Toh and Brody *Human Genomics* (2020) 14:36
https://doi.org/10.1186/s40246-020-00288-y

Human Genomics

PRIMARY RESEARCH Open Access

Evaluation of a genetic risk score for severity of COVID-19 using human chromosomal-scale length variation

Christopher Toh and James P. Brody*

Abstract

Introduction: The course of COVID-19 varies from asymptomatic to severe in patients. The basis for this range in symptoms is unknown. One possibility is that genetic variation is partly responsible for the highly variable response. We evaluated how well a genetic risk score based on chromosomal-scale length variation and machine learning classification algorithms could predict severity of response to SARS-CoV-2 infection.

Methods: We compared 981 patients from the UK Biobank dataset who had a severe reaction to SARS-CoV-2 infection before 27 April 2020 to a similar number of age-matched patients drawn for the general UK Biobank population. For each patient, we built a profile of 88 numbers characterizing the chromosomal-scale length variability of their germ line DNA. Each number represented one quarter of the 22 autosomes. We used the machine learning algorithm XGBoost to build a classifier that could predict whether a person would have a severe reaction to COVID-19 based only on their 88-number classification.

Results: We found that the XGBoost classifier could differentiate between the two classes at a significant level ($p = 2 \cdot 10^{-11}$) as measured against a randomized control and ($p = 3 \cdot 10^{-14}$) as measured against the expected value of a random guessing algorithm (AUC = 0.5). However, we found that the AUC of the classifier was only 0.51, too low for a clinically useful test.

Conclusion: Genetics play a role in the severity of COVID-19, but we cannot yet develop a useful genetic test to predict severity.

Keywords: COVID-19, Genetic risk score, UK biobank, Machine learning

Introduction

The course of COVID-19 varies from asymptomatic to severe (acute respiratory distress, cytokine storms, and death) in patients. The basis for this range in symptoms is unknown. One possibility is that genetic variation is partly responsible for the highly variable response to infection.

Human genetic variation can affect susceptibility and resistance to viral infections [1]. For instance, variants in the gene IFITM3 affect the severity of seasonal influenza [2]. Patients hospitalized from seasonal influenza had a particular allele of the gene IFITM3 at a higher rate than expected from the general population. Laboratory work determined that this particular allele can alter the course of the influenza virus infection.

We have previously shown that chromosomal-scale length variation is a powerful tool to analyze genome-wide associations [3]. This method is particularly appealing for genetic risk scores because it includes epistatic effects that might be missed with conventional genome-wide association studies. Others have used machine

* Correspondence: jpbrody@uci.edu
Department of Biomedical Engineering, University of California, Irvine, USA

learning in combination with copy number variation to predict cancer risk [4].

The purpose of this paper is to evaluate how well a genetic risk score based on chromosomal-scale length variation and machine learning classification algorithms can predict severity of response to SARS-CoV-2 infection. We evaluated this approach on a dataset of 931 patients who had a severe reaction to COVID-19 in the early part of the 2020 global pandemic. These patients had been previously genotyped as part of the UK Biobank.

Methods

Data was obtained from the UK Biobank under Application Number 47850. First, we downloaded the "l2r" files from the UK Biobank. Each chromosome has a separate "l2r" file. Each "l2r" file contained 488,377 columns and a variable number of rows. Each column represented a unique patient in the dataset, who is only identified by an encoded identification number. Each row represented a measurement at a different location in the genome. The values in the file represent the log (base 2) of the ratio of measured intensity measured in a microarray relative to the expected two copies at that location in the genome.

After downloading the "l2r" data from the UK Biobank, we computed the mean l2r value for a portion, we chose 25%, of the chromosome for each patient in the dataset. This process produced a dataset where each person was represented by a series of 88 numbers. Each number represents the length variation for 25% of the 22 non-sex chromosomes. A value of 0 ($\log_2$ ration) represents the nominal average length of that portion of the particular chromosome. We call this dataset the chromosomal-scale length variation (CSLV) dataset.

This CSLV dataset was matched with the UK Biobank COVID-19 dataset. The COVID-19 data were provided to UK Biobank by Public Health England. UK Biobank matched the person in the Public Health England data with UK Biobank's internal records to produce the person's encoded participant identification number. The dataset we have provided by UK Biobank contains the participant ID, date the specimen was taken, laboratory that processed the sample, whether the patient was an inpatient when the sample was taken, and the result (positive/negative) of the test. The UK Biobank continues to update the data approximately biweekly.

The criteria for testing and interpretation of results in the UK Biobank COVID-19 data has evolved. A positive test in this dataset earlier than 27 April 2020 was a good indication that the person had severe disease. During this initial period of the pandemic, SARS-CoV-2 testing was only performed on symptomatic people and this particular dataset only includes people tested in a hospital. After 27 April 2020, NHS instructed hospitals to test all non-elective patients admitted, including asymptomatic patients. The UK Biobank dataset released after 27 May 2020 includes "pillar 2" positive test results. These "pillar 2" tests include people in hospitals for non-elective procedures and staff screening. These results can include asymptomatic patients.

Using the CSLV-COVID-19 dataset, we selected all people who tested positive before 27 April 2020 and labeled these as people having a severe reaction to COVID-19. We segmented these into three overlapping datasets, as shown in Table 1. We constructed an age-matched control group of the same size that had an identical age profile as those in the severe reaction group. The age-matched control group was selected from the entire UK Biobank dataset, excepting those few who had a severe reaction to COVID-19. Since only a small fraction of the people in the UK Biobank had a severe reaction to COVID-19, we could rerun the analysis with a different age-matched control group many times to build up statistics. We chose this method of selecting the control group based on the finding that severe reactions to COVID-19 are both a strong function of age and uncommon (only about 20% of those infected with SARS-CoV-2 require ICU admission even among those in their 70s) [5, 6].

We used the H2O machine learning package in R to create XGBoos t[7] models that were trained to classify a person in the dataset, consisting of those who had a severe reaction and age-matched controls, based solely on their chromosomal-scale length variation data.

Results

The results are presented in Fig. 1 and Table 2. As Fig. 1 shows, we found a significant difference between all three age groupings and their corresponding random controls. This finding indicates that germ line genetics of the infected patient, as represented by the set of chromosomal-scale length variation numbers, is correlated with the severity of COVID-19.

Fig. 1 and Table 3 also show that the AUC (area under the curve of the receiver operating characteristic curve) for the XGBoost classification model was about 0.51, but

Table 1 We segmented the dataset into three overlapping subsets. The first, which we called "1930" contained all UK Biobank participants born after 1930 who had a severe reaction to SARS-CoV-2 infection before 27 April 2020. The two subsets contained people born after 1940 and after 1950

Dataset	Number
1930 (< 90 years of age)	981
1940 (< 80 years of age)	880
1950 (< 70 years of age)	468

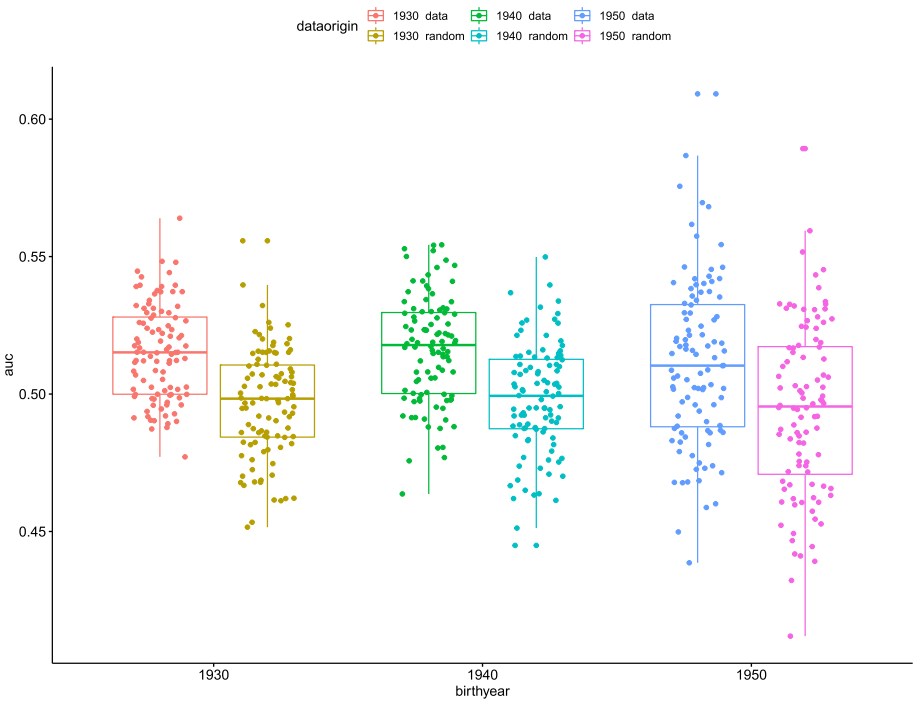

Fig. 1 This boxplot figure presents the results of the machine learning predictions. We created three different datasets, one which includes all patients less than 90 years old, the second includes every patient less than 80 years old, and the third with every patient less than 70 years old. These are indicated as the oldest birthyear "data." Each dataset included an equal number of patients with a "severe reaction" to COVID-19 and an equal number of age-matched people drawn from the general UK Biobank population, "normal." For comparison, we took those three datasets and randomly permuted the status ("severe reaction" or "normal") and repeated the process. This randomly permuted dataset is labeled oldest birthyear "random." For each dataset, we repeated the whole process 100 times, each time with a different set of age-matched people from the general UK Biobank population

Table 2 We compared the difference in mean AUC values between the various datasets using a *t* test. The datasets consisting of people born after 1930, 1940, and 1950 all showed significant differences with the corresponding random control. Those three datasets also showed significant differences between the mean AUC and 0.5. The three random controls did not show a significant difference between the mean AUC and 0.5, as expected. An AUC value of 0.5 represents a random classification test, one in which the algorithm is no better than guessing

		***p* value of *t* test**
1930 data	1930 random	$2 \cdot 10^{-11}$
1940 data	1940 random	$1 \cdot 10^{-9}$
1950 data	1950 random	$1 \cdot 10^{-4}$
0.5	1930 data	$3 \cdot 10^{-14}$
0.5	1940 data	$4 \cdot 10^{-13}$
0.5	1950 data	$3 \cdot 10^{-4}$
0.5	1930 random	0.1
0.5	1940 random	0.4
0.5	1950 random	0.08

still significantly greater than 0.50. A classification model with an AUC of 0.51 is just slightly better than guessing.

Discussion

The two conclusions of this study are divergent. First, a genetic difference exists between those who have the most severe course of COVID-19 and the general population. Second, we were not able to exploit this difference to develop a clinically useful test to distinguish between people who will experience a severe course of the disease and those who will not. We could only

Table 3 The mean and standard deviation of the area under the curve of the receiver operating characteristic curve was recorded after each of the 100 different XGBoost classification models. Each run used a different set of people who did not have a severe reaction to COVID-19. The mean AUC for all three datasets was well described by a normal distribution, as confirmed by a Shapiro normality test

	Mean AUC	**SD AUC**
1930 data	0.515	0.017
1940 data	0.516	0.019
1950 data	0.511	0.030

demonstrate a genetic risk test with an AUC of 0.51, just slightly above 0.50 which represents random guessing.

Although the AUC we found here is too low to be clinically useful, several avenues for improving the AUC exist. We were constrained by the data available to compare those who had a severe reaction to COVID-19 with the general population, but the general population probably contains a substantial number of people who would also have a severe reaction to COVID-19. A better approach would be to compare those who had a severe reaction to COVID-19 with those who were asymptomatic or had a mild reaction. Simply having a much larger number of patients who had a severe reaction might also lead to an increase in AUC.

Changes in our feature selection and classification algorithm might also improve the AUC. Our feature selection algorithm that transformed "l2r" data into our final chromosomal-scale length variation data took averages over each quarter of a chromosome. We could instead include smaller chromosome segments. Generally, we need the number of features to be much less than the number of observations (patients). So, an increase in the number of observations would allow an increase in the number of features. Also, an alternative machine learning algorithm might improve the AUC. Different algorithms perform differently on different classes of problems and XGBoost generally performs well on tabular data [8]. We did a brief test of different algorithms before choosing XGBoost as the best solution for this problem. But, for instance, a deep learning algorithm might have better performance with proper tuning.

Our results add to the recent work done by others on the link between genetics and severity of COVID-19. For instance, one study from the Netherlands identified four young men from two different families who had severe symptoms of COVID-19 and no preexisting medical conditions. Detailed genetic studies revealed that these four men all had a rare loss of function variant of TLR7, which lies on the X-chromosome [9].

A detailed study of this UK Biobank COVID-19 dataset found that Black and Asian patients were at a significantly higher risk of testing positive compared to white patients [10]. This study also attempted to derive a polygenic risk score. However, when they applied the polygenic risk score to a hold-out group, they found that the mean score was indistinguishable between the group of people who had tested positive and the group that had no positive test. In comparison, our work found that these two groups are distinguishable with a genetic risk score, but only very slightly. We measured the AUC at 0.51. They [10] do not report an AUC, but an indistinguishable test is the equivalent of an AUC of 0.50.

Other more comprehensive metastudies have identified one specific genetic component behind the severity of COVID-19. For instance, one study of COVID-19 patients who experienced respiratory failure at seven hospitals in Italy and Spain found a fairly strong association in a cluster of genes lying on part of chromosome 3 and a borderline association in chromosome 9 encompassing the ABO blood group locus [11]. The "ANA_B2" June 2020 results posted by the COVID-19 Host Genetics Initiative [12, 13] also indicate a strong association in chromosome 3 but fail to reproduce the association in chromosome 9. The COVID-19 Host Genetics Initiative "ANA_B2" study compares hospitalized COVID-19 patients to the general population and are mostly derived from patients in Europe and Brazil. Neither study attempted to derive a genetic risk score.

This study has several weaknesses. First, we cannot attribute the severity of COVID-19 to particular genetic variants. This study only finds correlations and does not establish a cause and effect. Second, while it is possible that these correlations relate to underlying biology, it is also possible that the correlations are related to ancestral differences that translate to socio-economic differences. COVID-19 severity is known to be correlated with racial/ethnic background [14, 15]. The small effect that we measured might be simply due to the larger complex effect of racial/ethnic disparities in COVID-19 severity.

Conclusion

In conclusion, we found a significant difference exists between the structural genomics of those patients in the UK Biobank who had a severe reaction to the SARS-CoV-2 virus and the general UK Biobank population. However, a test based upon this difference would not be clinically useful in its present state since it had an AUC of 0.51.

Acknowledgements

The data used in this study was obtained from the UK Biobank under Application Number 47850.

Authors' contributions

CT and JB analyzed the UK Biobank data. CT and JB contributed to the manuscript. All authors read and approved the final manuscript.

Funding

No external funding supported this research.

Availability of data and materials

The datasets analyzed during the current study are available from UK Biobank at https://www.ukbiobank.ac.uk/

Ethics approval and consent to participate

Ethical approval and participant consent was collected by UK Biobank at the time participants enrolled. This paper is an analysis of anonymized data provided by UK Biobank. According to UC Irvine's IRB, analysis of anonymized data does not constitute Human Subjects Research.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Received: 17 July 2020 Accepted: 30 September 2020
Published online: 09 October 2020

References

1. Kenney AD, Dowdle JA, Bozzacco L, McMichael TM, St Gelais C, Panfil AR, et al. Human genetic determinants of viral diseases. Annual review of genetics [Internet]. Annu Rev Genet; 2017 [cited 2020 Jun 15];51:241–63. Available from: http://www.ncbi.nlm.nih.gov/pubmed/28853921.
2. Everitt AR, Clare S, Pertel T, John SP, Wash RS, Smith SE, et al. IFITM3 restricts the morbidity and mortality associated with influenza. Nature [Internet]. Nature; 2012 [cited 2020 Jun 15];484:519–23. Available from: http://www.ncbi.nlm.nih.gov/pubmed/22446628.
3. Toh C, Brody JP. Analysis of copy number variation from germline DNA can predict individual cancer risk. bioRxiv [Internet]. Cold Spring Harbor Laboratory; 2018 [cited 2018 Jun 3];303339. Available from: https://www.biorxiv.org/content/early/2018/04/17/303339.
4. Ding X, Tsang S-Y, Ng S-K, Xue H. Application of machine learning to development of copy number variation-based prediction of cancer risk. Genomics Insights [Internet]. SAGE PublicationsSage UK: London, England; 2014 [cited 2020 Sep 14];7:GEI.S15002. Available from: http://journals.sagepub.com/doi/10.4137/GEI.S15002.
5. Davies NG, Klepac P, Liu Y, Prem K, Jit M, Eggo RM. Age-dependent effects in the transmission and control of COVID-19 epidemics. Nature Medicine [Internet]. Nature Publishing Group; 2020 [cited 2020 22];1–7. Available from: http://www.nature.com/articles/s41591-020-0962-9.
6. Bialek S, Boundy E, Bowen V, Chow N, Cohn A, Dowling N, et al. Severe outcomes among patients with coronavirus disease 2019 (COVID-19) — United States, February 12–March 16, 2020. MMWR Morbidity and Mortality Weekly Report [Internet]. 2020 [cited 2020 22];69:343–6. Available from: http://www.cdc.gov/mmwr/volumes/69/wr/mm6912e2.htm?s_cid=mm6912e2_w.
7. Chen T, Guestrin C. XGBoost. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '16 [Internet]. New York, New York, USA: ACM Press; 2016 [cited 2018 Jan 28]. p. 785–94. Available from: http://dl.acm.org/citation.cfm?doid = 2939672.2939785.
8. Olson RS, Cava W la, Mustahsan Z, Varik A, Moore JH. Data-driven advice for applying machine learning to bioinformatics problems. Pacific Symposium on Biocomputing Pacific Symposium on Biocomputing [Internet]. NIH Public Access; 2018 [cited 2020 Jun 16];23:192–203. Available from: http://www.ncbi.nlm.nih.gov/pubmed/29218881.
9. van der Made CI, Simons A, Schuurs-Hoeijmakers J, van den Heuvel G, Mantere T, Kersten S, et al. Presence of genetic variants among young men with severe COVID-19. JAMA [Internet]. American Medical Association; 2020 [cited 2020 18];324:663. Available from: https://jamanetwork.com/journals/jama/fullarticle/2768926.
10. Kolin DA, Kulm S, Elemento O. Clinical and genetic characteristics of Covid-19 patients from UK Biobank. medRxiv [Internet]. Cold Spring Harbor Laboratory Press; 2020 [cited 2020 Jun 20];2020.05.05.20075507. Available from: https://www.medrxiv.org/content/10.1101/2020.05.05.20075507v1.
11. Ellinghaus D, Degenhardt F, Bujanda L, Buti M, Albillos A, Invernizzi P, et al. Genomewide association study of severe Covid-19 with respiratory failure. New England Journal of Medicine [Internet]. Massachusetts Medical Society; 2020 [cited 2020 20];NEJMoa2020283. Available from: http://www.nejm.org/doi/10.1056/NEJMoa2020283.
12. 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. European Journal of Human Genetics [Internet]. Nature Publishing Group; 2020 [cited 2020 20];28:715–8. Available from: http://www.nature.com/articles/s41431-020-0636-6.
13. Covid-19 Host Genetics Initiative Results [Internet]. [cited 2020 Jun 29]. Available from: https://www.covid19hg.org/results/.
14. Webb Hooper M, Nápoles AM, Pérez-Stable EJ. COVID-19 and Racial/Ethnic Disparities. JAMA [Internet]. American Medical Association; 2020 [cited 2020 14];323:2466. Available from: https://jamanetwork.com/journals/jama/fullarticle/2766098.
15. Garg S, Kim L, Whitaker M, O'Halloran A, Cummings C, Holstein R, et al. Hospitalization rates and characteristics of patients hospitalized with laboratory-confirmed coronavirus disease 2019 — COVID-NET, 14 States, March 1–30, 2020. MMWR Morbidity and Mortality Weekly Report [Internet]. 2020 [cited 2020 14];69:458–64. Available from: http://www.cdc.gov/mmwr/volumes/69/wr/mm6915e3.htm?s_cid=mm6915e3_w.

Publisher's Note