relative proportion in the samples for which genomic sequences were uploaded in GISAID database. We assumed that similar proportions between variants exist in the actual population. Secondly, there has been inconsistency in reporting and uploading of the genomic sequences, which constrained examining a daily trend in the spread of the variants. The paucity of the genomic sequences and inconsistency in their uploading on the used databases for some states/union territories the variant dominance difficult. Thirdly, we didn't include data of SARS-CoV-2 genomic sequences beyond 30th April, 2020, when the second wave was supposedly reaching a peak. There had been sudden interruption in the genomic sequencing and uploading in GISAID database in the initial weeks of May, 2021 from the involved laboratories, probably due to the emergency situations created by the second wave, hence sufficient data couldn't availed for this period. However, this couldn't have significantly impacted the analyses as all epidemiological trends were clear by the end of April, 2021 defining a second COVID-19 wave.

Based on the findings of this study, we conclude that genomic surveillance of the variants augmented with epidemiological data can be a promising tool for predicting eminent COVID-19 waves in advance as early as it starts rising. However, the accuracy of the prediction would largely dependent on the population matched viral genomic sequencing and consistency in

medRxiv preprint doi: https://doi.org/10.1101/2021.06.09.21258612; this version posted June 12, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

uploading of the data from all geographical regions, which currently seems a big hindrance restricting timely predictions.

Online Methods:

Monthly distribution of SARS-CoV-2 variants and new cases and deaths due to COVID-19

The genomic sequence data for SARS-CoV-2 and official epidemiological data for COVID-19 for the period from 1st December 2020 to 30th April 2021 from India were downloaded from EpiCoV™ database of Global Initiative on Sharing All Influenza Data (GISAID) and Worldometer: https://www.worldometers.info/coronavirus/coronavirus/country/india/) respectively. A total of 9994 SARS-CoV-2 genomic sequences were analyzed, which were uploaded in GISAID database (last reporting date 2.06.2021) showing dates of collection for the samples falling under in aimed period of study.

The number of sequences for each SARS-CoV-2 variants was retrieved using automatic search function feeding information for the lineage/sub-lineage and collection dates in EpiCoV™ database of GISAID. Total number of sequence per month for the studied time period was noted for each variant and their relative proportions were calculated (in percent). Data was tabulated and monthly distribution of each variant was charted against the COVID-19 epidemiological data (total new case and deaths per month) and graphs were plotted to visualize the trends. Further genomic sequence of SARS-CoV-2 variants were analyzed for the individual states and union territories, to check if there has been any deviation from the collective data trends.

Tracing spread of B.1.617 lineage variants since first reported case

A tracing of the spread of B.1.617 lineage variants (B.1.617, and B.1.617.1-3) was performed since first case for which genomic sequence data were uploaded on GISAID database. Monthly distributions of the numbers of reported genomic sequence for the each B.1.617 lineage variant were charted and graphs were plotted to visualize the trends.

Acknowledgements: The study has used SARS-CoV-2 genomic sequence and epidemiological data from GISAID (https://www.gisaid.org/) and Worldometer (https://www.worldometers.info/coronavirus/coronavirus/country/india) databases respectively.

Author contributions: A.K., P.D., G.K., collected samples, and performed data analysis. A.K. wrote first draft. A.K., R.K.N., R.K.J., R. P., C.S., and S. N. P. edited paper. All authors consented for submission of the final draft.

Ethics statement: The study used publically available open access data hence an ethical approval was precluded.

Funding declaration: None

medRxiv preprint doi: https://doi.org/10.1101/2021.06.09.21258612; this version posted June 12, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

Conflict of interest: None

References:

1. Mahesh S Dhar, et al. "Genomic characterization and Epidemiology of an emerging SARS-CoV-2 variant in Delhi, India." *medRxiv* (2021).doi: https://doi.org/10.1101/2021.06.02.21258076
2. Tracking SARS-CoV-2 variants. https://www.who.int/en/activities/tracking-SARS-CoV-2-variants/
3. Alaa Abdel Latif, et al. "B.1.617 Lineage Report."outbreak.info, available at https://outbreak.info/situation-reports?pango=B.1.617. Accessed 6 June 2021.
4. WHO, Weekly epidemiological update on COVID-19 - 11 May 2021. https://www.who.int/publications/m/item/weekly-epidemiological-update-on-covid-19---11-may-2021
5. Tada, Takuya, et al. "The Spike Proteins of SARS-CoV-2 B. 1.617 and B. 1.618 Variants Identified in India Provide Partial Resistance to Vaccine-elicited and Therapeutic Monoclonal Antibodies." *bioRxiv* (2021).https://doi.org/10.1101/2021.05.14.444076
6. Hoffmann, Markus, et al. "SARS-CoV-2 variant B. 1.617 is resistant to Bamlanivimab and evades antibodies induced by infection and vaccination." *bioRxiv* (2021).https://doi.org/10.1101/2021.05.04.442663
7. Ferreira, Isabella, et al. "SARS-CoV-2 B. 1.617 emergence and sensitivity to vaccine-elicited antibodies." *bioRxiv* (2021).https://doi.org/10.1101/2021.05.08.443253
8. Yadav, PragyaDhruv, et al. "SARS CoV-2 variant B. 1.617. 1 is highly pathogenic in hamsters than B. 1 variant." *bioRxiv* (2021).https://doi.org/10.1101/2021.05.05.442760
9. Planas, Delphine, et al. "Reduced sensitivity of infectious SARS-CoV-2 variant B. 1.617. 2 to monoclonal antibodies and sera from convalescent and vaccinated individuals." *bioRxiv* (2021). https://doi.org/10.1101/2021.05.26.445838
10. Bernal, Jamie Lopez, et al. "Effectiveness of COVID-19 vaccines against the B. 1.617. 2 variant." *medRxiv* (2021).https://doi.org/10.1101/2021.05.22.21257658

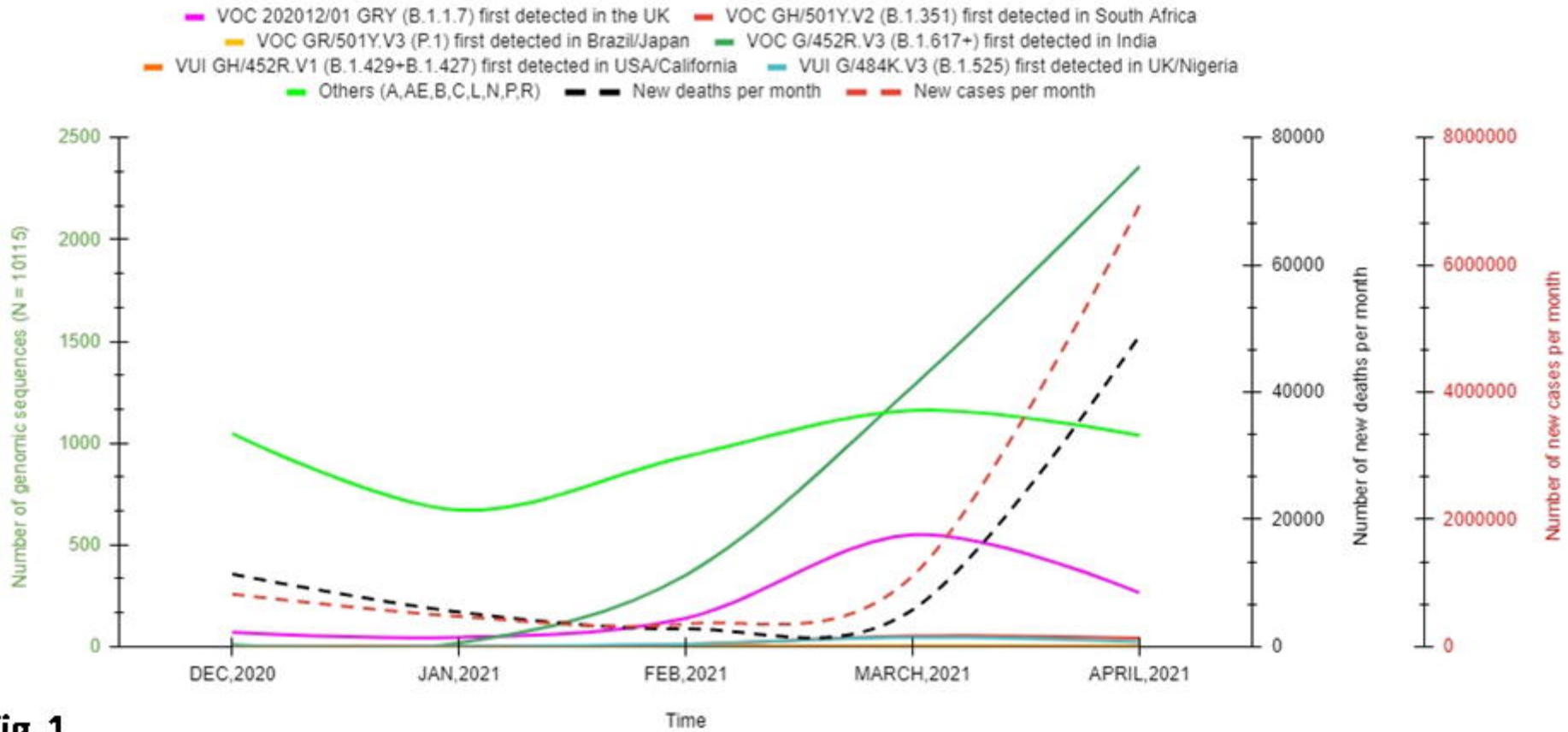

VOC 202012/01 GRY (B.1.1.7) first detected in the UK
VOC GH/501Y.V2 (B.1.351) first detected in South Africa
VOC GR/501Y.V3 (P.1) first detected in Brazil/Japan
VOC G/452R.V3 (B.1.617+) first detected in India
VUI GH/452R.V1 (B.1.429+B.1.427) first detected in USA/California
VUI G/484K.V3 (B.1.525) first detected in UK/Nigeria
Others (A,AE,B,C,L,N,P,R)
New deaths per month
New cases per month
Number of genomic sequences (N = 10115)
2500
2000
1500
1000
500
0
80000
60000
40000
20000
0
Number of new deaths per month
8000000
6000000
4000000
2000000
0
Number of new cases per month
DEC,2020
JAN,2021
FEB,2021
MARCH,2021
APRIL,2021
Time

Fig. 1

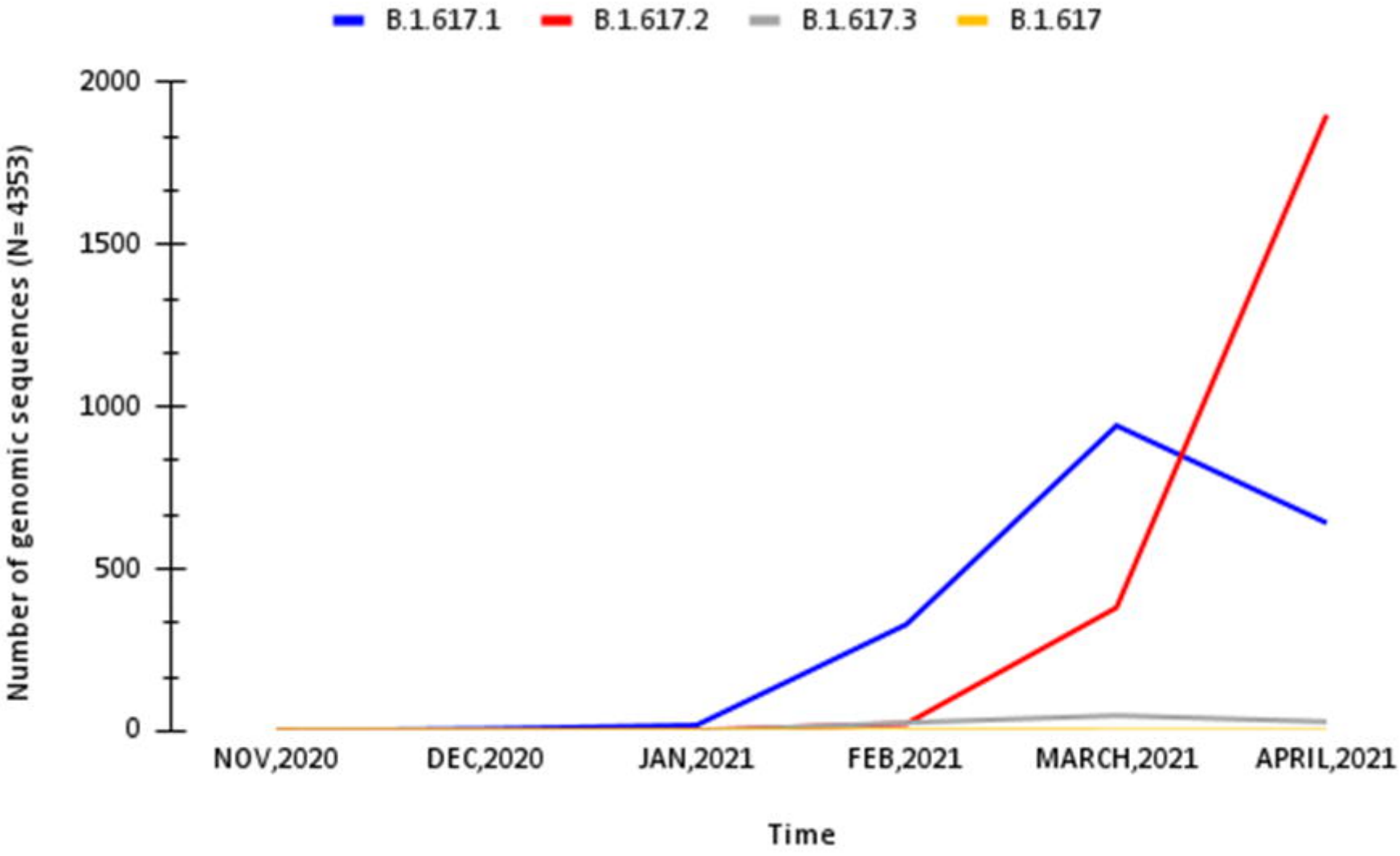

B.1.617.1
B.1.617.2
B.1.617.3
B.1.617
2000
1500
1000
500
0
Number of genomic sequences (N= 4353)
NOV,2020
DEC,2020
JAN,2021
FEB,2021
MARCH,2021
APRIL,2021
Time

Fig. 2