Fine-scale population structure confounds genetic risk scores in the ascertainment population

Holly Trochet[1] and Julie Hussin[1]

[1]Institut de Cardiologie de Montréal (Centre de Recherche), Faculté de Médecine, Université de Montréal, Montréal, Québec, Canada

Correspondence can be sent to holly.trochet@umontreal.ca.

August 9, 2020

Abstract

Genetic risk scores (GRS), also known as polygenic risk scores, are a tool to estimate individuals' liabilities to a disease or trait measurement based solely on genetic information. They have value in clinical applications [1] as well as for assessing relationships between traits and discovering causal determinants of complex disease [2, 3]. However, it has been shown that these scores are not robust to differences across continental populations [4, 5] and may not be portable within them either [6]. Even within a single population, they may have variable predictive ability across sexes and socioeconomic strata [7], raising questions about their potential biases. In this paper, we investigated the accuracy of two different GRS across population strata of the UK Biobank [8], separated along principal component (PC) axes, considering different approaches to account for social and environmental confounders. We found that these scores did not predict the real differences in phenotypes observed along the first principal component, with evidence of discrepancies on axes as high as PC45. These results demonstrate that the measures currently taken for correcting for population structure are not sufficient, and the need for social and environmental confounders to be factored into the creation of GRS.

Main

There have been a number of genetic scores created for traits ranging from risk of coronary artery disease [9, 10] to educational attainment [11]. These scores are used more and more as tools in research studies to help uncover links between traits and mechanisms of disease susceptibility. For instance, they have been used as the genetic instruments in Mendelian randomization studies to establish the causal relationship between an exposure and an outcome. They also have a potential clinical application—namely the stratification of individuals according to their risk of disease as predicted by their genetics, allowing for those at high risk to be monitored more closely or to be given medical interventions before the onset of the disease [1].

Population structure has been a concern in medical, statistical, and population genetics for years, as it may lead to spurious results in association studies, and GRS inherit this problem. It was shown that scores developed in UK Biobank (UKB) were confounded by population structure when applied in the Finnish population [6], but to our knowledge, the extent to which population structure in the ascertainment population affects the predictions remains unexplored, including in the papers that introduce them.

To investigate this question explicitly, we used two different GRS, one for coronary artery disease (CAD) called the metaGRS [10], and one for body mass index (BMI) [12], which we chose for several reasons. First, we wanted to investigate outcomes relating to a binary trait (CAD) and a quantitative one (BMI). Second, both scores were constructed using parameters tuned in subsets of the UKB, in which they were validated using the rest of the cohort. Third, the scores were generated in different ways, though both were in line with best practices at the time they were published: for CAD, markers contributing to the metaGRS were selected from a meta-analysis of several previously-published genetic risk scores, and their weights were estimated using UKB data. For BMI, the GRS was constructed from a previous meta-analysis of BMI genome-wide association studies and the algorithm LDpred [13], and validated using the UKB. The metaGRS was developed using all UKB participants, while the BMI GRS was created in the white British subset only, comprising 81.45% of the cohort. Finally, these scores are representative examples of GRS that assume a highly polygenic genetic architecture, with millions markers—the majority of which do not have validated associations with the trait in question—contributing to the calculation of the scores. The large number of markers potentially makes these GRS vulnerable to confounding due to population structure.

GRS—including the ones used in this study—are often assessed by dividing individuals according to quantiles of their GRS, with the lowest and highest quantiles being of particular interest. They are also assessed through regression (**Methods**, section M6): the trait is used as the outcome and the GRS is included as one of the predictor variables. Here, the values of interest are the significance of association between the GRS and the trait, and the regression coefficient of the GRS, which can be interpreted as the average per standard error effect on the measurement of a quantitative trait, or on the log odds ratio of having a binary trait. We can use this information to estimate the expected difference in trait mean or the odds ratio of its prevalence between two arbitrary groups of people (**Methods**, section M7).

All of our analyses are restricted to the white British subset (Figure 1a), a population which shows fine-scale structure [14]. To explore the effect of this structure on GRS predictions, we divided the cohort into groups based on where they fell along the genetic principal component (PC) axes calculated for the white British subsample (**Methods**, section M2). It has been shown that demographic processes relate directly to the PC projection, providing a way of summarizing the underlying genealogical history of the samples [15]. If the scores are confounded by population structure in the very cohort in which they were built, this will result in a mismatch between real and estimated differences in phenotype measurement or prevalence between groups. We calculated the mean GRS for each PC group and designated the one with the higher mean GRS as G_{high}, the high risk group, and the other group as G_{low}, the low risk group (Figure 1b). The distributions of the scores in both groups are similar, but shifted from one another, as shown Figure 2ac.

We then calculated the predicted differences in BMI mean and CAD prevalence between G_{high} and G_{low} by the GRS and compared them to the actual difference in prevalence observed in the cohort. For both BMI and CAD, the predicted score underestimates the true differences between G_{high} and G_{low} along PC1 (Figures 2b and 2d, respectively). The mean BMI of G_{high} is predicted to be 0.0321 $^{\mathrm{kg}}/_{m^2}$ higher than that of G_{low}, but in reality, we observed that it is 0.2859 $^{\mathrm{kg}}/_{m^2}$ higher. For CAD, the score predicts that G_{high} should have a prevalence of CAD that is 1.25% higher than that of G_{low}, but we observe that it is actually 7.81% higher.

To confirm that the discrepancy was driven by PC1, we estimated null distributions of the difference in mean BMI and the odds ratio of CAD prevalence (**Methods**, section M7). Briefly, we randomly sampled two groups, G'_{high} and G'_{low} so that for a given risk score, the distribution of the GRS matched those of G_{high} and G_{low}, respectively. For each risk score on each PC, we performed 1 million resamplings of G'_{high} and G'_{low}, each time recording the difference in BMI/the odds ratio of CAD prevalence between them. This generated an empirical distribution of BMI differences/odds ratios of CAD prevalences, given a risk score distribution of G_{low} and G_{high} to which we compare the true difference between G_{high} and G_{low} (Figures 2bd, S2 and S3). We found that while the mean differences of our null distributions coincided with scores' predictions, the observed difference deviated significantly ($p < 0.005$ in all cases) from the null (**Supplement**, section S3.1).

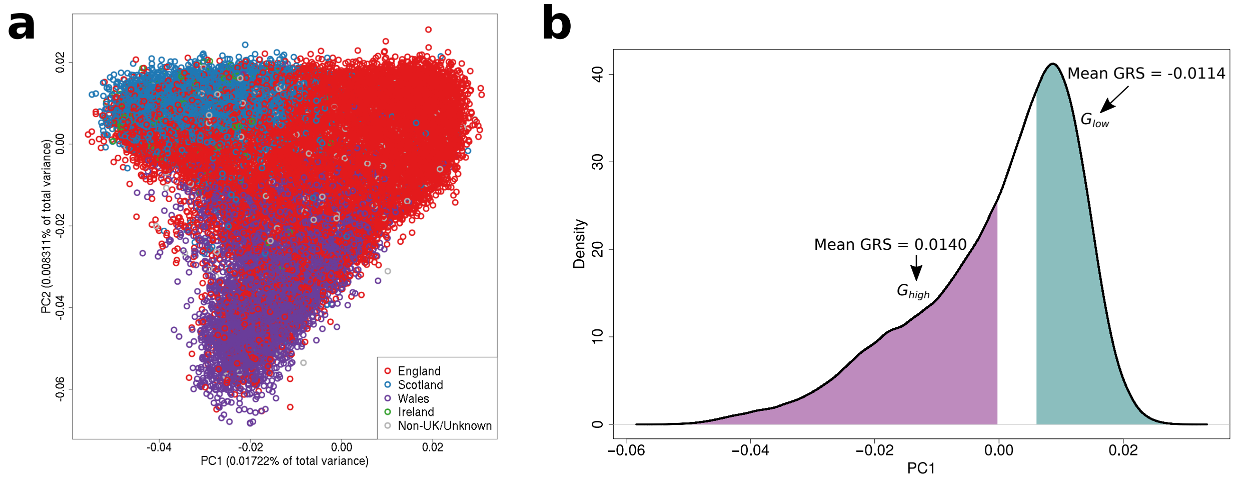

Figure 1: Stratifying the UKB white British subset according to population structure: a) First two principal components of the white British-only PCA. Each point is an individual, colored by his or her country of origin within in the UK or the Republic of Ireland. Since there were only 228 individuals born in the Republic of Ireland, we combined them with the 1,888 who were born in Northern Ireland, for purposes of this plot. **b)** Density plot of the distribution of PC1 measurements, with the lower and upper 40% highlighted in purple and teal, respectively. We also show the mean metaGRS score (for CAD) for each group. Because the lower 40% have a mean GRS higher than that of the upper 40%, they are predicted to be at a higher risk of CAD and thus are labeled G_{high}. Analogously, the upper 40% group is label G_{low}.

Both BMI and CAD risk are affected by environmental and lifestyle factors that could, in turn, vary along PCs. We adjusted for these potential confounders in different sub-analyses (**Methods**, section M9 and **Supplement** section S3 for details). Briefly, the first way was by matching individuals from G_{low} to individuals in G_{high} for age, sex, and smoking habits, as well as for lifestyle variables (sub-analysis M_1) and pollution variables (M_2) (Table S3). The lifestyle variables include Townsend deprivation index, alcohol consumption, and exercise habits. The pollution variables include covariates pertaining to nitrogen dioxide, nitrogen oxides, and particulate matter pollution. Individuals who could not be matched were excluded from the analysis. The second way of adjusting for environmental and lifestyle confounding was to create corrected, or modified PCs (mPCs), which we then used in place of the genetic PCs in our analyses. We used two approaches here. In the first, mPCs were the residuals of the regression of 22 environmental and lifestyle covariates on the original genetic PCs (R_1). In the second, we performed PC analysis on the same 22 covariates, and the projections of these covariates were then regressed onto the genetic PCs to generate the mPCs (R_2). We observed very high correlations between the mPCs and genetic PCs ($\rho > 0.97$ in all cases) (Table S6, **Supplement**, section S4.4 for details).

The results on PC1 hold for all adjustment techniques, for both CAD and BMI scores (Figure 3). While the results are attenuated by matching and by using mPCs, clear differences between the score predictions and true observations remained. The observed difference

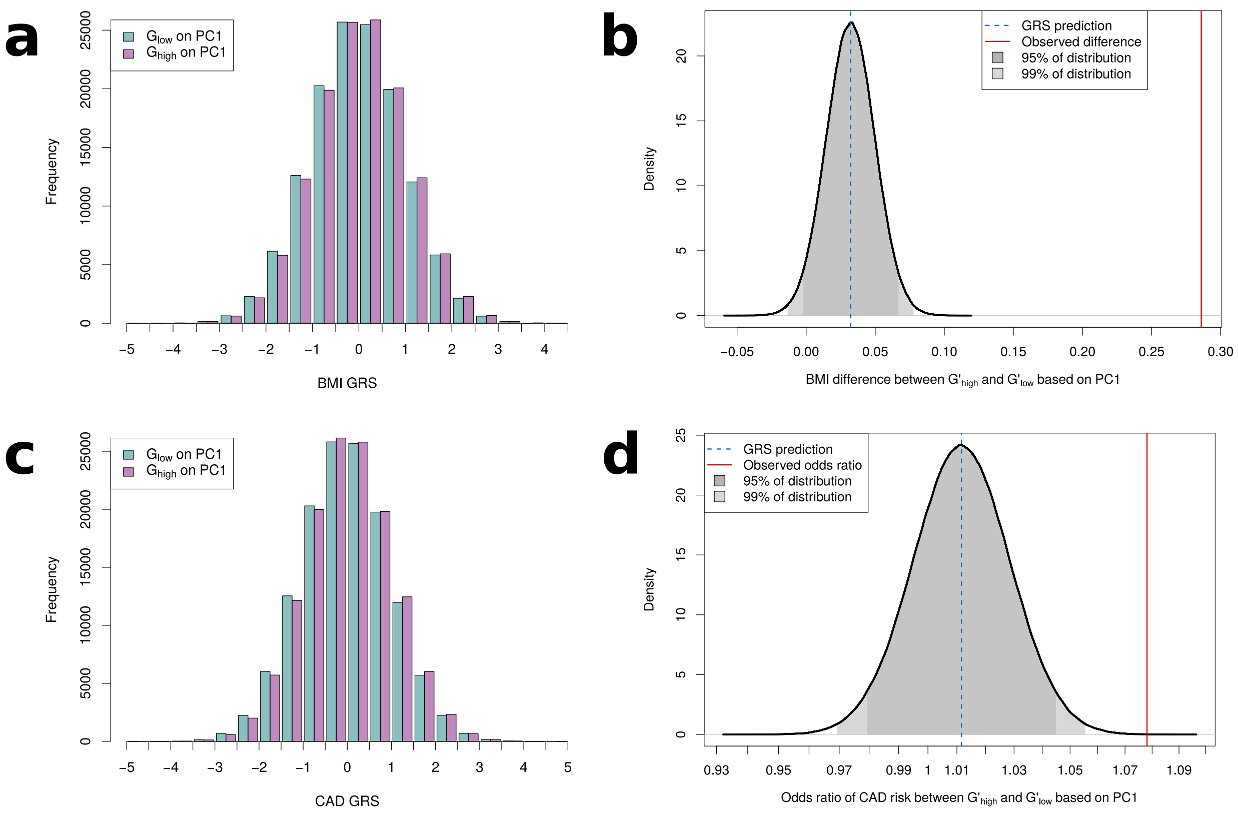

Figure 2: Distributions of risk scores in low- and high-risk groups and of the differences in phenotype/prevalence: (a) Histogram of the BMI risk score for G_{high} (light purple) and G_{low} (blue) defined for PC1, with the overlap shown in dark purple. **(b)** Density plot of the difference in mean BMI in groups that were resampled 1 million times so that their distributions matched that in **a**. Dark grey shows 95% of the distribution, with the light grey extension of this showing 99%. The vertical blue dotted line shows the difference in mean BMI predicted by the BMI GRS between G_{high} and G_{low} on PC1. The vertical red line shows the observed BMI difference between G_{high} and G_{low} on PC1. **(c)** Histogram of the CAD risk score for G_{high} and G_{low} defined for PC1. Coloring is the same as for **a**. **(d)** Density plot of the difference in CAD prevalence in groups that were resampled so that their distributions matched that in **c**. This plot is analogous to **b**, but for CAD prevalence. The same plots for PC2, PC3 and PC45 are shown in Figures S1 and S3.

in mean BMI (Figure 3a) differed from the predicted difference more significantly than CAD did (Figure 3b, Table S2), possibly due to the lower standard error for the quantitative trait (**Supplement**, section S5).

Discrepancies between observed and predicted differences, before adjusting for potential confounders, were also seen for population strata defined on PC2, PC3 and PC45, for both GRS (Figure S3). For CAD, the 95% confidence intervals for the observed prevalences after accounting for confounders often overlapped with the point estimate for the GRS predictions, suggesting that the prevalence predicted by the scores is a plausible value for the true prevalence (Figure 4b). This is especially true in the mPC analysis, which suggests the difference with the GRS predictions at baseline was due to several socio-economic factors.

The results for BMI on PC2 (Figure 4a) show stable estimations for predicted differences between mean BMI in G_{low} and G_{high} across all sub-analyses, but the observed results differed strongly between the mPC analyses and the non-mPC analyses. At baseline

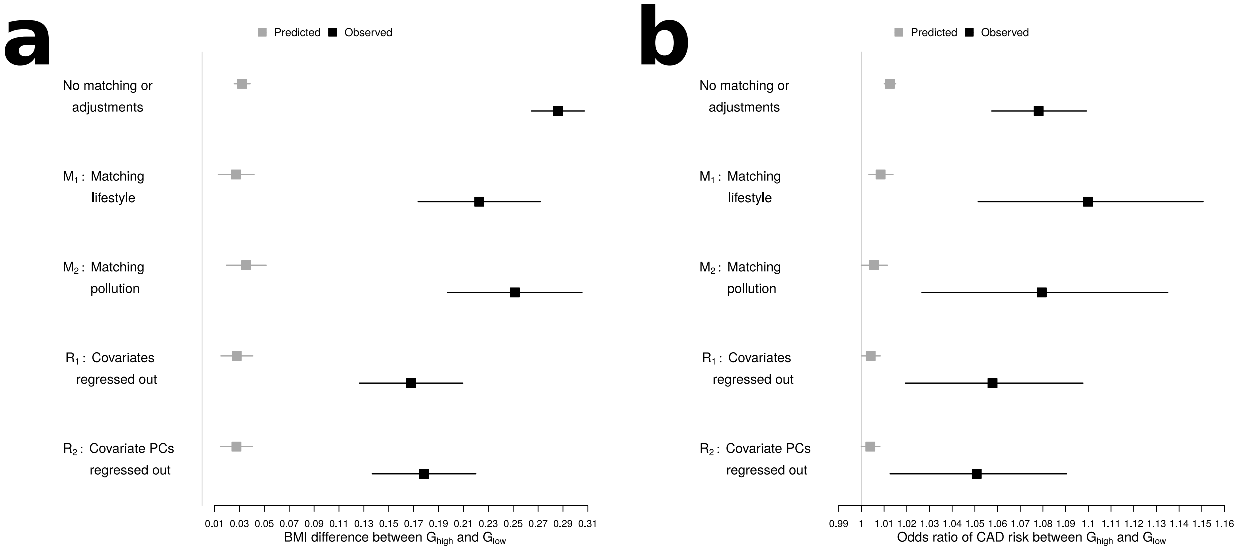

Figure 3: Differences between predicted and observed differences in phenotype/prevalence on PC1: Point estimates (boxes) and 95% confidence intervals (lines) of predicted (in grey) vs. observed (in black) differences in **(a)** mean BMI and **(b)** CAD prevalence between G_{high} and G_{low} on PC1. From top to bottom, the subsets of the UK Biobank used were all white British individuals, after matching for lifestyle variables (factors in M_1), after matching for pollution variables (factors in M_2), regressing out all lifestyle and pollution variables out of the genetic PCs (R_1), and after regressing the PCs of all the lifestyle and pollution variables out of the genetic PCs (R_2). Note that the predicted differences in prevalence/mean were recalculated for each analysis, using the individuals who were available for the observed analyses.

(no adjustment/no matching), the differences in BMI between the groups is small and not statistically different from zero (Figure S3a). In the matched subsamples (M_1, M_2), the observed differences were lower than what was predicted, but both contained the predicted difference in their 95% confidence intervals. When correcting for the 22 environmental and lifestyle covariates simultaneously (R_1, R_2), the result is the opposite: the GRS underestimates the observed differences between groups split along the mPCs. This suggests that the GRS fails to properly capture the reality of the phenotype heterogeneity when population structure and environmental variables co-occur.

Indeed, we find that there were statistically significant differences ($p < 0.05$, Figure S4) between G_{high} and G_{low} in age, Townsend deprivation index, nitrogen dioxide air pollution, and amount of exercise for PC2 and mPC2 for both mPC analyses (Table S4), suggesting that regressing out environmental factors does not succeed in completely removing their effects. These results suggest that the GRS falsely captures differences in susceptibility between groups separated on PC2, which appears to separate individuals born in Wales from those born in the rest of the British isles. In the UKB, individuals born in Wales had the lowest Townsend deprivation index (indicating less deprivation) among the white British born in the UK or elsewhere. This persistence of effects even after regression occurs on PC3 as well (Table S5), where there are statistically significant differences in age, Townsend deprivation index, smoking, and exercise across the genetic PC and the mPCs (**Supplement**, section S4.3).

For PC3, we observe another interesting phenomenon in the case of BMI: the differences

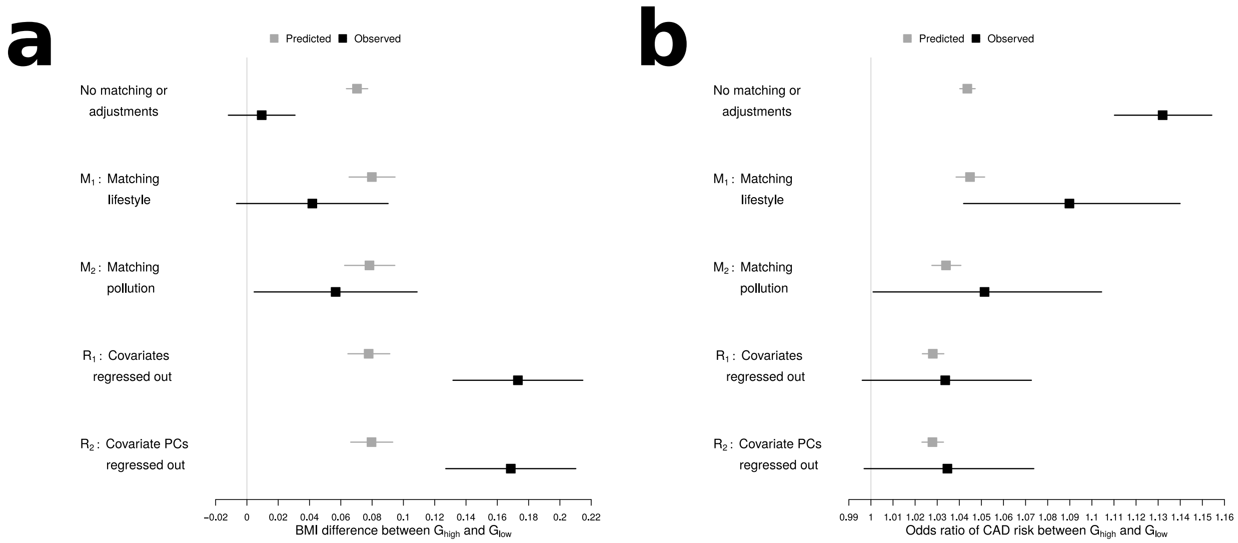

Figure 4: Differences between predicted and observed differences in phenotype/prevalence on PC2: Point estimates (boxes) and 95% confidence intervals (lines) of predicted (in grey) vs. observed (in black) of **(a)** the mean difference in BMI between G_{high} and G_{low} and **(b)** the odds ratio of CAD prevalence between $G_{h}igh$ and G_{low}. Groups are defined along the PC2 axis, and the analyses shown are at baseline (no adjustments), matching (M_1, M_2) and mPC (R_1, R_2).

predicted by the GRS between groups along PC3 are very small or non-significant (Figure S3cd), especially at baseline, but the G_{high} group has an observed mean BMI that is between 0.1 and 0.24 lower than that of the G_{low} group (depending on the sub-analysis, Figure 5a). This is also observed for CAD on PC45: G_{high} actually has a lower prevalence of CAD than G_{low} but the GRS predicts the opposite (Figure S3ef), although we note that the disease prevalence between G_{high} and G_{low} is not significantly different for M_2 and the mPC analyses (Figure 5d). Despite the fact that there is a high amount of uncertainty in our estimates for both observed CAD prevalence and mean BMI when stratifying along PC45, the fact that this PC is correlated with the traits and the risk scores at all, for both phenotypes, is remarkable, given how little genetic variance is explained by this axis (0.0036%). It also stands in contrast to the PCs provided by the UKB, which only go up to 40, and which summarize the genetic variation in the whole dataset, rather than the white British subset, and suggests that even the smallest PCs, representing very fine-scale population structure, may need to be taken into account in risk prediction.

Except for PC1, for which the GRS underestimates CAD prevalence differences between groups even after accounting for covariates, all of our adjusted analyses for CAD show odds ratios that are increasingly close to the ones predicted by the GRS, compared to baseline. This result illustrates the importance of adjusting for lifestyle and environmental factors when applying GRS, and demonstrates how these covariates can vary across PC axes independently of one's genetic risk, despite PCs being constructed entirely from genetic information. Though we have used only two GRS in our analysis, we do not believe the concerns are restricted to these specific scores, as they were generated using different methodologies and for different traits—one binary and one quantitative. We also doubt that this issue is restricted to construction of GRS in the UKB, but as this is one of the largest cohorts available right now to build these scores, an appropriate first step was to check how much population structure can affect risk prediction in that cohort. We also highlight that there may be other examples of confounded PCs that we missed due to the fact that we only investigated the top 50 PCs.

The use of two different methods to account for environmental differences is appropriate here, as benefits and drawbacks exist for both. In our matching strategy, we were not able to match on all the variables that might be relevant to the trait, much less to adjust for all of them simultaneously, as we could with the mPC analyses. However, as we saw in our results for PC2 and PC3, regressing out the relevant covariates does not always remove the differences between the groups. Additionally, we adjusted for the same set of covariates for both traits, even though risk factors like smoking do not necessarily have direct causal effects on BMI. This is not a problem in the matching analysis (except for unnecessarily restricting our sample size), but may introduce collider biases in the mPC analyses. There is also an interpretability problem inherent to the mPC analyses: regular PCs are pure summaries of the genetic data, but mPCs are not, and what they retain from the population stratification that truly exists in the cohort is an open question.

In this paper, we have shown that population structure can cause a GRS to over- or underestimate the phenotype differences between population strata. Because the scores take information on the genome-wide variability, phenotype prediction using GRS are intrinsically confounded by population structure in the ascertainment cohort, and we can hardly expect them to be robust to biases relating to population structure when applied to a new population. Previous research has shown that this kind of confounding can lead to overestimation of polygenic adaptation on height [16, 17]. Furthermore, the solution cannot be to create "ancestry-corrected polygenic scores" [12], which are the residuals from a regression of a certain number of genetic PCs on GRS, because this has the potential to remove real effects. Finally, we highlight that a score's association with population structure is not a problem in and of itself. There is no reason why a genetic locus that has a causal effect on a trait could not also have alleles whose frequencies vary across populations or subpopulations. The issue of concern to us is that population structure causes the score to predict greater or smaller differences across the population than actually exist. This can lead to problems such as inaccurate assessment of an individual's disease risk, or falsely attributing a genetic cause to a subpopulation's elevated rate of disease compared to another, when the true cause might be social, economic, or environmental.

Acknowledgments

We would like to thank Dr Na Cai, Dr Marie-Julie Favé, and Jean-Christophe Grenier for taking the time to read and comment on our manuscript. Additionally, we would like to thank Dr Ryan Christ for discussion and feedback on some of our statistical analyses. Julie Hussin is a FRQS Junior 1 Research Scholar also supported by the Institute for Data Valorization (IVADO). This project has been funded by the Molson Foundation and the Montreal Heart Institute Foundation.

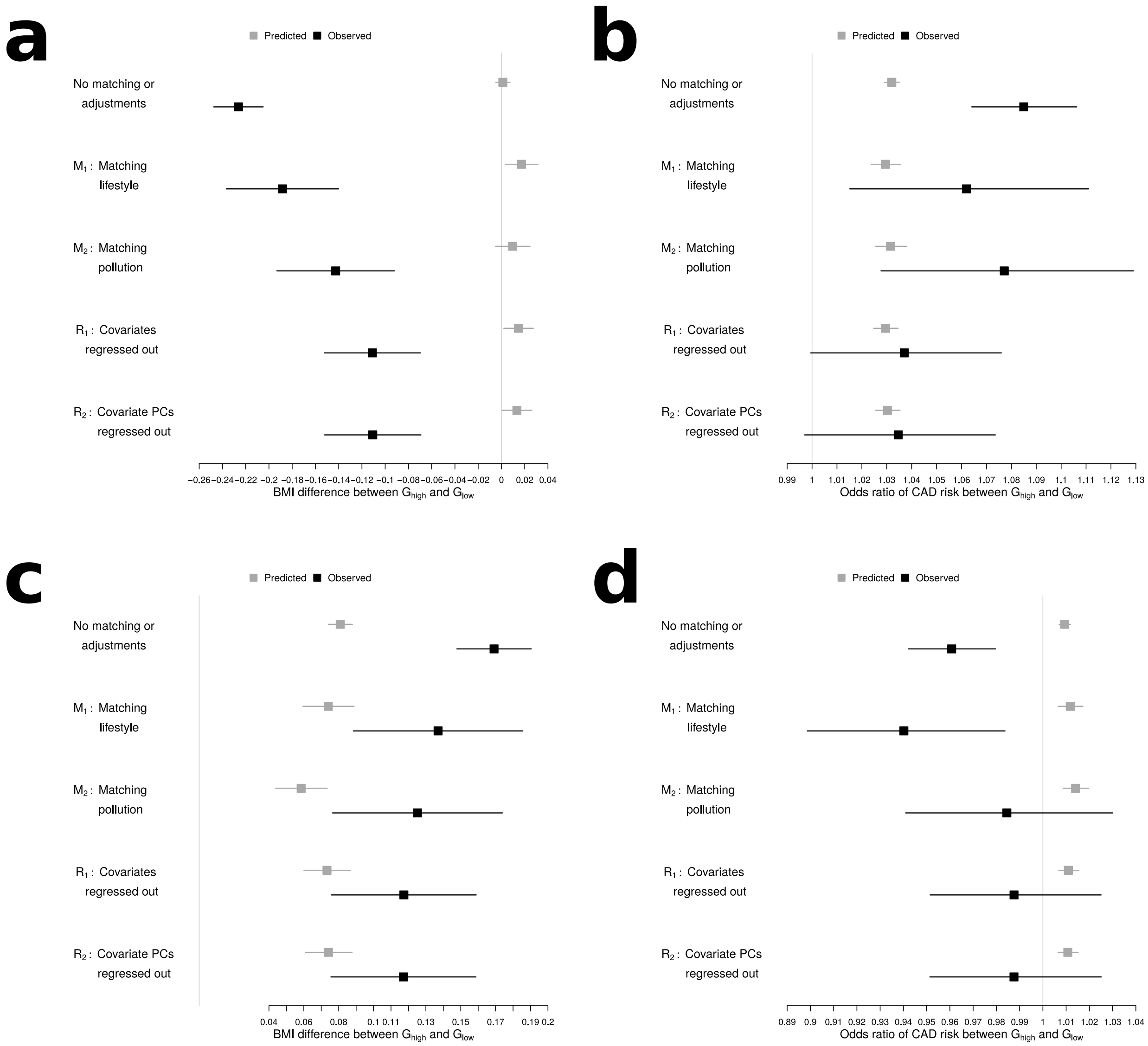

Figure 5: Differences between predicted and observed differences in phenotype/prevalence on PCs 3 and 45: Point estimates (boxes) and 95% confidence intervals (lines) of predicted (in grey) vs. observed (in black) along for G_{high} and G_{low} defined along **(a)** and **(b)** PC3, and **(c)** and **(d)** PC45. **(a)** and **(c)** show the difference in mean BMI, while **(b)** and **(d)** show the odds ratio of CAD prevalence. As before, we show each analysis: no matching or adjustments (baseline), then the matching analyses, and finally the mPC analyses.

References

[1] Ali Torkamani, Nathan E. Wineinger, and Eric J. Topol. The personal and clinical utility of polygenic risk scores. *Nature Reviews Genetics*, 19(9):581–590, September 2018.

[2] Stephen Burgess and Simon G Thompson. Use of allele scores as instrumental variables for Mendelian randomization. *International Journal of Epidemiology*, 42(4):1134–1144, August 2013.

[3] Tom G Richardson, Sean Harrison, Gibran Hemani, and George Davey Smith. An atlas of polygenic risk score associations to highlight putative causal relationships across the human phenome. *eLife*, 8:e43657, March 2019.

[4] Sulev Reisberg, Tatjana Iljasenko, Kristi Läll, Krista Fischer, and Jaak Vilo. Comparing distributions of polygenic risk scores of type 2 diabetes and coronary heart disease within different populations. *PLOS ONE*, 12(7):e0179238, July 2017.

[5] Alicia R. Martin, Christopher R. Gignoux, Raymond K. Walters, et al. Human Demographic History Impacts Genetic Risk Prediction across Diverse Populations. *The American Journal of Human Genetics*, 100(4):635–649, April 2017.

[6] Sini Kerminen, Alicia R. Martin, Jukka Koskela, et al. Geographic Variation and Bias in the Polygenic Scores of Complex Diseases and Traits in Finland. *The American Journal of Human Genetics*, 104(6):1169–1181, June 2019.

[7] Hakhamanesh Mostafavi, Arbel Harpak, Ipsita Agarwal, et al. Variable prediction accuracy of polygenic scores within an ancestry group. *eLife*, 9:e48376, January 2020.

[8] Clare Bycroft, Colin Freeman, Desislava Petkova, et al. The UK Biobank resource with deep phenotyping and genomic data. *Nature*, 562(7726):203–209, October 2018.

[9] Amit V. Khera, Mark Chaffin, Krishna G. Aragam, et al. Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations. *Nature Genetics*, 50(9):1219–1224, September 2018.

[10] Michael Inouye, Gad Abraham, Christopher P. Nelson, et al. Genomic Risk Prediction of Coronary Artery Disease in 480,000 Adults. *Journal of the American College of Cardiology*, 72(16):1883–1893, October 2018.

[11] 23andMe Research Team, COGENT (Cognitive Genomics Consortium), Social Science Genetic Association Consortium, et al. Gene discovery and polygenic prediction from a genome-wide association study of educational attainment in 1.1 million individuals. *Nature Genetics*, 50(8):1112–1121, August 2018.

[12] Amit V. Khera, Mark Chaffin, Kaitlin H. Wade, et al. Polygenic Prediction of Weight and Obesity Trajectories from Birth to Adulthood. *Cell*, 177(3):587–596.e9, April 2019.

[13] Bjarni J. Vilhjálmsson, Jian Yang, Hilary K. Finucane, et al. Modeling Linkage Disequilibrium Increases Accuracy of Polygenic Risk Scores. *The American Journal of Human Genetics*, 97(4):576–592, October 2015.

[14] Stephen Leslie, Bruce Winney, Garrett Hellenthal, et al. The fine-scale genetic structure of the British population. *Nature*, 519(7543):309–314, March 2015.

[15] G. McVean. A genealogical interpretation of principal components analysis. *PLoS Genet*, 5(10):e1000686, 2009.

[16] Mashaal Sohail, Robert M Maier, Andrea Ganna, et al. Polygenic adaptation on height is overestimated due to uncorrected stratification in genome-wide association studies. *eLife*, 8:e39702, March 2019. Publisher: eLife Sciences Publications, Ltd.

[17] Jeremy J Berg, Arbel Harpak, Nasa Sinnott-Armstrong, et al. Reduced signal for polygenic adaptation of height in UK Biobank. *eLife*, 8:e39725, March 2019.

[18] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2020.

[19] Ada Hui, Asam Latif, Kathryn Hinsliff-Smith, and Timothy Chen. Exploring the impacts of organisational structure, policy and practice on the health inequalities of marginalised communities: Illustrative cases from the UK healthcare system. *Health Policy*, 124(3):298–302, March 2020.

[20] Gad Abraham, Yixuan Qiu, and Michael Inouye. FlashPCA2: principal component analysis of Biobank-scale genotype datasets. *Bioinformatics*, 33(17):2776–2778, September 2017.

[21] Christopher C Chang, Carson C Chow, Laurent CAM Tellier, et al. Second-generation PLINK: rising to the challenge of larger and richer datasets. *GigaScience*, 4(1):7, December 2015.

[22] Shaun Purcell and Christopher Chang. Plink 1.9b_5.2.

[23] Samuel A. Lambert, Laurent Gil, Simon Jupp, et al. The Polygenic Score (PGS) Catalog: an open database to enable reproducibility and systematic evaluation.

Online Methods

Except where otherwise noted, all analyses were performed in R version 4.0.0. [18].

M1 Study population

The UK Biobank (UKB) is a prospective cohort of of about half a million individuals from the United Kingdom, recruited between the ages of 40 and 69 [8]. The full dataset is multiethnic, but our analyses were concentrated on the subset of "White British" individuals, that were defined as those who identified as "British" on the ethnicity question (field 21000) and who clustered together in the UKB principal component analysis (PCA) on PCs 1 and 2, for a total of 409,308 individuals. These people were also identified as "Caucasian" in field 22006 (genetic ethnic grouping). We selected this subset as we wished to avoid confounding due to systemic biases affecting access to and quality of healthcare in the UK [19]. Given that it represents 81.45% of the whole of the UKB, the genetic architecture of a given trait in this population will have a heavy influence on the results of genetic analyses that use the full UKB cohort. The analyses shown here were conducted under UK Biobank project number 49731.

M2 Principal component analysis of the white British subset

We used flashPCA [20] to calculate the top 50 PCs on the unrelated white British UKB participants, using the imputed genotype data, QCed so that all SNPs had a minor allele frequence (MAF) ≥ 0.01, have genotypes available for at least 99% of samples, a posterior probably of at least 0.9 on the imputed genotype, and whose p-values for being out of Hardy-Weinberg equilibrium were $\geq 10^{-6}$. We removed the four regions of high LD/known inversions suggested by the authors of flashPCA and used the `--indep-pairwise` function in Plink v1.9b_5.2 [21, 22] to prune the SNPs using the suggested parameters of a 1000 kilobase window, a step size of 50 variants, and an r^2 of 0.05.

In order to create this subset of unrelated people for the PCA, we removed one individual from each pair of related individuals identified in a file provided by the UKB, yielding 335,088 unrelated participants. We then used the loadings to project all 409,308 white British onto these 50 PCs. We computed the Pearson correlation coefficient between the top 40 principal components provided by the UKB over the whole dataset and our PCs computed on the white British, with strong correlation between our PC 1 and the UKB's PC 5 (correlation coefficient -0.961) and between our PC2 and the UKB's PC 9 (correlation coefficient of 0.917).

M3 Calculating genetic risk scores

We selected two Genetic Risk Scores (GRS) from the literature, one for a quantitative trait and another for a binary trait, Body Mass Index (BMI)[12] and Coronary Artery Disease

(CAD) [10], respectively. Both GRS are available at The Polygenic Score (PGS) Catalog [23], where we accessed the necessary information on the SNPs used in the scores, including their respective effect alleles and weights. We downloaded the data contained in this repository and calculated both scores in Plink v1.9b_5.2 [21, 22] with the `--score` function using the imputed UKB genetic data for each individual from the white British subset.

M4 Creating the risk groups

For each PC axis, we split the data into two groups: G_ℓ, individuals who were in the bottom 40% of the PC measurement; and G_u, individuals whose were in the top 40% for it. People who fell in the middle 20% of the PC measurements were removed from analysis for that PC in order to facilitate matching (see below). For a given risk score, we calculate the mean GRS in G_ℓ and G_u and assign the label G_{high} to whichever of the two has the higher mean GRS and the other group is correspondingly relabelled as G_{low} (**Supplement** section S2 for details).

M5 Trait definitions

Body mass index measurements were taken from field 21001. Coronary artery disease was defined in the same way as it was in Inouye *et al.*'s paper [10], using UKB fields 6150, 20002, and 20004. In the linked medical and death records, we looked for ICD9 codes 410-412, ICD10 codes I21-I24 and I25.2. Among the surgical procedure data, we looked for OPCS-4 codes K40-K46, K49, K50.1, and K75. In the self-reported data, the relevant surgical procedures were recorded as 1087, 1095, and 1581. Unlike the study's authors, we did not differentiate between incident and prevalent cases. Of the 408,729 white British individuals for whom these data were available, 23,375 (5.72%) met the above criteria for CAD.

M6 Estimating the effect of the score on the trait

We created a regression model for the trait—logistic regression for CAD and linear regression for BMI, following Inouye *et al.* [10]. With the trait as the outcome, we calculated the effects of the risk score while simultaneously adjusting for age, sex, UKB genotyping array, and the first 10 principal components calculated by the UKB, following Inouye *et al.* The regression coefficient of the risk score, $\hat{\beta}$ can be interpreted as the effect of the score on outcome risk (binary trait) or on phenotype measurement (quantitative trait), per standard deviation increase. Because the value of $\hat{\beta}$ can vary depending on which combination of covariates were included in the models, we explored the effect on the combination of covariates included in the model using a quintile approach (Table S1, **Supplement** section S1 for details). For all our analysis of CAD and its corresponding risk score, we use the $\hat{\beta}$ from the regression that used genotyping array and the risk score as its only covariates, which yielded a regression coefficient of 0.4878. Meanwhile, we kept regression coefficient from the full model on BMI (using age, sex, genotyping array, and first 10 UK Biobank provided principal components as the other covariates), for a $\hat{\beta}$ of 1.3710. We also performed sex-stratified analyses (*Supplement* section S6) in which case we removed sex from the regression model covariates but otherwise kept them the same.

M7 Predicted and observed differences

We define $\bar{s}_{low}$, and $\bar{s}_{high}$ as the mean GRSs for individuals in G_{low} and G_{high}, respectively. Using these values, along with $\hat{\beta}$, we calculate D_{pred}, the predicted odds ratio of CAD prevalence in the G_{high} compared to G_{low} as

$$D_{pred} = \exp\left(\hat{\beta}\left(\bar{s}_{high} - \bar{s}_{low}\right)\right). \tag{M1}$$

For BMI, D_{pred} is the predicted increase in mean BMI for G_{high}compared to G_{low} is

$$D_{pred} = \hat{\beta}\left(\bar{s}_{high} - \bar{s}_{low}\right). \tag{M2}$$

Next, we computed the actual differences in CAD prevalence and mean BMI between G_{low} and G_{high}. To assess how significant this observed value between PC groups was, we computed an empirical distribution of the observed difference using a resampling strategy (**Supplement**, section S3.1 for details). We resampled individuals in our dataset without regard to the PCs (or any other covariate, including age or sex) so that we created two new groups, G'_{low}, whose distributions of the given risk score matched that of G_{low}, and G'_{high}, which is defined analogously. We checked that distribution of risk scores matched using Kolmogorov-Smirnov test, requiring a p-value such that $p \geq 0.5$ before proceeding. Otherwise, the sample was rejected and redrawn. Sampling was performed without replacement so that the two groups would always be mutually exclusive. Once G'_{low} and G'_{high} were chosen, we calculated the difference in mean BMI or CAD prevalence of the groups and recorded them. We performed this sampling 1 million times to create a null distribution of these differences, given the distribution of risk scores in the cohort. We then compared the differences we find between G_{high} and G_{low} to this null distribution to get an empirical p-value (Table S2). These empirical distributions, along with the predicted and observed differences between G_{high} and G_{low} for both traits on PC1 are shown in Figure 2.

M8 Matching individuals

Because one can expect significant differences between G_{low} and G_{high} in terms of environmental and/or socioeconomic risk factors affecting the trait, we have to account for these factors. For each individual in G_{high}, we search for someone who matches them for age, sex, smoking behavior, drinking behavior, exercise, socioeconomic characteristics, and pollution exposures. If there is no sufficiently similar person in G_{low}, then the proband from G_{high} is removed from our analyses, as is anyone from G_{low} who is unmatched to someone in G_{high} once matching is finished. Matching is one-to-one—that is, every person who has a match, has exactly one match. When an individual had multiple potential matches, we selected one at random, leaving the others in the pool of potential matches. To avoid creating matches among people who were not very far apart on the PC, we remove the middle 20% of the

PC distribution, which forces a minimum PC distance between the members of a matched pair. Because it was not possible to match on every variable at once and still have a large enough sample on which to perform analysis, we matched on two sets of variables, M_1 and M_2, and the variables used in each are reported in Table S3 (**Supplement** section S4.1 for details). The thresholds for matches on each variable were found by balancing the need to keep samples with the need to ensure that there were no differences between G_{high} and G_{low} in the distributions of the variables in M_1 and M_2. Matching on the variables in M_1 typically removed about two thirds of the datapoints, leaving between $106,000 - 110,000$ individuals on which to compare CAD prevalence or mean BMI. The M_2 criteria removed more individuals (up to 72%) of the cohort, leaving between 91,500 and 107,500 individuals in both cases.

M9 Modified principal components

Another way of solving the problem of differences in covariate distributions between G_1 and G_2 is to regress the pollution and socioeconomic covariates out of the principal components and then perform our analyses on these modified PCs (mPCs). This solves the problem of finding enough suitable matches to retain enough samples for further analysis when considering all variables at once. Additionally, this process allows us to account for variables that were not included in the previous matching. The full list of variables used is reported in section S4.2 of the **Supplement**. While we do lose individuals due to missing data, we are able to retain 168,607 individuals for each mPC analysis, which is more samples than with matching. We performed an initial regression, which we will call R_1, where the mPC measurements were the residuals from the linear regression of all the above variables on the original PC. However, the above variables are not all independent of each other, and are in some cases—as with nitrogen dioxide air pollution measures—highly correlated with one another. As an alternate way of generating the mPCs, which we will call R_2, we performed a PCA on the matrix of covariate measures to remap them into a space where each variable was totally independent of all the others. We emphasize that the goal of this analysis was to remove the correlation structure among the covariates, and not to reduce the number of covariates tested. The remapped covariates were used in the regression to create the mPCs of genetic data. The PCA of the environmental and socioeconomic factors used in the second regression was performed using a singular value decomposition, implemented using the `prcomp` function in R, with the parameters set to scale and center the matrix before performing the PCA.