medRxiv preprint doi: https://doi.org/10.1101/2021.11.17.21266463; this version posted February 10, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.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 4.0 International license .

How many relevant SARS-CoV-2 variants might we expect in the future?

Roberto Littera[1,2], Maurizio Melis[2,*]

ABSTRACT

Objectives. The emergence of new SARS-CoV-2 variants is a major challenge in the management of Covid-19 pandemic. A crucial issue is to quantify the number of variants which may represent a potential risk for public health in the future.

Methods. We fitted the data on the most relevant SARS-CoV-2 variants recorded by the World Health Organization (WHO). The function exploited for the fit is related to the total number of infected subjects in the world since the start of the epidemic.

Results. We found that the relevant SARS-CoV-2 variants up to November 2021 were about 34. Moreover, the number of new relevant variants per ten million cases turned out to be 1.25 in November 2021, slightly decreased in comparison to the value of 1.75 in March 2020.

Conclusions. Our simple mathematical model can evaluate the number of relevant SARS-CoV-2 variants as the cumulative number of cases increases worldwide and may represent a useful tool in planning strategies to effectively contrast the pandemic.

Keywords: Covid-19; SARS-CoV-2 variants; infectious diseases; epidemiology.

NOTE: This preprint reports new research that has not been certified by peer review and should not be used to guide clinical practice.

[1] Complex Structure of Medical Genetics, R. Binaghi Hospital, Local Public Health and Social Care Unit (ASSL) of Cagliari, Sardinian Regional Company for the Protection of Health (ATS Sardegna), Cagliari, Italy.
[2] Association for the Advancement of Research on Transplantation O.d.V., non-profit organisation, Cagliari, Italy.
[*] Corresponding author.
E-mail addresses: roby.litter@gmail.com (R. Littera), maurizio.melis@gmail.com (M. Melis).

medRxiv preprint doi: https://doi.org/10.1101/2021.11.17.21266463; this version posted February 10, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.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 4.0 International license .

Introduction

Most mutations in the genome of the severe acute respiratory syndrome coronavirus (SARS-CoV-2) are neutral or only mildly deleterious. However, a small proportion of mutations can increase infectivity and promote virus-host interactions that are critical to the establishment of persistent and more severe infection [1, 2]. For example, mutations in the spike protein, which mediates attachment of the virus to host cell-surface receptors [3], can have significant effects on virus behaviour. In order to effectively control the pandemic, it is imperative to investigate the emergence and spread of variants with an impact on disease transmission and human health [4].

SARS-CoV-2 sequences are shared daily on public databases such as the Global Initiative on Sharing All Influenza Data (GISAID) [5] or the European Centre for Disease Prevention and Control (ECDC) [6], which significantly contribute to surveillance of the pandemic.

The World Health Organization has established that SARS-CoV-2 variants representing a possible risk to public health can be divided into three distinct categories [1]: variants under monitoring (VUMs), variants of interest (VOIs) and variants of concern (VOCs).

Variants Under Monitoring (VUMs) are associated with genetic mutations which alter virus characteristics, although evidence of phenotypic or epidemiological impact is still unclear.

Variants of Interest (VOIs) are associated with: *i)* genetic mutations which affect transmissibility, disease course, diagnostic or therapeutic escape; *ii)* relevant community transmission with an emerging risk to global public health.

Variants of Concern (VOCs) are associated with one or more of the following characteristics: *i)* increase in transmissibility; *ii)* increase in virulence or change in disease severity; *iii)* decrease in effectiveness of social measures, diagnostics, vaccines and therapeutics.

Given the continuous evolution of the SARS-CoV-2 virus, variants may be reclassified over time.

In the present study, we fitted the WHO data [1, 2] by exploiting a function which exclusively depends on the number of infected cases worldwide. Our fit allows for a fairly good estimate of the number of relevant variants that can be expected to appear for a given number of infected subjects throughout the world.

medRxiv preprint doi: https://doi.org/10.1101/2021.11.17.21266463; this version posted February 10, 2022. 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 4.0 International license .

Furthermore, our approach can predict the number of new relevant variants per ten million cases in any epidemiological situation.

The number of new relevant variants per ten million cases decreases very slowly as the cumulative number of Covid-19 cases increases. Therefore, it becomes crucial to carefully monitor and reduce virus circulation in order to avoid the emergence of new variants which may not be suitably covered by the vaccines and drugs currently available [7].

Despite the huge efforts put forth by healthcare services in many countries, vaccination campaigns have not achieved a population coverage that is high enough to prevent the spread of SARS-CoV-2. The WHO estimate [8] is that in Europe and Central Asia alone the current fourth wave of the Covid-19 pandemic is likely to cause more than half a million deaths.

Although it is obvious that the number of relevant variants increases with the cases throughout the world, it is extremely difficult to find out the precise relationship between these two variables. Our model is a simple attempt to make a fairly reliable estimate of the risk of new variants that can impact public health as the virus continues to spread.

Methods

By means of Wolfram Mathematica [9] we fitted the data on SARS-CoV-2 variants in order to evaluate the cumulative number v of relevant SARS-CoV-2 variants versus the cumulative number N of infected subjects worldwide.

The function v fitting the WHO data must satisfy the following conditions:

1) The function v varies from zero to infinity. If there is no infection, the number of variants is zero; vice versa, if the virus replicates infinite times, the cumulative number of variants is also infinite.

2) The function v increases monotonically, therefore the first derivative $v'(N)$ is positive. The cumulative number of variants increases with the number of infections, i.e. as the virus replications increase.

3) The first derivative of v decreases monotonically, therefore the second derivative $v''(N)$ is negative. As the cumulative number of variants increases with the total cases in the world, the emergence of new virus mutations turns out to be slightly less frequent.

medRxiv preprint doi: https://doi.org/10.1101/2021.11.17.21266463; this version posted February 10, 2022. 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 4.0 International license .

The fit of WHO data was obtained by means of the following function:

$$v(N) = k \cdot \frac{N}{\log N}$$

where k is the constant of the numerical fit and "$\log N$" represents the natural logarithm of N. This function satisfies all the previous conditions, as shown below:

1) $\lim_{N \to 0} v = 0$ and $\lim_{N \to \infty} v = \infty$.

2) $v'(N) = k \cdot \frac{\log N - 1}{(\log N)^2} > 0$ for $N > e$, where $e \cong 2.72$ is the Euler's number.

3) $v''(N) = k \cdot \frac{2 - \log N}{(\log N)^3} < 0$ for $N > e^2$, with $e^2 \cong 7.39$.

The number n of new relevant variants per ten million cases ($\Delta N = 10^7$) turns out to be:

$$n \cong v'(N) \cdot \Delta N = 10^7 \cdot k \cdot \frac{\log N - 1}{(\log N)^2}$$

The relative variation $|n'(N)|$ of new relevant variants per ten million cases decreases as the number N of cases increases: $|n'(N)| \cong |v''(N)| \cdot \Delta N = 10^7 \cdot k \cdot \left|\frac{2 - \log N}{(\log N)^3}\right|$.

The choice of the function $v = k \cdot N / \log N$ exploited in the fit can be justified through the heuristic arguments discussed in Appendix A.

Results

We fitted the data recorded by WHO [1, 2] up to November 2021 by means of a specific code written with Wolfram Mathematica 12.1.3 [9].

Table 1 lists the characteristics of SARS-CoV-2 variants reported by WHO [1, 2]: date and country of the earliest detection, PANGO (Phylogenetic Assignment of Named Global Outbreak) and WHO classification, current relevance (variants of concern, variants of interest or under monitoring), total number of cases in the world at the end of the month of detection and cumulative number of variants. PANGO is a rule-based nomenclature system for naming and tracking SARS-CoV-2 genetic lineages [10]. The numerical fit of the WHO data was obtained by means of the function $v(N) = k \cdot N / \log N$, where the constant of the numerical fit is $k = 2.55 \cdot 10^{-6}$. The 95% confidence interval (CI) of k is given by 95% CI $= (2.08 - 3.03) \cdot 10^{-6}$. The adjusted R-squared, measuring the goodness of the fit, turned out to

medRxiv preprint doi: https://doi.org/10.1101/2021.11.17.21266463; this version posted February 10, 2022. 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 4.0 International license .

be $R^2 = 0.92$. Further technical details on the fit of WHO data are presented in Appendix B, where we also considered other possible functions satisfying the three conditions listed in the Methods section.

Table 1
Characteristics of SARS-CoV-2 variants recorded by WHO [1, 2]: date and country of the earliest detection, PANGO and WHO nomenclature, current relevance (concern, interest or under monitoring) and cumulative number of cases in the world at the end of the month of detection. The last column summarises the cumulative number of observed relevant variants.

Earliest detection	Country	PANGO lineage[#]	WHO name	Current relevance[§]	Cases in the world	Cumulative number of variants
Mar-2020	USA	B.1.427 B.1.429	Epsilon[*]	-	750890	1
Apr-2020	Brazil	P.2	Zeta[*]	-	3090445	2
May-2020	South Africa	B.1.351	Beta	VOC	5934936	5
	Multiple countries	B.1.1.523	-	-		
	Multiple countries	B.1.619	-	-		
Sep-2020	UK	B.1.1.7	Alpha	VOC	32730945	6
Oct-2020	India	B.1.617.2	Delta	VOC	44888869	8
	India	B.1.617.1	Kappa[*]	-		
Nov-2020	Brazil	P.1	Gamma	VOC	62195274	14
	USA	B.1.526	Iota[*]	-		
	Indonesia	B.1.466.2	-	-		
	Multiple countries	B.1.214.2	-	-		
	Multiple countries	B.1.1.519	-	-		
	Multiple countries	B.1.620	-	-		
Dec-2020	Multiple countries	B.1.525	Eta[*]	-	79231893	16
	Peru	C.37	Lambda	VOI		
Jan-2021	Philippines	P.3	Theta[*]	-	98925221	22
	Multiple countries	R.1	-	-		
	Colombia	B.1.621	Mu	VOI		
	Multiple countries	AZ.5	-	VUM		
	Russian Federation	AT.1	-	-		
	Multiple countries	C.36.3	-	-		
Mar-2021	UK	AV.1	-	-	126697603	24
	Domenican Republic	B.1.630	-	-		
May-2021	South Africa	C.1.2	-	VUM	169597415	25
Sep-2021	Republic of Congo	B.1.640	-	VUM	231416660	26
Nov-2021	Multiple countries	B.1.1.529	Omicron[**]	VOC	260493573	27

[#] It includes all descendent lineages [1].
[§] Variants with no current relevance were formerly VOIs (Epsilon, Theta and Zeta) or VUMs [1].
[*] Former VOIs [1].
[**] Omicron includes lineage B.1.1.529 and descendent lineages BA.1, BA.1.1, BA.2 and BA.3 [1].

medRxiv preprint doi: https://doi.org/10.1101/2021.11.17.21266463; this version posted February 10, 2022. 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 4.0 International license .

Figure 1 represents the cumulative number v of relevant SARS-CoV-2 variants versus the cumulative number N of cases in the world. The dots from 1 to 12 correspond to the data reported by WHO [1, 2] from March 2020 to November 2021; the solid line represents the function $v = k \cdot N/\log N$ used in the fit with Wolfram Mathematica.

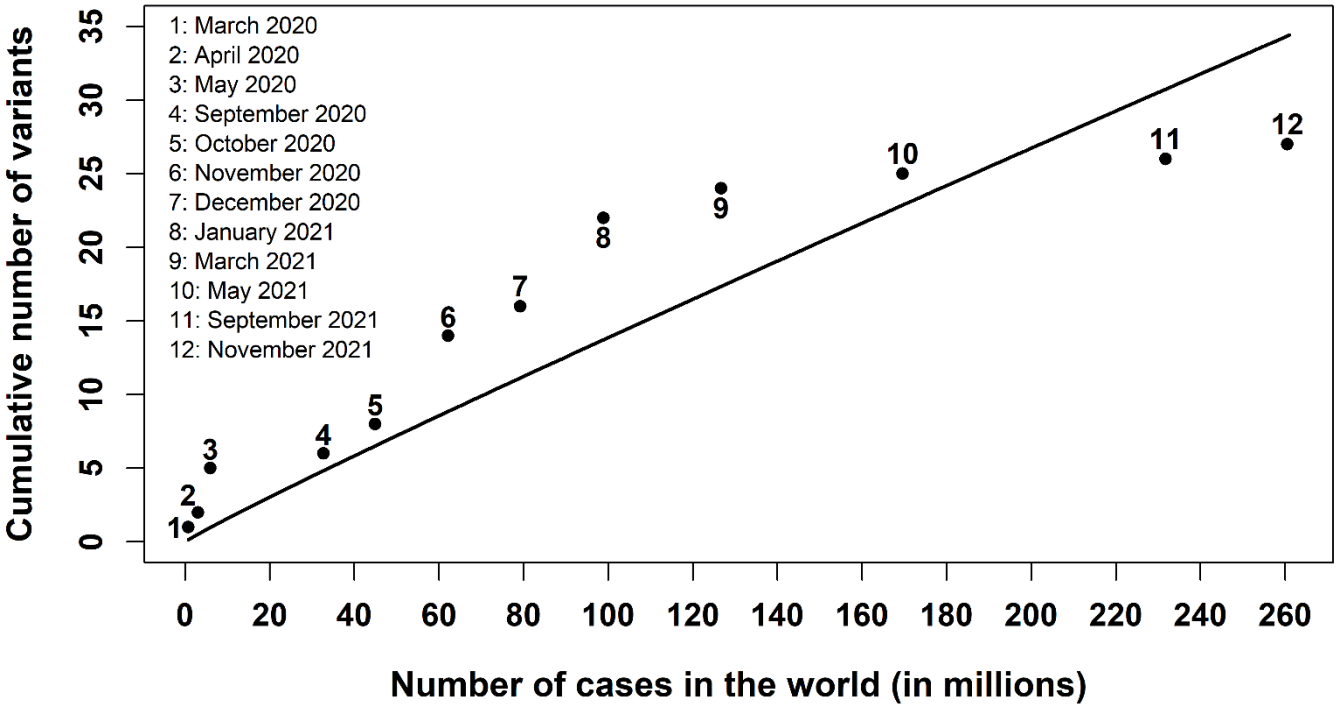

Figure 1. Cumulative number of relevant SARS-CoV-2 variants versus the cumulative number of cases in the world. The dots from 1 to 12 indicate the data reported by WHO [1, 2] from March 2020 to November 2021; the solid line represents the function $v = k \cdot N/\log N$ used in the numerical fit with Wolfram Mathematica.

The analytical formula v exploited in the fit allows to predict both the cumulative number of relevant variants and the new relevant variants for a given number of cases in the world.

The total cases in the world up to the 28th of November 2021 were 260493573 [2]. The corresponding cumulative number of relevant SARS-CoV-2 variants was 34.3, i.e. seven variants more than the last WHO update [1] in November 2021, when 27 relevant variants were recorded.

As discussed in the Methods section, the number n of new relevant variants per ten million (10^7) cases is $n = 10^7 \cdot k \cdot \frac{\log N - 1}{(\log N)^2}$, which becomes $n = 25.5 \cdot \frac{\log N - 1}{(\log N)^2}$ by substituting the numerical value of k. From March 2020 to November 2021 the number of new variants per ten million cases decreased only by 28.6%, from 1.75 to 1.25.

medRxiv preprint doi: https://doi.org/10.1101/2021.11.17.21266463; this version posted February 10, 2022. 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 4.0 International license .

Figure 2 reports the number n of new relevant SARS-CoV-2 variants per ten million cases versus the cumulative number N of cases in the world. The analytical behaviour represented by the plot is given by the formula $n = 25.5 \cdot \frac{\log N - 1}{(\log N)^2}$.

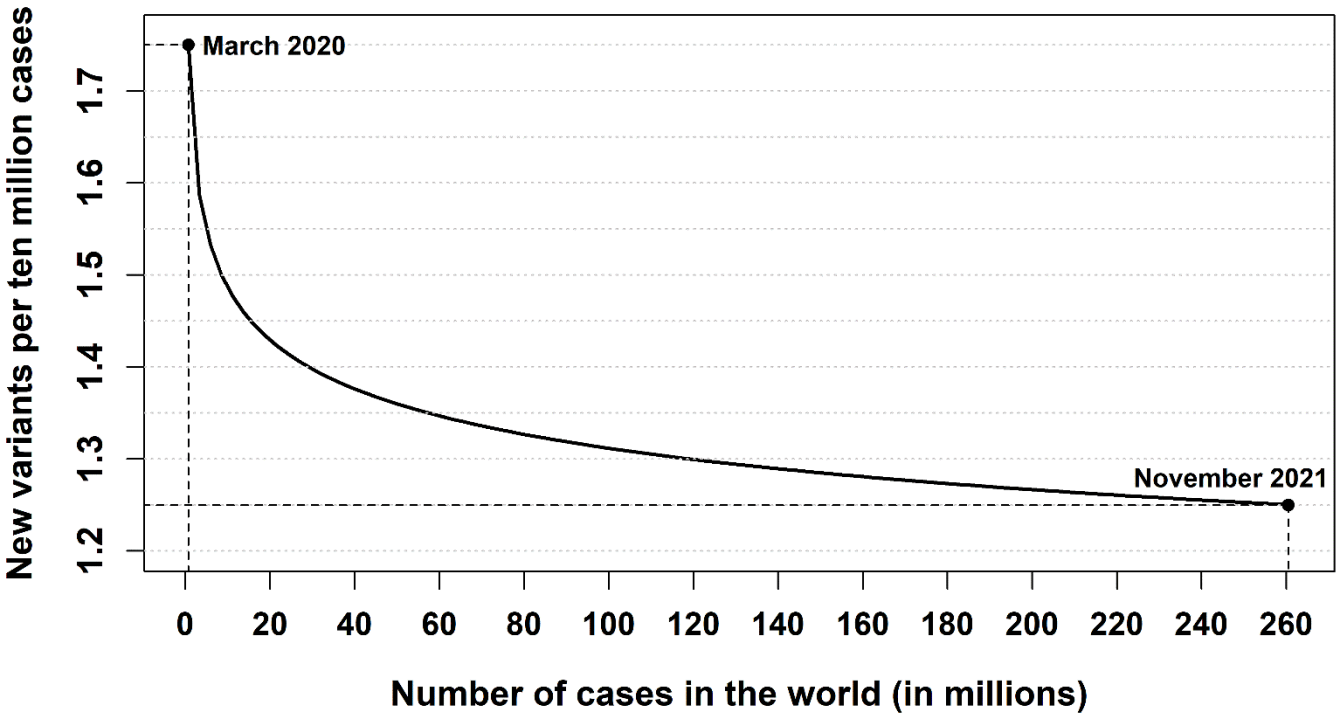

Figure 2. Number n of new relevant SARS-CoV-2 variants per ten million cases versus the cumulative number of cases in the world. From March 2020 to November 2021 n decreased from 1.75 to 1.25.

The relative variation $|n'(N)|$ of the number n of new relevant variants per ten million cases is given by $|n'(N)| = 25.5 \cdot \left|\frac{2 - \log N}{(\log N)^3}\right|$, with $|n'(N)| \ll 1$ for $N \gg 1$. This analytical result implies that the number n of new relevant variants per ten million cases decreases very slowly as the virus continues to circulate. For instance, n will remain above 1.10 as long as the cumulative cases in the world increase to about 4.2 billion.

Figure 3 represents the predicted increase of the cumulative number of relevant variants per each step of ten million infections, from 260 to 400 million cumulative cases in the world. In this range of cases, the number of relevant variants is predicted to increase from 34.3 to 51.6.

As shown in Figure 3, the total infected cases were about 260 million at the end of November 2021.

medRxiv preprint doi: https://doi.org/10.1101/2021.11.17.21266463; this version posted February 10, 2022. 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 4.0 International license .

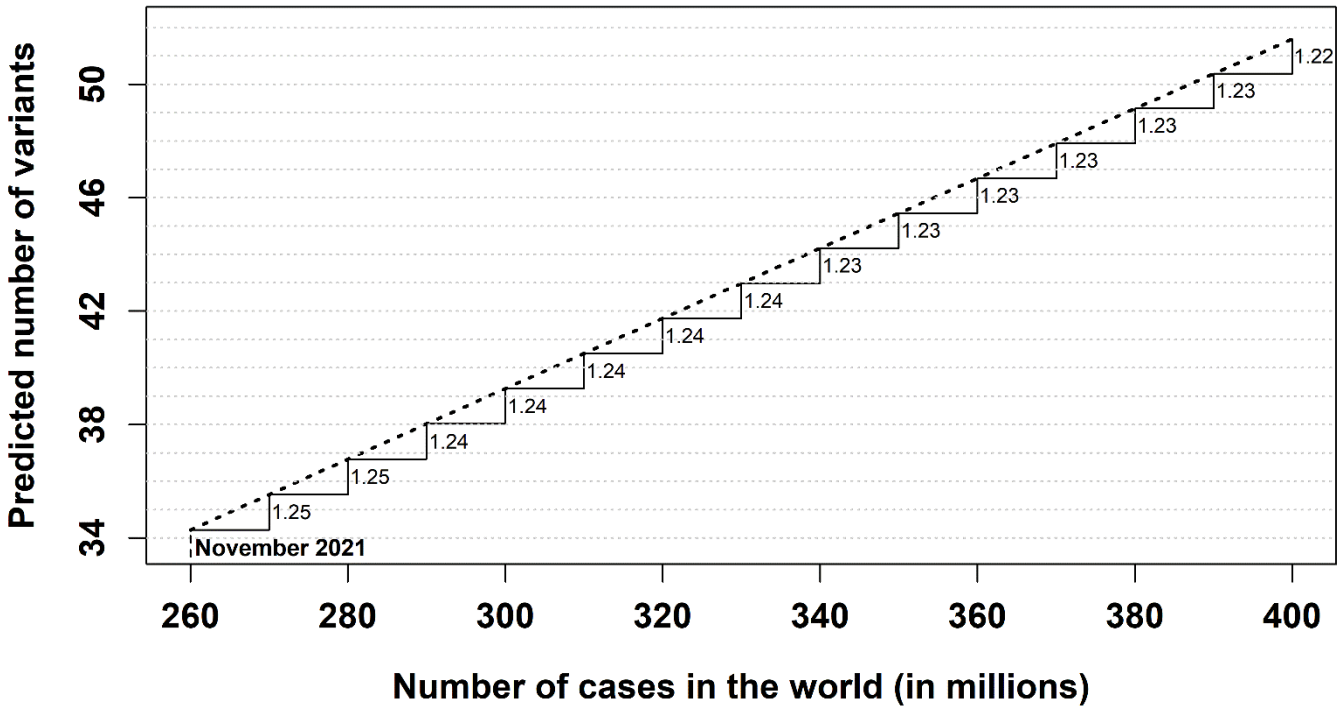

Figure 3. Prediction of the cumulative number of SARS-CoV-2 variants from 260 to 400 million cases in the world. The dotted line represents the function $v = k \cdot N/\log N$, while each step shown in the plot corresponds to the predicted number of new relevant variants per ten million cases.

Our model only focuses on the relationship between the number of virus replications and the emergence of relevant variants. All other factors involved in the diffusion of new variants were not taken into account. For this reason, we supposed that the parameter k in the fit $v = k \cdot N/\log N$ is constant, although it actually varies with the other factors affecting the emergence of relevant variants.

Discussion

Since the start of the Covid-19 pandemic there has been an impressive global effort in investigating every aspect of the coronavirus epidemic [11], including immunogenetic [12, 13] and epidemiological [14] issues.

In this study, we built a simple mathematical model to calculate the number of relevant SARS-CoV-2 variants from the number of infected cases in the world. By fitting the WHO data listed in Table 1, we obtained an analytical formula which allows to predict both the cumulative number of variants and the number of new relevant variants in a given epidemiological situation. For example, up to the 28th of November 2021, the cumulative number of cases worldwide were 260493573 [2], corresponding to 34.3 relevant variants, i.e. seven variants more than the last WHO update in November 2021 [1]. Analogously,

medRxiv preprint doi: https://doi.org/10.1101/2021.11.17.21266463; this version posted February 10, 2022. 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 4.0 International license .

we found that the number of new relevant variants per ten million cases was 1.25 in November 2021, decreased only by 28.6% in comparison to March 2020 when the number was 1.75.

Our method depends critically on the WHO efficiency in tracking the most relevant SARS-CoV-2 variants. A different approach would be to consider the whole number of variants detected by genomic sequencing of SARS-CoV-2 and recorded in public databases such as GISAID [5]. This choice would be independent from the WHO targeting of the most relevant variants but would be less interesting from a clinical viewpoint since only the variants affecting virus transmission or disease severity are important to the control and management of the pandemic.

As shown in Figure 2, the number of new relevant variants per ten million cases decreases very slowly as the cumulative number of cases increases. Therefore, the persistence of virus circulation will always cause the emergence of new relevant SARS-CoV-2 variants.

Our model does not take into account the fact that the number of virus replications is different in each infected subject. However, the average number of replications can be assumed to be constant over a large number of infected cases, such as those recorded worldwide.

The diffusion of new relevant variants depends on a large variety of factors. For instance, the ability to monitor all virus mutations, the effectiveness of containment measures, vaccination campaigns, the evolution patterns of the virus [15], and so on. All these factors were ignored in our model, which only focuses on the number of virus replications. The parameter k appearing in the fit $v = k \cdot N/\log N$ is not actually constant, as supposed in our model, but varies with all the other factors which can affect the emergence of relevant variants. None the less, our model provides a fairly good estimate of what we can expect in the future.

SARS-CoV-2 vaccination has led to a decrease in hospitalisations and disease severity. Nevertheless, the current number of infections is still too high to prevent the appearance of new variants potentially dangerous to public health.

The number of variants increases with the number of cases *in the world*: this result underlines the urgency of making every effort to reduce the impact of virus replication in all geographical areas. The risk that

medRxiv preprint doi: https://doi.org/10.1101/2021.11.17.21266463; this version posted February 10, 2022. 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 4.0 International license .

new relevant variants may emerge anywhere in the world indicates that the winning strategy is not to leave any country behind in the battle against the virus.

The possibility to predict the number of new relevant SARS-CoV-2 variants will become increasingly important in future to ensure optimal planning of vaccination campaigns by healthcare services, united in the awareness that new variants can change the characteristics of the virus and greatly influence the global management of the pandemic.

Authors' contributions

The authors contributed equally to the article.

Conflicts of Interest

The authors declare that no competing interests exist.

Funding

The authors received no specific funding for this work.

Ethical approval

Not applicable.

Acknowledgments

The authors are grateful to Anna Maria Koopmans for translations, professional writing assistance and preparation of the manuscript.

References

[1] WHO (World Health Organization). Tracking SARS-CoV-2 variants. Working Definitions and Actions Taken. https://www.who.int/en/activities/tracking-SARS-CoV-2-variants/. Accessed 30 Nov 2021.

medRxiv preprint doi: https://doi.org/10.1101/2021.11.17.21266463; this version posted February 10, 2022. 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 4.0 International license .

[2] WHO (World Health Organization). Covid-19 Weekly Epidemiological and Operational Update. https://www.who.int/emergencies/diseases/novel-coronavirus-2019/situation-reports/. Accessed 30 Nov 2021.

[3] Harvey WT, Carabelli AM, Jackson B, Gupta RK, Thomson EC, Harrison EM, et al. SARS-CoV-2 variants, spike mutations and immune escape. *Nat Rev Microbiol* 2021; 19: 409-24. https://doi.org/10.1038/s41579-021-00573-0.

[4] Otto SP, Day T, Arino J, et al. The origins and potential future of SARS-CoV-2 variants of concern in the evolving COVID-19 pandemic. *Current Biology* 2021; 31: R918–R929. https://doi.org/10.1016/j.cub.2021.06.049.

[5] GISAID (Global Initiative on Sharing All Influenza Data). hCoV-19 tracking of variants. https://www.gisaid.org/.

[6] ECDC (European Centre for Disease Prevention and Control). SARS-CoV-2 variants of concern. https://www.ecdc.europa.eu/en/covid-19/variants-concern.

[7] Krause PR, Fleming TR, Longini IM, Peto R, Briand S, Heymann DL, et al. SARS-CoV-2 Variants and Vaccines. *N Engl J Med* 2021; 385: 179-86. https://doi.org/10.1056/NEJMsr2105280.

[8] WHO (World Health Organization) Regional Office for Europe. Update on COVID-19: Europe and Central Asia again at the epicentre of the pandemic. https://www.euro.who.int/en/media-centre/sections/statements/2021/statement-update-on-covid-19-europe-and-central-asia-again-at-the-epicentre-of-the-pandemic. Accessed 4 Nov 2021.

[9] Wolfram Research, Inc. Mathematica 12.1.3 (Trial Version). Champaign, Illinois, US. Released in July 2021.

[10] Rambaut A, Holmes EC, O'Toole Á, Hill V, McCrone JT, Ruis C, et al. A dynamic nomenclature proposal for SARS-CoV-2 lineages to assist genomic epidemiology. *Nat Microbiol* 2020; 5: 1403-7. https://doi.org/10.1038/s41564-020-0770-5.

[11] Lavezzo E, Franchin E, Ciavarella C, Cuomo-Dannenburg G, Luisa Barzon L, Del Vecchio C, et al. Suppression of a SARS-CoV-2 outbreak in the Italian municipality of Vo'. *Nature* 2020; 584: 425-9. https://doi.org/10.1038/s41586-020-2488-1.

medRxiv preprint doi: https://doi.org/10.1101/2021.11.17.21266463; this version posted February 10, 2022. 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 4.0 International license .

[12] Littera R, Campagna M, Deidda S, Angioni G, Cipri S, Melis M, et al. Human Leukocyte Antigen Complex and Other Immunogenetic and Clinical Factors Influence Susceptibility or Protection to SARS-CoV-2 Infection and Severity of the Disease Course. The Sardinian Experience. *Front Immunol* 2020; 11: 605688. https://doi.org/10.3389/fimmu.2020.605688.

[13] Littera R, Chessa L, Deidda S, Angioni G, Campagna M, Lai S, et al. Natural killer-cell immunoglobulin-like receptors trigger differences in immune response to SARS-CoV-2 infection. *PLoS ONE* 2021; 16(8): e0255608. https://doi.org/10.1371/journal.pone.0255608.

[14] Melis M, Littera R. Undetected infectives in the Covid-19 pandemic. *Int J Inf Dis* 2021; 104: 262-8. https://doi.org/10.1016/j.ijid.2021.01.010.

[15] Giovanetti M, Benedetti F, Campisi G, et al. Evolution patterns of SARS-CoV-2: Snapshot on its genome variants. *Biochem Biophys Res Commun* 2021; 538: 88-91. https://doi.org/10.1016/j.bbrc.2020.10.102.