SARS-CoV-2 genome-based severity predictions correspond to lower qPCR values and higher viral load

Martin Skarzynski¹, Erin M. McAuley¹, Ezekiel J. Maier¹, Anthony C. Fries², Jameson D. Voss³, and Richard R. Chapleau^{2*}

¹Booz Allen Hamilton, Bethesda, MD 20814, USA

²US Air Force School of Aerospace Medicine, Wright Patterson AFB, OH 45433, USA

³US Air Force Medical Readiness Agency, Falls Church, VA 22042, USA

*Correspondence

The views expressed are those of the authors and do not reflect the official guidance or position of the United States Government, the Department of Defense or of the United States Air Force.

Abstract

The 2019 coronavirus disease (COVID-19) pandemic has demonstrated the importance of predicting, identifying, and tracking mutations throughout a pandemic event. As the COVID-19 global pandemic surpassed one year, several variants had emerged resulting in increased severity and transmissibility. In order to reduce the impact on human life, it is critical to rapidly identify which genetic variants result in increased virulence or transmission. To address the former, we evaluated if a genome-based predictive algorithm designed to predict clinical severity could predict polymerase chain reaction (PCR) results, as a surrogate for viral load and severity. Using a previously published algorithm, we compared the viral genome-based severity predictions to

clinically-derived PCR-based viral load of 716 viral genomes. For those samples predicted to be “severe” (predicted severity score > 0.5), we observed an average cycle threshold (Ct) of 18.3, whereas those in the “mild” category (severity prediction < 0.5) had an average Ct of 20.4 ($P = 0.0017$). We found a non-trivial correlation between predicted severity probability and cycle threshold ($r = -0.199$). Additionally, when divided into quartiles by prediction severity probability, the most probable quartile ($\geq 75\%$ probability) had a Ct of 16.6 ($n=10$) as compared to those least probable to be severe ($< 25\%$) of 21.4 ($n=350$) ($P = 0.0045$). Taken together, our results suggest that the severity predicted by a genome-based algorithm can be related to the metrics from the clinical diagnostic test, and that relative severity may be inferred from diagnostic values.

Keywords

SARS-CoV-2, COVID-19 severity, predictive modeling, genetic methods, viral epidemiology

Introduction

A classic model of disease causation is the epidemiologic triad of host, agent, and environment. Stated differently, the severity of an illness is based on an interplay between method of exposure (environment), pathogenicity of the organism (agent), and the host (host susceptibility and host response to the infectious agent). The recent SARS-CoV-2 pandemic has demonstrated that substantial diversity in both the host and virus can lead to a wide spectrum of clinical outcomes. Early research into symptom severity largely focused on host phenotypes, such as blood type [Goker 2020], age [O'Driscoll 2020], gender [Pradhan 2020], etc. As the scale of the pandemic grew, the role of geographic region and viral mutations in severe clinical outcomes began to emerge [Zhao 2020], followed by additional insights on host genetic susceptibility [COVID19 Host Genetics Initiative 2020]. Finally, efforts to predict a patient's outcomes using computational models developed with phenotypic, genetic, and demographic data are being pursued in order to tailor patient care and manage resources [Voss, 2021; Dong 2020, Marin 2021].

Early identification of patients at increased risk of developing severe symptoms can help preserve life and health. Estimates of viral load upon admission have been shown to be correlated with higher mortality [Choudhuri et al, 2021; Pujadas et al, 2021]. Using real-time PCR data, Choudhuri and colleagues demonstrated that increased cycle thresholds were associated with 9% reduction in the odds of in-hospital mortality [Choudhuri et al, 2021]. The greatest difference was found for those patients reporting to the hospital with a cycle threshold below 23; these patients encountered 3.9-fold increased odds of in-hospital mortality compared to patients with cycle thresholds above 33. Overall, using PCR cycle thresholds as a predictor of

outcome the area under the curve was found to be 0.68, suggesting an effective but limited discriminative ability for severity classification by cycle threshold at the time of admission.

The use of SARS-CoV-2 genome-wide sequencing has identified numerous variants that the World Health Organization has subsequently declared “variants of concern” [Mercatelli 2020]. Many of these variants are in the spike protein necessary for host recognition [Becerra-Florez 2020], but many are also scattered throughout the remainder of the viral genome [Nazario-Toole 2021]. Using data from the first year of the pandemic, we previously developed an algorithm to predict severity based upon viral mutations [Voss et al, 2021]. In addition, we reported 17 variants associated with severe clinical outcomes ($OR \geq 2$) and 67 variants associated with mild clinical outcomes ($OR \leq 0.5$). The area under the curve for our predictive algorithm was 0.91, suggesting a strong discriminative ability for classifying severe patients. Here we present results comparing the computationally predicted probability of a severe outcome to real-world laboratory measured PCR values.

Materials and methods

This study was approved as a portion of the study FWR20190037N, reviewed and approved by the Air Force Research Laboratory’s Institutional Review Board. Genome sequences were downloaded from the public access GISAID database [Elbe, 2017; Shu, 2017] restricted to sequences uploaded by the US Air Force School of Aerospace Medicine (USAFSAM) as of 16 April 2021 (see supplemental information for a complete list of accession numbers and references). Only the USAFSAM sequences were used as we also have access to PCR cycle thresholds for those data. Clinical rt-PCR diagnostic tests were performed using US

Centers for Disease Control and Prevention assay as described elsewhere [Chapleau et al, 2021] and the cycle threshold (Ct) for the N1 gene was used as the comparator.

A predictive algorithm based upon SARS-CoV-2 genomic variants, the geographic region of collection, and the age and gender of the patient has been reported to have been significantly predictive of clinical severity [Voss et al, 2021]. This algorithm was used to generate a predicted severity score from 0 to 1 in the Python environment. The genomic sequences were first downloaded and aligned to the Wuhan reference strain (NCBI: NC_045512.2; GISAID: EPI_ISL_402125) using MiniMap2 (version 2.17) [Li, 2018]. Sequences were aligned using MAFFT [Katoh 2002] and variants were called using SNP-sites [Page 2016]. Within our dataset, only 662 of the 4,502 variants used in the algorithm were observed in at least one viral genome, so the remaining variants were assigned columns of 0 within the Pandas dataframe (Figure 1).

Fig 1: Process flow chart for orthogonal validation of a severity prediction algorithm.

Sequences were downloaded from GISAID, processed through reference alignment and variant calling, predicted based upon variants identified by Voss et al, and compared to observed PCR threshold data by T-test and correlations. The matrix used for predictions was 716 samples with 662 variants overlapping the Voss et al variant list (all other variants were assigned 0 for Python predictions).

The sample IDs were divided into classes based upon severity prediction probabilities generated here for analysis. The mean observed Ct values for each class were compared using unequal variance unpaired t test. We used Welch's corrections because of unequal sample sizes. A second approach tested for correlation between the predicted probability of a viral sequence

belonging to the severe class and the observed Ct value. No modification to the original algorithm using PCR measurements was performed. Linear correlation was tested using Pearson's correlation. We used GraphPad Prism 7.0c for performing the t tests and Pearson's correlation.

Results

We obtained 1,314 SARS-CoV-2 sequences published to GISAID by the USAFSAM Epidemiology Surveillance Laboratory (collection dates between 22 February 2020 and 11 June 2021). Of those sequences, 716 had corresponding N gene PCR data and were used for analysis. Using the previously published SARS-CoV-2 genome severity prediction model [Voss et al], 636 (89%) specimens had a $\leq 50\%$ chance of being severe ("predicted mild" group) and 80 (11%) had a probability $> 50\%$ ("predicted severe" group). The "predicted severe" group had a significantly lower observed N gene Ct value (18.3 ± 0.6) as compared to the "predicted mild" group (20.4 ± 0.2). This difference of 2.1 ± 0.7 Ct (95% CI: 0.81-3.4) is consistent with differences seen in other studies signifying highly transmissible variants of concern [Roquebert 2021] and is significant based upon the unpaired t test (Figure 2, $P = 0.0017$).

Fig 2: PCR Ct values for viral specimens predicted to be from patients with mild (open box, n=636) or severe (dotted box, n=80) outcomes. $**P = 0.0017$.

A similar analysis on the upper and lower quartiles of severity prediction revealed a more substantial difference (Figure 3). Those specimens with the greatest probability of severe outcomes ($\geq 75\%$ probability) had a Ct of 16.6 ± 1.3 ($n=10$, 1.4% of samples), whereas those least likely to be associated with severe outcomes ($<25\%$) had a Ct of 21.4 ± 0.3 ($n=350$, 49% of samples) ($P = 0.0045$). The difference between the means and 95% confidence interval was 4.9 (CI: 1.9-7.8).

Fig 3: PCR Ct values for viral specimens least likely (probability $<25\%$, open box, $n=350$) or most likely (probability $\geq 75\%$, dotted box, $n=10$) outcomes. $**P = 0.0045$.

Discussion

Thoroughly validating predictions from machine learning models is necessary for establishing the credibility and power of analytics for public health. For biology and medicine, this often means finding a testable measurement that represents the predicted outputs. In our case, we used an outgroup sample with an orthogonal severity marker to provide strong evidence that the algorithm can identify strains that are biologically unique and present meaningful clinical differences.

Research on the MERS coronavirus showed that viral load, as inferred from PCR cycle threshold, is associated with risk of severe disease and death; with a 17% higher risk of death for each 1 point drop in Ct after adjusting for age and underlying illness [Feiken, et al., 2014]. In the context of SARS-CoV-2, data from Brazil also suggests cycle threshold can predict severity in a manner that is not explained by age [Faíco-Filho et al., 2020]. Indeed, the relationship between cycle threshold and COVID-19 severity has also been identified in a prior systematic review of 18 studies with a positive association in a majority [Rao et al., 2020], while also being described in a more recent narrative review [Rabaan et al., 2021]. Typically a Ct value of less than 20 is considered “highly infective” [Englemann 2021]. Here, our results are consistent with these previous observations in that those viral genomes with the greatest predicted probability of severe clinical outcomes had, on average, a 5 Ct difference. This difference corresponds to viral titers of approximately 1.1 million viable viral particles per mL for those most likely to be severe versus only 33,000 particles per mL for those least probable to be severe.

The practical implications of Ct differences are critical to consider when determining future public health responses. In a recent study of military basic trainees, Marcus and colleagues found three interesting differences for index patients: those index patients who were symptomatic had a 7 cycle lower Ct than those who were asymptomatic (20 vs. 27.2 cycles), symptomatic index patients had a 4 cycle lower Ct than symptomatic patients who did not occur as part of a cluster (20 vs. 24.4 cycles), and those who were asymptomatic had a 9 cycle difference (27.2 for asymptomatic index patients vs. 36.4 for asymptomatic non-index patients) [Marcus 2021]. Another study found a significant difference in viral loads in a cohort of military beneficiaries associated with obesity and inpatient or outpatient status [Epsi 2021]. When inpatients and outpatients were combined, the authors did not observe any significant difference

in viral load. The greatest difference observed by the authors was in severely obese outpatients, who had 77-fold higher viral loads than non-obese patients, suggesting that in mildly ill patients, those who are more obese can tolerate greater viral loads. Consequently, it becomes apparent that viral load alone is not the most powerful predictor of disease severity.

Ultimately, while it is likely that there are other factors (biological or methodological) contributing to severity, it is possible these balance out in a large enough population or where these factors are less likely to be different. For example, if there are measurement errors in the Ct associated with random differences in sample acquisition techniques, time of day a participant was tested, or sample type (nasopharyngeal vs. anterior nasal vs. nasal wash vs. oropharyngeal), these factors are likely to make the association between Ct and predicted risk appear smaller than it is. Alternatively, it remains plausible that predicted severity from the genome does not differ by time of collection or body weight category, as both the biological and methodological confounders primarily bias toward the null [Fosgate 2006]. Consequently, it is possible the difference we found is explained by factors like those mentioned above, especially considering the constrained metadata regarding the lack of patient outcomes in our dataset. Regardless, because we observed, in an independent patient sample set, a decrease in cycle thresholds (a surrogate for increased disease severity) between groups of patients predicted to experience severe clinical outcomes like hospitalization, it remains compelling that the predictions generated from the computational algorithm are reflecting biological mechanisms and are clinically meaningful for patients, providers, and public health leaders. Future work is required to take the results of these two papers from statistically significant research findings to practical tools for use in clinical practice and public health management.

In conclusion, here we report a correlation between a computationally predicted severity and a clinically measured surrogate for severity, PCR cycle threshold. Our results show that using viral genetic information combined with patient demographics could aid clinical triage and public health surveillance. While whole viral genome sequencing of every patient upon admission is cost prohibitive and results are unlikely to be timely enough to improve patient treatment, rapid diagnostic tests developed to identify variants critical to severity may provide insights to improve outcomes. Additionally, using sequence data along with laboratory validated, *in silico*-derived severity markers could help prioritize new variant vaccine development and in cases where there is clinical equipoise (for example, in cases where monoclonal antibody treatment is being considered or where post-exposure pre-symptomatic prophylaxis is being considered for known contacts, among others). Finally, public health officials employing large scale surveillance methods could benefit from these *in silico* markers by watching for the emergence of markers and gain a few days' lead time for preparedness actions.

Limitations

One limitation of this study is the use of PCR cycle threshold as a surrogate for clinical severity. As mentioned in the main text, the ability of PCR to predict clinical outcome only has an AUC of 0.68, making it a fair, but not ideal analog for clinical outcomes. Clinical outcome data were not available for this study. Another limitation related to using the PCR Ct value is the variability in sample collection per patient as some patients report to clinics early in illnesses whereas others may have reported only after symptoms became more severe. Regardless of these limitations, the high predictive ability of the algorithm along with the significant correlation with PCR data suggest that our results can be replicated in other populations and by other laboratories.

Another limitation is the relatively small number of viral genotypes. A post-hoc power analysis for the predicted severe (Ct mean = 18.33 ± 5.514 , n=80) versus predicted mild (20.44 ± 5.633 , n=636) with a 5% type I error rate revealed 89.6% power to detect the observed difference, indicating that though the sample sizes were relatively small, the study had sufficient power to draw the conclusions from the results. Similarly, the post-hoc power analysis for the most and least likely to be severe groups had a 95.3% power.

Declarations

Ethics approval and consent to participate

This study was approved by the Air Force Research Laboratory's Institutional Review Board (FWR20190037N).

Acknowledgments

The authors thank the laboratory technicians of the USAFSAM Applied Technology and Genomics Division for their support in generating the SARS-CoV-2 genomic sequences. The authors also thank the staff of the USAFSAM Public Health Epidemiology Surveillance Laboratory for their support and coordination of sample acquisition and PCR cycle threshold data.

Supplemental file

Excel file, two tabs. Tab 1: list of GISAID identifiers used; Tab 2: results of prediction (binary), probability of severity, and observed Ct values (if any).

References

1. Roquebert B, Haim-Boukoba S, Trombert-Paolantoni S, Lecorche E, Verdurme L, Foulongne V, Burrel S, Alizon S, Sofonea MT. SARS-CoV-2 variants of concern are associated with lower RT-PCR amplification cycles between January and March 2021 in France. medRxiv pre-print. <https://doi.org/10.1101/2021.03.19.21253971>
2. Voss JD, Skarzynski M, McAuley EM, Maier EJ, Gibbons T, Fries AC, Chapleau RR. Variants in SARS-CoV-2 associated with mild or severe outcome. *Evol Med Public Health*. 2021;ePub ahead of print. <https://doi.org/10.1093/emph/eoab019>
3. Chapleau RR, Christian M, Connors B, Premo C, Chao TC, Rodriguez J, Huntsberger S, Meyer J, Javorina A, Reynolds K, Riddle D, Lisanby M, Starr C. Early identification of SARS-CoV-2 emergence in the Department of Defense via retrospective analysis of 2019-2020 upper respiratory illness samples. *MSMR*. 2021;28(6):#-#.
4. Elbe S, Buckland-Merrett G; Data, disease and diplomacy: GISAID's innovative contribution to global health. *Global Challenges* 2017;1(1):33-46.
5. Shu Y, McCauley J; GISAID: Global initiative on sharing all influenza data—from vision to reality. *Eurosurveillance* 2017;22(13):30494.
6. Li H; Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* 2018;34(18):3094-3100.
7. Katoh, Misawa, Kuma and Miyata (Nucleic Acids Res. 30:3059-3066, 2002) MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform
8. "SNP-sites: rapid efficient extraction of SNPs from multi-FASTA alignments", Andrew J. Page, Ben Taylor, Aidan J. Delaney, Jorge Soares, Torsten Seemann, Jacqueline A. Keane, Simon R. Harris, *Microbial Genomics* 2(4), (2016)

9. Choudhuri J, Carter J, Nelson R, Skalina K, Osterbur-Badhey M, et al. (2020) SARS-CoV-2 PCR cycle threshold at hospital admission associated with patient mortality. PLOS ONE 15(12): e0244777. <https://doi.org/10.1371/journal.pone.0244777>
10. Pujadas E, Chaudhry F, McBride R, Richter F, Zhao S, Wajnberg A, Nadkarni G, Glicksberg BS, Houldsworth J, Cordon-Cardo C. 2020. SARS-CoV-2 viral load predicts COVID-19 mortality. The Lancet Respiratory Medicine 8(9):e70. [https://doi.org/10.1016/S2213-2600\(20\)30354-4](https://doi.org/10.1016/S2213-2600(20)30354-4)
11. O'Driscoll M, Ribeiro Dos Santos G, Wang L, Cummings DAT, Azman AS, Paireau J, Fontanet A, Cauchemez S, Salje H. Age-specific mortality and immunity patterns of SARS-CoV-2. Nature. 2021 Feb;590(7844):140-145. doi: 10.1038/s41586-020-2918-0. Epub 2020 Nov 2. PMID: 33137809.
12. Göker H, Aladağ Karakulak E, Demiroğlu H, Ayaz Ceylan ÇM, Büyükaşık Y, Inkaya AÇ, Aksu S, Sayinalp N, Haznedaroğlu IC, Uzun Ö, Akova M, Özcebe OI, Ünal S. The effects of blood group types on the risk of COVID-19 infection and its clinical outcome. Turk J Med Sci. 2020 Jun 23;50(4):679-683. doi: 10.3906/sag-2005-395. PMID: 32496734; PMCID: PMC7379446.
13. Pradhan A, Olsson PE. Sex differences in severity and mortality from COVID-19: are males more vulnerable? Biol Sex Differ 2020;11:53. doi: 10.1186/s13293-020-00330-7
14. Dong Y, Zhou H, Li M, Zhang Z, Guo W, Yu T, Gui Y, Wang Q, Zhao L, Luo S, Fan H, Hu D. A novel simple scoring model for predicting severity of patients with SARS-CoV-2 infection. Transbound Emerg Dis. 2020 Nov;67(6):2823-2829. doi: 10.1111/tbed.13651. Epub 2020 Jun 13. PMID: 32469137; PMCID: PMC7283685.

15. Gallo Marin B, Aghagoli G, Lavine K, Yang L, Siff EJ, Chiang SS, Salazar-Mather TP, Dumenco L, Savaria MC, Aung SN, Flanigan T, Michelow IC. Predictors of COVID-19 severity: A literature review. *Rev Med Virol*. 2021 Jan;31(1):1-10. doi: 10.1002/rmv.2146. Epub 2020 Jul 30. PMID: 32845042; PMCID: PMC7855377.
16. Nazario-Toole AE, Xia H, Gibbons TF. Whole-genome Sequencing of SARS-CoV-2: Using Phylogeny and Structural Modeling to Contextualize Local Viral Evolution. *Mil Med*. 2021 Feb 20;usab031. doi: 10.1093/milmed/usab031. Epub ahead of print. PMID: 33609027; PMCID: PMC7928784.
17. Mercatelli D, Giorgi FM: Geographic and genomic distribution of SARS-CoV-2 mutations. *Front Microbiol* 2020; 11: 1800. Published July 22, 2020.doi: 10.3389/fmicb.2020.01800
18. Becerra-Flores M, Cardozo T: SARS-CoV-2 viral spike G614 mutation exhibits higher case fatality rate. *Int J Clin Pract* 2020; 74(8): e13525. doi: 10.1111/ijcp.13525
19. Faíco-Filho KS, Passarelli VC, Bellei N. Is Higher Viral Load in SARS-CoV-2 Associated with Death? *Am J Trop Med Hyg*. 2020 Nov;103(5):2019-2021. doi: 10.4269/ajtmh.20-0954. PMID: 32996443; PMCID: PMC7646800.
20. Feikin DR, Alraddadi BM, Qutub M, Shabouni O, Curns A, Oboho IK, et al. Association of Higher MERS-CoV Virus Load with Severe Disease and Death, Saudi Arabia, 2014. *Emerg Infect Dis*. 2015;21(11):2029-2035.
21. Rao SN, Manissero D, Steele VR, Pareja J. A Systematic Review of the Clinical Utility of Cycle Threshold Values in the Context of COVID-19. *Infect Dis Ther*. 2020 Sep;9(3):573-586. doi: 10.1007/s40121-020-00324-3. Epub 2020 Jul 28. Erratum in: *Infect Dis Ther*. 2020 Aug 18;; PMID: 32725536; PMCID: PMC7386165.

22. Engelmann I, Alidjinou EK, Ogiez J, Pagneux Q, Miloudi S, Benhalima I, Ouafi M, Sane F, Hober D, Roussel A, Cambillau C, Devos D, Boukherroub R, Szunerits S. Preanalytical Issues and Cycle Threshold Values in SARS-CoV-2 Real-Time RT-PCR Testing: Should Test Results Include These? ACS Omega. 2021 Mar 6;6(10):6528-6536. doi: 10.1021/acsomega.1c00166. PMID: 33748564; PMCID: PMC7970463.
23. Marcus JE, Frankel DN, Pawlak MT, Casey TM, Cybulski RJ Jr, Enriquez E, Okulicz JF, Yun HC. Risk Factors Associated With COVID-19 Transmission Among US Air Force Trainees in a Congregant Setting. JAMA Netw Open. 2021 Feb 1;4(2):e210202. doi: 10.1001/jamanetworkopen.2021.0202. PMID: 33630090; PMCID: PMC7907953.
24. Epsi NJ, Richard SA, Laing ED, Fries AC, Millar E, Simons MP, English C, Colombo CJ, Colombo RE, Lindholm DA, Ganesan A, Maves RC, Huprikar N, Larson D, Mende K, Chi SW, Madar C, Lalani T, Broder CC, Tribble D, Agan BK, Burgess TH, Pollett SD; EPICC COVID-19 Cohort Study Group. Clinical, immunological and virological SARS-CoV-2 phenotypes in obese and non-obese military health system beneficiaries. J Infect Dis. 2021 Jul 31;jiab396. doi: 10.1093/infdis/jiab396. Epub ahead of print. PMID: 34331541.
25. Fosgate GT. Non-differential measurement error does not always bias diagnostic likelihood ratios towards the null. Emerg Themes Epidemiol. 2006 Jul 17;3:7. doi: 10.1186/1742-7622-3-7. PMID: 16846512; PMCID: PMC1550225.
26. Zhou B, Thao TTN, Hoffmann D, Taddeo A, Ebert N, Labroussaa F, Pohlmann A, King J, Portmann J, Halwe NJ, Ulrich L, Trüeb BS, Kelly JN, Fan X, Hoffmann B, Steiner S, Wang L, Thomann L, Lin X, Stalder H, Pozzi B, de Brot S, Jiang N, Cui D, Hossain J, Wilson M, Keller M, Stark TJ, Barnes JR, Dijkman R, Jores J, Benarafa C, Wentworth DE, Thiel V, Beer M. SARS-CoV-2 spike D614G variant confers enhanced replication and

transmissibility. bioRxiv [Preprint]. 2020 Oct 27:2020.10.27.357558. doi:

10.1101/2020.10.27.357558. PMID: 33140052; PMCID: PMC7605563.

27. COVID-19 Host Genetics Initiative. The COVID-19 Host Genetics Initiative, a global initiative to elucidate the role of host genetic factors in susceptibility and severity of the

SARS-CoV-2 virus pandemic. Eur J Hum Genet. 2020 Jun;28(6):715-718. doi:

10.1038/s41431-020-0636-6. Epub 2020 May 13. PMID: 32404885; PMCID: PMC7220587.