The PRIMED Consortium: Reducing disparities in polygenic risk assessment

Iftikhar J. Kullo,[1,16,*] Matthew P. Conomos,[2,16] Sarah C. Nelson,[2,16] Sally N. Adebamowo,[3] Ananyo Choudhury,[4] David Conti,[5] Stephanie M. Fullerton,[6] Stephanie M. Gogarten,[2] Ben Heavner,[2] Whitney E. Hornsby,[7,11] Eimear E. Kenny,[8] Alyna Khan,[2] Amit V. Khera,[7] Yun Li,[9] Iman Martin,[10] Josep M. Mercader,[11] Maggie Ng,[12] Laura M. Raffield,[9] Alex Reiner,[13] Robb Rowley,[10] Daniel Schaid,[14] Adrienne Stilp,[2] Ken Wiley,[10] Riley Wilson,[10] John S. Witte,[15] Pradeep Natarajan,[7,11] and Polygenic Risk Methods in Diverse Populations (PRIMED) Consortium

Summary

By improving disease risk prediction, polygenic risk scores (PRSs) could have a significant impact on health promotion and disease prevention. Due to the historical oversampling of populations with European ancestry for genome-wide association studies, PRSs perform less well in other, understudied populations, leading to concerns that clinical use in their current forms could widen health care disparities. The PRIMED Consortium was established to develop methods to improve the performance of PRSs in global populations and individuals of diverse genetic ancestry. To this end, PRIMED is aggregating and harmonizing multiple phenotype and genotype datasets on AnVIL, an interoperable secure cloud-based platform, to perform individual- and summary-level analyses using population and statistical genetics approaches. Study sites, the coordinating center, and representatives from the NIH work alongside other NHGRI and global consortia to achieve these goals. PRIMED is also evaluating ethical and social implications of PRS implementation and investigating the joint modeling of social determinants of health and PRS in computing disease risk. The phenotypes of interest are primarily cardiometabolic diseases and cancer, the leading causes of death and disability worldwide. Early deliverables of the consortium include methods for data sharing on AnVIL, development of a common data model to harmonize phenotype and genotype data from cohort studies as well as electronic health records, adaptation of recent guidelines for population descriptors to global cohorts, and sharing of PRS methods/tools. As a multisite collaboration, PRIMED aims to foster equity in the development and use of polygenic risk assessment.

Introduction

As genomic technologies spur progress in several areas of precision medicine, inequity in translating these advances to diverse groups has become evident. The historical bias in sampling for genome-wide association studies (GWASs) hinders application of the results to diverse groups, leading to inequity and bypassing potential scientific opportunities such as identifying causal variants, improving genetic risk prediction, and enhancing understanding of the genetic architecture of disease. Consequently, one of the guiding principles of the National Human Genome Research Institute (NHGRI)'s strategic vision[1] is "… to commit to systematic inclusion of underrepresented groups in future NHGRI programs and projects." The American Society for Human Genetics has also issued guidance stating that "benefits of genomic medicine should be accessible to all people, and this requires focused efforts to address health inequities and remove barriers to increase representation of diverse communities in genetics and genomics research."[2]

Polygenic risk scores (PRSs) represent a summation of genetic predisposition to disease susceptibility.[3] While there are many potential research and clinical applications of PRSs, a major focus is on the use of PRSs to refine risk estimation of common diseases beyond clinical risk factors and family history.[3] PRSs could have relevance for most individuals in a population by refining risk estimates for common diseases early in the life course when prevention may be most effective.[4,5] A PRS value in the top 5th percentile of the population distribution for several adult-onset cardiometabolic diseases and cancer may pose a relative risk comparable to monogenic etiology for these conditions.[6] Individuals with high predicted polygenic risk may benefit from heightened surveillance or risk mitigation interventions, a concept that is being prospectively

[1]Department of Cardiovascular Medicine, Mayo Clinic, Rochester, MN, USA; [2]Department of Biostatistics, University of Washington, Seattle, WA, USA; [3]Department of Epidemiology and Public Health, University of Maryland, Baltimore, MD, USA; [4]Sydney Brenner Institute of Molecular Bioscience, University of Witwatersrand, Johannesburg, South Africa; [5]Department of Population and Public Health Sciences, University of Southern California, Los Angeles, CA, USA; [6]Department of Bioethics and Humanities, University of Washington School of Medicine, Seattle, WA, USA; [7]Center for Genomic Medicine, Massachusetts General Hospital, Boston, MA, USA; [8]Institute of Genomic Health, Icahn School of Medicine at Mount Sinai, New York, NY, USA; [9]Department of Genetics, University of North Carolina Chapel Hill, Chapel Hill, NC, USA; [10]National Human Genome Research Institute, National Institutes of Health, Baltimore, MD, USA; [11]Programs in Metabolism and Medical & Population Genetics, Broad Institute of Harvard and MIT, Cambridge, MA, USA; [12]Division of Genetic Medicine, Vanderbilt University Medical Center, Nashville, TN, USA; [13]Department of Epidemiology, Fred Hutchinson Cancer Center, Seattle, WA, USA; [14]Department of Quantitative Health Sciences, Mayo Clinic, Rochester, MN, USA; [15]Department of Epidemiology and Population Health, Stanford University, Stanford, CA, USA
[16]These authors contributed equally
*Correspondence: kullo.iftikhar@mayo.edu
https://doi.org/10.1016/j.ajhg.2024.10.010.

tested for several conditions across geographically diverse healthcare systems by the NHGRI's electronic Medical Records and Genomics (eMERGE) Network.[7] Given the likelihood of widespread availability of PRSs in the near future, some medical professional societies have already identified potential clinical implementation opportunities as well as outstanding challenges.[8,9]

A major hurdle in the clinical use of PRSs is the lower predictive performance among individuals from understudied groups, such as those of African, admixed, or other diverse ancestry.[10–14] For example, a 1-standard deviation increase in a PRS for coronary heart disease was associated with an odds ratio of 1.53 in European ancestry individuals but only 1.27 in African ancestry individuals.[10] Despite several recent advances in PRS methodology, these differences in PRS performance remain, and are largely due to significant underrepresentation of human diversity in GWASs, the primary source data for PRS development. Population differences in allele frequencies, linkage disequilibrium, causal variant effect sizes, and gene-gene or gene-environment interactions[15] reduce the portability of PRSs across groups.[16] Several global GWAS meta-analyses have demonstrated the utility of large cohorts to discover "genomic" associations[17,18]; however, individuals of diverse genetic ancestries remain inadequately represented in such analyses. Thus, deployment of PRSs in their present form carries the risk of exacerbating health disparities.[19] In response, and given the ethical and scientific imperatives, funders such as the NHGRI and National Cancer Institute (NCI) are taking steps to address this concern.[20]

One such step is the creation of the Polygenic Risk Methods in Diverse Populations (PRIMED) Consortium by the NHGRI and the NCI to incorporate new GWAS data and leverage methodologic and computational advances to bridge the performance gap of PRSs in diverse groups. The PRIMED Consortium has two primary objectives: (1) to bring together and harmonize extant datasets with genotype and phenotype measures from diverse ancestry groups in support of PRS development and evaluation and (2) to develop new methods to improve polygenic risk prediction across diverse groups for a broad range of health and disease outcomes. The consortium is aggregating data on the secure, scalable Analysis Visualization and Informatics Lab-space (AnVIL) platform established by the NHGRI[21] for centralized analyses, as well as implementing coordinated analysis protocols for federated analyses across affiliated studies and biobanks. Through the complementary expertise of consortium members, the use of data and other resources generated by programs such as the Clinical Genome Resource (ClinGen)[22] and All of Us,[23] and collaborations with partner programs such as eMERGE,[7] PRIMED is developing, testing, and refining PRSs for use in diverse groups. In this perspective, we describe the PRIMED Consortium and organization of its activities and highlight methodological innovations and early products, with the goal of facilitating additional initiatives that aim to reduce inequity in genomic medicine.

Consortium structure and approach

PRIMED was established under a "diversity first" principle, emphasizing use of non-European ancestry data in PRS methods development and refinement, even if European ancestry datasets are larger and more frequently used (see RFA-HG-20-001 in web resources). The consortium comprises seven study sites, a coordinating center (CC), NIH program staff, and affiliate members (Figure 1A; Table S1). Core and affiliate members span 49 institutions in 12 countries (Figure 1B), collectively providing access to data from >75 existing studies and consortia. Study sites were selected based on ability to contribute datasets of diverse, non-European genetic ancestry with a broad range of phenotypes and genomic data, as well as member expertise in population genetics and statistical genetics relevant to PRS methodology. PRIMED investigators are affiliated with and making use of data from other large-scale precision health research programs focused on diversity, including the Million Veteran Program, the All of Us Research Program, and the NHLBI Trans-Omics for Precision Medicine (TOPMed) program. Additionally, PRIMED is fostering collaborations with multiple biobanks across the world.[24] The CC serves the PRIMED Consortium in supporting four main areas: data sharing, data harmonization, analysis and methods, and program coordination. NIH personnel are closely involved in shaping scientific direction and goals and work with the PRIMED investigators and the CC to facilitate the programmatic and scientific activities of the consortium. External investigators with complementary expertise and/or datasets to contribute are eligible to join the consortium by application as affiliate members (see web resources). An external scientific panel of experts assists the NIH in assessing and guiding the consortium.

Cloud-based data sharing and analyses

To achieve its goals related to PRS methods development and evaluation, PRIMED is bringing together many large datasets with genomic and phenotypic measures from diverse ancestry groups. The scale and scope of data sharing among PRIMED investigators within the consortium poses both technical and policy challenges. As a "consortium of consortia," PRIMED is analyzing individual- and summary-level data from numerous datasets, many of which are publicly available via repositories such as dbGaP, while others are contributed by the study investigators for consortium use. Collectively, participants in these studies and consortia identified for primary use in PRIMED are from over 40 countries around the globe (Figure S1; Table S2) and include many different ancestries, nationalities, and racial and ethnic groups (see Table S3). These data cover a broad range of phenotypes, genomic

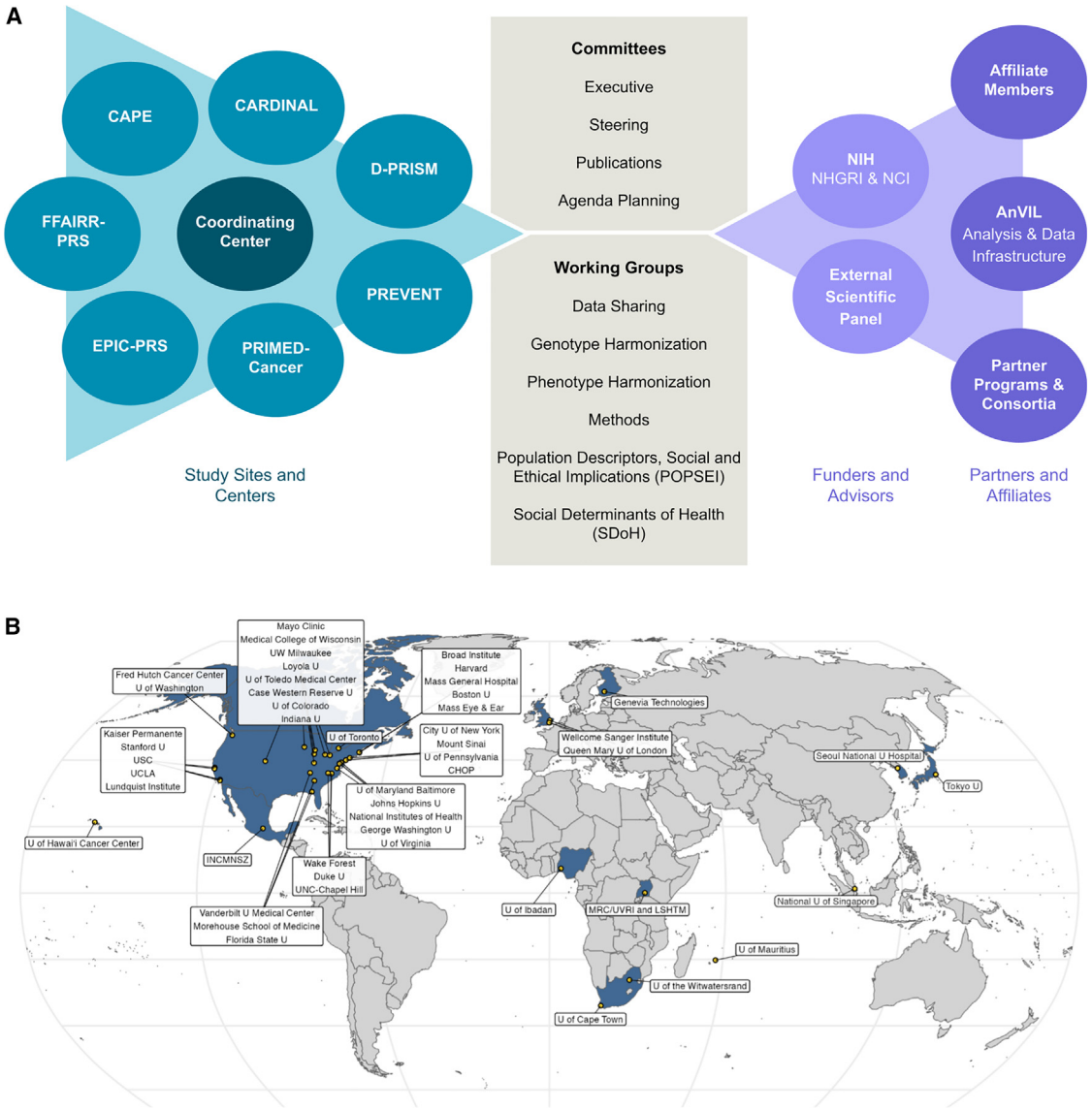

Figure 1. Overview of the PRIMED Consortium structure and investigators
(A) The consortium includes study sites, coordinating center, NIH funders, committees, working groups, external advisors, affiliates, and other partner programs/consortia.
(B) PRIMED investigators are located in 49 institutions across 12 countries.

data types, and diverse genetic ancestries, and they have varying data use restrictions and access procedures.

Due to the complexity and heterogeneity of data sharing in PRIMED arising from the various sources of data, a data sharing working group (WG) was created to develop policies and processes for data sharing both within the consortium and with the broader scientific community. The WG developed a data sharing policy (see web resources) with two primary mechanisms for intra-consortium data sharing: (1) coordinated dbGaP applications for data accessible via dbGaP and (2) a consortium data sharing Agreement for studies and datasets provided by study investigators for consortium use. Each of these mechanisms establish a data sharing circle (i.e., a group of investigators with permission to access the same data such that the data can be shared among them collectively) that can be implemented on the AnVIL cloud platform via consortium shared data workspaces, and collaborative analyses can be performed using a single, centralized copy of the data. However, some data are not shareable in centralized consortium workspaces on the AnVIL cloud platform, e.g., individual-level data from biobanks or datasets under specific privacy laws such as the General Data Protection Regulation in the European Union. In these cases, coordinated analysis protocols are distributed for federated analyses to generate unrestricted summary-level data that can be shared within the consortium on AnVIL (see the supplemental methods for additional information).

The data sharing WG also facilitates sharing of consortium-generated data products with the broader scientific community via AnVIL and/or specialized repositories in alignment with NIH data sharing policies. The WG

has engaged in policy discussions with NHGRI aimed at clarifying the definition of genomic summary results (GSRs) in the context of PRS development and is exploring ways to share GSRs that require controlled access. These conversations have prompted further internal policy discussion at NIH. When feasible, the consortium is releasing GSRs, including PRS scoring files and GWAS statistics, on the open access PGS Catalog[25] and GWAS Catalog,[26] respectively, noting in some cases this is not permissible when contributing studies are designated as sensitive for the purposes of GSR sharing (see web resources, NIH GSR sharing policy). PRS performance metrics are also posted to the PGS Catalog and a dynamic report of all PRIMED PRS development or evaluation submissions to the PGS Catalog is available on the public-facing consortium website (see web resources). While the WG is exploring ways to release individual-level derived consortium data products (e.g., harmonized phenotypes, imputed genotypes), as a consortium making secondary use of pre-existing datasets, this is generally not allowable, e.g., due to the non-transferability clause of the standard dbGaP Data Use Certification Agreement.

Phenotype harmonization

Extensive phenotype data are available from studies and cohorts contributed by the study sites and derive from two main sources: prospective cohort studies and clinic- or hospital-based biobanks with linked electronic health record (EHR) data, such as those in the eMERGE Network. Harmonizing such data is essential to conducting *de novo* GWASs across genetically diverse groups for PRS construction and methods comparison, keeping in mind different ascertainment strategies (self-report, diagnosis codes, physician adjudication) or phenotype definitions such as those cataloged by eMERGE in the PheKB.[27] Other considerations include use of repeat measurements to reduce noise in quantitative phenotypes, allowing better capture of true genetic effects,[28] and leveraging longitudinal data to develop PRSs for incident (rather than prevalent) disease.

The consortium established a phenotype harmonization WG to address the complexities that arise from combining and harmonizing heterogeneous phenotype data across many studies. The WG defined phenotype data standards, formats, ontologies, and metadata requirements to enable consortium-wide phenotype data collection and subsequent harmonization, taking into account procedures developed in other consortia; e.g., from NHLBI's TOPMed program,[29] the NHLBI Pooled Cohorts Study,[30] and eMERGE.[27] The WG identified priority traits and diseases based on public health burden and study site expertise, which led to four initial phenotype-domain sub-WGs (cancer, cardiometabolic quantitative traits, cardiovascular disease, and diabetes outcomes and complications) tasked with establishing phenotype definitions and harmonization algorithms. The WG coordinates harmonization efforts across studies, and study sites upload harmonized phenotype data to the AnVIL data workspaces maintained by the CC for sharing with the broader consortium. The consortium developed its own common data model for data uploaded to AnVIL data workspaces to accommodate ongoing phenotype harmonization efforts, heterogeneous study data, and longitudinal observations (see initial consortium products section).

Environmental factors and social determinants of health

Environmental exposure differences and gene-environment interactions may influence the generalizability of PRSs. The availability of data on social determinants of health (SDoH) and environmental measures is expected to vary depending on the cohort type and whether the cohort includes legacy data only or has ongoing data collection. Prospectively assembled cohorts may include lifestyle and exposure measures derived from surveys, but such variables are difficult to ascertain in EHR-linked biobank cohorts, and income, education, and geocode-based linkage with area-level metrics have been used as surrogates for "environmental" exposures.[31] PRIMED investigators are exploring methods to incorporate environmental measures and SDoH into both the development and contextualization of PRSs[32,33] and the implementation of multivariable risk models that include PRSs to evaluate whether such efforts can improve PRS accuracy within and across groups.[33]

Due to the complexity of measuring SDoH and environment and the anticipated importance of these factors for risk prediction, whether as effect modifiers, mediators, and/or separate risk factors, the consortium established a social determinants of health (SDoH) WG to identify and define relevant SDoH and other environmental phenotypic measures to be considered for use in PRIMED. However, due to differences in availability and definitions of SDoH and environmental measures across the legacy datasets used in PRIMED, harmonization is challenging. Therefore, the focus of this effort has turned to identifying and harmonizing variables within select cohorts in support of specific projects. The WG oversees the harmonization and documentation of SDoH variables selected for use in these analyses and collaborates with the phenotype harmonization WG to meet the consortium data standards. The SDoH WG is also tasked with developing a conceptual framework to integrate PRSs and SDoH in risk prediction models for diverse groups. The WG is working closely with the methods WG to incorporate SDoH variables into PRS analyses and explore their impact on construction, evaluation, and translation of PRSs in diverse groups, with the intent of developing recommendations on best practices for selection and use of these measures.

Genotype harmonization

Genomic data available from studies and cohorts contributed by the study sites include whole-genome sequencing, genotyping array, genome-wide imputed data, and GSRs including GWAS summary statistics, allele frequencies, and genetic ancestry models—e.g., SNP-loadings from principal components analysis (PCA). Harmonizing these

data is essential to ensure consistent representation of variant information and to merge genomic information across the various studies and cohorts for GWAS, meta-analysis, and PRS development and evaluation. Harmonizing genetic data requires careful consideration to mitigate potential biases from heterogeneous genotyping technologies, imputation approaches, and data types.

The genotype harmonization WG was established to address the complexities of aggregating heterogeneous genetic datasets by formulating quality control (QC) and harmonization plans that include data standards, formats, and metadata requirements to enable consortium-wide genotype data collection and integration. All individual-level genotype data are available in build GRCh38 (liftover is implemented as necessary) and stored in VCF files (though multiple copies of the data are allowable in different builds or file formats). Individual-level genotype datasets are accompanied by variant quality metrics (e.g., allele frequencies, missing call rates, read depth, GQ scores, imputation quality, etc.) and sample quality metrics (e.g., missing call rates, heterozygosity, average read depth, coverage metrics, etc.). Array datasets are imputed to the NHLBI's TOPMed reference panel,[34–37] and accompanying info files with imputation quality metrics are provided. All GWAS summary statistics are available in text files that follow a specified data dictionary inspired by the data format in the GWAS Catalog.[38] The WG coordinates harmonization efforts across studies, and study sites upload genotype data to AnVIL data workspaces maintained by the CC for sharing with the broader consortium.

Ethical and social implications

Ethical and social considerations—such as when to use genetically inferred ancestry versus self-reported/ascribed non-genetic population descriptors (e.g., race, ethnicity, ancestry, background, tribal affiliation, primary language, and/or religious heritage) and how best to account for the effects of social, structural, and political factors—provide an important context for developing and evaluating PRS methods. Recent work suggests that evaluating PRSs within broad socially defined race/ethnicity categories such as "Hispanic/Latino" may mask disparities in PRS performance by genetic ancestry,[39] highlighting the importance of careful consideration of population designation and description in PRS development. Additionally, PRSs constructed using data stratified based on continental ancestry may have variable performance within such groups[12] and, problematically, foster misconceptions regarding the biological basis of racial identity.[40]

The population descriptors, social and ethical implications (POPSEI) WG was established to identify, investigate, and respond to ethical and social issues relevant to developing and evaluating PRS methods across groups of diverse genetic ancestry and raised by the integration of heterogeneous datasets (including those ascertained outside of the US) encompassed by the consortium. Given the ethical and methodological consequences of how populations are defined, described, and used, the POPSEI WG also discusses best practices and provides guidance and support to the consortium on the selection, use, and discussion of population descriptors in analyses and manuscripts (see initial consortium products section).

While PRIMED is not directly investigating the clinical implementation of specific PRSs, the data harmonization, methodological, and analytical approaches recommended by the consortium have broader scientific and social consequences for how PRSs will be implemented and interpreted in clinical settings and understood by affected individuals and the public. The POPSEI WG works closely with other PRIMED WGs and complementary groups within the eMERGE Network to anticipate and address potential consequences that might arise during PRS development, validation, and clinical translation. Examples include inappropriate use of PRSs in the commercial setting (pre-implantation diagnosis, embryo selection, nutrition counseling) or premature use of PRSs in the clinical setting. To surface social and ethical considerations across multiple contexts, the POPSEI WG is collecting use cases for a manuscript focused on common applications of PRSs in research and healthcare settings, highlighting best practices and practical concerns, with added emphasis on the conceptualization and use of ancestry and on downstream clinical implementation. Additionally, POPSEI is working on a collaborative manuscript with eMERGE investigators that addresses the considerations and concerns of PRS development and implementation, drawing from the experiences of both consortia. Addressing these issues will also require educational programs for the public, affected individuals, and providers. The consortium also provides public-facing educational materials on the consortium website and does regular public outreach through social media, managed by the CC.

PRS methods development

A primary goal of the PRIMED Consortium is to develop new methods to improve polygenic risk prediction, with emphasis on diverse ancestry and admixed groups. Investigators use innovations in statistical and population genetics to improve PRS methods, develop new PRS models, and explore PRS-trait/disease associations across different age groups, environmental contexts, and diverse groups. The methods WG was established as a central forum in the consortium where members review existing literature and methods, as well as introduce, develop, collaborate on, and disseminate results from methods developed across the consortium. The methods WG has identified key methodological areas for discussion, development, and innovation[41]; a brief description follows.

Methodological innovations

PRIMED investigators are testing a broad range of PRS methods including pruning and thresholding (P&T) PRSs,

which include a set of non-correlated genetic variants meeting a certain significance threshold, and genome-wide PRSs, which may include thousands to millions of variants.[3] The latter may perform better for highly polygenic traits and typically involve "shrinkage" of regression coefficients using Bayesian methods or "penalized" methods such as Ridge regression or Lasso.[42–44] The consortium is investigating numerous approaches to improve PRS performance that include incorporating functional annotations to improve the identification of causal variants and thus transferability,[45–48] incorporating rare variants, statistical fine-mapping to overcome some of the barriers related to linkage disequilibrium (LD) differences across groups,[46,49] and the use of genetic correlation and pleiotropy with related traits.[50–52]

A major focus of PRIMED has been to develop different approaches to incorporate genetic ancestry information. PRS methods that integrate GWAS results from groups of diverse ancestries provide more accurate variant effect size estimates by sharing information and leveraging ancestral differences in LD patterns across groups to fine map the likely causal variants. These multi-ancestry methods improve predictive accuracy compared with methods utilizing GWAS results from a single ancestry source.[42] PRIMED is also developing PRSs that are applicable to recently admixed individuals, such as those who self-identify as African Americans and Latinos, in whom admixture proportions can vary significantly. PRS methods that utilize measures of local genetic ancestry to allow for modeling of different effect sizes by inferred ancestral haplotype[53] are being explored as a means of estimating PRSs in the background of varying degrees of admixture across groups/individuals.[54] Furthermore, as admixture is pervasive and individuals may not discretely map onto distinct continental ancestry groups, recently developed methods that account for continuous representations of genetic ancestry in PRSs may prove useful. For example, PRIMED investigators[55] have demonstrated that PRS accuracy decreases individual-to-individual along the continuum of genetic ancestries, even within traditionally labeled "homogeneous" genetic ancestry groups. This trend can be represented as genetic distance from the PRS training data, a measure negatively correlated with PRS accuracy. These findings motivate the use of continuous genetic similarity/distance in PRS interpretation, rather than discrete genetic ancestry clusters.

Another important methodological consideration has been developing approaches to combine non-genetic information, such as environmental exposures and SDoH, with PRSs to create comprehensive risk prediction models. This includes statistical models that contain joint effects, model potential mediation pathways, or investigate interactions.[56–58] PRS associations with a trait may vary in different contexts, such as across age groups,[59] across strata defined by clinical variables such as adiposity or smoking,[60] or across strata defined by different environmental, cultural, or social factors.[32,33] PRIMED investigators are designing and performing analyses to evaluate such contextual interactions[61] and have introduced an approach (CalPred) to account for variation in PRS-trait associations across contexts by modeling all contexts jointly to produce prediction intervals that vary across contexts.[62] Because absolute risk of disease may be the most meaningful in informing clinical decision-making, PRIMED investigators are also developing methods to include PRSs in the computation of such risk estimates using existing validated clinical algorithms or based on epidemiological indices.[63]

Genetic ancestry inference

The methods WG established a focused ancestry sub-WG to develop analysis workflows for harmonized genetic ancestry inference based on genetic similarity,[64] with a focus on identifying the best methods and reference panels to support PRS development and evaluation. These genetic ancestry workflows are applied across study datasets by utilizing the harmonized individual-level genotype data. PCA is applied to reference panel genotype data to calculate SNP-loadings for ancestry PCs, and all study datasets are projected onto this harmonized PC space. Similarly, global[65] and local[66–68] ancestry inference models are trained on reference panel genotype data and then applied to the harmonized genotype data for each study.

The choice of reference panel data is critical to the performance of genetic ancestry analysis and has therefore been a point of focus. From a global perspective, there are limitations to existing genotype reference panels such as large gaps in representation for many groups, including African, Middle Eastern, Native American, and South Asian. This sub-WG is currently working on comparisons of genetic ancestry inference based on genetic similarity measures obtained with commonly used reference panels (e.g., 1000 Genomes,[69] Human Genome Diversity Project,[70] PAGE global reference panel[71]) and evaluation of their downstream impact on PRS development and performance. Additionally, the sub-WG is constructing and evaluating an improved reference panel with better representation of global genetic diversity by combining public and private datasets. Genetic ancestry inference models developed with this reference panel will be shared publicly.

Standardized PRS evaluation, comparison, and reporting

PRIMED is building on existing frameworks to organize standardized comparisons across different PRS methods and models using harmonized data contributed to AnVIL as well as via analysis protocols distributed to biobank partners. Given the breadth of investigator expertise and access to data from many diverse studies and biobanks, PRIMED is uniquely positioned to perform standardized comparisons of PRS methods and models across numerous traits, contexts, and groups. To encourage rigor in PRS development, validation, evaluation, and reporting, and to facilitate benchmarking that ensures robust PRSs are available for diverse groups, PRIMED has adopted the polygenic risk score reporting standards (PRS-RS)[72] developed jointly by

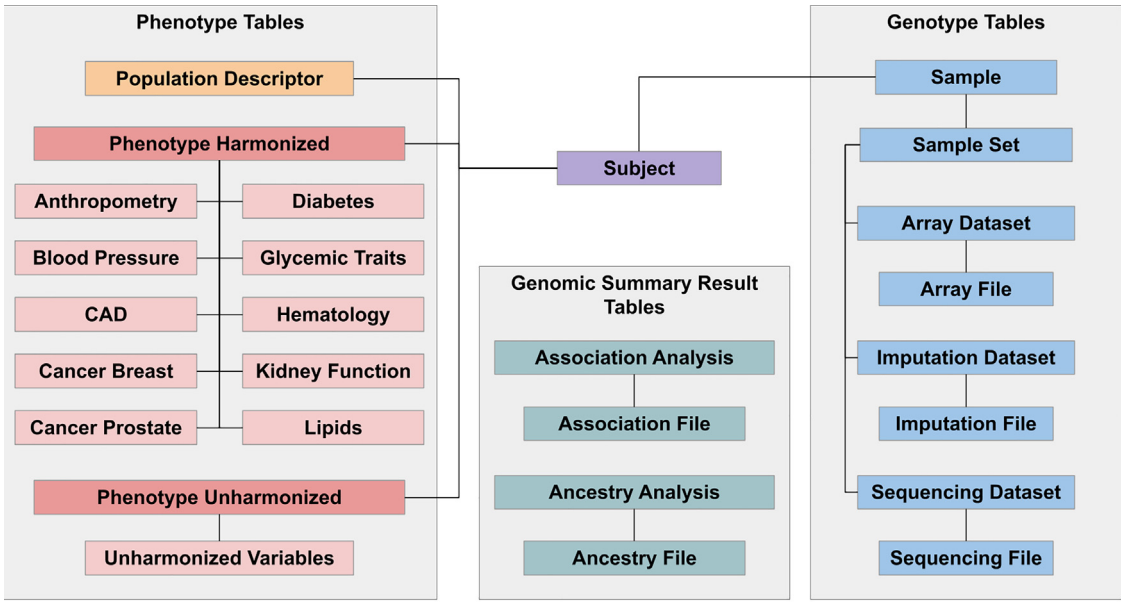

Figure 2. The PRIMED common data model

Each colored box represents a table, and lines represent links between tables. The subject table (purple) captures information on each subject/participant and is the linking point to the other components of the data model. The population descriptor table (orange) captures detailed population descriptor information on each subject. The phenotype dataset tables (dark pink) provide phenotype dataset metadata and provide links to the phenotype domain tables (light pink) which are tabular data files containing individual-level phenotype data for specified harmonized variables in a wide-format familiar to cohort studies; unharmonized phenotype data can be shared in tabular data files pre-harmonization. The genotype tables (blue) capture sample metadata and group samples into sets, corresponding to genotype datasets. Genotyping technology-specific dataset tables provide metadata describing key features of the genotype dataset, and file tables provide file paths to individual-level genotype data files (e.g., VCFs) linked to datasets. Genomic summary result tables (teal) include analysis tables that capture metadata about analyses that generated the GSRs and file tables provide file paths to tabular data files containing the GSR data linked to analyses.

ClinGen[22] and the PGS Catalog.[25] Further, building off of the PRS-RS, the methods WG aims to develop improved metrics for PRS evaluation and benchmarking that are applicable to diverse groups and individuals.

Initial consortium products

PRIMED common data model

To enable rapid data harmonization and analysis for non-EHR linked cohort studies, the consortium developed a data model (Figure 2) that stores phenotype data following the more familiar structure of cohort data (i.e., text files where rows are participant observations and columns are variables). For EHR sources, data are extracted and transformed into cohort datasets and subsequently formatted to the PRIMED common data model in an analysis-ready structure. This proved most practical to support analyses on AnVIL and simplify data sharing agreements for EHR-based studies and biobanks, as they could extract a harmonized cohort dataset with the required variables, rather than share their entire EHR data. A future effort to support other programs and consortia would be to design cloud-based platforms that natively support the deidentified observational medical outcomes partnership (OMOP, see web resources) data model, which is used by the All of Us program and UK Biobank. The PRIMED common data model also captures structured metadata that ensures that individual-level genotype datasets and GSR can easily be combined across studies for subsequent analyses. File paths to data files in the AnVIL data workspaces are provided to enable programmatically passing files into analysis workflows. The PRIMED common data model is available on GitHub (see web resources) encoded as JSON files. The CC maintains version updates to the data model as well as companion workflow description language (WDL) workflows in a collection on Dockstore (see web resources) that can be used to validate that data uploaded to AnVIL workspaces conform to the specifications.

Adaptation of recent guidelines for population descriptors to global cohorts

The PRIMED POPSEI WG produced two key products, initially for internal use only and now being prepared for dissemination: (1) recommendations for the use of population descriptors in PRIMED and (2) a data model designed to accommodate heterogeneity of descriptor availability and use among the extant data collections being aggregated in the consortium. The WG recommendations incorporate those recently advanced by the National Academies of Sciences, Engineering, and Medicine (NASEM, see web resources), including how to report inferred genetic ancestry or similarity and distinguish those inferences from self-reported (or ascriptively assigned) identifiers such as race/ethnicity, the latter associated with social,

cultural, environmental, and political factors. Additionally, the WG developed a component of the PRIMED data model (Figure 2) that provides a flexible approach to collating and combining datasets with different population descriptors. This model uses the NASEM framework for distinguishing descriptors (e.g., "country of recruitment") from labels (e.g., "US," "France"; see also Table S3), which encourages analysts to describe and combine groups in a consistent way. Notably, this data model allows an individual to have multiple descriptors and labels and does not enforce a single harmonization strategy for the entire consortium, which would be impractical given the heterogeneity of data sources in PRIMED. Furthermore, retention of detailed population information and deferring harmonization until the analysis stage provides flexibility and is intended to reduce the tendency to conceptualize all populations in terms of continental ancestry groups. GSRs are reported with a single population descriptor used for analysis and multiple labels to reflect the diversity of individuals included. Both products could inform the practices of other consortia working with similar biomedical data.

Analytic tools

PRIMED members are developing analysis workflows written in WDL that implement tools and pipelines for PRS methods and related analyses such as genetic ancestry inference and GWAS. The methods WG has developed an integrated toolkit and pipeline (admix-kit) for generating simulated admixed genotype data[73] and has shared synthetic cohorts in an AnVIL workspace for use in methods development and testing. Additionally, the harmonization WGs have developed data processing workflows for QC, harmonization, and imputation. The workflows for analysis, data processing, and simulation developed in PRIMED are openly available to the broader research community via the PRIMED Organization Dockstore and/or GitHub repositories (see Table S4 and web resources), from which they can be deployed directly on AnVIL or in other computing environments compatible with WDL.

Future directions

Increasing representation of diverse groups in GWASs requires significant investment in community engagement and cohort assembly.[24] In this perspective we provide insights into the PRIMED Consortium design and organization of activities to facilitate initiatives that aim to reduce inequity in genomic medicine and highlight early results from consortium activities. Future directions of the consortium include continuing collaborations among consortium study sites as well as with outside groups to increase the size of available datasets for individuals of diverse genetic ancestries. Collaborative projects with the eMERGE Network include incorporating PRSs into multivariable models that include environmental factors/SDoH and other clinical risk factors, generating absolute disease risk estimates in individuals from diverse groups, articulating ethical and social considerations for PRS development and implementation, and adapting to the availability of new PRSs as well as revisions of prior PRS interpretation. Methods development will continue to be an active area for innovation, and examples of proposed/ongoing work include combining dynamically updated lifetime risk factor trajectories with PRSs to better inform lifetime risk, incorporating information from relevant endophenotypes to improve PRS predictive performance for outcomes, and developing methods to generate PRSs that integrate rare and common variants. The consortium is also exploring ways to measure and integrate multiple axes of diversity, for example by developing reference panels from expanded and well-curated diverse ancestry groups that better represent the full spectrum of global genetic diversity and can be dynamically updated to incorporate future data; using genetic distance and genetic neighbors to estimate uncertainty in individual predictions from PRSs, regardless of genetically inferred ancestry and self-identified ethnicity; and benchmarking different PRS methods that utilize local or global genetic ancestry measures. Finally, the consortium will continue to explore current practices, ongoing issues, and potential solutions for the use of race, ethnicity, and genetic ancestry in PRSs through the lens of ethical and social implications.

Conclusion

The PRIMED Consortium has established infrastructure for secure, collaborative genomic research across datasets, institutions, and investigators to accelerate the development of methods for generating PRSs that can quantify polygenic risk across diverse genetic ancestries, including individuals who are recently admixed. Early deliverables of the consortium include methods for data sharing on AnVIL, a secure cloud-based environment; development of a common data model to harmonize phenotype and genotype data from cohort studies as well as EHRs; adaptation of recent guidelines for population descriptors to global cohorts; and sharing of analyses and tools. The consortium serves as a template for multisite collectives that aim to lessen health and healthcare disparities and extend advances in genomic medicine to diverse groups in the US and across the world.[24]

Consortia

The PRIMED Consortium banner authors are Sally Adebamowo, Clement Adebamowo, Nicholette Allred, Paul Auer, Jennifer Below, Palwende Romuald Boua, Kristin Boulier, Michael Bowers, Joseph Breeyear, Nilanjan Chatterjee, Tinashe Chikowore, Jaewon Choi, Ananyo Choudhury, Matthew Conomos, David Conti, Nancy Cox, Sinead Cullina, Burcu Darst, Aaron Deutsch, Yi Ding, Todd Edwards, Eleazar Eskin, Segun Fatumo, Jose Florez, Nelson Freimer, Stephanie Fullerton, Tian Ge, Daniel Geschwind, Chris

Gignoux, Stephanie Gogarten, Mark Goodarzi, Xiuqing Guo, Christopher Haiman, Neil Hanchard, Scott Hazelhurst, Ben Heavner, Susan Heckbert, Jibril Hirbo, Whitney Hornsby, Kangcheng Hou, Qinqin Huang, Alicia Huerta, Guoqian Jiang, Katherine Johnston, Linda Kachuri, Takashi Kadowaki, Abram Bunya Kamiza, Eimear Kenny, Sarah Kerns, Alyna Khan, Joohyun Kim, Iain Konigsberg, Charles Kooperberg, Matt Kosel, Peter Kraft, Iftikhar Kullo, Soo-Heon Kwak, Leslie Lange, Ethan Lange, Loic Le Marchand, Hyunsuk Lee, Aaron Leong, Yun Li, Meng Lin, Kirk Lohmueller, Ruth Loos, Kevin Lu, Ravi Mandia, Alisa Manning, Alicia Martin, Iman Martin, Hilary Martin, Rasika Mathias, James Meigs, Josep Mercader, Rachel Mester, Mariah Meyer, Tyne Miller-Fleming, Braxton Mitchell, Nicola Mulder, Jie Na, Pradeep Natarajan, Sarah Nelson, Maggie Ng, Kristjan Norland, Loes Olde Loohuis, Suna Onengut-Gumuscu, Ebuka Oneyobi, Roel Ophoff, Paivi Pajukanta, Bogdan Pasaniuc, Aniruddh Patel, Ulrike Peters, Jimmy Phuong, Michael Preuss, Bruce Psaty, Laura Raffield, Michele Ramsay, Alexander Reiner, Kenneth Rice, Stephen Rich, Jerome Rotter, Bryce Rowan, Robb Rowley, Yunfeng Ruan, Lori Sakoda, Siram Sankararaman, Dan Schaid, Dan Schrider, Philip Schroeder, Ruhoilah Shemirani, Jonathan Shortt, Megan Shuey, Xueling Sim, Roelof A.J. Smit, Johanna Smith, Lucia Sobrin, Lauren Stalbow, Adrienne Stilp, Daniel Stram, Ken Suzuki, Lukasz Szczerbinski, Ran Tao, Bamidele Tayo, Timothy Thornton, Buu Truong, Teresa Tusie, Miriam Udler, David van Heel, Luciana B. Vargas, Vidhya Venkateswaran, Ying Wang, Jennifer Wessel, Laura Wiley, Lynne Wilkens, Riley Wilson, John Witte, Genevieve Wojcik, Quenna Wong, Toshimasa Yamauchi, Lisa Yanek, Yue Yu, Haoyu Zhang, Yuji Zhang, and Michael Zhong.

Data and code availability

The workflows developed by PRIMED (see Table S4) are available in the PRIMED Dockstore Organization: https://dockstore.org/organizations/PRIMED. Each workflow listed in the Dockstore organization has a link to the corresponding GitHub repository with code available. No new source datasets are associated with this paper. dbGaP accessions for the pre-existing source data associated with this paper are listed in Table S3. When permissible under NIH data sharing policy, PRS models and evaluations generated by the consortium are deposited into the PGS Catalog. Links to PGS Catalog records associated with PRIMED publications are available on the PRIMED website: https://primedconsortium.org/publications/published.

Acknowledgments

The authors would like to thank Drs. Lucia Hindorff and Teri Manolio for their expertise and assistance with the PRIMED Consortium activities and initial preparation of this manuscript. The PRIMED Consortium is funded by the following grants from the National Human Genome Research Institute and the National Cancer Institute: CARDINAL (U01HG011717), CAPE (U01HG011715), D-PRISM (U01HG011723), EPIC-PRS (U01HG011720), FFAIRR-PRS (U01HG011719), PRIMED-Cancer (U01CA261339), PREVENT (U01HG011710), and Coordinating Center (U01HG011697). The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Author contributions

I.J.K., M.P.C., S.C.N., I.M., and P.N. cowrote the initial draft and revised the manuscript. S.N.A., A.C., and S.M.F. contributed to the initial draft. S.C.N., D.C., W.E.H., B.H., E.E.K., A.V.K., Y.L., J.M.M., M.N., L.M.R., A.R., R.R., D.S., A.S., K.W., R.W., and J.S.W., revised the manuscript. M.P.C. and S.M.G. contributed to visualization and revised the manuscript. A.K. contributed to visualization.

Declaration of interests

P.N. reports research grants from Allelica, Apple, Amgen, Boston Scientific, Genentech/Roche, and Novartis; personal fees from Allelica, Apple, AstraZeneca, Blackstone Life Sciences, Foresite Labs, Genentech/Roche, GV, HeartFlow, Magnet Biomedicine, and Novartis; scientific advisory board membership of Esperion Therapeutics, Preciseli, and TenSixteen Bio; scientific co-founder of TenSixteen Bio; equity in MyOme, Preciseli, and TenSixteen Bio; and spousal employment at Vertex Pharmaceuticals, all unrelated to the present work. E.E.K. has received personal fees from Regeneron Pharmaceuticals, 23&Me, Allelica, and Illumina; has received research funding from Allelica; and serves on the advisory boards for Encompass Biosciences, Overtone, and Galateo Bio.

Web resources

Dockstore, PRIMED Consortium, https://dockstore.org/organizations/PRIMED

Dockstore, PRIMED Consortium data validation, https://dockstore.org/organizations/PRIMED/collections/data-validation

GitHub, PRIMED data model, https://github.com/UW-GAC/primed_data_models

NASEM report, Population Descriptors in Genetics and Genomics Research, https://www.nationalacademies.org/our-work/use-of-race-ethnicity-and-ancestry-as-population-descriptors-in-genomics-research

NIH GSR sharing policy, https://grants.nih.gov/grants/guide/notice-files/NOT-OD-19-023.html

OMOP, https://ohdsi.github.io/CommonDataModel/index.html

PRIMED Consortium, https://primedconsortium.org/

PRIMED Consortium, affiliate membership policy, https://primedconsortium.org/about/policies/affiliate-membership-policy

PRIMED Consortium, data sharing policy, https://primedconsortium.org/about/policies/data-sharing-policy

PRIMED Consortium, publications, https://primedconsortium.org/publications/published

RFA-HG-20-001, Polygenic Risk Score (PRS) Methods and Analysis for Populations of Diverse Ancestry Study Sites, https://grants.nih.gov/grants/guide/rfa-files/RFA-HG-20-001.html

Supplemental information

Supplemental information can be found online at https://doi.org/10.1016/j.ajhg.2024.10.010.

References

1. Green, E.D., Gunter, C., Biesecker, L.G., Di Francesco, V., Easter, C.L., Feingold, E.A., Felsenfeld, A.L., Kaufman, D.J., Ostrander, E.A., Pavan, W.J., et al. (2020). Strategic vision for improving human health at The Forefront of Genomics. Nature *586*, 683–692. https://doi.org/10.1038/s41586-020-2817-4.
2. Lemke, A.A., Esplin, E.D., Goldenberg, A.J., Gonzaga-Jauregui, C., Hanchard, N.A., Harris-Wai, J., Ideozu, J.E., Isasi, R., Landstrom, A.P., Prince, A.E.R., et al. (2022). Addressing underrepresentation in genomics research through community engagement. Am. J. Hum. Genet. *109*, 1563–1571. https://doi.org/10.1016/j.ajhg.2022.08.005.
3. Kullo, I.J., Lewis, C.M., Inouye, M., Martin, A.R., Ripatti, S., and Chatterjee, N. (2022). Polygenic scores in biomedical research. Nat. Rev. Genet. *23*, 524–532. https://doi.org/10.1038/s41576-022-00470-z.
4. Mars, N., Koskela, J.T., Ripatti, P., Kiiskinen, T.T.J., Havulinna, A.S., Lindbohm, J.V., Ahola-Olli, A., Kurki, M., Karjalainen, J., Palta, P., et al. (2020). Polygenic and clinical risk scores and their impact on age at onset and prediction of cardiometabolic diseases and common cancers. Nat. Med. *26*, 549–557. https://doi.org/10.1038/s41591-020-0800-0.
5. Manikpurage, H.D., Eslami, A., Perrot, N., Li, Z., Couture, C., Mathieu, P., Bossé, Y., Arsenault, B.J., and Thériault, S. (2021). Polygenic risk score for coronary artery disease improves the prediction of early-onset myocardial infarction and mortality in men. Circ. Genom. Precis. Med. *14*, e003452. https://doi.org/10.1161/circgen.121.003452.
6. 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., and Kathiresan, S. (2018). Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations. Nat. Genet. *50*, 1219–1224. https://doi.org/10.1038/s41588-018-0183-z.
7. Linder, J.E., Allworth, A., Bland, H.T., Caraballo, P.J., Chisholm, R.L., Clayton, E.W., Crosslin, D.R., Dikilitas, O., DiVietro, A., Esplin, E.D., et al. (2023). Returning integrated genomic risk and clinical recommendations: The eMERGE study. Genet. Med. *25*, 100006. https://doi.org/10.1016/j.gim.2023.100006.
8. O'Sullivan, J.W., Raghavan, S., Marquez-Luna, C., Luzum, J.A., Damrauer, S.M., Ashley, E.A., O'Donnell, C.J., Willer, C.J., Natarajan, P.; and American Heart Association Council on Genomic and Precision Medicine; Council on Clinical Cardiology; Council on Arteriosclerosis, Thrombosis and Vascular Biology; Council on Cardiovascular Radiology and Intervention; Council on Lifestyle and Cardiometabolic Health; and Council on Peripheral Vascular Disease (2022). Polygenic risk scores for cardiovascular disease: A scientific statement from the American Heart Association. Circulation *146*, e93–e118. https://doi.org/10.1161/CIR.0000000000001077.
9. Abu-El-Haija, A., Reddi, H.V., Wand, H., Rose, N.C., Mori, M., Qian, E., Murray, M.F.; and ACMG Professional Practice and Guidelines Committee (2023). The clinical application of polygenic risk scores: A points to consider statement of the American College of Medical Genetics and Genomics (ACMG). Genet. Med. *25*, 100803. https://doi.org/10.1016/j.gim.2023.100803.
10. Dikilitas, O., Schaid, D.J., Kosel, M.L., Carroll, R.J., Chute, C.G., Hakonarson, H., Jarvik, G.P., Denny, J.A., Fedotov, A., Feng, Q., et al. (2020). Predictive utility of polygenic risk scores for coronary heart disease in three major racial and ethnic groups. Am. J. Hum. Genet. *106*, 707–716. https://doi.org/10.1016/j.ajhg.2020.04.002.
11. Fatumo, S., Sathan, D., Samtal, C., Isewon, I., Tamuhla, T., Soremekun, C., Jafali, J., Panji, S., Tiffin, N., and Fakim, Y.J. (2023). Polygenic risk scores for disease risk prediction in Africa: current challenges and future directions. Genome Med. *15*, 87. https://doi.org/10.1186/s13073-023-01245-9.
12. Kamiza, A.B., Toure, S.M., Vujkovic, M., Machipisa, T., Soremekun, O.S., Kintu, C., Corpas, M., Pirie, F., Young, E., Gill, D., et al. (2022). Transferability of genetic risk scores in African populations. Nat. Med. *28*, 1163–1166. https://doi.org/10.1038/s41591-022-01835-x.
13. Duncan, L., Shen, H., Gelaye, B., Meijsen, J., Ressler, K., Feldman, M., Peterson, R., and Domingue, B. (2019). Analysis of polygenic risk score usage and performance in diverse human populations. Nat. Commun. *10*, 3328. https://doi.org/10.1038/s41467-019-11112-0.
14. Wang, Y., Tsuo, K., Kanai, M., Neale, B.M., and Martin, A.R. (2022). Challenges and opportunities for developing more generalizable polygenic risk scores. Annu. Rev. Biomed. Data Sci. *5*, 293–320. https://doi.org/10.1146/annurev-biodatasci-111721-074830.
15. Hou, K., Ding, Y., Xu, Z., Wu, Y., Bhattacharya, A., Mester, R., Belbin, G.M., Buyske, S., Conti, D.V., Darst, B.F., et al. (2023). Causal effects on complex traits are similar for common variants across segments of different continental ancestries within admixed individuals. Nat. Genet. *55*, 549–558. https://doi.org/10.1038/s41588-023-01338-6.
16. Martin, A.R., Gignoux, C.R., Walters, R.K., Wojcik, G.L., Neale, B.M., Gravel, S., Daly, M.J., Bustamante, C.D., and Kenny, E.E. (2017). Human demographic history impacts genetic risk prediction across diverse populations. Am. J. Hum. Genet. *100*, 635–649. https://doi.org/10.1016/j.ajhg.2017.03.004.
17. Wang, Y., Namba, S., Lopera, E., Kerminen, S., Tsuo, K., Läll, K., Kanai, M., Zhou, W., Wu, K.H., Favé, M.J., et al. (2023). Global Biobank analyses provide lessons for developing polygenic risk scores across diverse cohorts. Cell Genom. *3*, 100241. https://doi.org/10.1016/j.xgen.2022.100241.
18. Zhou, W., Kanai, M., Wu, K.H.H., Rasheed, H., Tsuo, K., Hirbo, J.B., Wang, Y., Bhattacharya, A., Zhao, H., Namba, S., et al. (2022). Global Biobank Meta-analysis Initiative: Powering genetic discovery across human disease. Cell Genom. *2*, 100192. https://doi.org/10.1016/j.xgen.2022.100192.
19. Martin, A.R., Kanai, M., Kamatani, Y., Okada, Y., Neale, B.M., and Daly, M.J. (2019). Clinical use of current polygenic risk scores may exacerbate health disparities. Nat. Genet. *51*, 584–591. https://doi.org/10.1038/s41588-019-0379-x.
20. Hindorff, L.A., Bonham, V.L., Brody, L.C., Ginoza, M.E.C., Hutter, C.M., Manolio, T.A., and Green, E.D. (2018). Prioritizing diversity in human genomics research. Nat. Rev. Genet. *19*, 175–185. https://doi.org/10.1038/nrg.2017.89.
21. Schatz, M.C., Philippakis, A.A., Afgan, E., Banks, E., Carey, V.J., Carroll, R.J., Culotti, A., Ellrott, K., Goecks, J., Grossman, R.L., et al. (2022). Inverting the model of genomics data sharing with the NHGRI Genomic Data Science Analysis, Visualization, and Informatics Lab-space. Cell Genom. *2*, 100085. https://doi.org/10.1016/j.xgen.2021.100085.
22. Rehm, H.L., Berg, J.S., Brooks, L.D., Bustamante, C.D., Evans, J.P., Landrum, M.J., Ledbetter, D.H., Maglott, D.R., Martin, C.L., Nussbaum, R.L., et al. (2015). ClinGen–the Clinical

Genome Resource. N. Engl. J. Med. *372*, 2235–2242. https://doi.org/10.1056/NEJMsr1406261.

23. All of Us Research Program Investigators, Denny, J.C., Rutter, J.L., Goldstein, D.B., Philippakis, A., Smoller, J.W., Jenkins, G., and Dishman, E. (2019). The "All of Us" Research Program. N. Engl. J. Med. *381*, 668–676. https://doi.org/10.1056/NEJMsr1809937.
24. Kullo, I.J. (2024). Promoting equity in polygenic risk assessment through global collaboration. Nat. Genet. *56*, 1780–1787. https://doi.org/10.1038/s41588-024-01843-2.
25. Lambert, S.A., Gil, L., Jupp, S., Ritchie, S.C., Xu, Y., Buniello, A., McMahon, A., Abraham, G., Chapman, M., Parkinson, H., et al. (2021). The Polygenic Score Catalog as an open database for reproducibility and systematic evaluation. Nat. Genet. *53*, 420–425. https://doi.org/10.1038/s41588-021-00783-5.
26. Sollis, E., Mosaku, A., Abid, A., Buniello, A., Cerezo, M., Gil, L., Groza, T., Güneş, O., Hall, P., Hayhurst, J., et al. (2023). The NHGRI-EBI GWAS Catalog: knowledgebase and deposition resource. Nucleic Acids Res. *51*. D977-d985. https://doi.org/10.1093/nar/gkac1010.
27. Kirby, J.C., Speltz, P., Rasmussen, L.V., Basford, M., Gottesman, O., Peissig, P.L., Pacheco, J.A., Tromp, G., Pathak, J., Carrell, D.S., et al. (2016). PheKB: a catalog and workflow for creating electronic phenotype algorithms for transportability. J. Am. Med. Inform. Assoc. *23*, 1046–1052. https://doi.org/10.1093/jamia/ocv202.
28. Ganesh, S.K., Chasman, D.I., Larson, M.G., Guo, X., Verwoert, G., Bis, J.C., Gu, X., Smith, A.V., Yang, M.L., Zhang, Y., et al. (2014). Effects of long-term averaging of quantitative blood pressure traits on the detection of genetic associations. Am. J. Hum. Genet. *95*, 49–65. https://doi.org/10.1016/j.ajhg.2014.06.002.
29. Stilp, A.M., Emery, L.S., Broome, J.G., Buth, E.J., Khan, A.T., Laurie, C.A., Wang, F.F., Wong, Q., Chen, D., D'Augustine, C.M., et al. (2021). A system for phenotype harmonization in the National Heart, Lung, and Blood Institute Trans-Omics for Precision Medicine (TOPMed) Program. Am. J. Epidemiol. *190*, 1977–1992. https://doi.org/10.1093/aje/kwab115.
30. Oelsner, E.C., Balte, P.P., Cassano, P.A., Couper, D., Enright, P.L., Folsom, A.R., Hankinson, J., Jacobs, D.R., Jr., Kalhan, R., Kaplan, R., et al. (2018). Harmonization of respiratory data From 9 US population-based cohorts: The NHLBI Pooled Cohorts Study. Am. J. Epidemiol. *187*, 2265–2278. https://doi.org/10.1093/aje/kwy139.
31. Faure, E., Danjou, A.M.N., Clavel-Chapelon, F., Boutron-Ruault, M.-C., Dossus, L., and Fervers, B. (2017). Accuracy of two geocoding methods for geographic information system-based exposure assessment in epidemiological studies. Environ. Health. *16*, 15. https://doi.org/10.1186/s12940-017-0217-5.
32. Cromer, S.J., Lakhani, C.M., Mercader, J.M., Majarian, T.D., Schroeder, P., Cole, J.B., Florez, J.C., Patel, C.J., Manning, A.K., Burnett-Bowie, S.A.M., et al. (2023). Association and interaction of genetics and area-level socioeconomic factors on the prevalence of type 2 diabetes and obesity. Diabetes Care *46*, 944–952. https://doi.org/10.2337/dc22-1954.
33. Norland, K., Schaid, D.J., Naderian, M., Na, J., and Kullo, I.J. (2024). Associations of self-reported race, social determinants of health, and polygenic risk with coronary heart disease. Preprint at medRxiv, 2024.01.10.24301105. https://doi.org/10.1101/2024.01.10.24301105.
34. Taliun, D., Harris, D.N., Kessler, M.D., Carlson, J., Szpiech, Z.A., Torres, R., Taliun, S.A.G., Corvelo, A., Gogarten, S.M., Kang, H.M., et al. (2021). Sequencing of 53,831 diverse genomes from the NHLBI TOPMed Program. Nature *590*, 290–299. https://doi.org/10.1038/s41586-021-03205-y.
35. Kowalski, M.H., Qian, H., Hou, Z., Rosen, J.D., Tapia, A.L., Shan, Y., Jain, D., Argos, M., Arnett, D.K., Avery, C., et al. (2019). Use of >100,000 NHLBI Trans-Omics for Precision Medicine (TOPMed) Consortium whole genome sequences improves imputation quality and detection of rare variant associations in admixed African and Hispanic/Latino populations. PLoS Genet. *15*, e1008500. https://doi.org/10.1371/journal.pgen.1008500.
36. Hanks, S.C., Forer, L., Schönherr, S., LeFaive, J., Martins, T., Welch, R., Gagliano Taliun, S.A., Braff, D., Johnsen, J.M., Kenny, E.E., et al. (2022). Extent to which array genotyping and imputation with large reference panels approximate deep whole-genome sequencing. Am. J. Hum. Genet. *109*, 1653–1666. https://doi.org/10.1016/j.ajhg.2022.07.012.
37. Huerta-Chagoya, A., Schroeder, P., Mandla, R., Deutsch, A.J., Zhu, W., Petty, L., Yi, X., Cole, J.B., Udler, M.S., Dornbos, P., et al. (2023). The power of TOPMed imputation for the discovery of Latino-enriched rare variants associated with type 2 diabetes. Diabetologia *66*, 1273–1288. https://doi.org/10.1007/s00125-023-05912-9.
38. MacArthur, J.A.L., Buniello, A., Harris, L.W., Hayhurst, J., McMahon, A., Sollis, E., Cerezo, M., Hall, P., Lewis, E., Whetzel, P.L., et al. (2021). Workshop proceedings: GWAS summary statistics standards and sharing. Cell Genom. *1*, 100004. https://doi.org/10.1016/j.xgen.2021.100004.
39. Clarke, S.L., Huang, R.D.L., Hilliard, A.T., Tcheandjieu, C., Lynch, J., Damrauer, S.M., Chang, K.M., Tsao, P.S., and Assimes, T.L. (2022). Race and ethnicity stratification for polygenic risk score analyses may mask disparities in Hispanics. Circulation *146*, 265–267. https://doi.org/10.1161/circulationaha.122.059162.
40. Lewis, A.C.F., Molina, S.J., Appelbaum, P.S., Dauda, B., Di Rienzo, A., Fuentes, A., Fullerton, S.M., Garrison, N.A., Ghosh, N., Hammonds, E.M., et al. (2022). Getting genetic ancestry right for science and society. Science *376*, 250–252. https://doi.org/10.1126/science.abm7530.
41. Kachuri, L., Chatterjee, N., Hirbo, J., Schaid, D.J., Martin, I., Kullo, I.J., Kenny, E.E., Pasaniuc, B., Polygenic Risk Methods in Diverse Populations PRIMED Consortium Methods Working Group, Witte, J.S., and Ge, T. (2024). Principles and methods for polygenic risk scores (PRS) across global populations. Nat. Rev. Genet. *25*, 8–25. https://doi.org/10.1038/s41576-023-00637-2.
42. Ruan, Y., Lin, Y.F., Feng, Y.C.A., Chen, C.Y., Lam, M., Guo, Z., Stanley Global Asia Initiatives, He, L., Sawa, A., Martin, A.R., et al. (2022). Improving polygenic prediction in ancestrally diverse populations. Nat. Genet. *54*, 573–580. https://doi.org/10.1038/s41588-022-01054-7.
43. Privé, F., Arbel, J., and Vilhjálmsson, B.J. (2020). LDpred2: better, faster, stronger. Bioinformatics *36*, 5424–5431. https://doi.org/10.1093/bioinformatics/btaa1029.
44. Ge, T., Chen, C.Y., Ni, Y., Feng, Y.C.A., and Smoller, J.W. (2019). Polygenic prediction via Bayesian regression and continuous shrinkage priors. Nat. Commun. *10*, 1776. https://doi.org/10.1038/s41467-019-09718-5.

45. Márquez-Luna, C., Gazal, S., Loh, P.R., Kim, S.S., Furlotte, N., Auton, A., 23andMe Research Team, and Price, A.L. (2021). Incorporating functional priors improves polygenic prediction accuracy in UK Biobank and 23andMe data sets. Nat. Commun. *12*, 6052. https://doi.org/10.1038/s41467-021-25171-9.
46. Weissbrod, O., Hormozdiari, F., Benner, C., Cui, R., Ulirsch, J., Gazal, S., Schoech, A.P., van de Geijn, B., Reshef, Y., Márquez-Luna, C., et al. (2020). Functionally informed fine-mapping and polygenic localization of complex trait heritability. Nat. Genet. *52*, 1355–1363. https://doi.org/10.1038/s41588-020-00735-5.
47. Amariuta, T., Ishigaki, K., Sugishita, H., Ohta, T., Koido, M., Dey, K.K., Matsuda, K., Murakami, Y., Price, A.L., Kawakami, E., et al. (2020). Improving the trans-ancestry portability of polygenic risk scores by prioritizing variants in predicted cell-type-specific regulatory elements. Nat. Genet. *52*, 1346–1354. https://doi.org/10.1038/s41588-020-00740-8.
48. Hu, Y., Lu, Q., Powles, R., Yao, X., Yang, C., Fang, F., Xu, X., and Zhao, H. (2017). Leveraging functional annotations in genetic risk prediction for human complex diseases. PLoS Comput. Biol. *13*, e1005589. https://doi.org/10.1371/journal.pcbi.1005589.
49. Schaid, D.J., Chen, W., and Larson, N.B. (2018). From genome-wide associations to candidate causal variants by statistical fine-mapping. Nat. Rev. Genet. *19*, 491–504. https://doi.org/10.1038/s41576-018-0016-z.
50. Urbut, S.M., Wang, G., Carbonetto, P., and Stephens, M. (2019). Flexible statistical methods for estimating and testing effects in genomic studies with multiple conditions. Nat. Genet. *51*, 187–195. https://doi.org/10.1038/s41588-018-0268-8.
51. Norland, K., Schaid, D.J., and Kullo, I.J. (2024). A linear weighted combination of polygenic scores for a broad range of traits improves prediction of coronary heart disease. Eur. J. Hum. Genet. *32*, 209–214. https://doi.org/10.1038/s41431-023-01463-0.
52. Truong, B., Hull, L.E., Ruan, Y., Huang, Q.Q., Hornsby, W., Martin, H., van Heel, D.A., Wang, Y., Martin, A.R., Lee, S.H., and Natarajan, P. (2024). Integrative polygenic risk score improves the prediction accuracy of complex traits and diseases. Cell Genom. *4*, 100523. https://doi.org/10.1016/j.xgen.2024.100523.
53. Sun, Q., Rowland, B.T., Chen, J., Mikhaylova, A.V., Avery, C., Peters, U., Lundin, J., Matise, T., Buyske, S., Tao, R., et al. (2024). Improving polygenic risk prediction in admixed populations by explicitly modeling ancestral-differential effects via GAUDI. Nat. Commun. *15*, 1016. https://doi.org/10.1038/s41467-024-45135-z.
54. Marnetto, D., Pärna, K., Läll, K., Molinaro, L., Montinaro, F., Haller, T., Metspalu, M., Mägi, R., Fischer, K., and Pagani, L. (2020). Ancestry deconvolution and partial polygenic score can improve susceptibility predictions in recently admixed individuals. Nat. Commun. *11*, 1628. https://doi.org/10.1038/s41467-020-15464-w.
55. Ding, Y., Hou, K., Xu, Z., Pimplaskar, A., Petter, E., Boulier, K., Privé, F., Vilhjálmsson, B.J., Olde Loohuis, L.M., and Pasaniuc, B. (2023). Polygenic scoring accuracy varies across the genetic ancestry continuum. Nature *618*, 774–781. https://doi.org/10.1038/s41586-023-06079-4.
56. Thomas, D. (2010). Methods for investigating gene-environment interactions in candidate pathway and genome-wide association studies. Annu. Rev. Public Health *31*, 21–36. https://doi.org/10.1146/annurev.publhealth.012809.103619.
57. Durvasula, A., and Price, A.L. (2024). Distinct explanations underlie gene-environment interactions in the UK Biobank. Preprint at medRxiv, 2023.09.22.23295969. https://doi.org/10.1101/2023.09.22.23295969.
58. VanderWeele, T.J., and Robinson, W.R. (2014). On the causal interpretation of race in regressions adjusting for confounding and mediating variables. Epidemiology *25*, 473–484. https://doi.org/10.1097/ede.0000000000000105.
59. Jiang, X., Holmes, C., and McVean, G. (2021). The impact of age on genetic risk for common diseases. PLoS Genet. *17*, e1009723. https://doi.org/10.1371/journal.pgen.1009723.
60. Hui, D., Dudek, S., Kiryluk, K., Walunas, T.L., Kullo, I.J., Wei, W.Q., Tiwari, H.K., Peterson, J.F., Chung, W.K., Davis, B., et al. (2024). Risk factors affecting polygenic score performance across diverse cohorts. Preprint at medRxiv, 2023.05.10.23289777. https://doi.org/10.1101/2023.05.10.23289777.
61. Mandla, R., Schroeder, P., Porneala, B., Florez, J.C., Meigs, J.B., Mercader, J.M., and Leong, A. (2024). Polygenic scores for longitudinal prediction of incident type 2 diabetes in an ancestrally and medically diverse primary care physician network: a patient cohort study. Genome Med. *16*, 63. https://doi.org/10.1186/s13073-024-01337-0.
62. Hou, K., Xu, Z., Ding, Y., Mandla, R., Shi, Z., Boulier, K., Harpak, A., and Pasaniuc, B. (2024). Calibrated prediction intervals for polygenic scores across diverse contexts. Nat. Genet. *56*, 1386–1396. https://doi.org/10.1038/s41588-024-01792-w.
63. Choudhury, P.P., Maas, P., Wilcox, A., Wheeler, W., Brook, M., Check, D., Garcia-Closas, M., and Chatterjee, N. (2020). iCARE: R package to build, validate and apply absolute risk models. PLoS One *15*, e0228198. https://doi.org/10.1371/journal.pone.0228198.
64. Mathieson, I., and Scally, A. (2020). What is ancestry? PLoS Genet. *16*, e1008624. https://doi.org/10.1371/journal.pgen.1008624.
65. Alexander, D.H., Novembre, J., and Lange, K. (2009). Fast model-based estimation of ancestry in unrelated individuals. Genome Res. *19*, 1655–1664. https://doi.org/10.1101/gr.094052.109.
66. Browning, S.R., Waples, R.K., and Browning, B.L. (2023). Fast, accurate local ancestry inference with FLARE. Am. J. Hum. Genet. *110*, 326–335. https://doi.org/10.1016/j.ajhg.2022.12.010.
67. Hilmarsson, H., Kumar, A.S., Rastogi, R., Bustamante, C.D., Montserrat, D.M., and Ioannidis, A.G. (2021). High resolution ancestry deconvolution for next generation genomic data. Preprint at bioRxiv. https://doi.org/10.1101/2021.09.19.460980.
68. Maples, B.K., Gravel, S., Kenny, E.E., and Bustamante, C.D. (2013). RFMix: a discriminative modeling approach for rapid and robust local-ancestry inference. Am. J. Hum. Genet. *93*, 278–288. https://doi.org/10.1016/j.ajhg.2013.06.020.
69. 1000 Genomes Project Consortium, Auton, A., Brooks, L.D., Durbin, R.M., Garrison, E.P., Kang, H.M., Korbel, J.O., Marchini, J.L., McCarthy, S., McVean, G.A., and Abecasis, G.R. (2015). A global reference for human genetic variation. Nature *526*, 68–74. https://doi.org/10.1038/nature15393.
70. Koenig, Z., Yohannes, M.T., Nkambule, L.L., Zhao, X., Goodrich, J.K., Kim, H.A., Wilson, M.W., Tiao, G., Hao, S.P., Sahakian, N., et al. (2024). A harmonized public resource of deeply sequenced diverse human genomes.

Genome Res. *34*, 796–809. https://doi.org/10.1101/gr.278378.123.

71. Wojcik, G.L., Graff, M., Nishimura, K.K., Tao, R., Haessler, J., Gignoux, C.R., Highland, H.M., Patel, Y.M., Sorokin, E.P., Avery, C.L., et al. (2019). Genetic analyses of diverse populations improves discovery for complex traits. Nature *570*, 514–518. https://doi.org/10.1038/s41586-019-1310-4.
72. 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., et al. (2021). Improving reporting standards for polygenic scores in risk prediction studies. Nature *591*, 211–219. https://doi.org/10.1038/s41586-021-03243-6.
73. Hou, K., Gogarten, S., Kim, J., Hua, X., Dias, J.-A., Sun, Q., Wang, Y., Tan, T., Polygenic Risk Methods in Diverse Populations PRIMED Consortium Methods Working Group, and Atkinson, E.G., et al. (2024). Admix-kit: an integrated toolkit and pipeline for genetic analyses of admixed populations. Bioinformatics *40*, btae148. https://doi.org/10.1093/bioinformatics/btae148.