Comparison of polygenic risk scores for heart disease highlights obstacles to overcome for clinical use

Holly Trochet[1], Justin Pelletier[1], Rafik Tadros[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 or to julie.hussin@umontreal.ca.

September 28, 2021

Abstract

Polygenic risk scores, or PRS, are a tool to estimate individuals' liabilities to a disease or trait measurement based solely on genetic information. One commonly discussed potential use is in the clinic to identify people who are at greater risk of developing a disease. In this paper, we investigate the suitability of three large PRS for coronary artery disease (CAD) for clinical use. In the UK Biobank, the cohort which was used in the creation of each score, we calculated the association between CAD, the scores, and population structure for the white British subset. After adjustment for geographic and socioeconomic factors, CAD was not associated with population structure; however all three scores were confounded by genetic ancestry, raising questions about how these biases would impact clinical application. Furthermore, we investigated the differences in risk stratification using four different UK Biobank assessment centers as separate cohorts and tested how missing genetic data affected risk stratification through simulation and comparisons to scores calculated in French Canadians. We show that missing data impact classification for extreme individuals for high- and low-risk, and quantiles of risk is sensitive to individual-level genotype missingness. Distributions of scores varied between assessment centers, revealing that thresholding based on quantiles can be problematic for consistency across centers and populations. We highlight three criteria that a genetic risk score must fulfill in order to be used to stratify patients in the clinic: 1) it must be robust to population structure 2) scores must provide absolute thresholds for risk levels, and 3) there must be an approach to compensate for missing genetic data. Finally, we propose potential avenues of improvements for determining individual's genetic liability to a complex trait such as CAD.

Introduction

In the last decades, the need for improvement of risk prediction for human disease has led to the development of many genetic risk scores, which are metrics that use solely genetic

data to predict the likelihood of an individual to develop disease. In this highly active area of research, most papers cite the potential clinical value of genetic risk scores in identifying individuals at high risk of developing the disease for early intervention[1]. With the proliferation of direct-to-consumer genotyping panels and as the cost of whole genome sequencing continues to fall, we may not be far from the day where one's genome sequence becomes a standard part of one's medical record.

In anticipation of this, a number of polygenic risk scores (PRS), which are the subset of genetic risk scores that explicitly use a large number of genetic markers (thousands or millions) from across the genome, have been published for many diseases, with coronary artery disease seeming to be a popular and well-respected choice for potential clinical use[2,3]. These PRS include variants that do not have well-established associations (either through genome-wide association studies or functional studies) with the trait. PRS are also 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[4,1]. Guidelines are now being put forward[5] for appropriate clinical application and reporting of these scores.

Population structure is a concern in medical and statistical genetics, as it may lead to spurious results in association studies, and PRS inherit this problem. It has been previously established that health outcomes in the UK Biobank (UKB) are associated with the population geographic distribution, as are their corresponding PRS[6]. It has also been shown that PRS can be biased by recent demographic history and environmental structure that cannot be corrected for using PCs[7]. Additionally, differences in technologies (sequencing, genotyping), quality control (QC) pipelines, imputation pipelines and need for

lift over from a genome assembly to another, are all sources of potential missing data in genetic data sets. This may result in the removal of markers that are defined in the score, especially of low-frequency variants[8]. Because these values are calculated within a cohort, markers that are kept to construct a score will vary between cohorts.

PRS are often assessed by testing how well they predict the phenotype (with or without other covariates in the model) or by dividing individuals according to quantiles, with the lowest and highest quantiles being of particular interest. They are also usually normalized within the cohort to assess the mean effect on risk per standard deviation increase in the score. These scores are generally touted for their ability to identify "high-risk" individuals for early intervention, however it is hard to define what the appropriate threshold would be for the PRS in order to determine who these high-risk people are. Therefore, important questions about data completeness, population stratification and appropriate threshold definition arise, highlighting the need for careful examination of what is exactly captured by the proposed scores, before their translation into clinical care can become a reality.

In this paper we examine three scores for coronary artery disease (CAD): the metaGRS by Inouye *et al.*[3], the K2018 score by Khera *et al.*[2], and the E2020 score by Elliott *et al.*[9]. All three scores tuned parameters in subsets of the UKB, and then validated on the rest of the cohort. The subset and the validation set of UKB used to develop the metaGRS and the E2020 scores included all UKB participants, while the K2018 score was restricted to the white British subset (81.45% of the cohort). For the metaGRS, random linkage disequilibrium (LD) pruning was used to generate candidate sets of markers in training data from the UKB (1000 randomly selected prevalent CAD cases and 2000 controls) and the set that had the highest hazard ratio in the training set was selected. Weights were chosen by combining the weights from three previously published CAD PRS plus the effect size estimate of the marker on CAD in the UKB. The K2018 score was constructed using the algorithm LDpred[10], while the E2020 score used the software lassosum[11]. Both

pieces of software were used to generate candidate scores, and in both cases, the authors selected the score with the highest predictive value, measured by area under the curve (AUC) of the receiver operating characteristic (ROC) curve in a logistic regression model with CAD as the outcome.

Since all three scores were built using the UKB[12], we revisited these scores in this cohort to investigate three important aspects that any genetic risk score should demonstrate robustness for before it could be successfully implemented for clinical use: population structure, absolute thresholding for high-risk/low-risk individuals, and missing data. We show that all three scores demonstrate limitations on the basis of these criteria and explain why one should be careful in applying PRS in clinical practice to identify future outcomes. If these scores are fundamentally biased by both biological and technical features intrinsic to the genetic data they use, then this demands methods that account for the effects we describe. We also discuss potential avenues of improvements for PRS methodologies before they should be cleared for clinical use.

Results

Robustness to Population Structure

Throughout time and evolution, all populations' genetic makeup is subject to change due to mutational events, genetic drift, natural selection and demographic events such as expansion or bottlenecks, migration and admixture[13]. These events lead to systematic differences within and between populations due to variations in patterns of genetic diversity, such as allele frequency[14] or LD[15]. This is known as population structure, and it can induce spurious results in association studies of genetic markers and traits[16,17,18]. Specifically, if one population (or subpopulation) has a different prevalence of a trait or different mean phenotype measurement than the other, then the genetic markers that differentiate the two groups can be found to be associated with the trait, even if they do not have any

biological relationship to it. A common method of examining population structure is principal component analysis (PCA)[19,20]. PCA is a statistical procedure which uses orthogonal transformation to a set of genotypes into a set of values of linearly uncorrelated variables called principal components[21], which are used as low-dimension summaries of the original genetic data. This method can be used in population genetics to visualize distance and relatedness between populations/individuals[14,22] relative to other populations/individuals included in the analysis. PCs are frequently included as covariates in genome-wide association studies as a way of accounting for population structure. Like PRS, PCs are summaries of genetic information. An individual's PC coordinates describe where individuals sit on axes that are constructed to preserve as much variance as possible.

Fine-scale population structure in the white British subset

By far the largest group in the UKB are those who self-identified as "British" on the ethnicity question (UKB field 21000) and who clustered together in the PCA provided by the UKB. These individuals, henceforth known as the "white British", comprise 81.45% of the data set. We performed a principal component analysis specific to this group (Figure 1a). We see some separation of those born in Wales from those born in Scotland/Ireland. The participants who were born in England are the overwhelming majority of the cohort, and they appear throughout the plot.

Unsurprisingly, the first three PCs are all associated with the geographic distribution of individuals, both at birth and at the time of assessment. For a description of these and other variables used in our analyses, see Section M1.8. Both the home location at assessment variables as well as the place of birth ones were associated with each PC at a level of $p < 2 \times 10^{-16}$ for all univariate associations and for multivariate associations fitting all three PCs at once. These PCs were also associated with each PRS (Table S4).

A relationship between a PRS and a PC is not inherently problematic if it reflects true differences in disease risk across the population. Indeed, CAD is associated with the

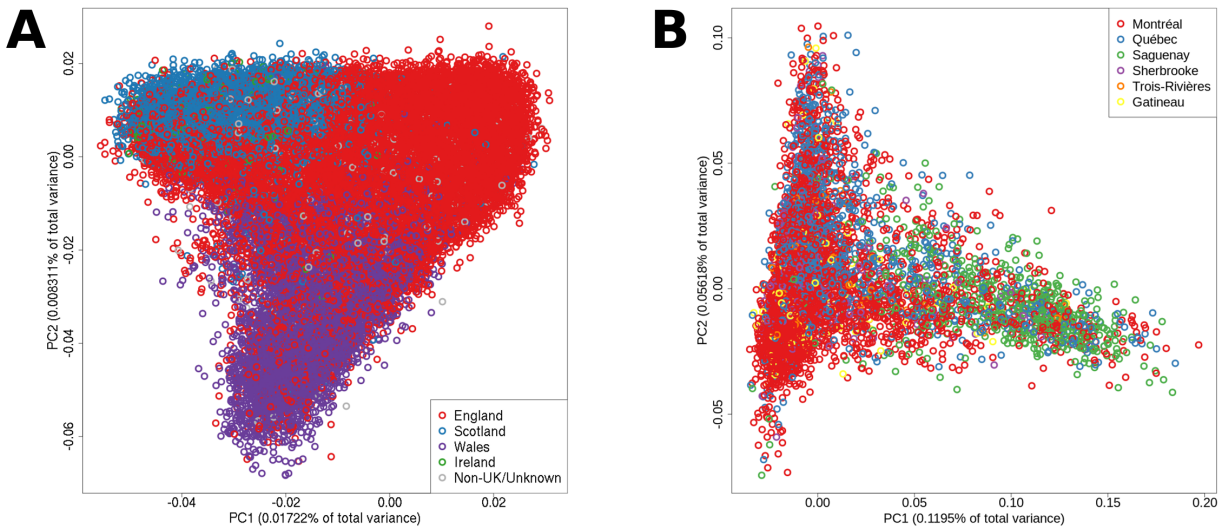

Figure 1: First two principal components of PCAs conducted in **(A)** the white British subset of the UK Biobank, where individuals are colored by their country of birth (England, Scotland, Wales, Ireland, or unknown), or **(B)** CARTaGENE, where individuals are colored by the region in which they were recruited.

first three PCs, however, several CAD risk factors, such as age, smoking behavior, and indices of deprivation are all with the PCs and could be responsible for association with the disease. Additionally, CAD prevalence shows an association with individuals' places of birth as well as their home locations, which are also associated with the first three PCs. Since population structure arises non-random mating within a population, often due to geographical and social barriers (for example, social class or religious sectarianism), it is unsurprising that the PCs are associated with these environmental variables. The question is whether or not the population differentiation captured by the PCs is a risk factor for CAD independently of these environmental factors, including geographic coordinates.

After adjustment for age, sex, place of birth, home location, Townsend deprivation index, income, age when completed full time education, and smoking status, we see that in all cases, at least one PC remains associated with each PRS at a significance level of $p \leq 1.5 \times 10^{-7}$, while none of the PCs is even nominally associated with CAD itself (Table S3). In Figure 2, we plot the p-value of association between the PCs and the different risk scores (Panels A, B, and C) under different linear regression models. The simplest model (on the far left in each panel) uses just the PCs, and progressively we add age and

sex, and various social, environmental, and geographic variables to the model. In all cases at least one PC remains strongly associated with each PRS, even after these adjustments. In contrast, in Panel D, we perform the same analysis, but as a logistic regression of CAD on these same models, again starting with just PCs and progressively adding more covariates. In this case, once all of these variables are added to the model, none of the PCs is even nominally associated with CAD. These results are also reported in Table S2, along with the Akaike information criterion, which shows that the addition of social and environmental coviarates improves the model fit. Therefore, the direct associations we observe between the PRS and the PCs (Table S4) do not reflect a true population cline of CAD risk along the PC axes. These results indicate that the scores capture differences in susceptibility to CAD within the white British population due to the correlation between population structure and environmental risk factors for the disease, rather than differences in the frequencies of alleles that directly affect genetic predisposition.

What we are highlighting is that all three PRS predict statistically significant differences across the white British population on the basis of genetic ancestry captured by PCs, even after adjustment for environmental variables. However, these PCs do not appear to be associated with the target trait, CAD, once these same environmental variables are accounted for, and as we observe in Table S2, it appears that the models that account for environmental effects fit the data better, according the the AIC. This suggests that the differences across the population predicted by the scores are spurious and this association between the scores and genetic PCs will result in the misclassification of individuals' genetic risk, as whose genetic PCs are in the extremes of the distributions are more likely to have scores that put them in the high- or low-risk category according to the PRS.

Risk scores in subpopulations of the UK Biobank

To date, most genetic research has been conducted in individuals of European ancestry, and so most genetic risk scores have been built based on data from European populations.

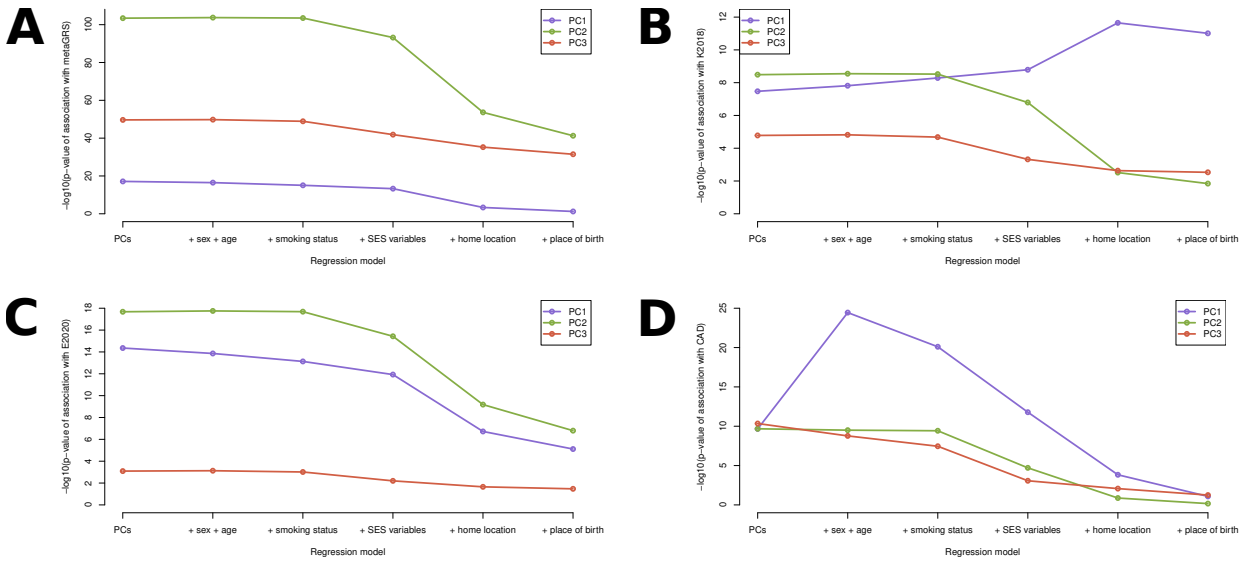

Figure 2: Significance of association between specific PCs (depicted by color) and **A)** the metaGRS, **B)** the K2018 score, **C)** the E2020 score, and **D)** CAD. $-\log_{10}(P\text{-values})$ are shown for different regression models, each successively adding covariates to the one that preceded it. The first model regressed the score or CAD on just the first three PCs; the second included the first three PCs and added sex and age; the third added smoking status (current, previous, or never); the fourth added socioeconomic variables (Townsend deprivation index, income, their interaction effect, and age completed education, adapted to include university); the fifth added home location at assessment (northern and eastern coordinates); and the sixth model added place of birth in the UK (northern and eastern coordinates). All regressions were performed on the same 334,181 white British individuals who had no missing data on any of the potential covariates.

For the three scores under investigation in our study, we observe that non-European subpopulations of the UKB (South Asian; Black British, African, or Caribbean; and Chinese) show different distributions of each CAD risk score compared to the European ones (white British, Irish, and other white) (Figure S1).

Furthermore, it has been demonstrated that PRS have much lower correlations with their target traits in African and East Asian populations than they do in Europeans[23,24]. We tested if these findings are replicated for the PRS under investigation (Table S1). We noticed that while the non-European subpopulations (especially the Chinese) are underpowered compared to the European ones, the point estimates for the effect sizes of each score are consistently lower for the non-Europeans than for the Europeans, except for the E2020 score, where South Asians have a higher point estimate and lower p-value than for the "other white" group. We also noticed that the estimated effect sizes on the scores are considerably lower in the Black, African, and Caribbean group than in all the others.

Population structure in CARTaGENE

It was shown that scores developed in UKB were also confounded by population structure when applied to other European populations, such as the Finnish population[25]. Whether this result is generalizable to other CAD scores and in other populations of European ancestry, which were not used in the development of these scores, remained to be demonstrated. To investigate this, we calculated each of these scores for 9,447 French-Canadian individuals with genetic data in CARTaGENE[26]. There are over 6 million French-Canadians today, most of whom are descended from a small founder population of 8,500 who colonized the province of Québec between the early seventeenth century and the mid eighteenth. In addition to this initial bottleneck, founder effects can be detected at the regional level due to the patterns of settlement around the province[27].

All three scores were strongly associated with CAD (defined following the definition from Wünnemann *et al.*[28]), with $p = 8.03 \times 10^{-15}$ for the E2020 score, and $p < 2 \times 10^{-16}$ for

the other two. Additionally, all three scores showed strong associations ($p \leq 7.12 \times 10^{-16}$) with PC1 which, in this cohort, differentiates individuals from the Saguenay-Lac-St-Jean (SLSJ) region from the rest of the population (see Figure 1b). SLSJ is a region of Québec whose population has gone through multiple bottlenecks and a rapid population expansion[29,30]. These results suggest that all three scores would assign a higher genetic risk of CAD to people whose ancestors come from this region. However, there is no association between this PC axis and CAD risk in this cohort, both before and after adjustment for age, sex, income, and educational attainment. We note, however, that the prevalence of CAD in this cohort is small, with only 292 cases (244 men, 48 women), and so may be underpowered to discover a true signal. It is however worrying that associations between the PRS and population structure appear even in a cohort that was not part of the scores' derivations.

Thresholds for risk

PRS are usually evaluated within a cohort, by dividing individuals according to quantiles of their PRS, with the lowest and highest quantiles being of particular interest. The ability to identify "high-risk" individuals for early intervention is often used as a selling point of these scores, although there is often no indication given on what an appropriate threshold would be for the PRS in order to determine who these high-risk people are. Furthermore, if scores vary across the population in relation to population structure and geography, then each institution will have a unique mean and standard error, as we observe in Table 1, where we calculated the means and standard deviations for each score in white British individuals who attended four different UKB assessment centers, and for the white British cohort as a whole. Additionally, we calculated the empirical thresholds for the top 10 and 5% of risk scores at each center. As expected from our examination of these scores and population structure, we see that each center has a subtly different distribution of each of the scores from the others, whose statistical significance we tested using

Kolmogorov-Smirnov tests between each pair of centers for all three scores. We indicate the ones that were significantly different ($p < 2.78 \times 10^{-3}$, to account for multiple testing) in Table 1 and report the p-values of these tests in Tables S8, S9, and S10.

Table 1: Sample size, mean, standard deviation (SD), and thresholds for the top 10% and top 5% of risk scores among white British participants from four different UK Biobank assessment centers, as well as the whole subset ("All"). For each center and each score, we calculated the mean and SD for the subset of white British participants who attended a given assessment center. We also calculated these values for the full cohort. We report significant difference of distributions (P-values of Kolmogorov-Smirnov tests < 0.00287, which is the Bonferroni-corrected 0.05 threshold) compared to Reading (^{r}), Cardiff (^{c}), Newcastle (^{n}) and Glasgow (^{g}) assessment centers. All P-values of Kolmogorov-Smirnov tests of the pairwise equivalence of the distributions between assessment centers are reported in Tables S8, S9 and S10.

Score		Reading	Cardiff	Newcastle	Glasgow	All
	Size	22,635	15,462	28,053	14,205	408,567
metaGRS	Mean	^{n,g}-0.7127	^{n,g}-0.7115	^{r,c}-0.6675	^{r,c}-0.6617	-0.6907
	SD	0.4478	0.4501	0.4474	0.4479	0.4483
	10%	-0.1380	-0.1329	-0.0893	-0.0871	-0.1153
	5%	0.0257	0.0453	0.0724	0.0753	0.0500
K2018	Mean	^{n}18.0323	18.0332	^{r}18.0363	18.0330	18.0340
	SD	0.0850	0.0856	0.0844	0.0848	0.0849
	10%	18.1406	18.1432	18.1456	18.1407	18.1427
	5%	18.1719	18.1747	18.1772	18.1736	18.1746
E2020	Mean	^{n,g}-3.0219	^{g}-3.0186	^{n}-3.0152	^{r,c}-3.0138	-3.0193
	SD	0.1453	0.1446	0.1450	0.1456	0.1451
	10%	-2.8346	-2.8332	-2.8300	-2.8264	-2.8330
	5%	-2.7843	-2.7763	-2.7773	-2.7739	-2.7806

Where differences in the distributions of the scores exist, they appear between the two northern cities (Glasgow and Newcastle) and the two southern ones (Cardiff and Reading). The metaGRS shows the strongest and most consistent differences between these groups. In contrast, the K2018 score has the most consistent distributions across the four

centers, with a statistically significant difference between Reading and Newcastle only, which may be detectable due to the relatively larger sample sizes in these two assessment centers. Finally, the E2020 score shows the strongest differences between Reading and the two northern cities, though there is also a statistically significant difference between Glasgow and Cardiff.

These differences are problematic because a person's genetic make-up is constant and so their PRS and the subsequent characterization of their genetic risk ought to be as well. In other words, a person's genetic predisposition does not change when they move to a new city. Additionally, normalizing within the available data means that individuals who are assessed at institutions where people are disproportionately high-risk may miss out on interventions that could be helpful to them. Conversely, people who are assessed at institutions where people are disproportionately low-risk may be given medical interventions inappropriately, exposing themselves to side-effects for minimal potential benefit.

Concordance of the scores

Because PRS tend to be defined in relative terms, one natural question when looking across multiple ones is how consistent the results are—that is, given a cohort, do different scores identify the same individuals as high-risk? To answer this question, we calculated the Spearman rank correlation between each of the three pairs of CAD PRS investigated here in the white British subset of the UKB. The highest correlation is between the metaGRS and the K2018 score, which is 0.7676. The next highest was between the K2018 and the E2020 score, at 0.6713. Finally, the correlation between the metaGRS and the E2020 score was 0.5797. These correlations are not as high as one might expect, suggesting that there is a good deal of variation among the scores who is identified as high- or low-risk.

Another way to interrogate this question is to look at the high- and low-risk individuals for each score to evaluate their overlap. The proportion of overlap of white British UKB participants who scored in the highest and lowest 10% of each score is reported in

Table S5. In the best case scenario, when comparing the metaGRS to the K2018 score, a little over half (55.30%) of individuals are identified as being in the top 10% of risk by both scores. The proportion of overlap among all three scores was 0.2917 for the low-risk group and 0.2912 for the high-risk one.

While there is some consistency across the scores in who is identified as high- or low-risk, the agreement among the scores is not as strong as we might expect or want for individuals in the extremes of the distribution. Additionally, there are a small number individuals who are defined by one score as being high-risk and by another as low-risk (highest or lowest decile of risk score), reported in Table S6. Consistent with the Spearman correlation results and the results in Table S5, the largest discrepancies were observed between the metaGRS and the E2020 scores. These analyses raise questions about which score is the most accurate or appropriate in a clinical setting; however assessing this is beyond the scope of this paper.

Impact of Missing Data

Genotyping chips, imputation pipelines, and quality control filters will vary between cohorts, which may cause difficulties in calculating PRS because of the removal of markers that are defined in the score, creating differences in genotyping data across cohorts, even when they use the same technology. In Table S11, we show this for UKB and CARTaGENE, in relation to the markers used in each PRS. Both cohorts were imputed using the Haplotype Reference Consortium[31] as the reference population, and in both cases, we used imputed markers with a genotype probability of at least 0.9 and removed all markers with a missingness rate greater than 0.01 across the cohort. We see that the imputed data that passed our QC for the UKB contains at least 99.7% of the markers used in each of the PRS. However, the imputed data for CARTaGENE contains between 73.9% and 86.0%. This leads to markedly different distributions of the scores (Figure 3). When we calculate the scores on the set of markers available in both cohorts, the distribution of

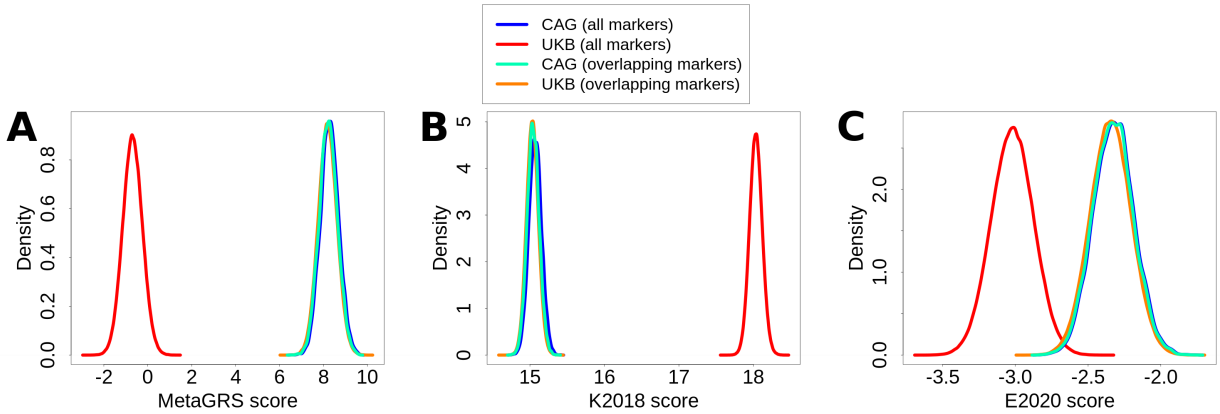

Figure 3: Distributions of raw risk scores for the three risk scores, calculated in the French Canadians of CARTaGENE (CAG) and the white British of the UK Biobank (UKB) on two different sets of markers in each cohort: the intersection of markers genotyped or imputed in both cohorts with high (> 0.9) genotype certainty (overlapping), and the full set of these markers available in each cohort individually (all).

scores for both cohorts are more similar to each other (though small, but statistically significant differences in their means remain). Since CARTaGENE contained information on fewer markers than UKB, the distributions on the overlapping marker set resemble that of CARTaGENE more than they do that of UKB. These results show how the set of markers available to calculate a PRS can drastically change the values of the raw scores, which will, in turn, affect the interpretation of the score. This problem is surmountable if suitable thresholds have been previously developed in a cohort that is similar in terms of population and data collection pipeline.

In addition to cohort-wide missing genotype data, most QC pipelines tolerate some amount of missing genotype data for each individual. This means that most, if not all, participants are missing genotypes at a small set of random markers, which changes from person to person. To handle this type of missingness when calculating the risk scores in each cohort, the mean effect allele dosage calculated in the rest of the cohort is often used. This discounts the effect of the marker on the score of that individual, but has downstream consequences on the interpretation of individuals' levels of risk.

To investigate the effect of individual missingness, for each PRS, we identified the 200 people with the most extreme scores (the highest 100 and the lowest 100). We then

randomly removed their genotypes at markers used in the scores. We repeated this process 10 times and took the average of these new scores over all replicates for these individuals. For a more detailed description of this process, see Section M1.7. In Table 2, we report how many individuals still had the 100 highest or 100 lowest scores for each level of missingness.

Table 2: Number of people who originally had the highest and lowest 100 scores for each risk score who, after setting a random group of genotypes missing, still have the most extreme risk scores.

Missingness	Extreme	metaGRS	K2018	E2020
0.01	Highest	89	92	91
	Lowest	96	94	91
0.02	Highest	82	87	85
	Lowest	90	85	86
0.05	Highest	60	69	64
	Lowest	71	70	66
0.10	Highest	35	49	39
	Lowest	45	45	35

Predictably, as individual missingness increases, the number of people whose scores remain in the extremes drops. In practice, it is unlikely that anyone with a genotype missing rate as high as 10% would be retained in analysis, but it is useful to see the trajectory. The reason for this phenomenon is illustrated in Figure S2, which shows that for each score, as missingness increases, the scores tend towards the mean.

When allele frequency data from a wider cohort is not available, there is no way to calculate a mean effect allele dosage. In this case, missing markers are effectively removed from the score, which is the equivalent of treating individuals as though they are homozygous for the non-effect allele at the missing marker. For scores that are defined solely in terms of risk increasing alleles—as is the case with the K2018 score—this will result in in-

dividuals with high missing genotype rates having lower risk estimates than those whose missing genotype rates are lower. When the score is defined as a mix of risk increasing and decreasing alleles—as the metaGRS and E2020 scores are—then individuals with higher missingness rates will tend to have less extreme scores. These results demonstrate that individual variation in missingness rates will impact the stratification of disease risk in the cohort.

Discussion

As a result of our findings, we propose that all GRS should fulfill the following criteria before being put into clinical use. Specifically, scores should

- demonstrate robustness to population structure.
- provide absolute thresholds for high-risk vs normal or low-risk.
- have a way of compensating for missing data.

It remains highly problematic that current PRS predict genetic differences in CAD risk across the PC axes that do not appear to exist in reality. In a clinical setting, this could lead to unintentional and erroneous pathologizing of genetic ancestry. There is also a risk that social problems, such as poverty and unequal access to quality education, food, and medical care remain unresolved due to the perception that the groups who suffer disproportionately from these problems are simply genetically more prone to disease. It is vital that authors of PRS investigate the degree to which their scores are associated with genetic population structure in their target populations. Ideally, they would demonstrate that any association between population structure and the PRS is in proportion to the association between population structure and the trait itself. However, in practice it may not be possible to completely separate a markers' effect on the trait from an effect it has that is mediated through social or environmental factors.

As with most other studies of PRS, we have neglected gene $\times$ gene and gene $\times$ envi-

ronment interactions. While one's genetic make-up does not change when one moves from one city to another, the interaction between one's environment and one's genetics might. Mostafavi *et al.* have previously demonstrated that PRS have varying predictive values across various social and environmental strata[32], which suggests that there may indeed be gene × environment effects that alter the interpretation of scores across different environments, but we have not investigated this in our own work.

We also advocate for absolute thresholds for risk, which would bring these scores in line with traditional CAD risk factors, such as measures of obesity (waist-hip ratio and body mass index), blood pressure, cholesterol levels. All of these have absolute thresholds that differentiate between low, normal and increased risk, even if some of those thresholds are not constant across ethnic groups or the sexes. To our knowledge, no one has ever advanced an argument for why genetic risk scores should be the exception to this. Furthermore, the current practice of normalizing within a cohort means that the score cannot be compared across cohorts or populations. This introduces a problem raised by Martin *et al.*[24]: where PRS can be used only in the majority ethnic group served by the institution, such that members of minority ethnic groups will miss out on the potential benefits, which will create or exacerbate existing healthcare disparities between these groups. Finally, absolute risk thresholds would make PRS more accessible to patients, since it means they could know their raw score and its interpretation the same way they do for measures like LDL cholesterol or BMI.

Missing data are an inevitability, and PRS—which incorporate such a large number of markers—need to provide ways of accounting for that. The imputation problems outlined in this study will be ameliorated somewhat once whole genome sequencing becomes standard, but sequencing data will still need to pass through quality control pipelines which will unavoidably lead to a loss of data at the cohort level, as well as for each individual. As a result, authors of PRS need to provide some guidance on how to deal with missing genetic data. For small genetic risk scores, it may be possible to use proxy mark-

ers in LD with the missing position, which are highly correlated with those in the risk score. However, as LD structure can vary even between closely related individuals (eg. French-Canadians vs. other European populations) this is a challenging endeavour. Furthermore, this becomes harder to do for larger scores where the best available proxies may also be included in the score. One possible area of future work could be in the development of algorithms that can readjust a score's weights based on markers are available.

We believe that improvement on current scores is possible. Part of the reason why the problems we have outlined arise is because of the way these scores are constructed and assessed. There is an implicit assumption that people who show a phenotype must have an increased genetic predisposition for it. While this is true for some traits—for example, eye color[33]—there is a large environmental component to CAD which could potentially overwhelm the underlying genetic predisposition. These environmental factors can create cases out of people at low genetic risk for the disease and prevent those at elevated genetic risk from developing it. The process of creating and validating these scores focuses on their ability to predict phenotypes from genotypes. For CAD, this might be a perverse incentive, since it rewards the PRS for including loci whose associations with the trait are mediated through a social or environmental covariates—that is, genetic artifacts of social and environmental risk. This means that if two individuals have the same true underlying genetic risk, but one of them develops CAD and the other does not, a potential score that assigns a higher value to the individual that develops CAD will be favored over on that assigns them the same value. One way of avoiding this problem may be to restrict the genetic risk scores only to loci with well-validated associations with CAD. These scores would use a much smaller number of markers, allowing for proxy SNPs to be used when genotype data is missing, and may show less spurious association with population structure.

Finally, it is worth questioning whether building a PRS for CAD itself is the best way of capturing the genetic liability for it, and whether it makes sense to have a sin-

gle score for both men and women. Traditional risk factors were highly associated ($p < 2 \times 10^{-16}$ with all three of our scores and a high PRS for any of the scores predicted hypertension, high cholesterol, diabetes, and obesity (measured by BMI and waist-hip ratio). Clearly the scores use markers associated with these risk factors, and it may make sense to build separate PRS for each of them, and then combine them into a risk model, as discussed previously[34]. One advantage of these approaches would be that pleiotropic effects could be accounted for. For instance, if a variant increases adiposity, but decreases HDL cholesterol, then these effects can both contribute to the final estimation of CAD via the separate scores, whereas in a single score for CAD, one of these effects may mask the other, or they might cancel each other out. Additionally, there might be more scope in using multiple scores to incorporate environmental and non-genetic biological effects, and specifically the effects of age and sex.

M1 Methods

Except where otherwise noted, all analyses were performed in R version 4.0.2.[35].

M1.1 Study populations

M1.2 UK Biobank

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[12]. 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[36]. 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.

M1.2.1 CARTaGENE

CARTaGENE is a Québec-based biobank of 43,000 participants, aged 40-60, and chosen to be representative of the general population of the province[26]. About 12,000 of these individuals have genotyping information available, and of this group, the majority (9,447) self-identified as "French-Canadian". It is in this group that we performed our analyses. The analyses shown here were conducted under CARTaGENE project number 406713.

M1.3 Principal component analysis of the white British subset of the UK Biobank

We used flashPCA[37] 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[38,39] 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).

M1.4 Principal component analysis of the French Canadians in CARTaGENE

We used flashPCA[37] to calculate the top 20 PCs on unrelated (IBS sharing proportion < 0.03 using King[40]) French-Canadian participants in CAG, using genetic data that had been imputed using the Haplotype Reference Consortium[31] as a reference panel. The resulting data was 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 $\geq 10^{-4}$. 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[38,39] 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.

M1.5 Calculating genetic risk scores

We selected three polygenic risk scores (PRS) from the literature, each predicting the risk of coronary artery disease (CAD)[3,41,9]. All three PRS are available at The Polygenic Score (PGS) Catalog[42], 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[38,39] with the `--score` function using the imputed UKB genetic data for each individual from the white British subset.

Unlike the other two scores, a high Elliott score is associated with a decreased risk of CAD. To aid in the comparison across scores, we used the negative of the Elliott score in all our analyses.

M1.6 Trait definitions

In the UK Biobank, Coronary artery disease was defined in the same way as it was in Inouye *et al.*'s paper[3], 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.

In CARTaGENE, we followed the definition of Wünnemann *et al.*, where CAD was defined by history of myocardial infarction and a keyword search among surgical procedures and other medical history fields for percutaneous coronary intervention or coronary artery bypass grafting.

M1.7 Simulating missing genotype data in the UK Biobank

For each score, we identified the people with the highest 100 and the lowest 100 scores. For each of these individuals, a random selection of genotypes at markers used by the score was set to missing so that. The score was then recalculated. This process was performed 10 times each for different rates of genotype missingness: 1%, 2%, 5%, and 10%. For each rate, we took the mean score for each individual across all 10 runs as their new score. These new scores replaced the individuals' original scores, and were renormalized with the rest of the cohort's original scores.

M1.8 Covariates used in regression analyses

We see in our regression analyses that the p-values of association between the PCs and CAD increase further with the addition of variables such as pack years of smoking (field 20161), measures of alcohol consumption, and exercise. However, the inclusion of these variables means the exclusion of increasing numbers of individuals, who are not evenly distributed throughout the dataset with respect to all the relevant variables. For instance, the individuals for whom there is no data on pack years of smoking are disproportionately from the "previous smoker" category. They are also older on average by almost a full year (0.9740) than the group for which these data are available. The variables included in the models in Table S2 were chosen specifically to retain as much of the data as possible and checked to ensure that biases (especially in the distribution of the PCs) were not introduced due to missing data.

It is for this reason that we approximate socioeconomic status using Townsend depri-

vation index, income, and age when completed full time education. The UK Biobank contains a number of potential measures, including indices of multiple deprivation that were calculated within each of England, Scotland, and Wales (fields 26410, 26427 and 26426, respectively), which were tempting to use, but whose inclusion altered the PC distributions of the remaining sample relative to the original. We report the full list of covariates in Table S7.

Data availability

All data used in this paper are available through the UK Biobank and CARTaGENE. Scripts for analysis can be made available upon request.

Acknowledgments

We would like to thank Dr Na Cai, Dr Marie-Julie Favé, Dr Ryan Christ, Dr Claude Bhérer and Jean-Christophe Grenier for their comments on our work. RT and JH are Fonds de la Recherche du Québec en Santé (FRQS) Junior 1 Scholars. This project was funded by grants from the Molson Foundation, the Montreal Heart Institute Foundation and the Institute for Data Valorization (IVADO) to JH.

Authors contributions

HT and JH designed this study. HT performed all analyses. JP contributed background research on the PRS used, and RT provided a clinical perspective for the improvement of these scores. All authors wrote the paper together and approved its final version.

Ethics Declaration

The data used in this study were approved by the UK Biobank under project 49731 and by CARTaGENE under project number 406713. Participants in these cohorts gave their general consent for their data to be used for research purposes. All individual-level data was anonymized and no efforts were made by the authors to deanonymize or recontact any of the participants from either cohort, in keeping with our agreements with the UK Biobank and CARTaGENE.

References

[1] Wray, N.R., Lin, T., Austin, J., McGrath, J.J., Hickie, I.B., Murray, G.K., Visscher, P.M.: From Basic Science to Clinical Application of Polygenic Risk Scores: A Primer. JAMA Psychiatry **78**(1), 101 (2021). doi:10.1001/jamapsychiatry.2020.3049. Accessed 2021-05-05

[2] Khera, A.V., Chaffin, M., Aragam, K.G., Haas, M.E., Roselli, C., Choi, S.H., Natarajan, P., Lander, E.S., Lubitz, S.A., Ellinor, P.T., Kathiresan, S.: Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations. Nature Genetics **50**(9), 1219–1224 (2018). doi:10.1038/s41588-018-0183-z. Number: 9

[3] Inouye, M., Abraham, G., Nelson, C.P., Wood, A.M., Sweeting, M.J., Dudbridge, F., Lai, F.Y., Kaptoge, S., Brozynska, M., Wang, T., Ye, S., Webb, T.R., Rutter, M.K., Tzoulaki, I., Patel, R.S., Loos, R.J.F., Keavney, B., Hemingway, H., Thompson, J., Watkins, H., Deloukas, P., Di Angelantonio, E., Butterworth, A.S., Danesh, J., Samani, N.J.: Genomic Risk Prediction of Coronary Artery Disease in 480,000 Adults. Journal of the American College of Cardiology **72**(16), 1883–1893 (2018). doi:10.1016/j.jacc.2018.07.079. Number: 16. Accessed 2018-10-31

[4] Torkamani, A., Wineinger, N.E., Topol, E.J.: The personal and clinical utility of polygenic risk scores. Nature Reviews Genetics **19**(9), 581–590 (2018). doi:10.1038/s41576-018-0018-x. Number: 9

[5] Wand, H., Lambert, S.A., Tamburro, C., Iacocca, M.A., O'Sullivan, J.W., Sillari, C., Kullo, I.J., Rowley, R., Dron, J.S., Brockman, D., Venner, E., McCarthy, M.I., Antoniou, A.C., Easton, D.F., Hegele, R.A., Khera, A.V., Chatterjee, N., Kooperberg, C., Edwards, K., Vlessis, K., Kinnear, K., Danesh, J.N., Parkinson, H., Ramos, E.M., Roberts, M.C., Ormond, K.E., Khoury, M.J., Janssens, A.C.J.W., Goddard,

K.A.B., Kraft, P., MacArthur, J.A.L., Inouye, M., Wojcik, G.L.: Improving reporting standards for polygenic scores in risk prediction studies. Nature **591**(7849), 211–219 (2021). doi:10.1038/s41586-021-03243-6. Accessed 2021-03-13

[6] Haworth, S., Mitchell, R., Corbin, L., Wade, K.H., Dudding, T., Budu-Aggrey, A., Carslake, D., Hemani, G., Paternoster, L., Smith, G.D., Davies, N., Lawson, D.J., J. Timpson, N.: Apparent latent structure within the UK Biobank sample has implications for epidemiological analysis. Nature Communications **10**(1), 333 (2019). doi:10.1038/s41467-018-08219-1. Number: 1. Accessed 2020-09-21

[7] Zaidi, A.A., Mathieson, I.: Demographic history mediates the effect of stratification on polygenic scores. eLife **9**, 61548 (2020). doi:10.7554/eLife.61548. Accessed 2021-01-22

[8] Anderson, C.A., Pettersson, F.H., Clarke, G.M., Cardon, L.R., Morris, A.P., Zondervan, K.T.: Data quality control in genetic case-control association studies. Nat Protoc **5**(9), 1564–73 (2010). doi:10.1038/nprot.2010.116. Edition: 2010/11/19

[9] Elliott, J., Bodinier, B., Bond, T.A., Chadeau-Hyam, M., Evangelou, E., Moons, K.G.M., Dehghan, A., Muller, D.C., Elliott, P., Tzoulaki, I.: Predictive Accuracy of a Polygenic Risk Score–Enhanced Prediction Model vs a Clinical Risk Score for Coronary Artery Disease. JAMA **323**(7), 636 (2020). doi:10.1001/jama.2019.22241. Number: 7. Accessed 2020-10-05

[10] Vilhjálmsson, B.J., Yang, J., Finucane, H.K., Gusev, A., Lindström, S., Ripke, S., Genovese, G., Loh, P.-R., Bhatia, G., Do, R., Hayeck, T., Won, H.-H., Kathiresan, S., Pato, M., Pato, C., Tamimi, R., Stahl, E., Zaitlen, N., Pasaniuc, B., Belbin, G., Kenny, E.E., Schierup, M.H., De Jager, P., Patsopoulos, N.A., McCarroll, S., Daly, M., Purcell, S., Chasman, D., Neale, B., Goddard, M., Visscher, P.M., Kraft, P., Patterson, N., Price, A.L., Ripke, S., Neale, B.M., Corvin, A., Walters, J.T.R., Farh, K.-

H., Holmans, P.A., Lee, P., Bulik-Sullivan, B., Collier, D.A., Huang, H., Pers, T.H., Agartz, I., Agerbo, E., Albus, M., Alexander, M., Amin, F., Bacanu, S.A., Begemann, M., Belliveau, R.A., Bene, J., Bergen, S.E., Bevilacqua, E., Bigdeli, T.B., Black, D.W., Bruggeman, R., Buccola, N.G., Buckner, R.L., Byerley, W., Cahn, W., Cai, G., Campion, D., Cantor, R.M., Carr, V.J., Carrera, N., Catts, S.V., Chambert, K.D., Chan, R.C.K., Chen, R.Y.L., Chen, E.Y.H., Cheng, W., Cheung, E.F.C., Chong, S.A., Cloninger, C.R., Cohen, D., Cohen, N., Cormican, P., Craddock, N., Crowley, J.J., Curtis, D., Davidson, M., Davis, K.L., Degenhardt, F., Del Favero, J., DeLisi, L.E., Demontis, D., Dikeos, D., Dinan, T., Djurovic, S., Donohoe, G., Drapeau, E., Duan, J., Dudbridge, F., Durmishi, N., Eichhammer, P., Eriksson, J., Escott-Price, V., Essioux, L., Fanous, A.H., Farrell, M.S., Frank, J., Franke, L., Freedman, R., Freimer, N.B., Friedl, M., Friedman, J.I., Fromer, M., Genovese, G., Georgieva, L., Gershon, E.S., Giegling, I., Giusti-Rodrguez, P., Godard, S., Goldstein, J.I., Golimbet, V., Gopal, S., Gratten, J., Grove, J., de Haan, L., Hammer, C., Hamshere, M.L., Hansen, M., Hansen, T., Haroutunian, V., Hartmann, A.M., Henskens, F.A., Herms, S., Hirschhorn, J.N., Hoffmann, P., Hofman, A., Hollegaard, M.V., Hougaard, D.M., Ikeda, M., Joa, I., Julia, A., Kahn, R.S., Kalaydjieva, L., Karachanak-Yankova, S., Karjalainen, J., Kavanagh, D., Keller, M.C., Kelly, B.J., Kennedy, J.L., Khrunin, A., Kim, Y., Klovins, J., Knowles, J.A., Konte, B., Kucinskas, V., Kucinskiene, Z.A., Kuzelova-Ptackova, H., Kahler, A.K., Laurent, C., Keong, J.L.C., Lee, S.H., Legge, S.E., Lerer, B., Li, M., Li, T., Liang, K.-Y., Lieberman, J., Limborska, S., Loughland, C.M., Lubinski, J., Lnnqvist, J., Macek, M., Magnusson, P.K.E., Maher, B.S., Maier, W., Mallet, J., Marsal, S., Mattheisen, M., Mattingsdal, M., McCarley, R.W., McDonald, C., McIntosh, A.M., Meier, S., Meijer, C.J., Melegh, B., Melle, I., Mesholam-Gately, R.I., Metspalu, A., Michie, P.T., Milani, L., Milanova, V., Mokrab, Y., Morris, D.W., Mors, O., Mortensen, P.B., Murphy, K.C., Murray, R.M., Myin-Germeys, I., Mller-Myhsok, B., Nelis, M., Nenadic, I., Nertney, D.A., Nestadt, G.,

Nicodemus, K.K., Nikitina-Zake, L., Nisenbaum, L., Nordin, A., O'Callaghan, E., O'Dushlaine, C., O'Neill, F.A., Oh, S.-Y., Olincy, A., Olsen, L., Van Os, J., Pantelis, C., Papadimitriou, G.N., Papiol, S., Parkhomenko, E., Pato, M.T., Paunio, T., Pejovic-Milovancevic, M., Perkins, D.O., Pietilinen, O., Pimm, J., Pocklington, A.J., Powell, J., Price, A., Pulver, A.E., Purcell, S.M., Quested, D., Rasmussen, H.B., Reichenberg, A., Reimers, M.A., Richards, A.L., Roffman, J.L., Roussos, P., Ruderfer, D.M., Salomaa, V., Sanders, A.R., Schall, U., Schubert, C.R., Schulze, T.G., Schwab, S.G., Scolnick, E.M., Scott, R.J., Seidman, L.J., Shi, J., Sigurdsson, E., Silagadze, T., Silverman, J.M., Sim, K., Slominsky, P., Smoller, J.W., So, H.-C., Spencer, C.C.A., Stahl, E.A., Stefansson, H., Steinberg, S., Stogmann, E., Straub, R.E., Strengman, E., Strohmaier, J., Stroup, T.S., Subramaniam, M., Suvisaari, J., Svrakic, D.M., Szatkiewicz, J.P., Sderman, E., Thirumalai, S., Toncheva, D., Tooney, P.A., Tosato, S., Veijola, J., Waddington, J., Walsh, D., Wang, D., Wang, Q., Webb, B.T., Weiser, M., Wildenauer, D.B., Williams, N.M., Williams, S., Witt, S.H., Wolen, A.R., Wong, E.H.M., Wormley, B.K., Wu, J.Q., Xi, H.S., Zai, C.C., Zheng, X., Zimprich, F., Wray, N.R., Stefansson, K., Visscher, P.M., Adolfsson, R., Andreassen, O.A., Blackwood, D.H.R., Bramon, E., Buxbaum, J.D., Børglum, A.D., Cichon, S., Darvasi, A., Domenici, E., Ehrenreich, H., Esko, T., Gejman, P.V., Gill, M., Gurling, H., Hultman, C.M., Iwata, N., Jablensky, A.V., Jonsson, E.G., Kendler, K.S., Kirov, G., Knight, J., Lencz, T., Levinson, D.F., Li, Q.S., Liu, J., Malhotra, A.K., McCarroll, S.A., McQuillin, A., Moran, J.L., Mortensen, P.B., Mowry, B.J., Nthen, M.M., Ophoff, R.A., Owen, M.J., Palotie, A., Pato, C.N., Petryshen, T.L., Posthuma, D., Rietschel, M., Riley, B.P., Rujescu, D., Sham, P.C., Sklar, P., St. Clair, D., Weinberger, D.R., Wendland, J.R., Werge, T., Daly, M.J., Sullivan, P.F., O'Donovan, M.C., Kraft, P., Hunter, D.J., Adank, M., Ahsan, H., Aittomäki, K., Baglietto, L., Berndt, S., Blomquist, C., Canzian, F., Chang-Claude, J., Chanock, S.J., Crisponi, L., Czene, K., Dahmen, N., Silva, I.d.S., Easton, D., Eliassen, A.H., Figueroa, J., Fletcher, O., Garcia-Closas,

M., Gaudet, M.M., Gibson, L., Haiman, C.A., Hall, P., Hazra, A., Hein, R., Henderson, B.E., Hofman, A., Hopper, J.L., Irwanto, A., Johansson, M., Kaaks, R., Kibriya, M.G., Lichtner, P., Lindström, S., Liu, J., Lund, E., Makalic, E., Meindl, A., Meijers-Heijboer, H., Müller-Myhsok, B., Muranen, T.A., Nevanlinna, H., Peeters, P.H., Peto, J., Prentice, R.L., Rahman, N., Sánchez, M.J., Schmidt, D.F., Schmutzler, R.K., Southey, M.C., Tamimi, R., Travis, R., Turnbull, C., Uitterlinden, A.G., van der Luijt, R.B., Waisfisz, Q., Wang, Z., Whittemore, A.S., Yang, R., Zheng, W.: Modeling Linkage Disequilibrium Increases Accuracy of Polygenic Risk Scores. The American Journal of Human Genetics **97**(4), 576–592 (2015). doi:10.1016/j.ajhg.2015.09.001. Number: 4. Accessed 2020-05-18

[11] Mak, T.S.H., Porsch, R.M., Choi, S.W., Zhou, X., Sham, P.C.: Polygenic scores via penalized regression on summary statistics: MAK et al. Genetic Epidemiology **41**(6), 469–480 (2017). doi:10.1002/gepi.22050. Accessed 2021-02-22

[12] Bycroft, C., Freeman, C., Petkova, D., Band, G., Elliott, L.T., Sharp, K., Motyer, A., Vukcevic, D., Delaneau, O., O'Connell, J., Cortes, A., Welsh, S., Young, A., Effingham, M., McVean, G., Leslie, S., Allen, N., Donnelly, P., Marchini, J.: The UK Biobank resource with deep phenotyping and genomic data. Nature **562**(7726), 203–209 (2018). doi:10.1038/s41586-018-0579-z. Number: 7726

[13] Provine, W.B.: The role of mathematical population geneticists in the evolutionary synthesis of the 1930s and 1940s. Studies in History of Biology **2**, 167–192 (1978)

[14] Price, A.L., Patterson, N.J., Plenge, R.M., Weinblatt, M.E., Shadick, N.A., Reich, D.: Principal components analysis corrects for stratification in genome-wide association studies. Nature Genetics **38**(8), 904–909 (2006). doi:10.1038/ng1847. Number: 8. Accessed 2020-12-11

[15] McVean, G.A.T.: A genealogical interpretation of linkage disequilibrium. Genetics **162**(2), 987–991 (2002). Number: 2

[16] Freedman, M.L., Reich, D., Penney, K.L., McDonald, G.J., Mignault, A.A., Patterson, N., Gabriel, S.B., Topol, E.J., Smoller, J.W., Pato, C.N., Pato, M.T., Petryshen, T.L., Kolonel, L.N., Lander, E.S., Sklar, P., Henderson, B., Hirschhorn, J.N., Altshuler, D.: Assessing the impact of population stratification on genetic association studies. Nature Genetics **36**(4), 388–393 (2004). doi:10.1038/ng1333. Number: 4. Accessed 2020-12-11

[17] Marchini, J., Cardon, L.R., Phillips, M.S., Donnelly, P.: The effects of human population structure on large genetic association studies. Nature Genetics **36**(5), 512–517 (2004). doi:10.1038/ng1337. Number: 5

[18] Hirschhorn, J.N., Daly, M.J.: Genome-wide association studies for common diseases and complex traits. Nature Reviews Genetics **6**(2), 95–108 (2005). doi:10.1038/nrg1521. Number: 2

[19] Bhattacharjee, S., Wang, Z., Ciampa, J., Kraft, P., Chanock, S., Yu, K., Chatterjee, N.: Using Principal Components of Genetic Variation for Robust and Powerful Detection of Gene-Gene Interactions in Case-Control and Case-Only Studies. The American Journal of Human Genetics **86**(3), 331–342 (2010). doi:10.1016/j.ajhg.2010.01.026. Number: 3. Accessed 2020-12-11

[20] Reich, D., Price, A.L., Patterson, N.: Principal component analysis of genetic data. Nature Genetics **40**(5), 491–492 (2008). Number: 5. Accessed 2014-07-18

[21] Menozzi, P., Piazza, A., Cavalli-Sforza, L.: Synthetic maps of human gene frequencies in Europeans. Science **201**(4358), 786–792 (1978). doi:10.1126/science.356262. Number: 4358. Accessed 2020-12-11

[22] McVean, G.: A Genealogical Interpretation of Principal Components Analysis. PLoS Genetics **5**(10), 1000686 (2009). doi:10.1371/journal.pgen.1000686. Number: 10. Accessed 2020-08-09

[23] Martin, A.R., Gignoux, C.R., Walters, R.K., Wojcik, G.L., Neale, B.M., Gravel, S., Daly, M.J., Bustamante, C.D., Kenny, E.E.: Human Demographic History Impacts Genetic Risk Prediction across Diverse Populations. The American Journal of Human Genetics **100**(4), 635–649 (2017). doi:10.1016/j.ajhg.2017.03.004. Number: 4. Accessed 2020-03-18

[24] Martin, A.R., Kanai, M., Kamatani, Y., Okada, Y., Neale, B.M., Daly, M.J.: Clinical use of current polygenic risk scores may exacerbate health disparities. Nature Genetics **51**(4), 584–591 (2019). doi:10.1038/s41588-019-0379-x. Accessed 2021-02-09

[25] Kerminen, S., Martin, A.R., Koskela, J., Ruotsalainen, S.E., Havulinna, A.S., Surakka, I., Palotie, A., Perola, M., Salomaa, V., Daly, M.J., Ripatti, S., Pirinen, M.: 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 (2019). doi:10.1016/j.ajhg.2019.05.001. Number: 6. Accessed 2019-07-05

[26] Awadalla, P., Boileau, C., Payette, Y., Idaghdour, Y., Goulet, J.-P., Knoppers, B., Hamet, P., Laberge, C.: Cohort profile of the CARTaGENE study: Quebec's population-based biobank for public health and personalized genomics. International Journal of Epidemiology **42**(5), 1285–1299 (2013). doi:10.1093/ije/dys160. Number: 5. Accessed 2018-11-20

[27] Laprise, C.: The Saguenay-Lac-Saint-Jean asthma familial collection: the genetics of asthma in a young founder population. Genes & Immunity **15**(4), 247–255 (2014). doi:10.1038/gene.2014.12. Accessed 2021-02-21

[28] Wünnemann, F., Sin Lo, K., Langford-Avelar, A., Busseuil, D., Dubé, M.-P., Tardif, J.-C., Lettre, G.: Validation of Genome-Wide Polygenic Risk Scores for Coronary Artery Disease in French Canadians. Circulation: Genomic and Precision Medicine **12**(6) (2019). doi:10.1161/CIRCGEN.119.002481. Number: 6. Accessed 2019-11-25

[29] Moreau, C., Bherer, C., Vezina, H., Jomphe, M., Labuda, D., Excoffier, L.: Deep Human Genealogies Reveal a Selective Advantage to Be on an Expanding Wave Front. Science **334**(6059), 1148–1150 (2011). doi:10.1126/science.1212880. Accessed 2021-02-23

[30] Morin, A., Madore, A.-M., Kwan, T., Ban, M., Partanen, J., Rönnblom, L., Syvänen, A.-C., Sawcer, S., Stunnenberg, H., Lathrop, M., Pastinen, T., Laprise, C.: Exploring rare and low-frequency variants in the Saguenay–Lac-Saint-Jean population identified genes associated with asthma and allergy traits. European Journal of Human Genetics **27**(1), 90–101 (2019). doi:10.1038/s41431-018-0266-4. Accessed 2021-02-21

[31] the Haplotype Reference Consortium, McCarthy, S., Das, S., Kretzschmar, W., Delaneau, O., Wood, A.R., Teumer, A., Kang, H.M., Fuchsberger, C., Danecek, P., Sharp, K., Luo, Y., Sidore, C., Kwong, A., Timpson, N., Koskinen, S., Vrieze, S., Scott, L.J., Zhang, H., Mahajan, A., Veldink, J., Peters, U., Pato, C., van Duijn, C.M., Gillies, C.E., Gandin, I., Mezzavilla, M., Gilly, A., Cocca, M., Traglia, M., Angius, A., Barrett, J.C., Boomsma, D., Branham, K., Breen, G., Brummett, C.M., Busonero, F., Campbell, H., Chan, A., Chen, S., Chew, E., Collins, F.S., Corbin, L.J., Smith, G.D., Dedoussis, G., Dorr, M., Farmaki, A.-E., Ferrucci, L., Forer, L., Fraser, R.M., Gabriel, S., Levy, S., Groop, L., Harrison, T., Hattersley, A., Holmen, O.L., Hveem, K., Kretzler, M., Lee, J.C., McGue, M., Meitinger, T., Melzer, D., Min, J.L., Mohlke, K.L., Vincent, J.B., Nauck, M., Nickerson, D., Palotie, A., Pato, M., Pirastu, N., McInnis, M., Richards, J.B., Sala, C., Salomaa, V., Schlessinger, D., Schoenherr, S., Slagboom, P.E., Small, K., Spector, T., Stambolian, D., Tuke, M., Tuomilehto, J.,

Van den Berg, L.H., Van Rheenen, W., Volker, U., Wijmenga, C., Toniolo, D., Zeggini, E., Gasparini, P., Sampson, M.G., Wilson, J.F., Frayling, T., de Bakker, P.I.W., Swertz, M.A., McCarroll, S., Kooperberg, C., Dekker, A., Altshuler, D., Willer, C., Iacono, W., Ripatti, S., Soranzo, N., Walter, K., Swaroop, A., Cucca, F., Anderson, C.A., Myers, R.M., Boehnke, M., McCarthy, M.I., Durbin, R., Abecasis, G., Marchini, J.: A reference panel of 64,976 haplotypes for genotype imputation. Nature Genetics **48**, 1279 (2016)

[32] Mostafavi, H., Harpak, A., Agarwal, I., Conley, D., Pritchard, J.K., Przeworski, M.: Variable prediction accuracy of polygenic scores within an ancestry group. eLife **9**, 48376 (2020). doi:10.7554/eLife.48376

[33] Simcoe, M., Valdes, A., Liu, F., Furlotte, N.A., Evans, D.M., Hemani, G., Ring, S.M., Smith, G.D., Duffy, D.L., Zhu, G., Gordon, S.D., Medland, S.E., Vuckovic, D., Girotto, G., Sala, C., Catamo, E., Concas, M.P., Brumat, M., Gasparini, P., Toniolo, D., Cocca, M., Robino, A., Yazar, S., Hewitt, A., Wu, W., Kraft, P., Hammond, C.J., Shi, Y., Chen, Y., Zeng, C., Klaver, C.C.W., Uitterlinden, A.G., Ikram, M.A., Hamer, M.A., van Duijn, C.M., Nijsten, T., Han, J., Mackey, D.A., Martin, N.G., Cheng, C.-Y., the 23andMe Research Team, the International Visible Trait Genetics Consortium, Hinds, D.A., Spector, T.D., Kayser, M., Hysi, P.G.: Genome-wide association study in almost 195,000 individuals identifies 50 previously unidentified genetic loci for eye color. Science Advances **7**(11), 1239 (2021). doi:10.1126/sciadv.abd1239. Accessed 2021-03-11

[34] Janssens, A.C.J.W.: Validity of polygenic risk scores: are we measuring what we think we are? Human Molecular Genetics **28**(R2), 143–150 (2019). doi:10.1093/hmg/ddz205. Accessed 2021-03-13

[35] R Core Team: R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria (2020). https://www.R-project.org

[36] Hui, A., Latif, A., Hinsliff-Smith, K., Chen, T.: 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 (2020). doi:10.1016/j.healthpol.2020.01.003. Number: 3. Accessed 2020-07-07

[37] Abraham, G., Qiu, Y., Inouye, M.: FlashPCA2: principal component analysis of Biobank-scale genotype datasets. Bioinformatics **33**(17), 2776–2778 (2017). doi:10.1093/bioinformatics/btx299. Number: 17. Accessed 2020-05-18

[38] Chang, C.C., Chow, C.C., Tellier, L.C., Vattikuti, S., Purcell, S.M., Lee, J.J.: Second-generation PLINK: rising to the challenge of larger and richer datasets. GigaScience **4**(1), 7 (2015). doi:10.1186/s13742-015-0047-8. Number: 1. Accessed 2020-07-31

[39] Purcell, S., Chang, C.: Plink 1.9b_5.2. Programmers: _:n17703 (2019). `www.cog-genomics.org/plink/1.9/`

[40] Manichaikul, A., Mychaleckyj, J.C., Rich, S.S., Daly, K., Sale, M., Chen, W.-M.: Robust relationship inference in genome-wide association studies. Bioinformatics **26**(22), 2867–2873 (2010). doi:10.1093/bioinformatics/btq559. Accessed 2021-02-16

[41] Khera, A.V., Chaffin, M., Wade, K.H., Zahid, S., Brancale, J., Xia, R., Distefano, M., Senol-Cosar, O., Haas, M.E., Bick, A., Aragam, K.G., Lander, E.S., Smith, G.D., Mason-Suares, H., Fornage, M., Lebo, M., Timpson, N.J., Kaplan, L.M., Kathiresan, S.: Polygenic Prediction of Weight and Obesity Trajectories from Birth to Adulthood. Cell **177**(3), 587–5969 (2019). doi:10.1016/j.cell.2019.03.028. Number: 3. Accessed 2020-05-12

[42] Lambert, S.A., Gil, L., Jupp, S., Chapman, M., Parkinson, H., Danesh, J., MacArthur, J.A.L., Inouye, M.: The Polygenic Score (PGS) Catalog: an open database to enable reproducibility and systematic evaluation. `www.pgscatalog.org`

Conflicts of interest

The authors have no conflicts of interest to declare.

Supplementary Information

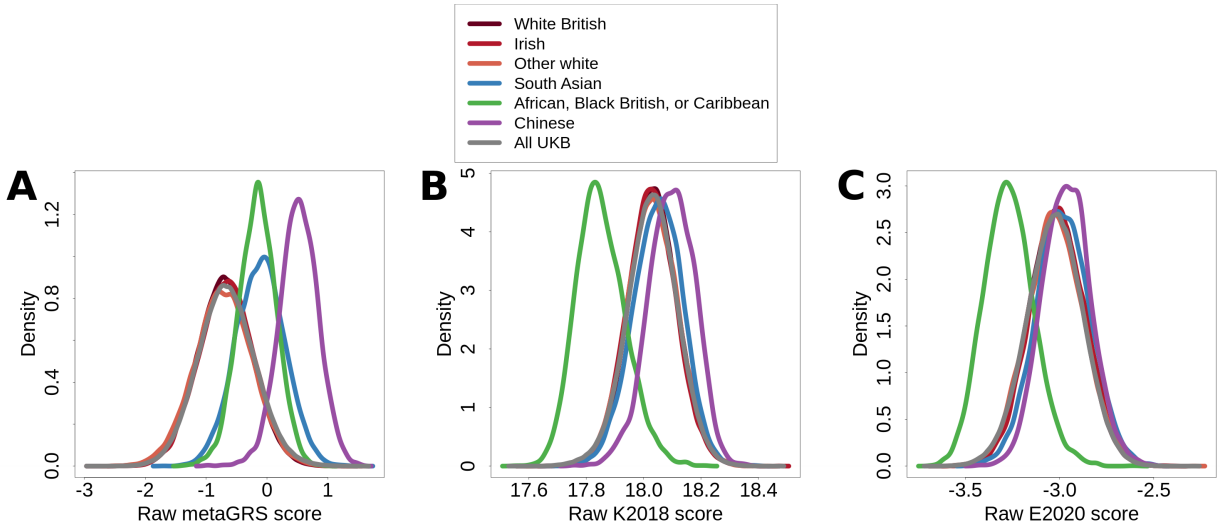

Figure S1: Distributions of raw risk scores for the three risk scores: **(A)** metaGRS, **(B)** K2018 score, and **(C)** E2020 score for different subsets of the UK Biobank (UKB).

Table S1: Effect size estimates and p-values of association of each risk score with CAD, calculated for each subpopulation. Scores were normalized within each subpopulation and included with age, sex, and the first 10 principal components provided by the UK Biobank.

Population	**Size**	**metaGRS $\hat{\beta}$ (p)**	**K2018 $\hat{\beta}$ (p)**	**E2020 $\hat{\beta}$ (p)**
white British	408,572	0.527 ($< 10^{-100}$)	0.514 ($< 10^{-100}$)	0.404 ($< 10^{-100}$)
Irish	12,669	0.544 (1.1×10^{-42})	0.515 (2.5×10^{-39})	0.430 (3.5×10^{-28})
other white	16,210	0.555 (2.3×10^{-36})	0.481 (1.5×10^{-30})	0.310 (2.4×10^{-14})
South Asian	7,628	0.464 (5.0×10^{-27})	0.405 (1.2×10^{-21})	0.394 (7.4×10^{-21})
Black, Afr., or Carib.	7,648	0.230 (0.001)	0.187 (0.011)	0.084 (0.250)
Chinese	1,503	0.319 (0.070)	0.350 (0.047)	0.313 (0.058)

Table S2: Strength of association between PCs and CAD in the white British subset of the UK Biobank. We provide the p-values and the Akaike information criterion (AIC) for each model of each PC in a logistic regression of the score (normalized within the white British subset) on the first three PCs calculated within the 334,181 white British individuals for whom all the relevant covariates were available, sequentially adding environmental covariates to the model—that is, the model described in each row includes the covariates listed in the preceding rows as well. SES variables were the Townsend deprivation index, income, and age when completed full time education.

Model	**PC1 p-val**	**PC2 p-val**	**PC3 p-val**	**AIC**
PCs	2.2×10^{-10}	2.1×10^{-10}	4.5×10^{-11}	144,801.2
" + sex + age	3.6×10^{-25}	3.2×10^{-10}	1.7×10^{-9}	130,894.4
" + smoking status	7.9×10^{-21}	3.8×10^{-10}	3.5×10^{-8}	129,636.0
" + SES variables	1.6×10^{-12}	2.0×10^{-5}	0.0009	128,001.1
" + home location (north and east)	0.0001	0.1340	0.0085	127,946.5
" + place of birth (north and east)	0.0827	0.6827	0.0562	127,917.5

Table S3: Strength of association between PCs and scores as well as the association between the PCs and CAD in the white British subset of the UK Biobank. We provide here the p-values of each PC in a linear regression of the score (normalized within the white British subset) on the first three PCs calculated within the white British subset, age (defined using fields 32 and 34), sex (field 31), place of birth coordinates (fields 129 and 130), home location coordinates (fields 20074 and 20075), smoking status (field 20116), Townsend deprivation index at recruitment (field 189), income (field 738 converted into numeric values), and age when completed full time education (field 845), supplemented by qualifications (field 6138), where people who indicated they had a university or college degree were assumed to have completed their degree at 22). The results for CAD were from a logistic regression that used the same set of covariates.

Outcome	**PC1 p-val**	**PC2 p-val**	**PC3 p-val**
metaGRS	0.0555	4.6×10^{-42}	3.8×10^{-32}
K2018	1.1×10^{-11}	0.0141	0.0030
E2020	7.1×10^{-6}	1.5×10^{-7}	0.0344
CAD	0.0741	0.6689	0.0594

Table S4: Strength of association between PCs and scores in the white British subset of the UK Biobank. We provide here the p-values of each PC in a linear regression of the score (normalized within the white British subset) on the first three PCs calculated within the white British subset.

Score	**PC1 p-val**	**PC2 p-val**	**PC3 p-val**
metaGRS	2.0×10^{-20}	3.8×10^{-117}	2.6×10^{-61}
K2018	1.5×10^{-11}	3.0×10^{-10}	9.5×10^{-6}
E2020	5.5×10^{-17}	1.5×10^{-18}	2.9×10^{-5}

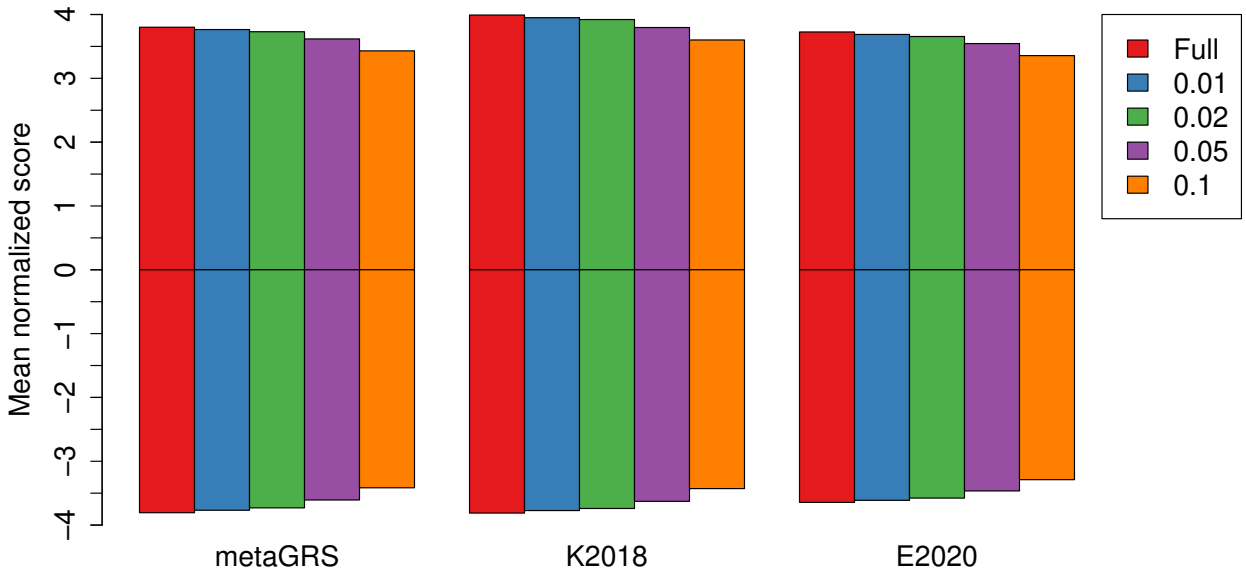

Figure S2: Mean normalized PRS for different rates of random missingness. The positive barplot shows the mean of the individuals who were identified in the original "full" score as having the highest 100 scores, and the negative shows the mean for those with the lowest 100 scores. The red bars show the values using the original "full" scores. Subsequent bars show the values of the scores after setting 1%, 2%, 5%, and 10% of the genotypes of these individuals at markers used in each score to missing.

Table S5: Proportion of overlap among the white British individuals identified by each score as being in the top 10% of CAD risk (upper triangle) to the bottom 10% (lower triangle).

	metaGRS	K2018	E2020
metaGRS	–	0.5530	0.3894
K2018	0.5381	–	0.4520
E2020	0.3906	0.4597	–

Table S6: Number of people who are assigned by the score in the left-most column a level of risk (highest or lowest decile) who were placed in the opposite extreme risk category by the other two scores. For example, the first line tells us that there were 17 people who were in the highest decile of metaGRS scores who were also in the lowest decile of K2018 scores. Similarly, there were 105 people who were in the highest decile of metaGRS scores who were in the lowest decile of E2020 scores. The next line tells us the inverse: how many people were in the lowest decile of risk for the metaGRS who were also in the highest decile of risk for the other two scores. For reference, each decile contains 40,858 individuals.

Score	**Decile**	**K2018**	**E2020**
metaGRS	Highest	17	105
	Lowest	0	90
K2018	Highest	–	37
	Lowest	–	18

Table S7: Description of variables used in our regression analyses.

Variable	Description
Age	Participants' ages (in years) at analysis. This is calculated from the month of birth provided by field 52 and year of birth from field 34. The day was set as the 15th of the month. The number of days between this date and 2020-08-31 (August being our most recent download of the UK Biobank data) is calculated and then divided by 365.25. In the case of deceased individuals, the age at death was used.
Sex	From field 31.
Smoking status	From field 20161. Possible statuses were "Never", "Previous", and "Current".
Townsend deprivation index	From field 189.
Income	From field 738 (average total household income before tax). Participants could choose among: "Less than 18,000", "18,000 to 30,999", "31,000 to 51,999", "52,000 to 100,000", "Greater than 100,000", "Do not know", "Prefer not to say". We set the last two responses to NA and converted the remaining ones to levels 1-6, where 1 was the lowest salary band and 6 was the highest.
Age completed full time education	Participants were asked to indicate their educational qualifications in Field 6138. Those who did not indicate a university degree were asked when they stopped their full time education in Field 845. We used Field 845 and, those who indicated they did have a university degree were assigned (by us) an age of 22 for this variable, in order to keep them in the analysis.
Place of birth in the UK	Fields 129 (north coordinate) and 130 (east coordinate). Coordinates are given in metre-grid map units.
Home location	Fields 20074 (east coordinate) and 20075 (north coordinate). Coordinates are given in metre-grid map units.

Table S8: *P*-values of Kolmogorov-Smirnov tests of the pairwise equivalence of the distributions of the meta-GRS scores calculated for white British individuals who attended one of the two indicated assessment centers. Low p-values indicate a rejection of the null hypothesis that the metaGRS score distributions are the same for the two assessment centers.

	Cardiff	Newcastle	Glasgow
Reading	0.9590	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$
Cardiff	–	1.55×10^{-15}	1.78×10^{-15}
Newcastle	–	–	0.2426

Table S9: *P*-values of Kolmogorov-Smirnov tests of the pairwise equivalence of the distributions of the K2018 scores calculated for white British individuals who attended one of the two indicated assessment centers. Low p-values indicate a rejection of the null hypothesis that the K2018 score distributions are the same for the two assessment centers.

	Cardiff	Newcastle	Glasgow
Reading	0.5549	0.0004	0.9053
Cardiff	–	0.0075	0.9708
Newcastle	–	–	0.0126

Table S10: *P*-values of Kolmogorov-Smirnov tests of the pairwise equivalence of the distributions of the E2020 scores calculated for white British individuals who attended one of the two indicated assessment centers. Low p-values indicate a rejection of the null hypothesis that the E2020 score distributions are the same for the two assessment centers.

	Cardiff	Newcastle	Glasgow
Reading	0.0485	7.17×10^{-6}	4.96×10^{-6}
Cardiff	–	0.0639	0.0003
Newcastle	–	–	0.0490

Table S11: Number of markers used in each score (Total), the number of markers used in the score for which we had genotypes in the UK Biobank (UKB), the number of markers used in the score for which we had genotypes in CARTaGENE (CAG), and the number of markers that were genotyped or imputed with high certainty in both UKB and CAG (Overlapping).

Score	**Total**	**UKB**	**CAG**	**Overlapping**
MetaGRS	1,745,179	1,742,121	1,291,113	1,290,327
K2018	6,630,150	6,619,599	5,698,895	5,694,174
E2020	40,079	39,960	33,034	32,994