

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

3
4 Arina V. Nostaeva¹, Valentin S. Shimansky^{1,2}, Svetlana V. Apalko^{1,2}, Ivan A. Kuznetsov³,
5 Natalya N. Sushentseva¹, Oleg S. Popov^{1,2}, Yurii S. Aulchenko^{4,5}, Sergey G. Shcherbak^{1,2}

6
7 ¹ St. Petersburg State Budgetary Healthcare Institution “City Hospital No. 40 of Kurortny
8 District”, Sestroretsk, Russia

9 ² St. Petersburg State University, St. Petersburg, Russia

10 ³ Skolkovo Institute of Science and Technology, Moscow, Russia

11 ⁴ Institute of Cytology and Genetics, Siberian Branch of Russian Academy of Sciences,
12 Novosibirsk, Russia;

13 ⁵ PolyKnomics BV, The Netherlands

14
15 Please address correspondence to:

16 Arina V. Nostaeva, MD

17 Sestroretsk, 9 Borisova str., Saint Petersburg, Russia, 197706

18 avnostaeva@gmail.com

19
20
21 **ABSTRACT**

22
23 The course of COVID-19 is characterized by wide variability, with genetics playing a
24 contributing role. Through large-scale genetic association studies, a significant link between
25 genetic variants and disease severity was established. However, individual genetic variants
26 identified thus far have shown modest effects, indicating a polygenic nature of this trait. To
27 address this, a polygenic risk score (PRS) can be employed to aggregate the effects of
28 multiple single nucleotide polymorphisms (SNPs) into a single number, allowing practical
29 application to individuals within a population. In this work, we investigated the performance
30 of a PRS model in the context of COVID-19 severity in 1085 Russian participants using low-
31 coverage NGS sequencing. By developing a genome-wide PRS model based on summary
32 statistics from the COVID-19 Host Genetics Initiative consortium, we demonstrated that the
33 PRS, which incorporates information from over a million common genetic variants, can
34 effectively identify individuals at significantly higher risk for severe COVID-19. The findings
35 revealed that individuals in the top 10% of the PRS distribution had a markedly elevated risk
36 of severe COVID-19, with an odds ratio (OR) of 2.2 (95% confidence interval (CI): 1.3-3.3, p-
37 value=0.0001). Furthermore, incorporating the PRS into the prediction model significantly
38 improved its accuracy compared to a model that solely relied on demographic information (p-
39 value < 0.0001). This study highlights the potential of PRS as a valuable tool for identifying
40 individuals at increased risk of severe COVID-19 based on their genetic profile.

41
42 **INTRODUCTION**

43
44 COVID-19, also known as coronavirus infection, is a contagious illness caused by the severe
45 acute respiratory syndrome-coronavirus-2 (SARS-CoV-2). The majority of individuals who
46 contract the virus exhibit mild to moderate respiratory symptoms and can recover without
47 requiring specific medical treatment. However, in certain cases, the disease can manifest in
48 a severe form, requiring medical intervention [1,2].

49

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

60

61 To gain insights into the aetiology of COVID-19, large-scale genetic association studies
62 incorporating both rare and common genetic variants have employed various study designs.
63 These investigations, along with subsequent follow-up studies, have expanded our
64 understanding of the disease and provided potential avenues for its treatment. The COVID-
65 19 Host Genetics Initiative (HGI) was established to identify genetic loci that impact the
66 severity and susceptibility of COVID-19 [8]. This global effort aims to conduct a meta-
67 analysis of multiple COVID-19 genome-wide association studies (GWAS), and to identify
68 significant single nucleotide polymorphisms (SNPs) associated with infection, hospitalization,
69 and mortality. Through comparisons of genomes of millions of COVID-19 patients and
70 healthy individuals, these studies have implicated genetic variants in 13 loci associated with
71 the severity of the disease [9]. The COVID-19-associated genetic variants could be related
72 to the regulation of processes such as innate antiviral defence signalling, regulation of
73 inflammatory organ damage, and upregulation of cell receptors [10]. Modulation of these
74 pathways can impact susceptibility to infection and subsequent disease manifestation [11].

75

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

86

87 Modern polygenic risk score models for human traits are typically estimated using summary
88 statistics obtained from a genome-wide association meta-analysis (GWAMA) and a
89 reference panel reflecting linkage disequilibrium (LD) in the population [13,14]. Over the past
90 decade, PRS predictive performance has significantly improved due to larger GWAS sample
91 sizes and advancements in methods for variable selection and effect estimation [15–24].
92 Polygenic scores can be utilized to rank individuals within a group based on their genetic
93 predisposition to a disease [25–27]. This approach considers an individual's genetic
94 predisposition relative to the genetic predisposition of others in the same group, often
95 expressed as a percentile representing where the individual's PRS falls within the overall
96 distribution of the group's PRS.

97

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

107

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

118

119 Our study aimed to investigate the performance of the PRS model in the Russian population.
120 The genomes of study participants (347 individuals with severe COVID-19 and 738 with
121 moderate or without disease) were assessed using low-coverage (with mean depth x3)
122 sequencing. Next, we developed a genome-wide PRS model for COVID-19 severity using
123 the summary statistics from the COVID-19 host genetics initiative consortium. We
124 demonstrated that PRS, incorporating information from more than a million common genetic
125 variants, for COVID-19 severity can identify individuals with markedly elevated risk of severe
126 COVID-19 course: OR=2.2 (95% confidence interval (CI): 1.3-3.3, p-value=0.0001) for
127 individuals in the top 10% of the PRS distribution, and produces a significant improvement in
128 the quality of prediction (p-value < 0.0001) compared to a model including only demographic
129 information.

130

131 **RESULTS**

132

133 **Participant Characteristics**

134

135 The participants of the study were the patients of the infectious disease department of the
136 St. Petersburg State Health Care Institution "City Hospital No. 40, Kurortny District" who
137 were admitted for treatment with coronavirus infection (confirmed by polymerase chain
138 reaction), and healthy individuals. Healthy individuals are defined as people who did not
139 require COVID-19 medical treatment at the time of the study (between April 2020 and March
140 2022).

141

142 Table 1 shows the participants' characteristics. Of the 1085 participants, 479 (44%) were
143 female, with a mean age of 60 years, while for 606 (56%) males the mean age was equal to
144 56 years. Overall, 895 (82%) of all participants had COVID-19, of which 347 (39%) had

145 severe COVID-19. Separation according to the severity of the disease was carried out
146 according to the following criteria: the case group included 347 patients (214 men and 133
147 women, 63±15 years) with lung damage more than 50% (computed tomography (CT)-3 and
148 CT-4), the control group included 738 patients (392 men and 346 women, 56±16 years), with
149 lung damage less than 50% or without COVID-19.

150

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

Characteristics	Male	Female
Mean age (sd)	56 (15)	60 (16)
Healthy individuals	116	74
Patients required treatment by severity	CT-1: 81 CT-2: 195 CT-3: 213 CT-4: 1	CT-1: 71 CT-2: 201 CT-3: 130 CT-4: 3
Outcome	death: 93 recovery or no disease: 513	death: 53 recovery or no disease: 426

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

156

157 **Low coverage sequencing and imputation**

158

159 For all samples low coverage sequencing, also called LP-WGS (low-pass whole genome
160 sequencing), was performed with a depth of x3 genome coverage. LP-WGS is the type of
161 WGS with genome coverage from x0.5 to x5 [36,37]. Due to low-coverage data often
162 having poor genotype quality and resulting in high missing genotype rates, the genotype
163 likelihoods (GL) need to be updated using a reference panel for more accurate genotype
164 imputation [38,39]. We used a recent method called GLIMPSE, which performs haplotype
165 phasing and genotype imputation for LP-WGS data through a Gibbs sampling procedure,
166 leading to improved accuracy [38]. As a reference panel, we used the 1000 Genomes data
167 [40]. To evaluate the efficiency of LP-WGS within PRS, we calculated PRS values for a
168 sample (not included in the study population) sequenced 45 times (in each of the batches to
169 control the quality of the sequencing process). The coefficient of variation (CV) for PRS
170 values was equal to 0.5% demonstrating a good method performance.

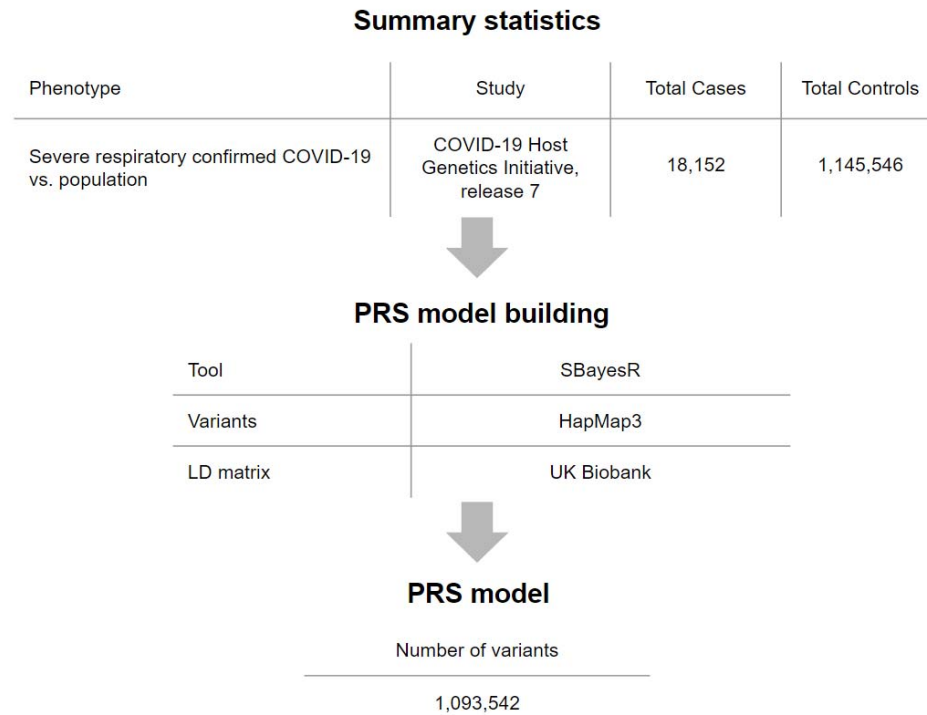
171

172 **Overview of the approach**

173

174 Computation of PRSs requires both genotype data of target individuals and the PRS model.
175 To build the PRS model we used summary statistics from the COVID-19 Host Genetics
176 Initiative consortium (release 7) [8]. These results were obtained by the meta-analysis, which
177 combined the results of 60 individual studies from 25 countries, with a total of 18,000 severe

178 cases of COVID-19 and more than a million controls who either did not have a severe
179 disease course or were not affected by COVID-19 during the study period. From the
180 obtained summary statistics, we generated the PRS model using the Bayesian approach
181 SBayesR with default parameters, implemented in the GCTB software [19,41,42]. Finally, we
182 calculated individual PRS values using the PRS model (Fig. 1, Methods).
183



184
185 **Figure 1. Study design and workflow.** The PRS model for COVID-19 severity was derived
186 by combining summary association statistics from the COVID-19 Host Genetics Initiative
187 consortium and a linkage disequilibrium reference panel of 50,000 individuals of European
188 ancestry from the UK Biobank data set. As a computational algorithm, SBayesR was used,
189 which is a Bayesian approach to calculate a posterior mean effect for all variants based on a
190 prior (effect size in the previous GWAS) and subsequent shrinkage based on linkage
191 disequilibrium. PRS model was restricted by a list of variants from HapMap3 and included
192 about one million variants.

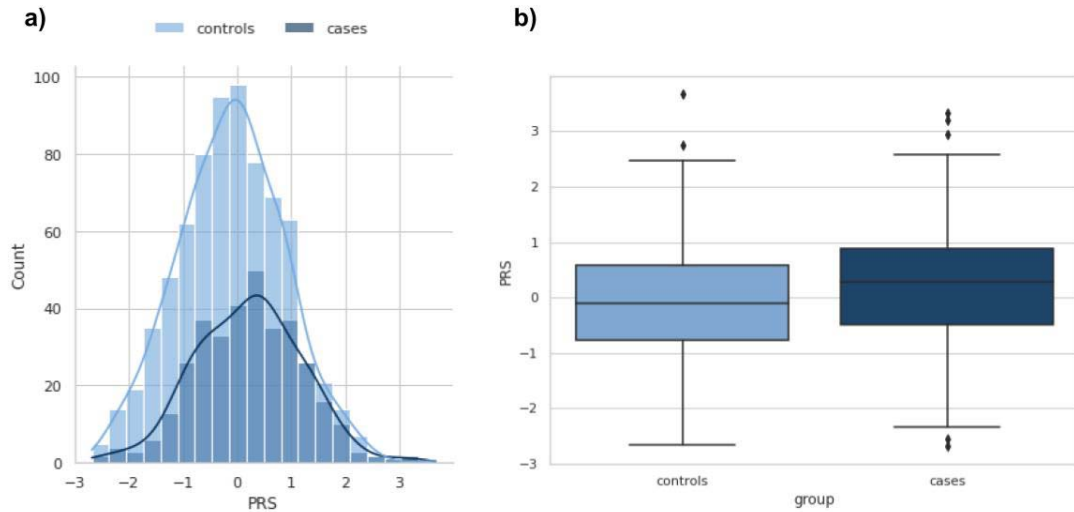
193

194 **Testing associations between PRS and severe COVID-19**

195

196 We compared the distributions of PRS values between severe cases and the control group
197 combining the milder forms of COVID-19 and healthy individuals (Fig. 2). Comparison of the
198 mean PRS values, performed using Student's t-test for two independent samples, showed
199 significant difference (p -value=1.2e-06).

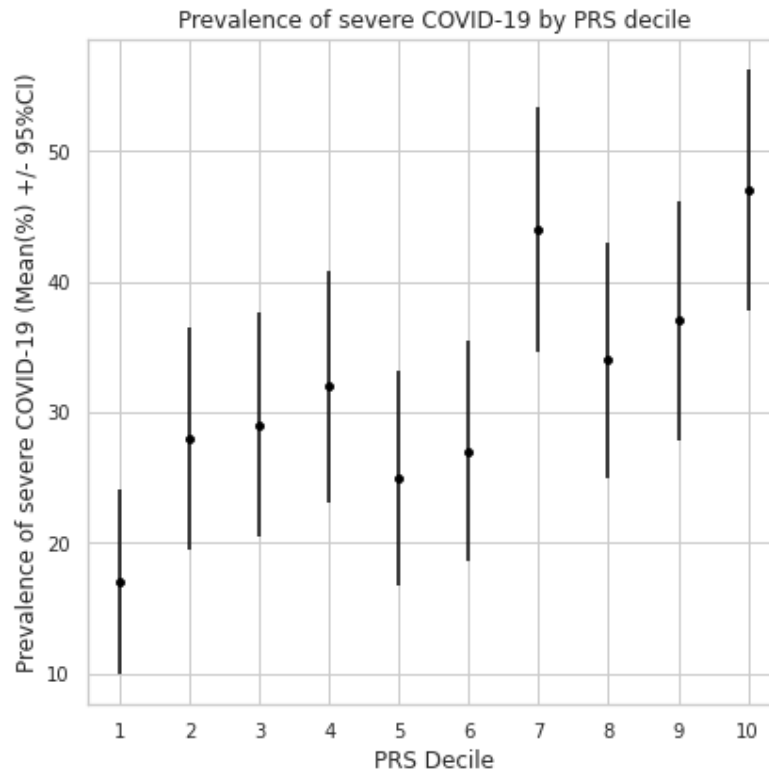
200



201
202
203
204
205
206
207
208
209
210
211
212
213

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 ($N_{\text{cases}}=347$) and without ($N_{\text{controls}}=738$) 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, PRS was normally distributed with the risk of severe COVID-19 rising in the right tail of the distribution, from 17% in the lowest decile to around 47% in the highest decile (Fig. 3).



214

215

216 **Figure 3. Prevalence of the severe COVID-19 according to PRS decile.** All participants
217 (N=1,085) were stratified by decile of the PRS distribution. The average prevalence in
218 percent and 95% CI within each decile are displayed.

218

219

220 Next, we found that 20% of the population with the highest PRS values had inherited a
221 genetic predisposition that conferred OR=1.8 for severe COVID-19 (95% CI: 1.3-2.4, p-
222 value=0.0003) in comparison with all others. The 10% of the population with the highest
223 PRS values had an OR=2.2 for COVID-19 (95% CI: 1.3-3.3, p-value=0.0001).

223

224

224 **Evaluating the relationship between PRS and COVID-19 outcome**

225

226

227

228

229

230

231

232

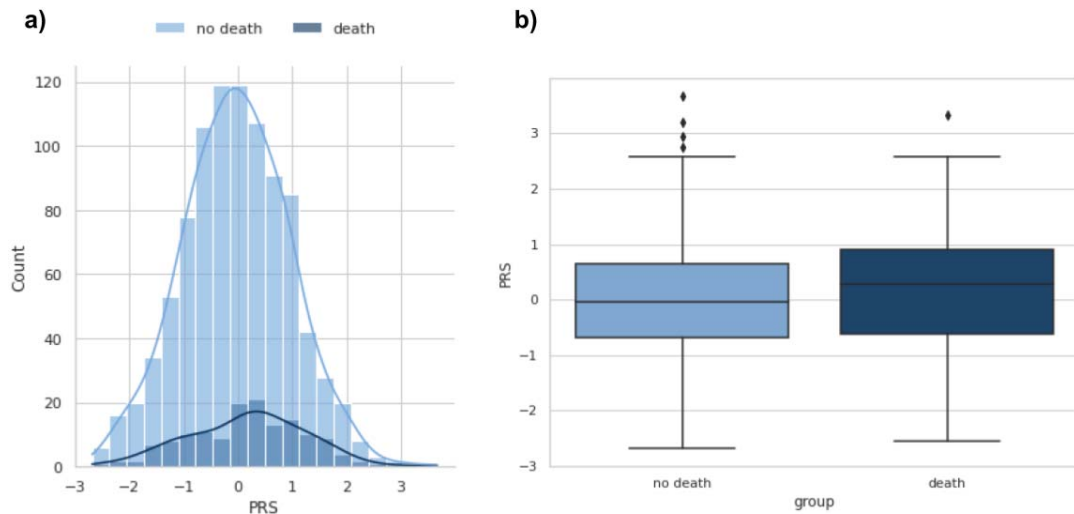
233

234

235

236

233 We also compared the mean PRS values for groups with different COVID-19 outcomes
234 (death vs no death or no disease). Results showed significant difference in mean PRS (p-
235 value = 0.02, Fig. 4).



237

238

Figure 4. Comparison of distributions of PRS values between the groups with and without death outcome.

a) Distribution of PRS in the groups with ($N_{\text{death}}=146$) and without ($N_{\text{no death}}=939$) 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”.

246

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

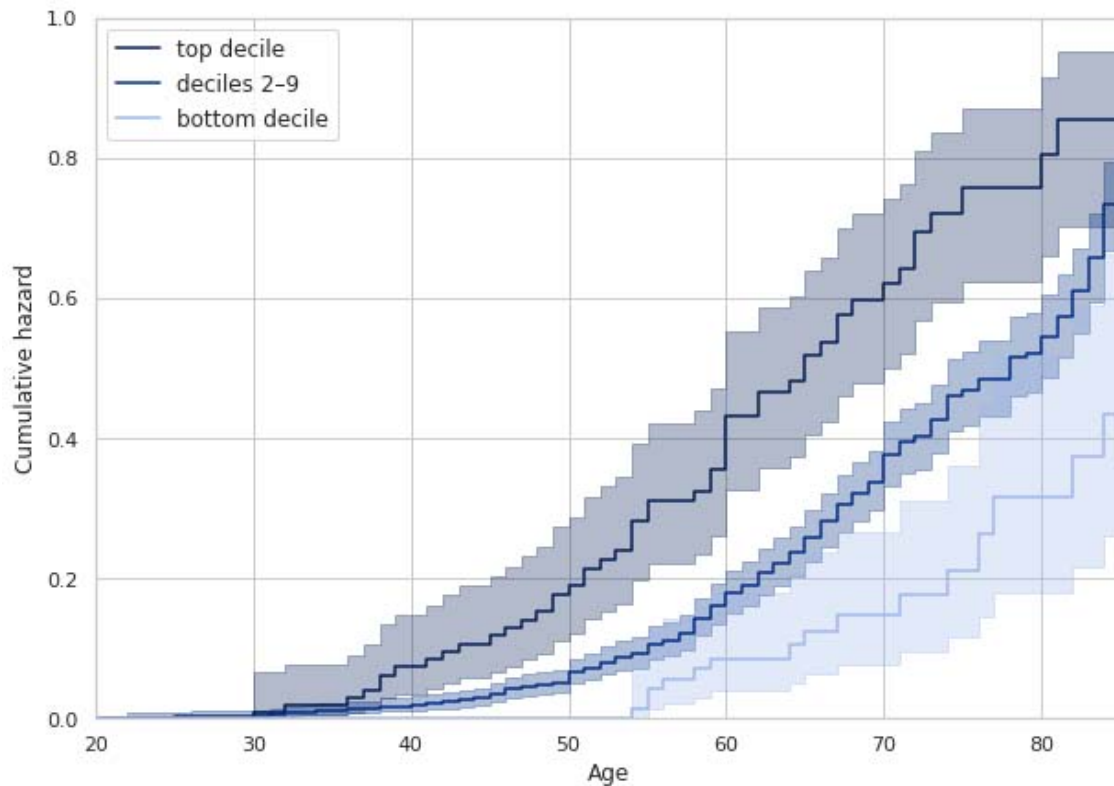
statistic used to estimate the survival function from lifetime data, we divided the sample into three groups: 10% of all individuals with the highest PRS values, 10% of all individuals with

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 ($p\text{-value}<3.7e-10$ 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 already at 50 years of age.

256



257

258

Figure 5. Association of PRS with Incident Severe COVID-19. All participants (N=1,085) were stratified, based on their PRS, into three categories: bottom decile, deciles 2–9, and top decile. Incident severe COVID-19 is plotted according to the PRS category.

260

261

262

Receiver Operating Curve (ROC) analysis

263

264

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=1.48 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=2.7 (95% CI: 1.8-4.2, p-value < 0.0001).

269

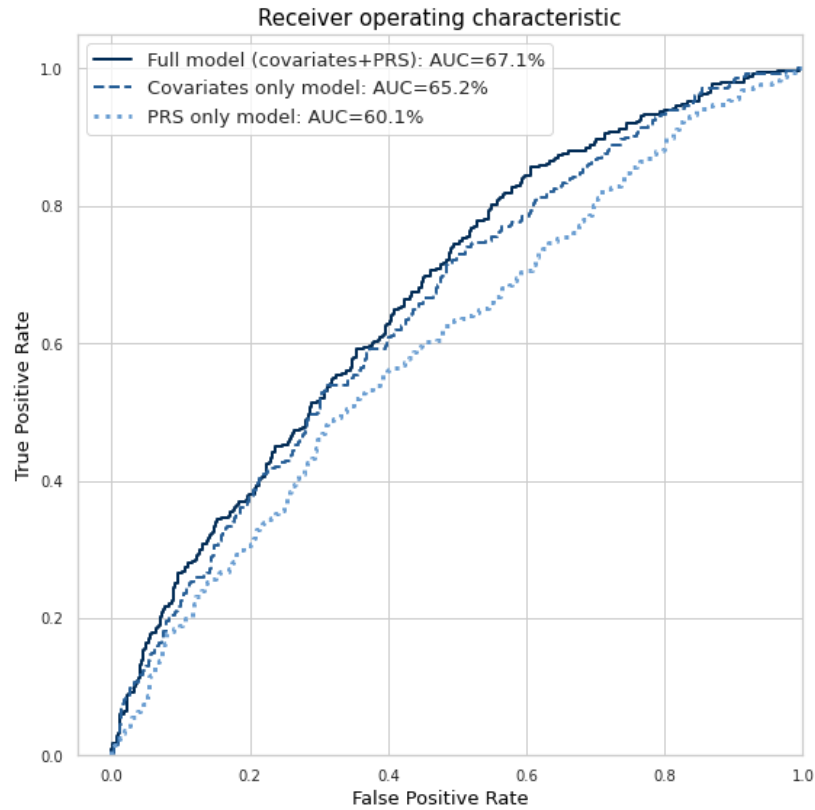
270

271

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.

275

276



277

278

279

280

281

282

283

284

285

286

287

288

289

290

291

292

293

294

295

296

297

298

299

300

301

302

Figure 6. The comparison of receiving operating curves for three logistic regression models. The full model included the demographic predictors (sex and age), PRS, and the first 10 principal components of genetic variation, while the covariates-only model excluded 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 1085 Russian participants. Comparing the distributions of PRS, incorporating information from one million common genetic variants, between the case and control groups revealed significant differences, indicating meaningful associations between PRS and corresponding COVID-19 outcomes. We also demonstrated the potential of LP-WGS with coverage less than $\times 5$ 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 PRS. Comparison between these models demonstrated that incorporating 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 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

303 attributed to utilisation of information on comorbidities (cardiovascular disease, hypertension,
304 diabetes, chronic respiratory infections, asthma, and chronic obstructive pulmonary disease).
305 Our PRS, based on approximately one million SNPs, gave a comparable improvement in
306 AUC (0.5% vs 2%, respectively). The higher contribution of PRS in our case can be
307 explained by the much larger number of genetic variants used but also by the absence of
308 clinical factors in our model. Indeed, it is often observed that adding a predictor to a model
309 having a high AUC improves it by an amount smaller than that that could be achieved by
310 adding the same factor to a poorer model.

311

312 Furthermore, stratifying individuals by PRS quantiles revealed an association with a
313 distinctive risk of severe COVID-19 in resulting groups. The highest PRS categories
314 generally exhibited higher (up to 2.2 for the top 10% PRS) odds ratios. This genetic basis for
315 differences in disease severity among individuals also extended to the occurrence of
316 fatalities due to COVID-19 (OR=1.9 for the top 10% PRS). These results demonstrate that
317 polygenic risks can be employed to stratify patients and assess their risk of severe disease
318 and mortality related to COVID-19.

319

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

325

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

334

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

341

342 **METHODS**

343

344 **Study population and genetic sequencing**

345

346 As part of the COVID-19 study, biomaterial (blood) and clinical data from COVID-19 patients
347 hospitalized in the infectious disease department of the St. Petersburg State Budgetary
348 Healthcare Institution "City Hospital No. 40 of Kurortny District" were collected. In this work,
349 low-coverage (x2-5) sequencing was performed for 1085 samples divided into 45 batch
350 sizes. Low-coverage sequencing, also called LP-WGS (low-pass whole genome

351 sequencing), is a low-cost, high-throughput DNA sequencing technology used to accurately
352 detect genetic variation in the genomes of multiple species [45]. Using imputation algorithms,
353 this technology provides high variant detection accuracy with very low sequence coverage.
354 LP-WGS and subsequent imputation yield more accurate genotypes than imputation using
355 genotyping data, allowing for increased power in GWAS studies and more accurate results
356 in polygenic risk studies [46].

357

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

361

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

368

369 **Variant calling, imputation, and quality control**

370

371 Quality analysis (FastQC) [47], alignment (BWA) [48], deduplication (samtools), and variant
372 collation (bcftools) were performed for the reads obtained from sequencing [49]. Imputation
373 of the resulting data was then performed using the GLIMPSE tool [38], which allows
374 imputation of low-coverage sequencing data. To improve imputation quality, only bi-allelic
375 sites were retained from the LP-WGS BAM data and processed with bcftools. Then iterative
376 refinement of GL using the reference panels with segmentation size of 2 Mb with buffer size
377 of 200 kb produced imputed dosages and multiple chunks within each chromosome were
378 ligated. A panel of 1000 Genomes with high coverage [40], including high-quality SNV- and
379 INDELS from over 3,000 samples, was used as a reference sample.

380

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

389

390 **Establishing COVID-19 outcomes**

391

392 The severity of the course was divided according to the following criteria: the case group
393 included samples with lung lesions greater than 50% (computed tomography (CT)-3 and CT-
394 4), while the control group included all other samples. As a result, the case group included
395 347 patients (214 men and 133 women, 63 ± 15 years) with lung damage more than 50%
396 (computed tomography (CT)-3 and CT-4), the control group included 738 patients (392 men
397 and 346 women, 56 ± 16 years), with lung damage less than 50% or without COVID-19.

398

399 Construction of PRS models

400

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

407

408 As the summary statistics, we used summary statistics from the COVID-19 Host Genetics
409 Initiative consortium (release 7). These results were obtained by the meta-analysis, which
410 combined the results of 60 individual studies from 25 countries.

411

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

419

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

$$PRS_{i,j} = \sum_{k=1}^M \beta_k G_{i,j,k}$$

421 with β_k the re-weighted effect size of the k th SNP, $G_{i,j,k}$ the genotype of the k th SNP for
422 i th individual. PLINK 1.90 software [52] was utilised for PRS calculation.

423

424 Statistical analysis and association testing

425

426 Logistic regression of PRS categories against COVID-19 severity outcomes was then
427 conducted using R [54] and Python3 [55], fully adjusted for covariates, such as sex and age.
428 Data on comorbidities were not available for the majority of patients, as well as other clinical
429 data, so parameters for these were not included in the model to cover as much data as
430 possible. The first 10 principal genetic components (PCs) were also included as covariates
431 to adjust for population genetic structures and avoid bias, as per current recommendations
432 [13].

433

434 The discriminative power of models in identifying high-risk individuals was then assessed
435 using receiver operating curve (ROC) analysis. Area under the ROC (AUC) was calculated
436 for full models (consisting of covariates and PRS) and base models (covariates only). The
437 confidence interval for AUC was calculated using the formula given by Hanley and McNeil
438 [43]. Increment in AUC (ΔAUC) was reported based on the difference between the two
439 models, reported as the discriminative or predictive power conferred by PRS. The
440 permutation test for differences between classifiers was used to estimate the significance (p-
441 value) of an increment in AUC.

442

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

450

451 **DATA AND CODE AVAILABILITY**

452

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

456

457 **ACKNOWLEDGEMENTS**

458 The authors are grateful to the study participants and the staff from the St. Petersburg State
459 Health Care Institution "City Hospital No. 40, Kurortny District" hospital. The authors would
460 like to thank all authors of the included studies for their valuable contributions to data
461 collection. This work was supported by Saint Petersburg State University, project ID:
462 94029859.

463

464 **CONFLICT OF INTEREST**

465

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

469

470 **REFERENCES**

471

- 472 1. Bose S, Adapa S, Aeddula NR, Roy S, Nandikanti D, Vupadhyayula PM, et al. Medical
473 Management of COVID-19: Evidence and Experience. *J Clin Med Res.* 2020;12: 329–
474 343.
- 475 2. Bevova MR, Netesov SV, Aulchenko YS. The New Coronavirus COVID-19 Infection.
476 *Mol Gen Microbiol Virol.* 2020;35: 53–60.
- 477 3. COVID-19 National Preparedness Collaborators. Pandemic preparedness and COVID-
478 19: an exploratory analysis of infection and fatality rates, and contextual factors
479 associated with preparedness in 177 countries, from Jan 1, 2020, to Sept 30, 2021.
480 *Lancet.* 2022;399: 1489–1512.
- 481 4. Biswas M, Rahaman S, Biswas TK, Haque Z, Ibrahim B. Association of Sex, Age, and
482 Comorbidities with Mortality in COVID-19 Patients: A Systematic Review and Meta-
483 Analysis. *Intervirology.* 2020; 1–12.
- 484 5. Wang Y, Wang Y, Chen Y, Qin Q. Unique epidemiological and clinical features of the
485 emerging 2019 novel coronavirus pneumonia (COVID-19) implicate special control
486 measures. *J Med Virol.* 2020;92: 568–576.
- 487 6. Fricke-Galindo I, Falfán-Valencia R. Genetics Insight for COVID-19 Susceptibility and

- 488 Severity: A Review. *Front Immunol.* 2021;12: 622176.
- 489 7. Yousefzadegan S, Rezaei N. Case Report: Death due to COVID-19 in Three Brothers.
490 *Am J Trop Med Hyg.* 2020;102: 1203–1204.
- 491 8. COVID-19 Host Genetics Initiative. The COVID-19 Host Genetics Initiative, a global
492 initiative to elucidate the role of host genetic factors in susceptibility and severity of the
493 SARS-CoV-2 virus pandemic. *Eur J Hum Genet.* 2020;28: 715–718.
- 494 9. Mapping the human genetic architecture of COVID-19. *Nature.* 2021;600: 472–477.
- 495 10. Pairo-Castineira E, Clohisey S, Klaric L, Bretherick AD, Rawlik K, Pasko D, et al.
496 Genetic mechanisms of critical illness in COVID-19. *Nature.* 2021;591: 92–98.
- 497 11. Velavan TP, Pallerla SR, Rüter J, Augustin Y, Kremsner PG, Krishna S, et al. Host
498 genetic factors determining COVID-19 susceptibility and severity. *EBioMedicine.*
499 2021;72: 103629.
- 500 12. Dudbridge F. Power and predictive accuracy of polygenic risk scores. *PLoS Genet.*
501 2013;9: e1003348.
- 502 13. Choi SW, Mak TS-H, O'Reilly PF. Tutorial: a guide to performing polygenic risk score
503 analyses. *Nat Protoc.* 2020;15: 2759–2772.
- 504 14. de los Campos G, Gianola D, Allison DB. Predicting genetic predisposition in humans:
505 the promise of whole-genome markers. *Nat Rev Genet.* 2010;11: 880–886.
- 506 15. Vilhjálmsson BJ, Yang J, Finucane HK, Gusev A, Lindström S, Ripke S, et al. Modeling
507 Linkage Disequilibrium Increases Accuracy of Polygenic Risk Scores. *Am J Hum Genet.*
508 2015;97: 576–592.
- 509 16. Zhu X, Stephens M. BAYESIAN LARGE-SCALE MULTIPLE REGRESSION WITH
510 SUMMARY STATISTICS FROM GENOME-WIDE ASSOCIATION STUDIES. *Ann Appl*
511 *Stat.* 2017;11: 1561–1592.
- 512 17. Khera AV, Chaffin M, Aragam KG, Haas ME, Roselli C, Choi SH, et al. Genome-wide
513 polygenic scores for common diseases identify individuals with risk equivalent to
514 monogenic mutations. *Nat Genet.* 2018;50: 1219–1224.
- 515 18. Choi SW, O'Reilly PF. PRSice-2: Polygenic Risk Score software for biobank-scale data.
516 *Gigascience.* 2019;8. doi:10.1093/gigascience/giz082
- 517 19. Lloyd-Jones LR, Zeng J, Sidorenko J, Yengo L, Moser G, Kemper KE, et al. Improved
518 polygenic prediction by Bayesian multiple regression on summary statistics. *Nat*
519 *Commun.* 2019;10: 5086.
- 520 20. Li R, Chang C, Tanigawa Y, Narasimhan B, Hastie T, Tibshirani R, et al. Fast numerical
521 optimization for genome sequencing data in population biobanks. *Bioinformatics.*
522 2021;37: 4148–4155.
- 523 21. Privé F, Arbel J, Vilhjálmsson BJ. LDpred2: better, faster, stronger. *Bioinformatics.*
524 2021;36: 5424–5431.
- 525 22. Ojavee SE, Kousathanas A, Trejo Banos D, Orliac EJ, Patxot M, Läll K, et al. Genomic
526 architecture and prediction of censored time-to-event phenotypes with a Bayesian
527 genome-wide analysis. *Nat Commun.* 2021;12: 2337.

- 528 23. Wand H, Lambert SA, Tamburro C, Iacocca MA, O'Sullivan JW, Sillari C, et al.
529 Improving reporting standards for polygenic scores in risk prediction studies. *Nature*.
530 2021;591: 211–219.
- 531 24. Lambert SA, Gil L, Jupp S, Ritchie SC, Xu Y, Buniello A, et al. The Polygenic Score
532 Catalog as an open database for reproducibility and systematic evaluation. *Nat Genet*.
533 2021;53: 420–425.
- 534 25. Mavaddat N, Michailidou K, Dennis J, Lush M, Fachal L, Lee A, et al. Polygenic Risk
535 Scores for Prediction of Breast Cancer and Breast Cancer Subtypes. *Am J Hum Genet*.
536 2019;104: 21–34.
- 537 26. Musliner KL, Mortensen PB, McGrath JJ, Suppli NP, Hougaard DM, Bybjerg-Grauholm
538 J, et al. Association of Polygenic Liabilities for Major Depression, Bipolar Disorder, and
539 Schizophrenia With Risk for Depression in the Danish Population. *JAMA Psychiatry*.
540 2019;76: 516–525.
- 541 27. Cupido AJ, Tromp TR, Hovingh GK. The clinical applicability of polygenic risk scores for
542 LDL-cholesterol: considerations, current evidence and future perspectives. *Curr Opin
543 Lipidol*. 2021;32: 112–116.
- 544 28. Lambert SA, Abraham G, Inouye M. Towards clinical utility of polygenic risk scores.
545 *Hum Mol Genet*. 2019;28: R133–R142.
- 546 29. Huang Q-M, Zhang P-D, Li Z-H, Zhou J-M, Liu D, Zhang X-R, et al. Genetic Risk and
547 Chronic Obstructive Pulmonary Disease Independently Predict the Risk of Incident
548 Severe COVID-19. *Ann Am Thorac Soc*. 2022;19: 58–65.
- 549 30. Dite GS, Murphy NM, Allman R. Development and validation of a clinical and genetic
550 model for predicting risk of severe COVID-19. *Epidemiol Infect*. 2021;149: e162.
- 551 31. Dite GS, Murphy NM, Allman R. An integrated clinical and genetic model for predicting
552 risk of severe COVID-19: A population-based case-control study. *PLoS One*. 2021;16:
553 e0247205.
- 554 32. Horowitz JE, Kosmicki JA, Damask A, Sharma D, Roberts GHL, Justice AE, et al.
555 Genome-wide analysis provides genetic evidence that ACE2 influences COVID-19 risk
556 and yields risk scores associated with severe disease. *Nat Genet*. 2022;54: 382–392.
- 557 33. Ahmetov II, Borisov OV, Semenova EA, Andryushchenko ON, Andryushchenko LB,
558 Generozov EV, et al. Team sport, power, and combat athletes are at high genetic risk
559 for coronavirus disease-2019 severity. *J Sport Health Sci*. 2020;9: 430–431.
- 560 34. Farooqi R, Kooner JS, Zhang W. Associations between polygenic risk score and covid-
561 19 susceptibility and severity across ethnic groups: UK Biobank analysis. *BMC Med
562 Genomics*. 2023;16: 150.
- 563 35. Sudlow C, Gallacher J, Allen N, Beral V, Burton P, Danesh J, et al. UK biobank: an
564 open access resource for identifying the causes of a wide range of complex diseases of
565 middle and old age. *PLoS Med*. 2015;12: e1001779.
- 566 36. Chaubey A, Shenoy S, Mathur A, Ma Z, Valencia CA, Reddy Nallamilli BR, et al. Low-
567 Pass Genome Sequencing: Validation and Diagnostic Utility from 409 Clinical Cases of
568 Low-Pass Genome Sequencing for the Detection of Copy Number Variants to Replace
569 Constitutional Microarray. *J Mol Diagn*. 2020;22: 823–840.

- 570 37. Li JH, Mazur CA, Berisa T, Pickrell JK. Low-pass sequencing increases the power of
571 GWAS and decreases measurement error of polygenic risk scores compared to
572 genotyping arrays. *Genome Res.* 2021;31: 529–537.
- 573 38. Rubinacci S, Ribeiro DM, Hofmeister RJ, Delaneau O. Efficient phasing and imputation
574 of low-coverage sequencing data using large reference panels. *Nat Genet.* 2021;53:
575 120–126.
- 576 39. Hui R, D’Atanasio E, Cassidy LM, Scheib CL, Kivisild T. Evaluating genotype imputation
577 pipeline for ultra-low coverage ancient genomes. *Sci Rep.* 2020;10: 18542.
- 578 40. 1000 Genomes Project Consortium, Auton A, Brooks LD, Durbin RM, Garrison EP,
579 Kang HM, et al. A global reference for human genetic variation. *Nature.* 2015;526: 68–
580 74.
- 581 41. Bycroft C, Freeman C, Petkova D, Band G, Elliott LT, Sharp K, et al. The UK Biobank
582 resource with deep phenotyping and genomic data. *Nature.* 2018;562: 203–209.
- 583 42. Jiang L, Zheng Z, Qi T, Kemper KE, Wray NR, Visscher PM, et al. A resource-efficient
584 tool for mixed model association analysis of large-scale data. *Nat Genet.* 2019;51:
585 1749–1755.
- 586 43. Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating
587 characteristic (ROC) curve. *Radiology.* 1982;143: 29–36.
- 588 44. Frieser MJ, Wilson S, Vrieze S. Behavioral impact of return of genetic test results for
589 complex disease: Systematic review and meta-analysis. *Health Psychol.* 2018;37:
590 1134–1144.
- 591 45. Alex Buerkle C, Gompert Z. Population genomics based on low coverage sequencing:
592 how low should we go? *Mol Ecol.* 2013;22: 3028–3035.
- 593 46. Homburger JR, Neben CL, Mishne G, Zhou AY, Kathiresan S, Khera AV. Low coverage
594 whole genome sequencing enables accurate assessment of common variants and
595 calculation of genome-wide polygenic scores. *Genome Med.* 2019;11: 74.
- 596 47. Andrews S. FastQC: A Quality Control Tool for High Throughput Sequence Data. In:
597 FastQC [Internet]. 2010 [cited 2010]. Available:
598 <http://www.bioinformatics.babraham.ac.uk/projects/fastqc/>
- 599 48. Li H, Durbin R. Fast and accurate short read alignment with Burrows–Wheeler
600 transform. *Bioinformatics.* 2009;25: 1754–1760.
- 601 49. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence
602 Alignment/Map format and SAMtools. *Bioinformatics.* 2009;25: 2078–2079.
- 603 50. Manichaikul A, Mychaleckyj JC, Rich SS, Daly K, Sale M, Chen W-M. Robust
604 relationship inference in genome-wide association studies. *Bioinformatics.* 2010;26:
605 2867–2873.
- 606 51. HapMap Project. HapMap Project. In: HapMap 3 [Internet]. Available:
607 <https://www.sanger.ac.uk/data/hapmap-3/>
- 608 52. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, et al. PLINK: a
609 tool set for whole-genome association and population-based linkage analyses. *Am J*
610 *Hum Genet.* 2007;81: 559–575.

- 611 53. Uffelmann E, Huang QQ, Munung NS, de Vries J, Okada Y, Martin AR, et al. Genome-
612 wide association studies. *Nature Reviews Methods Primers*. 2021;1: 1–21.
- 613 54. Team RDC. R: A language and environment for statistical computing. (No Title). 2010
614 [cited 20 Oct 2023]. Available: <https://cir.nii.ac.jp/crid/1370294721063650048>
- 615 55. Van Rossum G, Drake FL. Python 3 Reference Manual; CreateSpace: Scotts Valley,
616 CA, USA, 2009. Google Scholar.
- 617 56. Goel MK, Khanna P, Kishore J. Understanding survival analysis: Kaplan-Meier estimate.
618 *Int J Ayurveda Res*. 2010;1: 274–278.
- 619