Complex patterns of multimorbidity associated with severe COVID-19 and Long COVID
Maik Pietzner ^{1,2,3} , Spiros Denaxas ^{4,5,6,7} , Summaira Yasmeen ¹ , Maria A. Ulmer ⁸ , Tomoko
Nakanishi ² , Matthias Arnold ^{8,9} , Gabi Kastenmüller ⁸ , Harry Hemingway ^{4,5,7*} , Claudia
Langenberg ^{1,2,3,*}
¹ Computational Medicine, Berlin Institute of Health at Charité – Universitätsmedizin Berlin,
Berlin, Germany
² Precision Healthcare University Research Institute, Queen Mary University of London,
London, UK
^a NRC Epidemiology Unit, University of Cambridge, Cambridge, UK
Institute of Health Informatics, University College London, London, UK
³ Health Data Research UK, London, UK
^o British Heart Foundation Data Science Centre, London, UK
'National Institute of Health Research University College London Hospitals Biomedical
Research Centre
⁸ Institute of Computational Biology, Helmholtz Zentrum München - German Research Center
for Environmental Health, Neuherberg, Germany
⁹ Department of Psychiatry and Behavioral Sciences, Duke University, Durham, NC, USA
*these authors jointly supervised the work
Corresponding authors
Dr Maik Pietzner (<u>maik.pietzner@bih-charite.de</u>)
Prof Harry Hemingway (<u>h.hemingway@ucl.ac.uk</u>)
Prof Claudia Langenberg (<u>claudia.langenberg@qmul.ac.uk</u>)

29

NOTE: This preprint reports new research that has not been certified by peer review and should not be used to guide clinical practice.

30 ABSTRACT

31

32 Early evidence that patients with (multiple) pre-existing diseases are at highest risk for severe 33 COVID-19 has been instrumental in the pandemic to allocate critical care resources and later 34 vaccination schemes. However, systematic studies exploring the breadth of medical 35 diagnoses, including common, but non-fatal diseases are scarce, but may help to understand 36 severe COVID-19 among patients at supposedly low risk. Here, we systematically harmonized 37 >12 million primary care and hospitalisation health records from ~500,000 UK Biobank 38 participants into 1448 collated disease terms to systematically identify diseases predisposing 39 to severe COVID-19 (requiring hospitalisation or death) and its post-acute sequalae, Long 40 COVID. We identified a total of 679 diseases associated with an increased risk for severe 41 COVID-19 (n=672) and/or Long COVID (n=72) that spanned almost all clinical specialties and 42 were strongly enriched in clusters of cardio-respiratory and endocrine-renal diseases. For 57 diseases, we established consistent evidence to predispose to severe COVID-19 based on 43 44 survival and genetic susceptibility analyses. This included a possible role of symptoms of malaise and fatigue as a so far largely overlooked risk factor for severe COVID-19. We finally 45 46 observed partially opposing risk estimates at known risk loci for severe COVID-19 for 47 etiologically related diseases, such as post-inflammatory pulmonary fibrosis (e.g., MUC5B, 48 NPNT, and PSMD3) or rheumatoid arthritis (e.g., TYK2), possibly indicating a segregation of 49 disease mechanisms. Our results provide a unique reference that demonstrates how 1) 50 complex co-occurrence of multiple – including non-fatal – conditions predispose to increased COVID-19 severity and 2) how incorporating the whole breadth of medical diagnosis can guide 51 the interpretation of genetic risk loci. 52

54 INTRODUCTION

55 From the outset of the COVID-19 pandemic it was evident that underlying conditions were associated with both the risk of infection with SARS-CoV-2, the cause of COVID-19, and the 56 57 risk of it being severe, based on the risk of hospitalisation, to ventilation and death¹. Initial focus was on the small number of diseases known to put people at higher risk of other 58 59 respiratory viral infections, such as influenza. The Center for Disease Control in the US and other national bodies published lists of diseases associated with COVID-19 and in the UK more 60 than 1 million people were identified as clinically extremely vulnerable and required 61 'shielding' based on having one or more specified diseases². This included older individuals, 62 men, and those with the presence of multiple, pre-exiting long-term conditions, such as 63 64 impaired immunity, type 2 diabetes, hypertension, or chronic kidney disease (CKD)¹.

65 However, the vast body of COVID-19 risk factor studies were based on a candidate approach 66 (e.g., diseases known to be associated with immune compromise), studying common diseases in limited numbers (usually fewer than 100 diseases)^{3–6}. Studies that systematically 67 investigated diseases across clinical specialties, including those primarily managed and 68 69 treated in primary care are largely lacking, but are needed to understand why some patients 70 with COVID-19 suffer from a severe outcome or dead, albeit at supposedly low-risk. Such a 71 systematic, 'diseasome'-wide study can further improve our understanding of how variation in the host genome^{7,8} confers risk for severe COVID-19 and guide drug target prioritisation 72 73 strategies.

74 Here, we collated millions of health records from primary care, hospitalizations and cancer 75 registrations, and death records among ~500,000 participants of the UK Biobank (UKB) into medical diagnosis concept terms, so-called 'phecodes'⁹, to systematically assess the risk for 76 77 severe COVID-19 and its post-acute sequalae, Long COVID, across the breadth of medical 78 diagnosis. Apart from well-recognized high-risk patient groups, such as those with chronic 79 kidney disease or those with compromised immune function, we demonstrate consistent 80 evidence for the possible role of less recognized diseases and symptoms, including malaise and fatigue, based on survival and genetic susceptibility analyses. We finally observed that 81 82 some genomic regions conferring a higher risk for severe COVID-19 might be protective for 83 diseases that partially share pathomechanisms with COVID-19, or vice versa, with possible

84 implications for drug development programs, such as TYK2-inhibitors that may increase the85 risk for severe COVID-19.

86 **RESULTS**

87

88 Here, we systematically investigated the risk conferred by the presence and potential causal relevance of 1448 diseases for COVID-19 severity (hospitalisation, severe respiratory failure, 89 and death) and Long COVID (Fig. 1), based on medical disorder concepts^{10,11} defined and 90 91 collated from >12 million medical records from primary (general practice), secondary care 92 (hospital admissions), and disease registry (cancer registry), death certificates, and patient-93 reported conditions among 502,460 UKB participants (Fig. 1 and Supplemental Tab. 1). Incorporating primary care data more than doubled case numbers for more than half (n=817; 94 95 56.4%) of the diseases considered (Supplemental Tab. 1).

96 97

Disease risk profiles for COVID-19 and Long COVID

98

We collated EHRs up until 01/01/2020 to define pre-existing diseases at any timepoint before and defined severe COVID-19 based on hospital admissions and death certificates up until 31/12/2022 totalling 7,507 (hospitalisation), 662 (respiratory failure), and 1,546 cases (death), with first cases occurring end of January 2020. Due to restricted availability of primary care data, we only included records up until 30/09/2021 to identify 470 cases of Long COVID.

We identified 1,128 significant (p<1.1x10⁻⁵) disease – COVID-19 outcome associations, 104 including almost half (n=679) of the diseases considered with at least one of the four COVID-105 106 19 outcomes derived (Fig. 2 and Supplemental Tab. 2). Pre-existing diseases were almost 107 exclusively associated with a higher risk for COVID-19 endpoints (median hazard ratio (HR): 108 2.39, range: 0.59 - 17.3), only two diseases (benign neoplasm of skin and varicella infection) 109 were associated with a decreased risk. Associated diseases spanned almost all chapters of the 110 ICD-10 (17 out of 18) but were consistently enriched in the chapters 'respiratory' (odds ratio [OR]: 5.96; p-value: 2.7x10⁻⁸), 'circulatory' (OR: 2.95; p-value: 3.5x10⁻⁷), and 111 'endocrine/metabolic' diseases (OR: 2.76; p-value: 9.1x10⁻⁴) when associated with severe 112 113 COVID-19. In contrast, pre-existing disease-codes classified as 'symptoms' were more than 13fold enriched among diseases associated with an increased risk for Long COVID (OR: 13.2; p-114 value: 3.6x10⁻⁸) but also hospitalisation (OR: 5.53; p-value: 9.9x10⁻⁵) and death (OR: 3.06; p-115 116 value: 7.3x10⁻³).

117 For COVID-19 requiring hospitalisation, we replicated and refined known associations with serious pre-existing diseases that have been previously used to identify clinically extremely 118 119 vulnerable people. This included respiratory diseases like pseudomonal pneumonia (HR: 7.53, 95%-CI: 4.74-11.97; p-value<1.2x10⁻¹⁷), acute renal failure (HR: 4.02, 95%-CI: 3.74-4.32, p-120 value: <10⁻³⁰⁰) or type 2 diabetes with renal complications (HR: 7.44; 95%-CI: 5.67 – 9.76; p-121 value: 1.5x10⁻⁴⁷), as well as immune deficiencies (e.g., deficiency of humoral immunity HR: 122 6.02; 95%-CI: 4.36 – 8.31; p-value: 1.3×10^{-27}) or patients under immune suppression (e.g., liver 123 transplants HR: 7.25 95%-CI: 4.51 – 11.68, p-value: 3.4x10⁻¹⁶). However, we further observed 124 125 strong associations with so far less recognized pre-existing mental health and psychiatric 126 diseases and conditions with effect sizes comparable to those previously considered to 127 identify extremely vulnerable people. This included symptoms of malaise and fatigue (HR: 2.17, 95%-CI: 2.07 - 2.27, p-value: 4.4x10⁻²²²) or suicide attempts (HR 5.33, 95%-CI: 4.45 - 6.39, 128 p-value: 3.6x10⁻⁷³). Most diseases (n=641, 95.5%, p_{hetero}>10⁻³) associated with similar 129 130 magnitude across all three different definitions of COVID-19 severity, with different forms of 131 dementias (p_{hetero}<2.1x10⁻²⁴) being among the few exceptions, associating with hospitalisation 132 (HR: 3.83; 95%-CI: 3.38 - 4.34; p-value: 2.3x10⁻⁹⁷) and death (HR: 10.82; 95%-CI: 9.15 - 12.80; 133 p-value: 1.4x10⁻¹⁷⁰), but not severe respiratory failure (HR: 1.15; 95%-CI: 0.51 - 2.57; p-value: 134 0.74) due to COVID-19.

In contrast, pre-existing diseases associated with an increased risk for Long COVID only partially overlapped with those increasing the risk for severe COVID-19. Most notably, we replicated associations with anxiety disorders¹² (HR: 2.59; 95%-CI: 2.09 - 3.20; p-value:1.8x10⁻ 138 ¹⁸) and other mental health symptoms, but most prominently with symptoms of malaise and fatigue (HR: 2.78; 95%-CI: 2.29 - 3.37; p-value:1.5x10⁻²⁵) that are hallmarks of Long COVID and were also strongly associated with severe COVID-19.

Almost all significant associations (99.8%, n=1126) were consistent when considering all-cause 141 142 death as a competing event (Supplementary Tab. 3), and more than half (63.6%; n=718) remained statistically significant (p<4.4x10⁻⁵) when accounting for a large set of potential 143 confounders in multivariable Cox-models (Supplementary Tab. 3). This suggests that 144 145 potentially unreported associations, such as the increased risk for severe COVID-19 among 146 patients reporting symptoms of malaise and fatigue (adjusted HR: 1.66, 95%-CI: 1.58 - 1.74, 147 p-value=7.3x10⁻⁹²), are not just a reflection of a general disease burden or other chronic 148 diseases associated with a greater risk for severe COVID-19.

149 We observed limited evidence for effect modifications by sex (n=7), non-European ancestry (n=1), or age (n=8), but not social deprivation, with 16 disease – COVID 19 pairings showing 150 151 evidence of significant differences (**Supplementary Tab. 4**; p<3.6x10⁻⁶). All included stronger effects in women compared to men, e.g., gout for hospitalised COVID-19 (women: HR: 2.56, 152 95%-Cl 2.21 - 2.96, p-value: 1.3x10⁻³⁶; men: HR: 1.46, 95%-Cl: 1.34 – 1.58, p-value: 2.1x10⁻¹⁹), 153 among Europeans reporting vitamin D deficiencies (Europeans: HR: 2.31, 95%-CI: 2.13 – 2.51, 154 p-value: 2.1x10⁻⁸⁷; non-Europeans: HR: 1.31, 95%-CI: 1.08 – 1.60, p-value=5.5x10⁻³), or among 155 156 younger participants, e.g., disorders of magnesium metabolism and death with COVID-19 as 157 a likely result of renal failure (age ≤65 years: HR: 42.98, 95%-CI: 20.10 – 91.90, p-value: 3.0x10⁻ ²²; age >65 years: HR: 5.35, 95%-CI: 3.51 - 8.16, p-value: 5.9x10⁻¹⁵). 158

159

160 **Complex patterns of multimorbidity are associated with increased risk**

We next derived a disease-disease network¹³ (**Fig. 3A**) to understand, whether the large set of diseases associated with an increased risk for severe COVID-19 act independently or rather reflect an increased risk among participants suffering from multiple pre-existing conditions, i.e., multimorbidity. The network contained a total of 1,381 diseases connected through 5,212 edges based on non-random co-occurrence (**Supplementary Tab. 5a and b**; see **Methods**). Diseases segregated into 31 'communities' being more strongly connected to each other compared to the rest of the network (**Fig. 3B and C**).

Two disease communities were consistently and strongly enriched for diseases associated 168 with severe COVID-19. The first (e.g., OR: 5.20; p-value= 2.2x10⁻¹⁰; for severe respiratory 169 failure) community was strongly enriched for circulatory (OR: 17.6; p-value: 4.4x10⁻³⁹) and 170 171 respiratory (OR: 10.3; p-value: 7.8x10⁻¹⁶) diseases, closely resembling the cardio-respiratory 172 risk profile already described above (Fig. 3B). The second community consisted of diverse endocrine (OR: 6.19; p-value: 1.9x10⁻¹³) and circulatory disease (OR: 3.75; p-value: 5.4x10⁻⁸), 173 174 and largely reflected the renal-diabetic risk profile (Fig. 3C). Accordingly, for each disease acquired during lifetime within the latter disease community, participants' risk increased by 175 18% and 20% to be hospitalised (HR: 1.18; 95%-CI: 1.17 - 1.18; p-value: p<10⁻³⁰⁰) or die with 176 COVID-19 (HR: 1.20; 1.19 - 1.20; p-value<10⁻³⁰⁰), respectively. 177

Diseases increasing the risk for severe COVID-19, but not Long COVID further significantly correlated with hub status (e.g., hospitalisation: r=0.59; p-value: 2.8x10⁻¹²⁴) in the diseasedisease network (**Fig. 3D**), that is, diseases that connect a large cluster of diseases to the rest of the network and might hence be considered as multimorbidity hotspots. For example, acute renal failure, strongly associated with severe COVID-19 (**Fig. 3D**), showed strong partial correlations with 30 other diseases and patients are hence prone to complex multimorbidity. However, the imperfect correlation between hub status and disease-association profiles indicates that certain forms of multimorbidity, such those related secondary malignancies of lymph nodes, are possibly less related to severe COVID-19.

187

188 Convergence of associated disease risk and genetic liability

189 We next systematically characterised whether diseases identified to be associated with 190 COVID-19 severity or Long COVID shared genetic similarity with host genetic susceptibility to 191 severe COVID-19 to understand potential underlying causal mechanisms. We computed 192 genetic correlation estimates for all 1128 disease – COVID-19 outcome pairs (see Methods) and observed 75 pairs (6.6%) that showed evidence for significant ($p<4.4x10^{-5}$) and 193 194 directionally consistent genetic correlations (Fig. 4 and Supplemental. Tab 6), indicating a 195 putatively causal link of any of 57 unique diseases on severe COVID-19. We did not observe 196 evidence of convergence for Long COVID, which might likely be explained by the still low 197 statistical power for the respective genome-wide association study¹⁴.

198 The diseases with consistent evidence from survival and genetic analysis included well-199 described risk-increasing effects of pre-existing endocrine (e.g., type 2 diabetes), respiratory 200 (e.g., respiratory failure), or renal (e.g., chronic kidney disease) diseases, but also digestive 201 (e.g., gastritis and duodenitis), or musculoskeletal (e.g., rheumatoid arthritis) diseases, and further symptoms of malaise and fatigue ($r_G=0.26$; p-value=4.7x10⁻⁶) and abdominal pain 202 203 $(r_{G}=0.33; p=2.5 \times 10^{-11})$, as well as adverse reactions to drugs (e.g., poisoning by antibiotics: 204 r_{G} =0.38; p-value=2.2x10⁻⁶). Findings that collectively demonstrated the need for a 205 comprehensive assessment of disease-risk beyond few, selected common chronic conditions. 206 Among the 41 diseases for which we had sufficient genetic instruments to perform more 207 stringent Mendelian randomization (MR) analyses to assess causality (see Methods), we 208 observed at least nominally significant evidence for gout and hospitalisation (OR: 1.03; 95%-209 CI: 1.01 – 1.05, p-value: 0.03), as well as arthropathy not elsewhere specified (OR: 1.28; 95%-CI: 1.06 – 1.55; p-value: 0.02) and unspecified monoarthrtitis (OR: 1.21; 95%-CI: 1.04 – 1.41; 210 211 p-value: 0.02) for severe COVID-19 (Supplementary Tab. 7). While we might have been still underpowered for many diseases, this leaves the possibility that convergence of survival andgenetic correlation analysis might, in part, be explained by shared risk factors.

214

Evidence for partially opposing roles of shared molecular mechanisms between severe
 COVID-19 and related disorders

217 To finally understand possible molecular mechanisms linking the 'diseasome' to COVID-19, we systematically profiled disease associations across 49 independent genomic regions linked to 218 219 COVID-19 or Long COVID. We observed strong and robust evidence of a genetic signal shared 220 between severe COVID-19 and a total of 33 diseases at nine loci (posterior probability (PP) > 221 80%) (Fig. 5A and Supplemental Tab. 8). Apart from known pleiotropic loci, such as ABO and 222 FUT2 coding for blood group types, this included respiratory risk loci, albeit with contradicting 223 effect estimates for three loci (Fig. 5B). While COVID-19 risk increasing alleles at LZTFL1 and TRIM4 were consistently associated with a higher risk for viral pneumonia and post-224 225 inflammatory pulmonary fibrosis, respectively, risk-increasing alleles at MUC5B, NPNT, and 226 PSMD3 were inversely associated with post-inflammatory pulmonary fibrosis and asthma. An 227 observation that extended even beyond shared loci (Fig. 5C) illustrating a general trend of 228 phenotypic divergence of genetic effects on diseases that share pathological features with 229 severe COVID-19.

A notable observation was the TYK2 locus that has previously been suggested to indicate the 230 efficacy of successfully repurposed drugs for severe COVID-19¹⁵. Briefly, TYK2 encodes for 231 232 tyrosine kinase 2 (TYK2) a protein partially targeted by Janus kinase (JAK) inhibitors like 233 baricitinib, that have been approved for rheumatoid arthritis and successfully repurposed for 234 severe COVID-19, although predating possible evidence from genetic studies^{16–18}. Accordingly, 235 we observed that the same genetic variant, rs34536443 (PP=99.8%), associated with the risk 236 for severe COVID-19 was also associated with, amongst others, the risk of rheumatoid 237 arthritis, but in opposing effect directions (Fig. 5B). Rs34536443 is a loss-of-function missense 238 variant (p.Pro1104Ala) for TYK2 and the functionally impairing minor C allele was associated 239 with a 50% increased risk for severe COVID-19 (odds ratio: 1.50; 95%-CI: 1.40 - 1.62, pvalue=4.3x10⁻²⁹) but a 23% reduced risk for rheumatoid arthritis (odds ratio: 0.77; 95%-CI: 240 0.72 - 0.83; p-value=2.4x10⁻¹²) as well as other autoimmune diseases, in particular psoriasis 241 (Supplementary Tab. 8). While the discrepancy between the success of the drug and genetic 242 inference might be explained by the rather weak affinity of baricitinib for TYK2¹⁹, patients 243

- 244 undergoing trials with TYK2-inhibitors for psoriasis²⁰ might be at an elevated risk for severe
- 245 COVID-19. This observation seemingly aligns with studies on *Tyk2^{-/-}* mouse models reporting
- an impaired immune response to viral infections²¹.

248 **DISCUSSION**

249 An immediate understanding which patients are at greatest risk for severe COVID-19 and 250 possibly death has proven to be instrumental to triage patients early in the pandemic to 251 allocate critical care resources, such as ventilation or extracorporeal membrane oxygenation and, later, vaccination as well. The vast majority of studies^{3–6}, however, focussed on a rather 252 253 narrow set of common, usually chronic, conditions in the risk assessment leaving a 254 considerable number of severe COVID-19 cases unexplained. We demonstrate here how capitalizing on the whole breadth of medical diagnoses through electronic health record 255 256 linkage revealed 1) so far largely neglected patient populations at considerable risk, including 257 those reporting symptoms of malaise and fatigue, and 2) that patients with multiple pre-258 existing conditions, in particular cardio-respiratory and endocrine-renal diseases, are probably 259 at highest risk. Via integration of host genetics, we further provide evidence that a 260 considerable set of diverse diseases may causally drive, or at least share causal drivers with, the risk for severe COVID-19, and exemplify how disease-wide characterisation of specific risk 261 262 loci can inform disease mechanism and derivation of potentially druggable targets or adverse 263 effects.

264 Among the diseases for which we observed consistent evidence from survival and genetic 265 analysis to be linked to severe COVID-19 were multiple examples that have been rarely if at all 266 reported. For example, we observed consistent evidence that symptoms of malaise and 267 fatigue, as well as chronic fatigue, predispose to severe COVID-19. While the vast amount of 268 literature currently discusses or reported these symptoms and disease as characteristics for COVID-19 and its post-acute sequelae^{12,22}, little to nothing is known why patients reporting 269 270 fatigue might be at higher risk. While our definition of 'malaise and fatigue' covered a broad 271 range of partially unspecific medical codes with most cases (n=83,316 out of 87,908, 92.4%) 272 originating from primary care, we observed consistent evidence for the refined diagnosis of 273 chronic fatigue classified as post-viral fatigue symptom (Supplemental Tab. 2). A hypothesis 274 might be, that patients that are already suffering from post-viral symptoms are at a greater 275 risk in general to suffer from more severe courses of viral infections through yet to be 276 identified mechanisms, that may well comprise an altered immune response. However, the 277 evidence we provide does not preclude the existence of general, currently inaccessible, risk 278 factors that predispose to more severe long-term consequences of viral infections.

279 Our extensive genetic analysis revealed some partially contradicting findings that may point to a segregation of overall genetic susceptibility and risk conferred by specific loci and 280 281 mechanisms, replicating and augmenting findings from a previous study in the Million Veterans Study²³. For example, we observed consistent evidence that pre-existing post-282 inflammatory pulmonary fibrosis, likely representing cases of idiopathic pulmonary fibrosis, is 283 a strong risk factor for severe COVID-19 and death, and genome-wide effects were highly 284 correlated between both (r_G=0.45, p=2.3x10⁻⁵), but effects at one of the strongest risk loci for 285 286 post-inflammatory pulmonary fibrosis were protective for severe COVID-19. Our results thereby extend previous observations of misaligning effects at the MUC5B locus and 287 288 idiopathic pulmonary fibrosis^{24,25}. Results that might be explained by a latent, genome-wide 289 risk component (as genome-wide significant loci do not contribute to genetic correlation 290 analysis) that predisposes to severe lung fibrosis irrespective of the exact trigger, and specific 291 molecular pathways characteristic for each disease that differ based on the required immune 292 response to combat the infection. Cell-type and state-specific effects of shared genetic 293 variants or possible design artefacts of GWAS studies of infectious disease, by which certain 294 patient groups are 'underrepresented' due to tailored shielding efforts to minimize viral 295 exposure, are other possible explanations. A similar paradoxical effect at the TYK2 locus 296 highlights the unique potential of integrating electronic health care records with genetic data 297 to guide drug target identification and risk estimation, including emerging diseases and targets 298 in clinical trials.

299 There are a number of limitations that need to be taken in consideration when interpretating 300 our results. Firstly, the COVID-19 pandemic was characterised by strong disruptions of social 301 life and health care, with different waves of new SARS-CoV-2 variants of different 302 pathogenicity, lockdowns, and implementation of vaccines programs, all of which will have 303 influenced the general risk to develop severe COVID-19 for which we could not control for in 304 survival analysis. However, we observed generally little evidence of violation of the 305 proportional hazard assumptions and filtered associations with evidence for strong violations. 306 Secondly, we cannot exclude the possibility that the multitude of diseases associated with 307 severe COVID-19 might also be explained by shared, generic risk factors, such as obesity or 308 smoking, and we implemented sensitivity analysis and comprehensive genetic analysis to 309 mitigate possible confounding, although even larger genetic studies are needed to identify 310 robust genetic signals for diseases like chronic fatigue and other rare diseases that we linked to COVID-19. Thirdly, while we obtained little evidence that disease-risk patterns differ across 311 312 ancestries, the UK Biobank cohort is not a representative sample of the general population 313 and does not sufficiently cover underrepresented populations, e.g., ethnic minorities, and additional work is needed to verify our observations in other populations. Lastly, while our 314 effort to collate and harmonize electronic health records across various sources into medical 315 concept terms covered almost 1,500 diseases, it is still only an approximation of the 316 317 complexity of medical diagnosis and more work, using electronic health records at a national scale, is needed to refine and augment the space of diseases to investigate. 318

Our results demonstrate the unique potential of integrating health records from primary and secondary care with host genetic data to 1) rapidly identify patients at highest risk beyond commonly assessed risk groups, 2) understand pathological pathways, and 3) inform druggable strategies for emerging health threats, such as COVID-19.

323 METHODS

324

325 Study population

326 UK Biobank (UKB) is a prospective cohort study from the UK, which contains more than 500,000 volunteers between 40 and 69 years of age at inclusion. The study design, sample 327 characteristics and genome-wide genotype data have been described in Sudlow et al.²⁶ and 328 Bycroft et al.²⁷. The UKB was approved by the National Research Ethics Service Committee 329 330 Northwest Multi-Centre Haydock and all study procedures were performed in accordance 331 with the World Medical Association Declaration of Helsinki ethical principles for medical research. We included 502,460 individuals who had not withdrawn their consent. For survival 332 333 analysis we considered a set of 438,917 individuals who were still alive at the beginning of the COVID-19 pandemic (01/01/2020) and had genetically inferred ancestry also beyond white 334 335 Europeans. We chose the entire set of white Europeans (n=441,671) that passed standard quality control for genetic analysis to maximise statistical power. 336

337

338 COVID-19 and Long COVID outcome definitions

339 We defined a total of four different COVID-19 related outcomes closely aligned with previous 340 studies^{8,14,28}. We used hospital episode statistics to identify participants who had been 341 'hospitalised' with COVID-19 based on ICD-10 codes U07.1 and U07.2, and the same ICD-10 codes to identify participants who have died from/with COVID-19 based on death registries. 342 343 We did not require a positive PCR COVID-19 test due to differences in local reporting of test 344 results. We adopted a slightly more sophisticated definition for 'severe respiratory failure', demanding a positive COVID-19 test (based on test results released for England, Scotland, and 345 346 Wales provided by UKB through the COVID-19 Second Generation Surveillance System) within 347 a month of acute respiratory failure, defined by ICD-10 codes J80, J96.00, J96.09, Z99.1 from hospital episode statistics or E85.1 and E85.2 when admitted to the intensive care unit. To 348 349 define 'Long COVID we used primary care data released by UKB (covid19_emis_gp_clinical.txt, covid19_tpp_gp_clinical.txt) searching for codes indicating 350 351 suspected diagnosis [CTV3: Y2b89 – "Referral to post-COVID assessment clinic", Y2b8a – 352 "Referral to Your COVID Recovery rehabilitation platform", Y2b87 – "Post-COVID-19 syndrome", and Y2b88 – "Signposting to Your COVID Recovery"; SNOMED-CT: 353 354 1325161000000102 – "Post-COVID-19 syndrome", 1325031000000108 – "Referral to post-COVID assessment clinic", 1325041000000104 – "Newcastle post-COVID syndrome Follow-up 355 356 Screening Questionnaire", 1325181000000106 – "Referral to Your COVID Recovery 357 rehabilitation platform", 1325021000000106 – "Ongoing symptomatic disease caused by severe acute respiratory syndrome coronavirus 2", 1325141000000103 – "Signposting to Your 358 359 COVID Recovery", 1325081000000107 – "Assessment using Post-COVID-19 Functional Status Scale structured interview", 1325061000000103 – "Assessment using COVID-19 Yorkshire 360 361 Rehabilitation Screening tool", 1325071000000105 - "Assessment using Newcastle post-COVID syndrome Follow-up Screening Questionnaire", 1325051000000101 - "COVID-19 362 Yorkshire Rehabilitation Screening tool"]. For each event, we took the earliest record to define 363 364 disease onset.

365

366 Disease ascertainment

We collated electronic health records (EHRs) from primary and secondary care, cancer registries, and death certificates based on tables provided by UK Biobank (gp_clinical.txt, covid19_emis_gp_clinical.txt, covid19_tpp_gp_clinical.txt, hesin_diag.txt, death.txt) downloaded in June 2021. We parsed all records to exclude codes with a recorded date before or within the year of birth of the participant to minimize coding errors from EHRs. We used mappings provided by UK Biobank to include self-reported conditions based on ICD-10 codes. 373 For each data set separately we generated mapping tables that link ICD-10, ICD-9, Read version 2, Clinical Terms Version 3 (CTV3) terms, or SNOMED-CT codes to a set of 1560 374 summarized clinical entities called phecodes^{11,29} (Supplementary Tab. 1). For example, more 375 376 than 90 ICD-10 codes can indicate participants with type 1 diabetes that are here collectively summarized under the phecode 'type 1 diabetes'³⁰. We subsequently fused all data sources 377 based on a common set of phecodes and retained for each participant and each phecode only 378 379 the earliest entry across all EHR resources. We identified a total of 1448 phecodes with at least 380 100 cases in the overall UKB sample. For each participant and phecode, we kept only the earliest date as an indicator for disease onset and defined all events occurring before 381 382 01/01/2020 as prevalent, while we considered any event for genetic analysis. To increase the 383 accessibility of our results, we used the term 'disease' instead of 'phecode' throughout the 384 paper.

385

386 Survival analysis

387 We used Cox-proportional hazard models to estimate the risk associated with each disease 388 and any of the four COVID-19 related outcomes with age as the underlying scale, adjusting for 389 sex (omitted for sex-specific diseases) and genetically inferred ancestry. For each COVID-19 390 outcome, we defined controls separately as all those participants without a corresponding 391 record during the time course of the study. We repeated Cox-proportional hazard models 392 considering all-cause death as a competing event rather than censoring as a sensitivity 393 analysis. We selected 01/03/2020 as the starting point of our study and used 31/12/2022 394 (COVID-19 endpoints) or 30/09/2021 (Long COVID) as endpoints of the observation period 395 depending on the availability of health record linkage. We computed Schoenfeld residuals to 396 test for the proportional hazard assumption, and further computed time varying effects of 397 diseases by introducing 6 months breaks. For each disease – COVID-19 model, we considered 398 all participants that passed inclusion criteria. We applied stringent multiple testing correction (p<0.05/4*1448=4.8x10⁻⁸) and further filtered results for those possibly violating the 399 proportional hazard assumption ($p<10^{-3}$). To establish endpoint-specific associations, we 400 401 performed meta-analysis across disease associations for all three COVID-19 endpoints derived 402 using the R package *metafor* (v.3.8.1). We performed additional sensitivity analysis using an 403 extended set of confounders similar to previous work³¹, including self-reported smoking status 404 and alcohol consumption, body mass index, and Townsend deprivation index (all based on

baseline values), healthcare utilization in the five years before the pandemic (number of stays
and total days in hospital), as well as a variable indicating participants with two or more longterm conditions.

We tested for a potential modifying effect of sex, non-European ancestry, age (\leq 65 years vs >65 years), and social deprivation (Townsend index above median vs below median; median = -2.22) on the results by systematically performing interaction testing, i.e., introducing a disease – sex/non-European ancestry interaction term into Cox-models. For the latter, we requested to have at least 50 observations in each group to ensure model convergence. We subsequently corrected for a total of 13,728 tests (p<3.6x10⁻⁶). All statistical analysis were implemented using R v4.1.2.

415

416 Disease network

We computed a sex-aware disease network using partial correlations as implemented in the 417 R package *ppcor* (v.2.1.1) following previous work¹³. Briefly, partial correlations (r_P) account 418 419 for the fact, that a correlation, or co-occurrence, between two diseases might be driven by a 420 third or any other disease considered. We retained only partial correlations passing stringent 421 multiple testing ($p<4.9x10^{-8}$) and $r_P>0.02$ as we reasoned that a disease-disease network likely 422 exhibits scale-free properties³² with node degrees following a power law. The latter step 423 omitted many significant, but very weak and potentially artificial edges. The final network 424 contained 5212 edges connecting 1381 diseases. We then performed community detection 425 based on the Girvan-Newman algorithm to identify groups of diseases that were more closely 426 connected with each other compared to all other diseases in the network. We finally 427 computed different node characteristics to identify diseases with important roles in the 428 network. We implemented and visualized this analysis with the R package igraph (v.1.3.1).

429

430 Genotyping, quality control, and participant selection

Details on genotyping for UKBB have been reported in detail by Bycroft et al.²⁷. Briefly, we used data from the 'v3' release of UKBB containing the full set of Haplotype Reference Consortium (HRC) and 1000 Genomes imputed variants. We applied recommended sample exclusions by UKBB including low quality control values, sex mismatch, and heterozygosity outliers. We defined a subset of 'white European' ancestry by clustering participants based on the first four genetic principal component derived from the genotyped data using a k-means clustering approach with k=5. We classified all participants who belonged to the largest cluster
and self-identified as of being 'white,' 'British', 'Any other white background', or 'Irish' as 'white
European'. After application of quality control criteria and dropping participants who have
withdrawn their consent, a total of 441,671 UKBB participants were available for analysis with
genotype and phenotype data.

We used only called or imputed genotypes and short insertions/deletions (here commonly referred to as SNPs for simplicity) with a minor allele frequency (MAF) > 0.001%, imputation score >0.4 for common (MAF \ge 0.5%) and >0.9 for rare (MAF<0.5%), within Hardy-Weinberg equilibrium (p_{HWE}>10⁻¹⁵), and minor allele count (MAC) > 10. This left us with 15,519,342 autosomal and X-chromosomal variants for statistical analysis. GRCh37 was used as reference genome assembly.

448

449 *Genome-wide association studies*

450 We performed genome-wide association studies (GWAS) for a total of 1,445 diseases with at 451 least 80 cases (n>100 prior genetic exclusions; 3 diseases dropped out) using REGENIE v2.2.4 452 via a two-step procedure to account for population structure as described in detail 453 elsewhere³³. We used a set of high-quality genotyped variants (MAF>1%, MAC>100, missingness <10%, p_{HWE} >10⁻¹⁵) in the first step for individual trait predictions using the leave 454 455 one chromosome out (LOCO) scheme. These predictions were used in the second step as 456 offset to run logistic regression models with saddle point approximation to account for 457 case/control imbalance and rare variant associations. Each model was adjusted for age, sex, 458 genotyping batch, assessment centre, and the first ten genetic principal components. For 459 diseases reported in only one sex (n=113 in women, n=26 in men), we excluded the respective 460 sex from GWAS to avoid inflation by inappropriate controls. In general, we included all 461 participant with a disease in their records as case and treated all other participants as controls 462 to make best use of the computational efficacy of REGENIE. Testing for reported SNPs showed 463 highly consistent results whether related diseases were included as controls rather than 464 omitted. We used LD-score regression to test for genomic inflation (LDSC v1.0.1)³⁴.

465

466 COVID-19 genetic correlation and Mendelian randomization

467 We downloaded GWAS summary statistics for two different endpoints related to COVID-19

468 (A2 – critical illness; B2 – hospitalisation) and Long COVID (stringent case definition vs broad

control set) provided by the COVID-19 Host Genetics Initiative (release 7)^{8,14}. We used 469 470 summary statistics excluding UKB to avoid sample overlap. We computed genetic correlations as implemented by LD-score regression (LDSC v1.0.1)³⁴ with precomputed LD-scores, 471 excluding the extended MHC region. To test for potentially causal associations of diseases onto 472 473 COVID-19, we used genetic instruments identified in the present study for a total of 41 474 diseases with at least five genetic variants and evidence for significant genetic correlations in a two-sample MR setting. We used MR-PRESSO³⁵ as a first line tool as previously suggested³⁶ 475 476 to account for possible pleiotropy and subsequently report effect estimates from inverse-477 variance weighted analysis as the primary results. We flagged MR results that showed signs 478 of heterogeneity across instruments using Cochran Q statistic. We excluded any variants 479 mapping to the MHC regions for all analysis and implemented MR using the R packages MendelianRandomization (v0.6.0)³⁷ and TwoSampleMR (v0.5.6)³⁸. 480

481

482 Colocalisation at COVID-19 risk loci

483 We collected association statistics for a total of 49 independent risk loci for COVID-19 (selected 484 based on regional clumping (±500kb) of COVID-19 HGI GWAS statistics excluding UKB 485 participants, but SNPs available among imputed genetic data in UKB) across all 1445 diseases 486 included in the genetic analysis. For variant – disease pairings passing a moderate significance threshold (p<10⁻⁶), we implemented statistical colocalization³⁹ accounting for multiple causal 487 488 genetic variants via fine-mapping⁴⁰ using the R packages *coloc* (v.5.3.2) and *susieR* (v.0.11.92). We allowed for a maximum of five causal variants during fine-mapping of the disease and 489 490 linked COVID-19 outcome (via a potentially shared genetic variant) and subsequently tested 491 each credible set for colocalization. We applied a stringent prior to consider a shared signal $(p_{12}=5x10^{-6})$ and further filtered signals with evidence that the lead signal $(r^2$ with best 492 493 remaining signal >0.8) for COVID-19 was dropped from the set of overlapping genetic variants 494 between our UKB GWAS and the COVID-19 GWAS.

496 ACKNOWLEDGEMENT

The authors acknowledge the Scientific Computing of the IT Division at the Charité -497 498 Universitätsmedizin Berlin for providing computational resources that have contributed to the 499 research results reported in this paper (https:// 500 www.charite.de/en/research/research_support_services/research_infrastructure/science_it /#c30646061). This work was supported by funding of the German Centre for Cardiovascular 501 Research (DZHK) and the German Ministry of Education and Research (BMBF), and the 502 503 UKRI/NIHR Strategic Priorities Award in Multimorbidity Research for the Multimorbidity Mechanism and Therapeutics Research Collaborative (MR/V033867/1) to C.L.. H.H. and S.D. 504 505 are supported by Health Data Research UK and the National Institute for Health Research 506 (NIHR) Biomedical Research Centre at University College London NHS Hospitals Trust. M.A. 507 and G.K. are supported by National Institutes of Health/National Institute on Aging grants 508 RF1AG059093, U01AG061359, U19AG063744, R01AG069901, U19AG074879, and 509 R01AG081322. G.K. also received funding from the German Federal Ministry of Education and 510 Research (BMBF) (BiomarKid, 01EA2203B) under the umbrella of the European Joint 511 Programming Initiative "A Healthy Diet for a Healthy Life" (JPI HDHL) and of the ERA-NET 512 Cofund ERA-HDHL (GA N° 696295 of the EU Horizon 2020 Research and Innovation 513 Programme) and of the German Network for Mitochondrial Disorders (mitoNET, 01GM1906C). 514 This work was supported by the de.NBI Cloud within the German Network for Bioinformatics 515 Infrastructure (de.NBI) funded by the German Federal Ministry of Education and Research 516 (BMBF) (031A532B, 031A533A, 031A533B, 031A534A, 031A535A, 031A537A, 031A537B, 517 031A537C, 031A537D, 031A538A).

518

519 Data availability

520 Genome-wide summary statistics for diseases ('phecodes') in UK Biobank were generated, in 521 part, from primary care data released to UK Biobank specifically for the use of COVID-19 522 research only, according to COPI regulations, and can therefore not be made publicly available. 523 All Access to individual level data can be requested by bona fide researchers from the UK 524 Biobank (https://www.ukbiobank.ac.uk/). This research has been conducted under the 525 application 44448. Mapping of Read codes to phecodes can be downloaded from 526 https://github.com/spiros/ukbiobank-read-to-phecode.

528 Code availability

- 529 Associated code and scripts for the analysis will be made available upon publication.
- 530

531 Declaration of Interest

- 532 None of the authors have a conflict of interest.
- 533

534 Author contributions

- 535 Conceptualization: MP, HH, CL
- 536 Data curation/Software: MP, SD, YS, MU, MA
- 537 Formal Analysis: MP, YS
- 538 Methodology: MP, SD, TN
- 539 Visualization: MP
- 540 Funding acquisition: CL, HH
- 541 Project administration: CL, HH
- 542 Supervision: MP, CL, HH
- 543 Writing original draft: MP, CL, HH
- 544 Writing review & editing: SD, SY, MA, GK, TN
- 545

547 FIGURE LEGENDS

548

Figure 1 Outline of the study design. Scheme of the study design and analysis done, illustrating our workflow to define disease mechanisms that may causally contribute to severe COVID-19 or Long COVID. SNPs = single nucleotide polymorphisms; SPA = saddle point approximation; MAF = minor allele frequency; *COVID-19 HGI = COVID-19 Host Genetic Initiative, but excluding contributions from UK Biobank

554

Figure 2 Association results for three different COVID-19 outcomes and long COVID. Each panel contains association statistics, p-values, from Cox-proportional hazard ratios testing for an association between the disease on the x-axis and three different COVID-19 outcomes, as well as Long COVID. Disease associations passing the multiple testing correction (dotted line, p<1.1x10⁻⁵) are depicted by larger triangles of which facing up ones indicate positive, e.g., increased disease risk, associations and downward facing *vice versa*. The diseases are ordered by ICD-10 chapters (colours) and the top ten for each endpoint annotated. The underlying statistics can be found in Supplemental Table 2.

562

Figure 3 Disease-disease network and hub score. A Disease – disease network based on significant (p<4.8x10⁻⁸) positive partial correlations. Nodes (diseases) are coloured by ICD-10 chapters and strength of partial correlation depicted by width of the edges. **B/C** Same network, but only highlighting two disease communities strongly enriched for associations with severe COVID-19. **D** Hub score for the 30 diseases with highest values and associated association statistics, hazard ratios with 95%confidence intervals, from Cox-proportional hazard models. Significant associations are indicated by filled boxes. Colours according to ICD-10 chapters.

570

Figure 4 Convergence of Cox-models and genetic correlations. The first three panels show association statistics, hazard ratios (box) and 95%-confidence interval (lines), for 57 diseases with evidence of convergence with genetic correlation analysis, that are shown in the last two panels (box – genetic correlation; lines – 95%-confidence intervals). Disease have been grouped by ICD-10 chapters and coloured accordingly (see Fig. 2 or 3 for legend).

576

577 Figure 5 Shared genetic architecture at COVID-19 risk loci. A Network representation of significant 578 (PP>80%) colocalization results. Loci are depicted as white rectangles and diseases as coloured nodes 579 according to ICD-10 chapters. Edges represent strong evidence for colocalisation, and solid lines 580 indicate a risk-increasing effect of the COVID-19 risk increasing allele, whereas dashed lines indicate 581 protective effects. B Forest plot displaying hazard ratios with 95%-confidence intervals for each variant 582 and different COVID-19 and colocalising disease outcomes. C Heatmap of effect estimates across 49 583 independent genetic loci associated with increased risk for sever COVID-19 and corresponding effects 584 on six selected traits that showed evidence of colocalization at least one other locus. Black rectangles 585 indicate genome-wide significant effects ($p < 5x10^{-8}$).

587	7 F 5	REFERENCES
589) 1	Merad, M., Blish, C. A., Sallusto, F. & Iwasaki, A. The immunology and immunopathology
590	C	of COVID-19. Science 375 , 1122–1127 (2022).
593	1 2	2. GOV.uk. No Title. https://www.gov.uk/government/news/clinically-extremely-vulnerable-
592	2	receive-updated-guidance-in-line-with-new-national-restrictions.
593	3 3	8. Kompaniyets, L. et al. Underlying Medical Conditions and Severe Illness Among 540,667
594	4	Adults Hospitalized With COVID-19, March 2020-March 2021. Prev. Chronic. Dis. 18, E66
595	5	(2021).
596	5 4	Booth, A. et al. Population risk factors for severe disease and mortality in COVID-19: A
597	7	global systematic review and meta-analysis. <i>PloS One</i> 16 , e0247461 (2021).
598	3 5	. Williamson, E. J. et al. Factors associated with COVID-19-related death using
599	Э	OpenSAFELY. <i>Nature</i> 584 , 430–436 (2020).
600	5 6	5. McKeigue, P. M. et al. Rapid Epidemiological Analysis of Comorbidities and Treatments as
603	1	risk factors for COVID-19 in Scotland (REACT-SCOT): A population-based case-control
602	2	study. <i>PLoS Med.</i> 17 , 1–17 (2020).
603	3 7	7. Pairo-Castineira, E. et al. GWAS and meta-analysis identifies 49 genetic variants
604	4	underlying critical COVID-19. <i>Nature</i> (2023) doi:10.1038/s41586-023-06034-3.
605	5 8	8. COVID-19 Host Genetics Initiative. A second update on mapping the human genetic
606	õ	architecture of COVID-19. <i>Nature</i> 621, E7–E26 (2023).
607	7 9	9. Wei, WQ. et al. Evaluating phecodes, clinical classification software, and ICD-9-CM
608	3	codes for phenome-wide association studies in the electronic health record. (2017)
609	Ð	doi:10.1371/journal.pone.0175508.
610) 1	.0. Bastarache, L. Using Phecodes for Research with the Electronic Health Record: From
61:	1	PheWAS to PheRS. Annu. Rev. Biomed. Data Sci. 4, 1–19 (2021).

- 11. Wu, P. *et al.* Mapping ICD-10 and ICD-10-CM Codes to Phecodes: Workflow Development
- and Initial Evaluation. *JMIR Med. Inform.* **7**, e14325 (2019).
- 614 12. Su, S. et al. Epidemiology, clinical presentation, pathophysiology, and management of
- 615 long COVID: an update. *Mol. Psychiatry* (2023) doi:10.1038/s41380-023-02171-3.
- 616 13. Kuan, V. et al. Identifying and visualising multimorbidity and comorbidity patterns in
- 617 patients in the English National Health Service: a population-based study. *Lancet Digit.*
- 618 *Health* **5**, e16–e27 (2023).
- 619 14. Lammi, V. *et al.* Genome-wide Association Study of Long COVID Authors. 1–25 (2023).
- 620 15. Pairo-Castineira, E. et al. Genetic mechanisms of critical illness in COVID-19. Nature 591,
- 621 92–98 (2021).
- 622 16. Letter, T. M. An EUA for baricitinib (Olumiant) for COVID-19. *Med. Lett. Drugs Ther.* 62,
 623 202–203 (2020).
- 624 17. Kalil, A. C. et al. Baricitinib plus Remdesivir for Hospitalized Adults with Covid-19. N. Engl.
- 625 *J. Med.* **384**, 795–807 (2021).
- 626 18. Hoang, T. N. et al. Baricitinib treatment resolves lower-airway macrophage inflammation
- 627 and neutrophil recruitment in SARS-CoV-2-infected rhesus macaques. Cell 184, 460-
- 628 475.e21 (2021).
- 629 19. Chimalakonda, A. *et al.* Selectivity Profile of the Tyrosine Kinase 2 Inhibitor
- 630 Deucravacitinib Compared with Janus Kinase 1/2/3 Inhibitors. Dermatol. Ther. 11, 1763–
- 631 1776 (2021).
- 632 20. Elyoussfi, S., Rane, S. S., Eyre, S. & Warren, R. B. TYK2 as a novel therapeutic target in
 633 psoriasis. *Expert Rev. Clin. Pharmacol.* 16, 549–558 (2023).
- 634 21. Strobl, B. *et al.* Novel functions of tyrosine kinase 2 in the antiviral defense against
- 635 murine cytomegalovirus. J. Immunol. Baltim. Md 1950 **175**, 4000–8 (2005).

- 636 22. Davis, H. E., McCorkell, L., Vogel, J. M. & Topol, E. J. Long COVID: major findings,
- 637 mechanisms and recommendations. *Nat. Rev. Microbiol.* **21**, 133–146 (2023).
- 638 23. Verma, A. et al. A Phenome-Wide Association Study of genes associated with COVID-19
- 639 severity reveals shared genetics with complex diseases in the Million Veteran Program.
- 640 PLoS Genet. 18, e1010113 (2022).
- 641 24. Fadista, J. et al. Shared genetic etiology between idiopathic pulmonary fibrosis and
- 642 COVID-19 severity. *EBioMedicine* **65**, 103277 (2021).
- 643 25. Verma, A. et al. A MUC5B Gene Polymorphism, rs35705950-T, Confers Protective Effects
- 644 Against COVID-19 Hospitalization but Not Severe Disease or Mortality. Am. J. Respir. Crit.
- 645 *Care Med.* **206**, 1220–1229 (2022).
- 646 26. Sudlow, C. et al. UK biobank: an open access resource for identifying the causes of a
- 647 wide range of complex diseases of middle and old age. *PLoS Med.* **12**, e1001779 (2015).
- 648 27. Bycroft, C. *et al.* The UK Biobank resource with deep phenotyping and genomic data.
- 649 *Nature* **562**, 203–209 (2018).
- 650 28. Thygesen, J. H. et al. COVID-19 trajectories among 57 million adults in England: a cohort
- 651 study using electronic health records. *Lancet Digit. Health* **4**, e542–e557 (2022).
- 652 29. Denaxas, S. Mapping the Read2/CTV3 controlled clinical terminologies to Phecodes in UK
- 653 Biobank primary care electronic health records: implementation and evaluation. *Proc*
- 654 Am. Med. Inform. Assoc. Annu. Symp. 2021 (2021).
- 30. Bastarache, L. Using Phecodes for Research with the Electronic Health Record: From
- 656 PheWAS to PheRS. Annu. Rev. Biomed. Data Sci. 4, 1–19 (2021).
- 657 31. Zang, C. et al. Data-driven analysis to understand long COVID using electronic health
- records from the RECOVER initiative. *Nat. Commun.* **14**, 1948 (2023).

- 659 32. Barabasi, A. & Albert, R. Emergence of scaling in random networks. *Science* **286**, 509–12
- 660 (1999).
- 661 33. Mbatchou, J. et al. Computationally efficient whole-genome regression for quantitative
- 662 and binary traits. *Nat. Genet.* **53**, 1097–1103 (2021).
- 663 34. Bulik-Sullivan, B. K. et al. LD Score regression distinguishes confounding from
- 664 polygenicity in genome-wide association studies. *Nat. Genet.* **47**, 291–5 (2015).
- 665 35. Verbanck, M., Chen, C.-Y., Neale, B. & Do, R. Detection of widespread horizontal
- 666 pleiotropy in causal relationships inferred from Mendelian randomization between
- 667 complex traits and diseases. *Nat. Genet.* **50**, 693–698 (2018).
- 668 36. COVID-19 Host Genetics Initiative. Mapping the human genetic architecture of COVID-
- 669 19. *Nature* **600**, 472–477 (2021).
- 670 37. Yavorska, O. O. & Burgess, S. MendelianRandomization: an R package for performing
- 671 Mendelian randomization analyses using summarized data. Int. J. Epidemiol. 46, 1734–
- 672 1739 (2017).
- 673 38. Hemani, G. et al. The MR-Base platform supports systematic causal inference across the
- 674 human phenome. *eLife* **7**, (2018).
- 675 39. Giambartolomei, C. et al. Bayesian test for colocalisation between pairs of genetic
- association studies using summary statistics. *PLoS Genet.* **10**, e1004383 (2014).
- 40. Wallace, C. A more accurate method for colocalisation analysis allowing for multiple
- 678 causal variants. *PLoS Genet.* **17**, e1009440 (2021).











Α

С

Effect estimate -0.5 1