Clinical use of current polygenic risk scores may exacerbate health disparities

Alicia R. Martin [1,2,3*], Masahiro Kanai [1,2,3,4,5], Yoichiro Kamatani [5,6], Yukinori Okada [5,7,8], Benjamin M. Neale [1,2,3] and Mark J. Daly [1,2,3,9]

Polygenic risk scores (PRS) are poised to improve biomedical outcomes via precision medicine. However, the major ethical and scientific challenge surrounding clinical implementation of PRS is that those available today are several times more accurate in individuals of European ancestry than other ancestries. This disparity is an inescapable consequence of Eurocentric biases in genome-wide association studies, thus highlighting that—unlike clinical biomarkers and prescription drugs, which may individually work better in some populations but do not ubiquitously perform far better in European populations—clinical uses of PRS today would systematically afford greater improvement for European-descent populations. Early diversifying efforts show promise in leveling this vast imbalance, even when non-European sample sizes are considerably smaller than the largest studies to date. To realize the full and equitable potential of PRS, greater diversity must be prioritized in genetic studies, and summary statistics must be publically disseminated to ensure that health disparities are not increased for those individuals already most underserved.

PRS, which predict complex traits on the basis of genetic data, are of burgeoning interest to the clinical community, as researchers demonstrate their growing power to improve clinical care, genetic studies of a wide range of phenotypes increase in size and power, and genotyping costs plummet to less than US$50. Many earlier criticisms of limited predictive power are now recognized to have been chiefly an issue of insufficient sample size, which is no longer the case for many outcomes[1]. For example, PRS alone already predict the risk of breast cancer, prostate cancer and type 1 diabetes in individuals of European descent more accurately than current clinical models[2–4]. Additionally, integrated models of PRS together with other lifestyle and clinical factors have enabled clinicians to more accurately quantify the risk of heart attack for patients; consequently, they have more effectively targeted the decrease in low-density-lipoprotein cholesterol, and by extension heart attack, by prescribing statins to patients at the greatest overall risk of cardiovascular disease[5–9]. Promisingly, the return of genetic risk of complex disease to at-risk patients does not substantially induce self-reported negative behavior or psychological function, and some potentially positive behavioral changes have been detected[10]. Although we share enthusiasm about the potential of PRS to improve health outcomes through their eventual routine implementation as clinical biomarkers, we consider the consistent observation that they currently have far greater predictive value in individuals of recent European descent than of other ancestries to be the major ethical and scientific challenge surrounding clinical translation and, at present, the most critical limitation to genetics in precision medicine. The scientific basis of this imbalance has been demonstrated theoretically, in simulations and empirically across many traits and diseases[11–22].

All studies to date using well-powered genome-wide association studies (GWAS) to assess the predictive value of PRS across a range of traits and populations have made a consistent observation: PRS predict individual risk far more accurately in Europeans than non-Europeans[15,16,18–24]. Rather than being attributable to chance or biology, this consequence is predictable, given that the genetic discovery efforts to date heavily underrepresent non-European populations globally. The correlation between true and genetically predicted phenotypes decays with genetic divergence from the makeup of the discovery GWAS; therefore, the accuracy of polygenic scores in different populations is highly dependent on the representation of the study population in the largest existing 'training' GWAS. Here, we document study biases that underrepresent non-European populations in current GWAS and explain the fundamental concepts contributing to decreased phenotypic variance explained with increasing genetic divergence from populations included in GWAS.

Predictable basis of disparities in PRS accuracy

The poor generalizability of genetic studies across populations arises from the overwhelming abundance of European-descent studies and the dearth of well-powered studies in globally diverse populations[25–28]. According to the GWAS catalog, ~79% of all GWAS participants are of European descent despite making up only 16% of the global population (Fig. 1). This imbalance is especially problematic, because previous studies have shown that studies on Hispanic/Latino individuals and African Americans contribute an outsized number of associations relative to studies of similar sizes in Europeans[27]. More concerningly, the fraction of non-European individuals in GWAS has stagnated or declined since late 2014 (Fig. 1), thus suggesting the absence of a trajectory to correct this

[1]Analytic and Translational Genetics Unit, Massachusetts General Hospital, Boston, MA, USA. [2]Program in Medical and Population Genetics, Broad Institute of Harvard and MIT, Cambridge, MA, USA. [3]Stanley Center for Psychiatric Research, Broad Institute of Harvard and MIT, Cambridge, MA, USA. [4]Department of Biomedical Informatics, Harvard Medical School, Boston, MA, USA. [5]Laboratory for Statistical Analysis, RIKEN Center for Integrative Medical Sciences, Yokohama, Japan. [6]Kyoto–McGill International Collaborative School in Genomic Medicine, Kyoto University Graduate School of Medicine, Kyoto, Japan. [7]Department of Statistical Genetics, Osaka University Graduate School of Medicine, Suita, Japan. [8]Laboratory of Statistical Immunology, Immunology Frontier Research Center (WPI-IFReC), Osaka University, Suita, Japan. [9]Institute for Molecular Medicine Finland (FIMM), University of Helsinki, Helsinki, Finland. *e-mail: armartin@broadinstitute.org
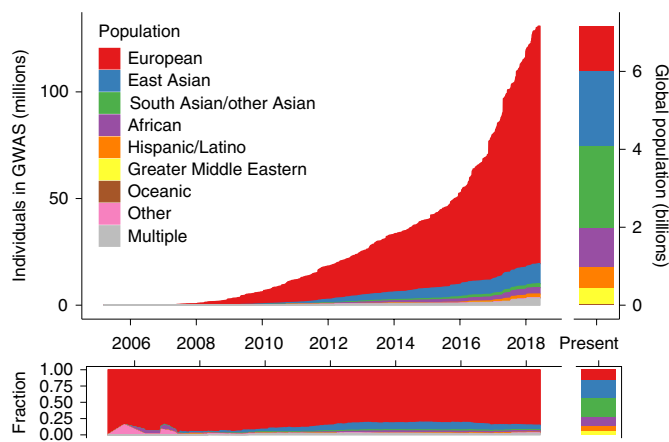
Fig. 1 | Ancestry of GWAS participants over time, as compared with the global population. Cumulative data, as reported by the GWAS catalog[76]. Individuals whose ancestry is 'not reported' are not shown.

imbalance. These numbers provide a composite metric of study availability, accessibility and use—cohorts that have been included in numerous GWAS are represented multiple times, and cohorts of European descent may be disproportionately included. However, whereas the average sample sizes of GWAS in Europeans continue to grow, those for other populations have stagnated and remain several fold smaller (Supplementary Fig. 1).

The relative sample compositions of GWAS result in highly predictable disparities in prediction accuracy; population genetics theory predicts that the accuracy of genetic-risk prediction will decay with increasing genetic divergence between the original GWAS sample and the target of prediction, a function of population history[13,14]. This pattern can be attributed to several statistical observations: (i) GWAS favor the discovery of genetic variants that are common in the study population; (ii) linkage disequilibrium (LD) differentiates marginal effect-size estimates for polygenic traits across populations, even when the causal variants are the same; and (iii) the environment and demography differ across populations. Notably, the first two phenomena substantially degrade prediction performance across populations even in the absence of biological, environmental or diagnostic differences, whereas the environment and demography may together drive differential forces of natural selection that in turn drive differences in causal genetic architecture. (We define the causal genetic architecture as the true effects of variants that affect a phenotype that would be identified in a population of infinite sample size. Unlike effect-size estimates, true effects are typically modeled as invariant with respect to LD and allele frequency differences across populations.)

Common variant discoveries and low-hanging fruit. First, the power to discover an association in a genetic study depends on the effect size and the frequency of the variant[29]. As a result of this dependence, the most significant associations tend to be more common in the populations in which they are discovered than elsewhere[13,30]. For example, GWAS-catalog variants are more common on average in European populations than in East Asian and African populations (Fig. 2b), an observation not representative of genomic variants at large. Understudied populations offer low-hanging fruit for genetic discovery, because variants that are common in these groups but rare or absent in European populations could not be discovered with even very large European sample sizes. Some examples include *SLC16A11* and *HNF1A* associations with type 2 diabetes in Latino populations, as well as *APOL1* associations with end-stage kidney disease and associations with prostate cancer in African-descent

populations[31–34]. If causal genetic variants are assumed to have an equal effect across all populations—an assumption with some empirical support that offers the best-case scenario for transferability[35–40]—Eurocentric GWAS biases would result in risk-associated variants being disproportionately common in European populations, and consequently accounting for a larger fraction of the phenotypic variance therein[13]. Furthermore, imputation reference panels share the same study biases as in GWAS[41], thus creating challenges for imputing sites that are rare in European populations but common elsewhere when the catalog of non-European haplotypes is substantially smaller. These issues are insurmountable through statistical methods alone[13], but they motivate substantial investments in more diverse populations to produce similar-sized GWAS of biomedical phenotypes in other populations.

Linkage disequilibrium. Second, LD, the correlation structure of the genome, varies across populations, owing to demographic history (Fig. 2a,c–e). These LD differences in turn drive differences in effect-size estimates (that is, predictors) from GWAS across populations in proportion to LD between tagging and causal SNP pairs, even when causal effects are the same[35,37–40] (Supplementary Note). Differences in effect-size estimates due to LD differences may typically be small for most regions of the genome (Fig. 2c–e), but PRS sum across these effects, also aggregating these population differences. Although causal effects would ideally be used rather than correlated effect-size estimates to calculate PRS, fine-mapping most variants to a single locus to solve issues of low generalizability may not be feasible, even with very large GWAS. This infeasibility is because complex traits are highly polygenic, and consequently most of the predictive power comes from small effects that do not meet genome-wide significance and/or cannot be fine-mapped, even in many of the best-powered GWAS to date[42].

Complexities of history, selection and the environment. Finally, other cohort considerations may further worsen prediction-accuracy differences across populations in less predictable ways. GWAS ancestry biases and LD differences across populations are extremely challenging to address, but these issues actually make many favorable assumptions that all causal loci have the same effect and are under equivalent selective pressure in all populations. In contrast, other effects on polygenic adaptation or risk scores, such as long-standing environmental differences across global populations that have resulted in differing responses of natural selection, can affect populations differently, depending on their unique histories. Additionally, residual uncorrected population stratification may affect risk-prediction accuracy across populations, but the magnitude of its effect is currently unclear. These effects are particularly challenging to disentangle, as has clearly been demonstrated for height, for which evidence of polygenic adaptation and/or its relative magnitude is under question[43,44]. Comparisons of geographically stratified phenotypes, such as height, across populations with highly divergent genetic backgrounds make environmental differences, such as differences in resource abundance during development across continents, especially prone to confounding from correlated environmental and genetic divergence[43,44]. This residual stratification can lead to over-predicted differences across geographical space[45].

Regarding stratification, most PRS methods do not explicitly address recent admixture, and none consider recently admixed individuals' unique local mosaics of ancestry; thus, further methodological development is needed. Additionally, comparing PRS across environmentally stratified cohorts, such as in some biobanks with healthy-volunteer effects versus disease-study datasets or hospital-based cohorts, requires careful consideration of technical differences, collider bias and variability in baseline health status among studies. Differences in definitions of clinical phenotypes
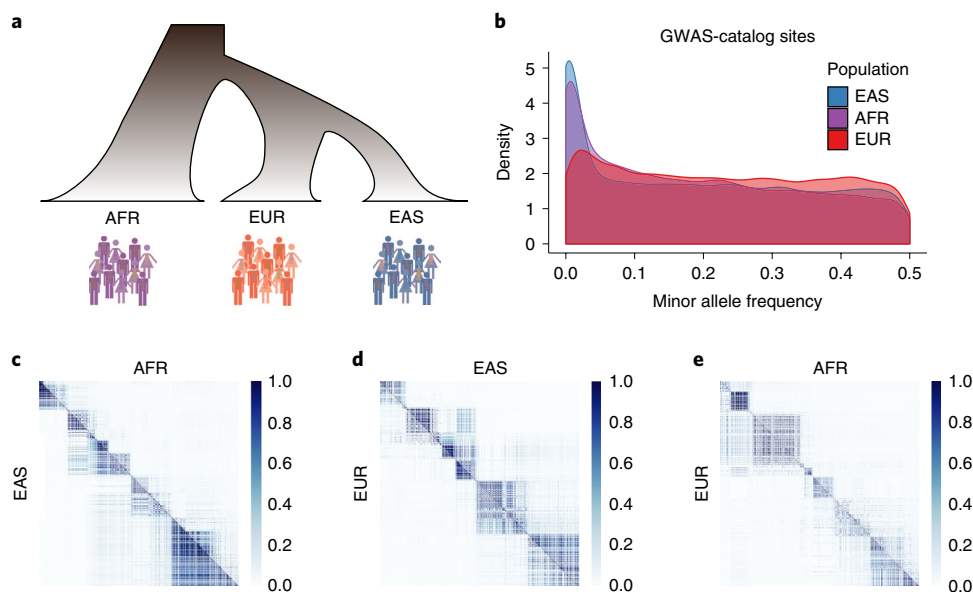
Fig. 2 | Demographic relationships, allele frequency differences and local LD patterns between population pairs. Data analyzed from 1000 Genomes. Population labels: AFR, continental African; EUR, European; EAS, East Asian. **a**, Cartoon relationships among AFR, EUR and EAS populations. **b**, Allele frequency distributions in AFR, EUR and EAS populations of variants from the GWAS catalog. **c–e**, Color axis shows LD scale (r^2) for the indicated LD comparisons between pairs of populations; the same region of the genome for each comparison (representative region is chromosome 1, 51572–52857 kilobases) among pairs of SNPs polymorphic in both populations is shown, illustrating that different SNPs are polymorphic across some population pairs and that these SNPs have variable LD patterns across populations.

and heterogeneity of sub-phenotypes among countries must also be considered.

Differences in environmental exposure, gene–gene interactions, gene–environment interactions, historical population-size dynamics, statistical noise, some potential causal effect differences and/or other factors further limit the generalizability of PRS in an unpredictable, trait-specific fashion[46–49]. Complex traits do not behave in a genetically deterministic manner: some environmental factors dwarf individual genetic effects, thus creating outsized issues of comparability across globally diverse populations. Among psychiatric disorders, for example, schizophrenia has a nearly identical genetic basis across East Asians and Europeans ($r_g = 0.98$) (ref. [40]), whereas the substantially different rates of alcohol-use disorder across populations are partially explained by differences in availability and genetic differences affecting alcohol metabolism[50]. Although nonlinear genetic factors explain little variation in complex traits beyond a purely additive model[51], some unrecognized nonlinearities and gene–gene interactions can also induce challenges to genetic-risk prediction, because pairwise interactions are likely to vary more across populations than individual SNPs. Mathematically, this scenario can simplistically be considered in terms of a two-SNP model, in which the sum of two SNP effects is likely to explain more phenotypic variance than the product of the same SNPs. Some machine-learning approaches may thus modestly improve PRS accuracy beyond current approaches for some phenotypes[52], but improvement is most likely for atypical traits with simpler architectures, known interactions and poor prediction generalizability across populations, such as skin pigmentation[53].

## Limited generalizability of PRS across diverse populations
To date, multi-ancestral work has been slow in most disease areas[54], thus limiting even the opportunity to assess PRS in non-European cohorts. Nonetheless, some previous work has assessed prediction accuracy across diverse populations in several traits and diseases for which GWAS summary statistics are available and has identified

large disparities across populations (Supplementary Note). These disparities are not simply methodological issues, because various approaches (for example, pruning and thresholding versus LDPred) and accuracy metrics (R^2 for quantitative traits and various pseudo-R^2 metrics for binary traits) illustrate this consistently poorer performance in populations distinct from discovery samples across a range of polygenic traits (Supplementary Table 1). These assessments are becoming increasingly feasible with the growth and public availability of global biobanks as well as diversifying priorities from funding agencies[55,56]. We assessed how prediction accuracy decayed across globally diverse populations for 17 anthropometric and blood-panel traits in the UK Biobank (UKBB) when European-derived summary statistics were used (Supplementary Note). In agreement with findings from previous studies, we found that the genetic prediction accuracy was far lower for other populations than for European populations: 1.6-fold lower in Hispanic/Latino Americans, 1.6-fold lower in South Asians, 2.0-fold lower in East Asians and 4.5-fold lower in Africans, on average (Fig. 3).

Prioritizing diversity shows early promise for PRS
Early diversifying GWAS efforts have been especially productive in addressing questions surrounding risk prediction. Rather than varying the prediction target dataset, some GWAS in diverse populations have increased the scale of non-European summary statistics and also varied the study dataset in multi-ancestral PRS studies[23,24,40]. These studies have shown that even when non-European cohorts are only a fraction of the size of the largest European study, they are likely to have disproportionate value for predicting polygenic traits in other individuals of similar ancestry.

Given this background, we performed a systematic evaluation of polygenic prediction accuracy across 17 quantitative anthropometric and blood-panel traits and five disease endpoints in British and Japanese individuals[23,57,58] by performing GWAS with the exact same sample sizes in each population. We symmetrically demonstrate that prediction accuracy is consistently higher with GWAS
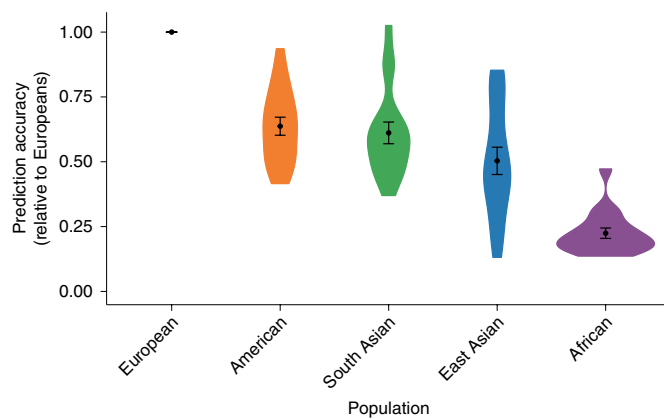
Fig. 3 | Prediction accuracy relative to European-ancestry individuals across 17 quantitative traits and 5 continental populations in the UKBB. All phenotypes shown here are quantitative anthropometric and blood-panel traits, as described in Supplementary Table 6, which includes discovery-cohort sample sizes. Prediction target individuals do not overlap with the discovery cohort and are unrelated; sample sizes are shown in Supplementary Table 7. Violin plots show distributions of relative prediction accuracies, points show mean values, and error bars show s.e.m. values. Prediction R^2 for each trait and population are shown in Supplementary Fig. 12.

summary statistics from ancestry-matched summary statistics (Fig. 4 and Supplementary Figs. 2–6). Keeping in mind the issues of comparability described above, we note that BioBank Japan (BBJ) is a hospital-based disease-ascertained cohort, whereas UKBB is a healthier-than-average[59] population-based cohort; thus, differences in the observed heritability among these cohorts (rather than among populations) due to differences in phenotype precision are likely to explain the lower prediction accuracy from the BBJ GWAS summary statistics for anthropometric and blood-panel traits but the higher prediction accuracy for five ascertained diseases (Supplementary Table 2). Indeed, other East Asian studies have estimated higher heritability for some quantitative traits than BBJ by using the same methods, such as for height ($h^2 = 0.48 \pm 0.04$ in Chinese women[60]). Some statistical fluctuations in the relative differences in prediction accuracy across populations are likely to be driven by differences in heritability measured in each population and/or trans-ancestral genetic correlation (that is, of common variant effect sizes at SNPs common in two populations, Supplementary Figs. 7–10 and Supplementary Tables 2–5). These trans-ancestral correlation estimates indicated that the effect sizes were mostly highly correlated across ancestries, and values for a few traits were somewhat lower than excepted (for example, height and body-mass index, with $\rho_{ge} = 0.69$ and 0.75, respectively). The prediction accuracy was far lower in individuals of African descent in the UKBB (Supplementary Figs. 4 and 11) when GWAS summary statistics from individuals of either European or Japanese ancestry were used, in agreement with decreased prediction accuracy with increasing genetic divergence (Figs. 3 and 4). These population studies demonstrate the power and utility of increasingly diverse GWAS for prediction, especially in populations of non-European descent.

Although many other traits and diseases have been studied in multi-ancestral settings, few studies have reported comparable metrics of prediction accuracy across populations. Cardiovascular research, for example, has led the charge toward clinical translation of PRS[1]. This enthusiasm has been driven by observations that a polygenic burden of coronary artery disease–increasing SNPs can confer monogenic-equivalent risk of cardiovascular disease, and PRS improve clinical models for risk assessment and statin prescription that can decrease coronary heart disease and improve the

efficiency of healthcare delivery[5–7]. However, many of these studies were conducted exclusively in European-descent populations, and few studies have rigorously evaluated population-level applicability to non-Europeans. Those existing findings indeed demonstrate a large decrease in predictive utility in non-European populations[11], though often with comparisons of odds ratios among arbitrary breakpoints in the risk distribution that make comparisons across studies challenging. To better clarify how polygenic prediction might be deployed in a clinical setting with diverse populations, more systematic and thorough evaluations of the utility of PRS within and across populations for many complex traits are still needed. These evaluations would benefit from rigorous evaluation of polygenic prediction accuracy, especially for diverse non-European patients[61–63].
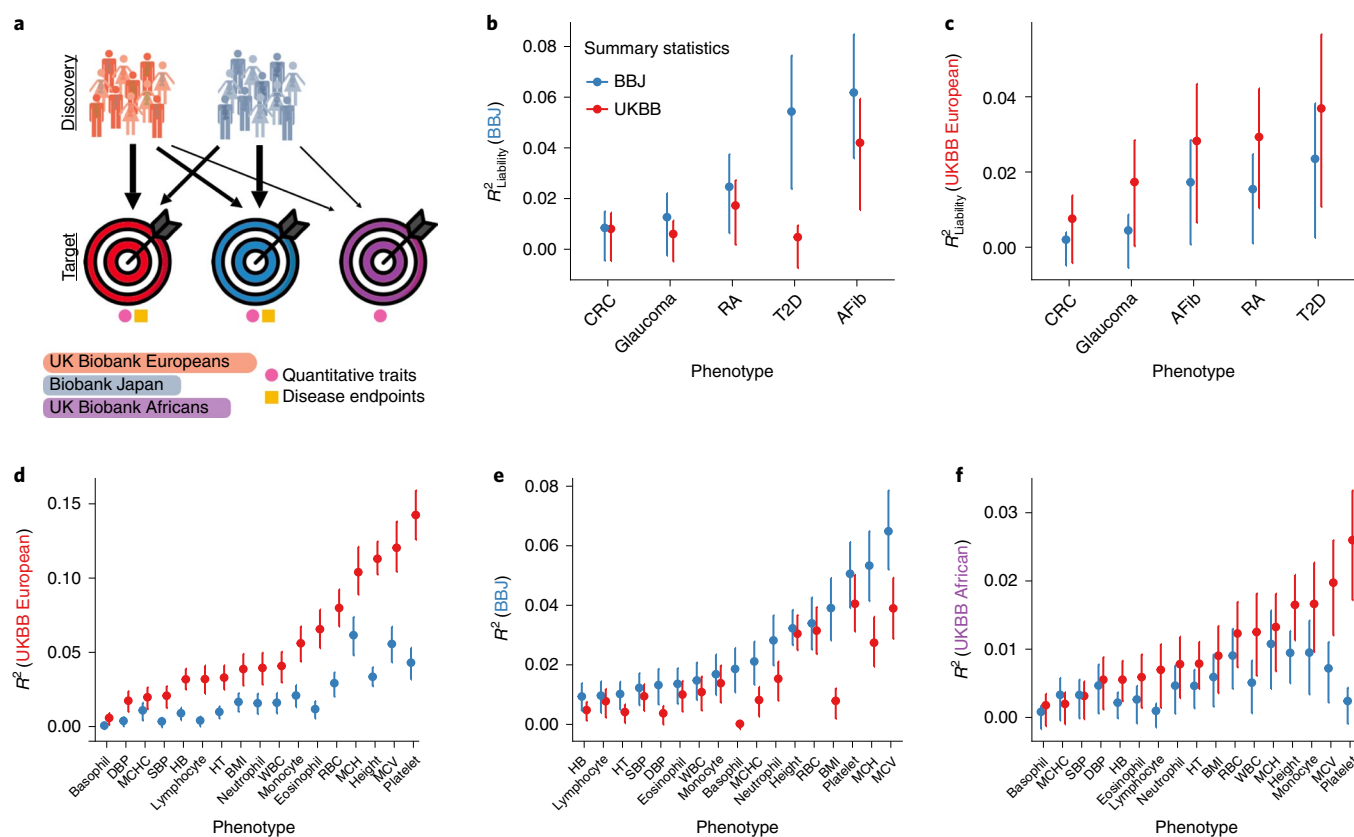
Clinical use of PRS may uniquely exacerbate disparities

Our impetus for raising these statistical issues limiting the generalizability of PRS across populations stems from our concerns that, although they are legitimately clinically promising for improving health outcomes for many biomedical phenotypes, they may have a larger potential to raise health disparities than other clinical factors for several reasons. Because they provide opportunities for improving health outcomes, they inevitably will and should be pursued in the near term, but we caution that a concerted prioritization to make GWAS summary statistics easily accessible for diverse populations and a variety of traits and diseases is imperative, even when they are a fraction of the size of the largest existing European datasets. Individual clinical tests, biomarkers and prescription-drug efficacy may vary across populations in their utility but be fundamentally informed by the same underlying biology[64,65]. Currently, guidelines state that as few as 120 individuals may be used to define reference intervals for clinical factors (though often smaller numbers from only one subpopulation are used), and there is no clear definition of who is 'normal'[64]. Consequently, reference intervals for biomarkers can sometimes deviate considerably by reported ancestry[66–68]. Defining ancestry-specific reference intervals is clearly an important problem that can provide fundamental interpretability gains with implications for some major health benefits (for example, the need for dialysis and the development of type 2 diabetes on the basis of ancestry-specific serum creatinine and hemoglobin A1C reference intervals, respectively)[67]. Therefore, some biomarkers or clinical tests scale directly with health outcomes independently of ancestry, and many others may have distributional differences by ancestry but be equally valid after centering with respect to a readily collected population reference.

In contrast, PRS are uniformly less useful in understudied populations, owing to differences in genomic variation and population history[13,14]. No analogous solution of defining ancestry-specific reference intervals would ameliorate the health-disparity implications for PRS or fundamentally aid in interpretability in non-European populations. Instead, as we and others have demonstrated, PRS are unique in that even with multi-ancestral population references, these scores are fundamentally less informative in populations more diverged from the GWAS study cohorts.

The clinical use and deployment of genetic-risk scores must be informed by the issues surrounding tests that currently would unequivocally provide much greater benefit to the subset of the world's population that is already on the favored end of health disparities. In contrast, predictions for African-descent populations, which already endure many of the largest health disparities globally, are often marginally better, if at all, than at random (Fig. 4f). This population is therefore least likely to benefit from improvements in precision healthcare delivery from PRS with existing data, owing to human population history and study biases. This phenomenon is a major concern globally and especially in the United States, which already leads other middle- and high-income countries in both real

Fig. 4 | Polygenic risk prediction accuracy in Japanese, British and African-descent individuals, on the basis of using independent GWAS of equal sample sizes in the BBJ and UKBB. a, Explanatory diagram showing the different discovery and target cohorts/populations and disease endpoints versus quantitative traits. **b**–**f**, Genetic prediction accuracy computed from independent BBJ and UKBB summary statistics with identical sample sizes (Supplementary Tables 6 and 8). The y axes differ, reflecting differences in prediction accuracy. **b,c**, PRS accuracy for five diseases in Japanese individuals in the BBJ (**b**) and British individuals in the UKBB (**c**). CRC, colorectal cancer; RA, rheumatoid arthritis; T2D, type II diabetes, AFib, atrial fibrillation. **d**–**f**, PRS accuracy for 17 anthropometric and blood panel traits in Japanese individuals in the BBJ (**d**), British individuals in the UKBB (**e**) and African-descent British individuals in the UKBB (**f**). BMI, body mass index; DBP, diastolic blood pressure; HB, hemoglobin; HT, hematocrit; MCH, mean corpuscular hemoglobin; MCHC, mean corpuscular hemoglobin concentration; MCV, mean corpuscular volume; RBC, red blood cell count; SBP, systolic blood pressure; WBC, white blood cell count. Each point shows the maximum R^2 (that is, the best predictor) across five P-value thresholds, and lines correspond to 95% confidence intervals calculated via bootstrap. R^2 values for all P-value thresholds tested are shown in Supplementary Figs. 2–6. Prediction accuracy tends to be higher in the UKBB for quantitative traits than in BBJ and vice versa for disease endpoints, probably because of concomitant phenotype precision and consequently observed heritability for these classes of traits (Supplementary Tables 2–4). Thalassemia and sickle cell disease are unlikely to explain a substantial fraction of the prediction accuracy differences for blood panels across populations, because few individuals have been diagnosed with these disorders via ICD-10 codes (that is, codes from the tenth revision of the International Statistical Classification of Diseases) (Supplementary Table 9).

and perceived health disparities[69,70]. Thus, we strongly urge that any discourse on clinical use of PRS include a careful, quantitative assessment of potential unintentionally introduced effects of economic and health disparities on underrepresented populations, and that awareness be raised regarding how to eliminate these disparities.

How do we even the ledger?

What can be done? The single most important step toward parity in PRS accuracy is vastly increasing the diversity of participants included and analyzed in genetic studies, which would improve utility for all groups, most rapidly for underrepresented groups. Regulatory protections against genetic discrimination are necessary to accompany calls for more diverse studies; although some already exist in the United States, including for health insurance and employment opportunities via the Genetic Information Nondiscrimination Act, stronger protections in these and other areas globally will be particularly important for minorities and/or marginalized groups. An equal investment in GWAS across all major ancestries and

global populations is the most obvious solution to generate a substrate for equally informative risk scores but is not likely to occur any time soon without a dramatic priority shift, given the current imbalance and stalled diversifying progress over the past five years (Fig. 1 and Supplementary Fig. 1). Although acquiring sufficiently large sample sizes for PRS to be equally informative in all populations may be challenging or in some cases infeasible, some much-needed efforts toward increasing diversity in genomics that support open sharing of GWAS summary data from multiple ancestries are underway. Examples include the All of Us Research Program, the Population Architecture using Genomics and Epidemiology (PAGE) Consortium and some disease-focused consortia, such as the T2D-GENES and Stanley Global initiatives on the genetics of type 2 diabetes and psychiatric disorders. Supporting data resources such as imputation panels, multi-ancestral genotyping arrays, gene expression datasets from genetically diverse individuals and other tools are necessary to similarly empower these diverse studies for all populations. The lack of supporting resources for diverse

ancestries creates financial challenges for association studies with limited resources, thus, for example, raising questions about whether to genotype samples on GWAS arrays that may favor European allele frequencies versus sequence samples, and how dense an array to choose or how deeply to sequence[71,72].

Additional leading global efforts also provide easy unified access linking genetic, clinical-record and national-registry data in more homogeneous continental ancestries, such as the UKBB, BBJ, China Kadoorie Biobank and Nordic efforts (for example, in Danish, Estonian, Finnish and other integrated biobanks). Notably, some of these biobanks such as UKBB have participants with considerable global genetic diversity that enables multi-ancestral comparisons; although minorities from this cohort provide the largest deeply phenotyped GWAS cohorts for several ancestries, these individuals are often excluded in current statistical analyses in favor of the simplicity afforded by analyzing only the largest genetically homogeneous European-ancestry data. These considerations notwithstanding, there are critical needs and challenges for expanding the scale of genetic studies of heritable traits in diverse populations; these challenges are especially apparent in Africa, where humans originated and have retained the most genetic diversity, because Africans are understudied but disproportionately informative for genetic analyses and evolutionary history[27,73]. The most notable investment here comes from the Human Heredity and Health in Africa (H3Africa) Initiative, increasing genomics research capacity in Africa through more than US$216 million in funding from the US National Institutes of Health and the Wellcome Trust (United Kingdom) for genetic research led by African investigators[55,74]. The increasing interest and scale of genetic studies in low- and middle-income countries (LMICs) raises ethical and logistical considerations about data generation, access, sharing, security and analysis, as well as clinical implementation, to ensure that these advances do not benefit only high-income countries. Frameworks such as the H3ABioNet, a pan-African bioinformatics network designed to build capacity to enable H3Africa researchers to analyze their data in Africa, provide cost-effective examples for training local scientists in LMICs[75].

The prerequisite data for dramatically increasing diversity also exist in several large-scale publicly funded datasets such as the Million Veterans Project and Trans-Omics for Precision Medicine (TOPMed), but with problematic data access issues in which even GWAS summary data within and across populations are not publicly shared. Existing GWAS consortia also must carefully consider the granularity of the summary statistics that they release, because finer-scale continental ancestries and phenotypes in large multi-ancestral projects enable ancestry-matched analyses that are not possible with a single set of summary statistics. Although an understandable patient-privacy balance must be struck when sharing individual-level data, GWAS summary statistics from all publicly funded and as many privately funded projects as possible should be made easily and publicly accessible to improve global health outcomes. Efforts to unify the phenotype definitions, normalization approaches and GWAS methods among studies will also improve comparability.

To enable progress toward parity, open data-sharing standards must critically be adopted for all ancestries and for genetic studies of all sample sizes, not just the largest European results. Locally appropriate and secure genetic-data-sharing techniques as well as equitable technology availability must be adopted widely in South America, Asia and Africa, as they are in Europe and North America, to ensure the achievement of maximum value from the existing and ongoing efforts being developed to help counter the current imbalance. Simultaneously, ethical considerations require that research capacity be increased in LMICs with simultaneous growth of diverse population studies to balance the benefits of these studies to scientists and patients globally versus locally, thereby ensuring

that everyone benefits. Methodological improvements to PRS by appropriately accounting for population allele frequency, LD and/or admixture differences are underway and may help considerably but will not by themselves bring equality. All these efforts are important and should be prioritized not just for risk prediction but more generally to maximize the use and applicability of genetics to provide information on the biology of disease. Given the acute recent attention paid to clinical use of PRS, we believe that it is paramount to recognize their potential to improve health outcomes for all individuals and many complex diseases. Simultaneously, the field must address the disparity in utility in an ethically thoughtful and scientifically rigorous fashion, lest genetic technologies inadvertently contribute to, rather than decrease, existing health disparities.

References

1. Knowles, J. W. & Ashley, E. A. Cardiovascular disease: the rise of the genetic risk score. *PLoS Med.* **15**, e1002546–e1002547 (2018).
2. Maas, P. et al. Breast cancer risk from modifiable and nonmodifiable risk factors among white women in the United States. *JAMA Oncol.* **2**, 1295–1302 (2016).
3. Schumacher, F. R. et al. Association analyses of more than 140,000 men identify 63 new prostate cancer susceptibility loci. *Nat. Genet.* **50**, 928–936 (2018).
4. Sharp, S. A. et al. Development and standardization of an improved type 1 diabetes genetic risk score for use in newborn screening and incident diagnosis. *Diabetes Care* **42**, 200–207 (2019).
5. Khera, A. V. et al. Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations. *Nat. Genet.* **50**, 1219–1224 (2018).
6. Kullo, I. J. et al. Incorporating a genetic risk score into coronary heart disease risk estimates: effect on low-density lipoprotein cholesterol levels (the MI-GENES Clinical Trial). *Circulation* **133**, 1181–1188 (2016).
7. Natarajan, P. et al. Polygenic risk score identifies subgroup with higher burden of atherosclerosis and greater relative benefit from statin therapy in the primary prevention setting. *Circulation* **135**, 2091–2101 (2017).
8. Paquette, M. et al. Polygenic risk score predicts prevalence of cardiovascular disease in patients with familial hypercholesterolemia. *J. Clin. Lipidol.* **11**, 725–732.e5 (2017).
9. Tikkanen, E., Havulinna, A. S., Palotie, A., Salomaa, V. & Ripatti, S. Genetic risk prediction and a 2-stage risk screening strategy for coronary heart disease. *Arterioscler. Thromb. Vasc. Biol.* **33**, 2261–2266 (2013).
10. Frieser, M. J., Wilson, S. & Vrieze, S. Behavioral impact of return of genetic test results for complex disease: systematic review and meta-analysis. *Health Psychol.* **37**, 1134–1144 (2018).
11. Khera, A. V. et al. Genetic risk, adherence to a healthy lifestyle, and coronary disease. *N. Engl. J. Med.* **375**, 2349–2358 (2016).
12. Khera, A. V. & Kathiresan, S. Genetics of coronary artery disease: discovery, biology and clinical translation. *Nat. Rev. Genet.* **18**, 331–344 (2017).
13. Martin, A. R. et al. Human demographic history impacts genetic risk prediction across diverse populations. *Am. J. Hum. Genet.* **100**, 635–649 (2017).
14. Scutari, M., Mackay, I. & Balding, D. Using genetic distance to infer the accuracy of genomic prediction. *PLoS Genet.* **12**, e1006288 (2016).
15. Vilhjálmsson, B. J. et al. Modeling linkage disequilibrium increases accuracy of polygenic risk scores. *Am. J. Hum. Genet.* **97**, 576–592 (2015).
16. Ware, E. B. et al. Heterogeneity in polygenic scores for common human traits. Preprint at https://www.biorxiv.org/content/10.1101/106062v1 (2017).
17. Curtis, D. Polygenic risk score for schizophrenia is more strongly associated with ancestry than with schizophrenia. *Psychiatr. Genet.* **28**, 85–89 (2018).
18. Schizophrenia Working Group of the Psychiatric Genomics Consortium. Biological insights from 108 schizophrenia-associated genetic loci. *Nature* **511**, 421–427 (2014).
19. Belsky, D. W. et al. Development and evaluation of a genetic risk score for obesity. *Biodemography Soc. Biol.* **59**, 85–100 (2013).
20. Domingue, B. W., Belsky, D., Conley, D., Harris, K. M. & Boardman, J. D. Polygenic influence on educational attainment: new evidence from The National Longitudinal Study of Adolescent to Adult Health. *AERA Open* **1**, 1–13 (2015).

21. Lee, J. J. et al. Gene discovery and polygenic prediction from a genome-wide association study of educational attainment in 1.1 million individuals. *Nat. Genet.* **50**, 1112–1121 (2018).

22. Vassos, E. et al. An examination of polygenic score risk prediction in individuals with first-episode psychosis. *Biol. Psychiatry* **81**, 470–477 (2017).

23. Akiyama, M. et al. Genome-wide association study identifies 112 new loci for body mass index in the Japanese population. *Nat. Genet.* **49**, 1458–1467 (2017).

24. Li, Z. et al. Genome-wide association analysis identifies 30 new susceptibility loci for schizophrenia. *Nat. Genet.* **49**, 1576–1583 (2017).

25. Need, A. C. & Goldstein, D. B. Next generation disparities in human genomics: concerns and remedies. *Trends Genet.* **25**, 489–494 (2009).

26. Popejoy, A. B. & Fullerton, S. M. Genomics is failing on diversity. *Nature* **538**, 161–164 (2016).

27. Morales, J. et al. A standardized framework for representation of ancestry data in genomics studies, with application to the NHGRI-EBI GWAS Catalog. *Genome Biol.* **19**, 21 (2018).

28. Rosenberg, N. A. et al. Genome-wide association studies in diverse populations. *Nat. Rev. Genet.* **11**, 356–366 (2010).

29. Sham, P. C., Cherny, S. S., Purcell, S. & Hewitt, J. K. Power of linkage versus association analysis of quantitative traits, by use of variance-components models, for sibship data. *Am. J. Hum. Genet.* **66**, 1616–1630 (2000).

30. 1000 Genomes Project Consortium. et al. A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).

31. Williams, A. L. et al. Sequence variants in SLC16A11 are a common risk factor for type 2 diabetes in Mexico. *Nature* **506**, 97–101 (2014).

32. Estrada, K. et al. Association of a low-frequency variant in HNF1A with type 2 diabetes in a Latino population. *JAMA* **311**, 2305–2314 (2014).

33. Haiman, C. A. et al. Genome-wide association study of prostate cancer in men of African ancestry identifies a susceptibility locus at 17q21. *Nat. Genet.* **43**, 570–573 (2011).

34. Genovese, G. et al. Association of trypanolytic ApoL1 variants with kidney disease in African Americans. *Science* **329**, 841–845 (2010).

35. Liu, J. Z. et al. Association analyses identify 38 susceptibility loci for inflammatory bowel disease and highlight shared genetic risk across populations. *Nat. Genet.* **47**, 979–986 (2015).

36. Carlson, C. S. et al. Generalization and dilution of association results from European GWAS in populations of non-European ancestry: the PAGE study. *PLoS Biol.* **11**, e1001661 (2013).

37. Easton, D. F. et al. Genome-wide association study identifies novel breast cancer susceptibility loci. *Nature* **447**, 1087–1093 (2007).

38. DIAbetes Genetics Replication And Meta-analysis (DIAGRAM) Consortium. et al. Genome-wide trans-ancestry meta-analysis provides insight into the genetic architecture of type 2 diabetes susceptibility. *Nat. Genet.* **46**, 234–244 (2014).

39. Waters, K. M. et al. Consistent association of type 2 diabetes risk variants found in europeans in diverse racial and ethnic groups. *PLoS Genet.* **6**, e1001078–e1001079 (2010).

40. Lam, M. et al. Comparative genetic architectures of schizophrenia in East Asian and European populations. Preprint at https://www.biorxiv.org/content/10.1101/445874v2 (2018).

41. McCarthy, S. et al. A reference panel of 64,976 haplotypes for genotype imputation. *Nat. Genet.* **48**, 1279–1283 (2016).

42. Huang, H. et al. Fine-mapping inflammatory bowel disease loci to single-variant resolution. *Nature* **547**, 173–178 (2017).

43. Sohail, M. et al. Signals of polygenic adaptation on height have been overestimated due to uncorrected population structure in genome-wide association studies. Preprint at https://www.biorxiv.org/content/10.1101/355057v3 (2018).

44. Berg, J. J. et al. Reduced signal for polygenic adaptation of height in UK Biobank. Preprint at https://www.biorxiv.org/content/10.1101/354951v4 (2018).

45. Kerminen, S. et al. Geographic variation and bias in polygenic scores of complex diseases and traits in Finland. Preprint at https://www.biorxiv.org/content/10.1101/485441v1 (2018).

46. Novembre, J. & Barton, N. H. Tread lightly interpreting polygenic tests of selection. *Genetics* **208**, 1351–1355 (2018).

47. Henn, B. M., Botigué, L. R., Bustamante, C. D., Clark, A. G. & Gravel, S. Estimating the mutation load in human genomes. *Nat. Rev. Genet.* **16**, 333–343 (2015).

48. Brown, B. C., Asian Genetic Epidemiology Network Type 2 Diabetes Consortium, Ye, C. J., Price, A. L. & Zaitlen, N. Transethnic genetic-correlation estimates from summary statistics. *Am. J. Hum. Genet.* **99**, 76–88 (2016).

49. Galinsky, K. J. et al. Estimating cross-population genetic correlations of causal effect sizes. *Genet. Epidemiol.* **43**, 180–188 (2019).

50. Li, D., Zhao, H. & Gelernter, J. Strong protective effect of the aldehyde dehydrogenase gene (ALDH2) 504lys (*2) allele against alcoholism and alcohol-induced medical diseases in Asians. *Hum. Genet.* **131**, 725–737 (2012).

51. Zhu, Z. et al. Dominance genetic variation contributes little to the missing heritability for human complex traits. *Am. J. Hum. Genet.* **96**, 377–385 (2015).

52. Paré, G., Mao, S. & Deng, W. Q. A machine-learning heuristic to improve gene score prediction of polygenic traits. *Sci. Rep.* **7**, 12665 (2017).

53. Martin, A. R. et al. An unexpectedly complex architecture for skin pigmentation in Africans. *Cell* **171**, 1340–1353.e14 (2017).

54. Duncan, L. E. et al. Largest GWAS of PTSD (N=20 070) yields genetic overlap with schizophrenia and sex differences in heritability. *Mol. Psychiatry* **23**, 666–673 (2018).

55. H3Africa Consortium. et al. Enabling the genomic revolution in Africa. *Science* **344**, 1346–1348 (2014).

56. Hindorff, L. A. et al. Prioritizing diversity in human genomics research. *Nat. Rev. Genet.* **19**, 175–185 (2018).

57. Kanai, M. et al. Genetic analysis of quantitative traits in the Japanese population links cell types to complex human diseases. *Nat. Genet.* **50**, 390–400 (2018).

58. Howrigan, D. *Details and Considerations of the UK Biobank GWAS*. http://www.nealelab.is/blog/2017/9/11/details-and-considerations-of-the-uk-biobank-gwas (accessed 9 November 2017)

59. Fry, A. et al. Comparison of sociodemographic and health-related characteristics of UK Biobank participants with those of the general population. *Am. J. Epidemiol.* **186**, 1026–1034 (2017).

60. Liu, S. et al. Genomic analyses from non-invasive prenatal testing reveal genetic associations, patterns of viral infections, and Chinese population history. *Cell* **175**, 347–359.e14 (2018).

61. Wray, N. R. et al. Pitfalls of predicting complex traits from SNPs. *Nat. Rev. Genet.* **14**, 507–515 (2013).

62. Wray, N. R. et al. Research review: polygenic methods and their application to psychiatric traits. *J. Child Psychol. Psychiatry* **55**, 1068–1087 (2014).

63. Torkamani, A., Wineinger, N. E. & Topol, E. J. The personal and clinical utility of polygenic risk scores. *Nat. Rev. Genet.* **19**, 581–590 (2018).

64. Manrai, A. K., Patel, C. J. & Ioannidis, J. P. A. In the era of precision medicine and big data, who is normal? *JAMA* **319**, 1981–1982 (2018).

65. Plenge, R. M., Scolnick, E. M. & Altshuler, D. Validating therapeutic targets through human genetics. *Nat. Rev. Drug Discov.* **12**, 581–594 (2013).

66. Carroll, M. D., Kit, B. K., Lacher, D. A., Shero, S. T. & Mussolino, M. E. Trends in lipids and lipoproteins in US adults, 1988–2010. *JAMA* **308**, 1545–1554 (2012).

67. Rappoport, N. et al. Comparing ethnicity-specific reference intervals for clinical laboratory tests from EHR data. *J. Appl. Lab. Med.* **3**, 366–377 (2018).

68. Lim, E., Miyamura, J. & Chen, J. J. Racial/ethnic-specific reference intervals for common laboratory tests: a comparison among Asians, Blacks, Hispanics, and White. *Hawaii J. Med. Public Health* **74**, 302–310 (2015).

69. Hero, J. O., Zaslavsky, A. M. & Blendon, R. J. The United States leads other nations in differences by income in perceptions of health and health care. *Health Aff. (Millwood)* **36**, 1032–1040 (2017).

70. Williams, D. R., Priest, N. & Anderson, N. B. Understanding associations among race, socioeconomic status, and health: Patterns and prospects. *Health Psychol.* **35**, 407–411 (2016).

71. Gilly, A. et al. Very low depth whole genome sequencing in complex trait association studies. *Bioinformatics* https://doi.org/10.1093/bioinformatics/bty1032 (2018).

72. Pasaniuc, B. et al. Extremely low-coverage sequencing and imputation increases power for genome-wide association studies. *Nat. Genet.* **44**, 631–635 (2012).

73. Martin, A. R., Teferra, S., Möller, M., Hoal, E. G. & Daly, M. J. The critical needs and challenges for genetic architecture studies in Africa. *Curr. Opin. Genet. Dev.* **53**, 113–120 (2018).

74. Coles, E. & Mensah, G. A. Geography of genetics and genomics research funding in Africa. *Glob. Heart* **12**, 173–176 (2017).

75. Mulder, N. J. et al. Development of bioinformatics infrastructure for genomics research. *Glob. Heart* **12**, 91–98 (2017).

76. MacArthur, J. et al. The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog). *Nucleic Acids Res.* **45**, D896–D901 (2017).

Author contributions

A.R.M. and M.J.D. conceived and designed the experiments. A.R.M. and M.K. performed statistical analysis. A.R.M. and M.K. analyzed the data. A.R.M., M.K., Y.K., Y.O., B.M.N. and M.J.D. contributed reagents/materials/analysis tools. A.R.M., M.K., B.M.N. and M.J.D. wrote the paper.

Competing interests

The authors declare no competing interests.

Additional information

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

nature research

Corresponding author(s):   Alicia R. Martin

Last updated by author(s):   Jan 21, 2019

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒	A description of all covariates tested
☐	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	N/A
Data analysis	All code used to perform analyses are accessible at the following site: https://github.com/armartin/prs_disparities. Hail software (https://hail.is/) was used to perform GWAS analyses (v0.1) and compute polygenic risk scores (v0.2).

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:
- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

UK Biobank analyses were conducted via application 31063. BBJ GWAS summary statistics are publicly available at our website (http://jenger.riken.jp/en/) and the National Bioscience Database Center (NBDC) Human Database (Research ID: hum0014). Genotype data from the BBJ subjects was deposited at the NBDC Human Database (Research ID: hum0014).

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	Our sample sizes were determined by the availability of UK Biobank and BioBank Japan data. GWAS for each of 17 phenotypes were selected by using the largest sample size available in BioBank Japan (a range of N=74287 for neutrophils to N=151,569 individuals per phenotype) with the exclusion of a small target cohort for prediction (N ~ 5,000), and matching the GWAS sample sizes in UK Biobank
Data exclusions	We excluded samples and variants based on the standard quality control for GWAS described in our manuscript.
Replication	We do not report or attempt to replicate individual genome-wide significant loci. Replicating generalizability of polygenic risk scores across similar populations is not feasible because similar scales of GWAS in non-overlapping samples for these phenotypes are not currently available.
Randomization	Our study was not experimental, so this question does not pertain to our study.
Blinding	N/A

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology
☒	☐ Animals and other organisms
☐	☒ Human research participants
☒	☐ Clinical data

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Human research participants

Policy information about studies involving human research participants

Population characteristics	UK Biobank recruited ~500,000 people aged between 40-69 years in 2006-2010 from across the country to take part in this project. BioBank Japan recruited ~200,000 patients of any of 47 target diseases between fiscal years of 2003 and 2007 from 66 cooperating hospitals across Japan. In our analysis, we separately included the individuals of European (for UKBB) or Japanese (for BBJ) descent that were not genetic outliers and passed the quality control tests as described previously. For each of the 17 studied quantitative traits, GWAS was separately conducted using the subjects whose phenotype of the trait was available.
Recruitment	UK Biobank and BioBank Japan participants were recruited to these biobanks as part of previous studies. Our study had no influence or control on recruitment.
Ethics oversight	All genetic and corresponding phenotype information was previously collected and de-identified, and this work was therefore deemed non-human subjects research.

Note that full information on the approval of the study protocol must also be provided in the manuscript.