Genetics in Medicine

Analysis of associations between polygenic risk score and COVID-19 severity in a Russian population using low-pass genome sequencing

--Manuscript Draft--

Manuscript Number:	
Full Title:	Analysis of associations between polygenic risk score and COVID-19 severity in a Russian population using low-pass genome sequencing
Article Type:	Article
Section/Category:	Public Health, Epidemiology and Personalized Medicine
Corresponding Author:	Arina Nostaeva St Petersburg University Sestroretsk, RUSSIAN FEDERATION
Corresponding Author's Institution:	St Petersburg University
Order of Authors:	Arina Nostaeva
	Valentin Shimansky
	Svetlana Apalko
	Ivan Kuznetsov
	Natalya Sushentseva
	Oleg Popov
	Anna Anisenkova
	Sergey Mosenko
	Lennart Karssen
	Yurii Aulchenko
	Sergey Shcherbak
Manuscript Region of Origin:	RUSSIAN FEDERATION
Abstract:	The course of COVID-19 is characterized by wide variability, with genetics playing a contributing role. Through large-scale genetic association studies, a significant link between genetic variants and disease severity was established. However, individual genetic variants identified thus far have shown modest effects, indicating a polygenic nature of this trait. To address this, a polygenic risk score (PRS) can be employed to aggregate the effects of multiple single nucleotide polymorphisms (SNPs) into a single number, allowing practical application to individuals within a population. In this work, we investigated the performance of a PRS model in the context of COVID-19 severity in 1,085 Russian participants using low-coverage NGS sequencing. By developing a genome-wide PRS model based on summary statistics from the COVID-19 Host Genetics Initiative consortium, we demonstrated that the PRS, which incorporates information from over a million common genetic variants, can effectively identify individuals at significantly higher risk for severe COVID-19. The findings revealed that individuals in the top 10% of the PRS distribution had a markedly elevated risk of severe COVID-19, with an odds ratio (OR) of 2.1 (95% confidence interval (CI): 1.4–3.2, p-value = 0.00046). Furthermore, incorporating the PRS into the prediction model significantly improved its accuracy compared to a model that solely relied on demographic information (p-value < 0.0001). This study highlights the potential of PRS as a valuable tool for identifying individuals at increased risk of severe COVID-19 based on their genetic profile.

Powered by Editorial Manager® and ProduXion Manager® from Aries Systems Corporation

St. Petersburg State Budgetary Healthcare Institution “City Hospital No. 40 of Kurortny District”
Sestroretsk
Russia
197706

December 13, 2023

Dear Dr. Editor,

We wish to submit an original research article entitled **Analysis of associations between polygenic risk score and COVID-19 severity in a Russian population using low-pass genome sequencing** for consideration by *Genetics in Medicine* journal. We confirm that this work is original and has not been published in another journal, nor is it currently under consideration for publication elsewhere.

In this study, we show the significant association between polygenic risk scores (PRS) and COVID-19 severity in a Russian population. We believe that the results may be of interest to the scientific community, firstly, from the perspective of PRS. It is known that a major aspect of scepticism about PRS is the under-representation of genetic data in people of non-European descent. This shift towards the European population makes the application of PRS for public health purposes less feasible in other populations, such as Asian or African populations. Previously published work has shown that using PRS as a predictor in a predictive model for non-European populations requires at least a calibration sample. Therefore, investigating the applicability of PRS among people of non-European ancestry is important. Secondly, despite the fact that the acute phase of the pandemic is behind us, COVID-19 still continues to sweep across the planet. At the same time, there are corresponding results demonstrating the diversity of the course of the disease. Earlier prediction of severe disease can greatly influence the treatment trajectory and reduce the risk of death. Our work covers all of the above aspects. Moreover, they are accompanied by significant results that can contribute to both PRS research and the field of COVID-19 research.

We have no conflicts of interest to disclose.

Please address all correspondence concerning this manuscript to me at avnostaeva@gmail.com.

Thank you for your consideration of this manuscript.

Sincerely,
Arina V. Nostaeva, MD
St. Petersburg State Budgetary Healthcare Institution “City Hospital No. 40 of Kurortny District”
avnostaeva@gmail.com

Analysis of associations between polygenic risk score and COVID-19 severity in a Russian population using low-pass genome sequencing

Arina V. Nostaeva[1], Valentin S. Shimansky[1,2], Svetlana V. Apalko[1,2], Ivan A. Kuznetsov[3], Natalya N. Sushentseva[1], Oleg S. Popov[1,2], Anna Y. Anisenkova[1,2], Sergey V. Mosenko[1,2], Lennart C. Karssen[4], Yurii S. Aulchenko[4,5], Sergey G. Shcherbak[1,2]

[1] St. Petersburg State Budgetary Healthcare Institution “City Hospital No. 40 of Kurortny District”, Sestroretsk, Russia
[2] St. Petersburg State University, St. Petersburg, Russia
[3] Skolkovo Institute of Science and Technology, Moscow, Russia
[4] PolyKnomics BV, The Netherlands
[5] Institute of Cytology and Genetics, Siberian Branch of Russian Academy of Sciences, Novosibirsk, Russia;

Please address correspondence to:
Arina V. Nostaeva, MD
Sestroretsk, 9 Borisova str., Saint Petersburg, Russia, 197706
avnostaeva@gmail.com

CONFLICT OF INTEREST

YSA is a co-owner of PolyKnomics BV, a private organization providing services, research, and development in the field of computational and statistical genomics. YSA is currently a full-time employee of GSK. The other authors declare that they have no competing interests.

Analysis of associations between polygenic risk score and COVID-19 severity in a Russian population using low-pass genome sequencing

Arina V. Nostaeva[1], Valentin S. Shimansky[1,2], Svetlana V. Apalko[1,2], Ivan A. Kuznetsov[3], Natalya N. Sushentseva[1], Oleg S. Popov[1,2], Anna Y. Anisenkova[1,2], Sergey V. Mosenko[1,2], Lennart C. Karssen[4], Yurii S. Aulchenko[4,5], Sergey G. Shcherbak[1,2]

[1] St. Petersburg State Budgetary Healthcare Institution "City Hospital No. 40 of Kurortny District", Sestroretsk, Russia
[2] St. Petersburg State University, St. Petersburg, Russia
[3] Skolkovo Institute of Science and Technology, Moscow, Russia
[4] PolyKnomics BV, The Netherlands
[5] Institute of Cytology and Genetics, Siberian Branch of Russian Academy of Sciences, Novosibirsk, Russia;

Please address correspondence to:
Arina V. Nostaeva, MD
Sestroretsk, 9 Borisova str., Saint Petersburg, Russia, 197706
avnostaeva@gmail.com

ABSTRACT

The course of COVID-19 is characterized by wide variability, with genetics playing a contributing role. Through large-scale genetic association studies, a significant link between genetic variants and disease severity was established. However, individual genetic variants identified thus far have shown modest effects, indicating a polygenic nature of this trait. To address this, a polygenic risk score (PRS) can be employed to aggregate the effects of multiple single nucleotide polymorphisms (SNPs) into a single number, allowing practical application to individuals within a population. In this work, we investigated the performance of a PRS model in the context of COVID-19 severity in 1,085 Russian participants using low-coverage NGS sequencing. By developing a genome-wide PRS model based on summary statistics from the COVID-19 Host Genetics Initiative consortium, we demonstrated that the PRS, which incorporates information from over a million common genetic variants, can effectively identify individuals at significantly higher risk for severe COVID-19. The findings revealed that individuals in the top 10% of the PRS distribution had a markedly elevated risk of severe COVID-19, with an odds ratio (OR) of 2.1 (95% confidence interval (CI): 1.4–3.2, p-value = 0.00046). Furthermore, incorporating the PRS into the prediction model significantly improved its accuracy compared to a model that solely relied on demographic information (p-value < 0.0001). This study highlights the potential of PRS as a valuable tool for identifying individuals at increased risk of severe COVID-19 based on their genetic profile.

INTRODUCTION

COVID-19, also known as coronavirus infection, is a contagious illness caused by the severe acute respiratory syndrome-coronavirus-2 (SARS-CoV-2). The majority of individuals who contract the virus exhibit mild to moderate respiratory symptoms and can recover without

requiring specific medical treatment. However, in certain cases, the disease can manifest in a severe form, requiring medical intervention [1,2].

Apart from external factors like virus characteristics and the effectiveness of public health, certain host-related factors such as older age, male gender, and pre-existing chronic diseases like hypertension and diabetes have been associated with susceptibility and severity of COVID-19 [3,4]. However, these risk factors alone cannot fully explain the wide variation observed in the disease severity. The course of COVID-19 can range from asymptomatic cases to acute respiratory distress and even death [5,6]. Early in the pandemic, it was noted that clinical factors alone were insufficient to account for the variability in disease severity across individuals, as severe cases were observed in young people without apparent predisposing factors, often within families [7]. This suggests that human genetics may play a role in the development of the disease.

To gain insights into the aetiology of COVID-19, large-scale genetic association studies incorporating both rare and common genetic variants have employed various study designs. These investigations, along with subsequent follow-up studies, have expanded our understanding of the disease and provided potential avenues for its treatment. The COVID-19 Host Genetics Initiative (HGI) was established to identify genetic loci that impact the severity and susceptibility of COVID-19 [8]. This global effort aims to conduct a meta-analysis of multiple COVID-19 genome-wide association studies (GWAS), and to identify significant single nucleotide polymorphisms (SNPs) associated with infection, hospitalization, and mortality. Through comparisons of genomes of millions of COVID-19 patients and healthy individuals, these studies have implicated genetic variants in 13 loci associated with the severity of the disease [9]. The COVID-19-associated genetic variants could be related to the regulation of processes such as innate antiviral defence signalling, regulation of inflammatory organ damage, and upregulation of cell receptors [10]. Modulation of these pathways can impact susceptibility to infection and subsequent disease manifestation [11].

The effects of individual genetic variants identified so far are generally small, consistent with the polygenic architecture of this trait. An individual who tests negative for a specific risk variant may still have a high genetic risk due to other unmeasured genetic factors. While each single variant only explains a small portion of the risk for severe COVID-19, combining multiple genetic variants into a polygenic risk score (PRS) can offer a better prediction of the risk. PRS allows for the aggregation of the effects of multiple SNPs into a single score, which can be practically applied to individuals within a population [12]. Conventionally, a polygenic score is defined as a weighted linear combination of allele counts for SNPs observed in an individual's genome. The PRS model consists of the weights of a set of SNPs, with the weights proportional to the estimated effects of the SNPs on the trait being studied [13].

Modern polygenic risk score models for human traits are typically estimated using summary statistics obtained from a genome-wide association meta-analysis (GWAMA) and a reference panel reflecting linkage disequilibrium (LD) in the population [13,14]. Over the past decade, PRS predictive performance has significantly improved due to larger GWAS sample sizes and advancements in methods for variable selection and effect estimation [15–24]. Polygenic scores can be utilized to rank individuals within a group based on their genetic predisposition to a disease [25–27]. This approach considers an individual's genetic

predisposition relative to the genetic predisposition of others in the same group, often expressed as a percentile representing where the individual's PRS falls within the overall distribution of the group's PRS.

Several studies have explored the development and application of PRS using variants associated with COVID-19, revealing clear associations between PRS and the risk of severe disease. However, most PRS models have been applied to cohorts consisting predominantly of individuals of Western European ancestry [28–31]. Using 1,582 SARS-CoV-2 positive participants from the UK Biobank (1,018 with severe COVID-19 and 564 without severe COVID-19) and 64 SNPs for PRS calculation, Dite et al. developed and validated a clinical and genetic model for predicting the risk of severe COVID-19. Only 13% of participants from this study were non-white, and PRS alone had an area under the receiver operating characteristic curve (AUC) of 68% [31].

While one recent study included African and South Asian groups, the associations with COVID-19 outcomes were limited by applying a PRS based on only six SNPs [32]. Another study that considered non-Western European populations was constrained by its focus on a specific Russian cohort (athletes) and also included only six genetic polymorphisms in the PRS assessment [33]. The multi-ethnic approach implemented in a very recent paper using UK biobank data, allowed the applicability of PRS, based on 17 SNPs, to diverse populations, with the severity model performing well within Black and Asian cohorts [34,35]. Overall, results highlight the potential of PRS as a predictive marker for disease severity and provide further support for its application in risk stratification and personalized healthcare approaches in the context of COVID-19.

Our study aimed to investigate the performance of the PRS model in the Russian population. The genomes of study participants (336 individuals with severe COVID-19 and 704 with moderate or without disease) were assessed using low-coverage (with mean depth x3) sequencing. Next, we developed a genome-wide PRS model for COVID-19 severity using the summary statistics from the COVID-19 host genetics initiative consortium. We demonstrated that PRS, incorporating information from more than a million common genetic variants, for COVID-19 severity can identify individuals with markedly elevated risk of severe COVID-19 course: OR = 2.1 (95% confidence interval (CI): 1.4–3.2, p-value = 0.00046) for individuals in the top 10% of the PRS distribution, and produces a significant improvement in the quality of prediction (p-value < 0.0001) compared to a model including only demographic information.

RESULTS

Participant Characteristics

The participants of the study were the patients of the infectious disease department of the St. Petersburg State Health Care Institution "City Hospital No. 40, Kurortny District" who were admitted for treatment with coronavirus infection (confirmed by polymerase chain reaction), and healthy individuals. Healthy individuals are defined as people who did not require COVID-19 medical treatment at the time of the study (between April 2020 and March 2022).

Table 1 shows the participants' characteristics. Of the 1040 participants, 458 (44%) were female, with a mean age of 60 years, while for 582 (56%) males the mean age was equal to 56 years. Overall, 859 (82%) of all participants had COVID-19, of which 336 (39% from participants with COVID-19) had severe COVID-19. Separation according to the severity of the disease was carried out according to the following criteria: the case group included 336 patients (208 men and 128 women, 63±15 years) with lung damage more than 50% (computed tomography (CT)-3 and CT-4), the control group included 704 patients (374 men and 330 women, 56±16 years), with lung damage less than 50% or without COVID-19.

Table 1. The demographic and clinical characteristics of the participants.

Characteristics	Male	Female
Mean age (sd)	56 (15)	60 (16)
Healthy individuals	110	71
Patients required treatment by severity	CT-1: 77 CT-2: 187 CT-3: 207 CT-4: 1	CT-1: 67 CT-2: 192 CT-3: 125 CT-4: 3
Outcome	death: 90 recovery or no disease: 492	death: 50 recovery or no disease: 408

Abbreviations: *CT* computed tomography, where CT-1 – mild form of pneumonia with areas of "frosted glass", the severity of pathological changes less than 25%; CT-2 – moderate pneumonia, 25–50% of lungs are affected; CT-3 – moderately severe pneumonia, 50–75% of lungs are affected; CT-4 – severe form of pneumonia, >75% of lungs are affected.

Low coverage sequencing and imputation

For all samples low coverage sequencing, also called LP-WGS (low-pass whole genome sequencing), was performed with a depth of 3x genome coverage. LP-WGS is the type of WGS with genome coverage from 0.5x to 5x [36,37]. Due to low-coverage data often having poor genotype quality and resulting in high missing genotype rates, the genotype likelihoods (GL) need to be updated using a reference panel for more accurate genotype imputation [38,39]. We used a recent method called GLIMPSE, which performs haplotype phasing and genotype imputation for LP-WGS data through a Gibbs sampling procedure, leading to improved accuracy [38]. As a reference panel, we used the 1000 Genomes data [40]. To evaluate the efficiency of LP-WGS within PRS, we calculated PRS values for a sample (not included in the study population) sequenced 41 times (in each of the batches to control the quality of the sequencing process). The coefficient of variation (CV) for PRS values was equal to 0.5% demonstrating a good method performance.

Overview of the approach

Computation of PRSs requires both genotype data of target individuals and the PRS model. To build the PRS model we used summary statistics from the COVID-19 Host Genetics Initiative consortium (release 7) [8]. These results were obtained by the meta-analysis, which

combined the results of 60 individual studies from 25 countries, with a total of 18,000 severe cases of COVID-19 and more than a million controls who either did not have a severe disease course or were not affected by COVID-19 during the study period. From the obtained summary statistics, we generated the PRS model using the Bayesian approach SBayesR with default parameters, implemented in the GCTB software [19,41,42]. Finally, we calculated individual PRS values using the PRS model (Fig. 1, Methods).

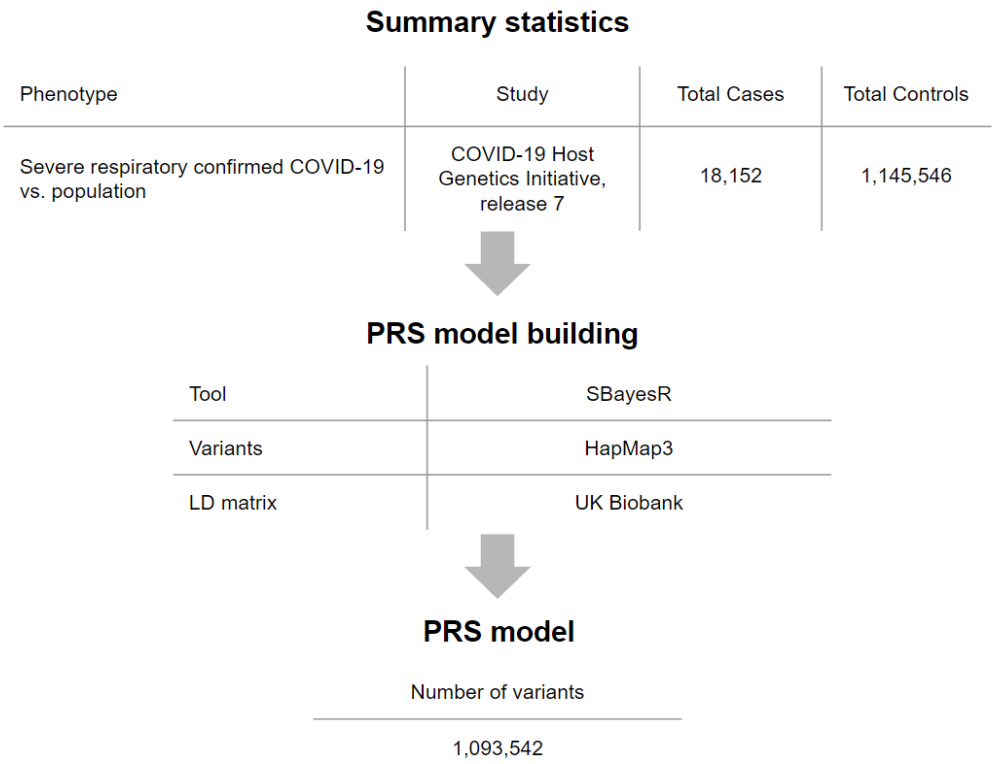

Figure 1. Study design and workflow. The PRS model for COVID-19 severity was derived by combining summary association statistics from the COVID-19 Host Genetics Initiative consortium and a linkage disequilibrium reference panel of 50,000 individuals of European ancestry from the UK Biobank data set. As a computational algorithm, SBayesR was used, which is a Bayesian approach to calculate a posterior mean effect for all variants based on a prior (effect size in the previous GWAS) and subsequent shrinkage based on linkage disequilibrium. The PRS model was restricted by the list of variants from HapMap3 and included about one million variants.

Testing associations between PRS and severe COVID-19

We compared the distributions of PRS values between severe cases and the control group combining the milder forms of COVID-19 and healthy individuals (Fig. 2). Comparison of the mean PRS values, performed using Student's t-test for two independent samples, showed significant difference (p-value = 1.8e-07).

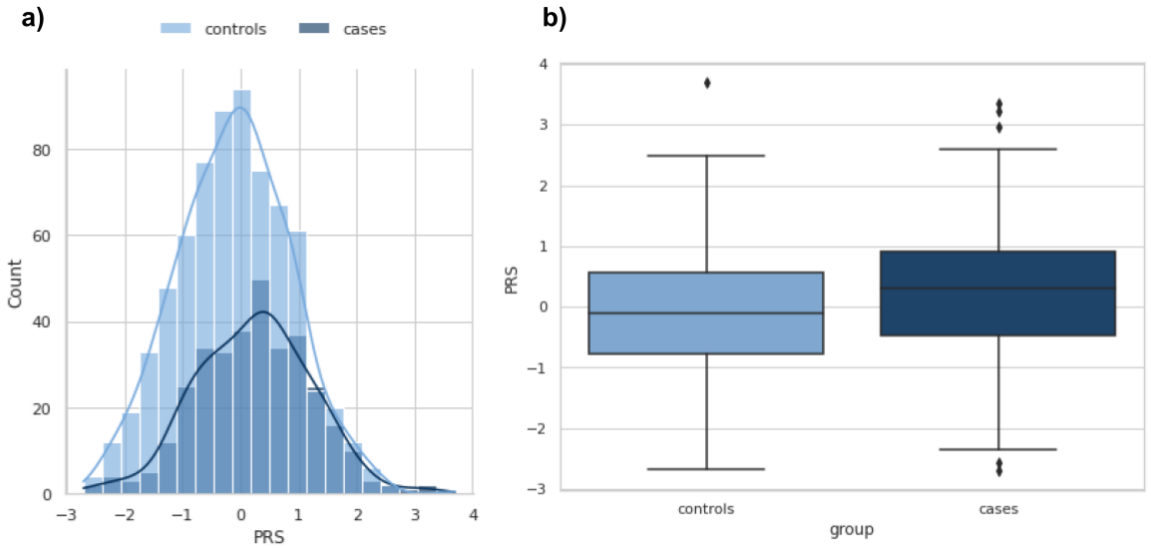

Figure 2. Comparison of distributions of PRS values between the groups with and without severe COVID-19. a) Distribution of PRS in the groups with (N_{cases} = 336) and without ($N_{controls}$ = 704) severe COVID-19. The x-axis represents PRS, with values scaled to a mean of 0 and a standard deviation of 1 (in the total sample) to facilitate interpretation. b) PRS values among cases versus controls. Within each box plot, the horizontal lines reflect the median, the top, and bottom of each box reflect the interquartile range, and the whiskers reflect the rest of the distribution, except for points that are determined to be “outliers”.

Across the study population, the PRS was normally distributed with the risk of severe COVID-19 rising in the right tail of the distribution, from 15% in the lowest decile to around 50% in the highest decile (Fig. 3).

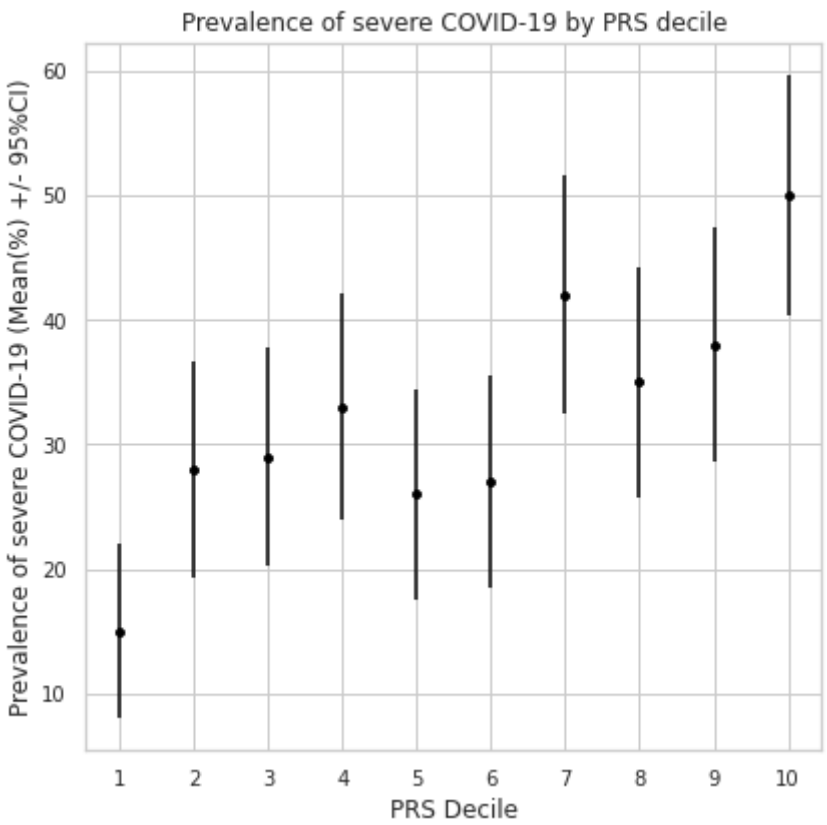

Figure 3. Prevalence of the severe COVID-19 according to PRS decile. All participants (N = 1,040) were stratified by decile of the PRS distribution. The average prevalence in percent and 95% CI within each decile are displayed.

Next, we found that 20% of the population with the highest PRS values had inherited a genetic predisposition that conferred OR = 1.66 for severe COVID-19 (95% CI: 1.2–2.3, p-value = 0.0013) in comparison with the whole cohort. The 10% of the population with the highest PRS values had an OR = 2.1 for COVID-19 (95% CI: 1.4–3.2, p-value = 0.00046).

Evaluating the relationship between PRS and COVID-19 outcome

The severe form of the disease is associated with an increased risk of death. To assess how much the risk of death is associated with an increased PRS value, we calculated the odds ratio (OR) for death between the group with the highest PRS values (10%) and all others. The resulting OR was 1.8 (95% CI: 1.05–3.04) with p-value = 0.026. Thus, in the group with the highest PRS values, the probability of death due to severe disease was almost doubled.

We also compared the mean PRS values for groups with different COVID-19 outcomes (death vs no death or no disease). Results showed significant difference in mean PRS (p-value = 0.005, Fig. 4).

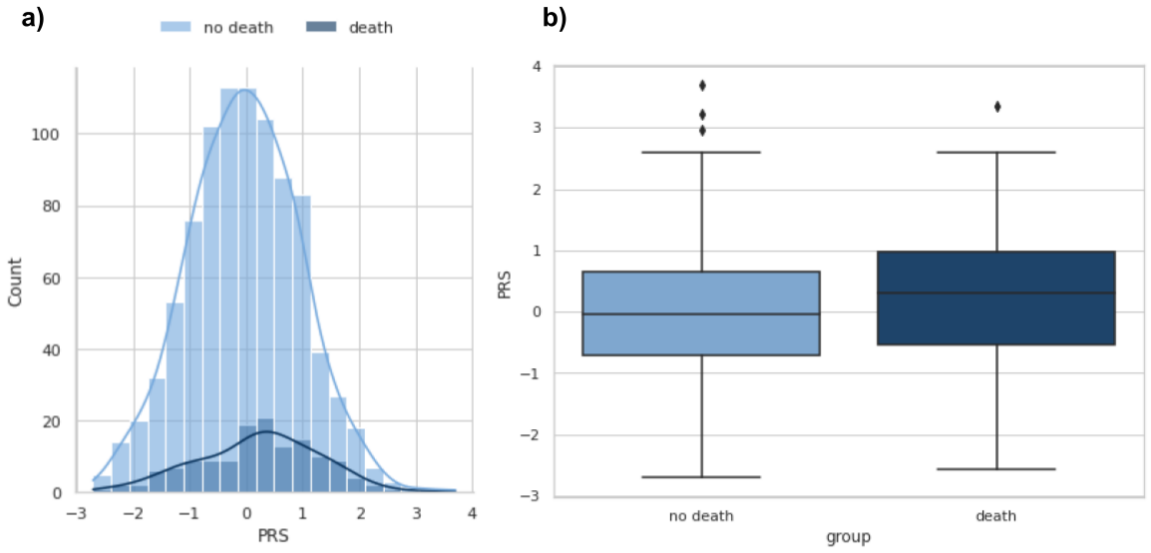

Figure 4. Comparison of distributions of PRS values between the groups with and without death outcome. a) Distribution of PRS in the groups with (N_{death} = 140) and without ($N_{no\ death}$ = 900) death outcome of COVID-19. The x-axis represents PRS, with values scaled to a mean of 0 and a standard deviation of 1 (in the total sample) to facilitate interpretation. b) PRS values among cases versus controls. Within each box plot, the horizontal lines reflect the median, the top, and bottom of each box reflect the interquartile range, and the whiskers reflect the rest of the distribution, except for points that are determined to be "outliers".

Next, we hypothesized that PRS for severe COVID-19 would be associated with a higher risk of severe COVID-19 in early age. In Kaplan–Meier analyses, which is a non-parametric statistic used to estimate the survival function from lifetime data, we divided the sample into three groups: 10% of all individuals with the highest PRS values, 10% of all individuals with the lowest PRS values and the rest (Fig. 5). The analysis showed that people from the group of high PRS values start to have increased risk in comparison with other groups already before the age of 40 years (p-value < 4.9e-9 for the log rank test). For example, the average risk of a severe course, which is reached at the age of 60 years, in the group with the highest PRS is reached before 50 years of age.

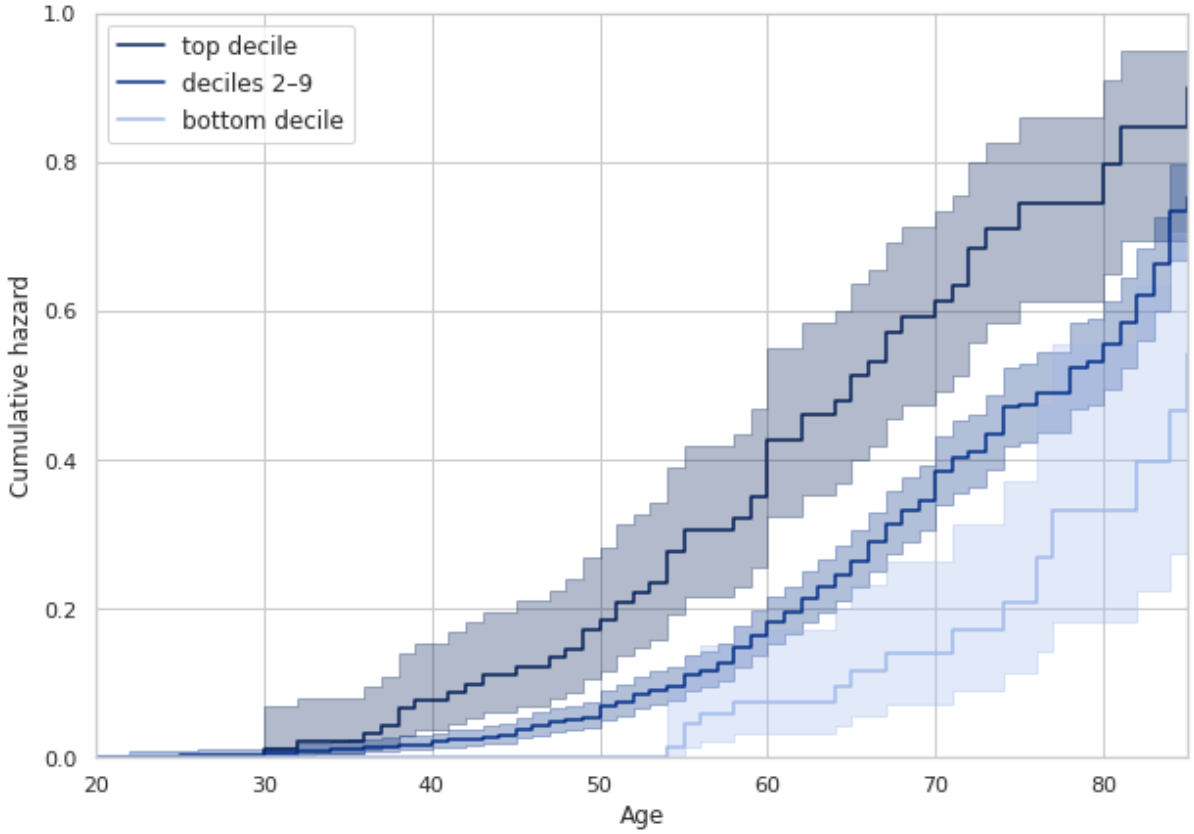

Figure 5. Association of PRS with Incident Severe COVID-19. All participants (N = 1,040) were stratified into three categories, based on their PRS: bottom decile, deciles 2–9, and top decile. Incident severe COVID-19 is plotted according to the PRS category.

Receiver Operating Curve (ROC) analysis

Next we analysed the association between PRS and severe COVID-19 using a multivariate logistic regression model adjusted for sex, age, and the first 10 principal components of genetic variation. In the adjusted model, a significant association between PRS and severe COVID-19 was found: OR = 1.51 per standard deviation (95% CI: 1.3–1.7 with p-value < 0.0001). High values of PRS (the 10% of PRS distribution) were associated with the adjusted OR = 2.5 (95% CI: 1.6–3.9, p-value < 0.0001, the reference was taken as the mean deciles 10%-90%). In comparison, unadjusted OR = 2.1 (95% CI: 1.4–3.2, p-value = 0.0004 with the reference taken as the mean deciles 10%-90%).

Analyses showed significant (p-value < 0.0001) improvements in AUC with the addition of PRS to the base model containing only the demographic predictors. Figure 6 shows that a model predicting the risk of severe COVID-19 had an AUC of 65% (95% CI: 62–69% by the formula given by Hanley and McNeil [43]) for a model excluding PRS, and it increased up to 67% (95% CI: 64–71%) when PRS was included.

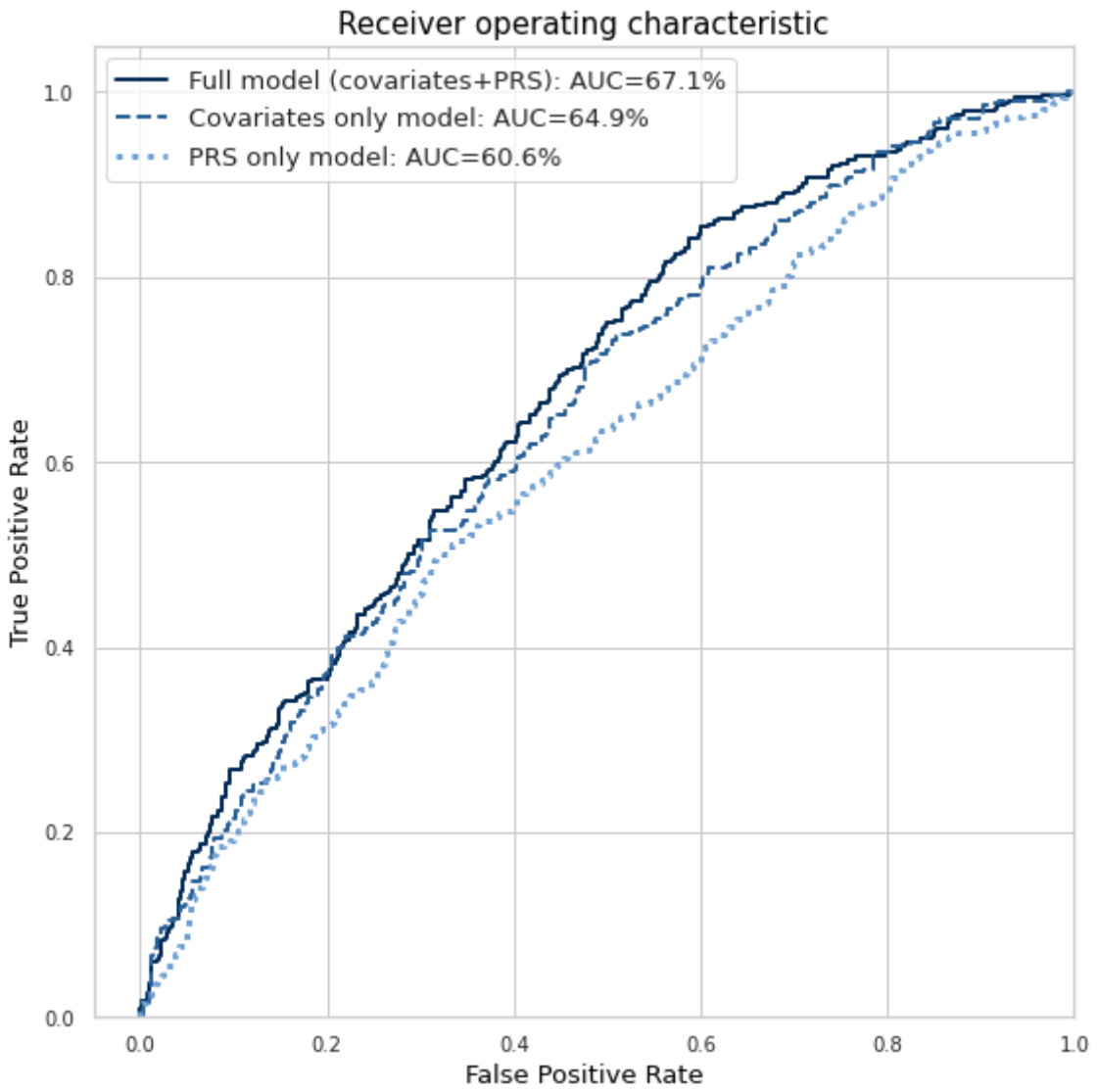

Figure 6. The comparison of receiving operating curves for three logistic regression models. The full model included the demographic predictors (sex and age), the PRS, and the first 10 principal components of genetic variation, while the covariates-only model excluded the PRS.

DISCUSSION

In this study, we constructed a polygenic risk model for the prediction of the severity of COVID-19 and applied it to a target cohort of 1,040 Russian participants. Comparing the distributions of a PRS, incorporating information from one million common genetic variants between the case and control groups revealed significant differences, indicating meaningful associations between the PRS and corresponding COVID-19 outcomes. We also demonstrated the potential of LP-WGS with coverage less than 5x to be used for predicting the severity of COVID-19.

Our main objective was to evaluate the predictive ability of PRS for COVID-19 severity. To achieve this, we developed a logistic regression model that included only demographic and technical covariates and the full model that also incorporated the PRS. Comparison between

these models demonstrated that incorporating our PRS significantly enhanced the predictive accuracy. These findings align with a previous analysis made by Huang et al., where PRS values for severe COVID-19 were constructed by using 112 SNPs in 430,582 participants from the UK Biobank study [29]. In this work, AUC was calculated for a model including only demographic and clinical parameters, and for the full model, which also included the PRS. For the first model, the AUC was 0.789, while in the full mode, the AUC was 0.794 (p-value = 0.002 for increment in AUC). Higher overall prediction accuracy of the model could be attributed to utilisation of information on comorbidities (cardiovascular disease, hypertension, diabetes, chronic respiratory infections, asthma, and chronic obstructive pulmonary disease). Our PRS, based on approximately one million SNPs, gave a comparable improvement in AUC (2% in our study vs 0.5% of Huang et al.). The higher contribution of the PRS in our case can be explained by the much larger number of genetic variants used but also by the absence of clinical factors in our model. Indeed, it is often observed that adding a predictor to a model having a high AUC improves it by an amount smaller than what could be achieved by adding the same factor to a poorer model.

Furthermore, stratifying individuals by PRS quantiles revealed an association with a distinctive risk of severe COVID-19 in resulting groups. The highest PRS categories generally exhibited higher (up to 2.1 for the top 10% PRS) odds ratios. This genetic basis for differences in disease severity among individuals also extended to the occurrence of fatalities due to COVID-19 (OR = 1.8 for the top 10% PRS). These results demonstrate that polygenic risks can be employed to stratify patients and assess their risk of severe disease and mortality related to COVID-19.

Additional survival analysis using the non-parametric Kaplan-Meier estimation revealed that the highest risk categories as defined by PRS not only exhibited higher odds ratios for COVID-19 severity but also experienced an earlier onset of increased risk compared to the mean- and low-risk categories. These findings provide insights into both the overall risk for severe COVID-19 and how the risk varies by age.

These results can have practical implications for protecting individuals with a greater genetic vulnerability during potential future outbreaks. Targeted public health interventions, such as shielding measures, closer monitoring, protection from high-risk frontline work, and prioritization for vaccination, could help to mitigate the associated risk. Hospital-based applications of PRS could facilitate the screening of COVID-19 patients and aid in the early detection of severe disease [28]. Moreover, informing patients about their increased polygenic risk has shown some evidence of positive behavioural impact [44], potentially leading to a decrease in risk-taking behaviours and promoting better outcomes.

A few limitations of the study should be noted. Firstly, despite the multi-ethnic and global nature of the HGI Release 7 meta-analysis, the participants were mostly of Western European descent, which may have affected the accuracy of the predictions in non-Western European populations [9]. Additionally, the lack of detailed clinical data led to the use of CT scans as a criterion for disease severity, which could have introduced some inaccuracy in the classification of the outcome measure for some participants.

METHODS

Study population and genetic sequencing

As part of the COVID-19 study, biomaterial (blood) and clinical data from COVID-19 patients hospitalized in the infectious disease department of the St. Petersburg State Budgetary Healthcare Institution "City Hospital No. 40 of Kurortny District" were collected. In this work, low-coverage (2–5x) sequencing was performed for 1040 samples divided into 41 batch sizes. Low-coverage sequencing, also called LP-WGS (low-pass whole genome sequencing), is a low-cost, high-throughput DNA sequencing technology used to accurately detect genetic variation in the genomes of multiple species [45]. Using imputation algorithms, this technology provides high variant detection accuracy with low sequence coverage. LP-WGS and subsequent imputation yield more accurate genotypes than imputation using array-based genotyping data, allowing for increased power in GWAS studies and more accurate results in polygenic risk studies [46].

Prior to sequencing, preliminary analysis and quality control of the case database were performed, and preliminary analysis of samples from each batch was performed to exclude bias for any of the sample characteristics: age, sex, and case/control.

Genome DNA isolation was performed with QIAcube, using a QIAamp DNA Blood Mini Kit. DNA concentration is measured with a Promega QuantiFluor dsDNA System. Library preparation was done using a Roche KAPA HyperPlus Kit. Quality control electrophoresis was done on a QIAxcel station using a QIAxcel High Resolution Kit. Circularization was made with an MGIEasy Circularization Kit. Sequencing was done on a MGISEQ-2000 sequencing machine with the DNBSEQ-G400RS High-throughput Sequencing Set (FCL PE150, 540 G).

Variant calling, imputation, and quality control

Quality analysis (FastQC, version 0.11.9) [47], alignment (BWA, version 0.7.17) [48], deduplication (samtools, version 1.16.1), and variant collation (bcftools, version 1.16) were performed for the reads obtained from sequencing [49]. Imputation of the resulting data was then performed using the GLIMPSE tool (version 1.1.1) [38], which allows imputation of low-coverage sequencing data. To improve imputation quality, only bi-allelic sites were retained from the LP-WGS BAM data and processed with bcftools. Then iterative refinement of the genotype likelihood using the reference panels with segmentation size of 2 Mb with buffer size of 200 kb produced imputed dosages and multiple chunks within each chromosome were ligated. A panel of the 1000 Genomes Project with high coverage [40], including high-quality single nucleotide variants and insertion-deletion mutations (SNV- and INDELs) from over 3,000 samples, was used as a reference sample.

Subsequently, we filtered imputed variants by an imputation INFO score, where variants with score < 0.7 and a minor allele frequency $< 0.1\%$ were removed from the analysis [9,13]. We focused on the variants and individuals with a call rate of more than 90%. Close relatives, identified using the KING-robust method [50], were removed from analysis. Using a threshold (kinship > 0.125), we found pairs of first- and second-degree relatives. We restricted our analyses to a list of variants from HapMap3 [51], which are included in the PRS models. PLINK 1.90 software [52] was utilised for all genotype extraction and quality control.

Establishing COVID-19 outcomes

The severity of the course was divided according to the following criteria: the case group included samples with lung lesions greater than 50% (computed tomography (CT)-3 and CT-4), while the control group included all other samples. As a result, the case group included 336 patients (208 men and 128 women, 63±15 years) with lung damage more than 50% (computed tomography (CT)-3 and CT-4), the control group included 704 patients (374 men and 330 women, 56±16 years), with lung damage less than 50% or without COVID-19.

Construction of PRS models

The calculation of PRS values relies on both genotype data from the target individuals and a PRS model. To derive a PRS model, GWAS are used to estimate the effect sizes of SNPs [53]. However, a GWAS gives the marginal effect size for each SNP estimated by a regression model that ignores linkage disequilibrium (LD) structure. As a result, to construct a PRS model that incorporates multiple SNPs, the SNP effects must be re-estimated while accounting for LD structure.

We used the GWAS from the COVID-19 Host Genetics Initiative consortium (release 7). These results were obtained by meta-analysis, which combined the results of 60 individual studies from 25 countries.

To re-weight the effect sizes we used SBayesR (version gctb_2.02), a software tool that has demonstrated superior performance compared to similar tools [19]. This tool re-weights the effects of each genetic variant based on the marginal estimate of its effect size, statistical strength of association, the degree of correlation between the variant and other variants nearby, and tuning parameters. It also requires a GCTB-compatible LD matrix file based on individual-level data from a reference population, and for this analysis, we used a shrunk sparse GCTB LD matrix from 50,000 individuals of European ancestry in the UK Biobank dataset [41].

PRS values were calculated as a weighted sum of allele counts:

$$PRS_i = \sum_j^N \beta_j G_{ij},$$

with β_j the re-weighted effect size of the j^{th}SNP, G_{ij} the genotype of the jth SNP for i^{th} individual, *N* the number of SNPs from the model. PLINK 1.90 software [52] was utilised for PRS calculation.

Statistical analysis and association testing

Logistic regression of PRS categories against COVID-19 severity outcomes was then conducted using R (version 4.3.2) [54] and Python (version 3.8.17) [55], fully adjusted for covariates, such as sex and age. Data on comorbidities were not available for the majority of patients, nor were other clinical data, so parameters for these were not included in the model to cover as much data as possible. The first 10 principal genetic components (PCs) were

also included as covariates to adjust for population genetic structures and avoid bias, as per current recommendations [13].

The discriminative power of models in identifying high-risk individuals was then assessed using receiver operating curve (ROC) analysis. The area under the ROC (AUC) was calculated for full models (consisting of covariates and PRS) and base models (covariates only). The confidence interval for the AUC was calculated using the formula given by Hanley and McNeil [43]. Increment in AUC (ΔAUC) was reported based on the difference between the two models, reported as the discriminative or predictive power conferred by PRS. The permutation test for differences between classifiers was used to estimate the significance (p-value) of an increment in AUC.

Once the PRS was calculated, individuals were separately stratified into quintiles for susceptibility and severity PRS, then categorised into low genetic risk (decile 1, bottom 10% of cohort), intermediate risk (decile 2–9, middle 80%) and high risk (decile 10, top 10%) for each outcome. In each group, we estimated the cumulative hazard curve using the non-parametric method called the Kaplan-Meier estimator [56]. For each pair of groups, the log rank test was applied, which is the statistical test for comparing the survival distributions of two or more groups.

DATA AND CODE AVAILABILITY

Personal genetic and clinical data are under restrictions and are available through collaboration with the St. Petersburg State Health Care Institution "City Hospital No. 40, Kurortny District" hospital.

ACKNOWLEDGEMENTS

The authors are grateful to the study participants and the staff from the St. Petersburg State Health Care Institution "City Hospital No. 40, Kurortny District" hospital. The authors would like to thank all authors of the included studies for their valuable contributions to data collection. This work was supported by Saint Petersburg State University, project ID: 94029859.

CONFLICT OF INTEREST

YSA is a co-owner of PolyKnomics BV, a private organization providing services, research, and development in the field of computational and statistical genomics. YSA is currently a full-time employee of GSK. The other authors declare that they have no competing interests.

ETHICS STATEMENT

Ethics committee of The Saint Petersburg State Health Care Establishment the City Hospital No 40 of the Resort District gave ethical approval for this work (protocol number 171, May 18, 2020).

REFERENCES

1. Bose S, Adapa S, Aeddula NR, Roy S, Nandikanti D, Vupadhyayula PM, et al. Medical Management of COVID-19: Evidence and Experience. J Clin Med Res. 2020;12: 329–343.

2. Bevova MR, Netesov SV, Aulchenko YS. The New Coronavirus COVID-19 Infection. Mol Gen Microbiol Virol. 2020;35: 53–60.

3. COVID-19 National Preparedness Collaborators. Pandemic preparedness and COVID-19: an exploratory analysis of infection and fatality rates, and contextual factors associated with preparedness in 177 countries, from Jan 1, 2020, to Sept 30, 2021. Lancet. 2022;399: 1489–1512.

4. Biswas M, Rahaman S, Biswas TK, Haque Z, Ibrahim B. Association of Sex, Age, and Comorbidities with Mortality in COVID-19 Patients: A Systematic Review and Meta-Analysis. Intervirology. 2020; 1–12.

5. Wang Y, Wang Y, Chen Y, Qin Q. Unique epidemiological and clinical features of the emerging 2019 novel coronavirus pneumonia (COVID-19) implicate special control measures. J Med Virol. 2020;92: 568–576.

6. Fricke-Galindo I, Falfán-Valencia R. Genetics Insight for COVID-19 Susceptibility and Severity: A Review. Front Immunol. 2021;12: 622176.

7. Yousefzadegan S, Rezaei N. Case Report: Death due to COVID-19 in Three Brothers. Am J Trop Med Hyg. 2020;102: 1203–1204.

8. 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;28: 715–718.

9. Mapping the human genetic architecture of COVID-19. Nature. 2021;600: 472–477.

10. Pairo-Castineira E, Clohisey S, Klaric L, Bretherick AD, Rawlik K, Pasko D, et al. Genetic mechanisms of critical illness in COVID-19. Nature. 2021;591: 92–98.

11. Velavan TP, Pallerla SR, Rüter J, Augustin Y, Kremsner PG, Krishna S, et al. Host genetic factors determining COVID-19 susceptibility and severity. EBioMedicine. 2021;72: 103629.

12. Dudbridge F. Power and predictive accuracy of polygenic risk scores. PLoS Genet. 2013;9: e1003348.

13. Choi SW, Mak TS-H, O'Reilly PF. Tutorial: a guide to performing polygenic risk score analyses. Nat Protoc. 2020;15: 2759–2772.

14. de los Campos G, Gianola D, Allison DB. Predicting genetic predisposition in humans: the promise of whole-genome markers. Nat Rev Genet. 2010;11: 880–886.

15. Vilhjálmsson BJ, Yang J, Finucane HK, Gusev A, Lindström S, Ripke S, et al. Modeling Linkage Disequilibrium Increases Accuracy of Polygenic Risk Scores. Am J Hum Genet. 2015;97: 576–592.

16. Zhu X, Stephens M. BAYESIAN LARGE-SCALE MULTIPLE REGRESSION WITH SUMMARY STATISTICS FROM GENOME-WIDE ASSOCIATION STUDIES. Ann Appl Stat. 2017;11: 1561–1592.

17. Khera AV, Chaffin M, Aragam KG, Haas ME, Roselli C, Choi SH, et al. Genome-wide

polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations. Nat Genet. 2018;50: 1219–1224.

18. Choi SW, O'Reilly PF. PRSice-2: Polygenic Risk Score software for biobank-scale data. Gigascience. 2019;8. doi:10.1093/gigascience/giz082

19. Lloyd-Jones LR, Zeng J, Sidorenko J, Yengo L, Moser G, Kemper KE, et al. Improved polygenic prediction by Bayesian multiple regression on summary statistics. Nat Commun. 2019;10: 5086.

20. Li R, Chang C, Tanigawa Y, Narasimhan B, Hastie T, Tibshirani R, et al. Fast numerical optimization for genome sequencing data in population biobanks. Bioinformatics. 2021;37: 4148–4155.

21. Privé F, Arbel J, Vilhjálmsson BJ. LDpred2: better, faster, stronger. Bioinformatics. 2021;36: 5424–5431.

22. Ojavee SE, Kousathanas A, Trejo Banos D, Orliac EJ, Patxot M, Läll K, et al. Genomic architecture and prediction of censored time-to-event phenotypes with a Bayesian genome-wide analysis. Nat Commun. 2021;12: 2337.

23. Wand H, Lambert SA, Tamburro C, Iacocca MA, O'Sullivan JW, Sillari C, et al. Improving reporting standards for polygenic scores in risk prediction studies. Nature. 2021;591: 211–219.

24. Lambert SA, Gil L, Jupp S, Ritchie SC, Xu Y, Buniello A, et al. The Polygenic Score Catalog as an open database for reproducibility and systematic evaluation. Nat Genet. 2021;53: 420–425.

25. Mavaddat N, Michailidou K, Dennis J, Lush M, Fachal L, Lee A, et al. Polygenic Risk Scores for Prediction of Breast Cancer and Breast Cancer Subtypes. Am J Hum Genet. 2019;104: 21–34.

26. Musliner KL, Mortensen PB, McGrath JJ, Suppli NP, Hougaard DM, Bybjerg-Grauholm J, et al. Association of Polygenic Liabilities for Major Depression, Bipolar Disorder, and Schizophrenia With Risk for Depression in the Danish Population. JAMA Psychiatry. 2019;76: 516–525.

27. Cupido AJ, Tromp TR, Hovingh GK. The clinical applicability of polygenic risk scores for LDL-cholesterol: considerations, current evidence and future perspectives. Curr Opin Lipidol. 2021;32: 112–116.

28. Lambert SA, Abraham G, Inouye M. Towards clinical utility of polygenic risk scores. Hum Mol Genet. 2019;28: R133–R142.

29. Huang Q-M, Zhang P-D, Li Z-H, Zhou J-M, Liu D, Zhang X-R, et al. Genetic Risk and Chronic Obstructive Pulmonary Disease Independently Predict the Risk of Incident Severe COVID-19. Ann Am Thorac Soc. 2022;19: 58–65.

30. Dite GS, Murphy NM, Allman R. Development and validation of a clinical and genetic model for predicting risk of severe COVID-19. Epidemiol Infect. 2021;149: e162.

31. Dite GS, Murphy NM, Allman R. An integrated clinical and genetic model for predicting risk of severe COVID-19: A population-based case-control study. PLoS One. 2021;16: e0247205.

32. Horowitz JE, Kosmicki JA, Damask A, Sharma D, Roberts GHL, Justice AE, et al.

Genome-wide analysis provides genetic evidence that ACE2 influences COVID-19 risk and yields risk scores associated with severe disease. Nat Genet. 2022;54: 382–392.

33. Ahmetov II, Borisov OV, Semenova EA, Andryushchenko ON, Andryushchenko LB, Generozov EV, et al. Team sport, power, and combat athletes are at high genetic risk for coronavirus disease-2019 severity. J Sport Health Sci. 2020;9: 430–431.

34. Farooqi R, Kooner JS, Zhang W. Associations between polygenic risk score and covid-19 susceptibility and severity across ethnic groups: UK Biobank analysis. BMC Med Genomics. 2023;16: 150.

35. Sudlow C, Gallacher J, Allen N, Beral V, Burton P, Danesh J, et al. UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. PLoS Med. 2015;12: e1001779.

36. Chaubey A, Shenoy S, Mathur A, Ma Z, Valencia CA, Reddy Nallamilli BR, et al. Low-Pass Genome Sequencing: Validation and Diagnostic Utility from 409 Clinical Cases of Low-Pass Genome Sequencing for the Detection of Copy Number Variants to Replace Constitutional Microarray. J Mol Diagn. 2020;22: 823–840.

37. Li JH, Mazur CA, Berisa T, Pickrell JK. Low-pass sequencing increases the power of GWAS and decreases measurement error of polygenic risk scores compared to genotyping arrays. Genome Res. 2021;31: 529–537.

38. Rubinacci S, Ribeiro DM, Hofmeister RJ, Delaneau O. Efficient phasing and imputation of low-coverage sequencing data using large reference panels. Nat Genet. 2021;53: 120–126.

39. Hui R, D'Atanasio E, Cassidy LM, Scheib CL, Kivisild T. Evaluating genotype imputation pipeline for ultra-low coverage ancient genomes. Sci Rep. 2020;10: 18542.

40. 1000 Genomes Project Consortium, Auton A, Brooks LD, Durbin RM, Garrison EP, Kang HM, et al. A global reference for human genetic variation. Nature. 2015;526: 68–74.

41. Bycroft C, Freeman C, Petkova D, Band G, Elliott LT, Sharp K, et al. The UK Biobank resource with deep phenotyping and genomic data. Nature. 2018;562: 203–209.

42. Jiang L, Zheng Z, Qi T, Kemper KE, Wray NR, Visscher PM, et al. A resource-efficient tool for mixed model association analysis of large-scale data. Nat Genet. 2019;51: 1749–1755.

43. Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology. 1982;143: 29–36.

44. Frieser MJ, Wilson S, Vrieze S. Behavioral impact of return of genetic test results for complex disease: Systematic review and meta-analysis. Health Psychol. 2018;37: 1134–1144.

45. Alex Buerkle C, Gompert Z. Population genomics based on low coverage sequencing: how low should we go? Mol Ecol. 2013;22: 3028–3035.

46. Homburger JR, Neben CL, Mishne G, Zhou AY, Kathiresan S, Khera AV. Low coverage whole genome sequencing enables accurate assessment of common variants and calculation of genome-wide polygenic scores. Genome Med. 2019;11: 74.

47. Andrews S. FastQC: A Quality Control Tool for High Throughput Sequence Data. In: FastQC [Internet]. 2010 [cited 2010]. Available:

http://www.bioinformatics.babraham.ac.uk/projects/fastqc/

48. Li H, Durbin R. Fast and accurate short read alignment with Burrows–Wheeler transform. Bioinformatics. 2009;25: 1754–1760.

49. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009;25: 2078–2079.

50. Manichaikul A, Mychaleckyj JC, Rich SS, Daly K, Sale M, Chen W-M. Robust relationship inference in genome-wide association studies. Bioinformatics. 2010;26: 2867–2873.

51. HapMap Project. HapMap Project. In: HapMap 3 [Internet]. Available: https://www.sanger.ac.uk/data/hapmap-3/

52. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet. 2007;81: 559–575.

53. Uffelmann E, Huang QQ, Munung NS, de Vries J, Okada Y, Martin AR, et al. Genome-wide association studies. Nature Reviews Methods Primers. 2021;1: 1–21.

54. Team RDC. R: A language and environment for statistical computing. (No Title). 2010 [cited 20 Oct 2023]. Available: https://cir.nii.ac.jp/crid/1370294721063650048

55. Van Rossum G, Drake FL. Python 3 Reference Manual; CreateSpace: Scotts Valley, CA, USA, 2009. Google Scholar.

56. Goel MK, Khanna P, Kishore J. Understanding survival analysis: Kaplan-Meier estimate. Int J Ayurveda Res. 2010;1: 274–278.

Supplementary Materials

These supplementary materials present additional details about the results.

LOGISTIC REGRESSION RESULTS

The full model

We analysed the association between PRS and severe COVID-19 using a multivariate logistic regression model adjusted for sex, age, and the first 10 principal components of genetic variation.

Table 1: Logistic regression results for the full model.

	coef	std err	z	P>|z|	[0.025	0.975]
intercept	-0.0495	0.219	-0.226	0.821	-0.478	0.379
age	0.5637	0.078	7.187	0.000	0.410	0.717
sex	-0.5737	0.145	-3.952	0.000	-0.858	-0.289
PRS	0.4135	0.073	5.651	0.000	0.270	0.557
PC1	-0.0025	0.095	-0.026	0.979	-0.189	0.184
PC2	0.2154	1.240	0.174	0.862	-2.216	2.647
PC3	-0.5792	1.203	-0.481	0.630	-2.938	1.779
PC4	0.3239	0.599	0.541	0.589	-0.850	1.498
PC5	0.3810	0.409	0.931	0.352	-0.421	1.183
PC6	0.2982	0.296	1.009	0.313	-0.281	0.878
PC7	-0.1118	0.109	-1.022	0.307	-0.326	0.103
PC8	-0.0367	0.072	-0.510	0.610	-0.178	0.104
PC9	0.0875	0.085	1.031	0.303	-0.079	0.254
PC10	0.0153	0.072	0.212	0.832	-0.127	0.157

- **coef**: the coefficient of the variable (log[odd ratio]);
- **std err**: the standard error of the coefficient;
- **Z**: the z value for the estimated coefficient and the standard error;
- **P>|z|**: the p-value associated with the value in the z value column;
- **[0.025, 0.975]**: 95% confident interval (CI) for the coefficient.

The covariates-only model

The base model contains only the demographic predictors (sex, age, and the first 10 principal components of genetic variation).

Table 2: Logistic regression results for the covariates-only model.

	coef	std err	z	P>|z|	[0.025	0.975]

intercept	-0.0549	0.214	-0.257	0.797	-0.474	0.364
age	0.5321	0.076	6.982	0.000	0.383	0.681
sex	-0.5383	0.143	-3.775	0.000	-0.818	-0.259
PC1	-0.0284	0.093	-0.305	0.760	-0.211	0.154
PC2	0.8426	1.226	0.687	0.492	-1.561	3.246
PC3	-1.0866	1.186	-0.916	0.360	-3.412	1.238
PC4	0.2649	0.513	0.516	0.606	-0.741	1.271
PC5	0.3364	0.354	0.950	0.342	-0.357	1.030
PC6	0.2184	0.250	0.874	0.382	-0.271	0.708
PC7	-0.0971	0.117	-0.832	0.405	-0.326	0.132
PC8	-0.0362	0.071	-0.508	0.612	-0.176	0.104
PC9	0.1146	0.083	1.377	0.169	-0.049	0.278
PC10	0.0180	0.071	0.255	0.799	-0.121	0.157

The PRS-only model

The base model contains only PRS predictor.

Table 3: Logistic regression results for the PRS-only model.

	coef	**std err**	**z**	**P>|z|**	**[0.025**	**0.975]**
intercept	-0.8025	0.081	-9.923	0.000	-0.961	-0.644
PRS	0.3727	0.071	5.278	0.000	0.234	0.511
PC1	-0.0165	0.092	-0.178	0.858	-0.197	0.164
PC2	0.4333	1.177	0.368	0.713	-1.874	2.740
PC3	-0.8196	1.152	-0.711	0.477	-3.078	1.439
PC4	0.2147	0.525	0.409	0.683	-0.815	1.244
PC5	0.3026	0.358	0.845	0.398	-0.399	1.005
PC6	0.2557	0.263	0.972	0.331	-0.260	0.771
PC7	-0.1390	0.111	-1.254	0.210	-0.356	0.078
PC8	-0.0669	0.077	-0.869	0.385	-0.218	0.084
PC9	0.0808	0.083	0.977	0.328	-0.081	0.243
PC10	-0.0271	0.070	-0.387	0.699	-0.164	0.110

The full model with PRS as a binary predictor

We analysed the association between PRS and severe COVID-19 using a multivariate logistic regression model adjusted for sex, age, and the first 10 principal components of genetic variation. PRS values were considered as a binary predictor, where the highest 10% of PRS were taken as 1, others as 0.

Table 4: Logistic regression results for the full model with PRS as a binary predictor.

	coef	std err	z	P>|z|	[0.025	0.975]
intercept	-0.0781	0.223	-0.350	0.726	-0.515	0.359
10% with the highest PRS values	0.9143	0.225	4.069	0.000	0.474	1.355
10% with the lowest PRS values	-1.0727	0.290	-3.703	0.000	-1.640	-0.505
age	0.5789	0.079	7.342	0.000	0.424	0.733
sex	-0.5430	0.145	-3.746	0.000	-0.827	-0.259
PC1	-0.0078	0.095	-0.082	0.935	-0.195	0.179
PC2	0.5737	1.249	0.459	0.646	-1.875	3.023
PC3	-0.8886	1.208	-0.736	0.462	-3.256	1.479
PC4	0.3058	0.612	0.500	0.617	-0.893	1.505
PC5	0.3584	0.417	0.859	0.390	-0.459	1.176
PC6	0.3072	0.300	1.023	0.306	-0.281	0.896
PC7	-0.0974	0.112	-0.871	0.384	-0.317	0.122
PC8	-0.0315	0.072	-0.436	0.663	-0.173	0.110
PC9	0.0818	0.085	0.959	0.338	-0.085	0.249
PC10	0.0190	0.073	0.261	0.794	-0.123	0.161