

1 Limited overlap between genetic effects on disease susceptibility and disease survival

2 Authors

3 Zhiyu Yang¹, Fanny-Dhelia Pajuste², Kristina Zguro³, Yipeng Cheng⁴, Danielle E. Kurant⁵, Andrea Eoli^{6,14},
4 Julian Wanner^{1,6}, Bradley Jermy¹, FinnGen, Estonian Biobank research team, Stavroula Kanoni⁷, David A.
5 van Heel⁸, Genes & Health Research Team⁸, Caroline Hayward⁴, Riccardo E Marioni⁴, Daniel L.
6 McCartney⁴, Alessandra Renieri^{3,9,10}, Simone Furini^{3,10}, Genomics England Research Consortium, Reedik
7 Mägi², Alexander Gusev⁵, Petros Drineas¹², Peristera Paschou¹³, Henrike Heyne^{1,6,14}, Samuli Ripatti^{1,15,16},
8 Nina Mars^{1,15}, Andrea Ganna^{1,15}

9 Affiliations

- 11 1. Institute for Molecular Medicine Finland, FIMM, HiLIFE, University of Helsinki, Helsinki,
12 Finland.
- 13 2. Estonian Genome Centre, Institute of Genomics, University of Tartu, Tartu, Estonia.
- 14 3. Med Biotech Hub and Competence Center, Department of Medical Biotechnologies, University of
15 Siena, Italy.
- 16 4. Centre for Genomic and Experimental Medicine, Institute of Genetics and Cancer, University of
17 Edinburgh, Edinburgh, UK.
- 18 5. Medical Oncology, Dana-Farber Cancer Institute, Harvard Medical School, Boston, MA, USA.
- 19 6. Hasso Plattner Institute, Digital Health Cluster, University of Potsdam, Germany.
- 20 7. William Harvey Research Institute, Barts and the London School of Medicine and Dentistry, Queen
21 Mary University of London, Charterhouse square, EC1M 6BQ, UK.
- 22 8. Blizzard Institute, Barts and The London School of Medicine, Queen Mary University of London,
23 London, UK
- 24 9. Medical Genetics, University of Siena, Siena, Italy.
- 25 10. Genetica Medica, Azienda Ospedaliera Universitaria Senese, Siena, Italy.
- 26 11. Department of Electrical, Electronic and Information Engineering "Guglielmo Marconi",
27 University of Bologna, Cesena (FC) 47521, Italy.
- 28 12. Department of Computer Science, Purdue University, West Lafayette, IN, USA.
- 29 13. Department of Biological Sciences, Purdue University, West Lafayette, IN, USA.
- 30 14. Hasso Plattner Institute, Mount Sinai School of Medicine, NY, US.
- 31 15. Analytic and Translational Genetics Unit, Massachusetts General Hospital, Boston, MA, USA.
- 32 16. Department of Public Health, University of Helsinki, Helsinki, Finland.

33 Abstract

34 Understanding disease progression is of a high biological and clinical interest. Unlike disease susceptibility
35 whose genetic basis has been abundantly studied, less is known about the genetics of disease progression
36 and its overlap with disease susceptibility. Considering ten common diseases (N cases ranging from 17,152
37 to 99,666) across seven biobanks, we systematically compared the genetic architecture of susceptibility and
38 progression, defined as disease-specific mortality. We identified only one locus significantly associated
39 with disease-specific mortality and show that, at a similar sample size, more genome-wide significant loci

40 can be identified in a GWAS of disease susceptibility. Variants that were significantly affecting disease
41 susceptibility were weakly or not associated with disease-specific mortality. Moreover, susceptibility
42 polygenic scores (PGSs) were weak predictor of disease-specific mortality while a PGS for general lifespan
43 was significantly associated with disease-specific mortality for five out of ten diseases. We used theoretical
44 derivation and simulation to propose plausible explanations for our empirical observations and account for
45 potential index-event bias. Overall, our findings point to little similarity in genetic effects between disease
46 susceptibility and disease-specific mortality and suggest that either larger sample sizes or different measures
47 of progression are needed to identify the genetic underpinning of disease progression.

48

49 **Introduction**

50 Genome-wide association studies (GWASs) have been successful in uncovering the genetic basis of human
51 diseases by employing a relatively simple study design that compares diseased individuals with controls
52 (Tcheandjieu et al., 2022; Wightman et al., 2021; H. Zhang et al., 2020). This approach is well suited to
53 identify loci associated with disease susceptibility, but it remains unclear whether these results can also
54 inform on the biology of disease progression. Studying the genetic basis of disease progression is relevant
55 for at least two reasons. First, biological insights from the study of disease progression can be more relevant
56 for drug target discovery since many medicines are developed to cure a disease rather than prevent its
57 occurrence. Second, most individuals approach the healthcare system once they develop a disease or its
58 symptoms, and predicting disease progression is in most diseases an important clinical challenge.

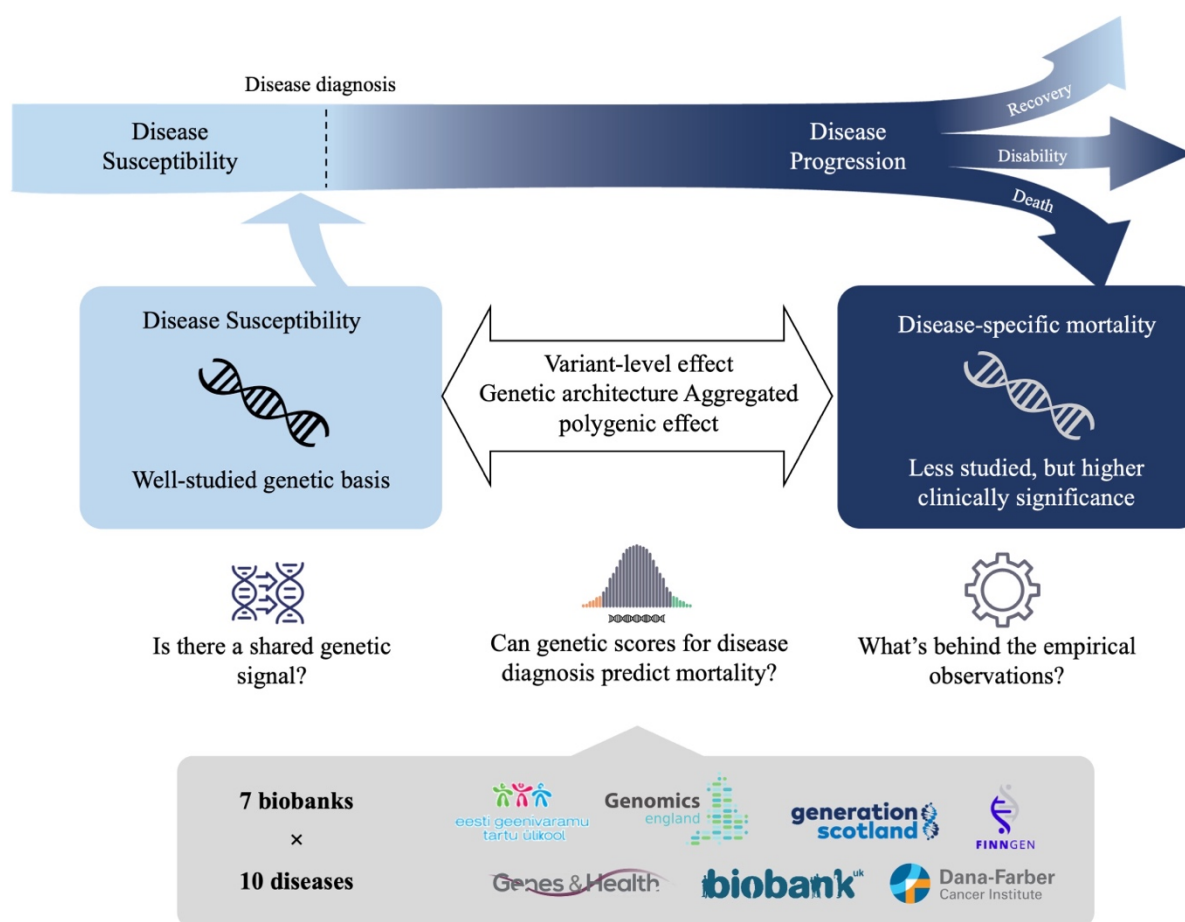
59 In the past years, several GWAS of disease progression have been performed (see **Table S1** for a detailed
60 review), but the number of progression-specific loci discovered has been limited.

61 In cancer, GWAS have focused on disease survival and have been generally unsuccessful in identifying
62 genome-wide significant signals. For example, a GWAS on breast-cancer survival in over 96,000 patients
63 did not identify any robust association (Escala-Garcia et al., 2019) and failed to replicate two loci found in
64 the previous largest GWAS of breast-cancer survival (Guo et al., 2015). Among neurological conditions,
65 GWAS have focused on disease survival as well as cognitive or motor decline. In one of the largest studies,
66 researchers have identified three novel loci associated with Parkinson's disease progression (Tan et al.,
67 2022). A recent study on multiple sclerosis progression has identified a locus pointing to involvement of
68 the central nervous system in disease outcome as opposed to the enrichment for immunological-related
69 signals observed for disease susceptibility (Harroud et al., 2023). However, it is worth noticing that older
70 studies of multiple sclerosis outcomes have failed to replicate in larger ones (Pan et al., 2016; Vandeborgh
71 et al., 2021). In cardiovascular diseases, studies have focused on disease recurrence, and initial results from
72 the GENIUS-CHD consortium showed the strongest GWAS signal for coronary artery disease was not
73 associated with subsequent events (Patel et al., 2019). In Crohn's disease, a study has identified four loci
74 for disease progression, indicating distinct genetic contribution from disease susceptibility (Lee et al.,
75 2017).

76 Apart from single-variant level effects, some studies examined the aggregate effect of many genetic
77 variants. Most of them suggested that polygenic scores (PGSs) for disease susceptibility do not transfer
78 well to disease progression (Barbieux et al., 2019; Lee et al., 2017; G. Liu et al., 2021), although they might
79 outperform other disease-specific biomarkers in the case of cardiovascular diseases (Cho et al., 2023).

80 Some authors have highlighted the challenges in interpreting results from genetic studies of disease
81 progression due to the bias induced when individuals are selected according to disease status. If common

82 causes of susceptibility and progression are not accounted for, association results can be unreliable due to
 83 what is called an index event bias (Yaghootkar et al., 2017) and several approaches to detect and correct
 84 for index event bias have been proposed (Dudbridge et al., 2019; Mahmoud et al., 2022).
 85 Large-scale biobanks linked with longitudinal electronic health records have accelerated the research into
 86 the genetic basis of disease progression and provide sufficient sample size to answer two key questions: 1)
 87 Do genetic predictors that influence disease susceptibility have a similar impact on disease progression? 2)
 88 Can we use PGSs for disease susceptibility to predict patients' disease progression? In this study, we aim
 89 to provide empirical and theoretical answers to these two questions by focusing on a specific, but commonly
 90 used definition of disease progression: disease-specific mortality (Figure 1). Through an international
 91 collaboration across multiple large-scale biobanks, we systematically compared genetic architecture of
 92 disease susceptibility and mortality for ten common diseases focusing on both single variant and aggregated
 93 polygenic effects.



94
 95 **Figure 1.** In this study, using data from seven biobanks, we investigated the genetic similarity between
 96 disease susceptibility and disease progression, defined as disease-specific mortality. We selected ten
 97 diseases and ran GWASs of disease-specific mortality among disease individuals. We then compared the
 98 genetic architecture of disease susceptibility and mortality focusing on both single variant and aggregated
 99 polygenic effects. Furthermore, we attempted to interpret our empirical observations with simulations and
 100 theoretical derivations.
 101

102 **Results**

103 **Participating biobanks and disease of interest**

104 We considered ten common diseases that substantially increase mortality risk in the general population and
 105 have a large public-health impact (**Table 1**). We confirmed disease association with mortality using nation-
 106 wide Finnish data and observed a hazard ratio (HR) for 20-years mortality ranging from 1.31 for Type 2
 107 diabetes to 3.61 for chronic kidney disease in females (Viippola et al., 2023) (**Table S2**). We identified
 108 diseased individuals based on consistent disease definitions captured via electronic health records or
 109 registry data across seven longitudinal studies: FinnGen (Kurki et al., 2023), UK biobank (Bycroft et al.,
 110 2018), Estonia biobank (Leitsalu et al., 2015), Generation Scotland (Smith et al., 2013), Genomics England
 111 (Turnbull, 2018), Genes & Health (Finer et al., 2020), Dana-Farber Cancer Institute (Gusev et al., 2021)
 112 and BioMe. The number of individuals included ranged from 99,666 individuals with type 2 diabetes to
 113 17,152 individuals with Alzheimer disease (**Table 1**). All diseased individuals were followed up for at least
 114 three months, with a maximum follow up of 63.29 years in FinnGen. We defined disease-specific mortality
 115 based on death certificates where the disease of interest was mentioned as primary or secondary cause of
 116 death. One participating biobank did not have information on causes of death, and we used overall death
 117 instead (**Supplementary Material**). We observed the highest cause-specific mortality rate for Alzheimer’s
 118 disease (28.3%) in FinnGen and the lowest for Type 2 diabetes (3%) in Estonia Biobank.

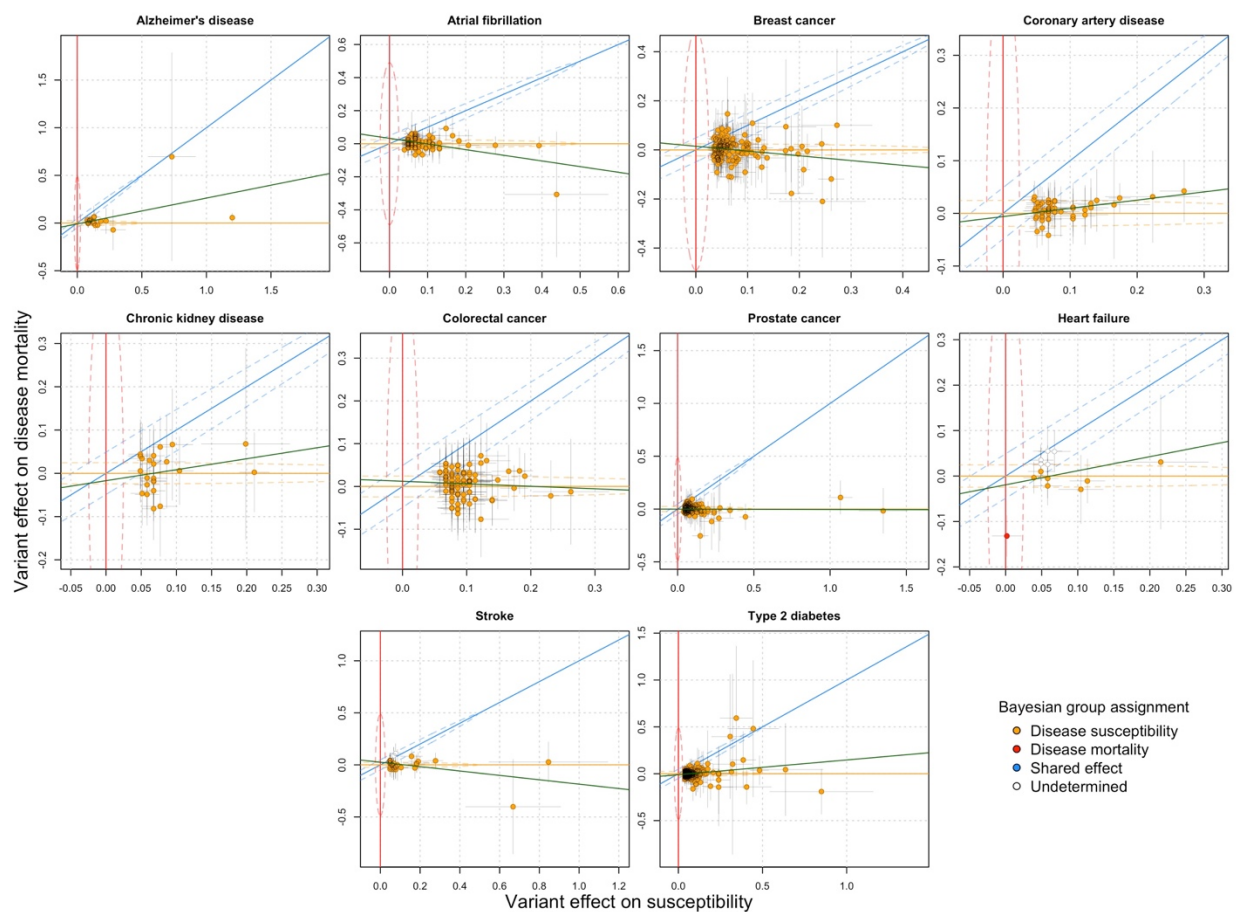
Disease	Sample Size		percentage of death within 2y	percentage of death within 5y	percentage of death within 10y
	N of disease specific deaths	N diseased individuals			
Prostate cancer	3045	27682	3.02 %	6.35 %	9.13 %
Breast cancer	2886	34849	1.70 %	4.50 %	6.59 %
Colorectal cancer	3635	17787	9.69 %	17.13 %	19.73 %
Coronary artery disease	10640	89249	1.74 %	3.86 %	6.79 %
Type 2 diabetes	4886	106405	0.45 %	1.20 %	2.50 %
Atrial fibrillation	3612	85824	0.92 %	1.99 %	3.14 %
Chronic kidney disease	1751	30143	1.98 %	3.94 %	5.47 %
Alzheimer’s disease	4659	17152	5.31 %	15.50 %	24.89 %
Heart failure	1757	35285	1.72 %	2.92 %	4.08 %
Stroke	6757	89435	1.26 %	2.89 %	5.05 %

119 **Table 1.** Total sample sizes for disease-specific mortality GWAS for each disease and percentage of
 120 mortality by year. Also see **Table S2** for details.

121
 122 **Variants affecting disease susceptibility do not have similar effects on disease-specific mortality**

123 For each disease, we carried out a GWAS of disease-specific mortality among disease individuals using
 124 Cox proportional hazard model as implemented in GATE (Dey et al., 2022) or SPACox (Bi et al., 2020)
 125 (**Figure S1-10**). On top of all common GWAS covariates, all analyses were also adjusted for age at disease
 126 diagnosis for two reasons: 1) Patients’ age is strong predictor of one’s mortality; 2) Age of onset has non-
 127 trivial genetic contribution partially overlapping with disease susceptibility (Feng et al., 2020) and we are
 128 instead interested in genetic effects on disease-specific mortality.

129 Out of all ten diseases studied, we only identified one locus associated with disease-specific mortality at
130 genome-wide significant level ($p < 5 \times 10^{-8}$). The locus (rs7360523) on chromosome 20, close to *SULF2*,
131 was associated with disease-specific mortality among patients with heart failure.
132 We asked whether well-established signals for disease susceptibility were associated with disease-specific
133 mortality (**Figure 2**). For each disease, we compared the effect sizes from the largest published GWAS
134 with the result from our GWAS of disease-specific mortality. Using a Bayesian approach (Pirinen, 2023)
135 we could not confidently assign any genetic variant as having the same magnitude of effect on disease
136 susceptibility and disease-specific mortality. In total 888 leading variants were reported from all
137 susceptibility GWAS, whereas none of them was significantly associated with disease-specific mortality
138 after multiple testing correction ($p < 0.05/888 = 5.63 \times 10^{-5}$). Nonetheless, 482 showed the same effect
139 direction, which is marginally more than expected by chance (probability of observing same direction of
140 effect direction 0.54 [95% CI: 0.51 - 0.58], binomial test against 0.5 $p = 0.01$).
141 The only disease-specific mortality locus identified for heart failure also did not show comparable effect
142 on heart failure susceptibility ($p = 0.87$ in susceptibility GWAS with opposite direction of effect). The low
143 number of genome-wide signals for disease-specific mortality were consistent with the lower estimated
144 heritability compared to the GWAS of disease susceptibility (**Table S2**).



145
146 **Figure 2.** Relationship between variant effects (one for each locus) on disease susceptibility (x-axis) and
147 disease-specific mortality (y-axis). Variants were selected either because genome-wide significance for
148 susceptibility in the largest disease specific GWAS or because genome-wide significance for disease-

149 specific mortality in the current study. Only one locus for heart failure mortality was genome-wide
 150 significant. Point colour indicates group assignment for variants (disease susceptibility, disease-specific
 151 mortality, or both). Variants with assignment posterior probability > 0.9 are assigned to the group. Variants
 152 in white indicate assignment posterior probability is < 0.9 for all the three groups. Posterior probabilities
 153 are estimated using R package linemodells (Pirinen, 2023). Red line: $x = 0$; blue line: $y = x$; orange line: y
 154 $= 0$; Green line: linear fit for all independent variants in the plot. Dashed lines represent 95% highest
 155 probability regions for each group. Also see **Table S3-S12** for quantitative results.

156
 157 **Statistical power does not explain the overall lack of genetic signals for disease-specific mortality**

158 To find out if the overall lack of significant genetic signals for disease-specific mortality was simply due
 159 to lower sample size compared to the GWAS of disease susceptibility, we performed a down-sampling
 160 experiment in FinnGen and UKBB by imposing the same effective sample size for both analyses. To further
 161 make the two analyses comparable, the GWAS of disease susceptibility was conducted using survival
 162 analysis with age as time scale and disease diagnosis as outcome. The GWAS of disease susceptibility
 163 returned 30 genome-wide significant loci across all diseases, except colorectal cancer and heart failure,
 164 while the GWAS of disease-specific mortality returned no genome-wide significant results (**Table 2**).

Disease	N. disease-specific deaths	N. diseased individuals	N. of GW-significant loci	
			Disease-specific mortality	Down-sampled disease susceptibility
Prostate cancer	2623	24070	0	8
Breast cancer	1584	27204	0	2
Colorectal cancer	2311	11986	0	0
Coronary artery disease	9067	67051	0	3
Type 2 diabetes	4169	85010	0	5
Atrial fibrillation	2943	71693	0	3
Chronic kidney disease	1283	23744	0	1
Alzheimer's disease	4659	17152	0	7
Heart failure	4203	62717	0 *	0
Stroke	1503	31414	0	1

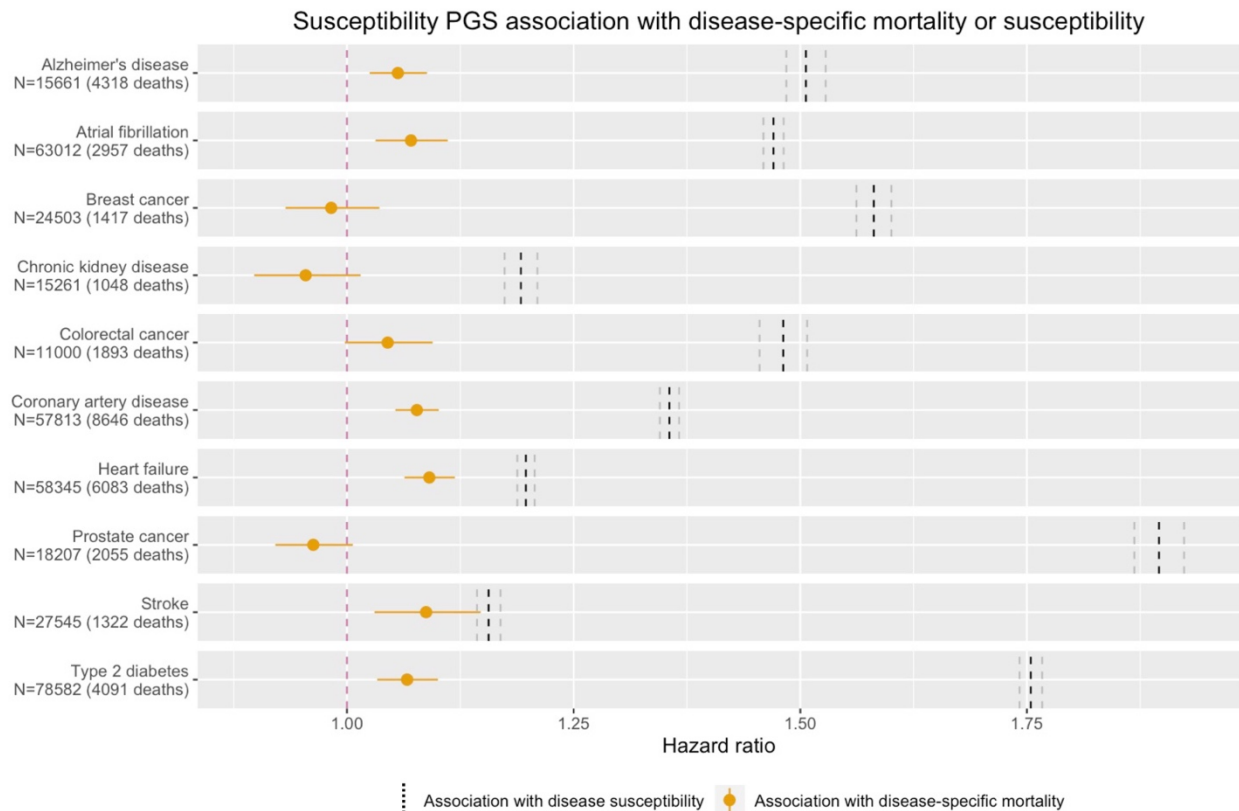
165 **Table 2.** GWAS power comparison between disease-specific mortality and disease susceptibility under the
 166 same sample size and GWAS model in FinnGen and UK biobank. Number of independently associated
 167 genome-wide significant loci. *We report no significant loci for Heart Failure in contrast to what reported in Figure 2 because the
 168 GWAS was conducted only in FinnGen and UK biobank.

169
 170 **Polygenic scores for disease susceptibility are weak predictors of disease-specific mortality**

171 We investigated the joint effects of genetic variants associated with disease susceptibility in predicting
 172 disease-specific mortality. For each disease we constructed a polygenic score (PGS) using results from the
 173 largest GWAS of disease susceptibility. All the PGSs were strongly associated with disease susceptibility.
 174 The hazard ratios (HR) for 1 standard deviation in the PGS ranged from 1.16 [1.14 - 1.17] for stroke to 1.90
 175 [1.87 - 1.92] for prostate cancer (dashed line in **Figure 3**). On the contrary, the same PGSs were weakly or
 176 not associated with disease-specific mortality (orange dots in **Figure 3**). For example, although strongly
 177 associated with disease susceptibility, a PGS for breast cancer showed no association with breast cancer
 178 mortality (HR = 0.98 [0.93-1.04]). The strongest association was observed between the heart failure PGS

179 and heart failure mortality (HR = 1.09 [1.06 - 1.12]), while the PGSs for chronic kidney disease and prostate
180 cancer trend towards having a protective effect on mortality (HR = 0.95 [0.90 - 1.01] and HR = 0.96 [0.92
181 - 1.00], respectively).

182 To understand the robustness of these results, we performed a variety of sensitivity analyses. First, we
183 assessed if using a less specific definition of disease progression, namely all-cause mortality, would impact
184 the observed results. We observed significantly larger correlation coefficients of susceptibility PGSs on
185 disease-specific mortality than on all-cause mortality in five out of ten diseases (**Figure S11, Table S15**).
186 Second, we only considered individuals who developed the disease after study enrollment (**Figure S12A,**
187 **Table S16**) as a way to account for survival bias, which might explain some of the negative associations
188 between PGSs and cause-specific mortality. Nonetheless, results were consistent (correlation coefficient r
189 between effect sizes β in main analysis and sensitivity analysis = 0.94), and we continued observing a
190 negative association between a PGS for prostate cancer and prostate cancer mortality. Third, we considered
191 different maximum follow-up lengths (2, 5 and 10 years) because we reasoned deaths occurring shortly
192 after disease diagnosis were more likely to be caused by the disease. However, results were overall
193 comparable across follow-up lengths (correlation coefficient r between effect sizes in main analysis and
194 sensitivity analysis = 0.68, 0.83 and 0.91 for 2, 5 and 10 years respectively. **S12B, Table S16**) and contrary
195 to our expectation, some diseases (e.g heart failure) showed a stronger association between the
196 susceptibility PGS and disease-specific mortality when considering longer rather than shorter follow-up
197 lengths (effect size $\beta = 9.61 \times 10^{-3}$, 0.03 and 0.06 for 2, 5, 10 years respectively). Finally, we evaluated if
198 adjusting the analyses for age at diagnosis could mask an age-specific effect of PGS on cause-specific
199 mortality, for example because such effect was only observed among young or old patients. We observed
200 a significant difference ($p < 0.005$ under Bonferroni correction) for PGS effect on disease-specific mortality
201 between lower and upper 50% quantile diagnosed age groups only for Alzheimer's disease (**Figure S13,**
202 **Table S17**). The association between Alzheimer disease PGS and mortality was significant only among
203 younger, but not older patient. Finally, we tested the effect using only unrelated individuals in FinnGen
204 and found the result to be robust (**Figure S14, Table S15**). We have also carried out the same analyses
205 using non-European individuals from Genes & Health. However, due to limited power no conclusion could
206 be drawn (**Figure S16**). See **Figure S15** for forest plot of effects from each participant European biobanks.



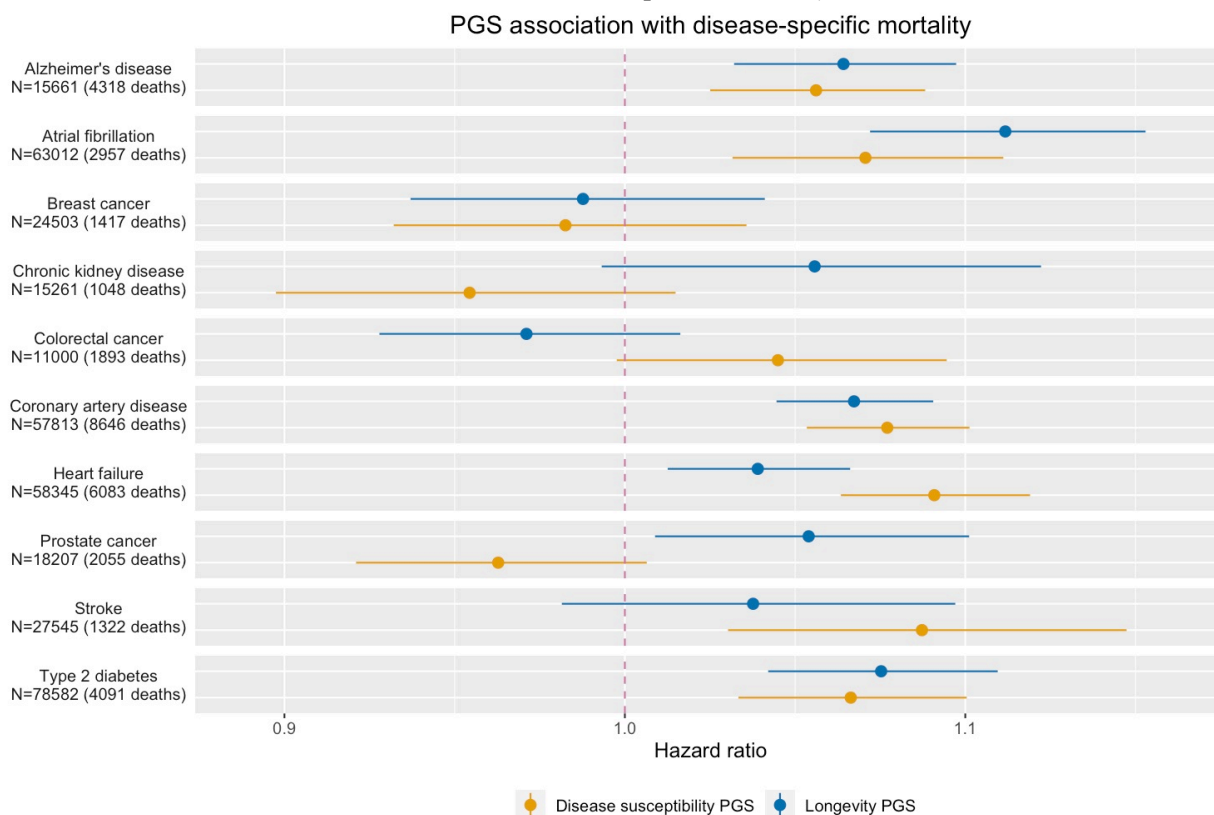
207
 208 **Figure 3.** Association between PGS for disease susceptibility and either disease-specific mortality (orange
 209 dot) or susceptibility (dashed line). Disease susceptibility PGS was derived from published large-scale
 210 GWAS for each disease. PGS associations with both disease susceptibility and disease-specific mortality
 211 were carried out using a Cox proportional hazard model. The sample size reported on the y axis refers to
 212 the disease-specific mortality analyses, the sample size for association with disease susceptibility can be
 213 found in **Table S13**. Horizontal solid lines represent 95% confidence interval (CI). The vertical dashed in
 214 black and grey represent association with disease susceptibility HR and 95% CI respectively.

215
 216 **A polygenic score for longevity was significantly associated with disease-specific mortality for five
 217 out of ten diseases and showed larger effects than the polygenic score for disease susceptibility**

218 Having established that susceptibility PGS are weakly associated with disease-specific mortality, we
 219 reasoned that other PGSs that are better proxies of disease-specific mortality could show stronger
 220 associations. First, we consider PGSs constructed directly from our GWASs of disease-specific mortality.
 221 For diseases where power allowed, we derived PGSs using weights from the meta-analysed GWAS results
 222 from all biobanks except for FinnGen and tested the association between PGS and disease-specific mortality
 223 within FinnGen. Surprisingly, none of the PGSs were associated with disease-specific mortality (**Figure
 224 S17, Table S18**).

225
 226 Second, we considered a PGS for general longevity derived from the largest lifespan (Timmers et al., 2019)
 227 under the assumption that it might capture some of the genetic effects related to disease survival. The
 228 longevity PGS was significantly associated with disease-specific mortality for five out of ten diseases ($p <$

229 0.005 accounting for the number of diseases tested) and it shows larger HR than a PGS for susceptibility
 230 for six out of ten diseases (**Figure 4, Table S14**). For prostate cancer, the association with disease-specific
 231 mortality was significantly larger for the longevity than the susceptibility PGS (HR = 1.05 [1.01 - 1.10] vs
 232 HR = 0.96 [0.92 - 1.01], t -test on effect size differences $p = 4.41 \times 10^{-3}$).



233
 234 **Figure 4.** Association between a PGS for disease susceptibility (orange dots) and longevity (blue dots) with
 235 disease-specific mortality. Disease susceptibility PGSs were derived from published large-scale GWAS for
 236 each disease. Also see **Table S14** for quantitative results. Longevity PGS was derived from (Timmers et
 237 al., 2019). Horizontal solid lines represent 95% CI.

238
 239 **Theoretical framework and results from simulation suggests the observed results are consistent with**
 240 **low heritability of disease-specific mortality and modest index event bias effect**

241 Towards better understanding of reasons behind our empirical observations, we proposed a simple
 242 framework to study the genetic effects on disease susceptibility and progression. We defined the liability
 243 to disease susceptibility under a polygenic risk model as random variable S

$$S = \beta_S^T g + \epsilon_S$$

Total genetic effect on susceptibility (points to $\beta_S^T g$)
Vector of individual genotype (points to g)
Environmental noise on susceptibility (points to ϵ_S)
Variant effect on susceptibility (points to β_S)

244
 245
 246
 247
 248
 249 where g is the random vector for standardised genotype and β_{g_i} is the random vector of their effect sizes
 250 on the diagnosis liability, ϵ_S is the zero mean residual independent to $\beta_S^T g$.

251 Next, we defined liability to disease progression as a random variable P which depends both on the causal
 252 effect of diseases susceptibility (c) and some unique genetic effect on disease progression ($\beta_p^T g$)

253

254

255

256

257

258

$$P = \beta_p^T g + cS + \epsilon_P$$

259 We define the heritability of disease susceptibility (h_{sus}) and *unique genetic components* of disease
 260 progression (h_{prog}) as:

261

262

263

$$h_{sus} = \frac{Var(\beta_s^T g)}{Var(S)}$$

$$h_{prog} = \frac{Var(\beta_p^T g)}{Var(P)}$$

264 Last, we define ρ as the correlation of the polygenic effect between disease susceptibility and disease
 265 progression:

266

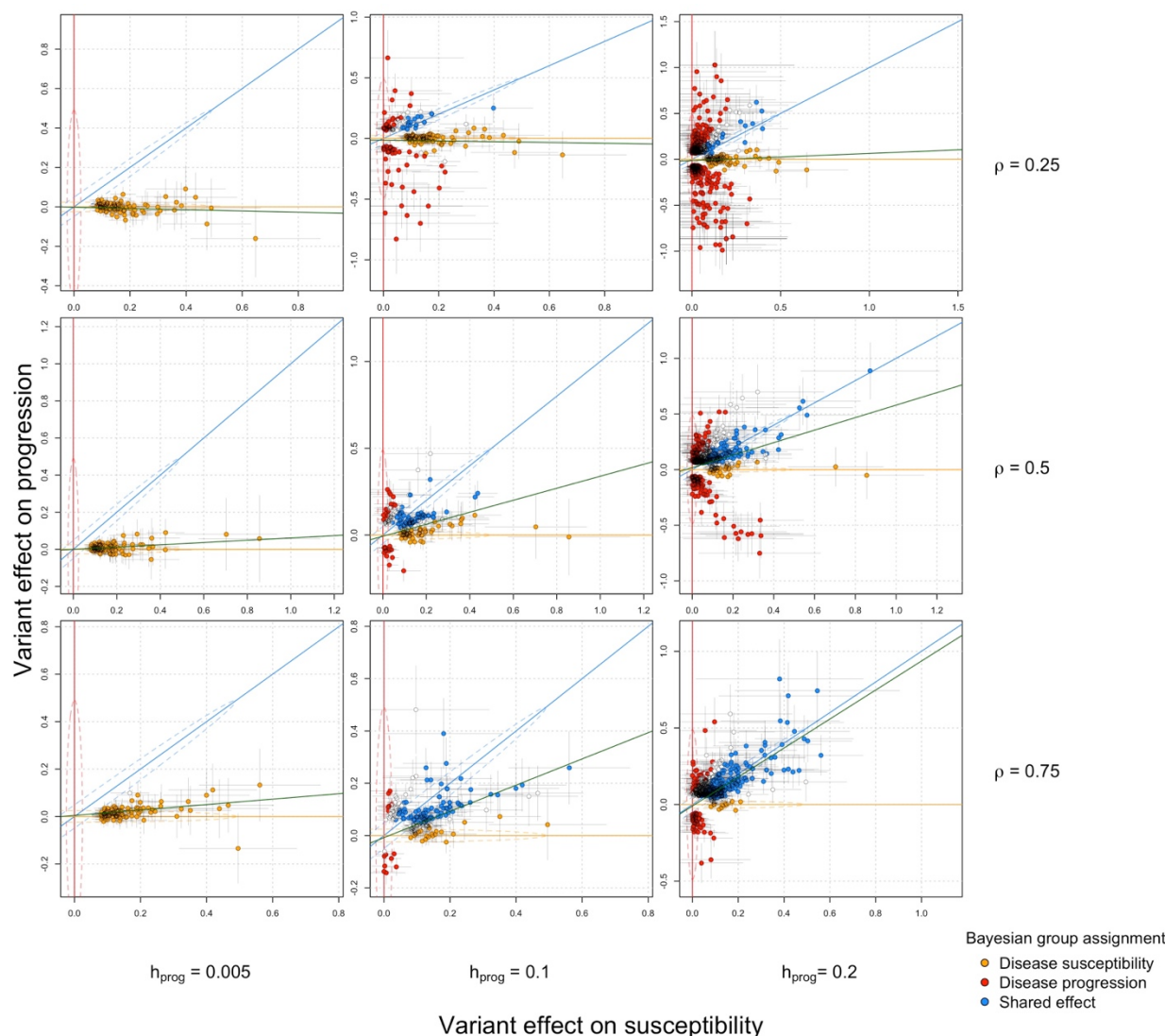
$$\rho = \frac{Cov(\beta_p^T g, \beta_s^T g)}{\sqrt{Var(\beta_p^T g)Var(\beta_s^T g)}}$$

267 First, we performed simulations by varying two main parameters: h_{prog} and ρ and compared the theoretical
 268 results with the empirical results of **Figure 2**. The empirical results match a scenario with very low
 269 heritability of disease progression (**Figure 5**) regardless of the correlation of genetic effects between
 270 susceptibility and progression.

271 Second, we derived the theoretical heritability of disease progression (h_{prog}) when measured within
 272 patients (**Supplementary material**) and showed a non-monotonic relationship with the genetic correlation
 273 (ρ) between disease susceptibility and progression (**Figure S18A**). We also derived how much a polygenic
 274 signal for susceptibility (e.g. PGS) can explain the variability in disease progression by calculating
 275 theoretical coefficient of determination (R^2) and showing, as expected, that it increases with ρ (**Figure**
 276 **S18B**). These two results indicate that high correlation in the polygenic effect between disease susceptibility
 277 and disease progression can decrease the heritability of disease progression while making the polygenic
 278 signals for disease susceptibility a stronger predictor of progression. Finally, we derived the heritability of
 279 disease progression outside the within-patient population (**Supplementary material**). This is relevant if
 280 one believes that traits measured in the general population (e.g. longevity) can be a proxy of disease
 281 progression.

282 Third, we explored the impact of index event bias by introducing a shared non-genetic risk factor accounting
 283 for various proportions of the liability in disease susceptibility and progression. We compared the simulated
 284 effect of causal genetic variants on progression with the observed effect from the progression GWAS and
 285 found larger differences when the shared non-genetic component accounted for higher liability variance,
 286 indicating higher impact of index event bias (**Figure S19**). A correction approach similar to slope-hunter
 287 (Mahmoud et al., 2022) reduced the bias improving the concordance with the true simulated effects.
 288 However, in the scenario of a low progression heritability (h_{prog}), which is consistent with our empirical

289 findings, index event bias correction showed a limited impact as we observed no genetic variants
 290 significantly associated with disease progression before or after bias correction (**Table S19**). Furthermore,
 291 we do not see the impact of this correction on posterior variant classification (**Figure S20**).



292
 293 **Figure 5.** Relationship between variant effects (one for each locus) on disease susceptibility (x-axis) and
 294 disease-specific mortality (y-axis) in simulations when varying the heritability of disease progression
 295 (h_{prog}) and the correlation of the polygenic effect between disease susceptibility and disease progression
 296 (ρ). Variants were plotted if they were genome-wide significant for susceptibility or progression. Point
 297 colour indicates group assignment for variants (susceptibility-specific, mortality-specific or both). Variants
 298 with assignment posterior probability > 0.9 are assigned to the group. Variants in white indicate assignment
 299 posterior probability is < 0.9 for all the three groups. Posterior probabilities are estimated using R package
 300 linemodels (Pirinen, 2023). Red line: $x = 0$; blue line: $y = x$; orange line: $y = 0$; Green line: linear fit for all
 301 independent variants in the plot. Dashed lines represent 95% highest probability regions for each group.
 302

303 **Discussion**

304 In this study we systematically explored the overlap of genetic effects on disease susceptibility and a
305 common measure of disease progression, disease-specific mortality, for 10 common diseases. By
306 conducting the largest within-patient GWAS of disease-specific mortality to date we found: 1) leading
307 variants affecting disease susceptibility do not have comparable effect sizes on disease mortality. Rather,
308 they show little effect and no significant association with disease-specific mortality in GWAS ; 2) at a
309 similar sample size, GWAS of disease-specific mortality identified fewer genome-wide significant loci than
310 GWAS of disease susceptibility, suggesting that GWAS of disease progression might require larger sample
311 size or more refined phenotypes than GWAS of disease susceptibility; 3) disease susceptibility PGSs do
312 not transfer well on disease-specific mortality, suggesting that current PGSs are more suitable for identify
313 individuals at high risk of developing a disease rather than those more likely to suffer from the worst
314 consequences. Given the interest in using PGS for optimising clinical trials (Fahed et al., 2022), our results
315 suggest PGS for disease susceptibility might not be the best choice if the trial main outcome is related to
316 disease progression.

317 Why do we observe limited overlap in genetic effects on disease susceptibility and disease progression?
318 There might be several explanations.

319 First, genetic influences on disease progression might be too small to detect. External environmental effects
320 such as treatment choice, treatment response, quality and access to care might have a disproportionate
321 impact on disease progression as compared to disease susceptibility, thus limiting the genetic influence.
322 Heterogeneity in patients and their treatments plays a big role in progression for many diseases and we are
323 not currently able to adjust for all that heterogeneity. Using data from clinical trials rather than observational
324 studies and including finer measurements, such as disease relevant biomarkers, can obviate these
325 shortcomings. We also notice that adjusting for age at disease diagnosis does reduce the overlap between
326 susceptibility and progression because variants increasing disease susceptibility are often associated with
327 earlier disease diagnosis (Feng et al., 2020). Previous studies have demonstrated impact of adjusting for
328 age in disease progression analyses (Houlahan et al., 2023) and suggested that association between PGS
329 and measurement of disease progression may be mediated by age.

330 Second, our definition of disease progression might be a poor proxy for the biological mechanisms
331 impacting disease progression. Our approach aims to compare progression across multiple diseases and
332 comes at expenses of a tailored definition of progression for each disease. Nonetheless, disease-specific
333 mortality has been widely used as a measure of progression (Hernesniemi, 2022; Jabbari et al., 2021; Tan
334 et al., 2022; Wu et al., 2014). In practice, biobank-based studies of disease progression often need to
335 converge to simple definitions to maximise sample size, and disease-specific mortality is an information
336 typically available across biobanks. However, a definition that permits the high data availability might not
337 be one that best reflects the genetic aetiology of a specific disease.

338 Third, as a common concern for all studies on disease progression, we explored the impact of index event
339 bias resulting from conditioning on diseased individuals. While this is not the main focus of this study, we
340 found that index event bias by itself does not fully explain the lack of concordance between genetic effects
341 on susceptibility and progression observed in our study. Our empirical observations in comparison to
342 various simulations indicate a relatively low heritability for disease progression, defined as disease-specific
343 mortality. Furthermore, heterogenous phenotypes like mortality, although constrained to be disease-
344 specific, can be highly polygenic. In this case, even a perfect correction for index event bias will only be
345 able to recover effect sizes that are not likely to be detected from a progression GWAS. The fact that only

346 one genome-wide locus was picked up from our progression GWAS, indicating low signal-to-noise ratio
347 in the progression GWAS, might be a bigger concern than index event bias. Furthermore, most methods to
348 correct for index-event bias rely on fitting the relationships between variant effects on susceptibility and
349 progression. In our case, this relationship is close to zero and thus the correction will be small and
350 insignificant.

351 Apart from the impact of index event bias, our theoretical framework reveals some other interesting
352 expectations. The heritability of disease progression is not monotonically increasing with the genetic
353 similarity between susceptibility which implies that the variance in the progression phenotype decreases
354 when disease susceptibility and progression get more genetically similar. This suggests that discovering
355 genetic signals that are specific to disease progression requires a fine balance in the genetic similarity
356 between disease susceptibility and progression and data availability. A simplified definition of disease
357 progression might be useful to increase the sample size but might not be relevant to capture genetic variants
358 specific to susceptibility. On the contrary, when disease susceptibility and progression highly overlap, we
359 can expect homogeneity in patients' progression, which reduces effective sample size. Once the similarity
360 between disease susceptibility and progression reaches a certain level, it might not be necessary carrying
361 out a GWAS of progression, since a susceptibility GWAS might already capture sufficient information to
362 infer the genetic bases of progression.

363 Given the aforementioned challenges when conducting GWAS of disease progression among diseased
364 individuals, one attractive alternative would be to study genetic signals for disease progression in a general
365 population and subsequently adapted for within-patients prognostic prediction. For example, PGS for
366 autoimmune conditions derived in the general population are correlated with immune-related adverse
367 events among cancer patients treated with immune checkpoint inhibitors (Groha et al., 2022; Khan et al.,
368 2020, 2021). In our analysis, a longevity PGS derived from GWAS of lifespan was significantly associated
369 with patients' survival for five out of ten diseases, suggesting patients' survival could be more affected by
370 general factors related to mortality than disease-specific factors. Methods for cross-trait PGS (Kember et
371 al., 2021) might be leveraged to obtain progression PGS based on existing GWAS results in the general
372 population.

373 The study has multiple limitations. First, while we explored the similarity in genetic effects between disease
374 susceptibility and disease-specific mortality, we cannot decisively conclude the biological underpinnings
375 to susceptibility and progression are distinct. For example, a phenotype that serves as poor proxy for disease
376 progression will result in attenuated effect sizes, despite genetic variants being causally associated with
377 both susceptibility and progression. Nonetheless the poor replication rate and opposite direction of effect
378 observed for susceptibility signals on disease-specific mortality are consistent with a scenario where at least
379 some variants have no shared effect on both susceptibility and progression. Second, our findings do not
380 necessarily extend outside the diseases explored in this study and further work is needed to confirm the
381 observed trends across more disease categories. Third, whether death certificates are accurately enough
382 capturing primary or contributing causes of death depends on the biobank and healthcare system. We tried
383 to address these concerns by restricting the follow-up duration in sensitivity analyses, reasoning that deaths
384 occurring shortly after disease diagnosis were more likely to be caused by the disease. Fourth, our
385 theoretical framework mostly focuses on the generic relationship between disease susceptibility and
386 progression and does not take the impact of non-genetic factors into account. Also, for simplicity we did
387 not use time-to-event model in this work, which might be more relatable to our empirical experiments.

388 In conclusion, our current results suggest there is a limited overlap in genetic effects on disease
389 susceptibility and progression, defined as patients' mortality. Further refinement in inclusion criteria among
390 the patient population and in the definitions of disease progressions can be considered in future studies to
391 robustly identify the genetic underpinning of disease progression.

392

393 **Methods (details)**

394 **Selection of diseases**

395 We selected ten common complex diseases spanning various disease categories for the analyses. The
396 diseases are selected to meet following criteria: 1. Have high epidemiological hazard ratio on mortality, so
397 that mortality can be viewed as a reasonable prognosis; 2. Constitute high global disease burden in terms
398 of disability adjusted life years (DALYs) (Abbas et al., 2020); 3. Relatively common (> 1% prevalence)
399 in population and have reasonable patient bodies in all biobanks; 4. Heritable and have large scale GWAS
400 available to construct PGS. All disease endpoints were defined as a composition of ICD-10 codes curated
401 by the clinical expert groups from FinnGen, Institute for Molecular Medicine Finland (FIMM) and Finnish
402 Institute for Health and Welfare (THL) (Kurki et al., 2023). Same disease definitions, in terms of ICD-10
403 codes, were adopted by all participating biobanks to the maximum possible extent. See **Table S2** for list of
404 disease and relevant descriptive statistics.

405

406 **Progression definition**

407 For all selected diseases, we defined mortality as our outcome. Precisely, we were interested in both *all-*
408 *cause mortalities*, namely simple death status of the patient regardless of relevance to the disease, and
409 *disease-specific mortalities*, meaning the death caused directly or indirectly by disease of interest
410 specifically. Disease progression was evaluated as patients' survival from each type of mortality after being
411 diagnosed with the disease. For all mortality GWASs, we consider only disease-specific mortality whenever
412 it is possible for each participating biobank. Whereas for the PGS analysis, both all-cause and disease-
413 specific mortalities were evaluated. Same as the disease endpoints, cause of death linked to each disease
414 was also curated by clinical expert groups and defined in terms of ICD 10 codes (World Health
415 Organization, 2004). The same definitions were systematically applied to all biobanks to the possible extent.
416 See **Table S2** for definition of cause-specific mortality for each disease of interest and available sample
417 sizes from each biobank.

418

419 **Within-patient mortality GWAS**

420 To achieve variant level effect comparison, for each selected disease, within-patient mortality GWAS was
421 carried out using GATE (Dey et al., 2022) for all biobanks but Generation Scotland, which used SPACox
422 (Bi et al., 2020) as an alternative. The event of interest in this GWAS was patients' survival after disease
423 diagnosis. For each disease of interest, GWAS was carried out separately within each ancestry group for
424 biobanks that have a cause-specific mortality event count of 50 at minimum after quality control. Eligible
425 individuals were restricted to patients having a follow-up time after diagnosis of three months (0.25 years)
426 at minimum. We used model below to examine SNP association with patients' survival:

427 $\text{surv}(\text{duration of follow up after diagnosis} \mid \text{disease-specific mortality}) \sim \text{SNP} + \text{patient's age of diagnosis}$
428 $+ \text{patient's birth year} + \text{sex} + \text{PCs} + \text{study specific covariates},$

429 where study specific covariates included other available non-heritable biobank specific covariates, such as
430 genotyping chip or batch.

431 For analyses in the UK biobank, to minimise potential impact of survivor bias, only patients with disease
432 diagnosed after enrollment were considered.

433

434 **Results quality control and meta-analysis**

435 After mortality GWAS for selected diseases were carried out within each contributed biobank, we then
436 filtered the resulting summary statistics by imputation INFO scores and minor allele counts. We kept only
437 variants showing an imputation INFO score > 0.7 and having at least 20 minor allele counts for each
438 summary statistics. For GWAS summary statistics with a different human genome build, we used the UCSC
439 LiftOver tool (Kuhn et al., 2013) to convert their genome coordinates into hg38 assembly. Subsequently,
440 for each disease, we meta-analysed GWAS results from each biobank using fixed-effect meta-analysis
441 implemented in METAL (Willer et al., 2010). With which, we also scanned for heterogeneity in effect sizes
442 across different biobanks using Cochran's Q test. We applied an inverse variance weighted meta-analysis
443 scheme whenever possible. However, since SPACox does not have effect size or standard error output, in
444 Generation Scotland, we estimated direction of effect under a logistic regression model using plink (Purcell
445 et al., 2007), and subsequently proceeded with a sample-size weighted meta-analysis using the z-scores.
446 This was done for four out of the ten diseases, for which Generation Scotland was one of the data sources:
447 atrial fibrillation, breast cancer, coronary artery disease and type 2 diabetes.

448

449 **Variant level effect size comparison**

450 We compared our mortality GWAS results for each disease of interest with large-scale published GWAS
451 on diagnosis of the same disease. For disease diagnosis GWAS, we extracted SNP effects of reported
452 genome-wide significant leading SNPs at independently associated loci from each study. For CKD, a large
453 GWAS on estimated glomerular filtration rate (eGFR) was considered (Wuttke et al., 2019). Specifically,
454 we looked at independent leading SNPs' effect sizes on binary CKD diagnosis reported from the study so
455 that the scale of measurement is more comparable. For our meta-analysed mortality GWAS, we identify
456 independent genome-wide loci using summary statistics based conditional analysis implemented in GCTA-
457 COJO (Yang et al., 2012). We merged 5,000 Finnish genomes, which is one of the largest GWAS cohorts
458 in this study, with EUR from 1000 Genome as LD reference for this step. Subsequently, for each leading
459 SNP from diagnosis or mortality GWAS, we classify them by linear relationships between their effect sizes
460 in the diagnosis and mortality GWAS into three groups: disease diagnosis specific (slope = 0), disease
461 mortality specific (slope = *inf*) and variants with effect on both diagnosis and mortality (slope = 1). The
462 classification was carried out using a Bayesian framework implemented in R package linemodels (Pirinen,
463 2023).

464

465 **Comparison of genetic architectures for disease diagnosis and mortality**

466 We compared genetic architectures between disease diagnosis and mortality in terms of SNP heritability
467 estimated from the meta-analysed mortality GWAS summary statistics using LD score regression (Bulik-
468 Sullivan et al., 2015). For eligible traits, i.e. traits with non-zero estimated SNP heritability, we further
469 analysed genetic correlation across disease diagnosis, mortality, and general longevity GWAS using the
470 same tool.

471

472 **Down-sampled GWAS on age of diagnosis**

473 To ensure heritability comparison between disease susceptibility and progression endpoints not being
474 subject to power issues resulted from difference in sample sizes and GWAS models, for each disease of
475 interest, we also ran time-to-event GWAS to find SNP association with age of diagnosis using a randomly
476 down-sampled cohort which had comparable number of total individuals and event counts as what was
477 available for the within-patient mortality GWAS. The down-sampled GWAS was carried out under model
478 below:

479
$$\text{surv}(\text{follow-up from birth until diagnosis} \mid \text{disease diagnosis}) \sim \text{SNP} + \text{patient's birth year}$$

480
$$+ \text{sex} + \text{PCs} + \text{study specific covariates}.$$

481 This analysis was also carried out using GATE (Dey et al., 2022) but in FinnGen and UKBB only, which
482 are two of the largest participating biobanks in this study (See **Table S2** for biobanks sample sizes).

483

484 **Computation of individual level PGS**

485 For each selected disease, we derived variant weights for PGS from GWAS summary statistics listed in
486 **Table S2** using MegaPRS (Q. Zhang et al., 2021). Heritability contributed by each variant was estimated
487 under the BLD-LDAK model as recommended. For weight estimation, we used the “mega” option which
488 leaves it to the software to decide the most appropriate model given the data. Since we studied mortality,
489 apart from the ten selected diseases, we also computed PGS weights for general longevity using the largest
490 GWAS on lifespan (Timmers et al., 2019). Due to the heterogeneous and polygenic nature of lifespan, for
491 this trait, we used the LDAK-Thin model for SNP level heritability estimation instead. Unlike the BLD-
492 LDAK model used in variant weighting for other diseases, LDAK-Thin model does not take functional
493 annotations into account but estimates SNP heritability only as functions of SNP allele frequencies and
494 local linkage structures. Variant weights were derived for 1,330,820 common SNPs (minor allele
495 frequency > 0.1) lying in the intersection of HapMap3 (International HapMap 3 Consortium et al., 2010)
496 and 1000 Genome (1000 Genomes Project Consortium, 2015) that are available for each GWAS summary
497 statistic.

498 Once the SNP weights were derived, individual level PGSs for each disease and general longevity were
499 subsequently computed as a weighted sum of effect allele counts using plink (Purcell et al., 2007). Scores
500 were standardised to have 0 mean and 1 as variance within each ancestry group.

501

502 **Association between PGS and disease of interest**

503 As a baseline, we first examined if the disease PGSs were associated with their diagnoses. For each selected
504 disease, the association was first tested using a general linear model on case-control status as below:

505
$$\text{logit}(\text{Pr}(\text{Individual is diagnosed})) \sim \text{disease PGS} + \text{birth year} + \text{sex} + \text{PC1-10},$$

506 To achieve a fairer comparison with the other experiments, we also evaluated such relationship using a
507 survival model on the age of diagnosis as below:

508
$$\text{surv}(\text{follow-up from birth until diagnosis} \mid \text{disease diagnosis}) \sim \text{disease PGS}$$

509
$$+ \text{birth year} + \text{sex} + \text{PC1-10}.$$

510 The two analyses above were carried out using all eligible individuals in the biobanks. Then for each
511 selected disease, we extracted only the patient group to further conduct the following analyses. To reduce
512 noise in measurements, we limited these within-patient analyses to individuals having a follow-up time of

513 at least three months (0.25 year) after the diagnosis. We tested the association of disease PGSs with our
514 defined prognosis, namely patient survival, using the model below:

515 $\text{surv}(\text{duration of follow up after diagnosis} \mid \text{mortality}) \sim \text{disease PGS} + \text{birth year}$
516 $+ \text{sex} + \text{PC1-10} + \text{age of diagnosis},$

517 as well as the association of general longevity PGS with patient survival as below:

518 $\text{surv}(\text{duration of follow up after diagnosis} \mid \text{mortality}) \sim \text{general longevity PGS} + \text{birth year}$
519 $+ \text{sex} + \text{PC1-10} + \text{age of diagnosis}.$

520 For both associations, we examined both all-cause mortality and cause-specific mortality within the patient
521 group. All analyses were corrected for gender, except in analyses for breast cancer and prostate cancer,
522 where only female/male individuals were used.

523 These analyses were carried out independently for each ancestry group within each participating biobank.
524 We only included biobanks where the count of events of interest in the analysed ancestry group was 50 or
525 more. We subsequently meta-analysed effect sizes for the same ancestry group across biobanks using the
526 inverse variance weighted approach.

527

528 **Construction of PGS from disease mortality GWAS and effect evaluation within FinnGen individuals**

529 For diseases with sufficient power, we derived mortality PGS weights using meta-analysed mortality
530 GWAS results of European populations from all available biobanks except for FinnGen or Generation
531 Scotland. Apart from FinnGen which was used as a test cohort, we also left out results from Generation
532 Scotland for this analysis because their summary statistics did not have effect size or standard error and
533 therefore cannot be used for inverse-variance weighted meta-analysis, which returns necessary statistics for
534 weight derivation. After deriving PGS weights using MegaPRS (Q. Zhang et al., 2021), we subsequently
535 computed individual level disease mortality PGS for patients of each corresponding disease within FinnGen
536 cohort. The weights and scores are computed in the same manner as mentioned in section **Computation of**
537 **individual level PGS**. We evaluated effects of these scores on predicting patients' disease mortality in
538 FinnGen using the model below:

539 $\text{surv}(\text{duration of follow up after diagnosis} \mid \text{mortality}) \sim \text{disease-mortality PGS} + \text{birth year}$
540 $+ \text{sex} + \text{PC1-10} + \text{age of diagnosis}$

542 **Sensitivity analyses for PGS experiments**

543 We ran a series of sensitivity analyses in eligible biobanks to ensure our observations on the PGSs
544 association were robust, under considerations listed below. Similarly, analyses were carried out per eligible
545 ancestry within each biobank and then meta-analysed.

546 First, to demonstrate the impact of relevance between disease progression and susceptibility as shown in
547 our theories, we examined the association between susceptibility PGS and all-cause mortality and compared
548 the results with disease-specific mortality in FinnGen. See **Figure S1** for this result.

549 We then consider other factors that may bias the results.

- 550 • Survivor bias

551 Depending on each biobank's recruitment scheme, some patients were diagnosed before the start of their
552 follow-up, which may lead to biased results due to survivor effect. Therefore, we also ran these analyses
553 for each disease using only samples enrolled before their first onset of the disease of interest. See **Figure**
554 **S12A** for this result.

555 • Relevance between cause of mortality in death certificate and disease diagnosis

556 In this study, we aimed to define disease progression as accurately as possible by focusing our analysis on
557 disease-caused mortality. However, some national death registries may not precisely capture the immediate
558 cause of death, and some mortalities, while documented with the disease as one of causes, may not be truly
559 relevant to the diagnosed disease. To address this concern, we ran the same analysis using only patients
560 with a restricted maximum follow-up length, since death taking place reasonably sooner after being
561 diagnosed might have more to do with the diagnosis, compared to death taking place decades after. Under
562 this consideration, we varied the maximum duration of follow-up after diagnosis by 2, 5 or 10 years. The
563 minimum is still 0.25 years for this analysis. See **Figure S12B** and **Table S16** for this result. Also see **Table**
564 **S2** for sample size breakdown by duration of follow-up in each biobank. As a measurement for
565 comparability between results, we reported the regression coefficients for PGS effect sizes on ten diseases
566 between each sensitivity analysis and main results.

567 • The effect of diagnosed age

568 As shown above, we have included age of diagnosis as one of the covariates in all within-patient main
569 analyses models in order to specifically investigate PGSs' unique genetic effect on disease progression by
570 correcting for the diagnosis. As one of our sensitivity analyses, we also analysed the role of these diagnosed
571 ages in more detail. We repeated all the within-patient analyses for each disease by stratifying patients into
572 early onset and late onset group using 50% age of diagnosis quantile as a cutoff and compared the PGS
573 effects across the two groups. See **Figure S13** and **Table S17** for the result.

574 • Sample relatedness

575 We included all eligible individuals of each biobank in our main analysis, and one may argue that could
576 impact our effect size estimates. Therefore, we ran the same analysis in FinnGen with up to second degree
577 relatives removed. See **Figure S14** and **Table S15** for this result.

578 • Results from non-European populations

579 Since only patients were considered for most of our analyses, although some of the biobanks, e.g. UK
580 biobank and BioMe, were known to be rather diverse, we wended up with enough power for main results
581 only for the European super population. Nevertheless, comparison of results with other less powered, but
582 available populations can be found in **Figure S15** for reference.

583 Forest plot for effects from each biobank is presented in **Figure S16**.

1 Theoretical framework

2 Setup and notations

We start by defining the liability of the endpoint of the disease susceptibility as the random variable S using a simple polygenic risk model, following the lines of (Hujouel et al., 2020):

$$S = \beta_S^T \mathbf{g} + \epsilon_S.$$

In the above expression, \mathbf{g} is an $m \times 1$ random vector of standardized genotypes¹ and β_S is a *sparse* $m \times 1$ zero-mean random vector of variant effect sizes on the liability of disease susceptibility. Additionally, ϵ_S is the zero mean *residual error* vector that is independent from $\beta_S \mathbf{g}$ and includes non-genetic effects and environmental noise in disease susceptibility liability. (We use the subscript S in the above vectors as a reminder that the respective variables correspond to the liability of disease susceptibility.) Therefore, $\beta_S \mathbf{g}$ is a zero mean random variable, from which it follows that its expectation is zero, namely

$$\mathbb{E}(S) = 0.$$

We now define the variance of $\beta_S \mathbf{g}$ as

$$\text{Var}(\beta_S \mathbf{g}) = h_{sus},$$

3 where $0 \leq h_{sus} \leq 1$. For simplicity, we normalize the variance of the random variable S to be equal to one,
 4 i.e., $\text{Var}(S) = 1$. Intuitively, this normalization implies that h_{sus} can be interpreted as the *heritability* of
 5 the disease susceptibility endpoint.

Similarly, we can define the liability of the disease progression endpoint, which, in our work, is the mortality due to the disease, as the random variable P :

$$P = \beta_P^T \mathbf{g} + f(S) + \epsilon_P.$$

In the above expression, β_P is the zero-mean random vector of direct variant effect sizes that play a role *specifically* on the disease progression liability. Also, ϵ_P is the zero mean random vector that models residual error that is independent of $\beta_P^T \mathbf{g}$ and any term of $f(S)$ in the liability of the disease progression. Since disease progression is clinically considered as a continuation of the development of the same disease, we believe that it is reasonable to assume that the liability of disease susceptibility will also play a role on its progression. Therefore, we added the term $f(S)$ in the liability of disease progression to introduce a *causal* contribution for the disease susceptibility liability. To the best of our knowledge, this function $f(S)$ has not been studied or quantified in prior work. Therefore, for the sake of simplicity and in the absence of prior models, we assume that $f(S)$ it as a simple linear function, namely

$$\begin{aligned} P &= \beta_P^T \mathbf{g} + cS + \epsilon_P \\ &= (\beta_P^T + c\beta_S^T) \mathbf{g} + (\epsilon_P + c\epsilon_S). \end{aligned}$$

In the above, c is a constant and we assume that the effect of disease susceptibility liability on the progression is non-negative, from which it follows that $c \geq 0$. Recall that β_P, ϵ_P , and S all have zero mean, implying that $\mathbb{E}(P) = 0$. Let the variance of the progression of the genetic component $\beta_P^T \mathbf{g}$ be defined as follows:

$$\text{Var}(\beta_P \mathbf{g}) = h_{prog},$$

6 where $0 \leq h_{prog} \leq 1$. Again, we normalize the variance of the random variable P to be equal to one, i.e.,
 7 $\text{Var}(P) = 1$, which can be achieved by placing constraints on the constant c and the error term ϵ_P . In this
 8 case, h_{prog} can be interpreted as the *unique heritability* of the disease progression. Given our normalization,
 9 it follows that $c \leq 1$. As a corner case, when $c = 1$, ϵ_P and h_{prog} must both be equal to zero, which implies
 10 $P = S$. This special case does not merit any further consideration and in the upcoming section we will only
 11 focus on $0 \leq c < 1$.

We note that the random variables S and P as defined in this section indicate the liability of the two endpoints of interest for a particular disease in *general* populations. However, in practice, the study of disease

¹Here m could be the total number of variants in the human genome.

progression focuses only on the patient group. Such “within-patient” measurements on a continuous scale can be viewed, at least conceptually, as the liability of disease progression conditioned on the liability of disease susceptibility. Therefore, we regress the liability of the progression (random variable P) on the liability of disease susceptibility (random variable S) to get a continuous, within-patient, progression measurement $P|S$:

$$P|S = P - \alpha S.$$

The regression coefficient α can be analytically derived as follows:

$$\begin{aligned} \alpha &= \text{Cov}(S, P) / \text{Var}(S) = \text{Cov}(S, P) \\ &= \text{Cov}(\beta_P^T \mathbf{g}, \beta_S^T \mathbf{g}) + c \text{Var}(\beta_S^T \mathbf{g}) + c \text{Var}(\epsilon_S) \\ &= \text{Cov}(\beta_P^T \mathbf{g}, \beta_S^T \mathbf{g}) + c \text{Var}(S) \\ &= \text{Cov}(\beta_P^T \mathbf{g}, \beta_S^T \mathbf{g}) + c. \end{aligned}$$

Therefore the within-patients progression liability can be expressed as follows:

$$\begin{aligned} P|S &= P - \alpha S \\ &= (\beta_P^T + c\beta_S^T \mathbf{g} + c\epsilon_S + \epsilon_P - (\text{Cov}(\beta_P^T \mathbf{g}, \beta_S^T \mathbf{g}) + c) \cdot (\beta_S^T \mathbf{g} + \epsilon_S)) \\ &= (\beta_P^T \mathbf{g} + \epsilon_P) - \text{Cov}(\beta_P^T \mathbf{g}, \beta_S^T \mathbf{g}) \cdot (\beta_S^T \mathbf{g} + \epsilon_S) \\ &= (\beta_P^T \mathbf{g} + \epsilon_P) - \text{Cov}(\beta_P^T \mathbf{g}, \beta_S^T \mathbf{g}) \cdot S. \end{aligned}$$

Let ρ denote the correlation coefficient between $\beta_P^T \mathbf{g}$ and $\beta_S^T \mathbf{g}$. Then,

$$\begin{aligned} \text{Cov}(\beta_P^T \mathbf{g}, \beta_S^T \mathbf{g}) &= \rho \sqrt{\text{Var}(\beta_P^T \mathbf{g}) \text{Var}(\beta_S^T \mathbf{g})} \\ &= \rho \sqrt{h_{prog} h_{sus}}. \end{aligned}$$

Finally, the within-patient disease progression liability $P|S$ can be simplified as follows:

$$P|S = (\beta_P^T \mathbf{g} + \epsilon_P) - \rho \sqrt{h_{prog} h_{sus}} \cdot S.$$

12 The $\rho = 0$ case

In order to gain intuition for the more complicated analyses that will follow, we start by looking at the simple (yet admittedly unrealistic) case where there is no genetic correlation between the disease susceptibility and the progression due to the *unique* genetic factors. Mathematically, we assume that the covariance between $\beta_P^T \mathbf{g}$ and $\beta_S^T \mathbf{g}$ is equal to zero, i.e., $\rho = 0$. In this case, the regression parameter α is equal to c and

$$P|S = \beta_P^T \mathbf{g} + \epsilon_P.$$

13 Heritability of within-patient disease progression

In this case, $P|S$ has a simple genetic component given by $G_{P|S} = \beta_P^T \mathbf{g}$. Therefore, its heritability can be expressed as

$$h_{prog|sus} = \frac{\text{Var}(G_{P|S})}{\text{Var}(P|S)}.$$

In the above,

$$\text{Var}(G_{P|S}) = \text{Var}(\beta_P^T \mathbf{g}) = h_{prog}$$

and

$$\begin{aligned} \text{Var}(P|S) &= \text{Var}(\beta_P^T \mathbf{g} + \epsilon_P) \\ &= h_{prog} + \text{Var}(\epsilon_P). \end{aligned}$$

Recall that P is normalized so that $\text{Var}(P) = 1$. Therefore, $\text{Var}(\beta_P^T \mathbf{g} + c\beta_S^T \mathbf{g} + c\epsilon_S + \epsilon_P) = 1$. Since we assumed that $\beta_P^T \mathbf{g}$ and $\beta_S^T \mathbf{g}$ are independent, all terms of P are pairwise independent and

$$\begin{aligned} 1 &= \text{Var}(P) = \text{Var}(\beta_P^T \mathbf{g} + c\beta_S^T \mathbf{g} + c\epsilon_S + \epsilon_P) \\ &= \text{Var}(\beta_P^T \mathbf{g}) + c^2 \text{Var}(\beta_S^T \mathbf{g}) + c^2 \text{Var}(\epsilon_S) + \text{Var}(\epsilon_P) \\ &= h_{prog} + c^2 \text{Var}(S) + \text{Var}(\epsilon_P) \\ &= h_{prog} + c^2 + \text{Var}(\epsilon_P). \end{aligned}$$

Thus,

$$\text{Var}(\epsilon_P) = 1 - h_{prog} - c^2$$

and

$$h_{prog|sus} = \frac{h_{prog}}{h_{prog} + 1 - h_{prog} - c^2} = (1/1-c^2) \cdot h_{prog}.$$

14 Recall that $0 \leq c < 1$, which implies that $1/(1-c^2) \geq 1$. Therefore, when c approaches one, the constant
15 before h_{prog} increases rapidly. Intuitively, in this case, the progression endpoint liability depends heavily
16 on susceptibility and little phenotypic variability among patients remains. However, the constraint on h_{prog}
17 ($1 - h_{prog} - c^2 > 0$) implies that h_{prog} will also have to be small in this case.

18 Using disease susceptibility genetics to understand “within-patients” progression

19 In our empirical evaluations using simulated data, we try to explore the association between between Poly-
20 genic Scores (PGS) derived from the respective disease susceptibility GWAS and the patient disease pro-
21 gression, as characterized by mortality. Equivalently, using the parlance of the previous sections, we are
22 trying to explore the extent to which the “within-patients” progression can be explained by the genetics of
23 the disease susceptibility endpoint.

Within our framework, we can theoretically answer this question. Let the genetic component of disease susceptibility liability be denoted by $G_S = \beta_S^T \mathbf{g}$. We can then look at the correlation coefficient $R^2(G_S, P|S)$ using the variance-covariance ratio:

$$R^2(G_S, P|S) = \left(\frac{\text{Cov}(G_S, P|S)}{\sqrt{\text{Var}(G_S)\text{Var}(P|S)}} \right)^2 = \frac{\text{Cov}^2(G_S, P|S)}{\text{Var}(G_S)\text{Var}(P|S)},$$

where $\text{Var}(G_S) = \text{Var}(\beta_S^T \mathbf{g}) = h_{sus}$ and $\text{Var}(P|S) = 1 - c^2$. Assuming $\rho = 0$, we get

$$\text{Cov}(G_S, P|S) = \text{Cov}(\beta_S^T \mathbf{g}, \beta_P^T \mathbf{g} + \epsilon_P) = 0.$$

24 As expected, in this setting, the genetics of disease susceptibility cannot be used to explain the “within-
25 patient” disease progression.

26 Using population-level disease progression genetics to understand “within-patients” progres- 27 sion

28 In our empirical evaluations using simulated data, we also tried to look at the behavior of general longevity
29 PGS derived from lifespan GWAS (Timmers et al., 2019). Recall that in our work we defined mortality as
30 the event of interest, which we will use as a proxy for genetics of the disease progression measure in the
31 general population instead of only within the patient group. We are aware that getting a real “liability of
32 disease progression measure in the general population” is usually impossible in practice: in most cases, a
33 disease progression is not defined for individuals who do not have a disease diagnosis in the first place. This
34 is only possible in special cases where the progression is defined in a general manner. For example, if instead
35 of disease specific mortality, mortality due to *any* cause is defined as disease progression², we can get the
36 genetic determinants of this progression from a longevity GWAS for the general population instead of just
37 the patient group.

²This measure has been widely used in prior work.

In our proposed framework, we can simulate such measurements by defining the liability of the disease progression in the general population as P . Then, the within-patient progression liability is defined as $P|S$. Therefore, the genetics of the disease progression measure in the general population can be analyzed via P , which includes both the unique component of the disease progression as well as the contribution of the disease susceptibility, namely

$$G_P = (\beta_P^T + c\beta_S^T)\mathbf{g}.$$

Association between the longevity PRS and patients' survival is akin to asking about the amount of variance of the within-patients disease progression that is explained by the genetics of the progression liability assessed in the general population. We can theoretically estimate this value using the correlation coefficient $R^2(G_P, P|S)$ as follows:

$$R^2(G_P, P|S) = \left(\frac{\text{Cov}(G_P, P|S)}{\sqrt{\text{Var}(G_P)\text{Var}(P|S)}} \right)^2 = \frac{\text{Cov}^2(G_P, P|S)}{\text{Var}(G_P)\text{Var}(P|S)}.$$

Assuming that $\beta_P^T\mathbf{g}$ and $\beta_S^T\mathbf{g}$ are independent, we get

$$\begin{aligned} \text{Var}(G_P) &= \text{Var}((\beta_P^T + c\beta_S^T)\mathbf{g}) \\ &= \text{Var}(\beta_P^T\mathbf{g}) + \text{Var}(c\beta_S^T\mathbf{g}) \\ &= h_{prog} + c^2h_{sus}. \end{aligned}$$

Similarly,

$$\begin{aligned} \text{Cov}(G_P, P|S) &= \text{Cov}((\beta_P^T + c\beta_S^T)\mathbf{g}, (\beta_P^T\mathbf{g} + \epsilon_P)) \\ &= \text{Cov}(\beta_P^T\mathbf{g}, \beta_P^T\mathbf{g}) \\ &= \text{Var}(\beta_P^T\mathbf{g}) = h_{prog}. \end{aligned}$$

Using the above, we can now express the correlation coefficient $R^2(G_P, P|S)$ as follows:

$$R_{G_P, P|S}^2 = \frac{\text{Cov}^2(G_P, P|S)}{\text{Var}(G_P)\text{Var}(P|S)} = \frac{h_{prog}^2}{(h_{prog} + c^2h_{sus}) \cdot (h_{prog} + \text{Var}(\epsilon_P))}.$$

In the extreme corner case $c = 0$, i.e., when disease susceptibility and progression are two completely independent endpoints, we get

$$R_{G_P, P|S}^2 = \frac{h_{prog}}{h_{prog} + \text{Var}(\epsilon_P)}.$$

38 In this case, we get perfect correlation ($R_{G_P, P|S}^2 = 1$) if $\text{Var}(\epsilon_P) = 0$, because assuming $c = 0$ and $\rho = 0$,
 39 disease susceptibility and progression are completely orthogonal events. Therefore, population-level disease
 40 progression genetics is equivalent to within-patient progression genetics, which predicts the progression status
 41 perfectly in the absence of noise.

42 Genetic correlation across within-patient disease progression, progression in general population 43 and disease susceptibility

Finally, we can estimate the correlation between the genetic component of the “within-patient” disease progression $G_{P|S} = \beta_P^T\mathbf{g}$, the population disease progression liability $G_P = (\beta_P^T + c \cdot \beta_S^T)\mathbf{g}$, and the genetics of the disease susceptibility $\beta_S^T\mathbf{g}$. In the case where $\rho = 0$, it is easy to show that the genetic correlation $\rho(P|S, S)$ between $P|S$ and S is

$$\rho(P|S, S) = \frac{\text{Cov}(G_{P|S}, G_S)}{\sqrt{\text{Var}(G_{P|S})\text{Var}(G_S)}} = 0.$$

We compute the genetic correlation $\rho(P|S, P)$ between $P|S$ and P as follows:

$$\begin{aligned}\rho(P|S, P) &= \frac{\text{Cov}(G_{P|S}, G_P)}{\sqrt{\text{Var}(G_{P|S})\text{Var}(G_P)}} \\ &= \frac{\text{Cov}(\beta_P^T \mathbf{g}, \beta_P^T \mathbf{g} + c\beta_S^T \mathbf{g})}{\sqrt{h_{prog} \cdot (h_{prog} + c^2 h_{sus})}} \\ &= \frac{h_{prog}}{\sqrt{h_{prog} \cdot (h_{prog} + c^2 h_{sus})}} \\ &= \sqrt{\frac{h_{prog}}{(h_{prog} + c^2 h_{sus})}}.\end{aligned}$$

The genetic correlation $\rho(P, S)$ between P and S can be computed as:

$$\begin{aligned}\rho(P, S) &= \frac{\text{Cov}(G_P, G_S)}{\sqrt{\text{Var}(G_P)\text{Var}(G_S)}} \\ &= \frac{\text{Cov}(\beta_P^T \mathbf{g} + c\beta_S^T \mathbf{g}, \beta_S^T \mathbf{g})}{\sqrt{(h_{prog} + c^2 h_{sus}) \cdot h_{sus}}} \\ &= \frac{ch_{sus}}{\sqrt{(h_{prog} + c^2 h_{sus}) \cdot h_{sus}}} \\ &= c \cdot \sqrt{\frac{h_{sus}}{(h_{prog} + c^2 h_{sus})}}.\end{aligned}$$

44 We note that when $c = 0$, i.e., the disease susceptibility is completely independent of the progression, we get
45 $\rho(P, S) = 0$, as expected. Similarly, when $c = 1$, it follows that $h_{prog} = \beta_P^T \mathbf{g} = \beta_S^T \mathbf{g}$ and we get $\rho(P, S) = 1$,
46 again as expected.

47 The $\rho \neq 0$ case

48 We now proceed to discuss the more interesting $\rho \neq 0$ case. The resulting formulas are analogs of the
49 formulas in Section , albeit more complicated to account for non-zero ρ .

50 Heritability of within-patient disease progression

Consider the case where $\beta_P^T \mathbf{g}$ and $\beta_S^T \mathbf{g}$ are not independent. In this setting, the genetic component for the within-patient disease progression will be

$$G_{P|S} = \beta_P^T \mathbf{g} - \rho \sqrt{h_{prog} h_{sus}} \cdot \beta_S^T \mathbf{g}.$$

Its variance can be computed as follows:

$$\begin{aligned}\text{Var}(G_{P|S}) &= \text{Var}(\beta_P^T \mathbf{g} - \rho \sqrt{h_{prog} h_{sus}} \cdot \beta_S^T \mathbf{g}) \\ &= \text{Var}(\beta_P^T \mathbf{g}) - 2\rho \sqrt{h_{prog} h_{sus}} \cdot \text{Cov}(\beta_P^T \mathbf{g}, \beta_S^T \mathbf{g}) + (\rho \sqrt{h_{prog} h_{sus}})^2 \text{Var}(\beta_S^T \mathbf{g}) \\ &= h_{prog} - 2(\rho \sqrt{h_{prog} h_{sus}})^2 + (\rho \sqrt{h_{prog} h_{sus}})^2 \cdot h_{sus} \\ &= h_{prog}(1 - 2\rho^2 h_{sus} + \rho^2 h_{sus}^2).\end{aligned}$$

³I.e., the disease susceptibility and the progression are exactly the same.

Similarly, its phenotypic variance will be:

$$\begin{aligned}
 \text{Var}(P|S) &= \text{Var}(\beta_P^T \mathbf{g} + \epsilon_P - \rho\sqrt{h_{prog}h_{sus}} \cdot S) \\
 &= h_{prog} + \text{Var}(\epsilon_P) + (\rho\sqrt{h_{prog}h_{sus}})^2 \cdot \text{Var}(S) - 2\rho\sqrt{h_{prog}h_{sus}} \cdot \text{Cov}(\beta_P^T \mathbf{g}, S) \\
 &= h_{prog} + \text{Var}(\epsilon_P) + (\rho\sqrt{h_{prog}h_{sus}})^2 - 2\rho\sqrt{h_{prog}h_{sus}} \cdot \text{Cov}(\beta_P^T \mathbf{g}, \beta_S^T \mathbf{g}) \\
 &= h_{prog} + \text{Var}(\epsilon_P) + (\rho\sqrt{h_{prog}h_{sus}})^2 - 2(\rho\sqrt{h_{prog}h_{sus}})^2 \\
 &= h_{prog} + \text{Var}(\epsilon_P) - \rho^2 h_{prog} h_{sus}.
 \end{aligned}$$

Next, in order to compute the variance of ϵ_P , i.e., $\text{Var}(\epsilon_P)$, we need to look at the variance of the disease progression liability defined in the population. Recall that, by our normalization assumptions, $\text{Var}(P) = 1$. Therefore,

$$\begin{aligned}
 1 = \text{Var}(P) &= \text{Var}(\beta_P^T \mathbf{g} + c\beta_S^T \mathbf{g} + c\epsilon_S + \epsilon_P) \\
 &= \text{Var}(\beta_P^T \mathbf{g}) + c^2 \text{Var}(\beta_S^T \mathbf{g}) + c^2 \text{Var}(\epsilon_S) + \text{Var}(\epsilon_P) + 2c \text{Cov}(\beta_P^T \mathbf{g}, \beta_S^T \mathbf{g}) \\
 &= h_{prog} + c^2 \text{Var}(S) + \text{Var}(\epsilon_P) + 2c\rho\sqrt{h_{prog}h_{sus}} \\
 &= h_{prog} + c^2 + \text{Var}(\epsilon_P) + 2c\rho\sqrt{h_{prog}h_{sus}}.
 \end{aligned}$$

We can now conclude that

$$\text{Var}(\epsilon_P) = 1 - h_{prog} - c^2 - 2c\rho\sqrt{h_{prog}h_{sus}}.$$

Using the above equation, the phenotypic variance of within-patient disease progression can then be expressed as:

$$\begin{aligned}
 \text{Var}(P|S) &= h_{prog} + 1 - h_{prog} - c^2 - 2c\rho\sqrt{h_{prog}h_{sus}} - \rho^2 h_{prog} h_{sus} \\
 &= 1 - c^2 - 2c\rho\sqrt{h_{prog}h_{sus}} - \rho^2 h_{prog} h_{sus}.
 \end{aligned}$$

Finally, its heritability can be expressed as

$$h_{prog|sus} = \frac{\text{Var}(G_{P|S})}{\text{Var}(P|S)} = \frac{h_{prog}(1 - 2\rho^2 h_{sus} + \rho^2 h_{sus}^2)}{1 - c^2 - 2c\rho\sqrt{h_{prog}h_{sus}} - \rho^2 h_{prog} h_{sus}}.$$

51 Using disease susceptibility genetics to understand “within-patients” progression

We again consider the correlation coefficient $R^2(G_S, P|S)$:

$$R^2(G_S, P|S) = \left(\frac{\text{Cov}(G_S, P|S)}{\sqrt{\text{Var}(G_S)\text{Var}(P|S)}} \right)^2 = \frac{\text{Cov}^2(G_S, P|S)}{\text{Var}(G_S)\text{Var}(P|S)}.$$

In the above, $\text{Var}(G_S) = \text{Var}(\beta_S^T \mathbf{g}) = h_{sus}$ and

$$\begin{aligned}
 \text{Cov}(G_S, P|S) &= \text{Cov}(\beta_S^T \mathbf{g}, (\beta_P^T \mathbf{g} + \epsilon_P) - \rho\sqrt{h_{prog}h_{sus}} \cdot S) \\
 &= \text{Cov}(\beta_S^T \mathbf{g}, \beta_P^T \mathbf{g}) - \rho\sqrt{h_{prog}h_{sus}} \cdot \text{Cov}(\beta_S^T \mathbf{g}, S) \\
 &= \rho\sqrt{h_{prog}h_{sus}} - \rho\sqrt{h_{prog}h_{sus}} \cdot h_{sus} \\
 &= \rho(1 - h_{sus})\sqrt{h_{prog}h_{sus}}.
 \end{aligned}$$

We can now use the expression for $\text{Var}(P|S)$ derived in the previous section to get:

$$\begin{aligned}
 R^2(G_S, P|S) &= \frac{\text{Cov}^2(G_S, P|S)}{\text{Var}(G_S)\text{Var}(P|S)} \\
 &= \frac{(\rho(1 - h_{sus})\sqrt{h_{prog}h_{sus}})^2}{h_{sus}(1 - c^2 - 2c\rho\sqrt{h_{prog}h_{sus}} - \rho^2 h_{prog} h_{sus})} \\
 &= \frac{\rho^2(1 - h_{sus})^2 h_{prog}}{1 - c^2 - 2c\rho\sqrt{h_{prog}h_{sus}} - \rho^2 h_{prog} h_{sus}}.
 \end{aligned}$$

52 **Using population-level disease progression genetics to understand “within-patients” progres-**
 53 **sion**

We now express the “within-patient” group disease progression variance that is explained by the genetics of the progression liability as measured in the population. Formally, $R^2(G_P, P|S)$ can be expressed as:

$$R^2(G_P, P|S) = \left(\frac{\text{Cov}(G_P, P|S)}{\sqrt{\text{Var}(G_P)\text{Var}(P|S)}} \right)^2 = \frac{\text{Cov}^2(G_P, P|S)}{\text{Var}(G_P)\text{Var}(P|S)}.$$

The variance of the genetic component for the “within-patient” group disease progression is:

$$\begin{aligned} \text{Var}(G_P) &= \text{Var}((\beta_P^T + c\beta_S^T)\mathbf{g}) \\ &= \text{Var}(\beta_P^T\mathbf{g}) + \text{Var}(c\beta_S^T\mathbf{g}) + 2\text{Cov}(\beta_P^T\mathbf{g}, c\beta_S^T\mathbf{g}) \\ &= h_{prog} + c^2h_{sus} + 2c\rho\sqrt{h_{prog}h_{sus}}. \end{aligned}$$

Similarly, the covariance $\text{Cov}(G_P, P|S)$ is

$$\begin{aligned} \text{Cov}(G_P, P|S) &= \text{Cov}((\beta_P^T + c\beta_S^T)\mathbf{g}, (\beta_P^T\mathbf{g} + \epsilon_P) - \rho\sqrt{h_{prog}h_{sus}} \cdot S) \\ &= \text{Cov}(\beta_P^T\mathbf{g}, \beta_P^T\mathbf{g}) - \rho\sqrt{h_{prog}h_{sus}} \cdot \text{Cov}(\beta_P^T\mathbf{g}, S) + \text{Cov}(c\beta_S^T\mathbf{g}, \beta_P^T\mathbf{g}) - \rho\sqrt{h_{prog}h_{sus}} \cdot \text{Cov}(c\beta_S^T\mathbf{g}, S) \\ &= \text{Var}(\beta_P^T\mathbf{g}) - \rho\sqrt{h_{prog}h_{sus}} \cdot \text{Cov}(\beta_P^T\mathbf{g}, \beta_S^T\mathbf{g}) + c \cdot \text{Cov}(\beta_S^T\mathbf{g}, \beta_P^T\mathbf{g}) - c\rho\sqrt{h_{prog}h_{sus}} \cdot \text{Var}(\beta_S^T\mathbf{g}) \\ &= h_{prog} - (\rho\sqrt{h_{prog}h_{sus}})^2 + c\rho\sqrt{h_{prog}h_{sus}} - c\rho\sqrt{h_{prog}h_{sus}}h_{sus} \\ &= h_{prog} + c(1 - h_{sus})\rho\sqrt{h_{prog}h_{sus}} - \rho^2h_{prog}h_{sus}. \end{aligned}$$

Finally, the correlation coefficient $R^2(G_P, P|S)$ can be expressed as follows:

$$\begin{aligned} R^2(G_P, P|S) &= \frac{\text{Cov}^2(G_P, P|S)}{\text{Var}(G_P)\text{Var}(P|S)} \\ &= \frac{(h_{prog} + c(1 - h_{sus})\rho\sqrt{h_{prog}h_{sus}} - \rho^2h_{prog}h_{sus})^2}{(h_{prog} + c^2h_{sus} + 2c\rho\sqrt{h_{prog}h_{sus}})(1 - c^2 - 2c\rho\sqrt{h_{prog}h_{sus}} - \rho^2h_{prog}h_{sus})}. \end{aligned}$$

54 Note that if $c = 0$ and $\rho = 0$, i.e., the liability of disease progression and the susceptibility are uncorrelated,
 55 the correlation coefficient $R^2(G_P, P|S) = h_{prog}$, which is exactly equal to the heritability of the progression
 56 liability.

57 **Genetic correlation across in-patient progression, progression endpoint in population, and**
 58 **susceptibility**

59 Similar to the $\rho = 0$ case, we can estimate the correlation coefficient between the genetic component of the
 60 “within-patient” group disease progression (in our notation, this component is $G_{P|S} = \beta_P^T\mathbf{g} - \rho\sqrt{h_{prog}h_{sus}} \cdot$
 61 $\beta_S^T\mathbf{g}$), the genetics for progression liability in the population (in our notation, $G_P = (\beta_P^T + c \cdot \beta_S^T)\mathbf{g}$), and the
 62 disease susceptibility genetics (in our notation, $\beta_S^T\mathbf{g}$).

Recall that, by definition, $G_{P|S}$ has a variance equal to

$$\text{Var}(G_{P|S}) = h_{prog}(1 - 2\rho^2h_{sus} + \rho^2h_{sus}^2).$$

Also,

$$\text{Var}(G_P) = h_{prog} + c^2h_{sus} + 2c\rho\sqrt{h_{prog}h_{sus}}.$$

Also, by definition, $\text{Var}(\beta_{\mathbf{g}i}\mathbf{g}) = h_{sus}$. We can now compute the genetic correlation $R(P|S, S)$ between $P|S$

and S as follows:

$$\begin{aligned}
 R(P|S, S) &= \frac{\text{Cov}(G_{P|S}, G_S)}{\sqrt{\text{Var}(G_{P|S})\text{Var}(G_S)}} \\
 &= \frac{\text{Cov}(\beta_P^T \mathbf{g} - \rho\sqrt{h_{prog}h_{sus}} \cdot \beta_S^T \mathbf{g}, \beta_S^T \mathbf{g})}{\sqrt{h_{prog}(1 - 2\rho^2h_{sus} + \rho^2h_{sus}^2)} \cdot h_{sus}} \\
 &= \frac{\rho\sqrt{h_{sus}h_{prog}} - \rho\sqrt{h_{sus}h_{prog}} \cdot h_{sus}}{\sqrt{h_{prog}(1 - 2\rho^2h_{sus} + \rho^2h_{sus}^2)} \cdot h_{sus}} \\
 &= \frac{\rho\sqrt{h_{sus}h_{prog}} \cdot (1 - h_{sus})}{\sqrt{h_{prog}(1 - 2\rho^2h_{sus} + \rho^2h_{sus}^2)} \cdot h_{sus}} \\
 &= \frac{\rho \cdot (1 - h_{sus})}{\sqrt{(1 - 2\rho^2h_{sus} + \rho^2h_{sus}^2)}}.
 \end{aligned}$$

The genetic correlation $R(P|S, P)$ between $P|S$ and P is:

$$\begin{aligned}
 R(P|S, P) &= \frac{\text{Cov}(G_{P|S}, G_P)}{\sqrt{\text{Var}(G_{P|S})\text{Var}(G_P)}} \\
 &= \frac{\text{Cov}(\beta_P^T \mathbf{g} - \rho\sqrt{h_{prog}h_{sus}} \cdot \beta_S^T \mathbf{g}, \beta_P^T \mathbf{g} + c\beta_S^T \mathbf{g})}{\sqrt{h_{prog}(1 - 2\rho^2h_{sus} + \rho^2h_{sus}^2)} \cdot (h_{prog} + c^2h_{sus} + 2c\rho\sqrt{h_{prog}h_{sus}})} \\
 &= \frac{\text{Var}(\beta_P^T \mathbf{g}) + c^2\rho\sqrt{h_{sus}h_{prog}} \cdot \text{Var}(\beta_S^T \mathbf{g}) + c(1 - \rho\sqrt{h_{sus}h_{prog}})\text{Cov}(\beta_P^T \mathbf{g}, \beta_P^T \mathbf{g})}{\sqrt{h_{prog}(1 - 2\rho^2h_{sus} + \rho^2h_{sus}^2)} \cdot (h_{prog} + c^2h_{sus} + 2c\rho\sqrt{h_{prog}h_{sus}})} \\
 &= \frac{h_{prog} + c^2\rho\sqrt{h_{sus}h_{prog}} \cdot h_{sus} + c(1 - \rho\sqrt{h_{sus}h_{prog}})\rho\sqrt{h_{sus}h_{prog}}}{\sqrt{h_{prog}(1 - 2\rho^2h_{sus} + \rho^2h_{sus}^2)} \cdot (h_{prog} + c^2h_{sus} + 2c\rho\sqrt{h_{prog}h_{sus}})} \\
 &= \frac{h_{prog} + c^2\rho\sqrt{h_{sus}h_{prog}} \cdot h_{sus} + c(1 - \rho\sqrt{h_{sus}h_{prog}})\rho\sqrt{h_{sus}h_{prog}}}{\sqrt{h_{prog}(1 - 2\rho^2h_{sus} + \rho^2h_{sus}^2)} \cdot (h_{prog} + c^2h_{sus} + 2c\rho\sqrt{h_{prog}h_{sus}})}.
 \end{aligned}$$

Finally, the genetic correlation $R(P, S)$ between P and S is:

$$\begin{aligned}
 R(P, S) &= \frac{\text{Cov}(G_P, G_S)}{\sqrt{\text{Var}(G_P)\text{Var}(G_S)}} \\
 &= \frac{\text{Cov}(\beta_P^T \mathbf{g} + c\beta_S^T \mathbf{g}, \beta_S^T \mathbf{g})}{\sqrt{(h_{prog} + c^2h_{sus} + 2c\rho\sqrt{h_{prog}h_{sus}})} \cdot h_{sus}} \\
 &= \frac{\rho\sqrt{h_{prog}h_{sus}} + ch_{sus}}{\sqrt{(h_{prog} + c^2h_{sus} + 2c\rho\sqrt{h_{prog}h_{sus}})} \cdot h_{sus}} \\
 &= \frac{\rho\sqrt{h_{prog}} + c\sqrt{h_{sus}}}{\sqrt{h_{prog} + c^2h_{sus} + 2c\rho\sqrt{h_{prog}h_{sus}}}}.
 \end{aligned}$$

63 We note that if $c = 0$, then $R(P, S) = \rho$; if, in addition, $\rho = 1$, then $R(P, S) = 1$ as well.

64 Estimating variant-level effect sizes for in-patient progression

65 As part of our empirical assessment using real and simulated data, we ran disease mortality GWAS within
 66 the patient groups in order to compare variant level effects between disease susceptibility and within-patient
 67 disease progression. In our empirical evaluations, we found it difficult to replicate SNPs that are strongly
 68 associated with disease susceptibility in the mortality GWAS. A naive explanation for such lack of concor-
 69 dance is the poor definition of the disease progression endpoint as well as the low genetic similarity between

70 disease susceptibility and disease-specific mortality. We believe that there are deeper and more significant
 71 reasons for the aforementioned issue. Indeed, in this section, we will show that beyond real genetic distinc-
 72 tions between the two endpoints, the so called *index event bias* (Yaghootkar et al., 2017) can also play an
 73 important role. Intuitively, this bias arises naturally from the design of prognostic research which focuses on
 74 within-patient assessments. This bias can result in attenuation of effects even for variables that have exactly
 75 the same underlying effect size on disease susceptibility and progression in the population.

76 We start by discussing two, somewhat minor, notational modifications that we will use in the remainder
 77 of this section compared to our work in the previous sections. *First*, our analysis will focus on a *single* variant
 78 instead of additive polygenic effects. Therefore, g is no longer a vector variable, and neither are β_{gp} and
 79 β_{gi} . Similarly to the previous sections, g is still standardized to satisfy $E(g) = 0$ and $\text{Var}(g) = 1$. *Second*,
 80 we introduce the standardized variable u with $E(u) = 0$ and $\text{Var}(u) = 1$ that is independent from g , as well
 81 as from the error terms ϵ_i and ϵ_p ; this variable u accounts for all other causal factors that could be shared
 82 by the disease susceptibility and the progression in the population. This new variable u has direct effects
 83 denoted by β_{ui} and β_{up} on the disease susceptibility and progression liability, respectively. As a composite
 84 variable, u can contain genetic effects from other variants beyond the one captured by g as well as common
 85 non-genetic effects.

In light of the above discussion and notations, we now model the random variables S , P as:

$$\begin{aligned} S &= \beta_{gS}g + \beta_{uS}u + \epsilon_S, \\ P &= \beta_{gP}g + \beta_{uP}u + cI + \epsilon_P \\ &= (\beta_{gP} + c\beta_{gS})g + (\beta_{uP} + c\beta_{uS})u + (\epsilon_P + c\epsilon_S). \end{aligned}$$

In a disease progression GWAS, we are usually interested in the unique genetic effect on disease progression, namely the quantity β_{gP} . However, in a within-patient GWAS, we are interested in measuring g 's observed genetic effect on disease progression, conditioned on disease susceptibility. We denote this within-patient observed genetic effect as β_{gP}^* . Then, the expectation of the within-patient disease progression can be expressed as

$$E(P|g, S) = \beta_{gP}^*g + c^*S.$$

In the above equation, c^* is the observed effect size from the disease susceptibility liability as modelled by the variable S . Using least-squares to estimate β_{gP}^* and c^* , we get:

$$\begin{bmatrix} \beta_{gP}^* \\ c^* \end{bmatrix} = \begin{bmatrix} \text{Var}(g) & \text{Cov}(g, S) \\ \text{Cov}(g, S) & \text{Var}(g) \end{bmatrix}^{-1} \begin{bmatrix} \text{Cov}(g, P) \\ \text{Cov}(S, P) \end{bmatrix}.$$

Recall that g , u , S , and P are all standardized random variables and that ϵ_S and ϵ_P satisfy $E(\epsilon_S) = E(\epsilon_P) = 0$. Moreover, g , u , ϵ_S , and ϵ_P are all pairwise independent. We now analytically compute each term in the above equation, starting with $\text{Cov}(g, S)$:

$$\begin{aligned} \text{Cov}(g, S) &= E(gS) - E(g)E(S) \\ &= E(\beta_{gS}g^2 + \beta_{uS}ug + \epsilon_Sg) - 0 \\ &= \beta_{gS}(E(g^2) - E^2(g)) + \beta_{uS}(E(gu) - E(g) \cdot E(u)) + \epsilon_S E(g) \\ &= \beta_{gS}\text{Var}(g) + \beta_{uS}\text{Cov}(g, u) + 0 \\ &= \beta_{gS}. \end{aligned}$$

Similarly,

$$\begin{aligned} \text{Cov}(g, P) &= E(gP) - E(g)E(P) \\ &= E[(\beta_{gP} + c\beta_{gS})g^2 + (\beta_{uP} + c\beta_{uS})ug + c\epsilon_Sg + \epsilon_Pg] \\ &= \beta_{gP} + c\beta_{gS}, \end{aligned}$$

and

$$\begin{aligned}
 \text{Cov}(S, P) &= \text{E}(SP) - \text{E}(S)\text{E}(P) \\
 &= \text{E}[(\beta_{gP} + c\beta_{gS})g + (\beta_{uP} + c\beta_{uS})u + c\epsilon_S + \epsilon_P](\beta_{gS}g + \beta_{uS}u + \epsilon_S) \\
 &= \text{E}[(\beta_{gP} + c\beta_{gS})\beta_{gS}g^2 + (\beta_{uP} + c\beta_{uS})\beta_{uS}u^2 + c\epsilon_S^2] \\
 &= (\beta_{gP} + c\beta_{gS})\beta_{gS} + (\beta_{uP} + c\beta_{uS})\beta_{uS} + c\text{Var}(\epsilon_S).
 \end{aligned}$$

We can now rewrite the least square solution for the observed effect sizes as follows:

$$\begin{aligned}
 \begin{bmatrix} \beta_{gP}^* \\ c^* \end{bmatrix} &= \begin{bmatrix} 1 & \beta_{gS} \\ \beta_{gS} & 1 \end{bmatrix}^{-1} \begin{bmatrix} \beta_{gP} + c\beta_{gS} \\ (\beta_{gP} + c\beta_{gS})\beta_{gS} + (\beta_{uP} + c\beta_{uS})\beta_{uS} + c\text{Var}(\epsilon_S) \end{bmatrix} \\
 &= \frac{1}{1 - \beta_{gS}^2} \begin{bmatrix} 1 & -\beta_{gS} \\ -\beta_{gS} & 1 \end{bmatrix} \begin{bmatrix} \beta_{gP} + c\beta_{gS} \\ (\beta_{gP} + c\beta_{gS})\beta_{gS} + (\beta_{uP} + c\beta_{uS})\beta_{uS} + c\text{Var}(\epsilon_S) \end{bmatrix}.
 \end{aligned}$$

Notice that in the above we analytically computed the inverse of the susceptibility matrix that appeared in the least-squares formulation. Focusing on β_{gP}^* , we get

$$\begin{aligned}
 \beta_{gP}^* &= \frac{1}{1 - \beta_{gS}^2} \cdot [(\beta_{gP} + c\beta_{gS}) - \beta_{gS}((\beta_{gP} + c\beta_{gS})\beta_{gS} + (\beta_{uP} + c\beta_{uS})\beta_{uS} + c\text{Var}(\epsilon_S))] \\
 &= (\beta_{gP} + c\beta_{gS}) - \frac{1}{1 - \beta_{gS}^2} \cdot [\beta_{gS}((\beta_{uP} + c\beta_{uS})\beta_{uS} + c\text{Var}(\epsilon_S))]. \tag{1}
 \end{aligned}$$

We can further compute the variance of ϵ_i , i.e., $\text{Var}(\epsilon_S)$, by looking at the variance of S , which is a sum of three independent components (recall that the variance of S is normalized and equal to one):

$$\begin{aligned}
 1 = \text{Var}(S) &= \text{Var}(\beta_{gS}g + \beta_{uS}u + \epsilon_S) \\
 &= \text{Var}(\beta_{gS}g) + \text{Var}(\beta_{uS}u) + \text{Var}(\epsilon_S) \\
 &= \beta_{gS}^2 + \beta_{uS}^2 + \text{Var}(\epsilon_S).
 \end{aligned}$$

Therefore,

$$\text{Var}(\epsilon_i) = 1 - \beta_{gS}^2 - \beta_{uS}^2.$$

Combining with eqn. (1), we get

$$\begin{aligned}
 \beta_{gP}^* &= (\beta_{gP} + c\beta_{gS}) - \frac{1}{1 - \beta_{gS}^2} \cdot [\beta_{gS}((\beta_{uP} + c\beta_{uS})\beta_{uS} + c(1 - \beta_{gS}^2 - \beta_{uS}^2))] \\
 &= (\beta_{gP} + c\beta_{gS}) - \frac{1}{1 - \beta_{gS}^2} \cdot [\beta_{gS}(\beta_{uP}\beta_{uS} + c(1 - \beta_{gS}^2))] \\
 &= (\beta_{gP} + c\beta_{gS}) - \frac{\beta_{gS}\beta_{uP}\beta_{uS}}{1 - \beta_{gS}^2} - c\beta_{gS} \\
 &= \beta_{gP} - \frac{\beta_{gS}\beta_{uP}\beta_{uS}}{1 - \beta_{gS}^2}.
 \end{aligned}$$

86 In order to further simplify and provide intuition for the above expression, recall that u is a composite variable
 87 accounting for all other causal factors shared by disease susceptibility and progression, including shared
 88 polygenic effects, with the exception of the variant of interest g . In order to proceed with our theoretical
 89 analysis, we will focus on the extreme case where u 's components only include the shared polygenic effect
 90 and there are no environmental risk factors. In other words, in our theoretical analysis, we care more about
 91 how the genetic relationship between disease susceptibility and progression affects downstream analysis.

Due to the polygenic nature of most complex diseases, a single variant usually contributes little to the total endpoint heritability. A reasonable way to model this assumption is to approximate β_{uP} as follows:

$$\beta_{uP}\beta_{uS} \approx \rho\sqrt{h_{sus}h_{prog}}.$$

Using this approximation, it follows that

$$\beta_{gP}^* \approx \beta_{gP} - \frac{\rho \sqrt{h_{sus} h_{prog}}}{1 - \beta_{gS}^2} \beta_{gS}.$$

92 The above relationship indicates that when there is no shared non-genetic risk factors, GWAS results could
93 suffer from index event bias depending on the genetic similarity between variants underlying disease sus-
94 ceptibility and the progression of the unique genetic components, and also the heritability of progression
95 endpoint. Thus, with fixed susceptibility genetics, when the progression endpoint has really low heritability,
96 the absolute biased effect size can also be relatively negligible.

584 **Simulation under proposed framework**

585 We carried out a simulation based on the genome of chromosome 21 (containing 111,212 HM3 SNPs) for
586 10,000 synthetic European individuals created using Hapgen2 (Su et al., 2011). In the simulation, we fixed
587 heritability of disease susceptibility at $h_{sus} = 0.2$, and impact of susceptibility liability on disease
588 progression liability at $c = 0.3$. We further fixed the compositions of causal SNPs for each of these two
589 endpoints so that 0.001 of total SNPs (ie. around 110 in this case) have direct effect on disease susceptibility
590 and 0.001 have direct effect on progression. We changed the proportion of overlap between these two
591 genetic components so that 25%, 50%, 75% of the causal SNPs were shared between susceptibility and
592 progression. To decide effect size for each causal SNP i , we first drew a base effect independently from a
593 standard univariate normal distribution

$$594 \beta_{i,base} \sim N(0, 1)$$

595 and multiply it to the square root of heritability this SNP accounts for to get its final effect size. Causal
596 SNPs shared by susceptibility and progression were simulated to have same base effect on susceptibility
597 and progression so that expected correlation of overall polygenic effects between two endpoints ρ will
598 approximately correspond to the proportion of shared causal SNPs, in this case $\rho = 0.25, 0.5$ and 0.75 .

599 We further vary heritability of disease progression $h_{prog} = 0.005, 0.1$, and 0.2 .

600 Under each simulation setup, we run standard GWAS correcting for top 10 PCs for both susceptibility and
601 progression liability. Note that just like we added age of diagnosis as a covariate in our empirical mortality
602 GWAS, in the progression GWAS, we also correct for susceptibility liability. Subsequently, we clump the
603 GWAS results using plink (Purcell et al., 2007) under parameters `--clump-p1 5e-8 --clump-r2 0.5 --clump-`
604 `kb 250` to extract independent genome-wide significant loci from each GWAS, ran linemodels (Pirinen,
605 2023) and plot the results just like we did in our empirical experiments.

606

607 **Impact of index event bias and Slope-Hunter-like adjustment**

608 To investigate the impact of index event bias, we think the most direct way would be to compare the
609 underlying simulated SNP effects to observed effects from GWAS for disease progression. Recall that
610 effect size for each causal SNP i is a standard normal variable multiplied by square root of its heritability,
611 which can be expressed as

$$612 \beta_i = \sqrt{\frac{h}{\text{Var}(\sum_{i=1}^n \beta_{i,base} g_i)}}$$

613 , where β_i is the underlying causal effect simulated, h is the endpoint heritability, g_i is the genotype of
614 causal SNP i , and n is the total number of causal SNPs for the endpoint. The same equation applies to both
615 susceptibility and progression causal genetic effects.

616 In this experiment as we are investigating the impact of index event bias, on top of shared polygenic effects,
617 we introduced another component u to account for any other shared non-genetic risk factor between the
618 two. Same as previous experiment, heritability of disease susceptibility and impact of susceptibility liability
619 on disease progression liability were still fixed as $h_{sus} = 0.2$ and $c = 0.3$. For this experiment, we further
620 fixed the heritability of disease progression at $h_{prog} = 0.005$. We vary $\rho = 0.25, 0.5$ and 0.75 , and
621 contribution of the non-genetic component on variance of susceptibility and progression liability among
622 $Var_u = 0, 10\%$, or 20% .

623 Under each setup, we ran GWAS on disease susceptibility and progression as described before, and for all
624 progression causal variants, we plotted simulated SNP effects against SNP effects observed from the
625 progression GWAS. We examined the residual sum of squares (rss) for the points around function $y = x$.
626 Furthermore, based on the theory behind of Slope-Hunter, we applied adjustment on SNPs that suffer from
627 index event bias through a procedure described as below:

- 628 1. Extract all susceptibility specific causal SNPs and regress their observed effect sizes from
629 susceptibility GWAS against progression GWAS to obtain the correction factor b .
- 630 2. For each causal SNP i shared between susceptibility and progression, compute the corrected
631 progression genetic effect $\widehat{\beta}_{i,prog}$ as below

$$632 \quad \widehat{\beta}_{i,prog} = \widehat{\beta}_{i,prog}^* - b\widehat{\beta}_{i,prog}$$

633 , where $\widehat{\beta}_{i,prog}^*$ is the observed effect for SNP i from the conditioned progression GWAS, and $\widehat{\beta}_{i,sus}$ is the
634 observed effect for SNP i from the susceptibility GWAS. Note that this experiment may demonstrate the
635 utility of Slope-Hunter-like correction in a nearly “perfect” scenario, where the classification of SNPs
636 (susceptibility specific, progression specific, shared or no effect in either) is given. In practice, a Bayesian
637 or comparable approach needs to be applied for posterior variant group assignment, which can result in
638 worse performance than shown in this manuscript. As a comparison, we show in the same plot the corrected
639 variants effect sizes against the simulated underlying effects and examined rss. See **Figure S19** for results.
640 Note the previous experiment shows impact of index event bias and correction on all underlying causal
641 variants for the progression, whereas in practice, such information is not a given. Therefore, subsequently
642 we tried to examine the real impact of Slope-Hunter-like correction on observed results from GWAS under
643 one of the conditions where the most severe index event bias could be observed ($h_{prog} = 0.005$, $Var_u =$
644 20% , $\rho = 0.5$). We chose $\rho = 0.5$ rather than $\rho = 0.75$, where more causal variants are shared, so that more
645 susceptibility specific SNPs were available for correction factor (b) estimation and a more accurate estimate
646 could be achieved. For this experiment, we first ran progression and susceptibility GWAS respectively as
647 mentioned before and clumped their results to identify independent signals for each. Then using
648 susceptibility specific SNPs, we fitted the correction factor and applied correction on all SNP effects in the
649 progression GWAS. We also corrected for the standard errors as mentioned in (Mahmoud et al., 2022) and
650 recomputed the p-values. Mindful of the difference from the previous experiment, that correction was not
651 only applied on shared causal SNPs but all variants, as that would be what is done empirically. Note that
652 here we still provided causal information for susceptibility specific variants to fit the correction factor,
653 which was again a rather ideal use case for the method. Then, as what was done empirically, we ran
654 linemodels (Pirinen, 2023) on SNP effect before and after the correction (**Figure S20**). None of the SNPs
655 became genome-wide significant in progression GWAS after correction (**Table S19**), and there was no
656 change in Bayesian classification of SNPs.

657 **Reference**

- 658 1000 Genomes Project Consortium. (2015). A global reference for human genetic variation. *Nature*.
659 <http://www.nature.com/nature/journal/v526/n7571/abs/nature15393.html>
- 660 Abbafati, C., Abbas, K. M., Abbasi-Kangevari, M., Abd-Allah, F., Abdelalim, A., Abdollahi, M.,
661 Abdollahpour, I., Abegaz, K. H., Abolhassani, H., Aboyans, V., Abreu, L. G., Abrigo, M. R. M.,
662 Abualhasan, A., Abu-Raddad, L. J., Abushouk, A. I., Adabi, M., Adekanmbi, V., Adeoye, A. M.,
663 Adetokunboh, O. O., ... Amini, S. (2020). Global burden of 369 diseases and injuries in 204
664 countries and territories, 1990–2019: a systematic analysis for the Global Burden of Disease Study
665 2019. *The Lancet*, 396(10258), 1204–1222. [https://doi.org/10.1016/S0140-6736\(20\)30925-](https://doi.org/10.1016/S0140-6736(20)30925-9/ATTACHMENT/1802C2B8-7CCC-467E-B4DD-92F466CF5E15/MMC2E.PDF)
666 [9/ATTACHMENT/1802C2B8-7CCC-467E-B4DD-92F466CF5E15/MMC2E.PDF](https://doi.org/10.1016/S0140-6736(20)30925-9/ATTACHMENT/1802C2B8-7CCC-467E-B4DD-92F466CF5E15/MMC2E.PDF)
- 667 Barbieux, P., György, B., Gand, E., Saulnier, P. J., Ducrocq, G., Halimi, J. M., Feigerlova, E., Hulin-
668 Delmotte, C., Llaty, P., Montaigne, D., Rigalleau, V., Roussel, R., Sosner, P., Zaoui, P., Ragot, S.,
669 Marre, M., Tregouët, D. A., & Hadjadj, S. (2019). No prognostic role of a GWAS-derived genetic
670 risk score in renal outcomes for patients from French cohorts with type 1 and type 2 diabetes.
671 *Diabetes & Metabolism*, 45(5), 494–497. <https://doi.org/10.1016/J.DIABET.2018.01.016>
- 672 Bellenguez, C., Strange, A., Freeman, C., Donnelly, P., & Spencer, C. C. A. (2012). A robust clustering
673 algorithm for identifying problematic samples in genome-wide association studies. *Bioinformatics*,
674 28(1), 134–135. <https://doi.org/10.1093/BIOINFORMATICS/BTR599>
- 675 Bi, W., Fritsche, L. G., Mukherjee, B., Kim, S., & Lee, S. (2020). A Fast and Accurate Method for
676 Genome-Wide Time-to-Event Data Analysis and Its Application to UK Biobank. *American Journal*
677 *of Human Genetics*, 107(2). <https://doi.org/10.1016/J.AJHG.2020.06.003>
- 678 Bien, S. A., Wojcik, G. L., Zubair, N., Gignoux, C. R., Martin, A. R., Kocarnik, J. M., Martin, L. W.,
679 Buyske, S., Haessler, J., Walker, R. W., Cheng, I., Graff, M., Xia, L., Franceschini, N., Matise, T.,
680 James, R., Hindorff, L., Marchand, L. Le, North, K. E., ... Carlson, C. S. (2016). Strategies for
681 Enriching Variant Coverage in Candidate Disease Loci on a Multiethnic Genotyping Array. *PLOS*
682 *ONE*, 11(12), e0167758. <https://doi.org/10.1371/JOURNAL.PONE.0167758>
- 683 Browning, B. L., Tian, X., Zhou, Y., & Browning, S. R. (2021). Fast two-stage phasing of large-scale
684 sequence data. *American Journal of Human Genetics*, 108(10), 1880–1890.
685 [https://doi.org/10.1016/J.AJHG.2021.08.005/ATTACHMENT/3B2363C6-FB5A-41B8-90D7-](https://doi.org/10.1016/J.AJHG.2021.08.005/ATTACHMENT/3B2363C6-FB5A-41B8-90D7-E8E171727BF0/MMC1.PDF)
686 [E8E171727BF0/MMC1.PDF](https://doi.org/10.1016/J.AJHG.2021.08.005/ATTACHMENT/3B2363C6-FB5A-41B8-90D7-E8E171727BF0/MMC1.PDF)
- 687 Bulik-Sullivan, B. K., Loh, P.-R., Finucane, H. K., Ripke, S., Yang, J., Patterson, N., Daly, M. J., Price,
688 A. L., & Neale, B. M. (2015). LD Score regression distinguishes confounding from polygenicity in
689 genome-wide association studies. *Nature Genetics*, 47(3), 291–295. <https://doi.org/10.1038/ng.3211>
- 690 Bycroft, C., Freeman, C., Petkova, D., Band, G., Elliott, L. T., Sharp, K., Motyer, A., Vukcevic, D.,
691 Delaneau, O., O'Connell, J., Cortes, A., Welsh, S., Young, A., Effingham, M., McVean, G., Leslie,
692 S., Allen, N., Donnelly, P., & Marchini, J. (2018). The UK Biobank resource with deep phenotyping
693 and genomic data. *Nature*, 562(7726), 203–209. <https://doi.org/10.1038/s41586-018-0579-z>
- 694 Chen, C. Y., Pollack, S., Hunter, D. J., Hirschhorn, J. N., Kraft, P., & Price, A. L. (2013). Improved
695 ancestry inference using weights from external reference panels. *Bioinformatics (Oxford, England)*,
696 29(11), 1399–1406. <https://doi.org/10.1093/BIOINFORMATICS/BTT144>
- 697 Cho, S. M. J., Koyama, S., Honigberg, M. C., Surakka, I., Haidermota, S., Ganesh, S., Patel, A. P.,
698 Bhattacharya, R., Lee, H., Kim, H. C., & Natarajan, P. (2023). Genetic, sociodemographic, lifestyle,
699 and clinical risk factors of recurrent coronary artery disease events: a population-based cohort study.
700 *European Heart Journal*. <https://doi.org/10.1093/EURHEARTJ/EHAD380>
- 701 Danecek, P., Auton, A., Abecasis, G., Albers, C. A., Banks, E., DePristo, M. A., Handsaker, R. E.,
702 Lunter, G., Marth, G. T., Sherry, S. T., McVean, G., & Durbin, R. (2011). The variant call format
703 and VCFtools. *Bioinformatics*, 27(15), 2156–2158.
704 <https://doi.org/10.1093/BIOINFORMATICS/BTR330>
- 705 Davies, R. W., Flint, J., Myers, S., & Mott, R. (2016). Rapid genotype imputation from sequence without
706 reference panels. *Nature Genetics* 2016 48:8, 48(8), 965–969. <https://doi.org/10.1038/ng.3594>

- 707 Davies, R. W., Kucka, M., Su, D., Shi, S., Flanagan, M., Cunniff, C. M., Chan, Y. F., & Myers, S. (2021).
708 Rapid genotype imputation from sequence with reference panels. *Nature Genetics* 2021 53:7, 53(7),
709 1104–1111. <https://doi.org/10.1038/s41588-021-00877-0>
- 710 Dey, R., Zhou, W., Kiiskinen, T., Havulinna, A., Elliott, A., Karjalainen, J., Kurki, M., Qin, A., Lee, S.,
711 Palotie, A., Neale, B., Daly, M., & Lin, X. (2022). Efficient and accurate frailty model approach for
712 genome-wide survival association analysis in large-scale biobanks. *Nature Communications* 2022
713 13:1, 13(1), 1–13. <https://doi.org/10.1038/s41467-022-32885-x>
- 714 Dudbridge, F., Allen, R. J., Sheehan, N. A., Schmidt, A. F., Lee, J. C., Jenkins, R. G., Wain, L. V.,
715 Hingorani, A. D., & Patel, R. S. (2019). Adjustment for index event bias in genome-wide
716 association studies of subsequent events. *Nature Communications* 2019 10:1, 10(1), 1–10.
717 <https://doi.org/10.1038/s41467-019-09381-w>
- 718 Escala-Garcia, M., Guo, Q., Dörk, T., Canisius, S., Keeman, R., Dennis, J., Beesley, J., Lecarpentier, J.,
719 Bolla, M. K., Wang, Q., Abraham, J., Andrulis, I. L., Anton-Culver, H., Arndt, V., Auer, P. L.,
720 Beckmann, M. W., Behrens, S., Benitez, J., Bermisheva, M., ... Schmidt, M. K. (2019). Genome-
721 wide association study of germline variants and breast cancer-specific mortality. *British Journal of*
722 *Cancer* 2019 120:6, 120(6), 647–657. <https://doi.org/10.1038/s41416-019-0393-x>
- 723 Fahed, A. C., Philippakis, A. A., & Khera, A. V. (2022). The potential of polygenic scores to improve
724 cost and efficiency of clinical trials. *Nature Communications*, 13(1).
725 <https://doi.org/10.1038/S41467-022-30675-Z>
- 726 Feng, Y.-C. A., Ge, T., Cordioli, M., FinnGen, Ganna, A., Smoller, J. W., & Neale, B. M. (2020).
727 Findings and insights from the genetic investigation of age of first reported occurrence for complex
728 disorders in the UK Biobank and FinnGen. *MedRxiv*, 2020.11.20.20234302.
729 <https://doi.org/10.1101/2020.11.20.20234302>
- 730 Finer, S., Martin, H. C., Khan, A., Hunt, K. A., Maclaughlin, B., Ahmed, Z., Ashcroft, R., Durham, C.,
731 Macarthur, D. G., McCarthy, M. I., Robson, J., Trivedi, B., Griffiths, C., Wright, J., Trembath, R.
732 C., & Van Heel, D. A. (2020). Cohort Profile: East London Genes & Health (ELGH), a community-
733 based population genomics and health study in British Bangladeshi and British Pakistani people.
734 *International Journal of Epidemiology*, 49(1), 20–21i. <https://doi.org/10.1093/IJE/DYZ174>
- 735 *FinnGen project | FinnGen*. (n.d.). Retrieved September 13, 2023, from
736 https://www.finnngen.fi/en/for_researchers
- 737 Finucane, H. K., Bulik-Sullivan, B., Gusev, A., Trynka, G., Reshef, Y., Loh, P.-R., Anttila, V., Xu, H.,
738 Zang, C., Farh, K., Ripke, S., Day, F. R., Purcell, S., Stahl, E., Lindstrom, S., Perry, J. R. B., Okada,
739 Y., Raychaudhuri, S., Daly, M. J., ... Price, A. L. (2015). Partitioning heritability by functional
740 annotation using genome-wide association summary statistics. *Nature Genetics*, 47(11), 1228–1235.
741 <https://doi.org/10.1038/ng.3404>
- 742 Gibbs, R. A., Belmont, J. W., Hardenbol, P., Willis, T. D., Yu, F., Zhang, H., Zeng, C., Matsuda, I.,
743 Fukushima, Y., Macer, D. R., Suda, E., Stein, L. D., Cunningham, F., Kanani, A., Thorisson, G. A.,
744 Chakravarti, A., Chen, P. E., Cutler, D. J., Kashuk, C. S., ... Tanaka, T. (2003). The International
745 HapMap Project. *Nature*, 426(6968), 789–796. <https://doi.org/10.1038/nature02168>
- 746 Groha, S., Alaiwi, S. A., Xu, W., Naranbhai, V., Nassar, A. H., Bakouny, Z., El Zarif, T., Saliby, R. M.,
747 Wan, G., Rajeh, A., Adib, E., Nuzzo, P. V., Schmidt, A. L., Labaki, C., Ricciuti, B., Alessi, J. V.,
748 Braun, D. A., Shukla, S. A., Keenan, T. E., ... Gusev, A. (2022). Germline variants associated with
749 toxicity to immune checkpoint blockade. *Nature Medicine*, 28(12), 2584–2591.
750 <https://doi.org/10.1038/S41591-022-02094-6>
- 751 Guo, Q., Schmidt, M. K., Kraft, P., Canisius, S., Chen, C., Khan, S., Tyrer, J., Bolla, M. K., Wang, Q.,
752 Dennis, J., Michailidou, K., Lush, M., Kar, S., Beesley, J., Dunning, A. M., Shah, M., Czene, K.,
753 Darabi, H., Eriksson, M., ... Pharoah, P. D. P. (2015). Identification of Novel Genetic Markers of
754 Breast Cancer Survival. *JNCI: Journal of the National Cancer Institute*, 107(5), 81.
755 <https://doi.org/10.1093/JNCI/DJV081>

- 756 Gusev, A., Groha, S., Taraszka, K., Semenov, Y. R., & Zaitlen, N. (2021). Constructing germline
757 research cohorts from the discarded reads of clinical tumor sequences. *Genome Medicine*, 13(1), 1–
758 14. <https://doi.org/10.1186/S13073-021-00999-4/FIGURES/6>
- 759 Harroud, A., Stridh, P., McCauley, J. L., Saarela, J., R van den Bosch, A. M., Engelenburg, H. J.,
760 Beecham, A. H., Alfredsson, L., Alikhani, K., Amezcua, L., M Andlauer, T. F., Ban, M., Barcellos,
761 L. F., Barizzone, N., Berge, T., Berthele, A., Bittner, S., Bos, S. D., S Briggs, F. B., ... Stefamp, ri.
762 (2023). Locus for severity implicates CNS resilience in progression of multiple sclerosis. *Nature*
763 2023, 1–9. <https://doi.org/10.1038/s41586-023-06250-x>
- 764 Hernesniemi, J. A. (2022). Dawn of the Era of Individualized Genetic Profiling in the Prevention of
765 Sudden Cardiac Death. *Journal of the American College of Cardiology*, 80(9), 884–886.
766 <https://doi.org/10.1016/J.JACC.2022.06.016>
- 767 Houlahan, K. E., Livingstone, J., Fox, N. S., Kurganovs, N., Zhu, H., Sietsma Penington, J., Jung, C.-H.,
768 Yamaguchi, T. N., Heisler, L. E., Jovelin, R., Costello, A. J., Pope, B. J., Kishan, A. U., Corcoran,
769 N. M., Bristow, R. G., Waszak, S. M., Weischenfeldt, J., He, H. H., Hung, R. J., ... Boutros, P. C.
770 (2023). A polygenic two-hit hypothesis for prostate cancer. *JNCI: Journal of the National Cancer*
771 *Institute*, 115(4), 468–472. <https://doi.org/10.1093/JNCI/DJAD001>
- 772 Howie, B. N., Donnelly, P., & Marchini, J. (2009). A Flexible and Accurate Genotype Imputation Method
773 for the Next Generation of Genome-Wide Association Studies. *PLOS Genetics*, 5(6), e1000529.
774 <https://doi.org/10.1371/JOURNAL.PGEN.1000529>
- 775 Hujoel, M. L. A., Gazal, S., Loh, P. R., Patterson, N., & Price, A. L. (2020). Liability threshold modeling
776 of case–control status and family history of disease increases association power. *Nature Genetics*
777 2020 52:5, 52(5), 541–547. <https://doi.org/10.1038/s41588-020-0613-6>
- 778 International HapMap 3 Consortium, D. M., Altshuler, D. M., Gibbs, R. A., Peltonen, L., Altshuler, D.
779 M., Gibbs, R. A., Peltonen, L., Dermitzakis, E., Schaffner, S. F., Yu, F., Peltonen, L., Dermitzakis,
780 E., Bonnen, P. E., Altshuler, D. M., Gibbs, R. A., de Bakker, P. I. W., Deloukas, P., Gabriel, S. B.,
781 Gwilliam, R., ... McEwen, J. E. (2010). Integrating common and rare genetic variation in diverse
782 human populations. *Nature*, 467(7311), 52–58. <https://doi.org/10.1038/nature09298>
- 783 Jabbari, E., Koga, S., Valentino, R. R., Reynolds, R. H., Ferrari, R., Tan, M. M. X., Rowe, J. B., Dalgard,
784 C. L., Scholz, S. W., Dickson, D. W., Warner, T. T., Revesz, T., Höglinger, G. U., Ross, O. A.,
785 Ryten, M., Hardy, J., Shoai, M., Morris, H. R., Mok, K. Y., ... T.M. Hu, M. (2021). Genetic
786 determinants of survival in progressive supranuclear palsy: a genome-wide association study. *The*
787 *Lancet. Neurology*, 20(2), 107–116. [https://doi.org/10.1016/S1474-4422\(20\)30394-X](https://doi.org/10.1016/S1474-4422(20)30394-X)
- 788 Kember, R. L., Merikangas, A. K., Verma, S. S., Verma, A., Judy, R., Abecasis, G., Baras, A., Cantor,
789 M., Coppola, G., Economides, A., Lotta, L., Overton, J. D., Reid, J. G., Shuldiner, A., Beechert, C.,
790 Forsythe, C., Fuller, E. D., Gu, Z., Lattari, M., ... Bućan, M. (2021). Polygenic Risk of Psychiatric
791 Disorders Exhibits Cross-trait Associations in Electronic Health Record Data From European
792 Ancestry Individuals. *Biological Psychiatry*, 89(3), 236–245.
793 <https://doi.org/10.1016/J.BIOPSYCH.2020.06.026>
- 794 Khan, Z., Di Nucci, F., Kwan, A., Hammer, C., Mariathasan, S., Rouilly, V., Carroll, J., Fontes, M.,
795 Acosta, S. L., Guardino, E., Chen-Harris, H., Bhangale, T., Mellman, I., Rosenberg, J., Powles, T.,
796 Hunkapiller, J., Chandler, G. S., & Albert, M. L. (2020). Polygenic risk for skin autoimmunity
797 impacts immune checkpoint blockade in bladder cancer. *Proceedings of the National Academy of*
798 *Sciences of the United States of America*, 117(22), 12288–12294.
799 <https://doi.org/10.1073/PNAS.1922867117/-/DCSUPPLEMENTAL>
- 800 Khan, Z., Hammer, C., Carroll, J., Di Nucci, F., Acosta, S. L., Maiya, V., Bhangale, T., Hunkapiller, J.,
801 Mellman, I., Albert, M. L., McCarthy, M. I., & Chandler, G. S. (2021). Genetic variation associated
802 with thyroid autoimmunity shapes the systemic immune response to PD-1 checkpoint blockade.
803 *Nature Communications*, 12(1). <https://doi.org/10.1038/S41467-021-23661-4>
- 804 Kuhn, R. M., Haussler, D., & James Kent, W. (2013). The UCSC genome browser and associated tools.
805 *Briefings in Bioinformatics*, 14(2), 144–161. <https://doi.org/10.1093/BIB/BBS038>

- 806 Kunkle, B. W., Grenier-Boley, B., Sims, R., Bis, J. C., Damotte, V., Naj, A. C., Boland, A., Vronskaya,
807 M., van der Lee, S. J., Amlie-Wolf, A., Bellenguez, C., Frizatti, A., Chouraki, V., Martin, E. R.,
808 Sleegers, K., Badarinarayan, N., Jakobsdottir, J., Hamilton-Nelson, K. L., Moreno-Grau, S., ...
809 Pericak-Vance, M. A. (2019). Genetic meta-analysis of diagnosed Alzheimer's disease identifies
810 new risk loci and implicates A β , tau, immunity and lipid processing. *Nature Genetics* 2019 51:3,
811 51(3), 414–430. <https://doi.org/10.1038/s41588-019-0358-2>
- 812 Kurki, M. I., Karjalainen, J., Palta, P., Sipilä, T. P., Kristiansson, K., Donner, K. M., Reeve, M. P.,
813 Laivuori, H., Aavikko, M., Kaunisto, M. A., Loukola, A., Lahtela, E., Mattsson, H., Laiho, P., Della
814 Briotta Parolo, P., Lehisto, A. A., Kanai, M., Mars, N., Rämö, J., ... Palotie, A. (2023). FinnGen
815 provides genetic insights from a well-phenotyped isolated population. *Nature* 2023 613:7944,
816 613(7944), 508–518. <https://doi.org/10.1038/s41586-022-05473-8>
- 817 Law, P. J., Timofeeva, M., Fernandez-Rozadilla, C., Broderick, P., Studd, J., Fernandez-Tajes, J.,
818 Farrington, S., Svinti, V., Palles, C., Orlando, G., Sud, A., Holroyd, A., Penegar, S., Theodoratou,
819 E., Vaughan-Shaw, P., Campbell, H., Zgaga, L., Hayward, C., Campbell, A., ... Dunlop, M. G.
820 (2019). Association analyses identify 31 new risk loci for colorectal cancer susceptibility. *Nature*
821 *Communications* 2019 10:1, 10(1), 1–15. <https://doi.org/10.1038/s41467-019-09775-w>
- 822 Lee, J. C., Biasci, D., Roberts, R., Gearry, R. B., Mansfield, J. C., Ahmad, T., Prescott, N. J., Satsangi, J.,
823 Wilson, D. C., Jostins, L., Anderson, C. A., Traherne, J. A., Lyons, P. A., Parkes, M., & Smith, K.
824 G. C. (2017). Genome-wide association study identifies distinct genetic contributions to prognosis
825 and susceptibility in Crohn's disease. *Nature Genetics* 2017 49:2, 49(2), 262–268.
826 <https://doi.org/10.1038/ng.3755>
- 827 Leitsalu, L., Haller, T., Esko, T., Tammesoo, M. L., Alavere, H., Snieder, H., Perola, M., Ng, P. C., Mägi,
828 R., Milani, L., Fischer, K., & Metspalu, A. (2015). Cohort Profile: Estonian Biobank of the Estonian
829 Genome Center, University of Tartu. *International Journal of Epidemiology*, 44(4), 1137–1147.
830 <https://doi.org/10.1093/IJE/DYT268>
- 831 Liu, G., Peng, J., Liao, Z., Locascio, J. J., Corvol, J. C., Zhu, F., Dong, X., Maple-Grødem, J., Campbell,
832 M. C., Elbaz, A., Lesage, S., Brice, A., Mangone, G., Growdon, J. H., Hung, A. Y., Schwarzschild,
833 M. A., Hayes, M. T., Wills, A. M., Herrington, T. M., ... Scherzer, C. R. (2021). Genome-wide
834 survival study identifies a novel synaptic locus and polygenic score for cognitive progression in
835 Parkinson's disease. *Nature Genetics*, 53(6), 787–793. <https://doi.org/10.1038/S41588-021-00847-6>
- 836 Liu, S., Huang, S., Chen, F., Zhao, L., Yuan, Y., Francis, S. S., Fang, L., Li, Z., Lin, L., Liu, R., Zhang,
837 Y., Xu, H., Li, S., Zhou, Y., Davies, R. W., Liu, Q., Walters, R. G., Lin, K., Ju, J., ... Xu, X. (2018).
838 Genomic Analyses from Non-invasive Prenatal Testing Reveal Genetic Associations, Patterns of
839 Viral Infections, and Chinese Population History. *Cell*, 175(2), 347-359.e14.
840 <https://doi.org/10.1016/J.CELL.2018.08.016>
- 841 Loh, P. R., Danecek, P., Palamara, P. F., Fuchsberger, C., Reshef, Y. A., Finucane, H. K., Schoenherr, S.,
842 Forer, L., McCarthy, S., Abecasis, G. R., Durbin, R., & Price, A. L. (2016). Reference-based
843 phasing using the Haplotype Reference Consortium panel. *Nature Genetics* 2016 48:11, 48(11),
844 1443–1448. <https://doi.org/10.1038/ng.3679>
- 845 Mahajan, A., Taliun, D., Thurner, M., Robertson, N. R., Torres, J. M., Rayner, N. W., Payne, A. J.,
846 Steinthorsdottir, V., Scott, R. A., Grarup, N., Cook, J. P., Schmidt, E. M., Wuttke, M., Sarnowski,
847 C., Mägi, R., Nano, J., Gieger, C., Trompet, S., Lecoeur, C., ... McCarthy, M. I. (2018). Fine-
848 mapping type 2 diabetes loci to single-variant resolution using high-density imputation and islet-
849 specific epigenome maps. *Nature Genetics* 2018 50:11, 50(11), 1505–1513.
850 <https://doi.org/10.1038/s41588-018-0241-6>
- 851 Mahmoud, O., Dudbridge, F., Davey Smith, G., Munafò, M., & Tilling, K. (2022). A robust method for
852 collider bias correction in conditional genome-wide association studies. *Nature Communications*
853 2022 13:1, 13(1), 1–13. <https://doi.org/10.1038/s41467-022-28119-9>
- 854 Malik, R., Chauhan, G., Traylor, M., Sargurupremraj, M., Okada, Y., Mishra, A., Rutten-Jacobs, L.,
855 Giese, A. K., Van Der Laan, S. W., Gretarsdottir, S., Anderson, C. D., Chong, M., Adams, H. H. H.,
856 Ago, T., Almgren, P., Amouyel, P., Ay, H., Bartz, T. M., Benavente, O. R., ... Yamaji, T. (2018).

- 857 Multiancestry genome-wide association study of 520,000 subjects identifies 32 loci associated with
858 stroke and stroke subtypes. *Nature Genetics* 2018 50:4, 50(4), 524–537.
859 <https://doi.org/10.1038/s41588-018-0058-3>
- 860 Manichaikul, A., Mychaleckyj, J. C., Rich, S. S., Daly, K., Sale, M., & Chen, W. M. (2010). Robust
861 relationship inference in genome-wide association studies. *Bioinformatics*, 26(22), 2867–2873.
862 <https://doi.org/10.1093/BIOINFORMATICS/BTQ559>
- 863 McCarthy, S., Das, S., Kretzschmar, W., Delaneau, O., Wood, A. R., Teumer, A., Kang, H. M.,
864 Fuchsberger, C., Danecek, P., Sharp, K., Luo, Y., Sidore, C., Kwong, A., Timpson, N., Koskinen,
865 S., Vrieze, S., Scott, L. J., Zhang, H., Mahajan, A., ... Marchini, J. (2016). A reference panel of
866 64,976 haplotypes for genotype imputation. *Nature Genetics* 2016 48:10, 48(10), 1279–1283.
867 <https://doi.org/10.1038/ng.3643>
- 868 Michailidou, K., Lindström, S., Dennis, J., Beesley, J., Hui, S., Kar, S., Lemaçon, A., Soucy, P., Glubb,
869 D., Rostamianfar, A., Bolla, M. K., Wang, Q., Tyrer, J., Dicks, E., Lee, A., Wang, Z., Allen, J.,
870 Keeman, R., Eilber, U., ... Easton, D. F. (2017). Association analysis identifies 65 new breast
871 cancer risk loci. *Nature*, 551(7678), 92–94. <https://doi.org/10.1038/nature24284>
- 872 Nelson, C. P., Goel, A., Butterworth, A. S., Kanoni, S., Webb, T. R., Marouli, E., Zeng, L., Ntalla, I., Lai,
873 F. Y., Hopewell, J. C., Giannakopoulou, O., Jiang, T., Hamby, S. E., Di Angelantonio, E., Assimes,
874 T. L., Bottinger, E. P., Chambers, J. C., Clarke, R., Palmer, C. N. A., ... Deloukas, P. (2017).
875 Association analyses based on false discovery rate implicate new loci for coronary artery disease.
876 *Nature Genetics* 2017 49:9, 49(9), 1385–1391. <https://doi.org/10.1038/ng.3913>
- 877 O’Connell, J., Gurdasani, D., Delaneau, O., Pirastu, N., Ulivi, S., Cocca, M., Traglia, M., Huang, J.,
878 Huffman, J. E., Rudan, I., McQuillan, R., Fraser, R. M., Campbell, H., Polasek, O., Asiki, G.,
879 Ekoru, K., Hayward, C., Wright, A. F., Vitart, V., ... Marchini, J. (2014). A General Approach for
880 Haplotype Phasing across the Full Spectrum of Relatedness. *PLOS Genetics*, 10(4), e1004234.
881 <https://doi.org/10.1371/JOURNAL.PGEN.1004234>
- 882 Pan, G., Simpson, S., Van Der Mei, I., Charlesworth, J. C., Lucas, R., Ponsonby, A. L., Zhou, Y., Wu, F.,
883 & Taylor, B. V. (2016). Role of genetic susceptibility variants in predicting clinical course in
884 multiple sclerosis: a cohort study. *Journal of Neurology, Neurosurgery, and Psychiatry*, 87(11),
885 1204–1211. <https://doi.org/10.1136/JNNP-2016-313722>
- 886 Patel, R. S., Schmidt, A. F., Tragante, V., McCubrey, R. O., Holmes, M. V., Howe, L. J., Direk, K.,
887 Åkerblom, A., Leander, K., Virani, S. S., Kaminski, K. A., Muehlschlegel, J. D., Dubé, M. P.,
888 Allayee, H., Almgren, P., Alver, M., Baranova, E. V., Behloul, H., Boeckx, B., ... Asselbergs, F.
889 W. (2019). Association of Chromosome 9p21 With Subsequent Coronary Heart Disease Events: A
890 GENIUS-CHD Study of Individual Participant Data. *Circulation. Cardiovascular Genetics*, 12(4),
891 e002471. <https://doi.org/10.1161/CIRCGEN.119.002471>
- 892 Pirinen, M. (2023). Genetics and population analysis linemodels: clustering effects based on linear
893 relationships. *Bioinformatics*, 39(3). <https://doi.org/10.1093/bioinformatics/btad115>
- 894 *Prediction within Ancestral Diversity*. (n.d.). Retrieved September 13, 2023, from
895 <https://opain.github.io/GenoPred/DiverseAncestry.html>
- 896 Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A. R., Bender, D., Maller, J., Sklar, P.,
897 de Bakker, P. I. W., Daly, M. J., & Sham, P. C. (2007). PLINK: A Tool Set for Whole-Genome
898 Association and Population-Based Linkage Analyses. *The American Journal of Human Genetics*,
899 81(3), 559–575. <https://doi.org/10.1086/519795>
- 900 *Quality Control (QC) | Pan UKBB*. (n.d.). Retrieved September 13, 2023, from [https://pan-](https://pan-dev.ukbb.broadinstitute.org/docs/qc/index.html)
901 [dev.ukbb.broadinstitute.org/docs/qc/index.html](https://pan-dev.ukbb.broadinstitute.org/docs/qc/index.html)
- 902 Roselli, C., Chaffin, M. D., Weng, L. C., Aeschbacher, S., Ahlberg, G., Albert, C. M., Almgren, P.,
903 Alonso, A., Anderson, C. D., Aragam, K. G., Arking, D. E., Barnard, J., Bartz, T. M., Benjamin, E.
904 J., Bihlmeyer, N. A., Bis, J. C., Bloom, H. L., Boerwinkle, E., Bottinger, E. B., ... Ellinor, P. T.
905 (2018). Multi-ethnic genome-wide association study for atrial fibrillation. *Nature Genetics* 2018
906 50:9, 50(9), 1225–1233. <https://doi.org/10.1038/s41588-018-0133-9>

- 907 Rubinacci, S., Ribeiro, D. M., Hofmeister, R. J., & Delaneau, O. (2021). Efficient phasing and imputation
908 of low-coverage sequencing data using large reference panels. *Nature Genetics* 2021 53:1, 53(1),
909 120–126. <https://doi.org/10.1038/s41588-020-00756-0>
- 910 Schumacher, F. R., Al Olama, A. A., Berndt, S. I., Benlloch, S., Ahmed, M., Saunders, E. J., Dadaev, T.,
911 Leongamornlert, D., Anokian, E., Cieza-Borrella, C., Goh, C., Brook, M. N., Sheng, X., Fachal, L.,
912 Dennis, J., Tyrer, J., Muir, K., Lophatananon, A., Stevens, V. L., ... Eeles, R. A. (2018).
913 Association analyses of more than 140,000 men identify 63 new prostate cancer susceptibility loci.
914 *Nature Genetics* 2018 50:7, 50(7), 928–936. <https://doi.org/10.1038/s41588-018-0142-8>
- 915 Shah, S., Henry, A., Roselli, C., Lin, H., Sveinbjörnsson, G., Fatemifar, G., Hedman, Å. K., Wilk, J. B.,
916 Morley, M. P., Chaffin, M. D., Helgadottir, A., Verweij, N., Dehghan, A., Almgren, P., Andersson,
917 C., Aragam, K. G., Ärnlöv, J., Backman, J. D., Biggs, M. L., ... Lumbers, R. T. (2020). Genome-
918 wide association and Mendelian randomisation analysis provide insights into the pathogenesis of
919 heart failure. *Nature Communications* 2020 11:1, 11(1), 1–12. <https://doi.org/10.1038/s41467-019-13690-5>
- 920
921 Smith, B. H., Campbell, A., Linksted, P., Fitzpatrick, B., Jackson, C., Kerr, S. M., Deary, I. J., MacIntyre,
922 D. J., Campbell, H., McGilchrist, M., Hocking, L. J., Wisely, L., Ford, I., Lindsay, R. S., Morton,
923 R., Palmer, C. N. A., Dominiczak, A. F., Porteous, D. J., & Morris, A. D. (2013). Cohort Profile:
924 Generation Scotland: Scottish Family Health Study (GS:SFHS). The study, its participants and their
925 potential for genetic research on health and illness. *International Journal of Epidemiology*, 42(3),
926 689–700. <https://doi.org/10.1093/IJE/DYS084>
- 927 Spiliopoulou, A., Colombo, M., Orchard, P., Agakov, F., & McKeigue, P. (2017). GeneImp: Fast
928 Imputation to Large Reference Panels Using Genotype Likelihoods from Ultralow Coverage
929 Sequencing. *Genetics*, 206(1), 91–104. <https://doi.org/10.1534/GENETICS.117.200063>
- 930 Su, Z., Marchini, J., & Donnelly, P. (2011). HAPGEN2: simulation of multiple disease SNPs.
931 *Bioinformatics*, 27(16), 2304–2305. <https://doi.org/10.1093/bioinformatics/btr341>
- 932 Taliun, D., Harris, D. N., Kessler, M. D., Carlson, J., Szpiech, Z. A., Torres, R., Taliun, S. A. G., Corvelo,
933 A., Gogarten, S. M., Kang, H. M., Pitsillides, A. N., LeFaive, J., Lee, S. been, Tian, X., Browning,
934 B. L., Das, S., Emde, A. K., Clarke, W. E., Loesch, D. P., ... Abecasis, G. R. (2021). Sequencing of
935 53,831 diverse genomes from the NHLBI TOPMed Program. *Nature* 2021 590:7845, 590(7845),
936 290–299. <https://doi.org/10.1038/s41586-021-03205-y>
- 937 Tan, M. M., Lawton, M. A., Pollard, M. I., Brown, E., Bekadar, S., Jabbari, E., Reynolds, R. H., Iwaki,
938 H., Blauwendraat, C., Kanavou, S., Hubbard, L., Malek, N., Grosset, K. A., Bajaj, N., Barker, R. A.,
939 Burn, D. J., Bresner, C., Foltynie, T., Wood, N. W., ... Morris, H. R. (2022). *Genome-wide
940 determinants of mortality and clinical progression in Parkinson's disease.*
941 <https://doi.org/10.1101/2022.07.07.22277297>
- 942 Tcheandjieu, C., Zhu, X., Hilliard, A. T., Clarke, S. L., Napolioni, V., Ma, S., Lee, K. M., Fang, H., Chen,
943 F., Lu, Y., Tsao, N. L., Raghavan, S., Koyama, S., Gorman, B. R., Vujkovic, M., Klarin, D., Levin,
944 M. G., Sinnott-Armstrong, N., Wojcik, G. L., ... Assimes, T. L. (2022). Large-scale genome-wide
945 association study of coronary artery disease in genetically diverse populations. *Nature Medicine*
946 2022 28:8, 28(8), 1679–1692. <https://doi.org/10.1038/s41591-022-01891-3>
- 947 Timmers, P. R. H. J., Mounier, N., Lall, K., Fischer, K., Ning, Z., Feng, X., Bretherick, A. D., Clark, D.
948 W., Shen, X., Esko, T., Kutalik, Z., Wilson, J. F., & Joshi, P. K. (2019). Genomics of 1 million
949 parent lifespans implicates novel pathways and common diseases and distinguishes survival
950 chances. *ELife*, 8, 1–40. <https://doi.org/10.7554/ELIFE.39856>
- 951 Turnbull, C. (2018). Introducing whole-genome sequencing into routine cancer care: the Genomics
952 England 100 000 Genomes Project. *Annals of Oncology : Official Journal of the European Society
953 for Medical Oncology*, 29(4), 784–787. <https://doi.org/10.1093/ANNONC/MDY054>
- 954 UK Biobank. (2015, October). *Genotyping and quality control of UK Biobank, a large-scale, extensively
955 phenotyped prospective resource.*
956 https://biobank.ctsu.ox.ac.uk/crystal/crystal/docs/genotyping_qc.pdf

- 957 Vandebergh, M., Andlauer, T. F. M., Zhou, Y., Mallants, K., Held, F., Aly, L., Taylor, B. V., Hemmer,
958 B., Dubois, B., & Goris, A. (2021). Genetic Variation in WNT9B Increases Relapse Hazard in
959 Multiple Sclerosis. *Annals of Neurology*, 89(5), 884–894. <https://doi.org/10.1002/ANA.26061>
- 960 Viippola, E., Kuitunen, S., Rodosthenous, R. S., Vabalas, A., Hartonen, T., Vartiainen, P., Demmler, J.,
961 Vuorinen, A.-L., Liu, A., Havulinna, A. S., Llorens, V., Detrois, K. E., Wang, F., Ferro, M.,
962 Karvanen, A., German, J., Jukarainen, S., Gracia-Tabuenca, J., Hiekkalinna, T., ... Perola, M.
963 (2023). Data Resource Profile: Nationwide registry data for high-throughput epidemiology and
964 machine learning (FinRegistry). *International Journal of Epidemiology*, 52(4), e195–e200.
965 <https://doi.org/10.1093/IJE/DYAD091>
- 966 Wightman, D. P., Jansen, I. E., Savage, J. E., Shadrin, A. A., Bahrami, S., Holland, D., Rongve, A.,
967 Børte, S., Winsvold, B. S., Drange, O. K., Martinsen, A. E., Skogholt, A. H., Willer, C., Bråthen, G.,
968 Bosnes, I., Nielsen, J. B., Fritsche, L. G., Thomas, L. F., Pedersen, L. M., ... Posthuma, D. (2021).
969 A genome-wide association study with 1,126,563 individuals identifies new risk loci for
970 Alzheimer’s disease. *Nature Genetics* 2021 53:9, 53(9), 1276–1282. [https://doi.org/10.1038/s41588-](https://doi.org/10.1038/s41588-021-00921-z)
971 [021-00921-z](https://doi.org/10.1038/s41588-021-00921-z)
- 972 Willer, C., Li, Y., & Abecasis, G. (2010). METAL: fast and efficient meta-analysis of genomewide
973 association scans. *Bioinformatics*. <http://bioinformatics.oxfordjournals.org/content/26/17/2190.short>
- 974 World Health Organization. (2004). *ICD-10 : international statistical classification of diseases and*
975 *related health problems : tenth revision*. <https://apps.who.int/iris/handle/10665/42980>
- 976 Wu, C., Kraft, P., Stolzenberg-Solomon, R., Steplowski, E., Brotzman, M., Xu, M., Mudgal, P.,
977 Amundadottir, L., Arslan, A. A., Bueno-De-Mesquita, H. B., Gross, M., Helzlsouer, K., Jacobs, E.
978 J., Kooperberg, C., Petersen, G. M., Zheng, W., Albanes, D., Boutron-Ruault, M. C., Buring, J.
979 E., ... Wolpin, B. M. (2014). Genome-wide association study of survival in patients with pancreatic
980 adenocarcinoma. *Gut*, 63(1), 152–160. <https://doi.org/10.1136/GUTJNL-2012-303477>
- 981 Wuttke, M., Li, Y., Li, M., Sieber, K. B., Feitosa, M. F., Gorski, M., Tin, A., Wang, L., Chu, A. Y.,
982 Hoppmann, A., Kirsten, H., Giri, A., Chai, J. F., Sveinbjornsson, G., Tayo, B. O., Nutile, T.,
983 Fuchsberger, C., Marten, J., Cocca, M., ... Pattaro, C. (2019). A catalog of genetic loci associated
984 with kidney function from analyses of a million individuals. *Nature Genetics* 2019 51:6, 51(6), 957–
985 972. <https://doi.org/10.1038/s41588-019-0407-x>
- 986 Yaghootkar, H., Bancks, M. P., Jones, S. E., McDaid, A., Beaumont, R., Donnelly, L., Wood, A. R.,
987 Campbell, A., Tyrrell, J., Hocking, L. J., Tuke, M. A., Ruth, K. S., Pearson, E. R., Murray, A.,
988 Freathy, R. M., Munroe, P. B., Hayward, C., Palmer, C., Weedon, M. N., ... Kutalik, Z. (2017).
989 Quantifying the extent to which index event biases influence large genetic association studies.
990 *Human Molecular Genetics*, 26(5), 1018–1030. <https://doi.org/10.1093/HMG/DDW433>
- 991 Zhang, H., Ahearn, T. U., Lecarpentier, J., Barnes, D., Beesley, J., Qi, G., Jiang, X., O’Mara, T. A., Zhao,
992 N., Bolla, M. K., Dunning, A. M., Dennis, J., Wang, Q., Ful, Z. A., Aittomäki, K., Andrulis, I. L.,
993 Anton-Culver, H., Arndt, V., Aronson, K. J., ... García-Closas, M. (2020). Genome-wide
994 association study identifies 32 novel breast cancer susceptibility loci from overall and subtype-
995 specific analyses. *Nature Genetics* 2020 52:6, 52(6), 572–581. [https://doi.org/10.1038/s41588-020-](https://doi.org/10.1038/s41588-020-0609-2)
996 [0609-2](https://doi.org/10.1038/s41588-020-0609-2)
- 997 Zhang, Q., Privé, F., Vilhjálmsson, B., & Speed, D. (2021). Improved genetic prediction of complex traits
998 from individual-level data or summary statistics. *Nature Communications* 2021 12:1, 12(1), 1–9.
999 <https://doi.org/10.1038/s41467-021-24485-y>

1000

1001

1002 Funding

1003 This study has received funding from the European Union’s Horizon 2020 research and innovation
1004 programme under grant agreement No 101016775, from the European Research Council (ERC) under the

1005 European Union’s Horizon 2020 research and innovation program (grant number 945733) and from
1006 Academy of Finland fellowship grant N. 323116.

1007

1008 **Acknowledgements**

1009 We would like to thank Prof. George Davey Smith for the invaluable comments on this manuscript.

1010 FinnGen

1011 We thank all those who contributed samples and data for the FinnGen scientific project; and P. VandeHaar
1012 for technical consultation on PheWeb. The FinnGen project is funded by two grants from Business Finland
1013 (HUS 4685/31/2016 and UH 4386/31/2016) and the following industry partners: AbbVie, AstraZeneca UK,
1014 Biogen, Bristol Myers Squibb (and Celgene Corporation & Celgene International II), Genentech, Merck
1015 Sharp & Dohme LLC, a subsidiary of Merck & Co., Inc., Rahway, NJ, USA, Pfizer, GlaxoSmithKline
1016 Intellectual Property Development, Sanofi US Services, Maze Therapeutics, Janssen Biotech, Novartis, and
1017 Boehringer Ingelheim. The following biobanks are acknowledged for delivering samples to FinnGen: Auria
1018 Biobank (<https://www.auria.fi/biopankki/>), THL Biobank (<https://www.thl.fi/biobank>), Helsinki Biobank
1019 (<https://www.helsinginbiopankki.fi>), Biobank Borealis of Northern Finland
1020 (<https://www.ppsHP.fi/Tutkimus-ja-opetus/Biopankki/Pages/Biobank-Borealis-briefly-in-English.aspx>),
1021 Finnish Clinical Biobank Tampere ([https://www.tays.fi/en-](https://www.tays.fi/en-US/Research_and_development/Finnish_Clinical_Biobank_Tampere)
1022 [US/Research_and_development/Finnish_Clinical_Biobank_Tampere](https://www.tays.fi/en-US/Research_and_development/Finnish_Clinical_Biobank_Tampere)), Biobank of Eastern Finland
1023 (<https://www.ita-suomenbiopankki.fi/en>), Central Finland Biobank ([https://www.ksshp.fi/fi-](https://www.ksshp.fi/fi-FI/Potilaalle/Biopankki)
1024 [FI/Potilaalle/Biopankki](https://www.ksshp.fi/fi-FI/Potilaalle/Biopankki)), Finnish Red Cross Blood Service Biobank
1025 (www.veripalvelu.fi/verenluovutus/biopankkitoiminta) and Terveystalo Biobank
1026 (<https://www.terveystalo.com/fi/Yritystietoa/Terveystalo-Biopankki/Biopankki/>). All Finnish biobanks are
1027 members of the BBMRI.fi infrastructure (<https://www.bbMRI.fi>). The FINBB (<https://finbb.fi/>) is the
1028 coordinator of BBMRI-ERIC operations in Finland. The Finnish biobank data can be accessed through the
1029 Fingenious services (<https://site.fingenious.fi/en/>) managed by FINBB.

1030 UK biobank

1031 Ethics approval for the UK Biobank study was obtained from the North West Centre for Research Ethics
1032 Committee (11/NW/0382). UK Biobank data used in this study were obtained under approved application
1033 78537.

1034 Estonia biobank

1035 The authors would also like to acknowledge all the recruiters and the participants of the EGCUT, the
1036 University of Tartu, the Ministry of Social Affairs, the Ministry of Science and Education, the Ministry of
1037 Economic Affairs and Communications, the Archimedes Foundation, the Estonian Biocentre, the Institute
1038 of Molecular and Cell Biology and the Centre for Ethics of the University of Tartu. The authors would also
1039 like to acknowledge the EGCUT technical personnel.

1040 Other contributors: Anneli Allik, Tarmo Annilo, Merli Hass, Atso-Heinar Jõks, Aidula-Taie Kaasik, Aime
1041 Keis, Erkki Leego, Merike Leego, Kadri Lilienthal, Kristjan Metsalu, Evelin Mihailov, Kairit Mikkil, Ene
1042 Mölder, Helja Niinemäe, Tiit Nikopensius, Mairo Puusepp, Steven Smit, Viljo Soo, Riin Tamm, Maris
1043 Teder-Laving and Maris Väli-Täht.

1044 Genomics England

1045 This research was made possible through access to data in the National Genomic Research Library, which
1046 is managed by Genomics England Limited (a wholly owned company of the Department of Health and
1047 Social Care). The National Genomic Research Library holds data provided by patients and collected by the

1048 NHS as part of their care and data collected as part of their participation in research. The National Genomic
1049 Research Library is funded by the National Institute for Health Research and NHS England. The Wellcome
1050 Trust, Cancer Research UK and the Medical Research Council have also funded research infrastructure.
1051 We acknowledge the contribution of the Genomics England Research Consortium. The members of this
1052 consortium are: John C. Ambrose¹, Prabhu Arumugam¹, Roel Bevers¹, Marta Bleda¹, Freya Boardman-
1053 Pretty^{1,2}, Christopher R. Boustred¹, Helen Brittain¹, Matt J. Brown¹, Mark J. Caulfield^{1,2}, Georgia C.
1054 Chan¹, Adam Giess¹, Angela Hamblin¹, Shirley Henderson^{1,2}, Tim J. P. Hubbard¹, Rob Jackson¹, Louise
1055 J. Jones^{1,2}, Dalia Kasperaviciute^{1,2}, Melis Kayikci¹, Athanasios Kousathanas¹, Lea Lahnstein¹, Sarah E.
1056 A. Leigh¹, Ivonne U. S. Leong¹, Javier F. Lopez¹, Fiona Maleady-Crowe¹, Meriel McEntagart¹, Federico
1057 Minneci¹, Jonathan Mitchell¹, Loukas Moutsianas^{1,2}, Michael Mueller^{1,2}, Nirupa Murugaesu¹, Anna C.
1058 Need^{1,2}, Peter O'Donovan¹, Chris A. Odhams¹, Christine Patch^{1,2}, Daniel Perez-Gill¹, Mariana
1059 Buongiorno Pereira¹, John Pullinger¹, Tahrima Rahim¹, Augusto Rendon¹, Tim Rogers¹, Kevin
1060 Savage¹, Kushmita Sawant¹, Richard H. Scott¹, Afshan Siddiq¹, Alexander Sieghart¹, Samuel C. Smith¹,
1061 Alona Sosinsky^{1,2}, Alexander Stuckey¹, Mélanie Tanguy¹, Ana Lisa Taylor Tavares¹, Ellen R. A.
1062 Thomas^{1,2}, Simon R. Thompson¹, Arianna Tucci^{1,2}, Matthew J. Welland¹, Eleanor Williams¹, Katarzyna
1063 Witkowska^{1,2}, Suzanne M. Wood^{1,2}, Magdalena Zarowiecki¹.

1064 1. Genomics England, London, UK

1065 2. William Harvey Research Institute, Queen Mary University of London, London, EC1M 6BQ, UK

1066 Generation Scotland

1067 Generation Scotland received core support from the Chief Scientist Office of the Scottish Government
1068 Health Directorates [CZD/16/6] and the Scottish Funding Council [HR03006]. Genotyping of the GS:SFHS
1069 samples was carried out by the Genetics Core Laboratory at the Wellcome Trust Clinical Research Facility,
1070 Edinburgh, Scotland and was funded by the Medical Research Council UK and the Wellcome Trust
1071 (Wellcome Trust Strategic Award “Stratifying Resilience and Depression Longitudinally” (STRADL)
1072 Reference 104036/Z/14/Z).

1073 We are grateful to all the families who took part, the general practitioners and the Scottish School of Primary
1074 Care for their help in recruiting them, and the whole Generation Scotland team, which includes
1075 interviewers, computer and laboratory technicians, clerical workers, research scientists, volunteers,
1076 managers, receptionists, healthcare assistants and nurses.

1077 Ethical approval for the GS:SFHS study was obtained from the Tayside Committee on Medical Research
1078 Ethics (on behalf of the National Health Service).

1079 Ethical approval for the GS:3D study was obtained from the Tayside Committee on Medical Research
1080 Ethics (on behalf of the National Health Service).

1081 Ethical approval for the GS:21CGH study was obtained from the Scotland A Research Ethics Committee.

1082 Genes & Health

1083 Genes & Health is/has recently been core-funded by Wellcome (WT102627, WT210561), the Medical
1084 Research Council (UK) (M009017, MR/X009777/1, MR/X009920/1), Higher Education Funding Council
1085 for England Catalyst, Barts Charity (845/1796), Health Data Research UK (for London substantive site),
1086 and research delivery support from the NHS National Institute for Health Research Clinical Research
1087 Network (North Thames). Genes & Health is/has recently been funded by Alnylam Pharmaceuticals,
1088 Genomics PLC; and a Life Sciences Industry Consortium of Astra Zeneca PLC, Bristol-Myers Squibb

1089 Company, GlaxoSmithKline Research and Development Limited, Maze Therapeutics Inc, Merck Sharp &
1090 Dohme LLC, Novo Nordisk A/S, Pfizer Inc, Takeda Development Centre Americas Inc.

1091 We thank Social Action for Health, Centre of The Cell, members of our Community Advisory Group, and
1092 staff who have recruited and collected data from volunteers. We thank the NIHR National Biosample Centre
1093 (UK Biocentre), the Social Genetic & Developmental Psychiatry Centre (King's College London),
1094 Wellcome Sanger Institute, and Broad Institute for sample processing, genotyping, sequencing and variant
1095 annotation.

1096 We thank: Barts Health NHS Trust, NHS Clinical Commissioning Groups (City and Hackney, Waltham
1097 Forest, Tower Hamlets, Newham, Redbridge, Havering, Barking and Dagenham), East London NHS
1098 Foundation Trust, Bradford Teaching Hospitals NHS Foundation Trust, Public Health England (especially
1099 David Wyllie), Discovery Data Service/Endeavour Health Charitable Trust (especially David Stables),
1100 Voror Health Technologies Ltd (especially Sophie Don), NHS England (for what was NHS Digital) - for
1101 GDPR-compliant data sharing backed by individual written informed consent.

1102 BioMe Biobank

1103 This work was supported in part through the computational and data resources and staff expertise provided
1104 by Scientific Computing and Data at the Icahn School of Medicine at Mount Sinai and supported by the
1105 Clinical and Translational Science Awards (CTSA) grant UL1TR004419 from the National Center for
1106 Advancing Translational Sciences. Additionally, this work was supported by the Office of Research
1107 Infrastructure of the National Institutes of Health under award number S10OD026880, which allowed us
1108 to use Mount Sinai Data Warehouse (MSDW) data. Regarding HPI.MS resources, funding was provided
1109 by the Hasso Plattner Foundation (HPF). The Mount Sinai BioMe Biobank has been supported by The
1110 Andrea and Charles Bronfman Philanthropies and in part by Federal funds from the NHLBI and NHGRI
1111 (U01HG00638001; U01HG007417; X01HL134588). We thank all participants in the Mount Sinai BioMe
1112 Biobank. We also thank all of our recruiters who have assisted in data collection and management and are
1113 grateful for the computational resources and staff expertise provided by Scientific Computing at the Icahn
1114 School of Medicine at Mount Sinai.

1115

1116 **Supplementary Method**

1117 **Data and resources**

1118 FinnGen

1119 *Genotyping and quality control*

1120 FinnGen consists of prospectively recruited samples and a series of legacy cohorts with genotypes already
1121 available (*FinnGen Project* | *FinnGen*, n.d.). Prospective samples were genotyped using the
1122 ThermoFisher Axiom custom array which tags a total of 655,973 variants. Genotype calling was performed
1123 using the Array Power Tools software. Legacy cohorts were genotyped using various Illumina arrays and
1124 genotype calling was performed using either GenCall or zCall algorithms.

1125 For both prospective and legacy cohorts the following quality control metrics were used.

1126 Samples were removed if:

- 1127 ● Pihat was > 0.9 and the samples were not monozygotic or replicates
- 1128 ● There was a discrepancy between reported sex and genetically determined sex (F-value ≤ 0.3 for
1129 females and ≥ 0.8 for males)
- 1130 ● Missingness was $\geq 5\%$

- 1131 • Heterozygosity was ± 4 standard deviations from the population average
- 1132 • Pihat was > 0.1 with 14 or more samples
- 1133 • Samples were ± 4 standard deviations away from the population average according to the first two
- 1134 genetic principal components.
- 1135 Samples were tagged should there be evidence of a mendelian error or contain replicate samples with over
- 1136 50,000 discrepancies.
- 1137 Variants were removed if:
- 1138 • The variant failed the Hardy-Weinberg Equilibrium test (p -value $< 10^{-6}$)
- 1139 • The variant had a call rate $< 98\%$

1140 *Imputation*

1141 Pre-phasing was performed using Eagle 2.3.5 (Loh et al., 2016) and samples were imputed using the SiSu

1142 v3 imputation reference panel. This reference panel is specific to the Finnish population, containing high-

1143 coverage (25-30x) whole-genome sequencing data from 3,775 Finns and 16,962,023 variants with minor

1144 allele count ≥ 3 . After imputation, 16,387,711 variants were imputed with high quality (INFO > 0.6).

1145 *Ancestry assignment*

1146 Firstly, the FinnGen samples were combined with the 1000 genomes phase 3 dataset. Genetic principal

1147 components were calculated using a subset of 49,451 pruned SNPs. Aberrant (Bellenguez et al., 2012) was

1148 used to identify and remove samples that deviated from the main cluster. A probability of belonging to

1149 either a North-Western European or Finnish population was calculated by firstly performing PCA with

1150 individuals belonging to these ancestries from 1000 genomes data. FinnGen samples were then projected

1151 onto this PCA space and Mahalanobis distances calculated for each sample against each of the two

1152 ancestries. Samples were retained if there was $\geq 95\%$ probability of belonging to the Finnish ancestry

1153 cluster.

1154 *Ethics statement*

1155 Patients and control subjects in FinnGen provided informed consent for biobank research, based on the

1156 Finnish Biobank Act. Alternatively, separate research cohorts, collected prior the Finnish Biobank Act

1157 came into effect (in September 2013) and start of FinnGen (August 2017), were collected based on study-

1158 specific consents and later transferred to the Finnish biobanks after approval by Fimea (Finnish Medicines

1159 Agency), the National Supervisory Authority for Welfare and Health. Recruitment protocols followed the

1160 biobank protocols approved by Fimea. The Coordinating Ethics Committee of the Hospital District of

1161 Helsinki and Uusimaa (HUS) statement number for the FinnGen study is Nr HUS/990/2017.

1162 The FinnGen study is approved by Finnish Institute for Health and Welfare (permit numbers:

1163 THL/2031/6.02.00/2017, THL/1101/5.05.00/2017, THL/341/6.02.00/2018, THL/2222/6.02.00/2018,

1164 THL/283/6.02.00/2019, THL/1721/5.05.00/2019 and THL/1524/5.05.00/2020), Digital and population

1165 data service agency (permit numbers: VRK43431/2017-3, VRK/6909/2018-3, VRK/4415/2019-3), the

1166 Social Insurance Institution (permit numbers: KELA 58/522/2017, KELA 131/522/2018, KELA

1167 70/522/2019, KELA 98/522/2019, KELA 134/522/2019, KELA 138/522/2019, KELA 2/522/2020, KELA

1168 16/522/2020), Findata permit numbers THL/2364/14.02/2020, THL/4055/14.06.00/2020,

1169 THL/3433/14.06.00/2020, THL/4432/14.06/2020, THL/5189/14.06/2020, THL/5894/14.06.00/2020,

1170 THL/6619/14.06.00/2020, THL/209/14.06.00/2021, THL/688/14.06.00/2021, THL/1284/14.06.00/2021,

1171 THL/1965/14.06.00/2021, THL/5546/14.02.00/2020, THL/2658/14.06.00/2021, THL/4235/14.06.00/2021,

1172 Statistics Finland (permit numbers: TK-53-1041-17 and TK/143/07.03.00/2020 (earlier TK-53-90-20)

1173 TK/1735/07.03.00/2021, TK/3112/07.03.00/2021) and Finnish Registry for Kidney Diseases
1174 permission/extract from the meeting minutes on 4th July 2019.

1175 The Biobank Access Decisions for FinnGen samples and data utilized in FinnGen Data Freeze 10 include:
1176 THL Biobank BB2017_55, BB2017_111, BB2018_19, BB_2018_34, BB_2018_67, BB2018_71,
1177 BB2019_7, BB2019_8, BB2019_26, BB2020_1, BB2021_65, Finnish Red Cross Blood Service Biobank
1178 7.12.2017, Helsinki Biobank HUS/359/2017, HUS/248/2020, HUS/150/2022 § 12, §13, §14, §15, §16, §17,
1179 §18, and §23, Auria Biobank AB17-5154 and amendment #1 (August 17 2020) and amendments BB_2021-
1180 0140, BB_2021-0156 (August 26 2021, Feb 2 2022), BB_2021-0169, BB_2021-0179, BB_2021-0161,
1181 AB20-5926 and amendment #1 (April 23 2020)and it's modification (Sep 22 2021), Biobank Borealis of
1182 Northern Finland_2017_1013, 2021_5010, 2021_5018, 2021_5015, 2021_5023, 2021_5017, 2022_6001,
1183 Biobank of Eastern Finland 1186/2018 and amendment 22 § /2020, 53§/2021, 13§/2022, 14§/2022,
1184 15§/2022, Finnish Clinical Biobank Tampere MH0004 and amendments (21.02.2020 & 06.10.2020),
1185 §8/2021, §9/2022, §10/2022, §12/2022, §20/2022, §21/2022, §22/2022, §23/2022, Central Finland Biobank
1186 1-2017, and Terveystalo Biobank STB 2018001 and amendment 25th Aug 2020, Finnish Hematological
1187 Registry and Clinical Biobank decision 18th June 2021, Arctic biobank P0844: ARC_2021_1001.

1188 UK Biobank

1189 *Genotyping and quality control*

1190 UK Biobank participants were genotyped by two genotyping arrays: The UK Biobank Lung Exome Variant
1191 Evaluation (UKBiLEVE) Axiom array was used to genotype 49,950 participants. The remaining 438,427
1192 participants were genotypes using the Applied Biosystems UK Biobank Axiom Array. Principal
1193 Component Analysis (PCA) was performed on the genetic data and centralised quality control (QC) on
1194 variants was performed on individuals identified to belong to the largest cluster (N=463,844) according to
1195 Aberrant - an unsupervised clustering algorithm (Bellenguez et al., 2012). Variants were assessed for
1196 evidence of allele frequency variation across batch, plate, sex or array and that genotypes were largely
1197 consistent with Hardy-Weinberg Equilibrium expectations (all p-value thresholds $< 10^{-12}$). If a variant failed
1198 one or more tests within a given batch it was set to missing. See (UK Biobank, 2015) for more detailed
1199 information on testing.

1200 *Imputation*

1201 For 487,442 individuals, imputation was performed using the IMPUTE4 (Howie et al., 2009) software.
1202 Genetic variation from the Haplotype Reference Consortium (HRC) (McCarthy et al., 2016) and merged
1203 UK10K+1000 Genomes (1000 Genomes Project Consortium, 2015) were used as a reference panel. Single
1204 Nucleotide Polymorphisms (SNPs) were only included in the final imputation if they were present in both
1205 reference panels, giving a total of 96,959,328 SNPs.

1206 *Ancestry assignment*

1207 Ancestry assignment uses methodology and scripts from GenoPred (*Prediction within Ancestral*
1208 *Diversity*, n.d.). Individuals were stratified into one of five super populations African (AFR), American
1209 (AMR), South Asian (SAS), East Asian (EAS) and European (EUR). The 1000 Genomes data (1000
1210 Genomes Project Consortium, 2015) acted as a reference given the individuals are known to belong to one
1211 of the 5 super populations. Only unambiguous SNPs also present in both the HapMap3 consortium (Gibbs
1212 et al., 2003) and the imputed UK Biobank data were retained for PCA. SNPs within both the reference
1213 (1000 Genomes) and target (UK Biobank) samples underwent quality control such that the minor allele
1214 frequency (MAF) $> 5\%$, variant missingness $> 2\%$ and Hardy-Weinberg Equilibrium p-value $> 1e^{-6}$.

1215 467,970 autosomal SNPs remained following QC and were in the intersection of the reference and target
1216 samples. Regions with long range linkage disequilibrium were excluded and independent SNPs (SNPs
1217 greater than 1000kb apart and $r^2 < 0.2$) retained. PCA was then performed in the reference sample using
1218 PLINK v2 (Purcell et al., 2007) and a multinomial elastic-net regression was trained using 5-fold cross
1219 validation, super population as the outcome and the first 10 PCs as covariates. PCs from the target sample
1220 were then projected into the reference space and prediction on super population made. Classifications were
1221 made according to the super population with the greatest probability. To be classified the max probability
1222 must be over 0.5, otherwise it was set to missing.

1223 PCA was performed using a random subset of 1000 individuals per super population and PC's from the rest
1224 of the super population sample projected onto this space. Distances from the centroid were calculated and
1225 outliers removed. Outliers were classified as having a distance > 75 percentile + $30 \times$ Interquartile Range.
1226 Following within-ancestry QC, 8,381, 1,063, 2,393, 447,332 and 9,435 individuals were allocated to AFR,
1227 AMR, EAS, EUR and SAS super populations respectively.

1228 Estonian Biobank

1229 *Genotyping and quality control*

1230 Estonian BioBank (EstBB) samples were genotyped with 4 sub-versions of Infinium Global Screening
1231 Array-24. Samples with less than 95% call-rate were excluded. Sample sex recorded in the EstBB database
1232 was compared with genetic sex. Samples with sex mismatch were further inspected for sex chromosome
1233 abnormalities (X0, XXY, etc.), and samples with confirmed database vs genetic sex mismatch were
1234 excluded. In total, 202 910 individuals passed sample quality control. SNP quality control was performed
1235 by excluding: (a) all SNPs with less than 95% call-rate, (b) SNPs showing more than 5% AF difference
1236 from the AF mean estimated using all genotyping batches with more than 10 000 samples per batch, (c)
1237 SNPs with Illumina GenTrain score < 0.6 or cluster separation score < 0.4 in any genotyping batch, (d)
1238 autosomal SNPs with HWE exact test p-value $< 1e-4$. In total, approximately 328K autosomal and X-
1239 chromosome SNPs with MAF $> 1\%$ passed quality control and were used in the imputation. All the variants
1240 were processed on the human genome assembly GRCh37.

1241 *Imputation*

1242 Imputation was performed using a local Estonian imputation reference panel made of 2056 WGS samples.
1243 Genotypes were pre-phased with Eagle v2.4.1 and imputed with Beagle 5.1 using default parameters.
1244 Multiallelic positions were excluded from imputation output. In total, 39 546 641 variants were used in the
1245 study.

1246 *Ancestry assignment*

1247 EstBB samples were combined with the 1000 genomes phase 3 dataset for ancestry analysis. Genetic
1248 principal components were calculated using a subset of quality controlled and pruned genotyped SNPs.
1249 This was further used to identify and remove samples that deviated from the main cluster via visual
1250 inspection. In total, 481 non-european ancestry individuals based on principal components were excluded
1251 from the analysis.

1252 Genomics England

1253 *Genotyping and quality control*

1254 Genome sequencing was performed in DNA samples from 78,195 individuals using Illumina HiSeq X
1255 systems (150bp paired-end format). Reads were aligned using the iSAAC Aligner (version 03.16.02.19)

1256 and small variants were called using Starling Small Variant Caller (version 2.4.7). Samples were aligned to
1257 the Homo Sapiens NCBI GRCh38 assembly with decoys.

1258 Aggregation of single-sample gVCFs was performed using the Illumina software gVCF genotyper (version
1259 2019). Variant normalisation and decomposition were implemented by vt (version 0.57721). Genomic
1260 annotation and calculation of allele statistics were performed using Ensembl VEP and bcftools respectively.
1261 The multi-sample VCF dataset (aggV2) was then split into 1,371 roughly equal chunks to allow faster
1262 processing. Only variants that passed all provided site quality control criteria were processed.

1263 *Imputation*

1264 The WGS genotypes (~722M variants) were filtered to a variant base list used for PGS model generation,
1265 which includes 18,421,839 variants. (For further information on how the variant list was derived see:
1266 <https://research-help.genomicsengland.co.uk/pages/viewpage.action?pageId=72351761>)

1267 Genotypes were phased and imputed using the 1000G reference panel (v5a) which was lifted-over from
1268 GRCh37 to GRCh38 using cross-map.

1269 *Ancestry assignment*

1270 The genetic ancestry of the patients was estimated using a random forest classifier and data from 1000
1271 genomes project phase 3 (1KGP3) dataset. Firstly, all unrelated samples from the 1KGP3 were selected and
1272 188,382 HQ SNPs were subsetted. After filtering for MAF > 0.05 in 1KGP3 (and GE data), the first 20 PCs
1273 were calculated using GCTA and the aggV2 data were projected onto the 1KGP3 PC loadings. The random
1274 forest model to predict ancestries was trained based on:

- 1275 A. First 8 1KGP3 PCs
- 1276 B. set Ntrees = 400
- 1277 C. Train and predict on 1KGP3 Admixed American, African, East Asian, European, and South Asian
1278 super-populations.

1279 Individuals were assigned for any one ancestry with a probability of > 0.8.

1280 Genes and Health

1281 *Genotyping and quality control*

1282 We used the latest 2021 July GNH data release including 44,190 individuals (26,537 British-Bangladeshi,
1283 17,653 British-Pakistani). Genotyping was performed on DNA samples from saliva, using the Illumina
1284 Infinium Global Screening Array v3, which contained 730,059 variants. GenomeStudio from Illumina was
1285 used to perform clustering and initial quality control on the genotype data. Variants were removed if they
1286 had low call rate, or were tagging structural variants, a positive HetExcess > 0.03, Hardy-Weinberg
1287 equilibrium P-value < 1.0×10^{-6} , cluster sep < 0.57, or automated clustering (GenTrain) score <= 0.7. A
1288 total of 637,829 variants remained with call rates of > 0.992 for female samples and > 0.995 for male
1289 samples (including X and Y chromosomes). Sample exclusion criteria included duplicate GSA genotypes
1290 that should not be sample duplicates, samples that should be duplicated but have not matching GSA
1291 genotypes, and a few late withdrawals of consent. Only chip genotyped samples with valid NHS numbers
1292 were preserved. When two chip genotype samples with the same NHS number were found, the samples
1293 with the highest call rate were retained.

1294 *Imputation*

1295 Monomorphic SNPs, non-ACGT, palindromic (A/T, T/A, C/G, G/C), and chr Y variants were excluded.
1296 Variants were evaluated by TOPMed QC to obtain SNPs that required strand flipping (performed in plink).
1297 Furthermore, variants with MAF < 0.0001 were excluded. The TOPMed-r2 Minimac4 Imputation Server

1298 (version 1.5.7, <https://imputation.biodatacatalyst.nhlbi.nih.gov/#!/pages/home>), created by the University of
1299 Michigan, was subsequently used to impute the genotypes. Rsq filter (imputation quality) of 0.3 was applied
1300 within the Imputation Server.

1301 *Ancestry assignment*

1302 A total of 44,396 individuals and 355,862 directly genotyped variants (retaining only autosomal variants,
1303 MAF>0.01, call rate >99% and those passing HWE in declared Bangladeshi individuals) were used with
1304 the KING software to estimate pairwise relationship up to 4 degrees. PCA was performed on GNH unrelated
1305 individuals, projecting related individuals into the PC, to obtain 50 PCs for all GNH samples. For the
1306 ancestry assignment, we used a reference cohort consisting of 3,433 individuals from 1000G and HGDP.
1307 A PCA up to 50 PCs was performed on the reference set (3,433 individuals and 104,552 variants) and
1308 subsequently the GNH samples were projected into the reference PCA. Using UMAP with 7 PCs, we
1309 genetically inferred Bangladeshi and Pakistani individuals and excluded 76 non South Asian outliers and
1310 130 South Asian outliers (not falling into the main clusters).

1311 Generation Scotland

1312 *Genotyping and quality control*

1313 Generation Scotland (GS) consists of ~24,000 individuals from across Scotland aged between 18-99 years.
1314 Phenotypic data were obtained at baseline along with whole blood samples for DNA quantification. Disease
1315 outcomes were ascertained through linkage to primary (GP) and secondary (hospital) healthcare records.

1316 Genotype data was assayed for 20,195 participants in two batches with 9,863 participants in the first batch
1317 and the remainder in the second. The genotyping was performed using the Illumina
1318 HumanOmniExpressExome-8 v1.0 BeadChip and the Illumina HumanOmniExpressExome-8 v1.2
1319 BeadChip, respectively. Individuals or SNPs with a low call rate (<98%) and SNPs with Hardy-Weinberg
1320 p -value< 1×10^{-6} were removed. Mendelian errors were removed by setting the individual-level genotypes at
1321 erroneous SNPs to missing.

1322 *Imputation*

1323 Genotyped data were imputed using the HRC panel v1.1 (McCarthy et al., 2016). Autosomal haplotypes
1324 were checked to ensure consistency with the reference panel (strand orientation, reference allele, position).
1325 Pre-phasing was performed using Shapeit2 v2r837 (O'Connell et al., 2014) using the Shapeit2 duohmm
1326 option11 (O'Connell et al., 2014) and cohort family structure in order to improve imputation quality
1327 (O'Connell et al., 2014). Variants with low imputation quality (INFO<0.4) as well as monogenic variants
1328 were removed from the imputed set resulting in 24,111,857 variants for downstream analysis.

1329 *Ancestry assignment*

1330 Ancestry outliers were removed from the dataset. These were defined as individuals who were more than
1331 six standard deviations away from the mean in a principal component analysis of GS merged with 1092
1332 participants from the 1000 Genomes Project (1000 Genomes Project Consortium, 2015).

1333 Dana Farber

1334 *Genotyping and quality control*

1335 DNA samples were processed from the whole blood and genotyped on either the Illumina Multi-Ethnic
1336 Genotyping Array (MEGA), the Expanded Multi-Ethnic Genotyping Array (MEGA Ex) array, or the Multi-
1337 Ethnic Global (MEG) BeadChip (Bien et al., 2016). All germline samples were imputed to the Haplotype
1338 Reference Consortium (HRC) reference panel (McCarthy et al., 2016) and then restricted to ~1.1 million
1339 HapMap3 variants that typically exhibit high imputation accuracy across genotyping platforms and

1340 uniformly tag common SNP variation (Finucane et al., 2015). Small indels were not available in the HRC
1341 reference panel due to sequencing ambiguity, and we additionally imputed small indels into the germline
1342 genotyped data using the 1000 Genomes Phase 3 reference panel (1000 Genomes Project Consortium,
1343 2015) and restricted to high-quality indels with INFO score (imputation confidence score) > 0.9.

1344 *Imputation*

1345 We assessed three imputation algorithms intended for low-coverage data: STITCH v1.5.3 (Davies et al.,
1346 2016), GLIMPSE v1.0.0 (Davies et al., 2021; Rubinacci et al., 2021), and QUILT v0.1.9 (Davies et al.,
1347 2021). For all analyses, OncoPanel data was aligned to hg19 using bwa and processed with the GATK
1348 IndelRealigner. The 1000 Genomes Phase 3 release was used as a haplotype reference, targeting variants
1349 with > 1% frequency in the European population. Tumor imputation was performed using the 1000
1350 Genomes reference (rather than the HRC reference) because the HRC panel is not publicly available and
1351 the HRC imputation server does not support raw sequencing data. We thus sought to use the best reference
1352 panels that were accessible for the two data types. We note that HRC largely improves imputation accuracy
1353 for low-frequency variants (McCarthy et al., 2016), which were not the target of our analysis.

1354 Imputation with STITCH was carried out on all samples using aligned reads in 5-MB batches (see the
1355 “Availability of data and materials” section for the detailed parameters and code). The potential influence
1356 of target cohort size was evaluated by randomly downsampling to a lower number of sequenced tumors.
1357 Imputation with QUILT was carried out using the same input and batching procedure, with default
1358 parameters. Imputation with GLIMPSE was carried out on all samples with default parameters as
1359 recommended in the documentation: calling genotype likelihoods from each raw BAM file, splitting the
1360 genome into chunks, performing imputation and phasing, and ligating the chunks. An alternative, reference-
1361 only version of GLIMPSE was kindly provided to us by the authors but could not be compiled in our
1362 computing environment. Lastly, we considered two other imputation approaches: GeneImp (Spiliopoulou
1363 et al., 2017) and BEAGLE (Browning et al., 2021), but found that their computational requirements were
1364 infeasible for sample sizes in the thousands. Identical reference panel data was used for all methods except
1365 small indels, structural variants, and multi-allelic polymorphisms were excluded from the STITCH and
1366 GLIMPSE analysis (which only allows biallelic single nucleotides). After imputation, variants were
1367 considered “filtered” if they had a minor allele frequency > 1% and an INFO score (imputation confidence
1368 score) > 0.4 (similar to parameters used previously (S. Liu et al., 2018)).

1369 *Ancestry assignment*

1370 Samples were projected into genetic ancestry principal components using the weights previously derived
1371 by the SNPWEIGHTS software (Chen et al., 2013) for the continental populations. Weights were
1372 constructed from the 1000 Genomes reference groups with ancestry from Northern/Western Europe (CEU),
1373 Western Africa (YRI), and China (CHB+CHD). In our data, each component was projected independently
1374 as a linear combination of the weights and individual sample dosages (using the plink2 “--score” command).
1375 Components were then linearly recalibrated by fitting to self-reported race as an outcome (note this linear
1376 recalibration is for interpretation purposes only and does not influence the significance of any downstream
1377 associations). To estimate ancestry fractions, we uniformly rescaled the African and Asian components to
1378 be between 0 and 1 and additionally uniformly scaled the ancestry of each individual to be between 0 and
1379 1.

1380 BioMe

1381 *Genotyping and quality control*

1382 BioMe participants have been genotyped using Illumina's Global Screening Array (GSA-24 v1). Samples
1383 flagged as being contaminated, possibly duplicated, having low coverage, a call rate < 95%, or showing
1384 genotype-exome discordance were removed. Sex discordant samples were either reconciled after a plate
1385 swap resolution or removed. Sample missingness and depth of coverage were calculated using vcfTools:
1386 mean missingness was 1.24×10^{-3} , mean depth of coverage for all samples was 36.4x. Variant missingness
1387 and depth of coverage were calculated using vcfTools (Danecek et al., 2011): mean missingness rate of 1.24
1388 $\times 10^{-3}$, mean depth of coverage for all coding sites was 36.4x. Sites with HWE P-values < $1e-6$ were retained
1389 but flagged.

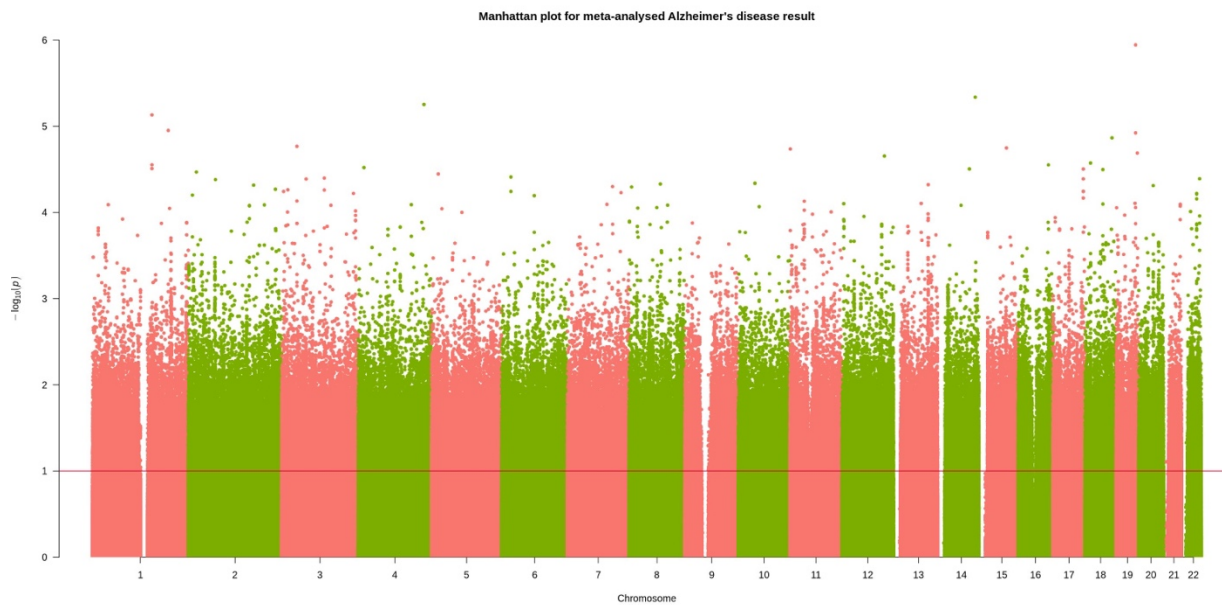
1390 *Imputation*

1391 Imputation was performed using the 1000G (1000 Genomes Project Consortium, 2015) and TOPMed
1392 (Taliun et al., 2021) reference panel, and the software packages Beagle (Browning et al., 2021) and Impute2
1393 (Howie et al., 2009). A filter of $r^2 > 0.7$ was applied. Approximately 31,700 samples and 7,8M variants
1394 passed QC and were used in downstream analyses.

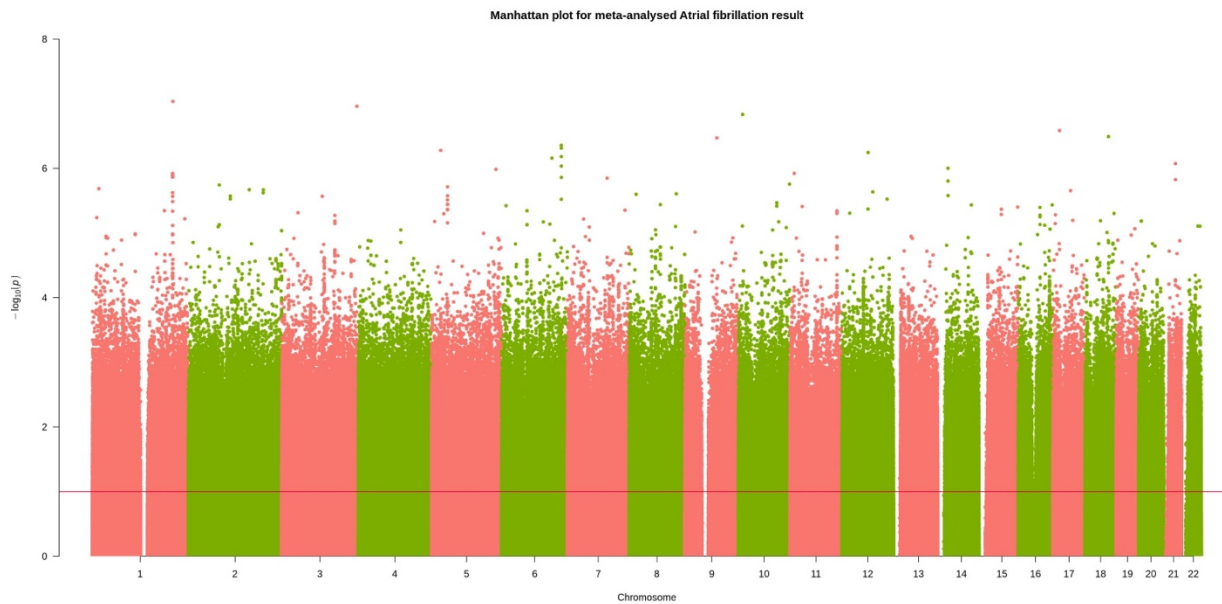
1395 *Ancestry assignment*

1396 We inferred the genetic ancestry following the guidelines of the Pan UKBB (*Quality Control (QC) | Pan*
1397 *UKBB*, n.d.). We performed a PCA using PLINK (Purcell et al., 2007), excluding relatives above 2nd-
1398 degree (kinship method, estimated using KING (Manichaikul et al., 2010)) and variants with MAF < 0.05.
1399 We trained a random forest classifier to infer the cohort's genetic ancestry using the 1000G labels as
1400 reference, removed outliers (by only including the quantiles 0.25-0.90) and participants with mixed ancestry
1401 (random forest probability ≤ 0.5). Inferred ancestry: AMR (n=5,336), AFR (n=5,660), EUR (n=7,447), SAS
1402 (n=613), and EAS (n=728).

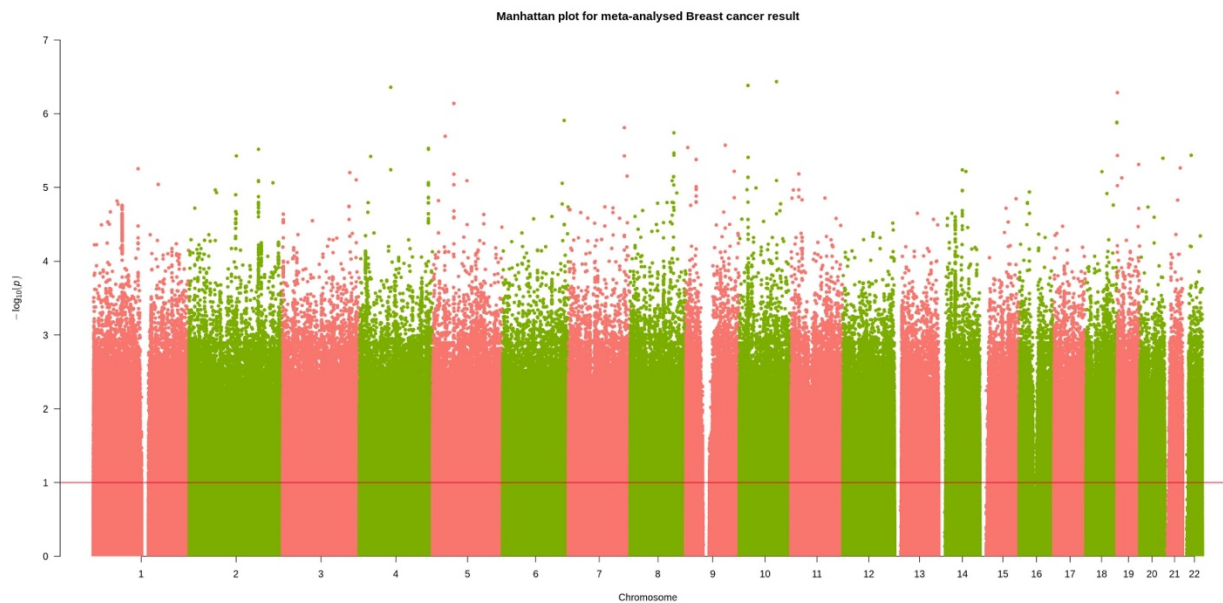
1403 **Supplementary figures**



1404 **Figure S1.** Manhattan plot for meta-analysed Alzheimer's disease mortality GWAS.
1405

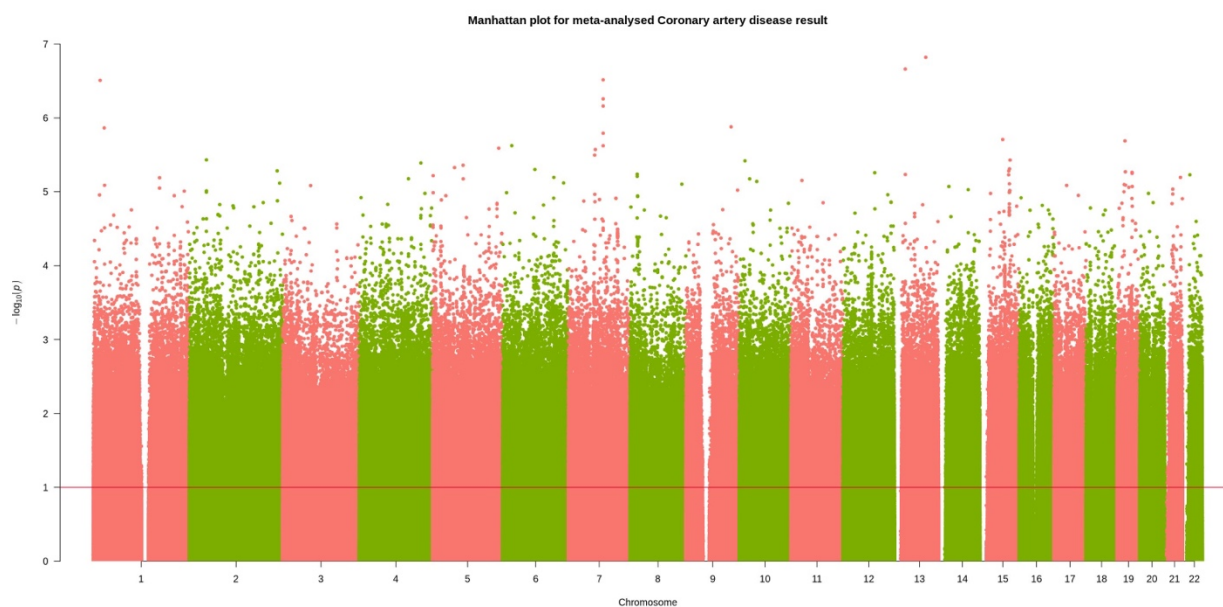


1406 **Figure S2.** Manhattan plot for meta-analysed atrial fibrillation mortality GWAS.
1407



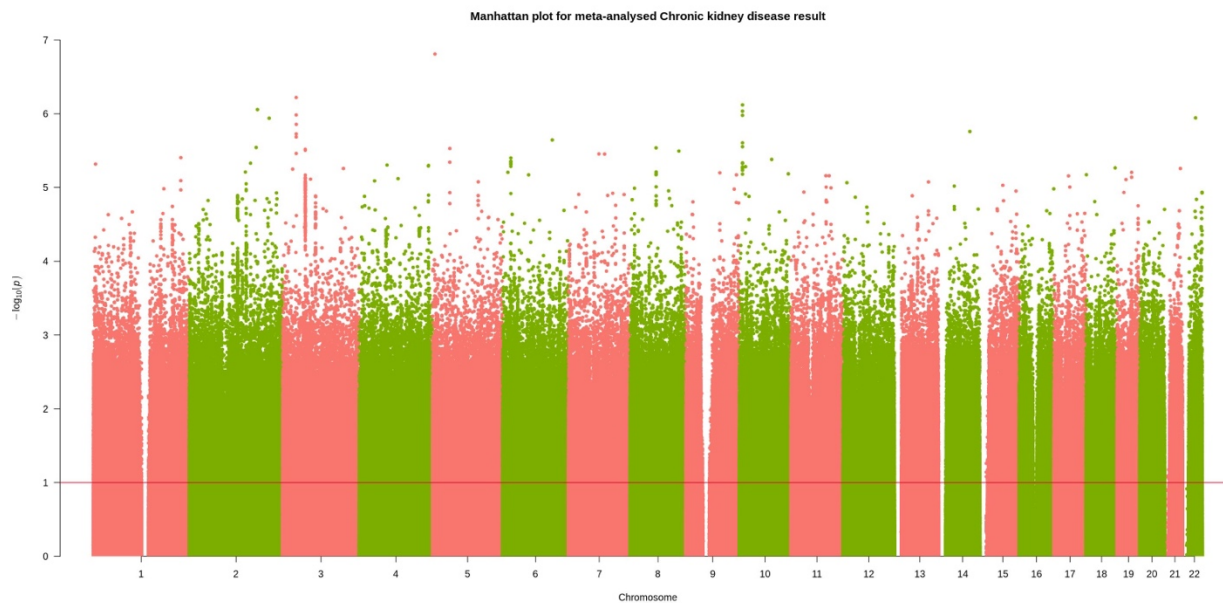
1408
1409

Figure S3. Manhattan plot for meta-analysed breast cancer mortality GWAS.



1410
1411

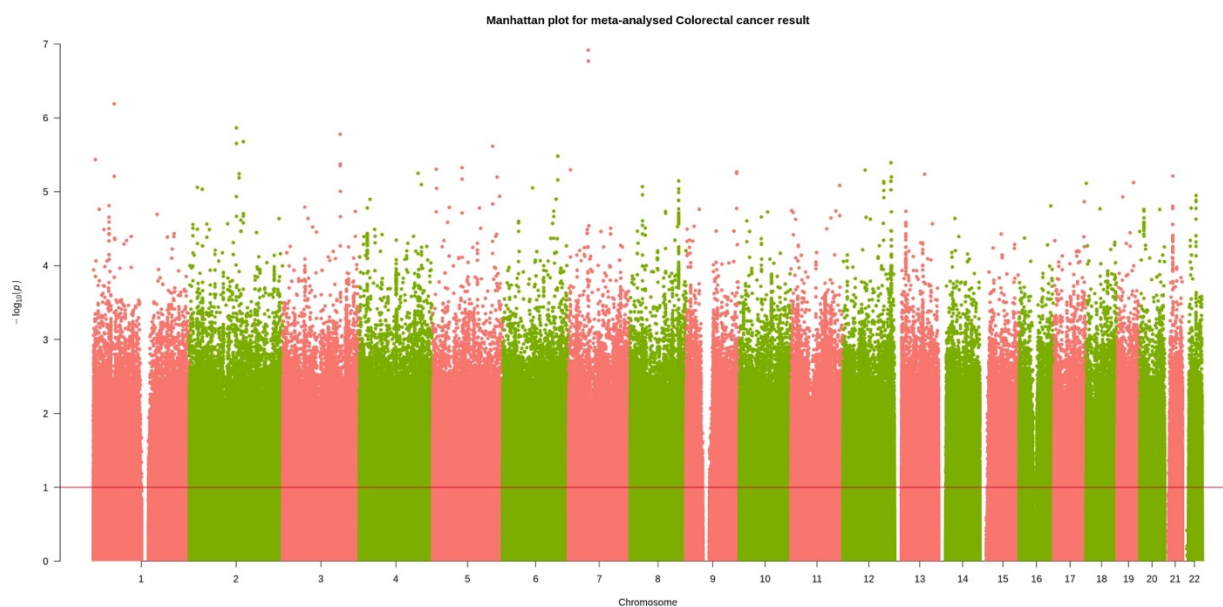
Figure S4. Manhattan plot for meta-analysed coronary artery disease mortality GWAS.



1412

1413

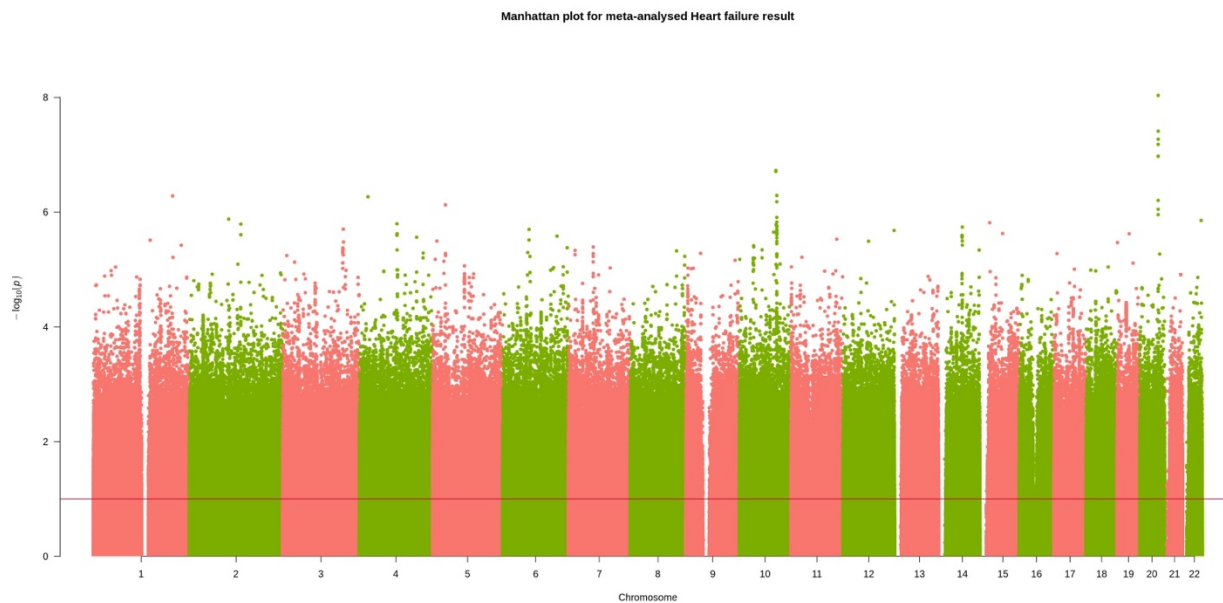
Figure S5. Manhattan plot for meta-analysed chronic kidney disease mortality GWAS.



1414

1415

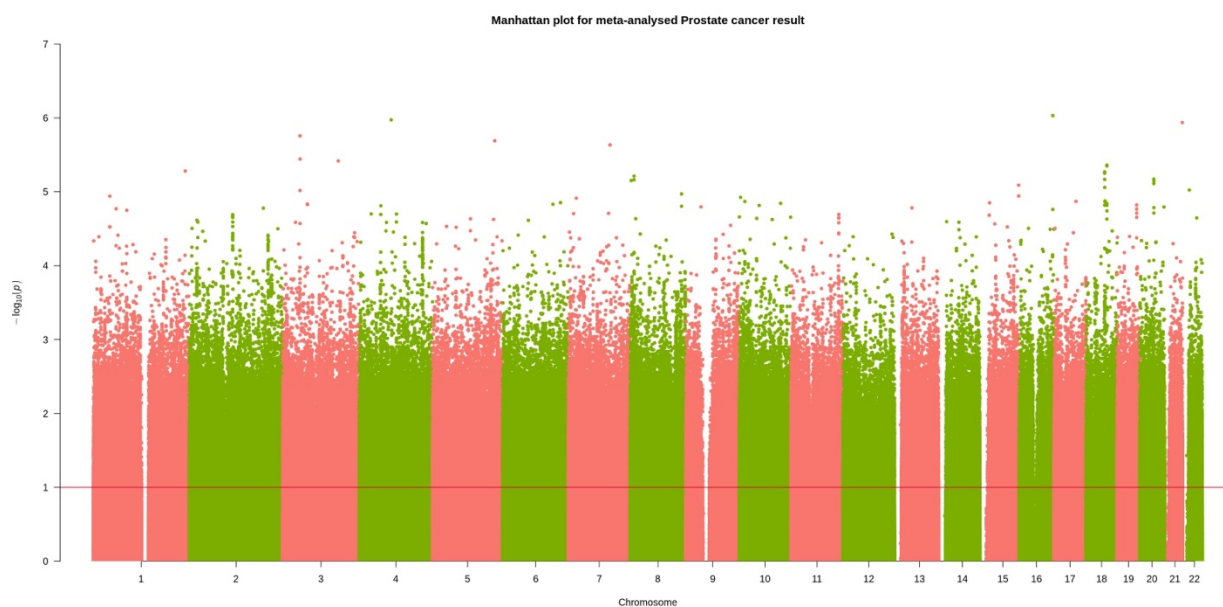
Figure S6. Manhattan plot for meta-analysed colorectal cancer mortality GWAS.



1416

1417

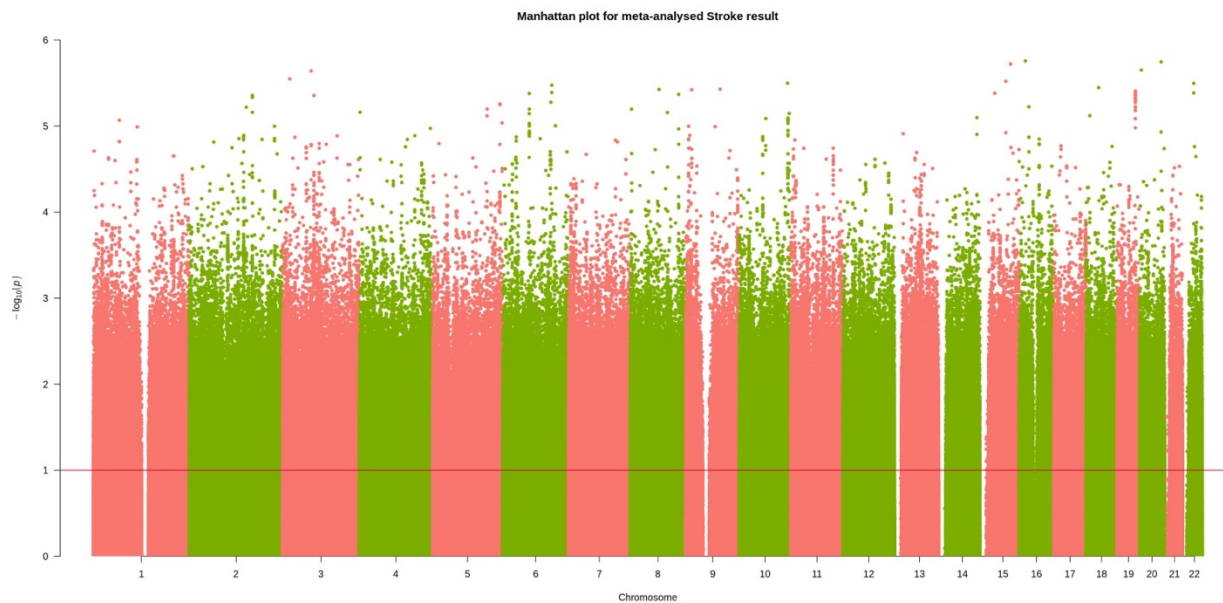
Figure S7. Manhattan plot for meta-analysed heart failure mortality GWAS.



1418

1419

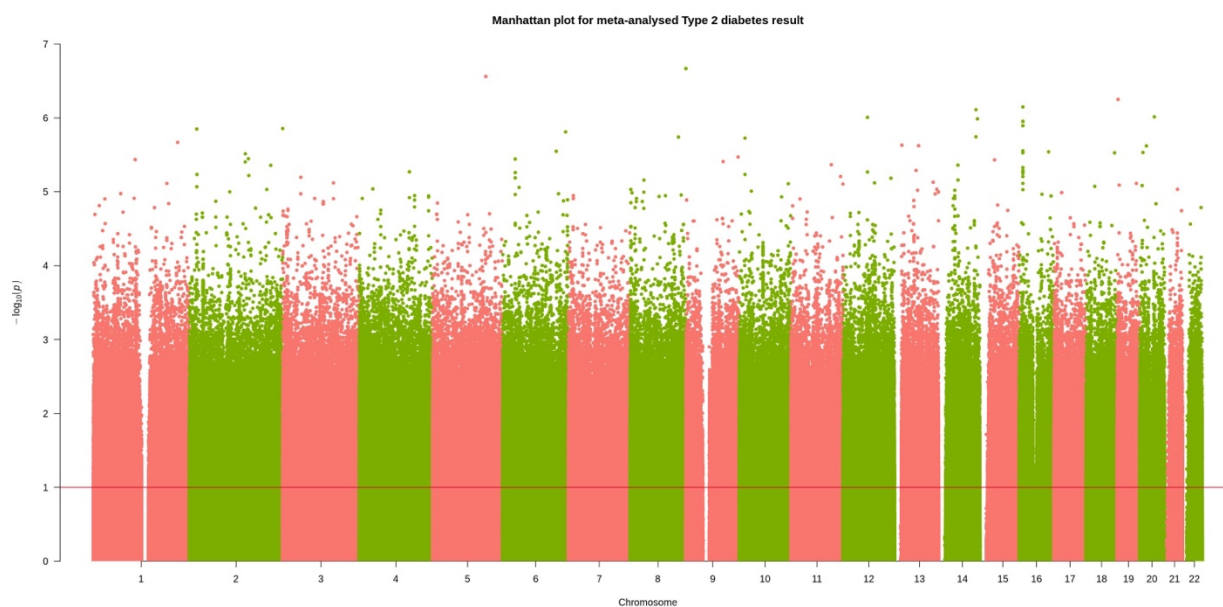
Figure S8. Manhattan plot for meta-analysed prostate cancer mortality GWAS.



1420

1421

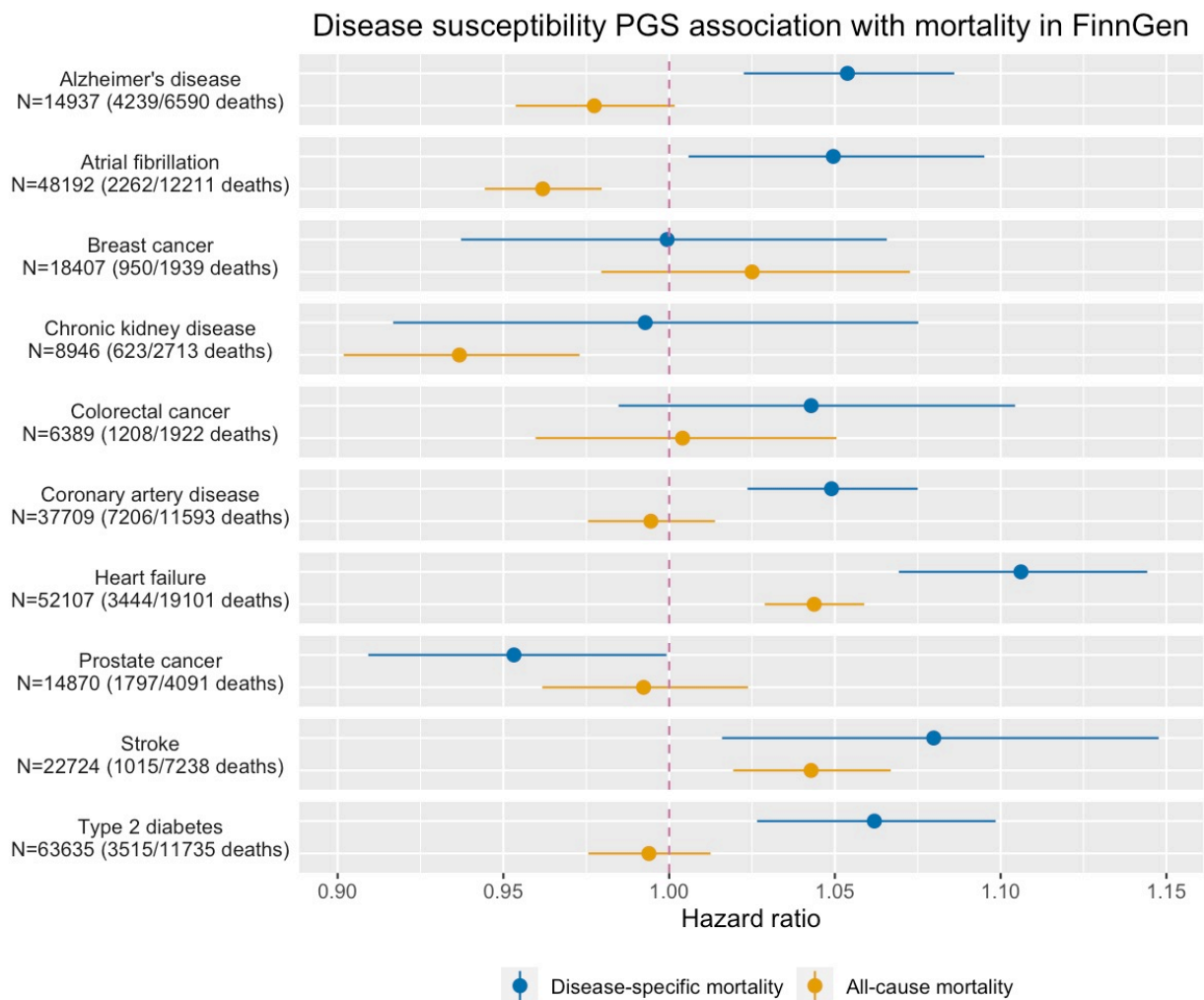
Figure S9. Manhattan plot for meta-analysed stroke mortality GWAS.



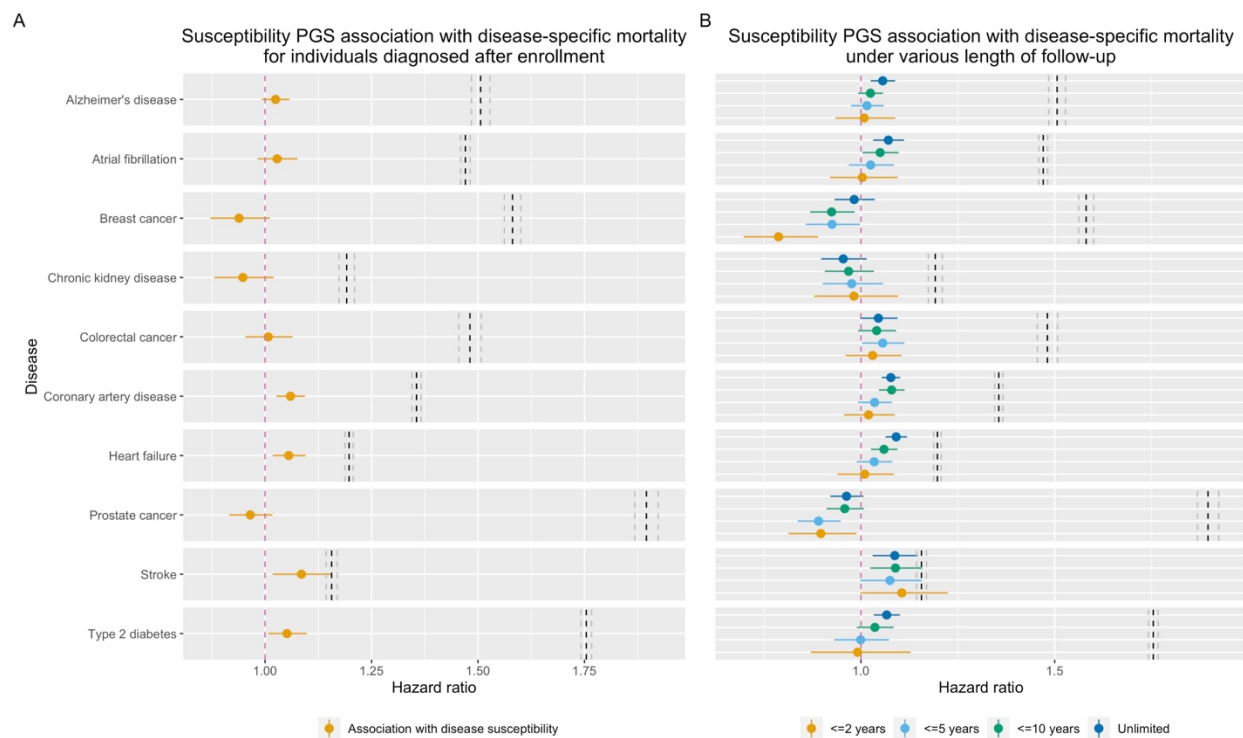
1422

1423

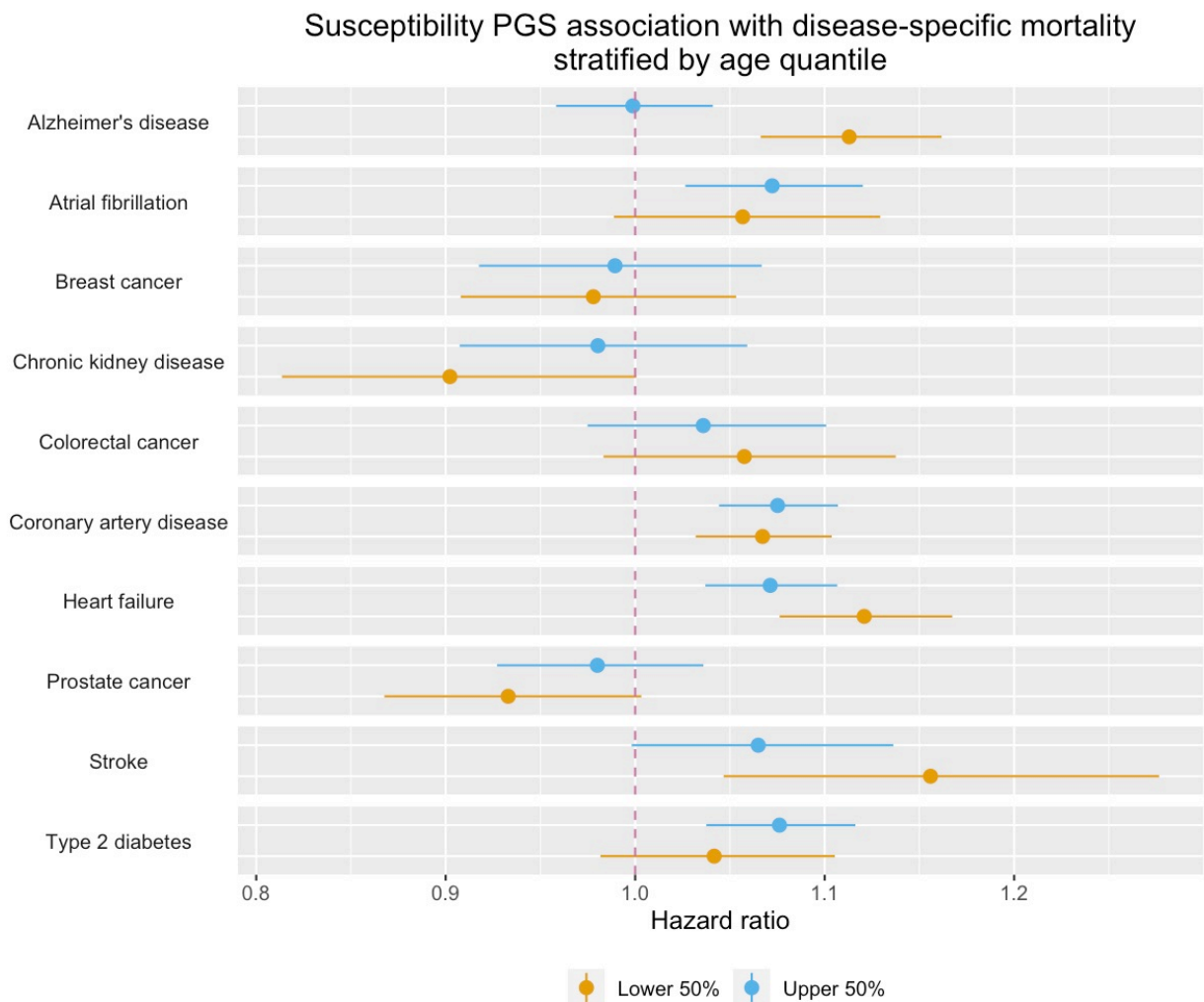
Figure S10. Manhattan plot for meta-analysed type ii diabetes mortality GWAS.



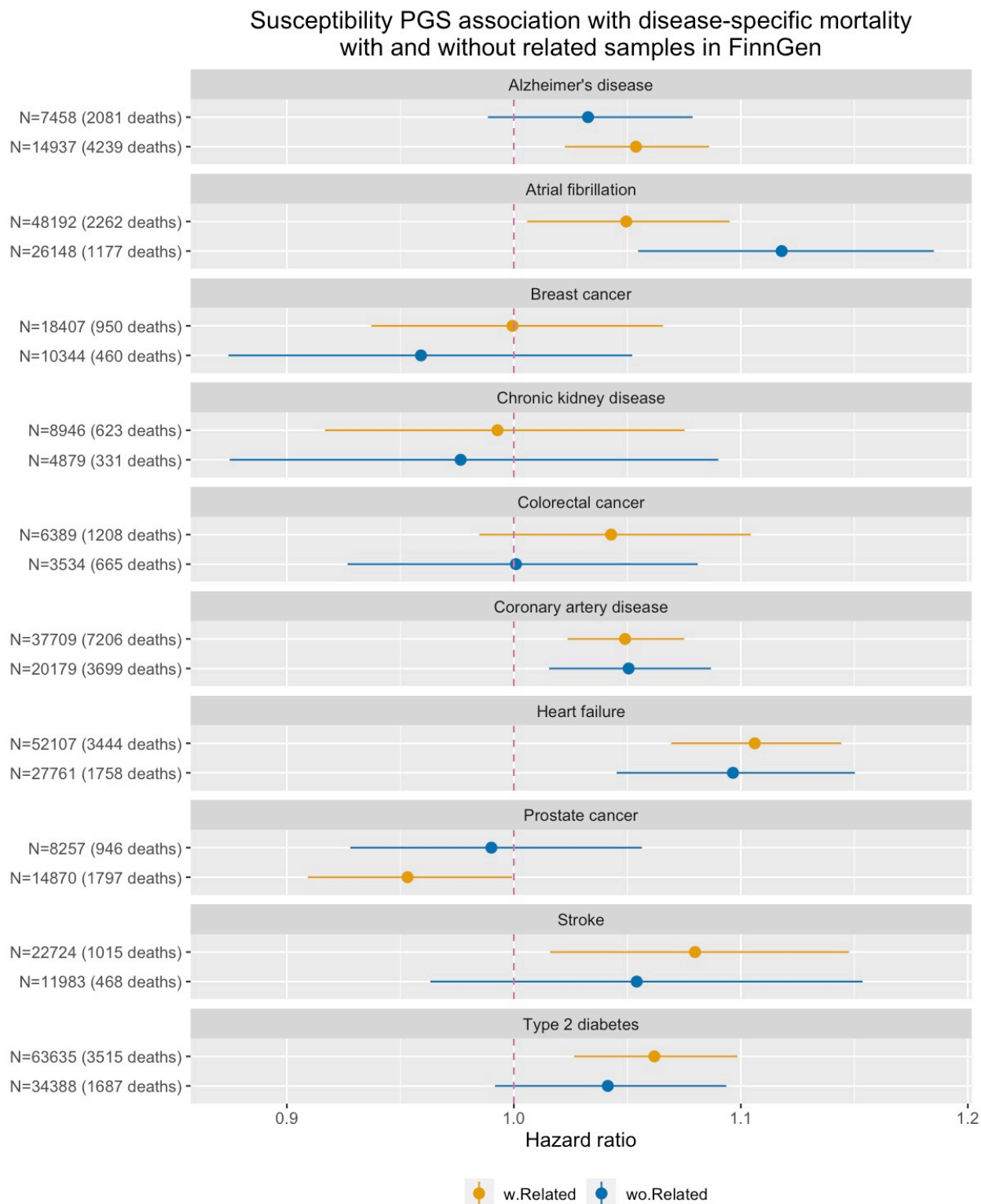
1424
 1425 **Figure S11.** Comparison of disease susceptibility PGS association with disease-specific mortality and all-
 1426 cause mortality in FinnGen. In parenthesis stated number of disease-specific mortalities/number of all-cause
 1427 mortality within the total number of patients (N). Horizontal solid lines represent 95% CI. Also see Table
 1428 S15 for quantitative results.



1429
1430 **Figure S12.** Sensitivity analyses. Left: Association between disease susceptibility PGS with disease
1431 specific mortality among only patients diagnosed after enrollment; Right: association between disease
1432 susceptibility PGS with disease specific mortality among patients with various lengths of follow-up after
1433 diagnosis. Horizontal solid lines represent 95% CI. Also see Table S16 for quantitative results.

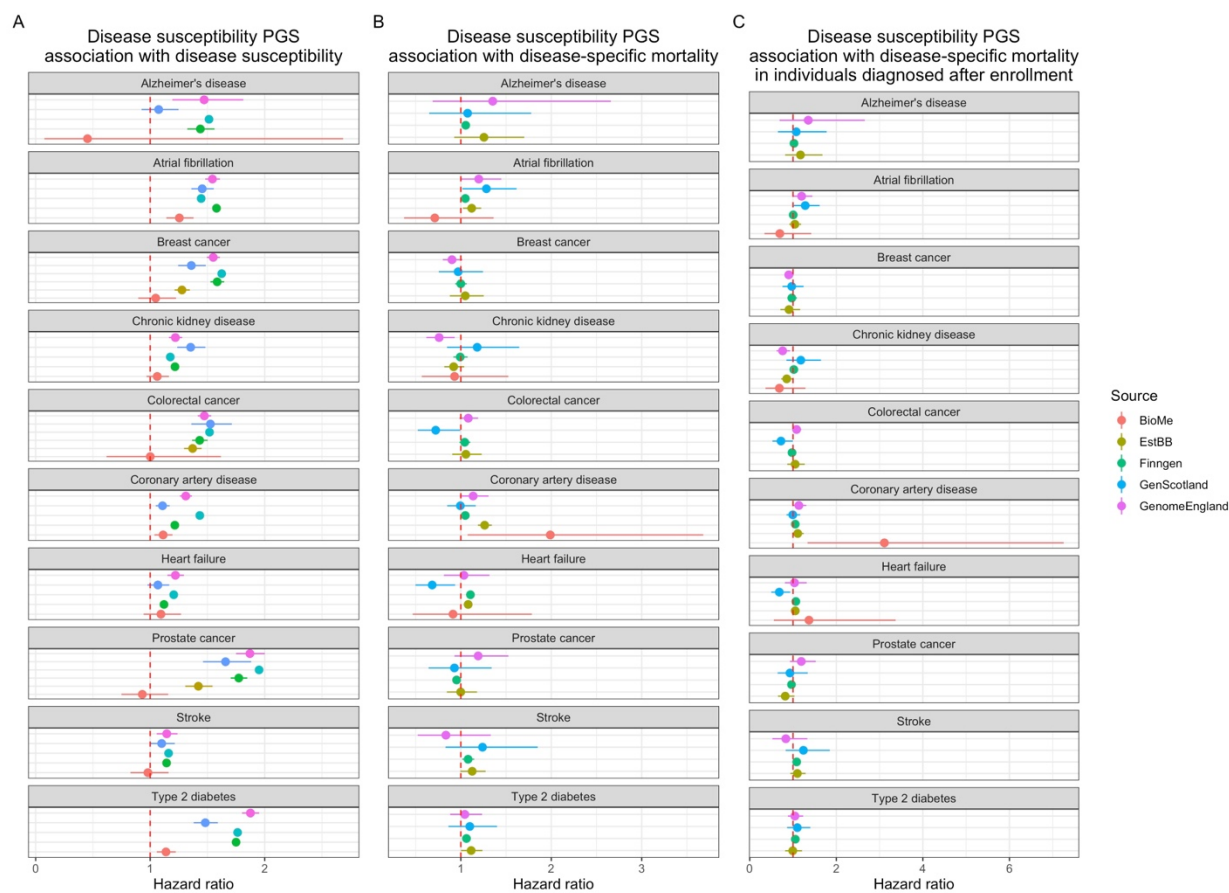


1434
1435 **Figure S13.** Sensitivity analyses. Susceptibility PGS association with disease-specific mortality stratified
1436 by age quantile. Horizontal solid lines represent 95% CI. Also see Table S17 for quantitative results.

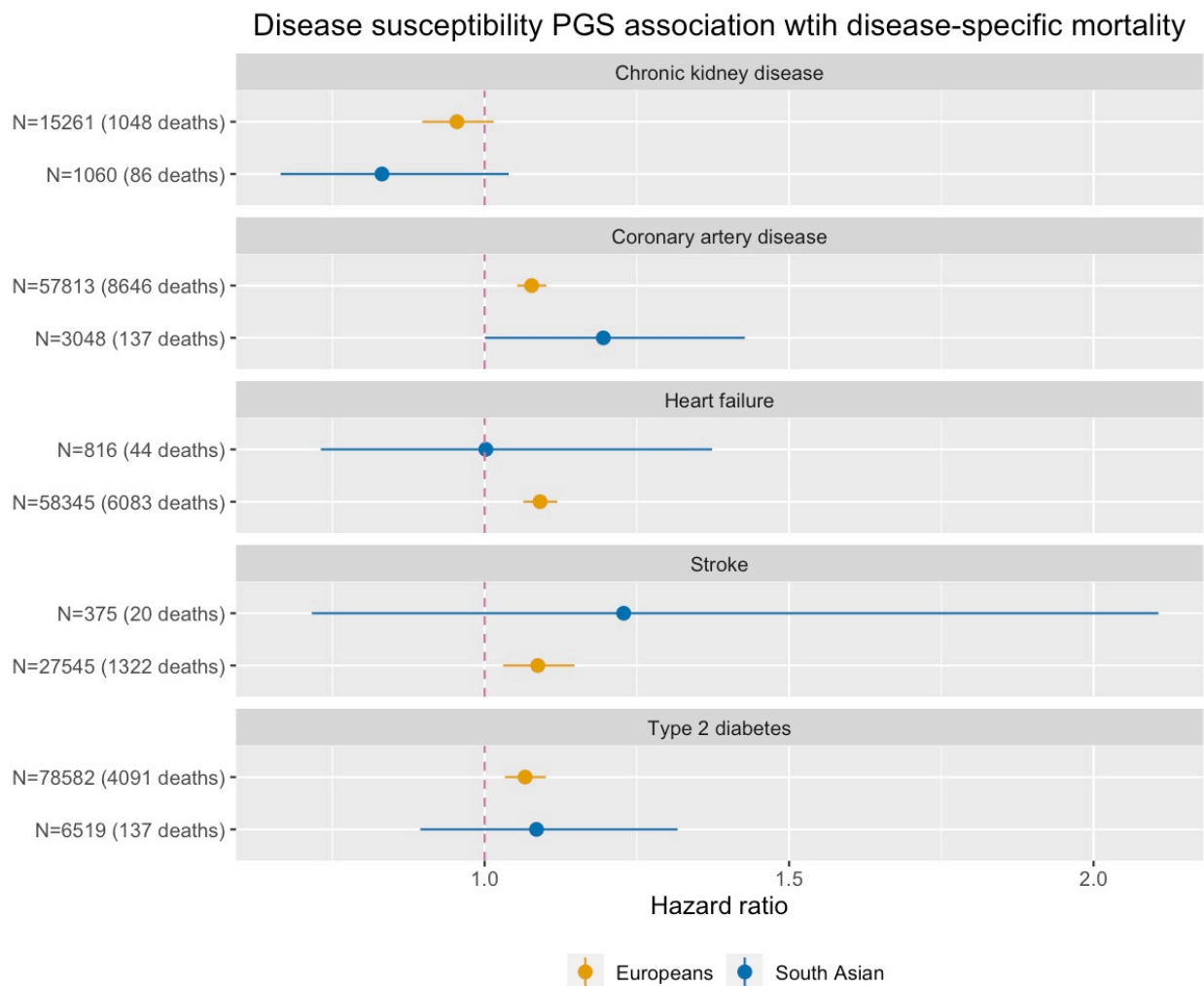


1437
1438
1439
1440
1441

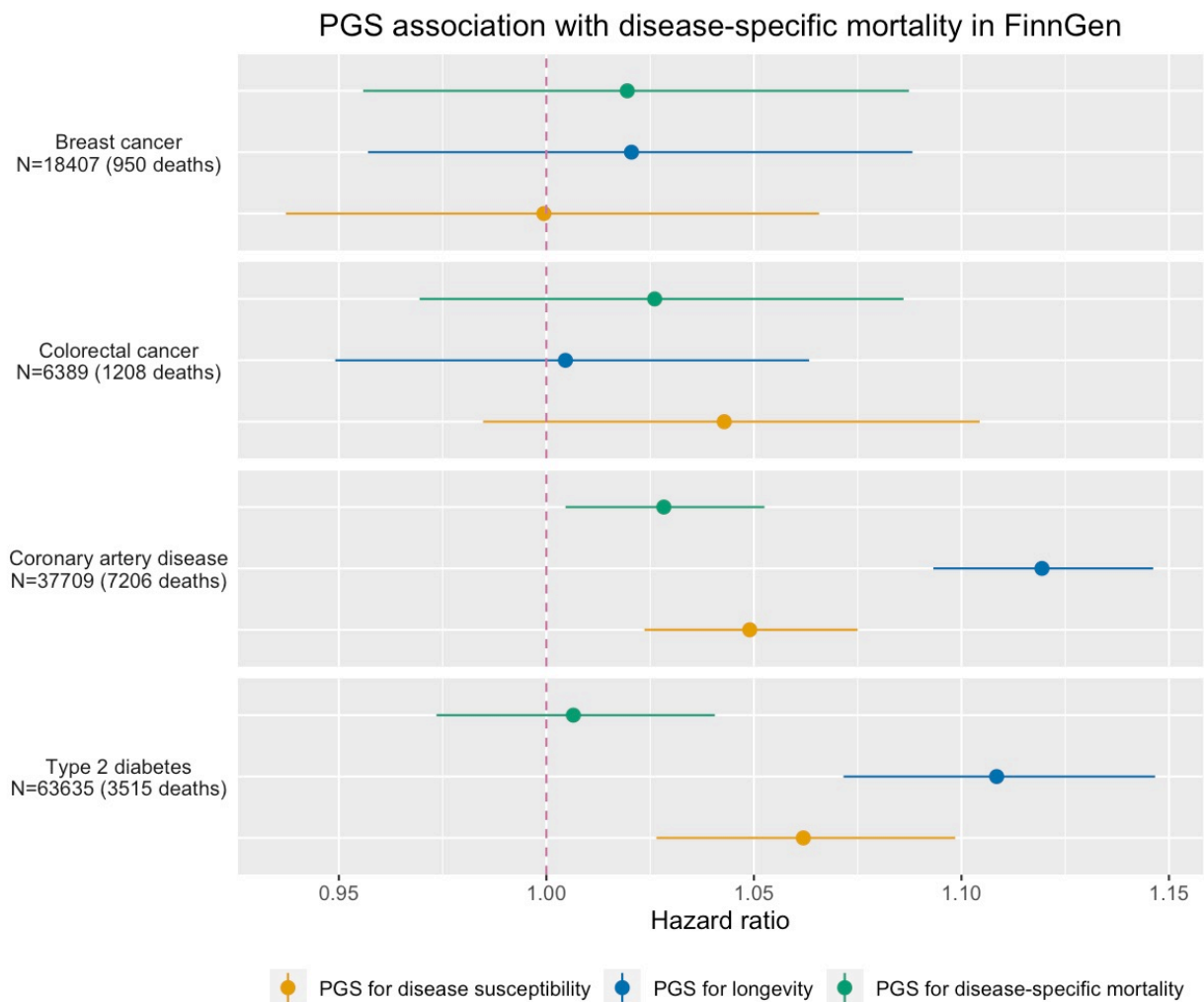
Figure S14. Sensitivity analyses. Susceptibility PGS association with disease-specific mortality in FinnGen with and without related individuals. For the without relatedness group (wo.Related), we removed up until second degree relatedness in the analyses. Horizontal solid lines represent 95% CI. Also see Table S15 for quantitative results.



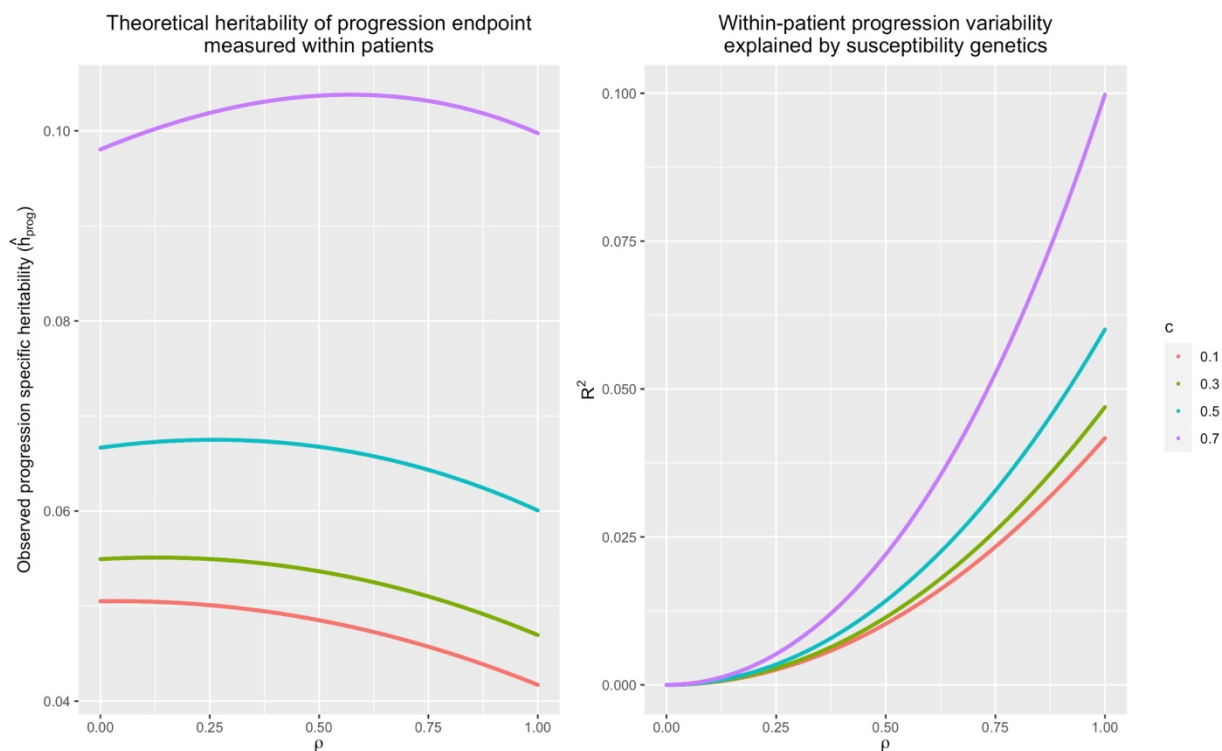
1442
 1443 **Figure S15.** Forest plot for effect sizes from each participant biobank. Horizontal solid lines represent 95%
 1444 CI.



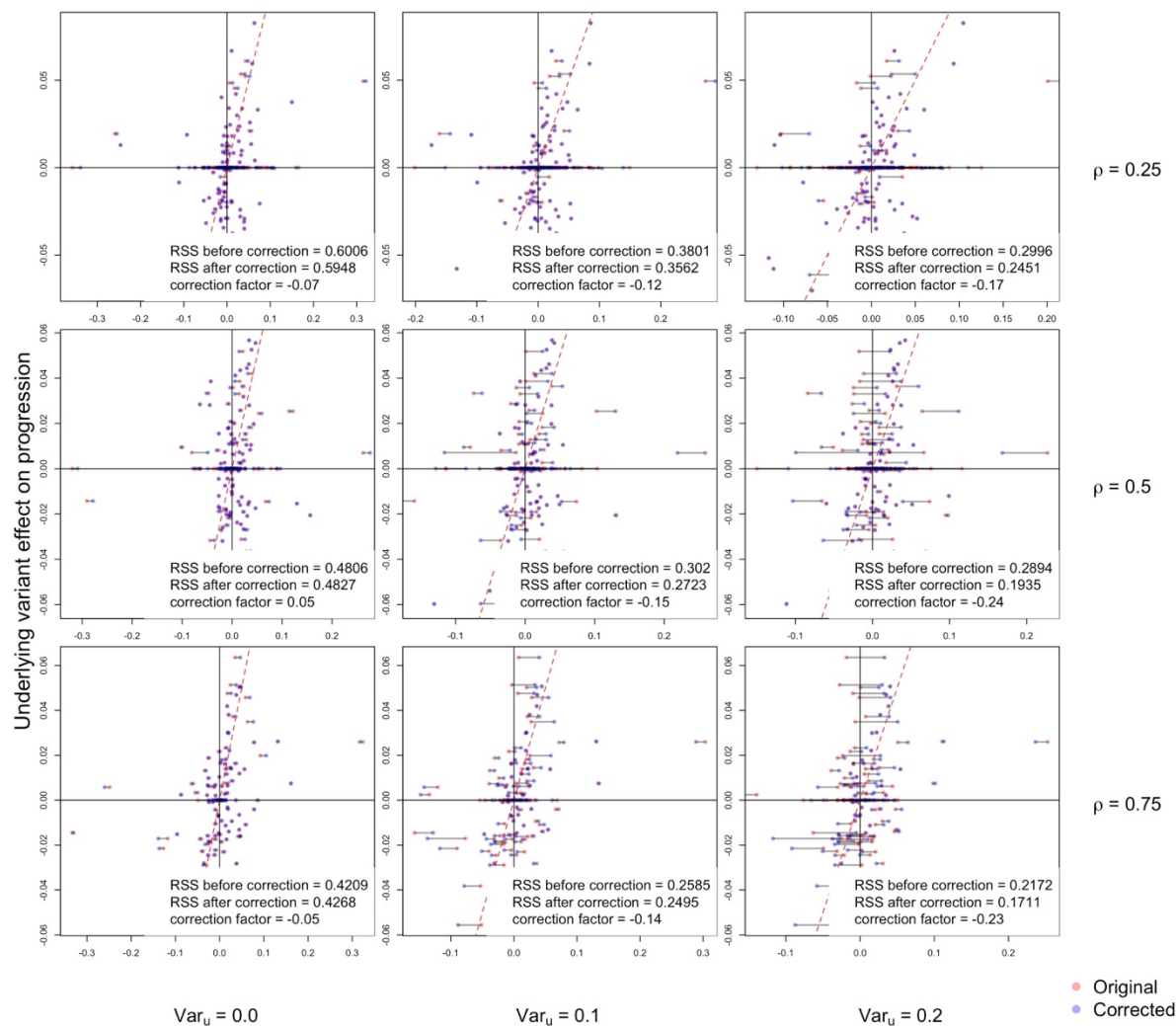
1445
1446 **Figure S16.** Disease susceptibility PGS association with disease-specific mortality in non-European
1447 population. As only patient cohorts are of interest in this study, for the non-European population, the only
1448 relatively powered results we had were associations for South Asians from biobank Genes & Health in a
1449 subset of diseases. Horizontal solid lines represent 95% CI.



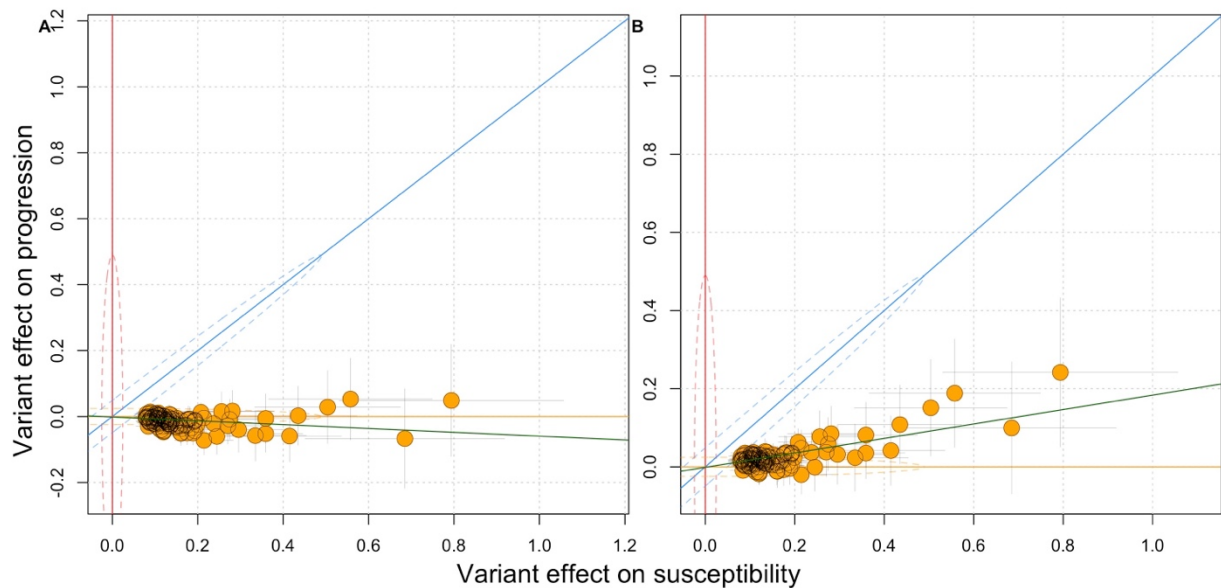
1450
 1451 **Figure S17.** Association between PGS and disease-specific mortality in FinnGen for eligible diseases. We
 1452 constructed disease mortality PGS using meta-analysed mortality GWAS results with FinnGen left out and
 1453 evaluated its association with disease specific mortality in FinnGen, comparing with disease diagnosis PGS
 1454 and longevity PGS. Horizontal solid lines represent 95% CI. Also see Table S18 for quantitative results.



1455
1456 **Figure S18.** Theoretical derivation of expected genetic measurements. For both derivations, heritability for
1457 the genetic component of disease susceptibility (h_{sus}) has been set to fixed value 0.3 and *unique genetic*
1458 *components* of progression endpoint in population (h_{prog}) has been set to fixed value 0.1. Under varying
1459 contribution of disease susceptibility on progression liability (c) and correlation of the susceptibility and
1460 progression specific genetic component (ρ), we derive A. Theoretical within-patient heritability of disease
1461 progression, corresponding to the expected heritability can be observed from a within-patient progression
1462 GWAS; B. Theoretical patient progression variance explained by susceptibility genetics, corresponding to
1463 the expected R^2 can be observed from disease susceptibility association with patients' progression.



1464
 1465 **Figure S19.** Impact of index event bias and slope-hunter-like correction under various conditions. In this
 1466 experiment, we fixed heritability of disease susceptibility ($h_{sus} = 0.2$) and progression ($h_{out} = 0.005$). Impact
 1467 of susceptibility liability on disease progression liability was also fixed at $c = 0.3$. Each panel corresponds
 1468 to a scenario under certain amount of shared causal variants (ρ) and amount of shared non-genetic factor
 1469 between the two endpoints (Var_u). Plot shows alignment of GWAS observed variant effects (x-axis) with
 1470 underlying causal effects (y-axis) on disease progression for all causal SNPs before and after slope-hunter-
 1471 like correction on shared and susceptibility specific causal variants (note susceptibility specific causal
 1472 variants are on $y = 0$ axis since their underlying effects on progression are 0). Also see Table S19 for
 1473 quantitative results.



1474
1475 **Figure S20.** Impact of index event bias and slope-hunter-like correction on GWAS observation and SNP
1476 classification. From all configurations demonstrated in Figure S7, we chose one of the settings where we
1477 see most severe impact of index event bias ($Var_u = 0.2$, $\rho = 0.5$) and compared the linemodel (Pirinen, 2023)
1478 classification on GWAS results before (left) and after (right) correction. We plot variant effects before and
1479 after correction on disease progression (y-axis) against their effect on susceptibility (x-axis). The regression
1480 line (green line) shifted after correction, whereas there was no change in variant classification. Also see
1481 Table S19 for quantitative results.