Genome Medicine

RESEARCH

Open Access

# Ancestry-aligned polygenic scores combined with conventional risk factors improve prediction of cardiometabolic outcomes in African populations

Michelle Kamp[1,2,3*], Oliver Pain[4], Cathryn M. Lewis[3,5] and Michèle Ramsay[1,2*]

## Abstract

**Background** Cardiovascular diseases (CVD) are a major health concern in Africa. Improved identification and treatment of high-risk individuals can reduce adverse health outcomes. Current CVD risk calculators are largely unvalidated in African populations and overlook genetic factors. Polygenic scores (PGS) can enhance risk prediction by measuring genetic susceptibility to CVD, but their effectiveness in genetically diverse populations is limited by a European-ancestry bias. To address this, we developed models integrating genetic data and conventional risk factors to assess the risk of developing cardiometabolic outcomes in African populations.

**Methods** We used summary statistics from a genome-wide association meta-analysis ($n = 14{,}126$) in African populations to derive novel genome-wide PGS for 14 cardiometabolic traits in an independent African target sample (Africa Wits-INDEPTH Partnership for Genomic Research (AWI-Gen), $n = 10{,}603$). Regression analyses assessed relationships between each PGS and corresponding cardiometabolic trait, and seven CVD outcomes (CVD, heart attack, stroke, diabetes mellitus, dyslipidaemia, hypertension, and obesity). The predictive utility of the genetic data was evaluated using elastic net models containing multiple PGS (MultiPGS) and reference-projected principal components of ancestry (PPCs). An integrated risk prediction model incorporating genetic and conventional risk factors was developed. Nested cross-validation was used when deriving elastic net models to enhance generalisability.

**Results** Our African-specific PGS displayed significant but variable within- and cross- trait prediction (max.$R^2 = 6.8\%$, $p = 1.86 \times 10^{-173}$). Significantly associated PGS with dyslipidaemia included the PGS for total cholesterol (logOR = 0.210, SE = 0.022, $p = 2.18 \times 10^{-21}$) and low-density lipoprotein (logOR = −0.141, SE = 0.022, $p = 1.30 \times 10^{-20}$); with hypertension, the systolic blood pressure PGS (logOR = 0.150, SE = 0.045, $p = 8.34 \times 10^{-4}$); and multiple PGS associated with obesity: body mass index (max. logOR = 0.131, SE = 0.031, $p = 2.22 \times 10^{-5}$), hip circumference (logOR = 0.122, SE = 0.029, $p = 2.28 \times 10^{-5}$), waist circumference (logOR = 0.013, SE = 0.098, $p = 8.13 \times 10^{-4}$) and weight (logOR = 0.103, SE = 0.029, $p = 4.89 \times 10^{-5}$). Elastic net models incorporating MultiPGS and PPCs significantly improved prediction over MultiPGS alone. Models including genetic data and conventional risk factors were more

*Correspondence:
Michelle Kamp
michelle.kamp@kcl.ac.uk
Michèle Ramsay
michele.ramsay@wits.ac.za
Full list of author information is available at the end of the article

predictive than conventional risk models alone (dyslipidaemia: $R^2$ increase = 2.6%, $p = 4.45 \times 10^{-12}$; hypertension: $R^2$ increase = 2.6%, $p = 2.37 \times 10^{-13}$; obesity: $R^2$ increase = 5.5%, $1.33 \times 10^{-34}$).

**Conclusions** In African populations, CVD and associated cardiometabolic trait prediction models can be improved by incorporating ancestry-aligned PGS and accounting for ancestry. Combining PGS with conventional risk factors further enhances prediction over traditional models based on conventional factors. Incorporating data from target populations can improve the generalisability of international predictive models for CVD and associated traits in African populations.

**Keywords** Prediction modelling, Polygenic scores, Cardiovascular diseases, Cardiometabolic diseases, African populations

## Background

Cardiovascular diseases (CVD) are the leading cause of mortality and morbidity worldwide. In 2019, approximately 17.9 million deaths were attributed to CVD, 75% of which occurred in low- and middle-income regions, including Africa [1]. Early identification of high-risk patients and timely initiation of appropriate treatment are crucial in mitigating adverse health outcomes associated with CVD. Clinical guidelines recommend using ten-year risk calculators, such as Framingham's and QRISK3, to identify individuals at increased risk of developing CVD [2, 3]. However, the application of these calculators in Africa is challenging due to limited longitudinal data and limited validation in African populations [4, 5]. Moreover, most calculators do not consider genetic risk factors, which have been shown to contribute to CVD development [6].

Polygenic scores (PGS) can provide an estimate of genetic risk for disease based on the aggregate effect of many common variants, as originally identified in case–control GWAS studies. PGS can potentially improve risk stratification, the systematic process of categorising individuals based on their likelihood of developing a disease, and better identify those at higher risk for developing disease [7, 8]. Furthermore, including PGS alongside conventional risk factors further improves CVD-risk prediction accuracy [9, 10]. However, more than 85% of genome-wide association studies (GWAS) have been performed in populations of European ancestry, substantially reducing their predictability in other ancestry groups [11–13]. This limited portability is likely due to differences in population allele frequencies, linkage disequilibrium (LD), population-specific causal variants or effects that significantly influence disease risk within a population, and potential variations in gene–gene and gene-environment interactions ([11], and as reviewed by 8 and 9).

To address these challenges, there is active investment in increasing the representation of diverse populations in GWAS and developing innovative methodological and computational approaches to data analysis [13, 14]. Research indicates that PGS perform optimally within ancestrally matched populations, including in continental African populations [15] due to the increased genetic diversity, low LD, and high population substructure in African populations. Whilst scores that work well across all populations are desired, developing scores that consider the unique epidemiological characteristics and genetic diversity of African populations is necessary and could inform trans-ancestry method developments. By integrating population-specific genetic risk factors, we can enhance the accuracy and precision of risk assessment, ultimately improving patient stratification and optimising the allocation of limited healthcare resources [9, 16, 17]. This study aimed to develop and assess an integrated risk score model, considering both genetic and conventional factors, for CVD and associated cardiometabolic traits in populations residing in Africa.

(See figure on next page.)

**Fig. 1** Study design and overview. This study employed two primary datasets, the base and target datasets. Base: The African Partnership for Chronic Disease Research (APCDR) dataset, encompassing 14,126 participants from South Africa, Kenya, Uganda, Ghana, and Nigeria, and the target was the Africa Wits- INDEPTH Partnership for Genomic Research (AWI-Gen) dataset which included 10,602 participants from Burkina Faso, Ghana, Kenya, and South Africa. Polygenic scores for 14 cardiometabolic traits were derived using the *p*-value thresholding method combined with linkage disequilibrium (LD) clumping. The African subset of the 1000 Genomes Project (1 KG) served as the reference panel. Predictive modelling evaluated the efficacy of genetic, non-genetic, and integrated models in forecasting disease outcomes. The map included in the figure visually represents the approximate geographical distribution of the cohorts, with the position circles indicating the location. Key acronyms: APCDR, African Partnership for Chronic Disease Research; AWI-Gen, Africa Wits- INDEPTH Partnership for Genomic Research; CVD, cardiovascular disease; GWAS, Genome-Wide Association Study; LD, Linkage disequilibrium; PGS, Polygenic score

## Methods

### Study design and overview

The design of this study is illustrated in Fig. 1. The objective of the present study was to determine and evaluate the predictive performance of an integrated risk score (IRS), which encompasses genetic and conventional risk factors, for CVD and related cardiometabolic outcomes in African populations. The genetic factors in the IRS consist of multiple PGS and genetic ancestry, as inferred by projected principal components of ancestry (PPCs). Conventional risk factors, referred to as non-genetic factors going forward, included sociodemographic and lifestyle risk factors such as age, sex, smoking status, alcohol consumption, diet-related factors, physical activity, and nightly sleep duration.

Summary statistics from the African Partnership for Chronic Disease Research (APCDR) GWAS meta-analysis for cardiometabolic traits were used to derive PGS. The APCDR cohort consists of 14,126 individuals from different African regions [18]. Scores were trained on data from the Africa Wits-INDEPTH Partnership for Genomic Research (AWI-Gen) [19] cohort comprising 10,603 individuals from four African countries. Study participants from AWI-Gen resided in Burkina Faso, Ghana, Kenya, and South Africa, while those from ACPDR were from Ghana, Kenya, Nigeria, South Africa, and Uganda. Despite a similar regional mix, there was no recorded sample overlap between the target cohort and the individuals analysed in the GWAS from which summary statistics were obtained.

First, we constructed distinct PGS for each of the fourteen cardiometabolic traits, all continuous phenotypes, using the *p*-value thresholding and LD clumping approach (pT + clump), and according to standard procedures outlined by Choi and colleagues [20], employing the GenoPred pipeline (https://github.com/opain/GenoPred/tree/master/GenoPredPipe) [21]. Scores were derived for the following traits: six anthropometric indices (body mass index (BMI), height, weight, hip circumference, waist circumference, and waist-to-hip ratio (W–H ratio)); two blood pressure measurements (diastolic blood pressure (DBP) and systolic blood pressure (SBP)); four lipid traits (low-density lipoprotein cholesterol (LDL), high-density lipoprotein (HDL), total cholesterol (TC) and triglycerides (TG)); and two liver function measures (albumin and bilirubin blood serum levels).

Subsequently, the associations of the PGS and non-genetic factors with thirteen of the cardiometabolic traits (excluding bilirubin, which was not measured in AWI-Gen) and seven CVD-associated outcomes, namely CVD, diabetes mellitus, dyslipidaemia, heart attack, hypertension, obesity, and stroke, were tested. Next, to derive and evaluate the predictive utility of genetic, non-genetic, and IRS models, elastic net regression with nested cross-validation (NCV) was used.

### Data sources

#### *Base dataset: APCDR meta-analysis*

The APCDR (African Partnership for Chronic Disease Research) is a genome-wide association meta-analysis of association statistics that encompasses association statistics derived from four African cohorts [18]. The cohorts included the Uganda Genome Resource (UGR) ($n = 6188$), the Africa-America Diabetes Mellitus Study (AADM) ($n = 5231$), the Durban Diabetes Study (DDS) ($n = 1165$), and the Durban Case Control (DCC) ($n = 1542$) [18, 22, 23]. In brief, the meta-analysis, as conducted by Gurdasani et al. (2019), investigated 34 cardiometabolic traits in up to 14,126 individuals aged 18 years and older residing in Ghana, Kenya, Nigeria, South Africa, and Uganda [18]. The APCDR data is publicly available and includes imputed dosage data for all individuals and ~ 96 million variants. These data were generated using METASOFT [24] with a composite reference panel developed by authors [18]. Summary statistics were downloaded from the NHGRI-EBI GWAS Catalog [25] on 01 Jun 2021 for studies GCST009042 to GCST009060 (details of studies are provided in Table S1) [18].

#### *Target dataset: AWI-Gen*

AWI-Gen is a cross-sectional cohort study undertaken across four sub-Saharan African countries: Burkina Faso, Ghana, Kenya, and South Africa [19]. This study's primary objective is to explore genetic and environmental factors associated with cardiometabolic diseases in Africans. It is part of the Human Heredity and Health in Africa Consortium (H3Africa). From 2012 to 2016, approximately 12,000 participants, primarily between the ages of 40 and 60 years, were enrolled across six study centres, and individual-level genetic, health-related, and phenotypic data relating to lifestyle was collected. Baseline data was used in this study. The study sites are from South Africa, the MRC/Wits Agincourt Health and Demographic Surveillance System Site (HDSS) (referred to as Agincourt), the Dikgale HDSS of the University of Limpopo, and the Soweto Centre which is coordinated by the South African Medical Research Council/Wits Developmental Pathways for Health Research Unit (DPHRU); in Kenya, the African Population and Health Research Center HDSS in Nairobi; in Ghana, the Navrongo HDSS in the Navrongo Health Research Centre; and in Burkina Faso, the Nanoro HDSS hosted by the Institut de Recherche en Sciences de la Santé Clinical Research Unit [19, 26].

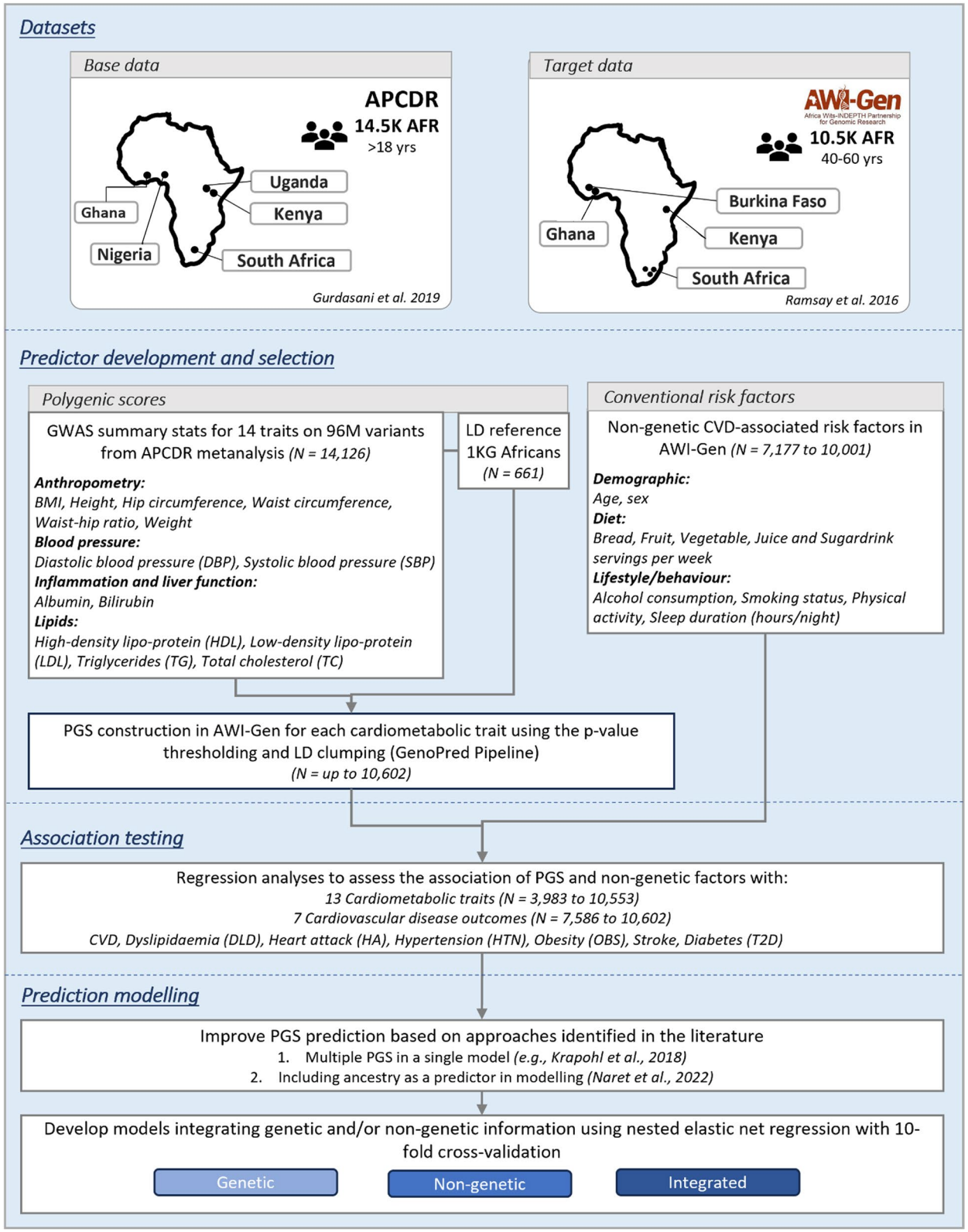


**Fig. 1** (See legend on previous page.)

#### *Genetic data*

Approximately 11,000 individuals were genotyped on the 2.3 M SNP H3Africa array at Illumina® FastTrackTM Microarray services (Illumina, San Diego, USA). Genotype calling was performed using the Illumina pipeline. Quality control (QC) was performed as described previously, but in summary, pre-imputation QC was performed using the H3ABioNet/H3Agwas pipeline (https://github.com/h3abionet/h3agwas) and variants with a minor allele frequency (MAF) < 0.01, missingness > 0.05 or Hardy–Weinberg equilibrium $p$-value $< 1 \times 10^{-3}$ were removed. Additionally, SNPs from the X and Y chromosomes, mitochondrial SNPs, and SNPs that did not match the GRCh37 reference alleles were removed. Samples that were potential duplicates (PIHAT > 0.9), had a missing SNP genotyping rate greater than 0.05, and reported vs. genetic sex inconsistencies were excluded. Population stratification was assessed using principal component (PC) analysis based on an LD-pruned subset of SNPs using the smartPCA program implemented in EIGENSTRAT. Imputation was performed using the African Genome Resources reference panel (EAGLE2 + PBWT pipeline) at the Sanger Imputation Server (https://imputation.sanger.ac.uk/). Post-imputation QC involved removing indels, rare SNPs (MAF ≤ 0.01), and poorly imputed SNPs (Info score ≤ 0.6), resulting in a final dataset containing 10,603 participants and 13.98 M SNPs.

#### *Phenotypic data*

In addition to demographic, general health and infection history variables, the AWI-Gen questionnaire provided information on diet, smoking status, alcohol use, physical activity, and sleep. The variables associated with CVDs and used in conventional CVD risk calculators were included in our models and referred to as non-genetic factors throughout our analyses [3, 27, 28]. The variables selected included age, sex, current smoking status, alcohol consumption status, sleep (hours/night), moderate and vigorous physical activity (minutes/week), juice (number per week) and sugar drinks (number per week). Current smoking status was obtained from "Yes", "No" responses to the following question, "Do you currently smoke tobacco?" Similarly, alcohol consumption status was determined from "Yes," "No" responses to the following question "Are you a current alcohol consumer?" Those who preferred not to answer or did not know, were excluded. The Global Physical Activity Questionnaire (GPAQ) was used to obtain self-reported physical activity. Total moderate-vigorous physical activity (MVPA) in minutes per week was calculated from the accumulation of occupation, travel-related and leisure time physical activity. Sitting time (minutes/week) is used as a proxy for sedentary behaviour [29]. Weekly consumption of bread (slices per week), fruit (servings per week), and vegetables (servings per week) was calculated by multiplying the individual's number of servings per day by the number of times a week each respective food group was consumed. Not all the selected variables were available in the Soweto sample; thus, these samples were excluded from the prediction modelling analyses. For the remaining participants, individuals with more than 5% of the data missing among these selected variables were removed. Additional details on the variables and their construction can be found in Supplementary Materials S2 and S3.

#### *Disease outcomes*

The outcome variables in this study included 13 cardiometabolic traits and seven CVD-associated outcomes. Cardiometabolic traits included BMI, height, weight, hip circumference, waist circumference, and W–H ratio, DBP, SBP, LDL, HDL, TC, TG, and albumin and bilirubin blood serum levels. CVD-associated disease outcomes assessed were CVD, diabetes mellitus (T2D), dyslipidaemia (DLD), heart attack (HA), hypertension (HTN), obesity (OBS), and stroke. CVD was defined as present if the participant reported having had a heart attack, stroke, or transient ischaemic attack. Participants previously diagnosed with congestive heart failure or angina were also classified as having CVD. Transient ischaemic attack, congestive heart failure and angina outcomes are not included as single disease endpoints in our analyses due to the small sample size. Further information regarding the outcome definitions can be found in Supplementary Materials S2 and S3. For disease traits, all cases were included. The maximum sample sizes for each phenotype in the prediction modelling analyses are shown in Table 1.

NA refers to participants who reported that they did not know their disease status, and those from the Soweto site—these participants were excluded from modelling analysis given the high level of missingness of non-genetic risk factors. CVD includes heart attack, stroke, or transient ischaemic attack, and participants previously diagnosed with congestive heart failure or angina. Limited case numbers for transient ischaemic attack, congestive heart failure and angina restricted their use as disease endpoints themselves. Some individuals experienced multiple CVD outcomes, and subsequently, a summation of individual outcomes does not equate to the total number of CVD cases reported.

### Polygenic scoring

#### *Quality control of datasets*

QC of base data: GWAS summary statistics.

GWAS summary statistics of traits for inclusion in the study were selected due to their relevance to

**Table 1** Sample sizes for each disease outcome (e.g., characteristics of the AWI-Gen cohort for the CVD traits and associated outcomes)

| Phenotype | Abbrev | Total sample size | NA | No. cases | No. controls |
|---|---|---|---|---|---|
| Cardiovascular disease | **CVD** | **7586** | **1517 (20%)** | **219 (2.9%)** | **5850 (77.1%)** |
| Dyslipidaemia | **DLD** | **10,602** | **2121 (20%)** | **5680 (53.6%)** | **2801 (26.4%)** |
| Heart attack | **HA** | **8598** | **1719 (20%)** | **46 (0.5%)** | **6833 (79.5%)** |
| Hypertension | **HTN** | **10,602** | **2121 (20%)** | **3183 (30%)** | **5298 (50%)** |
| Stroke | **Stroke** | **9737** | **1947 (20%)** | **113 (1.2%)** | **7677 (78.8%)** |
| Obesity | **OBS** | **10,602** | **2121 (20%)** | **1775 (16.7%)** | **6706 (63.3%)** |
| Type 2 diabetes | **T2D** | **10,537** | **2109 (20%)** | **568 (5.4%)** | **7861(74.6%)** |

cardiometabolic disease, presence in the AWI-Gen cohort, as well as sharing a similar ancestry to the target AWI-Gen population. The identified summary statistics underwent a series of standard quality control (QC) procedures [20], including the extraction of HapMap3 variants, and the removal of ambiguous variants, or where variants had missing data. Variants were flipped to match the 1000 Genomes Phase 3 (1 KG) reference, and then variants were retained if the MAF > 0.01 in the African subset of 1 KG (1 KG AFR), the MAF > 0.01 in the GWAS sample, and the INFO > 0.6. GWAS summary statistics variants and samples were removed if they (1) had a discordant MAF (> 0.2) between the reference and GWAS sample, (2) had reported *p*-values outside the range of 0 to 1, (3) were duplicates, or (4) had a sample size > 3 SD from the median sample size.

QC of target data: ancestry classification.

Individuals in AWI-Gen were assigned to the five super populations present in the 1000 Genomes phase 3 (1 KG) reference sample [30], namely European, East Asian, South Asian, African, and Admixed American. Super population membership was predicted using a 1 KG reference trained elastic net model consisting of the first six reference-projected genetic principal components (PPCs). Principal components were defined in the 1 KG reference using HapMap3 SNPs in common between the 1 KG and AWIGEN data with a minor allele frequency > 0.05, missingness < 0.02 and Hardy–Weinberg *p*-value > $1 \times 10^{-6}$. LD pruning for independent variants was performed in PLINK [31] after the removal of long-range LD regions [32], using a window size of 1000, step size of 5, and $r^2$ threshold of 0.2.

A multinomial elastic net model, created using the "glmnet" R package [33], predicted super population membership in the 1 KG reference with fivefold cross-validation. This model, along with reference-derived principal components, was applied to AWI-Gen for similar predictions. Participants with a predicted probability over 0.5 were assigned to a super population, with all being assigned to the AFR superpopulation as expected.

### *Score construction*

Typically, a PGS follows the form $\beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k + \ldots + \beta_n X_n$, where $\beta_k$ represents the effect size attributed to each allele for a given cardiometabolic trait associated with SNP *k*. $X_k$ is the number of effect alleles at SNP *k*, and *n* is the total number of SNPs in the PGS. To derive the PGS for each trait, we used (1) publicly available GWAS summary statistics described in Gurdasani et al. (2019); extracting the disease-associated variants, the effect allele, the estimated $\beta$-coefficient for the effect allele, and the *p*-value of each genetic variant, and (2) linkage disequilibrium (LD) between genetic variants from the African 1 KG LD reference panel (661 Africans) [30]. Scores were derived using the pT + clump approach. The pT + clump method is a robust approach that enhances the accuracy and relevance of PGS by selecting the most informative genetic variants while reducing redundancy due to LD. We used default LD-based clumping parameters ($r^2 = 0.1$, window = 250 kb) to retain only the single most significant variant within each locus, as overly aggressive LD thresholds can detrimentally affect the predictive power PGS [20]. The 1 KG AFR was used to estimate LD. Ten *p*-value thresholds were considered ($1 \times 10^{-8}$, $1 \times 10^{-6}$, $1 \times 10^{-4}$, $1 \times 10^{-2}$, 0.1, 0.2, 0.3, 0.4, 0.5 and 1). Polygenic scores were then calculated in AWI-Gen participants, imputing missing variants using the 1 KG AFR allele frequency. In the AWI-Gen sample, 1,104,4026 HapMap3 variants were present. Polygenic scores were standardised (scaled and centred) based on the mean and SD of PGS in the 1 KG AFR reference sample. The score calculations were performed using PLINK v1.9 as implemented in the GenoPred pipeline (https://github.com/opain/GenoPred).

#### *Association testing*

Following PGS development, regression analysis was used to assess the within- and cross-trait predictive utility of each PGS, and with the seven disease outcomes of interest, while accounting for confounders such as age, sex, and the first eight with-in sample principal components to avoid PGS associations being confounded by population structure [15]. Similarly, regression analyses were run to assess the association between conventional risk factors, including age, smoking status, alcohol consumption, diet-related variables, sleep, and physical activity, with selected disease outcomes. Given the differences by sex of these traits within African populations, the analyses were adjusted for sex [34–36]. The proportion of variance for a trait explained by the PGS and non-genetic factors was computed as the phenotypic variance explained, $R^2$. For PGS association testing, $R^2$ was obtained from a full model including both PGS and covariates (PCs, sex, age, and age-squared) minus the $R^2$ obtained from a model including covariates alone. $R^2$ was not adjusted to the liability threshold model due to limited disease prevalence estimates available across Africa and the substantial variation in prevalence noted across cohort sites. For multiple testing, results were corrected for the number of PGS tested for each outcome (i.e., applying a $p$-value threshold of 0.05/14). We did not correct for the number of $p$-value thresholds as they are correlated, and a Bonferroni correction would be overly conservative. The performance of each PGS was assessed as the Pearson correlation ($r$) between the observed and predicted outcome values and the Area under the receiver operating characteristic curve (AUC) statistics calculated. Correlation was used as the main test statistic as it is applicable for both binary and continuous outcomes and standard errors are easily computed.

#### *Derivation of genetic ancestry predictors*

In addition to PGS, reference-projected genetic principal components (PPCs) were included in prediction models to enhance prediction. Genetic principal components capture major axes of genetic variation, which primarily represent differences in genetic ancestry [37] and can be used to enhance prediction over PGS alone [38]. To prevent overfitting, the principal component SNP-weights should be derived independently of the target sample. Therefore, we used the PPCs described in Sect. 2.3.1.2, where the first six genetic PCs were derived from the 1 KG reference, and then projected these PCs into the AWI-Gen target sample.

#### Integrated risk score and prediction modelling

Elastic net regression with nested cross-validation (NCV) (https://github.com/opain/GenoPred/blob/master/Scripts/Model_builder/Model_builder_V2_nested.R) was used to develop and evaluate the predictive utility of three risk prediction models: genetic, non-genetic, and integrated:

a. Genetic:
    i. MultiPGS—assessed the predictive utility of utilising multiple PGS compared to single-trait PGS
    ii. MultiPGS + Ancestry—assessed the predictive utility of utilising multiple PGS and information relating to ancestry, specifically projected principal components.
b. Non-genetic—assessed the predictive utility of selected conventional risk factors.
c. Integrated—assessed the predictive utility when combining all genetic (MultiPGS and Ancestry) and non-genetic predictors.

Elastic net balances feature selection and regularisation to reduce over-fitting and address collinearity among predictors. It combines the properties of both ridge and lasso regression, where similar to ridge regression, elastic net applies a penalty to model coefficients which shrinks them towards zero, thus reducing the impact of less important predictors. And similar to lasso regression, elastic net performs variable selection by setting some coefficients to zero. By balancing the weight of ridge and lasso penalties, this regularisation removes the need for manual selection of predictors and selects and weights the most predictive variables appropriately, reducing redundancy and enhancing model interpretability [39]. NCV repeatedly partitioned the dataset into training, validation, and testing sets, and consisted of 5 outer folds with a 90–10 data split (90% training, 10% testing) to provide an unbiased estimate of the predictive utility of the model, and 10 inner folds (80% training, 20% testing) for hyperparameter tuning. The proportion of variance explained by a model was computed as $R^2$. Hyperparameters were determined using the "caret" R package, which optimises the RMSE for continuous outcomes and accuracy for binary outcomes.

The predictive utility of the models were defined as the correlation between observed and predicted values of each model, and the comparative performance of the models assessed using William's test (also known as the Hotelling–Williams test) as implemented by the "psych" R package's "paired. r" function. The code used to prepare

data and conduct analyses is available on the GenoPred Pipeline GitHub page (see Data and Code Availability).

For genetic (the MultiPGS and MultPGS + PPC models) and integrated models, for each cardiometabolic trait, rather than selecting the single best-performing PGS (based on max $R^2$), all PGS were retained for subsequent predictive modelling analyses and elastic net regression was utilised to simultaneously select and weight predictors [40, 41]. Genetically inferred ancestry was included in prediction models to account for population stratification and potentially improve prediction [42]. To reduce overfitting, ancestry was determined by fitting data to the 1 KG Phase 3 projected principal components (PPCs) of population structure and not to AWI-Gen sample PCs.

Non-genetic models included ten conventional risk factors selected based on data availability in AWI-Gen and their known association with CVDs. Integrated models included genetic (PGS and PPCs) and non-genetic factors. No data were available for diet-related variables for the Soweto study site for men and women, so this site was excluded from prediction modelling analyses.

#### Statistical analysis

All analyses were performed using PLINK v1.9 (https://www.cog-genomics.org/plink/1.9/) [43], and R version 3.4.4 (http://www.r-project.org/) [44] unless specified otherwise. Data are presented as percentages (%) or mean ± SD. Associations between non-genetic factors and PGS and CVD-associated outcomes were assessed by logistic regression with the adjustment of PCs, sex, age, and age squared as described previously [15].

#### Ethics statement

This study was approved by the Human Research Ethics Committee (Medical) of the University of the Witwatersrand (Wits)(protocol number M210355) as a substudy of the AWI-Gen project (protocol number M170880).

## Results

Fourteen PGS were derived, and their within- and cross-trait associations assessed using regression analyses. Elastic net regression with nested tenfold cross-validation was used to determine the predictive utility of models (genetic, non-genetic, and integrated), and the performance for each PGS and model was assessed using the correlation between observed and fitted values.

#### Derivation, validation, and association testing of PGS

All PGS, except those for albumin and waist-hip ratio, had at least one significant association after correcting for multiple testing (Fig. 2). There was extensive variability in variance explained across phenotypes, with the variance explained ranging ($R^2$) from 0.068 ($p = 1.86 \times 10^{-173}$) for LDL to 0.004 ($p = 4.70 \times 10^{-20}$) for height. The most significant associations were found amongst lipid traits (HDL, LDL, TC, and TG). Given the genetic correlation across traits, especially amongst anthropometric and lipid traits, significant cross-prediction was also noted (max $R^2 = 0.068$, $p = 6.23 \times 10^{-175}$ for TC PGS predicting LDL). Tables S4 and S5 list all significant within and cross-trait predictions after correcting for multiple testing.

Cardiometabolic outcomes of DLD, OBS and HTN were significantly associated with the PGS (Fig. 3). Four PGS (BMI, hip circumference, weight, and waist circumference) were associated with increased risk of OBS, with the largest increase in risk linked to the PGS for BMI (max. logOR = 0.131, SE = 0.031, $p = 2.22 \times 10^{-5}$). Similarly, three PGS (TC, LDL and HDL) were associated with DLD, with the greatest increase in risk linked to the PGS for HDL (max. logOR = 0.210, SE 0.022, $p = 2.18 \times 10^{-21}$). For HTN, only the PGS for SBP was associated (max. logOR = 0.150, SE 0.045, $p = 8.34 \times 10^{-4}$). No significant associations were found for CVD, HA, T2D and stroke disease outcomes. Distribution assessments (mean, standard deviation, interquartile range) of derived PGS across AWI-Gen sites could not be done as sample sizes were too small to accurately contrast effect sizes between populations.

#### Non-genetic factor associations with CVD disease outcomes

Factors previously identified as associated with CVD were selected from the AWI-Gen study: age, sex, smoking status, alcohol consumption, various diet factors, physical activity, and sleep (hours per night). The dietary variables included weekly consumption of fruit, vegetables, bread, fruit juice and sugar drinks. Using regression analysis, and accounting for confounders such as age, sex and principal components, the relationship between each factor and disease outcome was assessed. These non-genetic factors accounted for little to no variance explained in CVD and HA (Fig. 3 and supplementary material Table S7). In contrast, all factors were associated with HTN (max. logOR = 0.64, SE = 0.023, $p = 9.75 \times 10^{-173}$) and OBS (max. logOR = 0. 58, SE = 0.029, $p = 1.23 \times 10^{-92}$). Age was the most significant predictor of HTN ($\text{logOR}_{\text{AGE}} = 0.64$, SE = 0.023, $p = 9.75 \times 10^{-173}$), T2D ($\text{logOR}_{\text{AGE}} = 0.41$, SE = 0.037, $p = 5.32 \times 10^{-28}$), and stroke ($\text{logOR}_{\text{AGE}} = 0.43$, SE = 0.073, $p = 3.05 \times 10^{-9}$); whilst bread servings per week accounted for greatest increased odds in OBS ($\text{logOR}_{\text{BREAD}} = 0.58$, SE = 0.029, $p = 1.23 \times 10^{-92}$) and DLD ($\text{logOR}_{\text{BREAD}} = 0.12$, SE = 0.025, $p = 5.42 \times 10^{-06}$). Alcohol consumption was associated with reduced odds in all diseases except CVD. This effect was most pronounced in

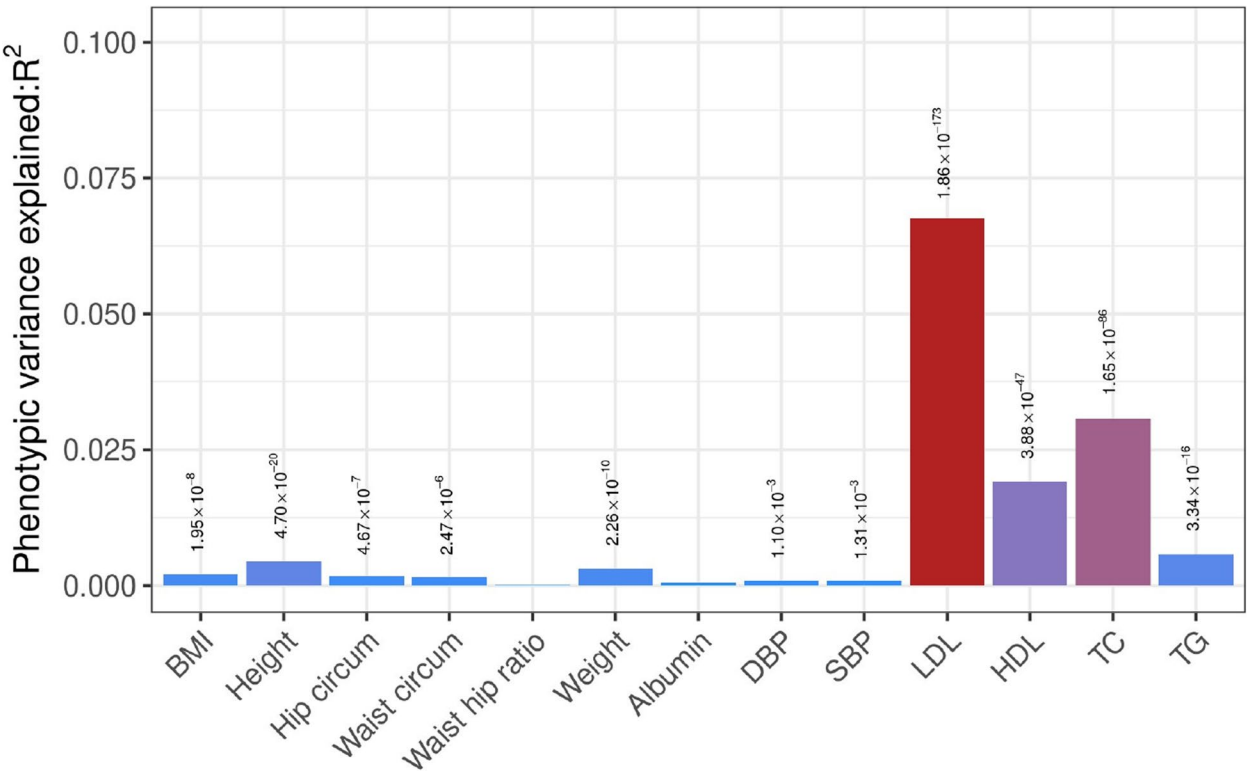


**Fig. 2** Within-trait prediction of derived PGS. Variance explained (R-squared) by the PGS for 13 cardiometabolic traits in the AWI-Gen target sample across *P*-value thresholds. Figure only shows the results for the PGS with the highest variance explained. The values above the bars are *P*-values indicating whether the variance explained is significantly different from zero. Within-trait predictive utility, described as the phenotypic variance explained ($R^2$), of 13 of the derived PGS. DBP, diastolic blood pressure; SBP, systolic blood pressure; LDL, low-density lipoprotein; HDL, high-density lipoprotein; TC, total cholesterol; TG, triglycerides

OBS ($\log OR_{ALC} = -0.72$, SE = 0.039, $p = 9.17 \times 10^{-77}$), T2D ($\log OR_{ALC} = -0.54$, SE = 0.054, $p = 6.98 \times 10^{-24}$), and HA ($\log OR_{ALC} = -0.53$, SE = 0.154, $p = 6.73 \times 10^{-04}$) (Fig. 3).

## Prediction modelling for CVD disease outcomes

### *Genetic prediction models*

Models consisting of all PGS combined (MultiPGS) were significantly associated with four outcomes: DLD ($p = 1.81 \times 10^{-37}$, AUC = 0.574), HTN ($p = 1.84 \times 10^{-162}$, AUC = 0.652), OBS ($1.69 \times 10^{-4}$, AUC = 0.690), and T2D ($p = 9.43 \times 10^{-17}$, AUC = 0.594) (Fig. 4 orange bar). The greatest improvement was noted in HTN, where phenotypic variance explained more than doubled from 2.8 to 6.7%, followed by OBS (4.0% to 7.5%) and DLD (1.2% to 1.5%). Prediction in the T2D MultiPGS model remained low, increasing from 0.3% to 0.7%. The MultiPGS model for CVD and HA did not significantly improve prediction (Supplementary materials Table S8).

To further improve prediction of PGS, we assessed whether accounting for population stratification through the incorporation of ancestry could improve performance. As per Fig. 5, including an ancestry predictor improved prediction, by approximately 2.5% for both HTN ($R^2_{PGS} = 6.5\%$, $R^2_{ANS} = 8.9\%$, $R^2_{PGS+ANS} = 9\%$, $AUC_{PGS+ANS} = 0.68$) and OBS ($R^2_{PGS} = 7.1\%$, $R^2_{ANS} = 7.5\%$, $R^2_{PGS+ANS} = 10\%$, $AUC_{PGS+ANS} = 0.735$), but had little to no effect for DLD, stroke, and T2D with improvements less than 0.1% for each. For DLD, the PGS accounted for greater variance explained and including ancestry reduced prediction performance by 0.1% ($R^2_{PGS} = 1.7\%$, $R^2_{ANS} = 0.1\%$, $R^2_{PGS+ANS} = 1.6\%$). Similarly, to the MultiPGS model, CVD and HA PGS + Ancestry models showed no significant prediction (Supplementary materials Table S9).

### *Integrated prediction models*

Lastly, we assessed whether including genetic (PGS and PPCs) and non-genetic factors in an IRS model could improve the predictive performance of models for selected disease outcomes in African populations. The integrated model consisted of 143 predictors and a maximum of 8057 individuals. As per Fig. 6, when we

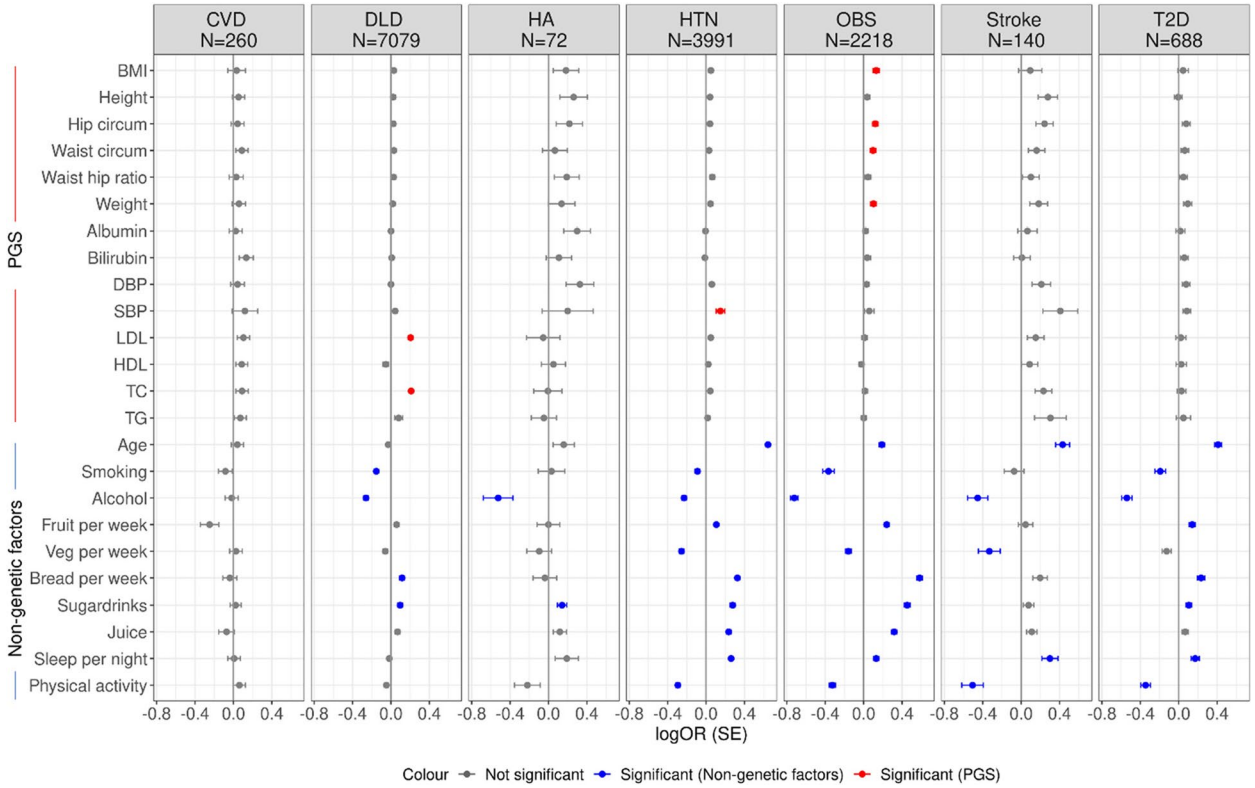


**Fig. 3** Association analysis of the PGS and non-genetic factors with seven CVD-associated outcomes. Logistic regression analyses assessing the association of each of the 14 derived PGS and 10 non-genetic factors with the seven selected CVD-disease outcomes. Significant associations with PGS are coloured in red, and significant associations with non-genetic factors are coloured in blue. Maximum case numbers are indicated by *N* = . CVD, cardiovascular disease; DLD, dyslipidaemia; HA, heart attack; HTN, hypertension; OBS, obesity; T2D, type 2 diabetes; DBP, diastolic blood pressure; SBP, systolic blood pressure; LDL, low-density lipoprotein; HDL, high-density lipoprotein; TC, total cholesterol; TG, triglycerides

examined the non-genetic and genetic contribution to total variation, the non-genetic factors (in blue) consistently accounted for greater variance explained than genetic factors, the greatest prediction from non-genetic factors seen in OBS ($R^2_{\text{Nongen}}=19\%$) and the lowest in stroke ($R^2_{\text{Nongen}}=0.3\%$). Within DLD, HTN and OBS, non-genetic risk factors explained 7.4% to 9% of the total variation in fully adjusted models depending on the outcome. The predictive utility of genetic models (in red) was variable across outcomes, with the highest $R^2$ noted in HTN ($R^2_{\text{Gen}}=8\%$) and OBS ($R^2_{\text{Gen}}=9\%$). For CVD, HA and stroke, genetic models yielded no significant phenotypic variance explained. In contrast, the IRS models (in green) showed significant prediction in all outcomes except CVD. Models including genetic predictors minimally reduced the prediction of Stroke, HA, and T2D (by 0.5% or less) but increased the prediction of OBS, HTN, and DLD. OBS prediction increased by 5.5% to reach 21% ($R^2_{\text{Gen}}=9.3\%$, $R^2_{\text{Nongen}}=15.3\%$, $R^2_{\text{IRS}}=20.9\%$, $\text{AUC}_{\text{IRS}}=0.830$), HTN by 2.6% to reach 14% ($R^2_{\text{Gen}}=8.3\%$, $R^2_{\text{Nongen}}=11.2\%$, $R^2_{\text{IRS}}=13.8\%$, $\text{AUC}_{\text{IRS}}=0.723$), and DLD by 2.6% to 15% ($R^2_{\text{Gen}}=2.9\%$, $R^2_{\text{Nongen}}=11.5\%$, $R^2_{\text{IRS}}=14.1\%$, $\text{AUC}_{\text{IRS}}=0.738$) (Supplementary materials Table S10A). A pairwise comparison of scores was performed for each model to show the difference in correlation within and between models for outcomes, with *p*-values for significant differences calculated using William's test results (Supplementary material S10B and S10C). The integrated models were significantly different to those including non-genetic factors alone for the same traits, suggesting genetic information provides independent and complementary information to non-genetic risk factors for risk prediction (DLD: $p=4.45\times10^{-12}$; HTN: $p=2.37\times10^{-13}$; OBS: $p=1.33\times10^{-34}$).

## Discussion

This study developed and evaluated the predictive utility of 14 cardiometabolic trait ancestry-aligned PGS for seven CVD-associated outcomes in continental African populations. We investigated whether modelling these PGS into a MultiPGS improved prediction and if integrating this MultiPGS with established non-genetic risk factors had greater predictive utility beyond non-genetic factors alone. To date, AWI-Gen provides the largest

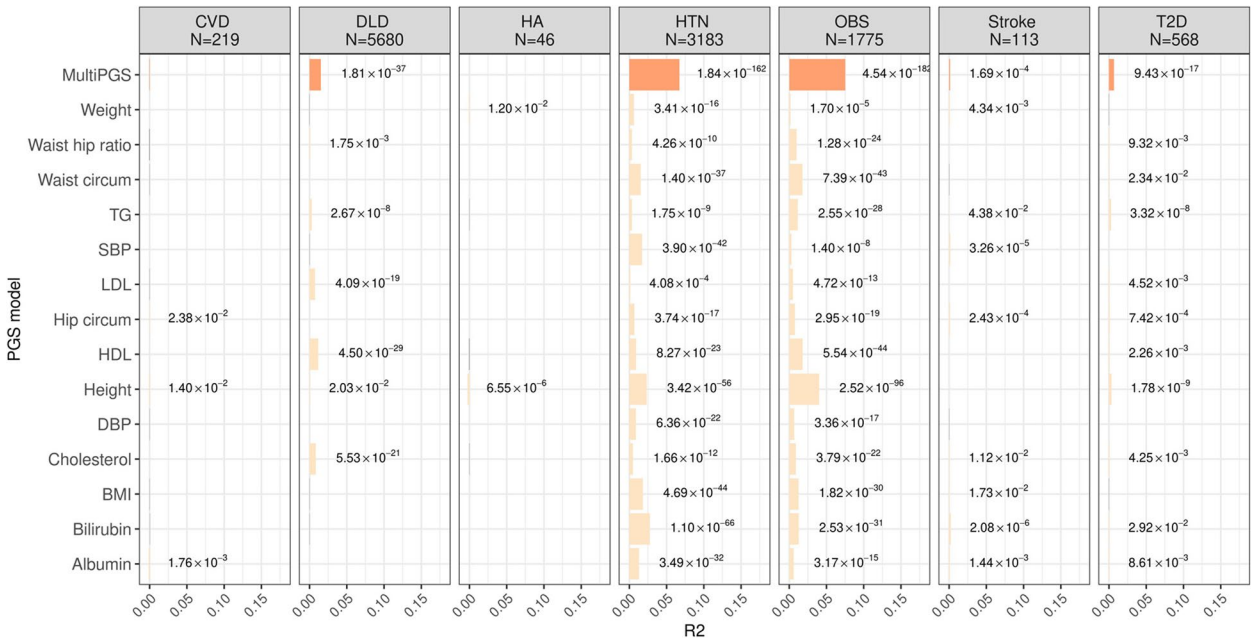


**Fig. 4** Predictive utility of the single best PGS and MultiPGS models for seven CVD-associated outcomes. The predictive utility of a single PGS was compared with that inclusive of all cardiometabolic trait PGS (MultiPGS) across seven selected CVD-disease outcomes. Single trait PGS are shown in light orange, and MultiPGS shown in dark orange. Significant associations have *p*-values displayed. Case numbers for each phenotype are indicated by *N* = . CVD, cardiovascular disease; DLD, dyslipidaemia; HA, heart attack; HTN, hypertension; OBS, obesity; T2D, type 2 diabetes; DBP, diastolic blood pressure; SBP, systolic blood pressure; LDL, low-density lipo-protein; HDL, high-density lipo-protein; TC, total cholesterol; TG, triglycerides

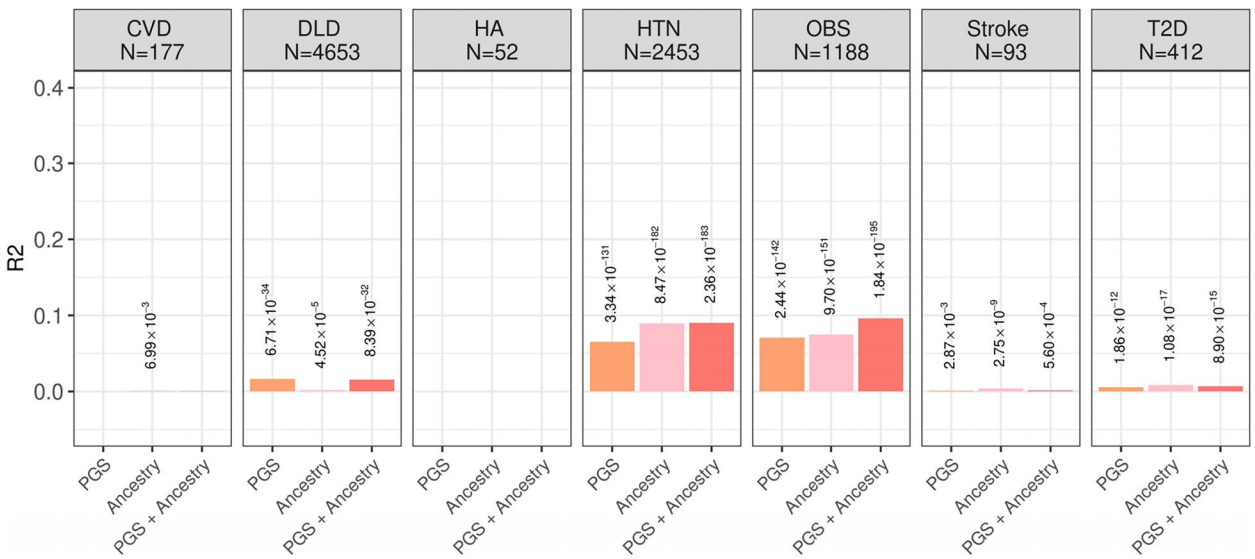


**Fig. 5** The predictive utility of genetic models (including PGS and projected principal components) for CVD-associated disease outcomes. Comparing the predictive utility of models inclusive of PGS (orange), projected ancestry (pink), and PGS + Projected Ancestry (red) across seven selected CVD-associated disease outcomes. Significant associations have *p*-values displayed. CVD, cardiovascular disease; DLD, dyslipidaemia; HA, heart attack; HTN, hypertension; OBS, obesity; T2D, type 2 diabetes; DBP, diastolic blood pressure; SBP, systolic blood pressure; LDL, low-density lipoprotein; HDL, high-density lipoprotein; TC, total cholesterol; TG, triglycerides

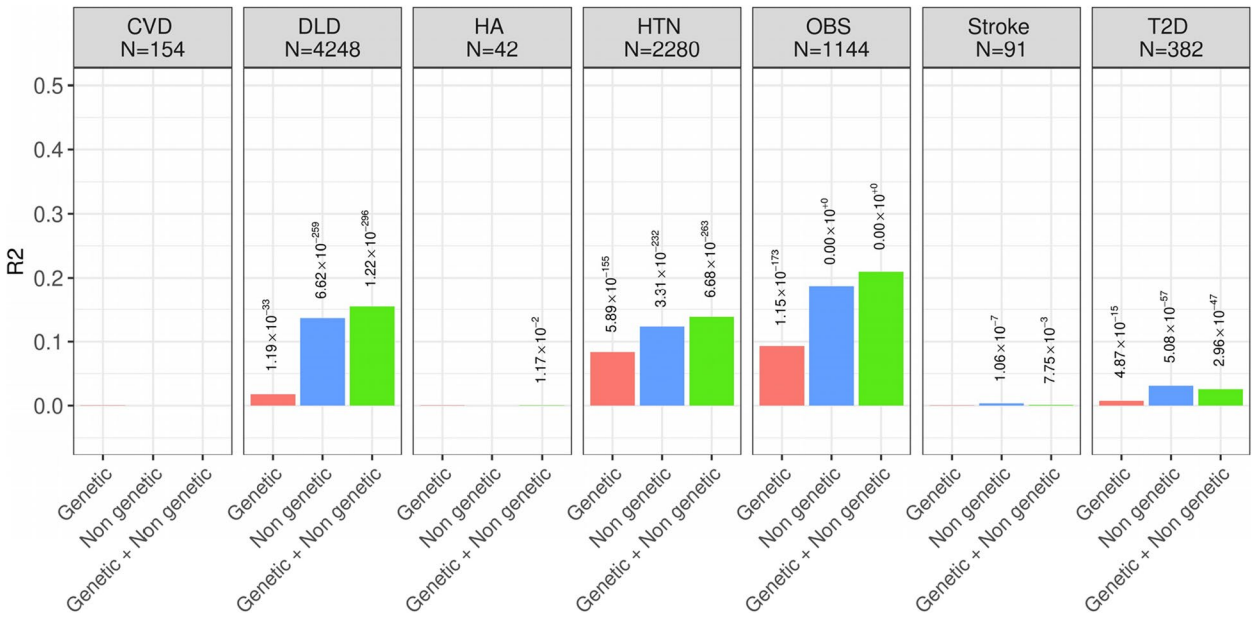


**Fig. 6** The predictive utility of integrated models (including genetic and non-genetic factors) for CVD-associated disease outcomes. Prediction modelling analyses assessing the predictive performance of integrated models across seven selected CVD-disease outcomes. Predictive performance, measured as $R^2$, of genetic models (PGS + Ancestry), non-genetic models, and integrated models, including genetic and non-genetic predictors, and denoted as "Genetic + Non genetic." Genetic models are shown in red, non-genetic in blue and integrated in green. Significant associations have *p*-values displayed. $0.00\times10^{+00}$ indicates a significance value of $\leq 1\times10^{-300}$. CVD, cardiovascular disease; DLD, dyslipidaemia; HA, heart attack; HTN, hypertension; OBS, obesity; T2D, type 2 diabetes; DBP, diastolic blood pressure; SBP, systolic blood pressure; LDL, low-density lipoprotein; HDL, high-density lipoprotein; TC, total cholesterol; TG, triglycerides

sample size for the prediction of CVD and associated cardiometabolic outcomes in continental Africa across genetic and traditional risk factors. The study supports the use of genomic information for enhanced CVD and associated disease outcome risk stratification in African populations.

PGS of cardiometabolic biomarkers showed significant yet variable prediction within and across traits in continental African populations. As previously observed in other populations, lipid measurements had the highest variance explained, likely due to higher reported heritability compared to obesity/anthropometric, and blood pressure traits [45]. Although cardiometabolic trait PGS predict several CVD-associated outcomes (dyslipidaemia, hypertension and obesity) in African populations, their predictive utility is relatively modest and typically explain less phenotypic variance than conventional non-genetic risk factors [17]. This is in part due to limited GWAS sample sizes within continental populations, although previous research has shown the substantial value of using ancestry-aligned GWAS compared to those derived from genetically distant populations [15, 46].

Nonetheless, biomarker PGS provide a potential preliminary step in PM-based risk stratification in Africa—particularly considering the limited availability of disease-specific cohorts in Africa and the potential to leverage more readily accessible and routinely collected biomarker data for identification of at-risk individuals. As disease-specific variant data becomes available in African populations e.g. CVD, Hypertension, etc., future research should investigate whether models including PGS of both disease outcome and biomarker/trait PGS enhance predictive accuracy and clinical application as seen in other studies [47].

Our research also revealed that combining the PGS of multiple related traits into a single model—MultiPGS—substantially increased the predictive performance for dyslipidaemia, hypertension, obesity, stroke and diabetes compared with single-score predictor models [40, 41, 47]. This improvement is partly due to genetic correlation among traits and leveraging the increased statistical power among discovery GWAS (variance explained and sample size), of either genetically proximal or distantly related traits. Sinnott-Armstrong et al. noted improved predictive accuracy of the aggregated PGS—trait PGS plus 35 cardiometabolic PGS—across multiple disease outcomes in European ancestry populations, albeit with limited generalisability to other populations [47].

By including predictors capturing genetic ancestry through projected principal components, we achieved enhanced risk prediction across our seven phenotypes to

varying degrees. Naret et al., who investigated the inclusion of ancestry in prediction models using a different approach, observed similar improvements and hypothesised that traits that show a greater gain in prediction are likely those that are influenced by more population-specific alleles [42]. This underscores the importance of accounting for the substantial genetic diversity and regional variation in Africa, since African populations are markedly heterogeneous and exhibit greater genetic diversity than any other ancestry group [48, 49]. Consequently, a prediction score developed in one African region may not be universally applicable across all African populations [15, 50]. This regional variation necessitates the development of region-specific models or the inclusion of a wider array of African genetic data in global models to reflect the genetic diversity and population structure of Africans more accurately. For example, in the 1 KG reference data, continental representation is limited to Gambian individuals from the Gambia, Esan and Yoruba in Nigeria, Luhya in Kenya, and Mende in Sierra Leone [30]. This approach inadequately captures genetic diversity among Africans and thus may skew prediction models, highlighting the need for broader genetic data representation.

Integrating genetic factors with non-genetic risk factors showed statistically significant improvement in predicting dyslipidaemia, hypertension, and obesity. We show that genetic factors provide independent but complementary information in risk prediction models, a finding supported by previous research assessing integrated models for CAD and CVD in primarily European populations [9, 10, 51].

This research demonstrates the potential of using genetic information, such as PGS, to improve CVD risk calculators in African populations. It also highlights the need for additional research investigating the generalisability of models across diverse African populations and other ancestries, noting that previous work showed that PGSs built from African Americans increased prediction in sub-Saharan Africans compared to using a European GWAS, but not consistently across African populations [50]. As data from continental African cohorts grow, the opportunities for validation and refinement of these scores is also expected to increase. However, robust assessment of the clinical, financial, and system benefits these scores provide is crucial to gauge their true translational value [52]. Translation will require an understanding of the sensitivity and specificity of scores in different African populations, whilst considering, cost, accessibility, and acceptance within the health ecosystem. In clinical contexts, particularly for PGS, distinguishing between relative and absolute risk is important. Relative risk assesses the connection between genetic traits and disease, contrasting disease incidence in individuals with specific genetic markers against those without. Absolute risk, incorporating factors like age and population disease prevalence, however, reflects the actual probability of disease development. While PGS may indicate increased relative risk, this does not necessarily equate to a high absolute risk. Tools to estimate absolute risk from relative risk [53] require accurate disease prevalence and other data, and their applicability remains limited in the African context due to the paucity of such data. In addition, considerations for resource-constrained settings are essential for the successful integration of genetic approaches into routine practice [54, 55].

This study is unique for its use of data solely from continental Africa, a region often underrepresented in genetic research. However, although the study made use of the largest dataset of continental African populations, the sample size is orders of magnitude smaller than many European and multi-ancestry studies [10, 15, 35]. Some limitations must be noted. Firstly, the APCDR dataset meta-analysis includes two population-based cohorts and two disease-based cohorts, specifically diabetes, which may potentially result to an overestimation of genetic factors associated with diabetes and related conditions in the GWAS results. Secondly, while the pT + clump method has been used given its robustness in calculating PGS in different ancestries, it would be valuable for future research to compare and evaluate the performance of multiple PGS construction methods in African ancestry populations. Thirdly, it is important to acknowledge the extensive diversity within the continent [48], which means that despite both the base and target datasets having representation from multiple African regions, and a similar regional mix, our base and target data are not necessarily ancestrally matched. The strength of this approach lies in its focus on African-specific genetic profiles, addressing a critical gap in current genomic research which often generalises findings from non-African populations, and contributes to a more tailored understanding of CVD prediction and prevention for African populations. However, the extensive variability means that a model developed for one African region may not generalise across the continent. Also, despite using a nested tenfold CV to reduce overfitting, validation in an appropriate continental African dataset was not possible given the scarcity of such cohorts. Also, the relatively small GWAS sample sizes and imputation efficiencies in African GWAS studies may affect the precision of the estimated impact of individual variants on disease risk. Similarly, case numbers were insufficient to examine CVD and associated disease outcomes such as stroke and heart attack. Lastly, previous studies have shown the added value of including PGS in clinical risk

scores, such as Framingham Risk Score and QRISK [16, 56]. AWI-Gen data does not yet have sufficient 10-year longitudinal data available to undertake such performance comparisons.

Future research should employ large, diverse multi-ancestry cohorts, once these become available, to overcome sample size limitations, reduce overfitting, and enhance generalisability. It is important to note existing multi-ancestry cohorts often disproportionately represent African ancestry through African American participants. To more accurately capture genetic diversity and enhance research applicability, it is essential to include a broader representation of African ancestry individuals from regions across Africa. In addition, moving to models that account for gene–gene and gene-environment interactions will further advance our understanding in this field.

## Conclusions

The integrated risk score derived in this study demonstrates the value of including genetic and non-genetic risk information for improving CVD risk prediction in African populations. This approach could provide more accurate and personalised risk assessment, tailoring prevention and treatment strategies more effectively. The inclusion of genetic information improves prediction performance over and above traditional non-genetic factors. In African populations, ancestry accounts for a substantial proportion of variance in CVD prediction models, and modelling this variance through principal components suggests a promising direction for model refinement. However, improving these models will require research across diverse African populations and the use of appropriately ancestry-aligned cohorts. Improving access to extensive African datasets is crucial for refining CVD prediction models and necessary to effectively address health disparities both on the African continent and among global African-ancestry populations.

Supplementary information.

S1. GWAS Catalog study accession numbers, S2. Non genetic factor definitions, S3. Disease outcome definitions, S4. Association testing: Significant within trait associations of derived PGS, S5. Association testing: Significant cross-trait associations of derived PGS, S6. Association testing: Significant associations of derived PGS with CVD-associated disease outcomes, S7. Association testing: Significant associations between non-genetic factors and CVD-associated disease outcomes, S8. Prediction Modelling: MultiPGS model, S9. Prediction Modelling: Genetic models; MultiPGS + Ancestry, S10. Prediction Modelling: Integrated models; Genetic (PGS + Ancestry) and Non-genetic.

**Abbreviations**

| | |
|---|---|
| AADM | African Ancestry Diabetes Mellitus Study |
| APCDR | African Partnership for Chronic Disease Research |
| AWI-Gen | Africa Wits-INDEPTH Partnership for Genomics Research |
| BMI | Body mass index |
| CVD | Cardiovascular disease |
| DBP | Diastolic blood pressure |
| DDS | Durban Diabetes Study |
| DCC | Durban Case–Control Study |
| DLD | Dyslipidaemia |
| GPAQ | Global Physical Activity Questionnaire |
| GWAS | Genome-Wide Association Study |
| HA | Heart attack |
| HDL-C | High-density lipoprotein cholesterol |
| HDSS | Health and Demographic Surveillance System |
| HTN | Hypertension |
| IRS | Integrated risk score |
| LD | Linkage disequilibrium |
| LDL-C | Low-density lipoprotein cholesterol |
| MAF | Minor allele frequency |
| NCV | Nested cross-validation |
| OBS | Obesity |
| PGS | Polygenic score |
| PPCs | Referenced-projected principal components of ancestry |
| pt + clump | *p*-Value thresholding + LD clumping |
| SBP | Systolic blood pressure |
| SNPs | Single nucleotide polymorphisms |
| T2D | Type 2 diabetes |
| TC | Total cholesterol |
| TG | Triglycerides |
| UGR | Uganda Genome Resource |
| W–H ratio | Waist-hip ratio |

## Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s13073-024-01377-6.

Additional file 1: S1. GWAS Catalog study accession numbers, S2. Non genetic factor definitions, S3. Disease outcome definitions, S4. Association testing: Significant within trait associations of derived PGS, S5. Association testing: Significant cross-trait associations of derived PGS, S6. Association testing: Significant associations of derived PGS with CVD-associated disease outcomes, S7. Association testing: Significant associations between non-genetic factors and CVD-associated disease outcomes, S8. Prediction Modelling: MultiPGS model, S9. Prediction Modelling: Genetic models; MultiPGS + Ancestry, S10. Prediction Modelling: Integrated models; Genetic (PGS+ Ancestry) and Non-genetic

**Acknowledgements**

We gratefully acknowledge the time and attention of the participants who generously provided data and donated their samples. The AWI-Gen team, from the field workers, project managers, data managers, laboratory scientists, and all other investigators are gratefully acknowledged for generating and sharing the extraordinary body of data used in this research.

**Authors' contributions**

Conceptualisation, MK, OP, CML, and MR; methodology, MK, OP, CML, and MR; formal analysis, MK, and OP; investigation, MR; resources, MR; data curation, MK; writing—original draft preparation, MK; writing—review and editing, MK, OP, CML, and MR, visualisation, MK; supervision, MR, CML, OP; project administration, MK, MR; funding acquisition, MR. All authors read and approved the final manuscript.

**Funding**

MR holds a South African Research Chair in Genomics and Bioinformatics of African populations hosted by the University of the Witwatersrand, funded by the Department of Science and Innovation, and administered by the National Research Foundation. The research is funded, in part, by a supplement to the

National Institutes of Health H3Africa Consortium Coordinating Centre grant U24HG009780 and the Fogarty International Center of the National Institutes of Health MADIVA grant U54TW012077. The AWI-Gen data and the APC were funded by the H3Africa AWI-Gen Collaborative Centre grant U54HG006938. CML is part-funded by the NIHR Maudsley Biomedical Research Centre at South London and Maudsley NHS Foundation Trust and King's College London. The views expressed are those of the authors and not necessarily those of the NHS, the NIHR, the Department of Health and Social Care in the UK, the US National Institutes of Health, or the South African National Research Foundation.

### Availability of data and materials

The AWI-Gen dataset used in this study is available in the European Genome-phenome Archive (EGA) database (https://ega-archive.org/) under the study accession code EGAD00001006425 (https://ega-archive.org/datasets/EGAD00001006425). The genotype dataset accession code is EGAD00010001996 (https://ega-archive.org/datasets/EGAD00010001996). The availability of these datasets is subject to controlled access through the H3Africa Data and Biospecimen Access Committee. The GWAS summary statistics for APCDR analysed during the current study can be accessed at GWAS catalog under the accession numbers described below. The GenoPred Pipeline (Pain, O, GenoPred Pipeline (https://opain.github.io/GenoPred/)) used to generate the polygenic scores and the code ca be found here (https://opain.github.io/GenoPred/). Similarly, the code used for nested cross validation models can be obtained here.
https://github.com/opain/GenoPred/blob/master/Scripts/Model_builder/Model_builder_V2_nested.R
APCDR GWAS accession numbers and associated GWAS catalog links:Total cholesterol: https://ftp.ebi.ac.uk/pub/databases/gwas/summary_statistics/GCST009001-GCST010000/GCST009042/
Low-density lipoprotein: https://ftp.ebi.ac.uk/pub/databases/gwas/summary_statistics/GCST009001-GCST010000/GCST009043/
High-density lipoprotein: https://ftp.ebi.ac.uk/pub/databases/gwas/summary_statistics/GCST009001-GCST010000/GCST009044/
Triglycerides: https://ftp.ebi.ac.uk/pub/databases/gwas/summary_statistics/GCST009001-GCST010000/GCST009045/
Serum albumin measurement: https://ftp.ebi.ac.uk/pub/databases/gwas/summary_statistics/GCST009001-GCST010000/GCST009048/
Bilirubin measurement: https://ftp.ebi.ac.uk/pub/databases/gwas/summary_statistics/GCST009001-GCST010000/GCST009051/
Diastolic blood pressure: https://ftp.ebi.ac.uk/pub/databases/gwas/summary_statistics/GCST009001-GCST010000/GCST009052/
Systolic blood pressure: https://ftp.ebi.ac.uk/pub/databases/gwas/summary_statistics/GCST009001-GCST010000/GCST009053/
Body height: https://ftp.ebi.ac.uk/pub/databases/gwas/summary_statistics/GCST009001-GCST010000/GCST009055/
Body weight: https://ftp.ebi.ac.uk/pub/databases/gwas/summary_statistics/GCST009001-GCST010000/GCST009056/
Body mass index: https://ftp.ebi.ac.uk/pub/databases/gwas/summary_statistics/GCST009001-GCST010000/GCST009057/
Waist circumference: https://ftp.ebi.ac.uk/pub/databases/gwas/summary_statistics/GCST009001-GCST010000/GCST009058/
Hip circumference: https://ftp.ebi.ac.uk/pub/databases/gwas/summary_statistics/GCST009001-GCST010000/GCST009059/
Waist-hip ratio: https://ftp.ebi.ac.uk/pub/databases/gwas/summary_statistics/GCST009001-GCST010000/GCST009060/

## Declarations

### Ethics approval and consent to participate

The study was conducted in accordance with the Declaration of Helsinki and approved by the Institutional Ethics Committee of the University of the Witwatersrand (HREC(Medical) protocol code M210355 and approval date 10-Jun-2021). Written informed consent was obtained from all participants involved in the study.

### Consent for publication

Not applicable.

### Competing interests

Cathryn M. Lewis is a Research and Development SAB member at Myriad Neuroscience. Oliver Pain provides consultancy services for UCB pharma company. The remaining authors declare that there are no competing interests. The funders had no role in the study's design; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

### Author details

[1]Division of Human Genetics, National Health Laboratory Service and School of Pathology, Faculty of Health Sciences, The University of the Witwatersrand, Johannesburg, South Africa. [2]Sydney Brenner Institute for Molecular Bioscience, Faculty of Health Sciences, University of the Witwatersrand, Johannesburg, South Africa. [3]Social, Genetic and Developmental Psychiatry Centre, Institute of Psychiatry, King's College London, Psychology & Neuroscience, London, UK. [4]Department of Basic and Clinical Neuroscience, Institute of Psychiatry, Psychology and Neuroscience, Maurice Wohl Clinical Neuroscience Institute, King's College London, London, UK. [5]Department of Medical & Molecular Genetics, Faculty of Life Sciences and Medicine, King's College London, London, UK.

Received: 19 December 2023 Accepted: 12 August 2024
Published online: 26 August 2024

### References

1. WHO. Cardiovascular diseases (CVDs) . 2017 . Available from: https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds)[cited 2020 Jan 28]
2. Goff DCJ, Lloyd-Jones DM, Bennett G, Coady S, D'Agostino RB, Gibbons R, et al. 2013 ACC/AHA guideline on the assessment of cardiovascular risk: a report of the American College of Cardiology/American Heart Association Task Force on Practice Guidelines. Circulation. 2014Jun;129(25 Suppl 2):S49-73.
3. Hippisley-Cox J, Coupland C, Brindle P. Development and validation of QRISK3 risk prediction algorithms to estimate future risk of cardiovascular disease: prospective cohort study. BMJ. 2017 May 23;357:j2099. Available from: http://www.bmj.com/content/357/bmj.j2099.abstract
4. Wagner RG, Crowther NJ, Micklesfield LK, Boua PR, Nonterah EA, Mashinya F, et al. Estimating the burden of cardiovascular risk in community dwellers over 40 years old in South Africa, Kenya, Burkina Faso and Ghana. BMJ Glob Heal. 2021 Jan 21 [cited 2022 May 10];6(1). Available from: https://pubmed.ncbi.nlm.nih.gov/33479017/
5. Boateng D, Agyemang C, Beune E, Meeks K, Smeeth L, Schulze MB, et al. Cardiovascular disease risk prediction in sub-Saharan African populations — Comparative analysis of risk algorithms in the RODAM study. Int J Cardiol. 2018;254:310–5. https://doi.org/10.1016/j.ijcard.2017.11.082.
6. Khera AV, Chaffin M, Aragam KG, Haas ME, Roselli C, Choi SH, et al. Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations. Nat Genet. 2018;50(9):1219–24. https://doi.org/10.1038/s41588-018-0183-z.
7. Kullo IJ, Lewis CM, Inouye M, Martin AR, Ripatti S, Chatterjee N. Polygenic scores in biomedical research. Nat Rev Genet. 2022Sep;23(9):524–32.
8. Torkamani A, Wineinger NE, Topol EJ. The personal and clinical utility of polygenic risk scores. Nat Rev Genet. 2018;19(9):581–90. https://doi.org/10.1038/s41576-018-0018-x.
9. Inouye M, Abraham G, Nelson CP, Wood AM, Sweeting MJ, Dudbridge F, et al. Genomic Risk Prediction of Coronary Artery Disease in 480,000 Adults: Implications for Primary Prevention. J Am Coll Cardiol. 2018 Oct 16 [cited 2022 May 11];72(16):1883–93. Available from: https://pubmed.ncbi.nlm.nih.gov/30309464/
10. Riveros-Mckay F, Weale ME, Moore R, Selzam S, Krapohl E, Sivley RM, et al. Integrated Polygenic Tool Substantially Enhances Coronary Artery Disease Prediction. Circ Genomic Precis Med [Internet]. 2021 [cited 2022 May 11];14:192–200. Available from: https://www.ahajournals.org/doi/abs/https://doi.org/10.1161/CIRCGEN.120.003304
11. Martin AR, Kanai M, Kamatani Y, Okada Y, Neale BM, Daly MJ. Clinical use of current polygenic risk scores may exacerbate health disparities. Nat Genet. 2019;51(4):584–91. Available from: https://doi.org/10.1038/s41588-019-0379-x

12. Fatumo S, Sathan D, Samtal C, Isewon I, Tamuhla T, Soremekun C, et al. Polygenic risk scores for disease risk prediction in Africa: current challenges and future directions. Genome Med. 2023;15(1):87. Available from: https://doi.org/10.1186/s13073-023-01245-9
13. Duncan L, Shen H, Gelaye B, Meijsen J, Ressler K, Feldman M, et al. Analysis of polygenic risk score usage and performance in diverse human populations. Nat Commun. 2019 Dec 1 [cited 2021 May 3];10(1):1–9. Available from: https://doi.org/10.1038/s41467-019-11112-0
14. Kachuri L, Chatterjee N, Hirbo J, Schaid DJ, Martin I, Kullo IJ, et al. Principles and methods for transferring polygenic risk scores across global populations. Nat Rev Genet. 2023 Aug;
15. Choudhury A, Brandenburg J-T, Chikowore T, Sengupta D, Boua PR, Crowther NJ, et al. Meta-analysis of sub-Saharan African studies provides insights into genetic architecture of lipid traits. Nat Commun. 2022May;13(1):2578.
16. Weale ME, Riveros-Mckay F, Selzam S, Seth P, Moore R, Tarran WA, et al. Validation of an Integrated Risk Tool, Including Polygenic Risk Score, for Atherosclerotic Cardiovascular Disease in Multiple Ethnicities and Ancestries. Am J Cardiol. 2021Jun;1(148):157–64.
17. Aragam KG, Dobbyn A, Judy R, Chaffin M, Chaudhary K, Hindy G, et al. Limitations of Contemporary Guidelines for Managing Patients at High Genetic Risk of Coronary Artery Disease. J Am Coll Cardiol. 2020Jun;75(22):2769–80.
18. Gurdasani D, Carstensen T, Fatumo S, Chen G, Franklin CS, Prado-Martinez J, et al. Uganda Genome Resource Enables Insights into Population History and Genomic Discovery in Africa. Cell. 2019;179(4).
19. Ramsay M, Crowther N, Tambo E, Agongo G, Baloyi V, Dikotope S, et al. H3Africa AWI-Gen Collaborative Centre: A resource to study the interplay between genomic and environmental risk factors for cardiometabolic diseases in four sub-Saharan African countries. Glob Heal Epidemiol Genomics. 2016;1.
20. Choi SW, O'Reilly PF, PRSice-2: Polygenic Risk Score software for biobank-scale data. Gigascience. 1;8(7):giz082. Available from: 2019J. https://doi.org/10.1093/gigascience/giz082.
21. Pain O, Al-Chalabi A, Lewis C. The GenoPred Pipeline: A Comprehensive and Scalable Pipeline for Polygenic Scoring. 2024.
22. Hird TR, Young EH, Pirie FJ, Riha J, Esterhuizen TM, O'Leary B, et al. Study profile: the Durban Diabetes Study (DDS): a platform for chronic disease research. Glob Heal Epidemiol genomics. 2016;1: e2.
23. Rotimi CN, Chen G, Adeyemo AA, Furbert-Harris P, Guass D, Zhou J, et al. A Genome-Wide Search for Type 2 Diabetes Susceptibility Genes in West Africans : The Africa America Diabetes Mellitus (AADM) Study. Diabetes [Internet]. 2004 Mar 1;53(3):838–41. Available from: https://doi.org/10.2337/diabetes.53.3.838
24. Han B, Eskin E. Random-Effects Model Aimed at Discovering Associations in Meta-Analysis of Genome-wide Association Studies. Am J Hum Genet. 2011;88(5):586–98. Available from: https://www.sciencedirect.com/science/article/pii/S0002929711001558
25. Sollis E, Mosaku A, Abid A, Buniello A, Cerezo M, Gil L, et al. The NHGRI-EBI GWAS Catalog: knowledgebase and deposition resource. Nucleic Acids Res. 2023Jan;51(D1):D977–85.
26. Ali SA, Soo C, Agongo G, Alberts M, Amenga-etego L, Boua, Romuald P. Choudhury A, et al. Genomic and environmental risk factors for cardiometabolic diseases in Africa : methods used for Phase 1 of the AWI-Gen population cross-sectional study. Glob Health Action [Internet]. 2018;11(2). Available from: https://doi.org/10.1080/16549716.2018.1507133
27. D'Agostino RB, Vasan RS, Pencina MJ, Wolf PA, Cobain M, Massaro JM. General cardiovascular risk profile for use in primary care: the Framingham heart study. Circulation. 2008;117(6):743–53.
28. Arnett DK, Blumenthal RS, Albert MA, Buroker AB, Goldberger ZD, Hahn EJ, et al. 2019 ACC/AHA Guideline on the Primary Prevention of Cardiovascular Disease: A Report of the American College of Cardiology/American Heart Association Task Force on Clinical Practice Guidelines. Circulation [Internet]. 2019 Sep 10 [cited 2022 May 10];140(11):e596–646. Available from: https://www.ahajournals.org/doi/abs/https://doi.org/10.1161/CIR.0000000000000678
29. Bull FC, Maslin TS, Armstrong T. Global physical activity questionnaire (GPAQ): nine country reliability and validity study. J Phys Act Health. 2009Nov;6(6):790–804.
30. Genomes Project Consortium. A global reference for human genetic variation. Nature. 2015;526(7571):68.
31. Chang CC, Chow CC, Tellier LCAM, Vattikuti S, Purcell SM, Lee JJ. Second-generation PLINK: rising to the challenge of larger and richer datasets. Gigascience. 2015 Dec 1;4(1):s13742–015–0047–8. Available from: https://doi.org/10.1186/s13742-015-0047-8
32. Price AL, Weale ME, Patterson N, Myers SR, Need AC, Shianna K V, et al. Long-range LD can confound genome scans in admixed populations. Vol. 83, American journal of human genetics. United States; 2008. p. 132–9.
33. Friedman JH, Hastie T, Tibshirani R. Regularization Paths for Generalized Linear Models via Coordinate Descent. J Stat Softw [Internet]. 2010 Feb 2;33(1 SE-Articles):1–22. Available from: https://www.jstatsoft.org/index.php/jss/article/view/v033i01
34. Ramsay M, Crowther NJ, Agongo G, Ali SA, Asiki G, Boua RP, et al. Regional and sex-specific variation in BMI distribution in four sub-Saharan African countries: The H3Africa AWI-Gen study. Glob Health Action. 2018;11(sup2):1556561.
35. George JA, Brandenburg JT, Fabian J, Crowther NJ, Agongo G, Alberts M, et al. Kidney damage and associated risk factors in rural and urban sub-Saharan Africa (AWI-Gen): a cross-sectional population study. Lancet Glob Heal [Internet]. 2019;7(12):e1632–43. Available from: https://doi.org/10.1016/S2214-109X(19)30443-7
36. Gómez-Olivé FX, Ali SA, Made F, Kyobutungi C, Nonterah E, Micklesfield L, et al. Regional and Sex Differences in the Prevalence and Awareness of Hypertension: An H3Africa AWI-Gen Study Across 6 Sites in Sub-Saharan Africa. Glob Heart. 2017;12(2):81–90.
37. Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D. Principal components analysis corrects for stratification in genome-wide association studies. Nat Genet. 2006Aug;38(8):904–9.
38. Chen C-Y, Han J, Hunter DJ, Kraft P, Price AL. Explicit Modeling of Ancestry Improves Polygenic Risk Scores and BLUP Prediction. Genet Epidemiol. 2015Sep;39(6):427–38.
39. Zou H, Hastie T. Regularization and Variable Selection Via the Elastic Net. J R Stat Soc Ser B Stat Methodol. 2005 Apr 1;67(2):301–20. Available from: https://doi.org/10.1111/j.1467-9868.2005.00503.x
40. Krapohl E, Patel H, Newhouse S, Curtis CJ, von Stumm S, Dale PS, et al. Multi-polygenic score approach to trait prediction. Mol Psychiatry. 2018May;23(5):1368–74.
41. Pain O, Glanville KP, Hagenaars SP, Selzam S, Fürtjes AE, Gaspar HA, Evaluation of polygenic prediction methodology within a reference-standardized framework. PLOS Genet ., et al. 4;17(5):e1009021. Available from: 2021May. https://doi.org/10.1371/journal.pgen.1009021.
42. Naret O, Kutalik Z, Hodel F, Xu ZM, Marques-Vidal P, Fellay J. Improving polygenic prediction with genetically inferred ancestry. Hum Genet Genomics Adv. 2022;3(3).
43. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet. 2007Sep;81(3):559–75.
44. R Core Team. R: A language and environment for statistical computing . Vienna, Austria: R Foundation for Statistical Computing; 2021. Available from: https://www.r-project.org/
45. Ekoru K, Adeyemo AA, Chen G, Doumatey AP, Zhou J, Bentley AR, et al. Genetic risk scores for cardiometabolic traits in sub-Saharan African populations. Int J Epidemiol. 2021Aug;50(4):1283–96.
46. Privé F, Aschard H, Carmi S, Folkersen L, Hoggart C, O'Reilly PF, et al. Portability of 245 polygenic scores when derived from the UK Biobank and applied to 9 ancestry groups from the same cohort. Am J Hum Genet. 2022Jan;109(1):12–23.
47. Sinnott-Armstrong N, Tanigawa Y, Amar D, Mars N, Benner C, Aguirre M, et al. Genetics of 35 blood and urine biomarkers in the UK Biobank. Nat Genet. 2021Feb;53(2):185–94.
48. Choudhury A, Aron S, Botigué LR, Sengupta D, Botha G, Bensellak T, et al. High-depth African genomes inform human migration and health. Nature. 2020 Oct 1 [cited 2021 May 11];586(7831):741–8. Available from: https://doi.org/10.1038/s41586-020-2859-7
49. Gurdasani D, Carstensen T, Tekola-Ayele F, Pagani L, Tachmazidou I, Hatzikotoulas K, et al. The African Genome Variation Project shapes medical genetics in Africa. Nature. 2015Jan 15;517(7534):327–32.
50. Kamiza AB, Toure SM, Vujkovic M, Machipisa T, Soremekun OS, Kintu C, et al. Transferability of genetic risk scores in African populations.

Nat Med. 2022;28(6):1163–6. Available from: https://doi.org/10.1038/s41591-022-01835-x
51. Elliott J, Bodinier B, Bond TA,Marc Chadeau-Hyam, Evangelos Evangelou KGMM, Dehghan A, Muller DC, et al. Predictive Accuracy of a Polygenic Risk Score-Enhanced Prediction Model vs a Clinical Risk Score for Coronary Artery Disease. JAMA. 2020;323(7):636–45.
52. Lewis ACF, Green RC, Vassy JL. Polygenic risk scores in the clinic: Translating risk into action. HGG Adv. 2021Oct;2(4): 100047.
53. Pain O, Gillett AC, Austin JC, Folkersen L, Lewis CM. A tool for translating polygenic scores onto the absolute scale using summary statistics. Eur J Hum Genet. 2022;30(3):339–48. Available from: https://doi.org/10.1038/s41431-021-01028-z
54. Kamp M, Krause A, Ramsay M. Has translational genomics come of age in Africa? Hum Mol Genet. 2021 Oct 1 [cited 2021 Oct 7];30(R2):R164–73. Available from: https://academic.oup.com/hmg/article/30/R2/R164/6316672
55. Chikowore T, Kamiza AB, Oduaran OH, Machipisa T, Fatumo S. Non-communicable diseases pandemic and precision medicine: Is Africa ready? EBioMedicine. 2021Mar;1(65): 103260.
56. Tamlander M, Mars N, Pirinen M, Widén E, Ripatti S. Integration of questionnaire-based risk factors improves polygenic risk scores for human coronary heart disease and type 2 diabetes. Commun Biol 2022 51. 2022 Feb 23 [cited 2022 May 10];5(1):1–13. Available from: https://www.nature.com/articles/s42003-021-02996-0

## Publisher's Note