nature genetics

Supplementary information

https://doi.org/10.1038/s41588-024-01792-w

Calibrated prediction intervals for polygenic scores across diverse contexts

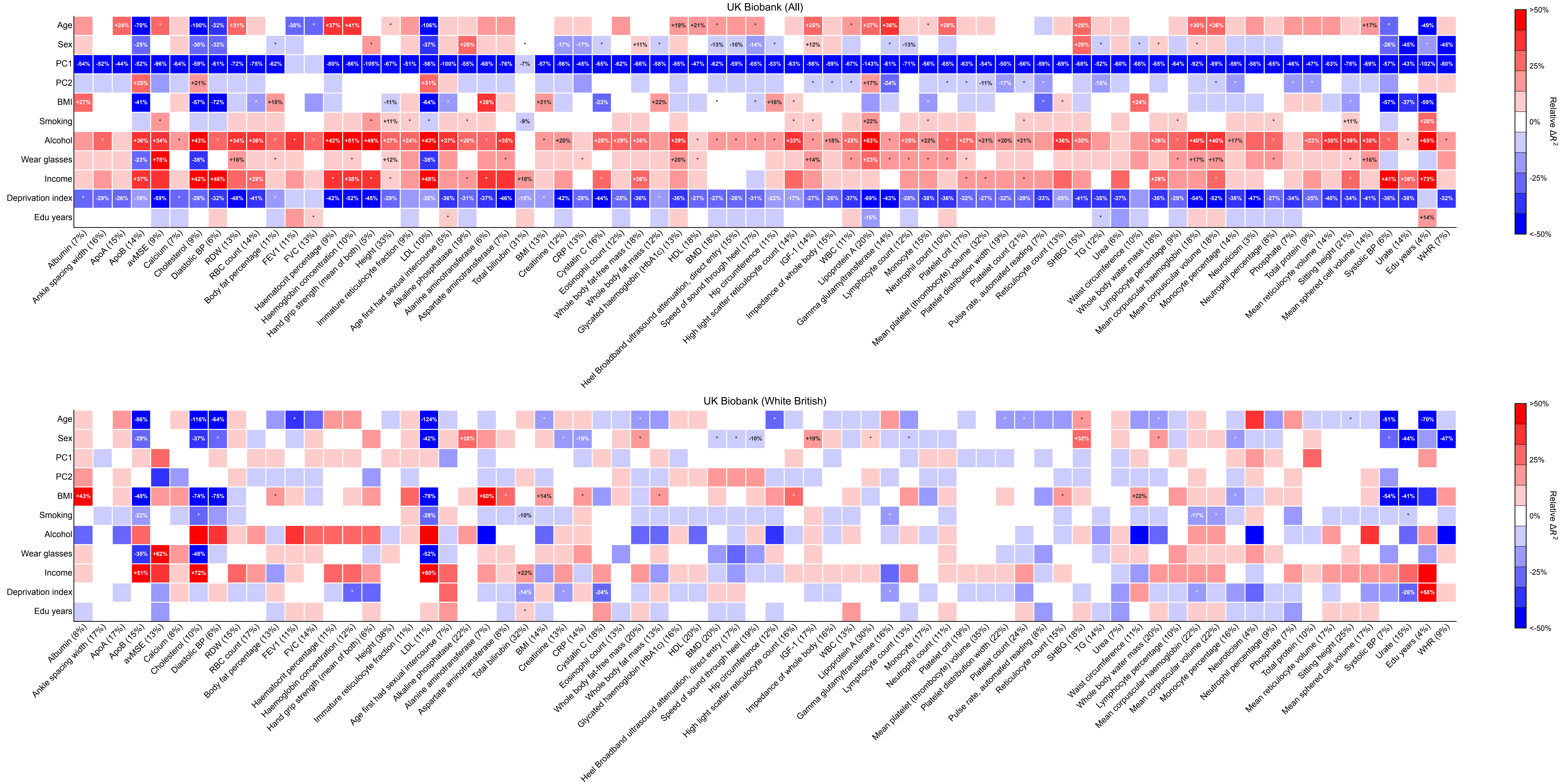
Supplementary Information

Calibrated prediction intervals for polygenic scores across diverse contexts

Supplementary Figures

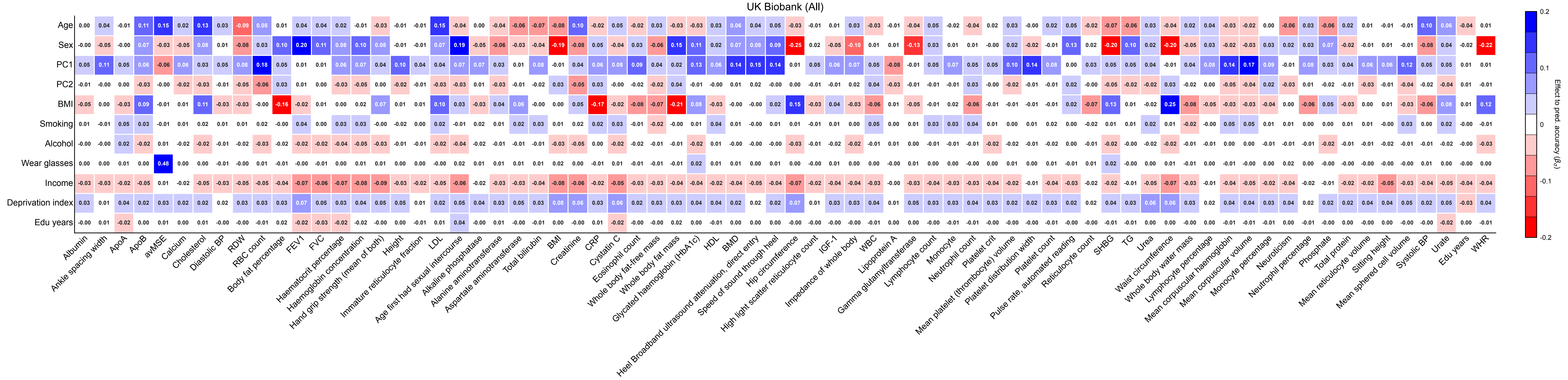
(see next page)

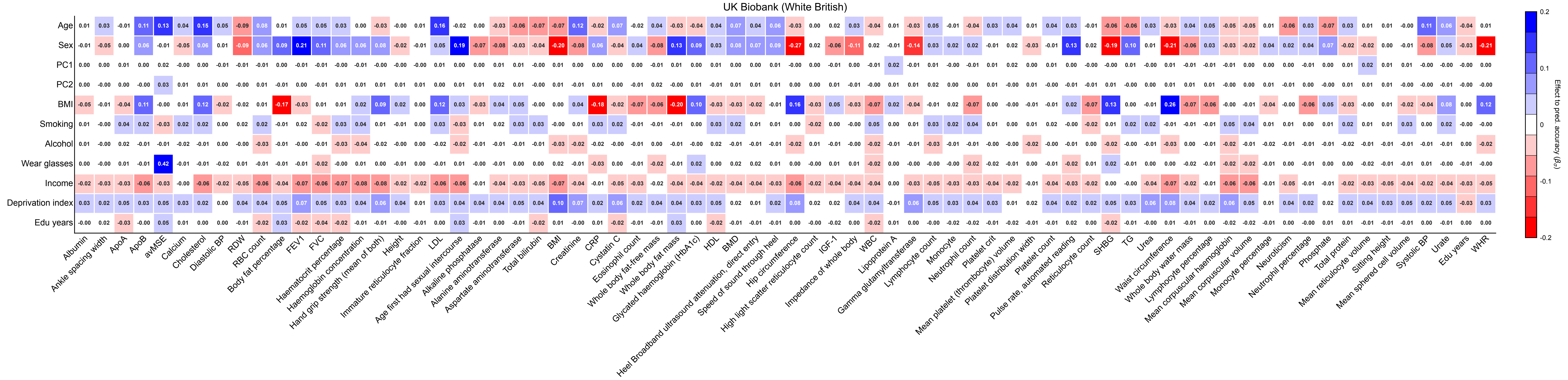
Supplementary Figure 1: Relative ΔR^2 for “white British” and all individuals in UK Biobank. Numerical values of relative ΔR^2 are displayed for trait-context pairs with statistically significant differences (multiple testing correction for all 72×11 trait-context pairs in this figure; two-sided $p < 0.05 / (72\times11)$). ‘*’ are displayed for context-trait pairs with nominally significant differences (multiple testing correction for 11 contexts; two-sided $p < 0.05 / 11$). See Fig. 2 caption for additional details. Numerical results are reported in Supplementary Table 2.

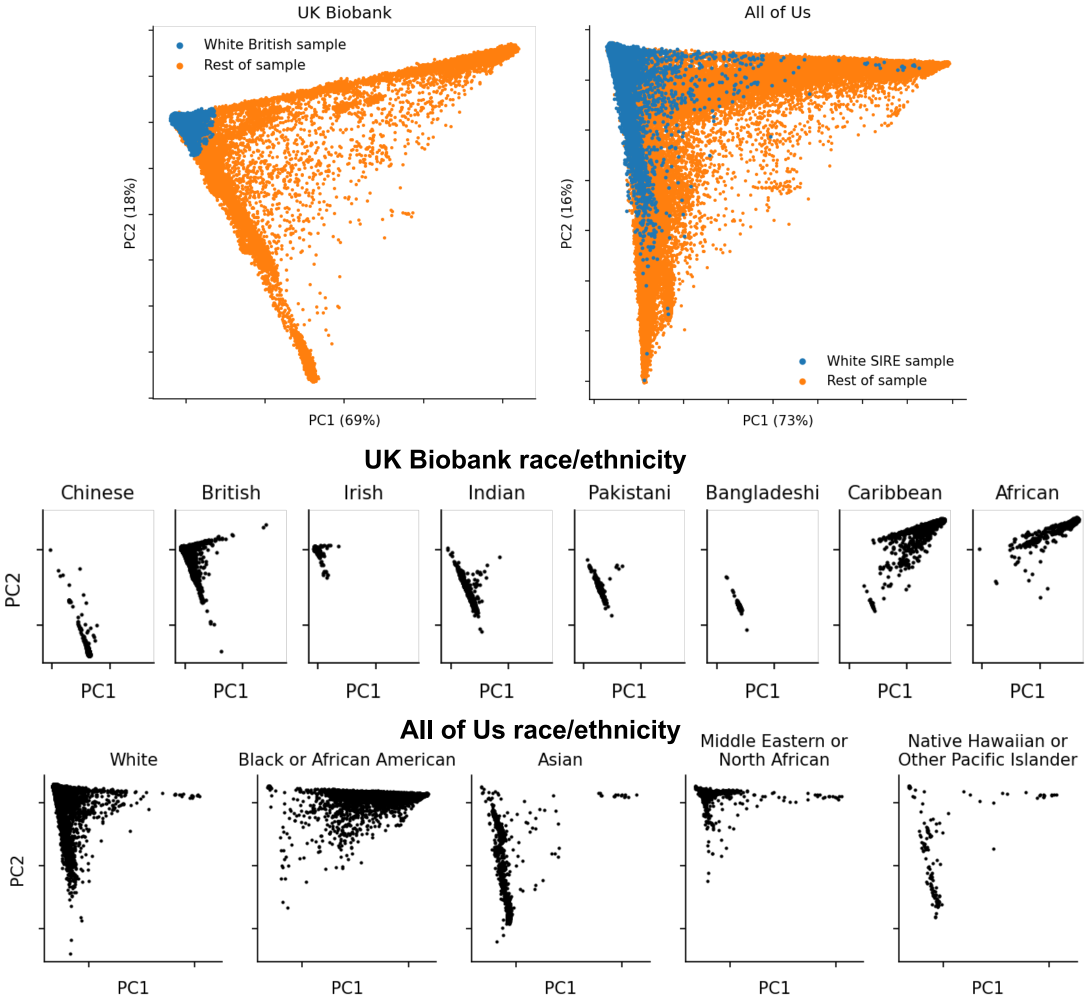

UK Biobank (All)
Age
Sex
PC1
PC2
BMI
Smoking
Alcohol
Wear glasses
Income
Deprivation index
Edu years
Albumin (7%)
Ankle spacing width (16%)
ApoA (15%)
ApoB (14%)
avMSE (9%)
Calcium (7%)
Cholesterol (9%)
Diastolic BP (6%)
RDW (13%)
RBC count (14%)
Body fat percentage (11%)
FEV1 (11%)
FVC (13%)
Haematocrit percentage (9%)
Haemoglobin concentration (10%)
Hand grip strength (mean of both) (5%)
Height (33%)
Immature reticulocyte fraction (9%)
LDL (10%)
Age first had sexual intercourse (5%)
Alkaline phosphatase (19%)
Alanine aminotransferase (6%)
Aspartate aminotransferase (7%)
Total bilirubin (31%)
BMI (13%)
Creatinine (12%)
CRP (13%)
Cystatin C (16%)
Eosinophil count (12%)
Whole body fat-free mass (18%)
Whole body fat mass (12%)
Glycated haemoglobin (HbA1c) (13%)
HDL (18%)
BMD (18%)
Heel Broadband ultrasound attenuation, direct entry (15%)
Speed of sound through heel (17%)
Hip circumference (11%)
High light scatter reticulocyte count (14%)
IGF-1 (14%)
Impedance of whole body (15%)
WBC (11%)
Lipoprotein A (20%)
Gamma glutamyltransferase (14%)
Lymphocyte count (12%)
Monocyte (15%)
Neutrophil count (10%)
Platelet crit (17%)
Mean platelet (thrombocyte) volume (32%)
Platelet distribution width (19%)
Platelet count (21%)
Pulse rate, automated reading (7%)
Reticulocyte count (13%)
SHBG (15%)
TG (12%)
Urea (6%)
Waist circumference (10%)
Whole body water mass (18%)
Lymphocyte percentage (9%)
Mean corpuscular haemoglobin (18%)
Mean corpuscular volume (18%)
Monocyte percentage (14%)
Neuroticism (3%)
Neutrophil percentage (8%)
Phosphate (7%)
Total protein (9%)
Mean reticulocyte volume (14%)
Sitting height (21%)
Mean sphered cell volume (14%)
Systolic BP (6%)
Urate (14%)
Edu years (4%)
WHR (7%)
>50%
25%
0%
-25%
<-50%
Relative ΔR²
UK Biobank (White British)
Age
Sex
PC1
PC2
BMI
Smoking
Alcohol
Wear glasses
Income
Deprivation index
Edu years
Albumin (8%)
Ankle spacing width (17%)
ApoA (17%)
ApoB (15%)
avMSE (13%)
Calcium (8%)
Cholesterol (10%)
Diastolic BP (6%)
RDW (15%)
RBC count (17%)
Body fat percentage (13%)
FEV1 (11%)
FVC (14%)
Haematocrit percentage (11%)
Haemoglobin concentration (12%)
Hand grip strength (mean of both) (6%)
Height (38%)
Immature reticulocyte fraction (11%)
LDL (11%)
Age first had sexual intercourse (7%)
Alkaline phosphatase (22%)
Alanine aminotransferase (7%)
Aspartate aminotransferase (8%)
Total bilirubin (32%)
BMI (14%)
Creatinine (13%)
CRP (14%)
Cystatin C (18%)
Eosinophil count (13%)
Whole body fat-free mass (20%)
Whole body fat mass (13%)
Glycated haemoglobin (HbA1c) (16%)
HDL (20%)
BMD (20%)
Heel Broadband ultrasound attenuation, direct entry (17%)
Speed of sound through heel (19%)
Hip circumference (12%)
High light scatter reticulocyte count (16%)
IGF-1 (17%)
Impedance of whole body (16%)
WBC (13%)
Lipoprotein A (30%)
Gamma glutamyltransferase (16%)
Lymphocyte count (13%)
Monocyte (17%)
Neutrophil count (11%)
Platelet crit (19%)
Mean platelet (thrombocyte) volume (35%)
Platelet distribution width (22%)
Platelet count (24%)
Pulse rate, automated reading (8%)
Reticulocyte count (15%)
SHBG (18%)
TG (14%)
Urea (7%)
Waist circumference (11%)
Whole body water mass (20%)
Lymphocyte percentage (10%)
Mean corpuscular haemoglobin (22%)
Mean corpuscular volume (22%)
Monocyte percentage (16%)
Neuroticism (4%)
Neutrophil percentage (9%)
Phosphate (7%)
Total protein (10%)
Mean reticulocyte volume (17%)
Sitting height (25%)
Mean sphered cell volume (17%)
Systolic BP (7%)
Urate (15%)
Edu years (4%)
WHR (9%)
>50%
25%
0%
-25%
<-50%
Relative ΔR²

(see next page)

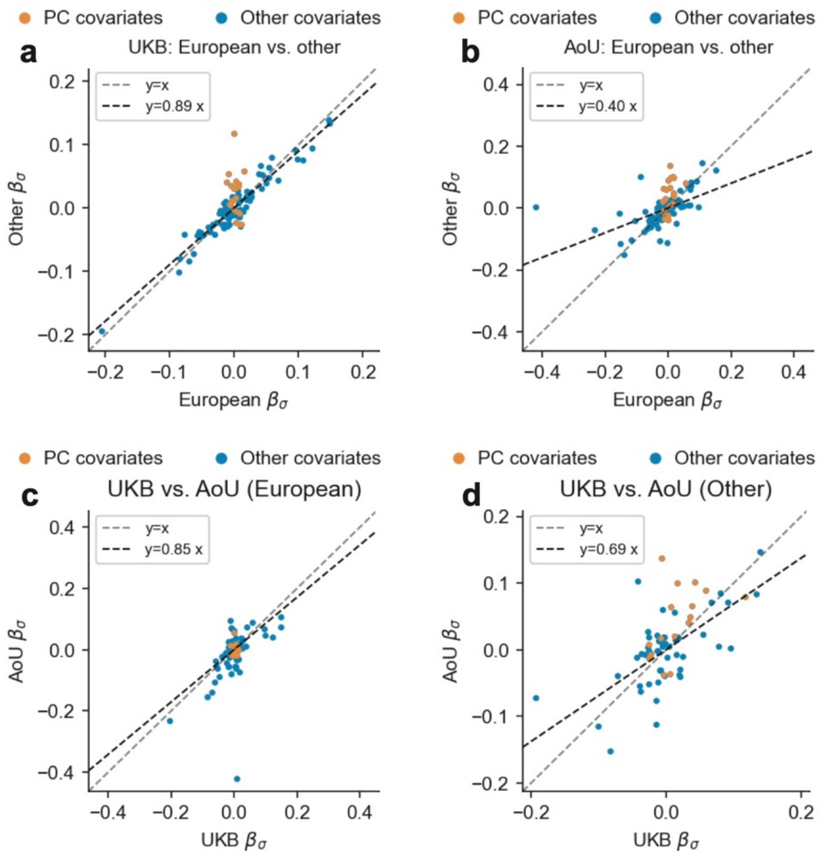
Supplementary Figure 2: Estimated β_{σ} for "white British" and all individuals in UK Biobank. We show estimated β_{σ} in CalPred model. Numerical results are reported in Supplementary Table 2.


UK Biobank (All)
Age
Sex
PC1
PC2
BMI
Smoking
Alcohol
Wear glasses
Income
Deprivation index
Edu years
Albumin
Ankle spacing width
ApoA
ApoB
avMSE
Calcium
Cholesterol
Diastolic BP
RDW
RBC count
Body fat percentage
FEV1
FVC
Haematocrit percentage
Haemoglobin concentration
Hand grip strength (mean of both)
Height
Immature reticulocyte fraction
Age first had sexual intercourse
LDL
Alkaline phosphatase
Alanine aminotransferase
Aspartate aminotransferase
Total bilirubin
BMI
Creatinine
CRP
Cystatin C
Eosinophil count
Whole body fat-free mass
Whole body fat mass
Glycated haemoglobin (HbA1c)
HDL
BMD
Heel Broadband ultrasound attenuation, direct entry
Speed of sound through heel
Hip circumference
High light scatter reticulocyte count
IGF-1
Impedance of whole body
WBC
Lipoprotein A
Gamma glutamyltransferase
Lymphocyte count
Monocyte
Neutrophil count
Platelet crit
Mean platelet (thrombocyte) volume
Platelet distribution width
Platelet count
Pulse rate, automated reading
Reticulocyte count
SHBG
TG
Urea
Waist circumference
Whole body water mass
Lymphocyte percentage
Mean corpuscular haemoglobin
Mean corpuscular volume
Monocyte percentage
Neuroticism
Neutrophil percentage
Phosphate
Total protein
Mean reticulocyte volume
Sitting height
Mean sphered cell volume
Systolic BP
Urate
Edu years
WHR
0.2
0.1
0
-0.1
-0.2
Effect to pred. accuracy (β_r)

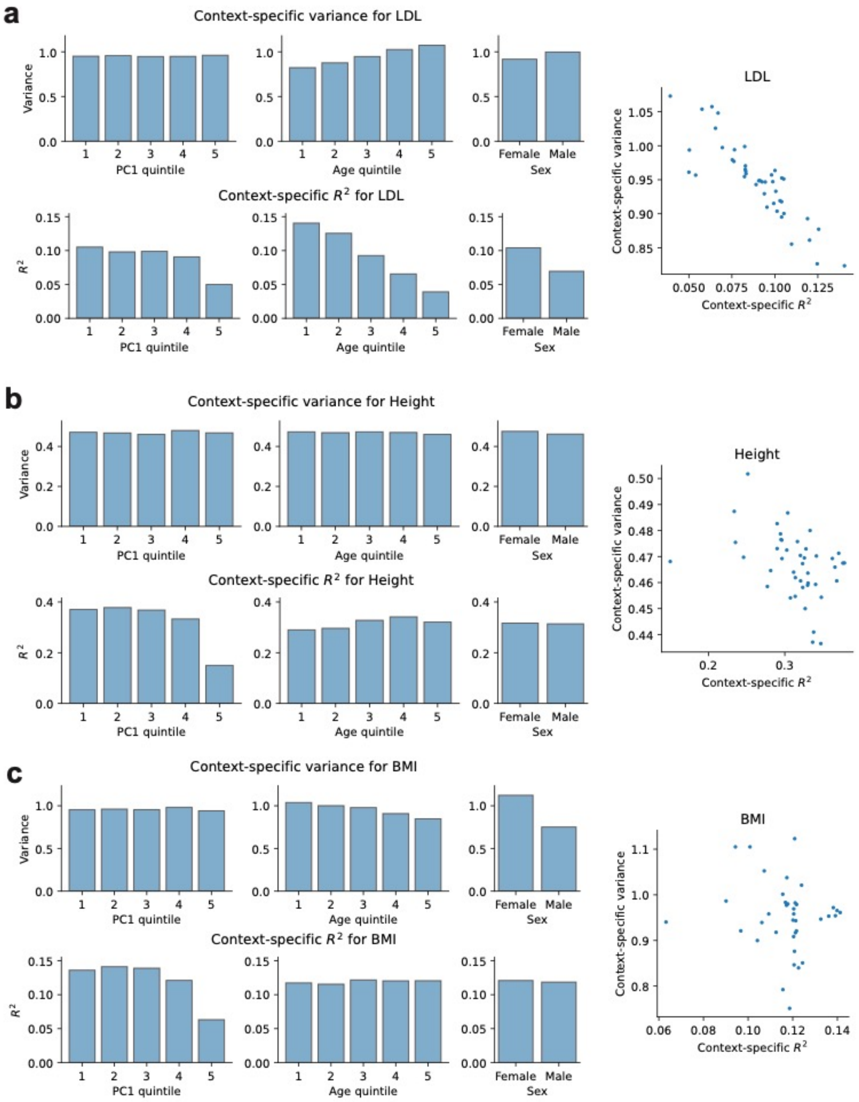

UK Biobank (White British)
Age
Sex
PC1
PC2
BMI
Smoking
Alcohol
Wear glasses
Income
Deprivation index
Edu years
Albumin
Ankle spacing width
ApoA
ApoB
avMSE
Calcium
Cholesterol
Diastolic BP
RDW
RBC count
Body fat percentage
FEV1
FVC
Haematocrit percentage
Haemoglobin concentration
Hand grip strength (mean of both)
Height
Immature reticulocyte fraction
Age first had sexual intercourse
LDL
Alkaline phosphatase
Alanine aminotransferase
Aspartate aminotransferase
Total bilirubin
BMI
Creatinine
CRP
Cystatin C
Eosinophil count
Whole body fat-free mass
Whole body fat mass
Glycated haemoglobin (HbA1c)
HDL
BMD
Heel Broadband ultrasound attenuation, direct entry
Speed of sound through heel
Hip circumference
High light scatter reticulocyte count
IGF-1
Impedance of whole body
WBC
Lipoprotein A
Gamma glutamyltransferase
Lymphocyte count
Monocyte
Neutrophil count
Platelet crit
Mean platelet (thrombocyte) volume
Platelet distribution width
Platelet count
Pulse rate, automated reading
Reticulocyte count
SHBG
TG
Urea
Waist circumference
Whole body water mass
Lymphocyte percentage
Mean corpuscular haemoglobin
Mean corpuscular volume
Monocyte percentage
Neuroticism
Neutrophil percentage
Phosphate
Total protein
Mean reticulocyte volume
Sitting height
Mean sphered cell volume
Systolic BP
Urate
Edu years
WHR
0.2
0.1
0
-0.1
-0.2
Effect to pred. accuracy (β_r)



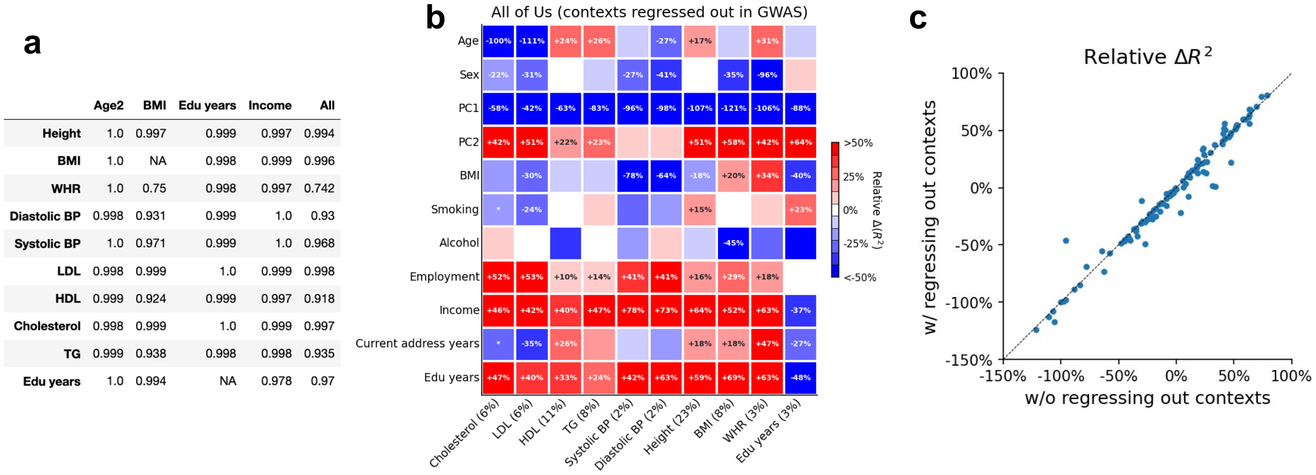
Supplementary Figure 3: Principal components calculated in UK Biobank and All of Us. PC1 and PC2 were calculated across all individuals separately in UK Biobank and All of Us. We show self-reported race/ethnicity in UK Biobank and All of Us to help interpret PC1/PC2. We show the proportion of variance explained by PC1 and PC2 out of top ten PCs calculated in each dataset in x-axis and y-axis labels.



Supplementary Figure 4: Comparison of fitted parameters across populations and biobanks. We compare estimated β_σ across populations **(a-b)** and biobanks **(c-d)**. Each dot denotes a trait-context pair. We separately annotate genetic principal component contexts and other contexts, because PC contexts only have small variations within Europeans ("white British" in UK Biobank or "white SIRE" in All of Us), therefore are not comparable between European and other populations. The regression slope is calculated across all estimated β_σ. For **(c-d)**, we include traits that are shared across biobanks. Overall, we find that β_σ are highly consistent across populations and biobanks.

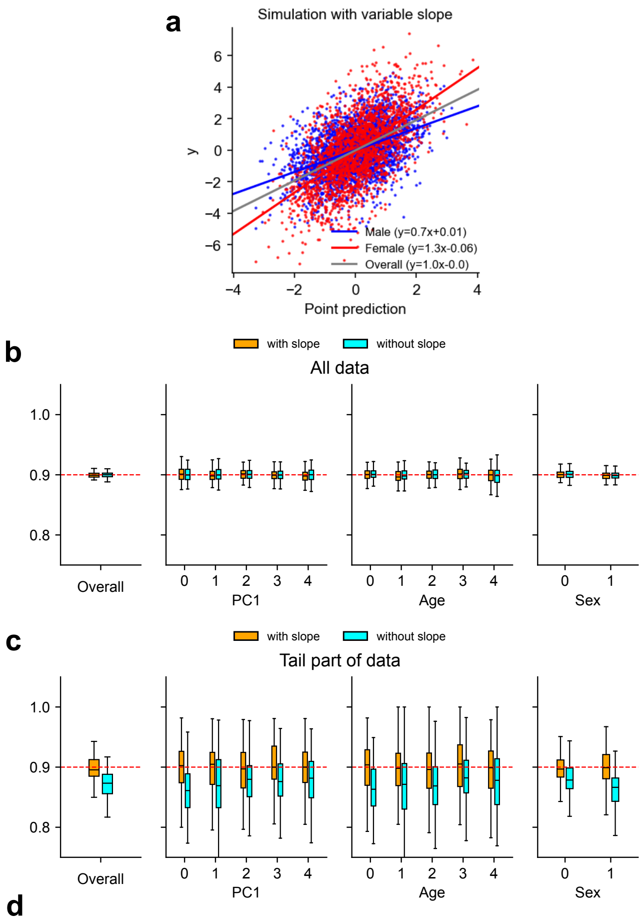


Supplementary Figure 5: R^2 and phenotypic variance across context groups for example traits. We calculate R^2 between PGS and covariate-adjusted phenotypes as well as phenotypic variance by PC1 quintile, age quintile and sex for LDL, height and BMI across all individuals in UK Biobank**.** We determined that variable R^2 across contexts were not solely driven by differences of phenotype variance in context strata. The relationship between R^2 and phenotypic variance depends on the specific trait-context being studied. For example, height R^2 varies across age quintiles while the phenotypic variance remains relatively constant; BMI R^2 stays relatively constant across age quintiles while the phenotypic variance varies.



	Age2	BMI	Edu years	Income	All
Height	1.0	0.997	0.999	0.997	0.994
BMI	1.0	NA	0.998	0.999	0.996
WHR	1.0	0.75	0.998	0.997	0.742
Diastolic BP	0.998	0.931	0.999	1.0	0.93
Systolic BP	1.0	0.971	0.999	1.0	0.968
LDL	0.998	0.999	1.0	0.999	0.998
HDL	0.999	0.924	0.999	0.997	0.918
Cholesterol	0.998	0.999	1.0	0.999	0.997
TG	0.999	0.938	0.998	0.998	0.935
Edu years	1.0	0.994	NA	0.978	0.97

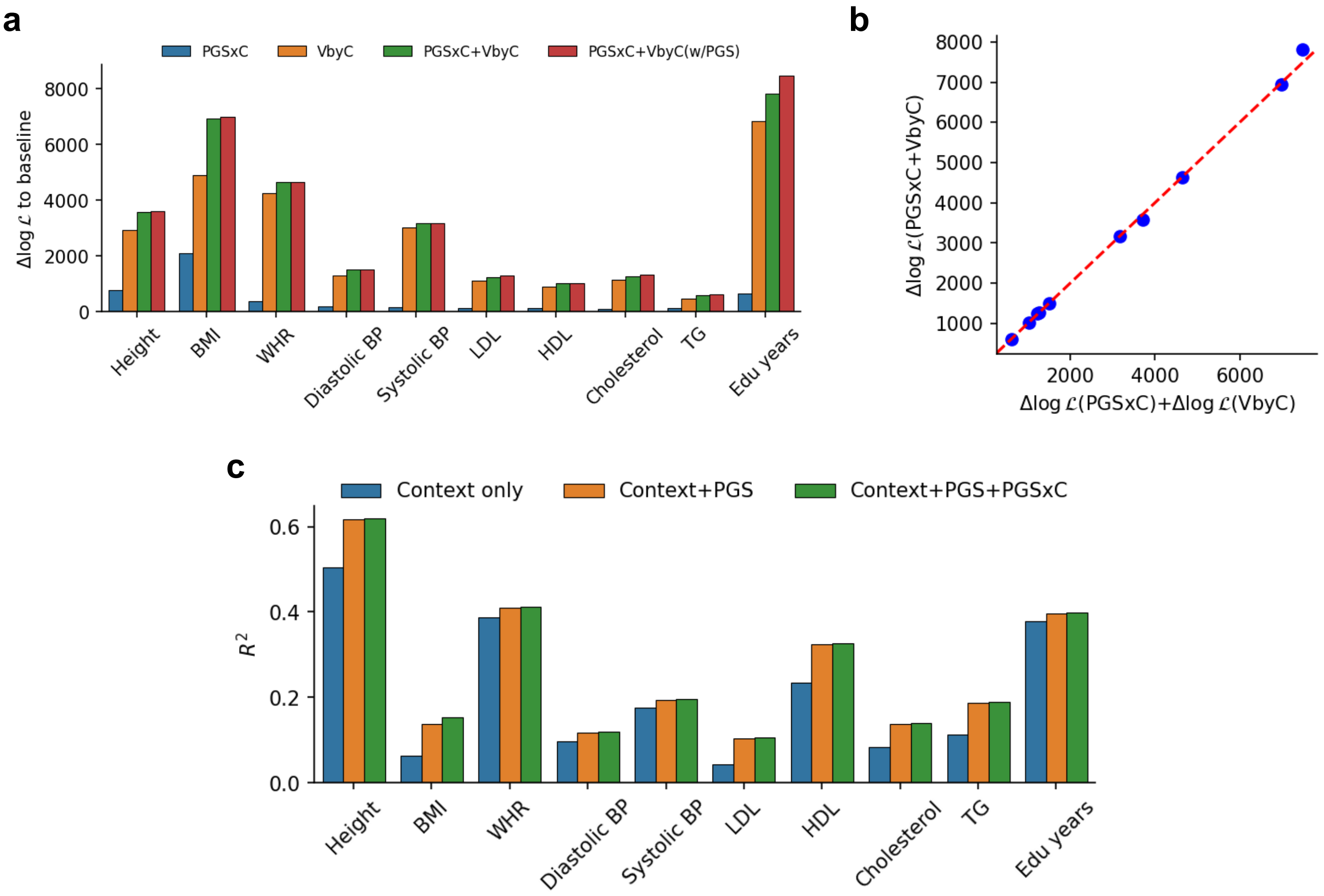
Supplementary Figure 6: (a) Correlation of PGS evaluated across All of Us individuals derived from UK Biobank GWAS when (1) regressing out age, sex, top 16 PCs and (2) separately regressing out age^2, BMI, education years, income, or all of these contexts in addition to age, sex, top 16 PCs. By comparing results of separately regressing out each context with results regressing out all contexts, we determined that regressing out BMI had the most impact to context-specific accuracy. An exception is WHR regressed out of BMI, where PGS for WHR is only weakly correlated with PGS for WHR regressed out of BMI (correlation=0.75). This is expected because WHR regressed out of BMI (WHRadjBMI) has been routinely used as a standard derived phenotype as a proxy of abdominal adiposity and it has been shown that WHRadjBMI had moderate genetic overlap with WHR (Locke 2015 Nature[1]). **(b)** Context-specificity patterns for PGS derived from GWAS regressing out contexts of age^2, BMI, education years, income. These context-specificity patterns were similar to those in Fig. 3 without regressing out these contexts. **(c)** Consistency of relative ΔR^2 for all context-trait pairs, with or without regressing out additional contexts when performing GWAS.



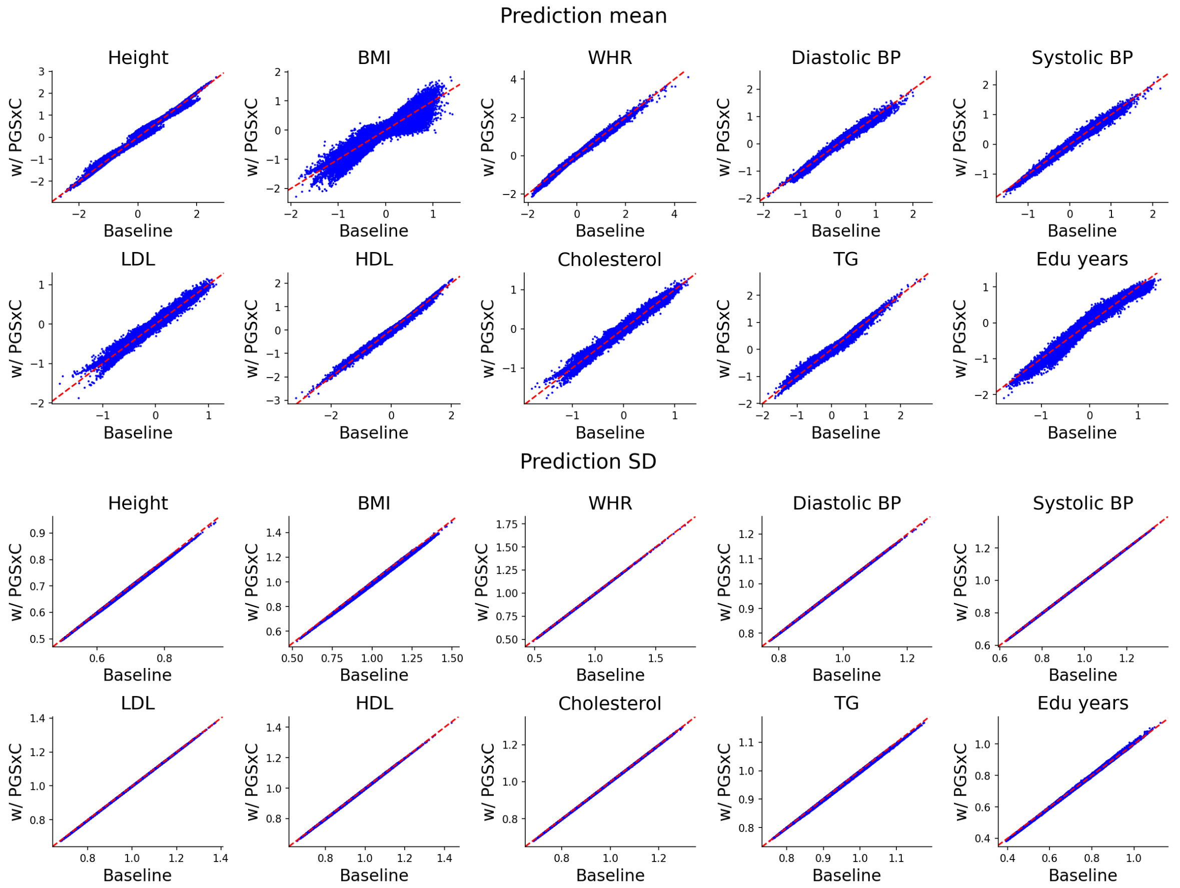
d

	Constant slope		Variable slope (with slope)		Variable slope (without slope)	
	True	95% CI	True	95% CI	True	95% CI
Age var	0.25	[0.244, 0.252]	0.25	[0.244, 0.252]	0.25	[0.234, 0.243]
Sex var	0.20	[0.192, 0.200]	0.20	[0.192, 0.200]	0.20	[0.209, 0.217]
PC1 var	0.15	[0.143, 0.153]	0.15	[0.143, 0.153]	0.15	[0.139, 0.148]
Age slope	0.00	[-0.005, 0.004]	0.00	[-0.005, 0.004]	0.00	NA
Sex slope	0.00	[-0.005, 0.004]	0.30	[0.296, 0.305]	0.30	NA
PC1 slope	0.00	[-0.006, 0.003]	-0.05	[-0.056, -0.047]	-0.05	NA
Prediction beta	1.00	[0.994, 1.003]	1.00	[0.994, 1.003]	1.00	[0.937, 0.946]

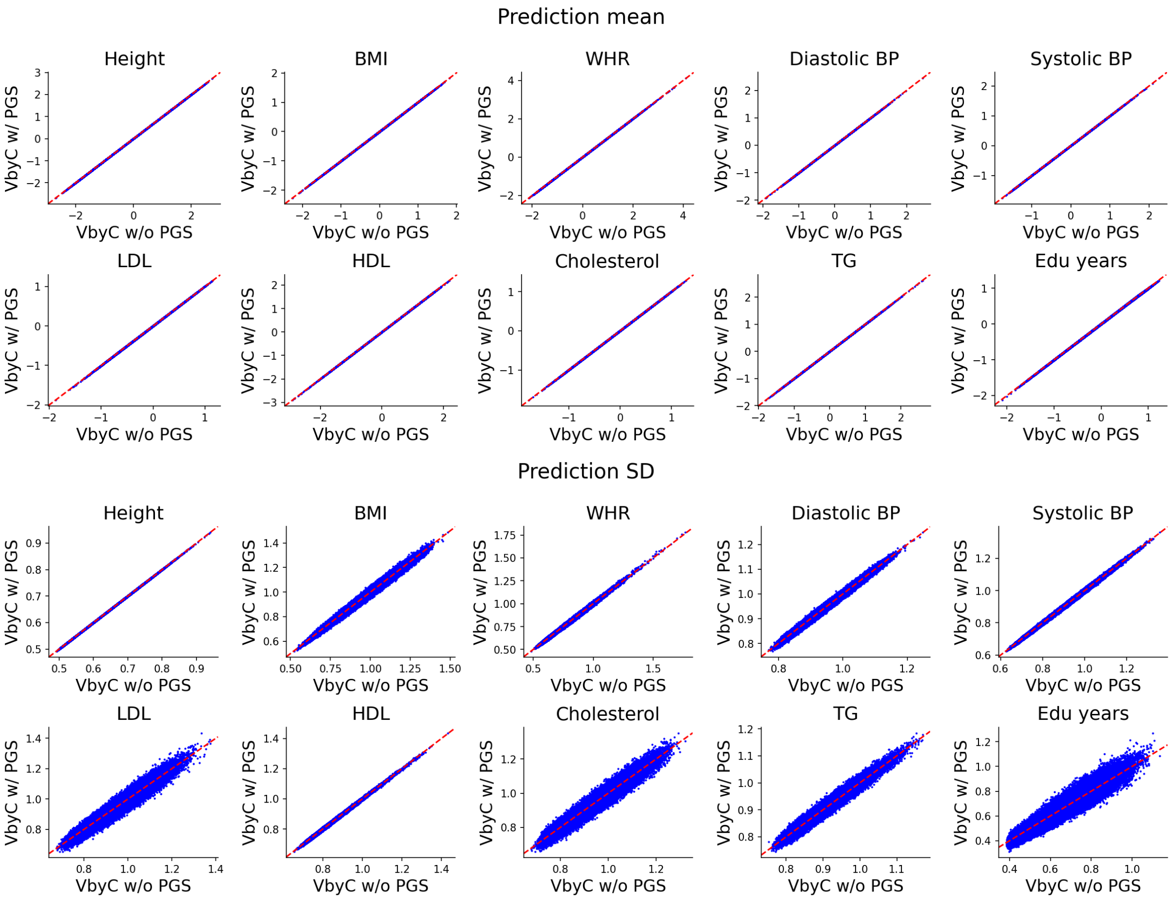
Supplementary Figure 7: Simulations with varying phenotype-prediction slopes. We simulated variable slopes (in regressing phenotypes against point predictions) in addition to variable prediction variances in Fig. 5. We applied CalPred with or without fitting the slope parameters (denoted by ‘with slope’ / ‘without slope’) and evaluated the coverage of prediction intervals. We simulate point predictions as $y \sim \mathcal{N}\left(\hat{y} \times \left(1 + \sum \beta_{\eta,c} \times c\right), \exp\left(\beta_{\sigma,0} + \sum_c \beta_{\sigma,c} \times c\right)\right.$, where true $\beta_{\eta,\text{age}} = 0, \beta_{\eta,\text{sex}} = 0.3, \beta_{\eta,\text{PC1}} = -0.05$. Other simulation settings were the same as in Fig. 5. **(a)** Example data with variable slope. Because $\beta_{\eta,\text{sex}} = 0.3$, slopes between phenotype and point predictions were different between male and female. When interaction parameters β_η were not fitted in the calibration, an overall fitted line was obtained. Consequently, at the extreme of point predictions, prediction intervals would have lower-than-expected coverages because of the shift in the mean predictions. **(b)** Coverage of prediction intervals evaluated across all 5,000 target individuals. CalPred produced prediction intervals with expected coverage level at the overall level regardless of whether interaction term is fitted or not. **(c)** Coverage was evaluated across 5% individuals at the tail distribution of point predictions (left tail~2.5%, right tail ~2.5%). When applied without fitted slope, prediction intervals had lower-than-expected coverage because of biased point predictions. When applied with fitted slope, prediction intervals resumed well-calibration at extremes of the distribution. **(d)** Numerical results of parameter estimation. We report the parameter values and 95% confidence intervals of estimated parameters across 100 simulations. ‘Constant slope’ column denotes simulations in Fig. 5. ‘Variable slope (with slope)’ / ‘Variable slope (without slope)’ denotes simulations in **(a-c)** with / without fitting interaction term. We determined that when model is correctly specified (‘Constant slope’ and ‘Variable slope (with slope)’), parameter estimation was unbiased. When model is mis-specified (‘Variable slope (without slope)’), parameter estimation remained robust. For (b) and (c), each box plot contains results across 100 simulations (each box contains n=100 points). For box plots, the center corresponds to the median; the box represents the first and third quartiles of the points; the whiskers represent the minimum and maximum points located within 1.5× interquartile ranges from the first and third quartiles, respectively.



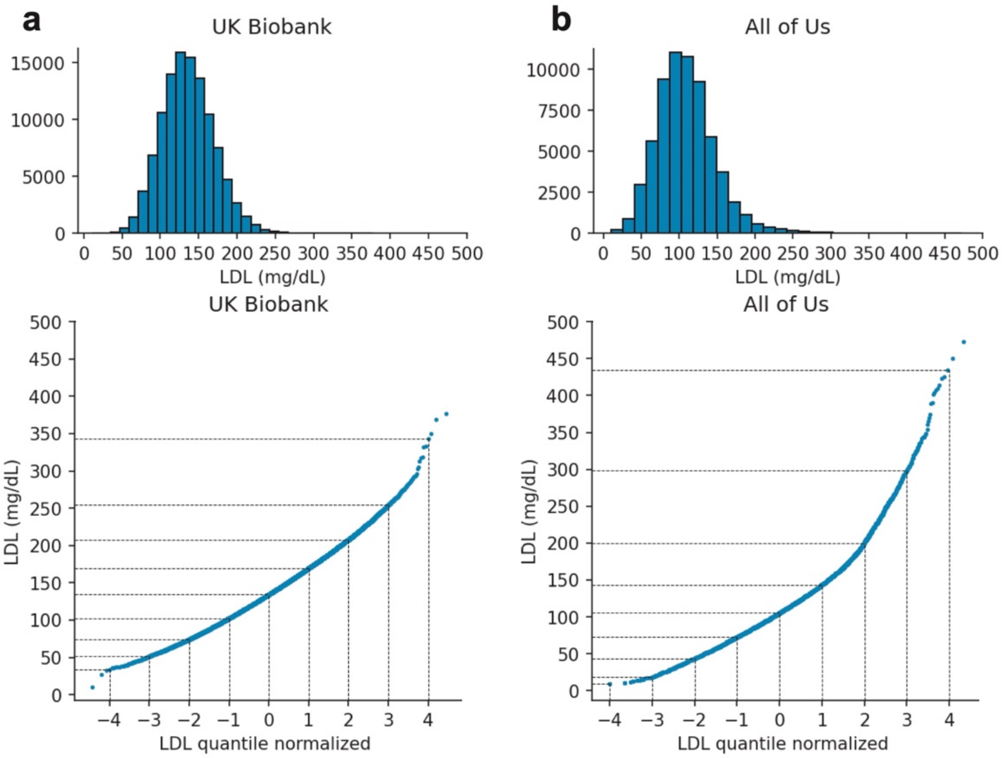
Supplementary Figure 8: Data likelihood comparison of different modeling strategies in ten quantitative traits in All of Us. (a) We evaluated four modeling strategies with regard to their log-likelihood increase (ΔlogL) over baseline model (using PGS, age, sex, age*sex, age^2, and top 10 PCs, BMI, smoking status, drinking, employment, income, current address years, education and constant prediction variance across individuals): (1) “PGSxC”: modeling PGSxC interaction in addition to baseline model; (2) “VbyC”: modeling context-specific variance which does not include PGS; (3) “PGSxC+VbyC”: modeling both PGSxC interaction and VbyC; (4) “PGSxC+VbyC(w/PGS)”: modeling PGSxC interaction and VbyC where VbyC also contains PGS. We found limited improvement of model fitting when including PGS in VbyC and therefore did not include PGS in the main analysis because model interpretation is more straightforward when prediction variance is solely a function of contexts. However, it is technically valid to include PGS in VbyC as genetic contexts may modify prediction precisions. **(b)** The summation of ΔlogL(PGSxC) and ΔlogL(VbyC) is approximately equal to ΔlogL(PGSxC+VbyC), indicating that PGSxC and VbyC components had independent contribution to data modeling. **(c)** Prediction R^2 between prediction and phenotype for “context only”: using contexts excluding PGS; “context+PGS” using contexts and PGS; “context+PGS+PGSxC” using contexts, PGS and PGSxC interaction terms. Results for “context+PGS+PGSxC+VbyC” were not shown as including VbyC terms did not impact prediction R^2.



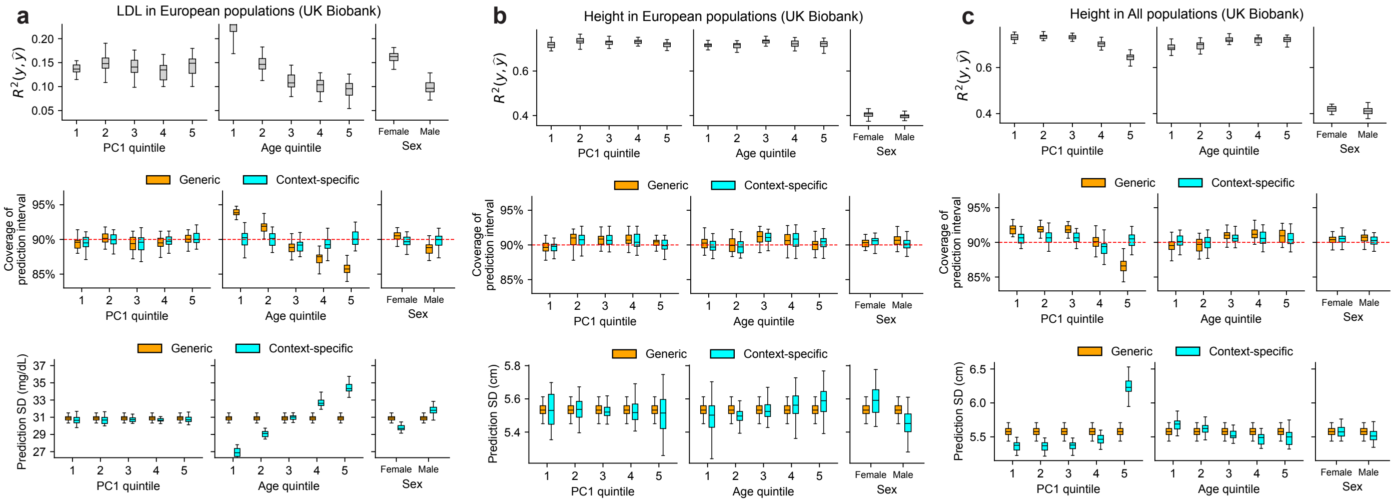
Supplementary Figure 9: Consistency of prediction mean and SD with/without PGSxC interaction terms. Evaluation was performed in ten quantitative traits across All of Us individuals. PGSxC had moderate impact to prediction mean and had no impact to prediction SD, suggesting the VbyC and PGSxC captured independent aspects of trait modeling.



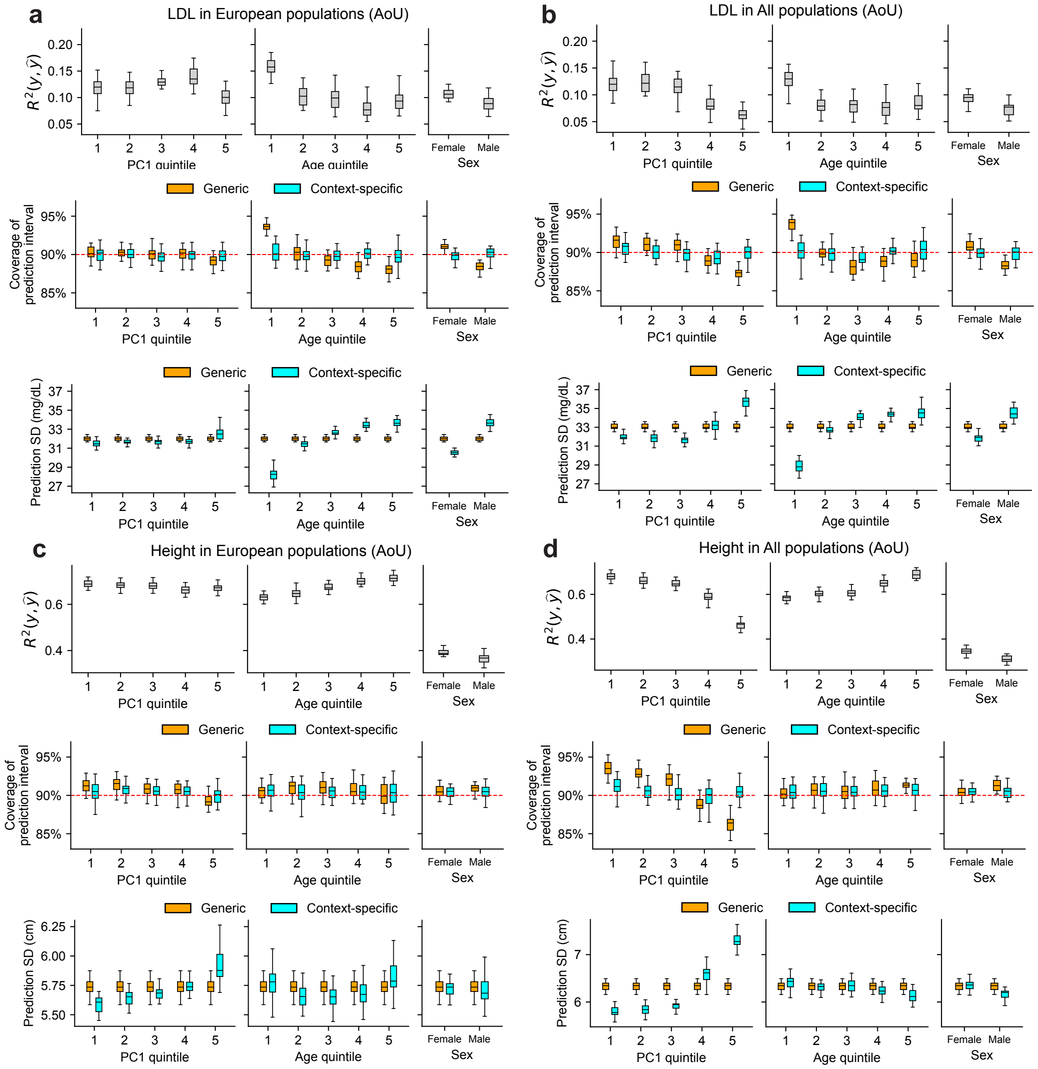
Supplementary Figure 10: consistency of prediction mean and SD with/without PGS in modeling variance by contexts (VbyC). Evaluation was performed across ten quantitative traits in all individuals from All of Us. We determined including PGS in VbyC had no impact to prediction mean and had minor impact to the prediction SD.



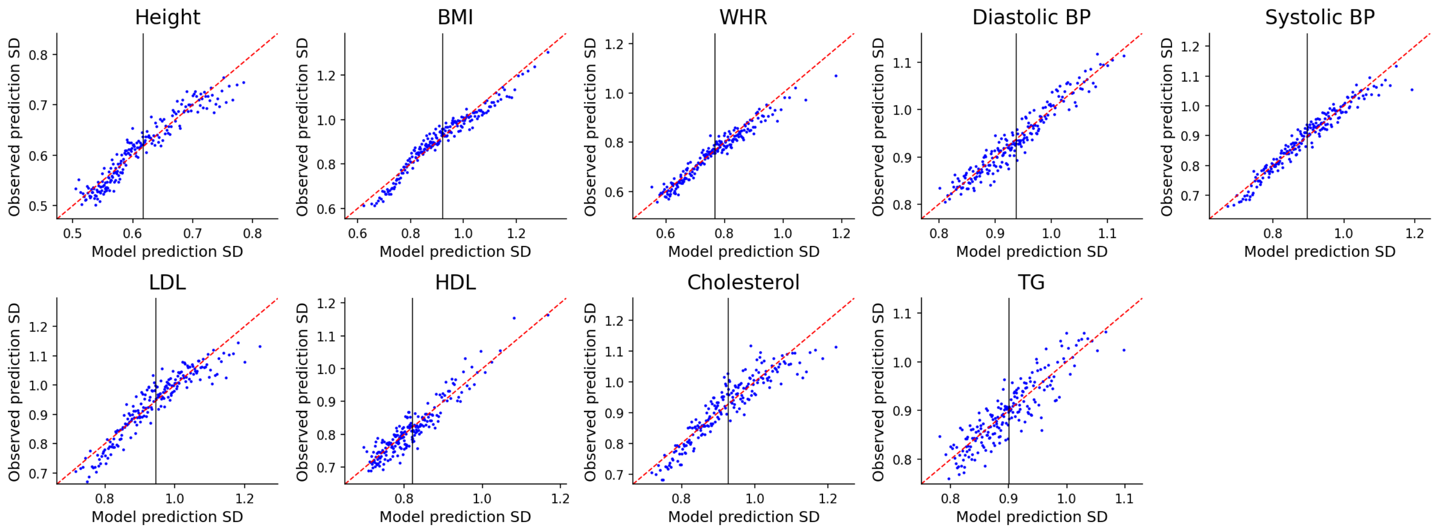
Supplementary Figure 11: correspondence between quantile normalized and raw LDL levels for UK Biobank (a) and All of Us (b). We plot distribution of LDL in raw unit of mg/dL in upper panels, and plot the correspondence between rank-based inverse normal transformed LDL and original LDL measurement in unit of mg/dL in lower panels. We note that distribution of LDL is right-skewed in both UK Biobank and All of Us. Therefore, the same unit increase in normalized scale at different LDL levels correspond to different amount of change in the original LDL measurement.



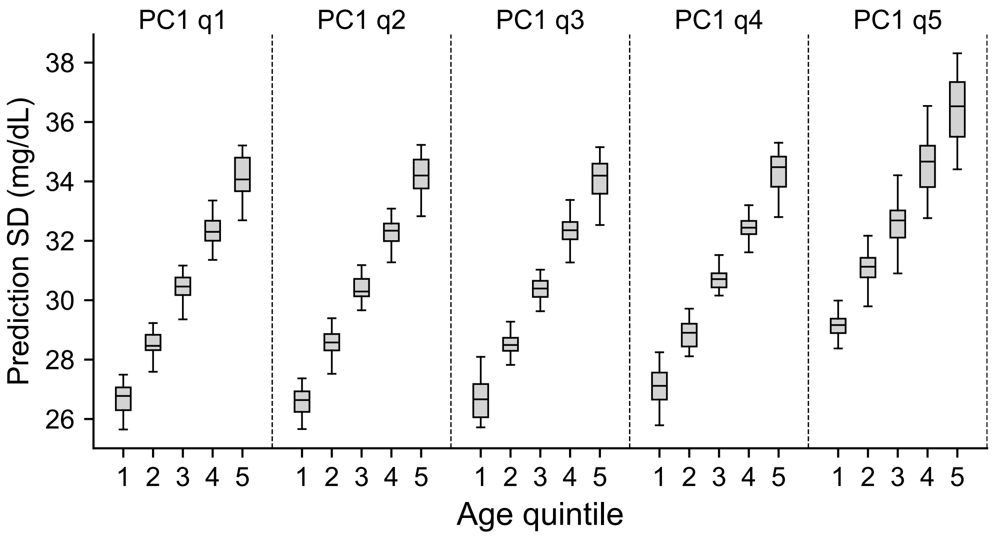
Supplementary Figure 12: Results for other populations and traits in UK Biobank. We plot results for LDL and height in "white British", and height in all individuals. **(top panel)** prediction R^2 between phenotype and point predictions. **(middle panel)** coverage of generic vs. context-specific 90% prediction intervals. **(bottom panel)** average length of generic vs. context-specific 90% prediction intervals in each context. See Fig. 6 caption for more details. For box plots, the center corresponds to the median; the box represents the first and third quartiles of the points; the whiskers represent the minimum and maximum points located within 1.5× interquartile ranges from the first and third quartiles, respectively. Each box contains results across 30 random samples with each sample of 5000 training and 5000 target individuals (n=30 points for each box).



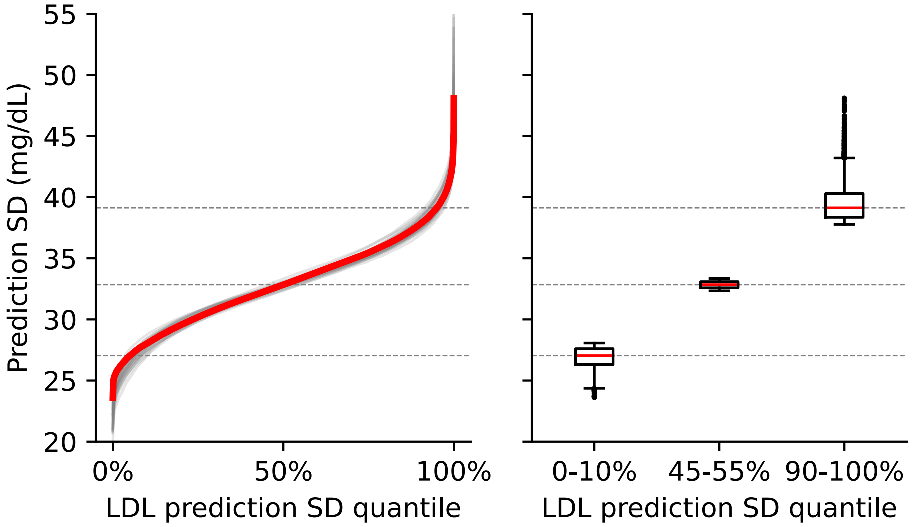
Supplementary Figure 13: Results for other traits in All of Us. We plot results for LDL and height in “white SIRE” and height in all individuals. **(top panel)** prediction R^2 between phenotype and point predictions. **(middle panel)** coverage of generic vs. context-specific 90% prediction intervals. **(bottom panel)** average length of generic vs. context-specific 90% prediction intervals in each context. See Fig. 6 caption for more details. For box plots, the center corresponds to the median; the box represents the first and third quartiles of the points; the whiskers represent the minimum and maximum points located within 1.5× interquartile ranges from the first and third quartiles, respectively. Each box contains results across 30 random samples with each sample of 5000 training and 5000 target individuals (n=30 points for each box).



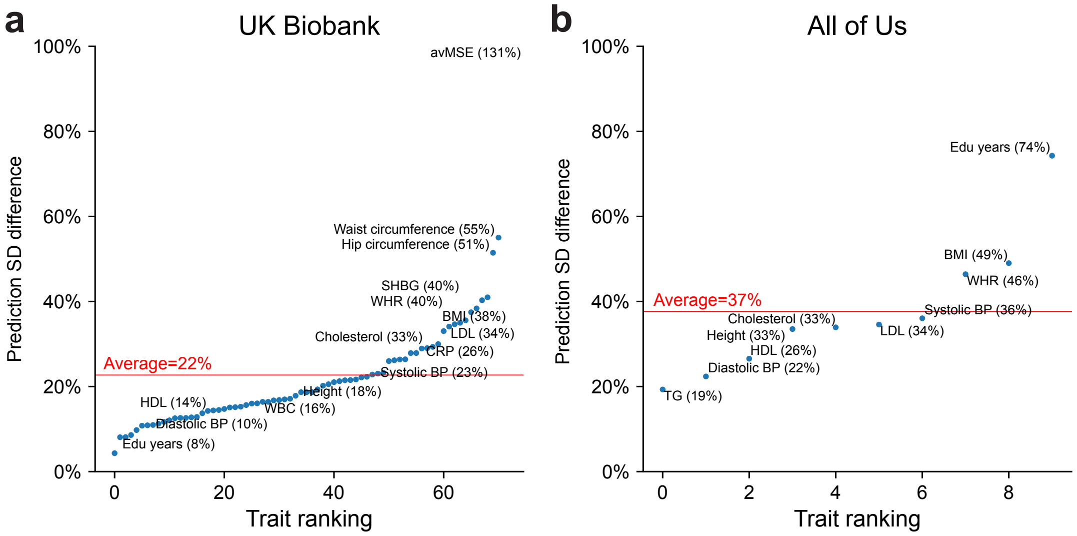
Supplementary Figure 14: Comparison of model-predicted and observed SD levels incorporating all contexts. We display model-predicted versus observed prediction SDs across traits incorporating effects of all contexts in All of Us. We observed high consistency between expected model-predicted SD and observed prediction SD (standard deviation of differences between observed and predicted phenotypes). For each trait, we also show the overall prediction SD produced by a model without modeling VbyC (vertical black lines). Here we did not show results for education years which was coded with only four discrete levels in our analysis and was not suitable to calculate prediction SD designed for continuous values.



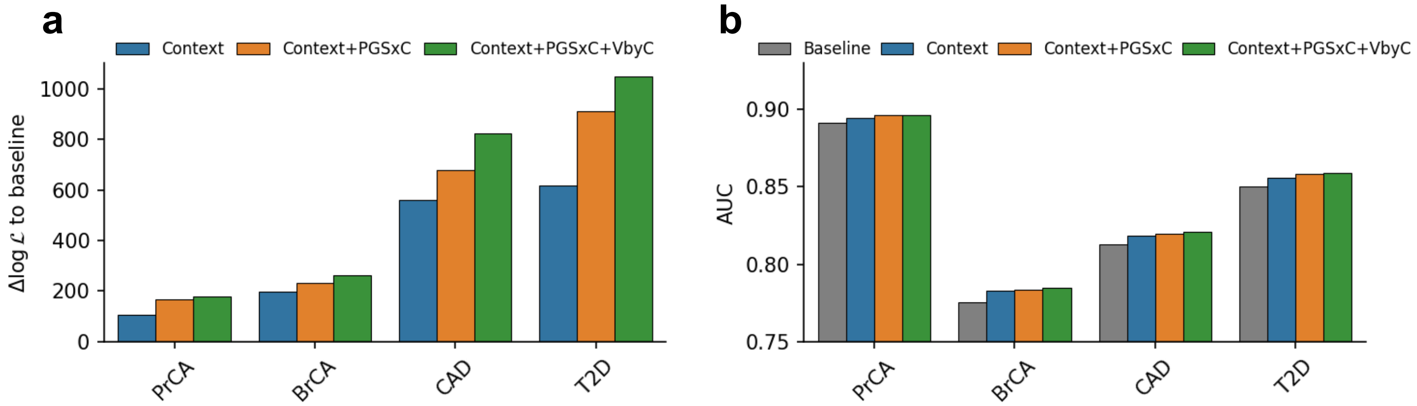
Supplementary Figure 15: Prediction interval length in predicting LDL as a function of PC1 and age. Average length of context-specific 90% prediction intervals stratified by both five quintiles of PC1 and five quintiles of age. By contrasting individuals with youngest age and smallest PC1 (leftmost) quintiles versus those of oldest age and largest PC1 quintiles (rightmost), we find larger differences of prediction interval length across these subgroups compared to results obtained when single context (either age or PC1) is considered as in Fig. 6. By considering the contribution of both age and PC1 (two largest contributors to context-specific accuracy), we detected larger differences for individuals with youngest age and smallest PC1 quintiles (more similar to European) versus those of oldest age and largest PC1 quintiles (less similar to European) (27.3 vs. 36.3 mg/dL, 33% difference). Each box plot contains data across 30 random samples with each sample of 5,000 training individuals and 5,000 target individuals (n=30 points for each box plot), the center corresponds to the median; the box represents the first and third quartiles of the points; the whiskers represent the minimum and maximum points located within 1.5× interquartile range from the first and third quartiles, respectively.



Supplementary Figure 16: Variation of prediction SD accounting for all contexts in All of Us. (a) Ordered LDL prediction SD in unit of mg/dL. Gray lines denote prediction SD obtained with random sample of 5,000 training and applied to 5,000 testing individuals. Red line denote prediction SD obtained from all individuals. **(b)** Box plots of results in (a) from individuals of LDL prediction SD quantile of 0-10%, 45-55%, 90-100% (n=96K individuals in total); the center corresponds to the median; the box represents the first and third quartiles of the points; the whiskers represent the minimum and maximum points located within 1.5× interquartile range from the first and third quartiles, respectively.



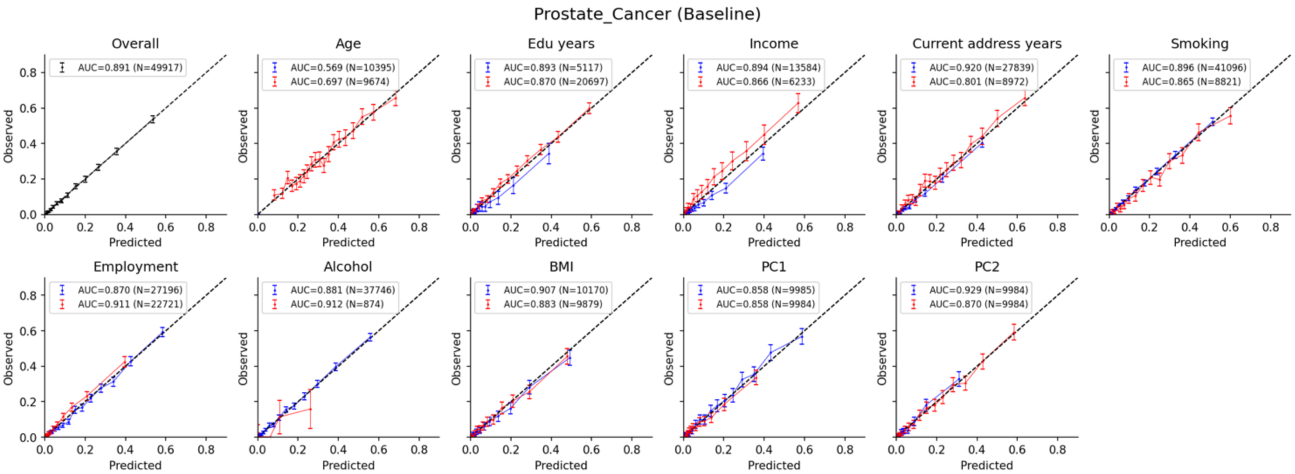
Supplementary Figure 17: Variation of prediction standard deviation (SD) accounting for all contexts across prediction SD quintiles. Relative difference of prediction SD between top and bottom prediction SD quintiles (80-100% vs. 0-20%) for all traits in UK Biobank **(a)** and All of Us **(b)**. Traits are ranked by prediction SD. See Fig. 7 for additional details.

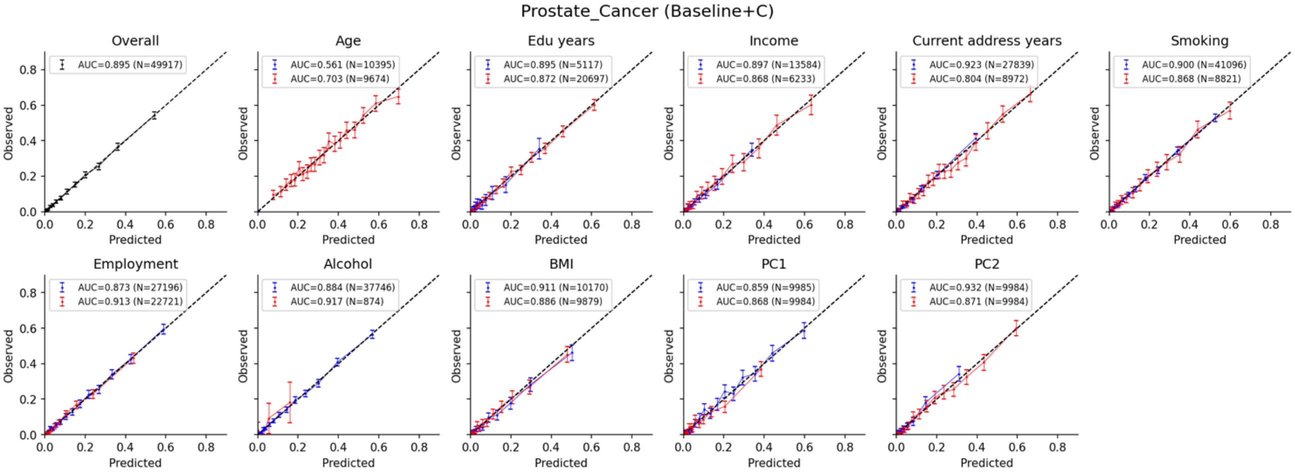


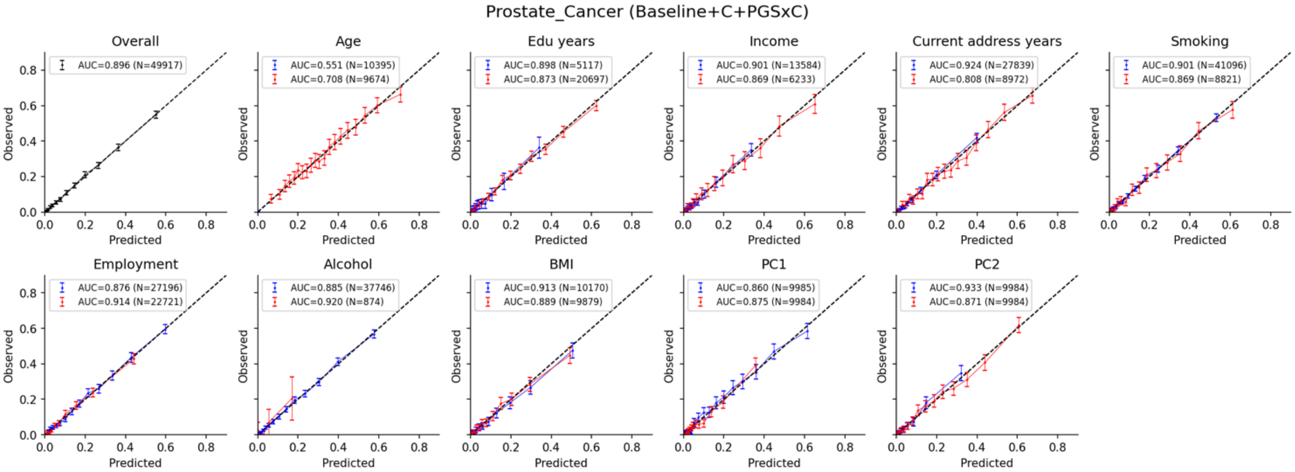
Supplementary Figure 18: Comparison of different types of models in predicting disease traits in All of Us. We evaluated three modeling strategies in their **(a)** incremental data likelihood and **(b)** area under the curve (AUC) compared to baseline models. "Baseline": a baseline model with prediction mean modeled using age, sex, age*sex, age^2, BMI, and top 10 PCs; "Context": additionally include contexts of smoking status, drinking, employment, income, current address years and education; "Context+PGSxC": additionally include PGSxC interaction terms; "Context+PGSxC+VbyC": additionally include VbyC terms.

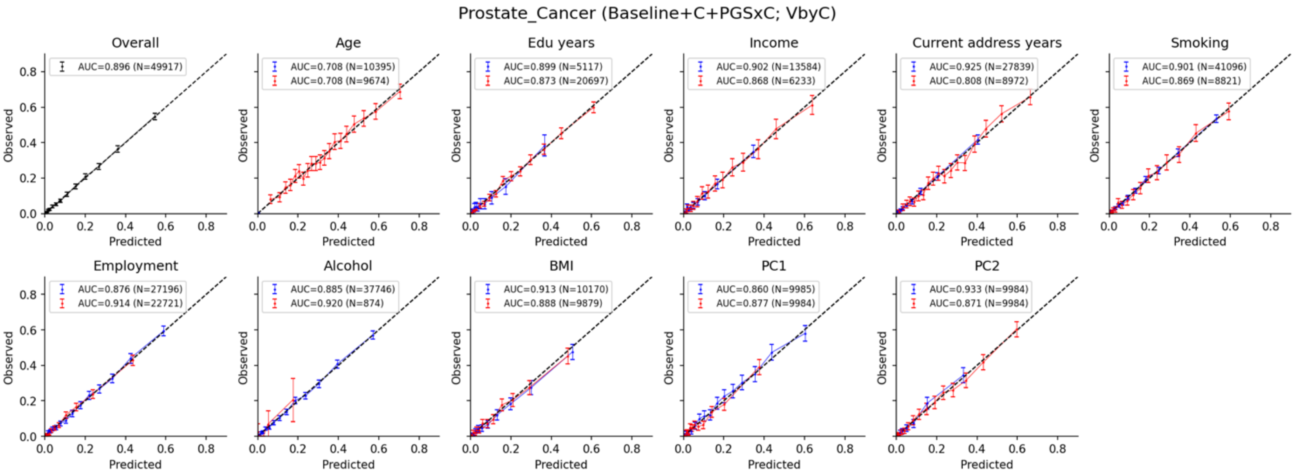
(see next four pages)

Supplementary Figure 19: Calibration of predicted probability. Calibration curve of observed proportions versus predicted probabilities for different context groups in prediction models for all individuals in All of Us. We fit four models using different sets of predictors in regression across all individuals: "Baseline": using PGS, age, sex, age*sex, age^2, BMI, and top 10 PCs; "Baseline+C": also including smoking status, drinking, employment, income, current address years and education; "Baseline+C+PGSxC": also including corresponding PGSxC interaction; "Baseline+C+PGSxC; VbyC": additionally model variance by contexts within a liability threshold model. For binary contexts, we show the calibration curve separately for two groups. For continuous context variables, we divided individuals into five quintiles and compared the calibration between top and bottom quintile groups. Error bars denote observed disease proportions and their 95% confidence intervals for each predicted probability bin (n="number of individuals shown in parenthesis" / 20 bins). See Fig. 8 caption for additional details.

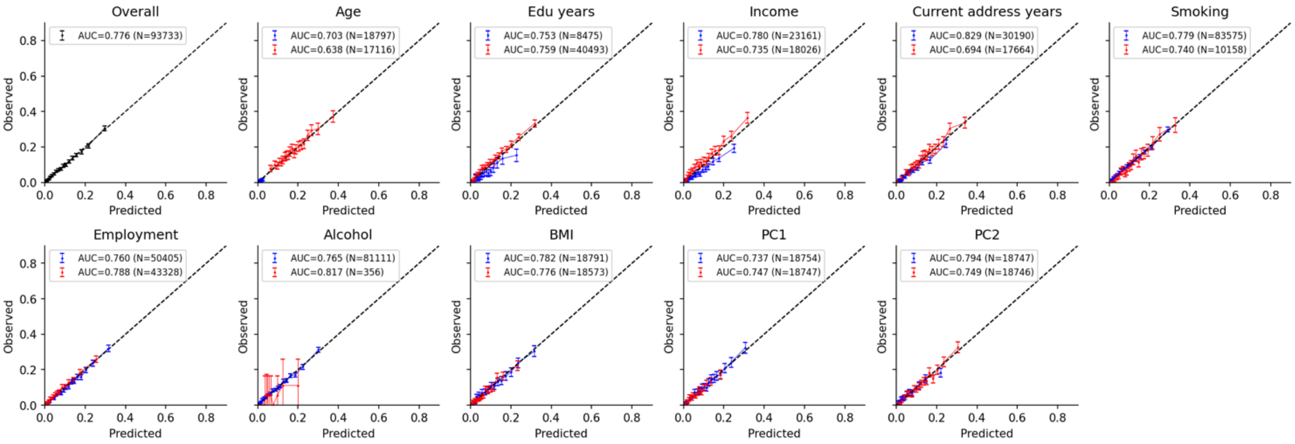

Prostate_Cancer (Baseline)
Overall
AUC=0.891 (N=49917)
Age
AUC=0.569 (N=10395)
AUC=0.697 (N=9674)
Edu years
AUC=0.893 (N=5117)
AUC=0.870 (N=20697)
Income
AUC=0.894 (N=13584)
AUC=0.866 (N=6233)
Current address years
AUC=0.920 (N=27839)
AUC=0.801 (N=8972)
Smoking
AUC=0.896 (N=41096)
AUC=0.865 (N=8821)
Employment
AUC=0.870 (N=27196)
AUC=0.911 (N=22721)
Alcohol
AUC=0.881 (N=37746)
AUC=0.912 (N=874)
BMI
AUC=0.907 (N=10170)
AUC=0.883 (N=9879)
PC1
AUC=0.858 (N=9985)
AUC=0.858 (N=9984)
PC2
AUC=0.929 (N=9984)
AUC=0.870 (N=9984)
Observed
Predicted
0.0
0.2
0.4
0.6
0.8


Prostate_Cancer (Baseline+C)
Overall
AUC=0.895 (N=49917)
Age
AUC=0.561 (N=10395)
AUC=0.703 (N=9674)
Edu years
AUC=0.895 (N=5117)
AUC=0.872 (N=20697)
Income
AUC=0.897 (N=13584)
AUC=0.868 (N=6233)
Current address years
AUC=0.923 (N=27839)
AUC=0.804 (N=8972)
Smoking
AUC=0.900 (N=41096)
AUC=0.868 (N=8821)
Employment
AUC=0.873 (N=27196)
AUC=0.913 (N=22721)
Alcohol
AUC=0.884 (N=37746)
AUC=0.917 (N=874)
BMI
AUC=0.911 (N=10170)
AUC=0.886 (N=9879)
PC1
AUC=0.859 (N=9985)
AUC=0.868 (N=9984)
PC2
AUC=0.932 (N=9984)
AUC=0.871 (N=9984)
Observed
Predicted
0.0
0.2
0.4
0.6
0.8

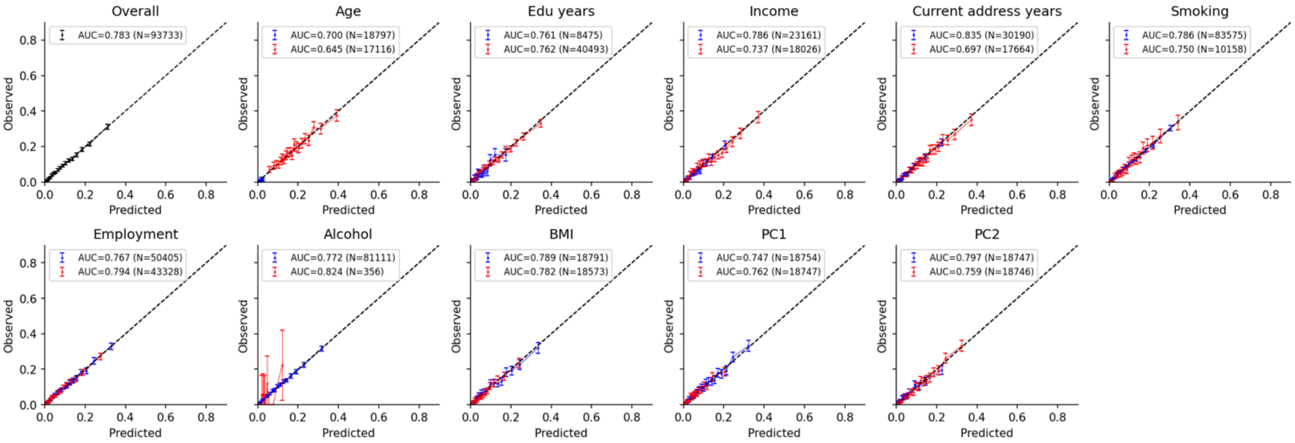

Prostate_Cancer (Baseline+C+PGSxC)
Overall
AUC=0.896 (N=49917)
Age
AUC=0.551 (N=10395)
AUC=0.708 (N=9674)
Edu years
AUC=0.898 (N=5117)
AUC=0.873 (N=20697)
Income
AUC=0.901 (N=13584)
AUC=0.869 (N=6233)
Current address years
AUC=0.924 (N=27839)
AUC=0.808 (N=8972)
Smoking
AUC=0.901 (N=41096)
AUC=0.869 (N=8821)
Employment
AUC=0.876 (N=27196)
AUC=0.914 (N=22721)
Alcohol
AUC=0.885 (N=37746)
AUC=0.920 (N=874)
BMI
AUC=0.913 (N=10170)
AUC=0.889 (N=9879)
PC1
AUC=0.860 (N=9985)
AUC=0.875 (N=9984)
PC2
AUC=0.933 (N=9984)
AUC=0.871 (N=9984)
Observed
Predicted
0.0
0.2
0.4
0.6
0.8


Prostate_Cancer (Baseline+C+PGSxC; VbyC)
Overall
AUC=0.896 (N=49917)
Age
AUC=0.708 (N=10395)
AUC=0.708 (N=9674)
Edu years
AUC=0.899 (N=5117)
AUC=0.873 (N=20697)
Income
AUC=0.902 (N=13584)
AUC=0.868 (N=6233)
Current address years
AUC=0.925 (N=27839)
AUC=0.808 (N=8972)
Smoking
AUC=0.901 (N=41096)
AUC=0.869 (N=8821)
Employment
AUC=0.876 (N=27196)
AUC=0.914 (N=22721)
Alcohol
AUC=0.885 (N=37746)
AUC=0.920 (N=874)
BMI
AUC=0.913 (N=10170)
AUC=0.888 (N=9879)
PC1
AUC=0.860 (N=9985)
AUC=0.877 (N=9984)
PC2
AUC=0.933 (N=9984)
AUC=0.871 (N=9984)
Observed
Predicted
0.0
0.2
0.4
0.6
0.8

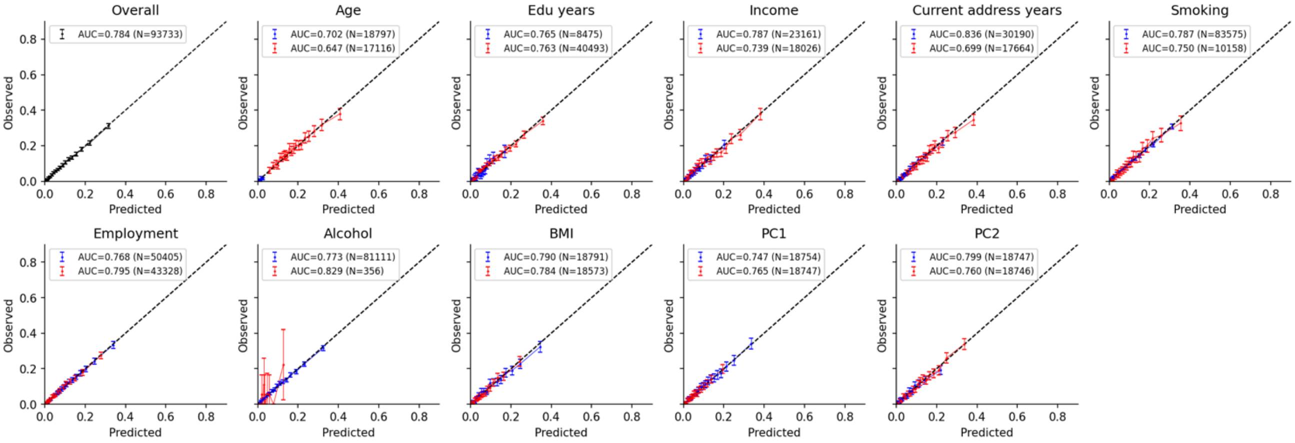
Breast_Cancer (Baseline)



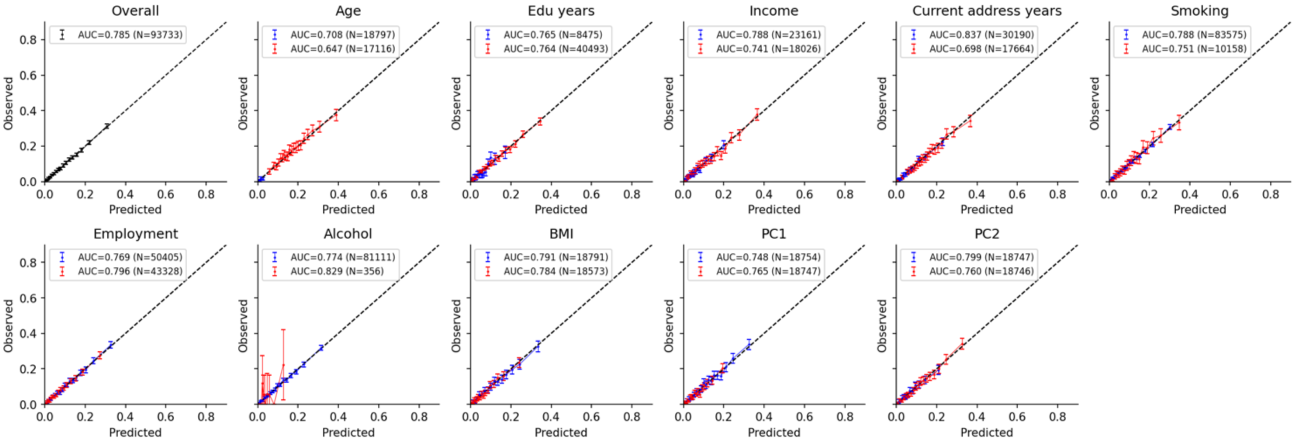
Breast_Cancer (Baseline+C)



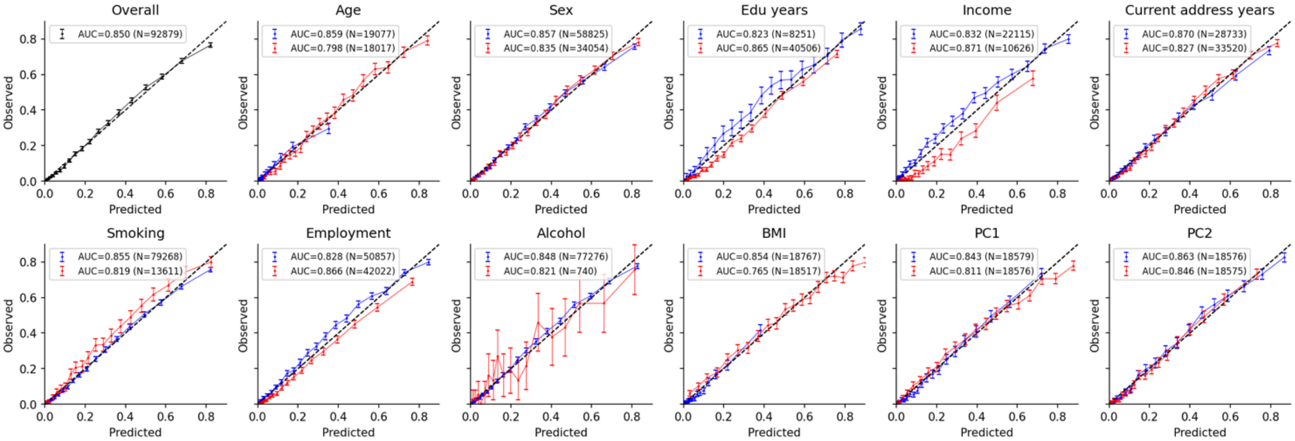
Breast_Cancer (Baseline+C+PGSxC)



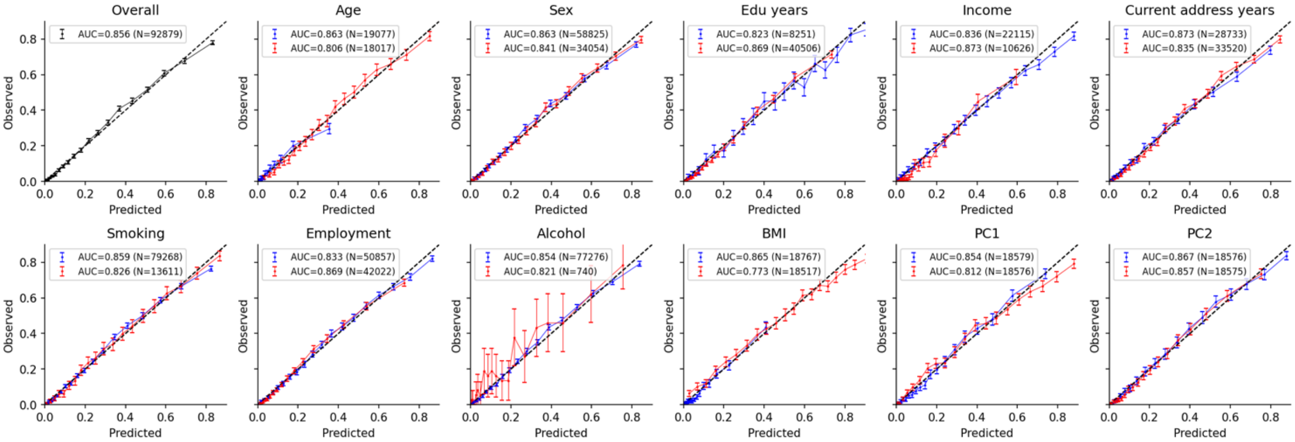
Breast_Cancer (Baseline+C+PGSxC; VbyC)



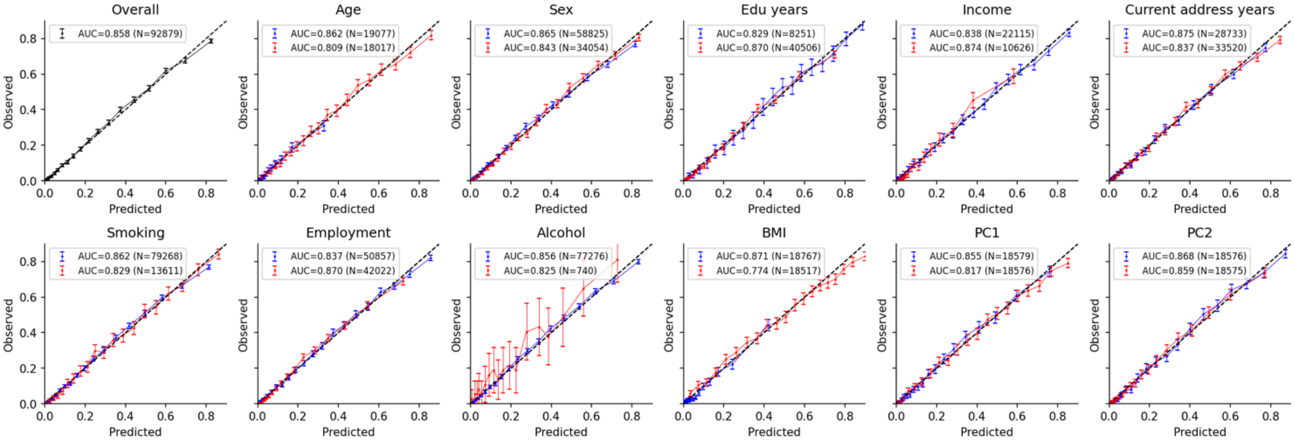
T2D (Baseline)



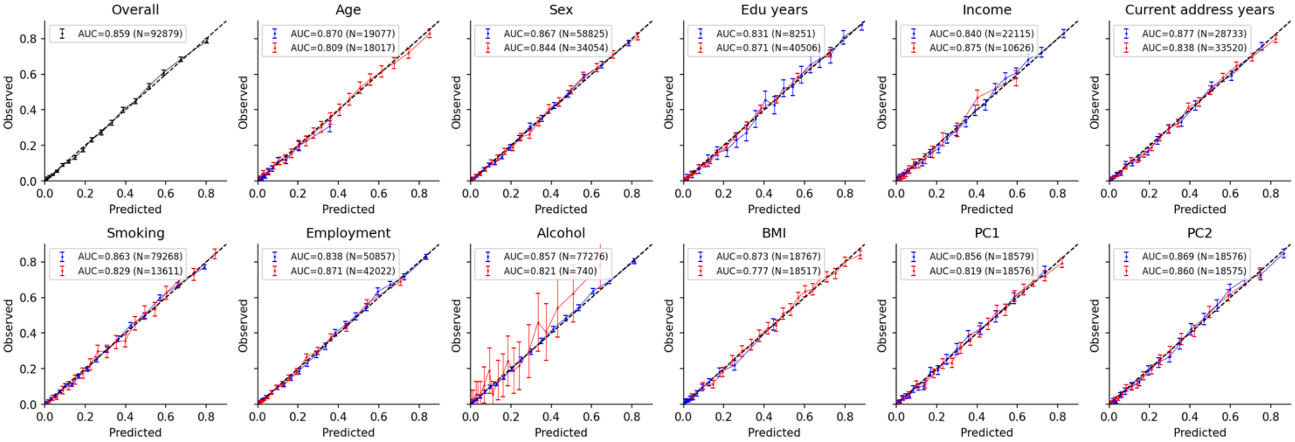
T2D (Baseline+C)

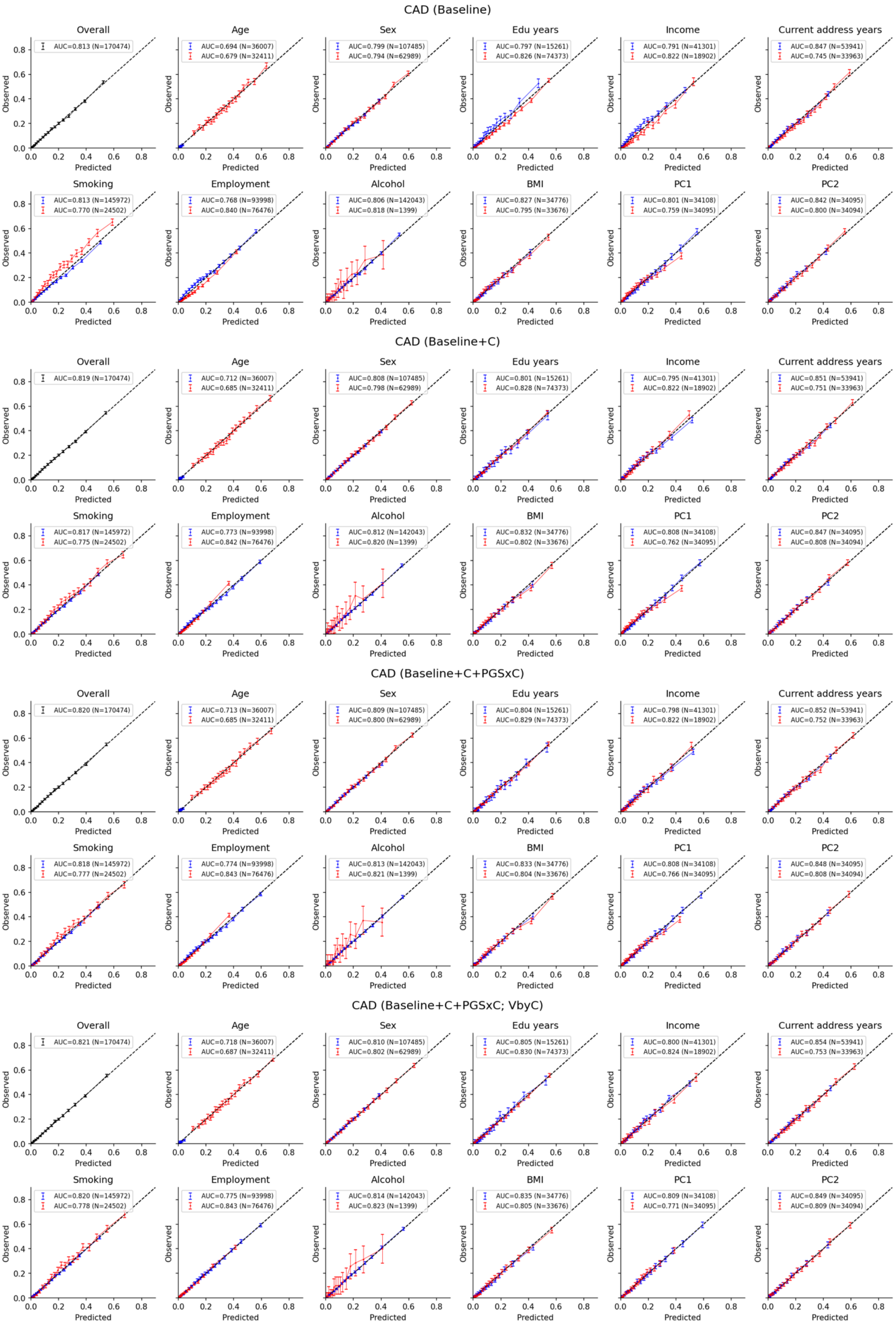


T2D (Baseline+C+PGSxC)



T2D (Baseline+C+PGSxC; VbyC)




CAD (Baseline)
Overall
AUC=0.813 (N=170474)
Age
AUC=0.694 (N=36007)
AUC=0.679 (N=32411)
Sex
AUC=0.799 (N=107485)
AUC=0.794 (N=62989)
Edu years
AUC=0.797 (N=15261)
AUC=0.826 (N=74373)
Income
AUC=0.791 (N=41301)
AUC=0.822 (N=18902)
Current address years
AUC=0.847 (N=53941)
AUC=0.745 (N=33963)
Smoking
AUC=0.813 (N=145972)
AUC=0.770 (N=24502)
Employment
AUC=0.768 (N=93998)
AUC=0.840 (N=76476)
Alcohol
AUC=0.806 (N=142043)
AUC=0.818 (N=1399)
BMI
AUC=0.827 (N=34776)
AUC=0.795 (N=33676)
PC1
AUC=0.801 (N=34108)
AUC=0.759 (N=34095)
PC2
AUC=0.842 (N=34095)
AUC=0.800 (N=34094)
CAD (Baseline+C)
Overall
AUC=0.819 (N=170474)
Age
AUC=0.712 (N=36007)
AUC=0.685 (N=32411)
Sex
AUC=0.808 (N=107485)
AUC=0.798 (N=62989)
Edu years
AUC=0.801 (N=15261)
AUC=0.828 (N=74373)
Income
AUC=0.795 (N=41301)
AUC=0.822 (N=18902)
Current address years
AUC=0.851 (N=53941)
AUC=0.751 (N=33963)
Smoking
AUC=0.817 (N=145972)
AUC=0.775 (N=24502)
Employment
AUC=0.773 (N=93998)
AUC=0.842 (N=76476)
Alcohol
AUC=0.812 (N=142043)
AUC=0.820 (N=1399)
BMI
AUC=0.832 (N=34776)
AUC=0.802 (N=33676)
PC1
AUC=0.808 (N=34108)
AUC=0.762 (N=34095)
PC2
AUC=0.847 (N=34095)
AUC=0.808 (N=34094)
CAD (Baseline+C+PGSxC)
Overall
AUC=0.820 (N=170474)
Age
AUC=0.713 (N=36007)
AUC=0.685 (N=32411)
Sex
AUC=0.809 (N=107485)
AUC=0.800 (N=62989)
Edu years
AUC=0.804 (N=15261)
AUC=0.829 (N=74373)
Income
AUC=0.798 (N=41301)
AUC=0.822 (N=18902)
Current address years
AUC=0.852 (N=53941)
AUC=0.752 (N=33963)
Smoking
AUC=0.818 (N=145972)
AUC=0.777 (N=24502)
Employment
AUC=0.774 (N=93998)
AUC=0.843 (N=76476)
Alcohol
AUC=0.813 (N=142043)
AUC=0.821 (N=1399)
BMI
AUC=0.833 (N=34776)
AUC=0.804 (N=33676)
PC1
AUC=0.808 (N=34108)
AUC=0.766 (N=34095)
PC2
AUC=0.848 (N=34095)
AUC=0.808 (N=34094)
CAD (Baseline+C+PGSxC; VbyC)
Overall
AUC=0.821 (N=170474)
Age
AUC=0.718 (N=36007)
AUC=0.687 (N=32411)
Sex
AUC=0.810 (N=107485)
AUC=0.802 (N=62989)
Edu years
AUC=0.805 (N=15261)
AUC=0.830 (N=74373)
Income
AUC=0.800 (N=41301)
AUC=0.824 (N=18902)
Current address years
AUC=0.854 (N=53941)
AUC=0.753 (N=33963)
Smoking
AUC=0.820 (N=145972)
AUC=0.778 (N=24502)
Employment
AUC=0.775 (N=93998)
AUC=0.843 (N=76476)
Alcohol
AUC=0.814 (N=142043)
AUC=0.823 (N=1399)
BMI
AUC=0.835 (N=34776)
AUC=0.805 (N=33676)
PC1
AUC=0.809 (N=34108)
AUC=0.771 (N=34095)
PC2
AUC=0.849 (N=34095)
AUC=0.809 (N=34094)
Observed
Predicted
0.0
0.2
0.4
0.6
0.8

Supplementary Tables

Supplementary Table 1: Trait information in UK Biobank. We report trait names, sample size used in training PGS weights, estimated heritability in the training sample, prediction R^2 (between PGS and covariate-adjusted phenotypes) and sample sizes used in testing populations, separately for white British and all populations. These traits are selected for their sufficient predictive power and/or biological importance.

Supplementary Table 2: Numerical results of relative ΔR^2 and estimated $\boldsymbol{\beta}_\sigma$ in UK Biobank. We report trait, context, relative ΔR^2 differences between groups, and z-score for the significance of R^2 differences between groups. We also report $\boldsymbol{\beta}_\sigma$ estimates and their standard errors.

Supplementary Table 3: Trait information in All of Us. We report trait names, prediction R^2 (between PGS and covariate-adjusted phenotypes) and sample size used in testing populations, separately for self-reported race of white and all populations.

Supplementary Table 4: Numerical results of relative ΔR^2 and estimated $\boldsymbol{\beta}_\sigma$ in All of Us. We report trait, context, relative ΔR^2 differences between groups, and z-score for the significance of R^2 differences between groups. We also report $\boldsymbol{\beta}_\sigma$ estimates and corresponding standard errors.

Supplementary Note

Comparison to other PGS calibration methods

The observation that PGS distribution differs across genetic ancestry continuum[2] motivates methods that regress out effects of variables representing genetic ancestry from PGS distribution to facilitate comparison across individuals locating at different positions in genetic ancestry continuum[3,4]. However, such approaches may unintentionally remove true biological differences of PGS distribution across genetic ancestry continuum (e.g., African Americans have reduced neutrophil count explained by the large effect of a single Duffy-null SNP[5]) as they do not consider phenotype value distribution in calibration procedure; in addition, these approaches do not represent different standard errors in PGS predictions for individuals across genetic ancestry continuum. Our method CalPred leverages a set of calibration data to adjust point predictions across contexts according to true phenotype distribution. Compared to other data-driven calibration methods[6], our approach provides a framework to incorporate context information.

Comparison between single-context and combined-context analyses

Here, we clarify differences between single-context and combined-context analyses. Overall, single-context analyses compared R^2 between covariate-adjusted phenotypes and PGS, while combined-context analyses assessed how residual phenotypic variance increases by contexts in a regression model. Technically, single-context analyses require discretization of continuous context variable into pre-specified groups to calculate R^2 differences across groups, while combined-context analyses use a regression model that naturally accommodate continuous context variable. Moreover, single-context analyses only consider one context variable at a time, while combined-context analyses consider multiple context variables jointly and can evaluate effect of each context variable conditional on other contexts.

To illustrate their differences, we consider a model where phenotype y is a function of predicted phenotype $\hat{y}$ multiplied by a slope s and environmental noise e: $y = s \cdot \hat{y} + e$, and we discuss how $R^2(y, \hat{y})$ varies as a function of both s and $\mathrm{Var}[e]$. The slope s can vary across contexts (e.g., phenotype-PGS regression slope s can vary across sex (see ref.[7]). Here prediction R^2 is

$$R^2(y, \hat{y}) = R^2(s \cdot \hat{y} + e, \hat{y}) = \frac{s^2 \cdot \mathrm{Var}[\hat{y}]}{s^2 \cdot \mathrm{Var}[\hat{y}] + \mathrm{Var}[e]}.$$

Holding $\mathrm{Var}[\hat{y}]$ as constant, $R^2(y, \hat{y})$ increases with increasing s and decreasing $\mathrm{Var}[e]$ – variable R^2 can result from variable slope s when residual variance $\mathrm{Var}[e]$ and prediction interval length are constant. Therefore, the distinction between single- and combined- context analyses is that single-context analyses models R^2 while combined-context analyses models $\mathrm{Var}[e]$.

Next, we performed case studies on two context-trait pairs (sex-BMI, sex-WHR) where we observed large differences between single-context and combined-context analyses. We focused on evaluation within white British individuals from UK Biobank to minimize confounding due to differences across ancestry groups. First, for sex-BMI, we observed almost no cross-sex difference in R^2_{PGS} between covariate-adjusted phenotype and PGS, while we observed a significant $\beta_\sigma(\mathrm{sex})$. We determined that this was because both $\mathrm{Var}[\hat{y}]$ and $\mathrm{Var}[\mathrm{e}]$ varied proportionally across sex, while the ratio between the two $R^2(y, \hat{y})$ did not vary substantially across sex.

Sex	R^2	R^2_{PGS}	SD(pred)	SD(resid)	AVG(y)	SD(y)
Female	0.159	0.137	0.423	0.992	-0.12	1.082
Male	0.146	0.136	0.343	0.809	0.139	0.875

Supplementary Note Table 1. Statistics of PGS-based predictions of BMI across sex in UK Biobank. R^2 denotes squared Pearson correlation between measured phenotype and predicted phenotype incorporating PGS, age, sex and top 10 PCs. R^2_{PGS} denotes squared Pearson correlation between covariate-adjusted phenotype (adjusted for age, sex and top 10 PCs) and PGS. SD(pred), SD(resid), SD(y) denote standard deviation of predicted phenotype, residual phenotype and phenotype, respectively. AVG(y) denotes average of phenotype.

Second, for sex-WHR, we observed lower R^2_{PGS} between covariate-adjusted phenotype and PGS in male compared to female in single-context analyses, while the sign of β_σ indicated lower prediction residual error in male compared to female, which seemed to be inconsistent with results from R^2_{PGS}. This was explained by that

β_σ captured the residual error level in trait prediction, and the overall R^2 between prediction (with contribution from including PGS and other contexts) and phenotype is higher in male compared to female. This discrepancy between R^2 and R^2_{PGS} was because BMI – the context with the highest prediction power to WHR had a much higher effect in male compared to female.

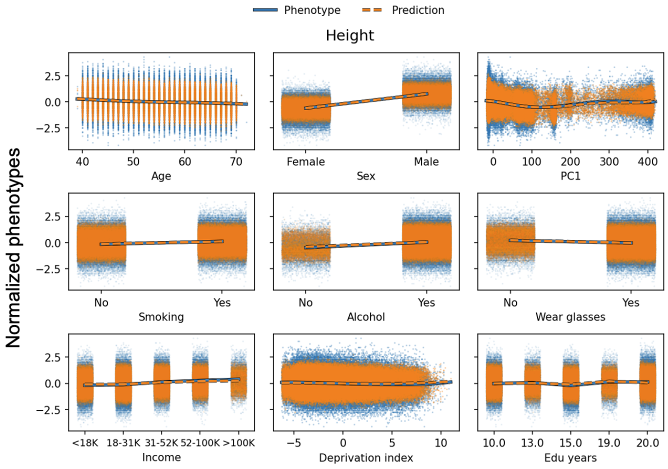
Sex	R^2	R^2_{PGS}	SD(pred)	SD(resid)	AVG(y)	SD(y)
Female	0.31	0.104	0.45	0.656	-0.603	0.79
Male	0.438	0.064	0.48	0.541	0.702	0.721

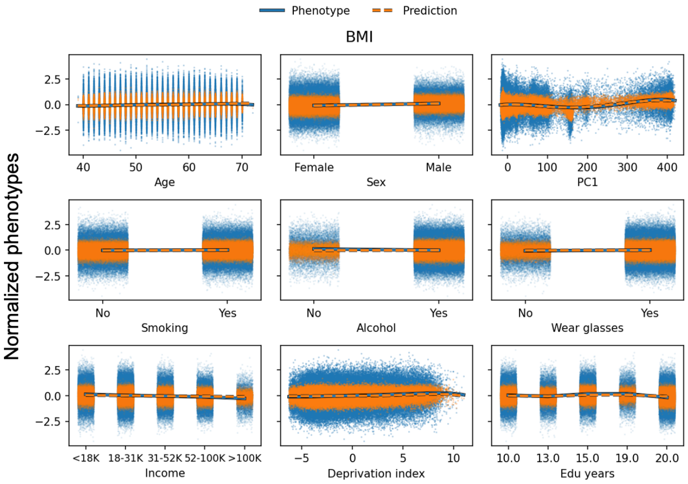
Supplementary Note Table 2. Statistics of PGS-based predictions of WHR across sex in UK Biobank. R^2 denotes squared Pearson correlation between phenotype and predicted phenotype incorporating PGS, age, sex and top 10 PCs. R^2_{PGS} denotes squared Pearson correlation between covariate-adjusted phenotype (adjusted for age, sex and top 10 PCs) and PGS. SD(pred), SD(resid), SD(y) denote standard deviation of predicted phenotype, residual phenotype and phenotype, respectively. AVG(y) denotes average of phenotype.

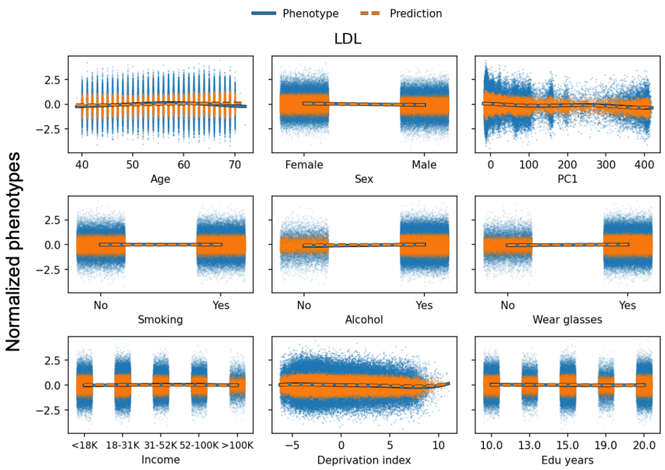
Overall, the commonly-used metric of R^2_{PGS} between covariate-adjusted phenotype and PGS and β_σ in this work are related but different quantities for PGS-based predictions. The extent of concordance of the two metrics depends on the trait. Notably, β_σ we estimated here directly relates to individual-specific residual noise level when PGS and other contexts are jointly used to predict quantitative trait values.

Investigating source of variable accuracy due to bias versus conditional variance

With $y_i = \mathcal{N}\left(\mu(\mathbf{c}_i), \sigma^2(\mathbf{c}_i)\right)$, CalPred models the variable conditional variance through $\sigma^2(\mathbf{c}_i)$. Therefore, CalPred performance relies on the unbiasedness of prediction mean $\mu(\mathbf{c}_i)$ to properly model the conditional variance term. We investigated the bias of prediction mean by comparing distribution of phenotype values versus prediction mean across contexts. We performed the comparison for three example traits of height, BMI and LDL across all individuals in UK Biobank.







Across three example traits, we observed that the prediction mean tracked well with the phenotypic mean, indicating that prediction biases were small across context groups and that CalPred model assumption was valid. Meanwhile, more fine-grained modeling of prediction factors that capture more variation, for example, by modeling interactions between prediction factors will benefit CalPred model via more precise, and shorter, prediction intervals.

References

1. Locke, A. E. *et al.* Genetic studies of body mass index yield new insights for obesity biology. *Nature* **518**, 197–206 (2015).
2. Martin, A. R. *et al.* Human demographic history impacts genetic risk prediction across diverse populations. *Am. J. Hum. Genet.* **107**, 788–789 (2020).
3. Hao, L. *et al.* Development of a clinical polygenic risk score assay and reporting workflow. *Nat. Med.* **28**, 1006–1013 (2022).
4. Khera, A. V. *et al.* Whole-genome sequencing to characterize monogenic and polygenic contributions in patients hospitalized with early-onset myocardial infarction. *Circulation* **139**, 1593–1602 (2019).
5. Reich, D. *et al.* Reduced neutrophil count in people of African descent is due to a regulatory variant in the Duffy antigen receptor for chemokines gene. *PLoS Genet.* **5**, e1000360 (2009).
6. Sun, J. *et al.* Translating polygenic risk scores for clinical use by estimating the confidence bounds of risk prediction. *Nat. Commun.* **12**, 5276 (2021).
7. Zhu, C., Ming, M. J., Cole, J. M., Kirkpatrick, M. & Harpak, A. Amplification is the primary mode of gene-by-sex interaction in complex human traits. (2022) doi:10.1101/2022.05.06.490973.