Title

Comparing ancestry calibration approaches for a trans-ancestry colorectal cancer polygenic risk score

Author List

Elisabeth A. Rosenthal[1]*, Li Hsu[2], Minta Thomas[2], Ulrike Peters[2], Christopher Kachulis[3], Karynne Patterson[4], Gail P. Jarvik[1,4]

*Correspondence: erosen@uw.edu

Affiliations:

1. Division Medical Genetics, School of Medicine, University of Washington, Seattle, WA
2. Public Health Sciences Division, Fred Hutchinson Cancer Research Center, Seattle, WA
3. Broad Institute of Harvard and MIT, Cambridge, MA
4. Department of Genome Sciences, University of Washington, Seattle, WA

Abstract

Background: Colorectal cancer (CRC) is a complex disease with monogenic, polygenic and environmental risk factors. Polygenic risk scores (PRS) are being developed to identify high polygenic risk individuals. Due to differences in genetic background, PRS distributions vary by ancestry, necessitating calibration.

Methods: We compared four calibration methods using the All of Us Research Program Whole Genome Sequence data for a CRC PRS previously developed in participants of European and East Asian ancestry. The methods contrasted results from linear models with A) the entire data set OR an ancestrally diverse

training set AND B) covariates including principal components of ancestry OR admixture. Calibration with the training set adjusted the variance in addition to the mean.

Results: All methods performed similarly within ancestry with OR(95% C.I.) per s.d. change in raw PRS: African 1.5(1.02,2.08), Admixed American 2.2(1.27,3.85), European 1.6(1.43,1.89), and Middle Eastern 1.1(0.71,1.63). Using admixture and an ancestrally diverse training set resulted in distributions closest to standard Normal with accurate upper tail frequencies, whereas the other combinations resulted in inconsistent tail frequencies.

Conclusions: Training a calibration model on ancestrally diverse participants to adjust both the mean and variance, using admixture as covariates, was best at identifying patients at high polygenic risk. Although the PRS is valid for most ancestries, its performance varies by ancestry, even after calibration. More diverse datasets are required to further develop and validate the PRS.

Introduction

Colorectal cancer (CRC) is the third most common cancer in the United States (U.S.), and is associated with the second most deaths from cancer [1]. CRC risk may be attributable to social determinants of health such as socioeconomic status, health care access, and food security, which are known to be associated with race [2]. Individuals who self-report as Black or Native American/Alaskan Native have the highest lifetime risk for CRC as well as higher mortality rates [1,3,4]. Additionally, risk for CRC is affected by other environmental factors such as diet, smoking and physical activity, as well as genetics [5]. Although 25% of CRC appear to be familial, only 20% of familial CRC is explained by variation at single genes [6–8]. In aggregate, low risk SNPs across the genome also contribute to risk of CRC. Polygenic risk scores (PRS), which aggregate the effect of genetic variants across the genome, are currently being developed to aid in identifying individuals at higher risk of developing CRC and those at risk of developing CRC at an early age [9–12].

Development of PRS involves estimation of SNP effects, which are influenced by the genetic ancestry of the cohorts used to develop the PRS through linkage disequilibrium (LD) across the genome and minor allele frequency (MAF) at the SNPs [13]. The PRS for CRC used here was developed with participants of European and East Asian ancestry, using Bayesian methods which estimates SNP effect sizes accounting for LD in both populations, while simultaneously estimating the number of causal SNPs given the heritability of CRC [10,14]. When participants from multiple ancestries are included in the development of PRS, the estimated SNP effects are a weighted average of the estimated effects in the different ancestry groups, dependent on the relative sample sizes of those groups [15,16].

As LD and MAF vary across genetic ancestries, the distribution of the PRS will vary by ancestry. Therefore, the raw scores need to be calibrated so that CRC risk estimation is valid and equitable regardless of genetic ancestry. Several mechanisms for calibrating the PRS for ancestry in the clinic

setting have been proposed [17,18]. These include using linear models, based on a training data set of diverse populations, to adjust the mean PRS using principal components (PCs) of ancestry. Additionally, methods using training models have been adapted to adjust for both the expected mean and variance of the PRS. One benefit of these approaches is that patients do not need to be binned within an ancestry. Furthermore, these approaches can be applied to admixed individuals, who make up an increasing proportion of U.S. residents.

We set out to determine which, if any, of these calibration methods is preferable in the context of this CRC PRS, by analyzing data from the All of Us Research Program (AOU) [19] (and AOU genome paper under review). The purpose of AOU is to collect survey, electronic health record (EHR), and genotype data on diverse participants who live in the U.S., for use in broad research and to help reduce inequities in healthcare in the U.S.. As this data set contains both ancestrally diverse and admixed participants, it is an ideal biobank to assess the ancestry calibration methods for use in the clinic. Since it is well known that PRS have less accurate risk estimates for individuals from ancestries different from the development cohorts [20–22], we also examined whether any of these calibration methods could improve risk estimates for participants with ancestry different from the development cohorts.

Subjects and Methods

Case/Control assignment

We updated a previously developed CRC case/control algorithm (see online materials) to define CRC cases and controls. This algorithm was created using International Classification of Disease (ICD) 9 codes and Current Procedural Terminology (CPT) codes. The AOU dataset contains a rich database of ICD 9 and 10 codes, CPT codes, and Logical Observation Identifiers Names and Codes (LOINC). As the completeness of this data varies among participants [23] we attempted to widen our net by creating concept sets based on the original ICD9 and CPT codes, incorporating the related ICD 10 codes and LOINC codes. Each

concept set relates to a different table from the word files provided by the algorithm's authors (Supplemental Tables S2-S10). The algorithm was developed to screen for CRC cases with potentially monogenic causes of CRC, and therefore excluded individuals with ulcerative colitis (UC) or Crohn's disease, two disorders associated with CRC. As we are interested in a polygenic component underlying CRC we did not exclude participants with UC or Crohn's disease. Additionally, we excluded all participants with a known monogenic (pathogenic or likely pathogenic variant in ClinVar) cause of CRC in any of *AKT1*, *APC*, *AXIN2*, *BMPR1A*, *CDH1*, *CHEK2*, *CTNNA1*, *EPCAM*, *GALNT12*, *GREM1*, *MLH1*, *MSH2*, *MSH3*, *MSH6*, *MUTYH*, *NTHL1*, *PDGFRA*, *PIK3CA*, *PMS2*, *POLD1*, *POLE*, *PTEN*, *RPS20*, *SMAD4*, *STK11*, or *TP53* (AOU genome paper under review).

Potential cases included all participants that had at least one CRC diagnosis in their medical record (Supplemental table S2) but did not have a known pathogenic variant for CRC or a diagnosis for monogenic disorders with increased risk of CRC (Supplemental table S3). The cleaned set of cases was derived from potential cases using the following ordered algorithm:

A. They had a surgical procedure related to CRC within a year of diagnosis (Supplemental Table S4)
B. If not (A), then they had chemotherapy or radiation (Supplemental Tables S5 and S6) within a year of CRC diagnosis <u>and</u> they did not have other types of cancers listed in the exclusion table (Supplemental Table S7)
C. If not (A) or (B) and they had at least two CRC diagnosis codes within 2 years of each other and no other cancers listed in the exclusion table

Potential controls were participants with no diagnosis codes for CRC or a monogenic disorder with increased risk of CRC, and no evidence of a pathogenic variant for CRC. Using guidance from the algorithm, the screened controls met either of the following criteria:

A. They had at least one sigmoidoscopy or colonoscopy and no positive pathology report (Supplemental Table S8)

B. They had at least two instances of a fecal immunochemical test (FIT) or a fecal occult blood test (FOBT) that were at least 5 years apart and never had a positive lab result from any of these tests (Supplemental Tables S9 and S10)

All other participants that were not assigned case/control status, or who were not excluded by the algorithm, were left unassigned. Unassigned participants are included in the analysis except when statistical testing involves case and control status. Some analyses include age and sex. Age is defined as observational age: age at onset of cases, age at last screening for controls, and age at consent to AOU for the unassigned participants. In these analyses we further restricted analysis to participants who were older than the minimum observed age of onset of CRC, 19, as CRC rarely presents so early. Sex is defined as sex assigned at birth.

Genetic Ancestry Clustering

We calculated 32 genome-wide PCs using an unrelated subset of reference genomes from 1000 Genomes (1KG) and Human Genome Diversity Project (HGDP) [24,25] (N = 4151). We then projected the AOU participants onto the PC space. We used EIGMIX to estimate admixture in the AOU participants for the continental ancestry groups represented by the reference genomes: African (AFR), Admixed American/Latino (AMR), East Asian (EAS), European (EUR), Middle Eastern (MID) and South Asian (SAS) [26]. See supplemental methods for details on PC and admixture calculation. We then clustered the AOU participants into ancestry groups, using the estimated percent ancestries, in order to compare the performance of the PRS across continental ancestry groups. As the amount of genetic diversity among different racial and ethnic groups in the United States differs (Supplemental Figures S1 and S2) [26–28], we

used an 80% cutoff to cluster individuals into EAS, EUR and SAS clusters, and a 60% cutoff to cluster individuals into AFR, AMR and MID clusters. All other participants are referred to as Other (OTH).

PRS calculation

Details of PRS calculation in AOU and the reference 1KG and HGDP data are given in the supplemental methods. In brief, we extracted genotype data from WGS using HAIL[29] on the Spark cluster. After performing quality control, we calculated the sum of genotype effects for each participant using the R package bigSNPR v. 1.10.8 . The allele effects for each SNP were obtained from the corresponding author of PMID: 36789420 [10] (under review).

PRS Calibration

We started with three calibration methods of the raw PRS (RAW) to account for ancestry: linear model to adjust for the first 5 PCs of ancestry in AOU (PC_μ), linear model to adjust for the estimated admixture percentages in AOU (AD_μ), and a previously published method which adjusts both mean and standard deviation (s.d.) as a function of PCs using a trained model on a reference dataset (PC.REF_μσ), [17,18]. This adjustment estimates the effects of the first 5 PCs of ancestry in the 1KG and HGDP reference genomes on the mean PRS and its variance. Given these estimated effects, the raw PRS from AOU can be adjusted by both the expected PRS and its expected variance for each participant given their first 5 PCs of ancestry. We chose to use 5 PCs of ancestry to make comparisons fair with AD_μ which has 5 degrees of freedom due to the six global ancestries considered in this analysis. We added in two more adjustments after observing the distribution of PC.REF_μσ. We modified the method by using a random subset (N=4,151) of the unassigned participants from AOU, called AOU.REF, to derive the prediction model (PC.AOU_μσ). Additionally, we modified the method, using the same random sample, AOU.REF, and used the admixture estimates to create the prediction model rather than the first 5 PCs of ancestry (AD.AOU_μσ).

We kept the sample size for AOU.REF the same as that in 1KG and HGDP so that we could make fair comparisons. As the AOU dataset is heavily weighted toward individuals of EUR ancestry, a simple random sample would also be heavily weighted in this way. Therefore, we attempted to randomly sample from the space of the first 5 PCs. First, we calculated the geometric median of the first five PCs. Second, we calculated the Euclidean distance to this median for each participant. We divided the distances into five equal ranges. We then randomly sampled an equal number of participants from each range, resulting in the following counts of unassigned participants from each ancestry cluster: 1108 AFR; 1459 AMR; 84 EAS; 820 EUR; 72 MID; 577 OTH; 31 SAS.

Comparing raw and calibrated PRSs

We compared the raw PRS and the adjusted PRSs in several ways, for each ancestry. First, we compared the odds ratio (OR) for a single s.d. change in the score, with and without adjusting for age and sex. Second, we compared the OR for each quantile compared to the middle quantile. Third, we compared the area under the receiver operating curve (AUC) which compares how well the scores can distinguish between cases and controls, overall. We used 500 bootstrap iterations, with replacement, to estimate the 95% confidence intervals (C.I.) for the AUC. We calculated both the unadjusted AUC and the AUC adjusted for age and sex. Finally, we compared the observed and expected upper 5, 7.5 and 10th percentiles, as these values could be used to select patients at higher risk for developing disease and therefore receiving increased/earlier screening. Observed percentiles were defined as the proportion of participants whose adjusted PRS was greater than the standard normal cutoffs (z=1.28, 1.44, or 1.64). In this case, RAW, PC_μ, and AD_μ were standardized using the variance from the entire data set. The C.I. for the percentiles was calculated using the R function binom.test().

Results

Case/Control Demographics and Ancestry

We excluded 8 CRC affected participants who had penetrant Lynch syndrome or Familial Adenomatous Polyposis reported in their EHR and 455 participants with a known pathogenic or likely pathogenic variant for these monogenic disorders (AOU genome paper, under review). There were 668 participants remaining with at least one CRC code in the EHR, 348 of whom were classified as cases by the algorithm (Table 1, Supplemental Tables S2 and S3). There were 97,588 participants who had no evidence for CRC. Of these 12,378 had been screened for CRC with negative results and were assigned as controls. Using the ancestry clustering rule, the majority of participants clustered with EUR ancestry (45%), followed by AFR (23%), OTH (16%), AMR (9.5%), and MID (4.1)% with EAS and SAS making up < 3% of the sample (Table 1). As the sample size for EAS and SAS was small, we do not report statistical tests for these ancestries.

The mean observed age of unassigned participants was about 10 years younger than that of the cases, overall (Table 2). The mean censored age of screened controls was similar to that of the cases, consistent with recommendations for screening in older patients. The observed age for both the cases and screened controls tended to be younger than their age at consent, indicating that most cases had CRC before entering the study and most screened controls were screened before entering the study.

Overall, males made up 49% of the cases and 40% of the controls. This sex difference may reflect sex differences in seeking healthcare and/or enrolling in genetic research studies. For AMR, EUR and OTH, the sex ratio difference was ≤ 10 percentage points. The sex ratio differences were > 20 percentage points for AFR and MID. We do not report exact counts in order to comply with AOU publication rules for cell counts less than 20.

PRS Distribution Comparison

The distributions of RAW differed by ancestry in both their mean and s.d. (Table 3, Figure 1). The mean ranged between 0.05 (MID) and 0.4 (AFR) and the standard deviation (s.d.) ranged between 0.41 (AFR) and 0.47 (most ancestries). The means were close to zero for adjustment PC_μ for all ancestries except MID and OTH. AD_μ resulted in means closer to zero, even for MID and OTH. It should be noted that when RAW, PC_μ and AD_μ are scaled to have a s.d. of 1, the non-zero means are scaled similarly so that the resulting means are further from zero (since the raw s.d. are less than one). PC.REF_μσ resulted in non-zero means for EAS, EUR, MID and SAS, prompting us to develop PC.AOU_μσ and AD.AOU_μσ. PC.AOU_μσ performed similarly to PC_μ and the scaling of the variance results in means further from zero, as expected. AD.AOU_μσ performed similarly to AD_μ, except that the mean for SAS was non-zero. This is likely due to the small number of SAS participants in AOU.REF. Given that we are unable to perform statistical tests in the SAS cluster due to low sample size, we chose to continue with these calibrations and make comparisons between RAW, PC_μ, AD_μ, PC.AOU_μσ, and AD.AOU_μσ.

PRS OR

Results for the OR per 1 s.d. change in the PRS were similar across calibration methods, within ancestries (Table 4). The OR estimate was highest for AMR (range 2.1-2.2), was statistically significant, and had wide confidence intervals. For EUR, the estimated OR ranged between 1.6 and 1.7, was statistically significant, and had narrow confidence intervals reflecting the large sample size. For OTH, the estimated OR ranged between 1.5 and 1.6 and was statistically significant. The estimated OR for AFR ranged between 1.4 and 1.5, and was not significant after adjusting for age and sex. As this lack of significance could be due to a sex imbalance among the cases and controls, we performed sex stratified analysis in AFR. Sex stratified analysis in AFR resulted in similar estimated ORs and they were not statistically significant (data not shown). The estimated OR for MID ranged between 1.05 and 1.12, and was not

statistically significant in any model. Although the ORs indicate that an increase in PRS is associated with increased risk of CRC, statistical significance for the uppermost quintile compared to the middle is observed only in EUR, due to sample size (Figures 2-6). For AMR, there were no cases in the lowest quantile.

PRS AUC

The estimated AUC and its 95% C.I. are given in Table 5. Overall, these results are similar to those for the OR, as expected[30]. The estimated AUC is greater than 0.5 for all ancestries and all adjustments. However, the AUC is not statistically significantly different from 0.5 for both AFR for some calibrations and when adjusting for age and sex, as well as for MID in all situations. The AUC is highest for AMR (0.68-0.72), with wide C.I.s. The AUC is similar for EUR and OTH (~0.64). EUR has the narrowest C.I. reflecting the larger sample size. Adjusting the AUC for age and sex does not change the results for AMR, EUR or OTH.

Observed Upper Percentiles

We compared the observed upper 5, 7.5 and 10th percentiles to their expected values for each PRS. (Figure 7). AD.AOU_μσ results in observed percentages close to expected, with confidence intervals overlapping the expected values for AMR, EUR and MID. For AFR and OTH, the observed values are slightly lower than expected for AD.AOU_μσ. There are no consistent patterns when comparing the other calibrations across ancestries. For example, all calibrations, except for AD.AOU_μσ result in larger percentages than expected for EUR, whereas they result in lower percentages than expected for AFR and AMR. The C.I.s for EAS and SAS are large, reflecting their smaller sample sizes.

Discussion

The PRS was statistically significantly associated with CRC status for EUR, AMR, and OTH, when evaluating the OR for one standard deviation change in PRS and the AUC, regardless of the ancestry calibration method. However, the strength of the effect in these groups differed. The estimated OR and AUC were highest for AMR, possibly due to the lack of cases in the lowest quantile of the PRS. The confidence interval for AMR was large, as the total sample size was small. The OR estimate was higher in EUR than in OTH, but they had similar AUC. Power was sufficient in only the large EUR cluster to show that participants with PRS in the upper quintile were at higher risk of CRC than those in the middle quantile. As we used whole genome sequence data, differences in OR across ancestries cannot be attributed to uncertainty in imputation of genotypes, but rather solely to differences in LD with causal variation. However, the availability of EHR data, and therefore the number of cases and controls, may be associated with self-identified race or ancestry, which further complicates interpretation of the results [23].

We observed marginal statistical significance for the OR per s.d. in PRS for AFR, but only when not adjusting for age and sex. The sex imbalance between cases and controls for AFR may contribute to this lack of significance, but sex stratified analysis did not have enough power to confirm this, due to low sample size. We also did not observe statistically significant AUC for AFR when adjusting for age and sex, as expected [30]. We did not observe statistically significant OR or AUC for MID. Most likely, the lack of participants with genetic similarity to AFR and MID reference populations in the development of the PRS contributes to this lack of significance and calibration for ancestry does not ameliorate this issue. The sample sizes were too small to perform testing with EAS and SAS.

It has been shown that the applicability of PRS declines with further genetic ancestry distance from the population used to develop a PRS [20]. However, this phenomenon has yet to be analyzed for PRS that are

developed in more than one population, as in this study. An open question may be whether the PRS analyzed here behaves similarly for individuals who lie along the EAS-EUR axis in PC1xPC2 space, such as the AMR cluster, as opposed to those that cluster within EUR or EAS.

Although each calibrated PRS performed similarly within each ancestry cluster, their mean and standard deviations were not consistent within ancestries. Furthermore, the observed upper percentiles were not consistent. Only AD.AOU_μσ consistently resulted in a mean of zero and a standard deviation of one, as well as observed percentiles close to expected, for the most ancestries. This may be due to the fact that, although the model contains only 5 degrees of freedom, the admixture estimates used in the model are based on 32 PCs of ancestry, and therefore provide more information than the first 5 PCs used in PC.AOU_μσ. In fact, when we used the PC.AOU_μσ method using 32 PCs of ancestry, the resulting distributions were similar to that of AD.AOU_μσ (data not shown). It should be noted that the mean AD.AOU_μσ adjusted PRS was not near zero for the SAS cluster. This may be due to too few participants in AOU with genetic similarity to 1KG and HGPD SAS reference populations at this time.

We believe that further developing the AD.AOU_μσ method would be useful. It can be applied to any patient and does not require the clinician to determine the patient's ancestry. It also uses reasonably sized reference data (N<10,000). However, it requires the use of two reference datasets. The first reference dataset, representing defined ancestral groups (e.g., 1KG and HGDP) is required to estimate PCs of ancestry and to estimate admixture for all participants. The second reference dataset represents the expected ancestral space of patients in the clinic. The selection of both reference datasets affect the validity of the calibration. Both the PCs and admixture estimates depend on which ancestries are included in the first reference dataset. For example, In this study we were not able to evaluate the PRS in Native American/Alaskan Native participants as the reference dataset did not contain samples from these populations. Methods to sample the second reference dataset from the PC ancestry space may need further development to ensure adequate representation for all potential patients. Furthermore,

both the patient and the second reference dataset must have genotypes for the same ancestry informative markers as those used in the first reference data. The second reference dataset also needs to contain genotypes at the same markers that are used to derive the PRS. As the PRS evolves, this may become more complicated. However, using WGS data, as is available in AOU may alleviate this issue. Further, estimating admixture is time consuming. One could potentially avoid estimating admixture and use the PC.AOU_μσ method with all PCs in the adjustment model. The validity of this approach needs further investigation.

In order to bring PRS to the clinic, we need to determine what percentile cutoff results in clinically relevant risk and if the same cutoff can be used regardless of ancestry. The current sample sizes in AOU are too small for ancestry groups other than EUR to do this. Furthermore, the estimated effect sizes observed in this study may be an underestimate as cases in AOU data may represent a healthier cohort than typical CRC cases due to survival bias. This is evidenced by the fact that the diagnosis of CRC occurred before entry into the study by a mean of 5 years with a maximum of 27 years. This phenomenon of biobank participants being healthier than the population they represent has been observed previously [31–33].

Ultimately, the PRS requires further development in order to be equitably applied in the clinical setting. Most pressing is the addition of samples with genetic similarity to AFR and MID ancestry in further development of the PRS. In the U.S., the lifetime risk for CRC is highest in self-identified black and Native American populations. A comprehensive score that reflects this difference in risk associated with ancestry, race and ethnicity may be preferable. One way to ensure this is to combine the PRS with the overall expected lifetime risk, given genetic ancestry, social determinants of health, and other known risk factors. A more comprehensive risk calculation incorporating these predictors is a topic for future research.

Declaration of Interests

The authors declare no competing interests.

Acknowledgments

The authors would like to thank Alyna T. Khan for her valuable comments and suggestions. This work was funded by the Office of the Director at the National Institute of Health, under award notice 1OT2OD002748-01 and by the NHGRI through the grant U01HG008657.

Web Resources

Colorectal Cancer case/control algorithm: phekb.org/phenotype/colorectal-cancer-crc

1000 Genomes Data: https://www.internationalgenome.org/home
https://www.internationalgenome.org/data-portal/search?q=1000%2Bgenomes

Human Genome Diversity Project: https://www.internationalgenome.org/data-portal/data-collection/hgdp

Data and code availability

Data from the NIH All of Us study are available via institutional data access for researchers who meet the criteria for access to confidential data. To register as a researcher with All of Us, researchers may use the following URL and complete the laid out steps: https://www.researchallofus.org/register/. Researchers can contact All of Us Researcher Workbench Support at support@researchallofus.org. Code used in this study is available at the pre-production Researcher Workbench "CRC PRS Analysis (Interactive notebooks)".

References

1. Siegel, R.L., Wagle, N.S., Cercek, A., Smith, R.A., and Jemal, A. (2023). Colorectal cancer statistics, 2023. CA Cancer J. Clin. *73*, 233–254.

2. Bundy, J.D., Mills, K.T., He, H., LaVeist, T.A., Ferdinand, K.C., Chen, J., and He, J. (2023). Social determinants of health and premature death among adults in the USA from 1999 to 2018: a national cohort study. Lancet Public Health *8*, e422–e431.

3. Siegel, R.L., Miller, K.D., Goding Sauer, A., Fedewa, S.A., Butterly, L.F., Anderson, J.C., Cercek, A., Smith, R.A., and Jemal, A. (2020). Colorectal cancer statistics, 2020. CA Cancer J. Clin. *70*, 145–164.

4. Haverkamp, D., Redwood, D., Roik, E., Vindigni, S., and Thomas, T. (2023). Elevated colorectal cancer incidence among American Indian/Alaska Native persons in Alaska compared to other populations worldwide. Int. J. Circumpolar Health *82*, 2184749.

5. Sawicki, T., Ruszkowska, M., Danielewicz, A., Niedźwiedzka, E., Arłukowicz, T., and Przybyłowicz, K.E. (2021). A Review of Colorectal Cancer in Terms of Epidemiology, Risk Factors, Development, Symptoms and Diagnosis. Cancers *13*,.

6. Rosenthal, E.A., Shirts, B.H., Amendola, L.M., Horike-Pyne, M., Robertson, P.D., Hisama, F.M., Bennett, R.L., Dorschner, M.O., Nickerson, D.A., Stanaway, I.B., et al. (2018). Rare loss of function variants in candidate genes and risk of colorectal cancer. Hum. Genet. *137*, 795–806.

7. Patel, S.G., and Ahnen, D.J. (2012). Familial colon cancer syndromes: an update of a rapidly evolving field. Curr. Gastroenterol. Rep. *14*, 428–438.

8. Lowery, J.T., Ahnen, D.J., Schroy, P.C., 3rd, Hampel, H., Baxter, N., Boland, C.R., Burt, R.W., Butterly, L., Doerr, M., Doroshenk, M., et al. (2016). Understanding the contribution of family history to colorectal cancer risk and its clinical implications: A state-of-the-science review. Cancer *122*, 2633–2645.

9. Archambault, A.N., Su, Y.-R., Jeon, J., Thomas, M., Lin, Y., Conti, D.V., Win, A.K., Sakoda, L.C., Lansdorp-Vogelaar, I., Peterse, E.F.P., et al. (2020). Cumulative Burden of Colorectal Cancer-Associated Genetic Variants Is More Strongly Associated With Early-Onset vs Late-Onset Cancer. Gastroenterology *158*, 1274–1286.e12.

10. Thomas, M., Su, Y.-R., Rosenthal, E.A., Sakoda, L.C., Schmit, S.L., Timofeeva, M.N., Chen, Z., Fernandez-Rozadilla, C., Law, P.J., Murphy, N., et al. (2023). Combining Asian-European Genome-Wide Association Studies of Colorectal Cancer Improves Risk Prediction Across Race and Ethnicity. medRxiv. Under Review

11. Thomas, M., Sakoda, L.C., Hoffmeister, M., Rosenthal, E.A., Lee, J.K., van Duijnhoven, F.J.B., Platz, E.A., Wu, A.H., Dampier, C.H., de la Chapelle, A., et al. (2020). Genome-wide Modeling of Polygenic Risk Score in Colorectal Cancer Risk. Am. J. Hum. Genet. *107*, 432–444.

12. Briggs, S.E.W., Law, P., East, J.E., Wordsworth, S., Dunlop, M., Houlston, R., Hippisley-Cox, J., and Tomlinson, I. (2022). Integrating genome-wide polygenic risk scores and non-genetic risk to predict colorectal cancer diagnosis using UK Biobank data: population based cohort study. BMJ *379*, e071707.

13. Choi, S.W., Mak, T.S.-H., and O'Reilly, P.F. (2020). Tutorial: a guide to performing polygenic risk score analyses. Nat. Protoc. *15*, 2759–2772.

14. Privé, F., Arbel, J., and Vilhjálmsson, B.J. (2021). LDpred2: better, faster, stronger. Bioinformatics *36*, 5424–5431.

15. Márquez-Luna, C., Loh, P.-R., South Asian Type 2 Diabetes (SAT2D) Consortium, SIGMA Type 2 Diabetes Consortium, and Price, A.L. (2017). Multiethnic polygenic risk scores improve risk prediction in diverse populations. Genet. Epidemiol. *41*, 811–823.

16. Privé, F., Arbel, J., Aschard, H., and Vilhjálmsson, B.J. (2022). Identifying and correcting for misspecifications in GWAS summary statistics and polygenic scores. HGG Adv *3*, 100136.

17. Ge, T., Irvin, M.R., Patki, A., Srinivasasainagendra, V., Lin, Y.-F., Tiwari, H.K., Armstrong, N.D., Benoit, B., Chen, C.-Y., Choi, K.W., et al. (2022). Development and validation of a trans-ancestry polygenic risk score for type 2 diabetes in diverse populations. Genome Med. *14*, 70.

18. Khera, A.V., Chaffin, M., Zekavat, S.M., Collins, R.L., Roselli, C., Natarajan, P., Lichtman, J.H., D'Onofrio, G., Mattera, J., Dreyer, R., et al. (2019). Whole-Genome Sequencing to Characterize Monogenic and Polygenic Contributions in Patients Hospitalized With Early-Onset Myocardial Infarction. Circulation *139*, 1593–1602.

19. All of Us Research Program Investigators, Denny, J.C., Rutter, J.L., Goldstein, D.B., Philippakis, A., Smoller, J.W., Jenkins, G., and Dishman, E. (2019). The "All of Us" Research Program. N. Engl. J. Med. *381*, 668–676.

20. Ding, Y., Hou, K., Burch, K.S., Lapinska, S., Privé, F., Vilhjálmsson, B., Sankararaman, S., and Pasaniuc, B. (2022). Large uncertainty in individual polygenic risk score estimation impacts PRS-based risk stratification. Nat. Genet. *54*, 30–39.

21. Martin, A.R., Gignoux, C.R., Walters, R.K., Wojcik, G.L., Neale, B.M., Gravel, S., Daly, M.J., Bustamante, C.D., and Kenny, E.E. (2017). Human Demographic History Impacts Genetic Risk Prediction across Diverse Populations. Am. J. Hum. Genet. *100*, 635–649.

22. Duncan, L., Shen, H., Gelaye, B., Meijsen, J., Ressler, K., Feldman, M., Peterson, R., and Domingue, B. (2019). Analysis of polygenic risk score usage and performance in diverse human populations. Nat. Commun. *10*, 3328.

23. Sulieman, L., Cronin, R.M., Carroll, R.J., Natarajan, K., Marginean, K., Mapes, B., Roden, D., Harris, P., and Ramirez, A. (2022). Comparing medical history data derived from electronic health records and survey answers in the All of Us Research Program. J. Am. Med. Inform. Assoc. *29*, 1131–1141.

24. 1000 Genomes Project Consortium, Auton, A., Brooks, L.D., Durbin, R.M., Garrison, E.P., Kang, H.M., Korbel, J.O., Marchini, J.L., McCarthy, S., McVean, G.A., et al. (2015). A global reference for human genetic variation. Nature *526*, 68–74.

25. Bergström, A., McCarthy, S.A., Hui, R., Almarri, M.A., Ayub, Q., Danecek, P., Chen, Y., Felkel, S., Hallast, P., Kamm, J., et al. (2020). Insights into human genetic variation and population history from 929 diverse genomes. Science *367*,.

26. Zheng, X., and Weir, B.S. (2016). Eigenanalysis of SNP data with an identity by descent interpretation. Theor. Popul. Biol. *107*, 65–76.

27. Hsu, C.-Y., Yang, W., Parikh, R.V., Anderson, A.H., Chen, T.K., Cohen, D.L., He, J., Mohanty, M.J., Lash, J.P., Mills, K.T., et al. (2021). Race, Genetic Ancestry, and Estimating Kidney Function in CKD. N. Engl. J. Med. *385*, 1750–1760.

28. Bryc, K., Durand, E.Y., Macpherson, J.M., Reich, D., and Mountain, J.L. (2015). The genetic ancestry of African Americans, Latinos, and European Americans across the United States. Am. J. Hum. Genet. *96*, 37–53.

29. Hail Team (2021). Hail 0.2.13-81ab564db2b4. https://github.com/hail-is/hail/releases/tag/0.2.13.

30. Pepe, M.S., Kerr, K.F., Longton, G., and Wang, Z. (2013). Testing for improvement in prediction model performance. Stat. Med. *32*, 1467–1482.

31. Fry, A., Littlejohns, T.J., Sudlow, C., Doherty, N., Adamska, L., Sprosen, T., Collins, R., and Allen, N.E. (2017). Comparison of Sociodemographic and Health-Related Characteristics of UK Biobank Participants With Those of the General Population. Am. J. Epidemiol. *186*, 1026–1034.

32. Schoeler, T., Speed, D., Porcu, E., Pirastu, N., Pingault, J.-B., and Kutalik, Z. (2023). Participation bias in the UK Biobank distorts genetic associations and downstream analyses. Nat Hum Behav.

33. Lee, Y.H., Thaweethai, T., Sheu, Y.-H., Feng, Y.-C.A., Karlson, E.W., Ge, T., Kraft, P., and Smoller, J.W. (2022). Impact of selection bias on polygenic risk score estimates in healthcare settings.

Figures

Figure 1: Each PRS by ancestry. A) RAW=original PRS. B) PC_μ=PRS adjusted by the first 5 PCs of ancestry in AOU. C) AD_μ=PRS adjusted by admixture estimates in AOU. D) PC.REF_μ.=PRS adjusted using a model trained on the first 5 PCs of ancestry in the 1KG and HGDP reference dataset. E) PC.AOU_μσ=PRS adjusted using a model trained on the first 5 PCs of ancestry in the subset of unassigned AOU participants (AOU.REF). F) AD.AOU_μσ=PRS adjusted using a model trained on admixture in AOU.REF. AFR=African, AMR=Admixed American, EAS=East Asian, EUR=European, MID=Middle Eastern, OTH=Other, SAS=South Asian.

Figure 2: OR by quantiles and calibration for African ancestry. A) RAW=raw PRS; B) PC_μ=raw PRS adjusted for first 5 PCs of ancestry; C) AD_μ=raw PRS adjusted for admixture; D) PC.AOU_μσ=raw PRS adjusted using a training model developed using random subset of unscreened controls (AOU.REF) E) AD.AOU_μσ=raw PRS adjusted using a training model developed using AOU.REF and admixture estimates rather than PCs.

Figure 3: OR by quantiles and calibration for Admixed American ancestry. A) RAW=raw PRS; B) PC_μ=raw PRS adjusted for first 5 PCs of ancestry; C) AD_μ=raw PRS adjusted for admixture; D) PC.AOU_μσ=raw PRS adjusted using a training model developed using random subset of unscreened controls (AOU.REF). E) AD.AOU_μσ=raw PRS adjusted using a training model developed using AOU.REF and admixture estimates rather than PCs.

Figure 4: OR by quantiles and calibration for European ancestry. RAW=raw PRS; PC_μ=raw PRS adjusted for first 5 PCs of ancestry; AD_μ=raw PRS adjusted for admixture; PC.AOU_μσ=raw PRS adjusted using a training model developed using random subset of unscreened controls (AOU.REF). AD.AOU_μσ=raw PRS adjusted using a training model developed using AOU.REF and admixture estimates rather than PCs.

Figure 5: OR by quantiles and calibration for Middle Easetern ancestry. A) RAW=raw PRS; B) PC_μ=raw PRS adjusted for first 5 PCs of ancestry; C) AD_μ=raw PRS adjusted for admixture; D) PC.AOU_μσ=raw PRS adjusted using a training model developed using random subset of unscreened controls (AOU.REF). E) AD.AOU_μσ=raw PRS adjusted using a training model developed using AOU.REF and admixture estimates rather than PCs.

Figure 6: OR by quantiles and calibration for Other ancestry. A) RAW=raw PRS; B) PC_μ=raw PRS adjusted for first 5 PCs of ancestry; C) AD_μ=raw PRS adjusted for admixture; D) PC.AOU_μσ=raw PRS adjusted using a training model developed using random subset of unscreened controls (AOU.REF.) E) AD.AOU_μσ=raw PRS adjusted using a training model developed using AOU.REF and admixture estimates rather than PCs.

Figure 7: Observed versus expected upper percentiles, by ancestry and PRS. A) African B) Admixed American C) East Asian D) European E)Middle Eastern F) Other G) South Asian genetic ancestries. For each panel, solid colored line is observed upper 10th percentile. Dashed colored line is observed upper 7.5 percentile. Dotted colored line is observed upper 5th percentile. 95% C.I. are indicated with error bars. Horizontal black solid, dashed and dotted lines indicate the expected percentiles, respectively. RAW=raw PRS; PC_μ=raw PRS adjusted for first 5 PCs of ancestry; AD_μ=raw PRS adjusted for admixture; PC.AOU_μσ=raw PRS adjusted using a training model developed using random subset of unscreened controls (AOU.REF). AD.AOU_μσ=raw PRS adjusted using a training model developed using AOU.REF and admixture estimates rather than PCs.

Tables

Table 1: Count (percent) by ancestry and case status. “Cleaned cases” refers to participants assigned case status according to the CRC phenotyping algorithm. “Screened controls” refers to participants assigned control status by the CRC phenotyping algorithm. “Unassigned” refers to participants who were not excluded from analysis and were not assigned to case or control status by the CRC phenotype algorithm. Ancestries are AFR=African, AMR=Admixed American, EAS=East Asian, EUR=European, MID=Middle Eastern, OTH=Other, SAS=South Asian. For AMR, EAS and SAS, only total counts are given as the number of cases was <20, to comply with AOU publication rules.

Ancestry	**Cleaned Cases**	**Screened Controls**	**Unassigned**	**N(%)**
AFR	**42**	2012	20296	22350 (23)
AMR				9360 (9.5)
EAS				2065 (2.1)
EUR	**204**	6625	37102	43931 (45)
MID	**23**	821	3179	4023 (4.1)
OTH	**57**	2085	13305	15447 (16)
SAS				760 (0.8)
All	348	12378	85210	97936

Table 2: Mean observed age (min, max) by ancestry within all cleaned cases (age of onset), screened controls (age of last negative screening) and unassigned participants (age of consent). AFR=African, AMR=Admixed American, EAS=East Asian, EUR=European, MID=Middle Eastern, OTH=Other, SAS=South Asian. Age diff = difference between age of consent and observed age. Distributions are not reported for AMR, EAS and SAS due to AOU publication rules for sample sizes < 20.

Ancestry	**Cleaned cases**	**Screened controls**	**Unscreened controls**
AFR	57(41,83)	57(18,90)	48(18,104)
EUR	62(26,89)	61(18,94)	54(18,103)
MID	65(34,84)	64(18,96)	59(18,101)
OTH	55(19,83)	58(18,88)	46(18,100)
All	60(19,89)	60(18,96)	50(18,104)
Age diff	-5.2 (-27,2.5)	-2.5 (-34, 3.7)	0

Table 3: Mean (s.d.) of PRS by ancestry. AFR=African, AMR=Admixed American, EAS=East Asian, EUR=European, MID=Middle Eastern, OTH=Other, SAS=South Asian. RAW=raw PRS; PC_μ=raw PRS adjusted for first 5 PCs of ancestry; AD_μ=raw PRS adjusted for admixture; PC.REF_μσ=raw PRS adjusted using a training model developed using 1KG and HGDP reference data. PC.AOU_μσ=raw PRS adjusted using a training model developed using random subset of unscreened controls (AOU.REF). AD.AOU_μσ=raw PRS adjusted using a training model developed using AOU.REF and admixture estimates rather than PCs.

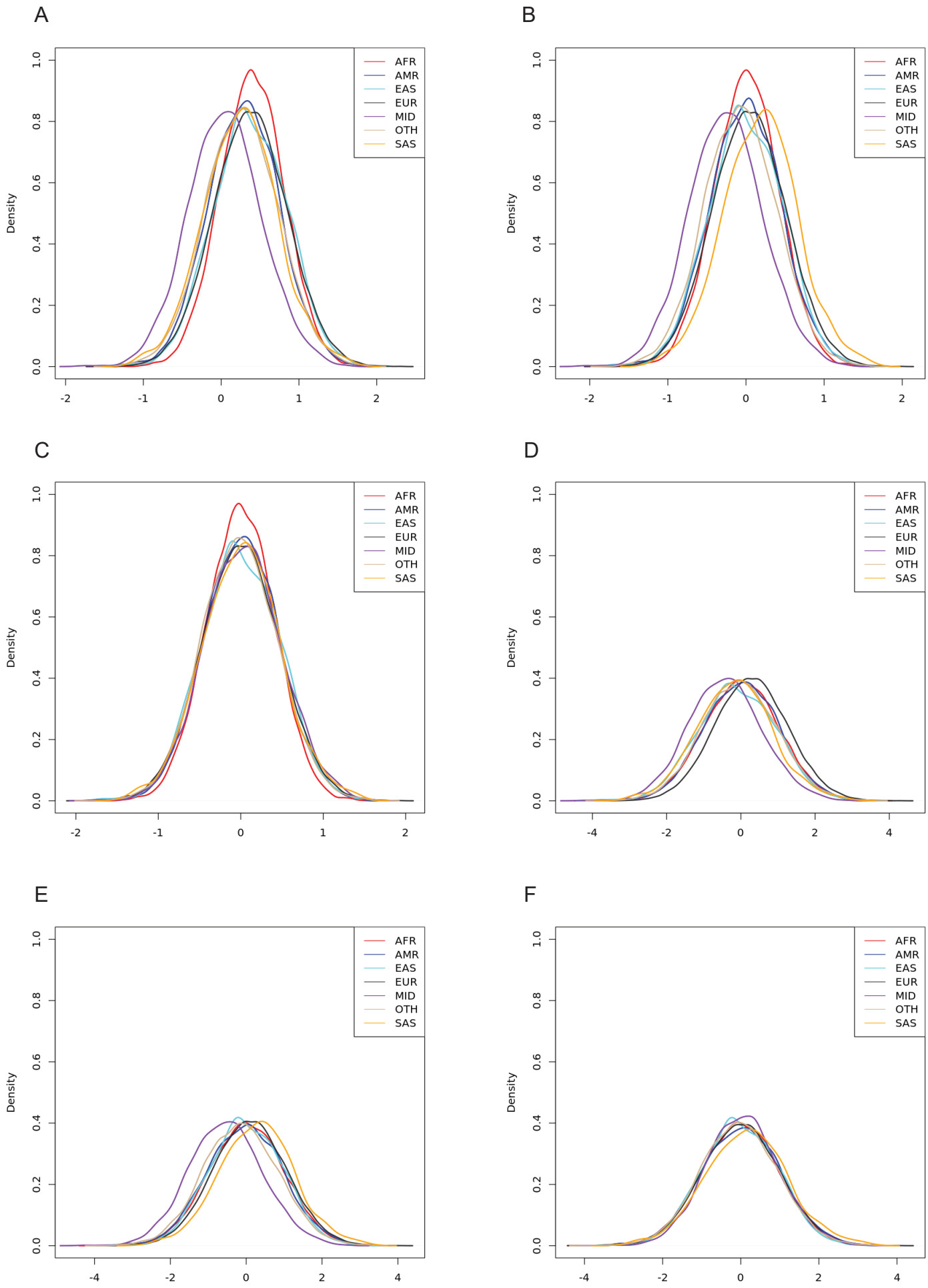
ANC	**N**	**RAW**	**PC_μ**	**AD_μ**	**PC.REF_μσ**	**PC.AOU_μσ**	**AD.AOU_μσ**
AFR	22385	0.40(0.41)	0.01(0.41)	0.00(0.41)	0.01(1.01)	0.01(0.99)	-0.01(0.99)
AMR	9382	0.29(0.45)	-0.01(0.45)	0.00(0.45)	0.01(1.01)	-0.03(1.00)	0.00(1.01)
EAS	2069	0.37(0.47)	0.00(0.47)	0.00(0.47)	-0.1(1.03)	-0.01(0.95)	-0.02(0.95)
EUR	44099	0.38(0.47)	0.04(0.47)	0.00(0.47)	0.28(0.99)	0.10(0.97)	0.01(1.00)
MID	4056	0.05(0.47)	-0.25(0.47)	0.03(0.47)	-0.43(0.99)	-0.53(0.97)	0.04(0.93)
OTH	15505	0.27(0.47)	-0.07(0.47)	-0.01(0.46)	-0.03(1.01)	-0.14(1.00)	-0.03(0.99)
SAS	760	0.25(0.47)	0.18(0.47)	0.01(0.47)	-0.17(1.01)	0.28(0.98)	0.14(1.05)

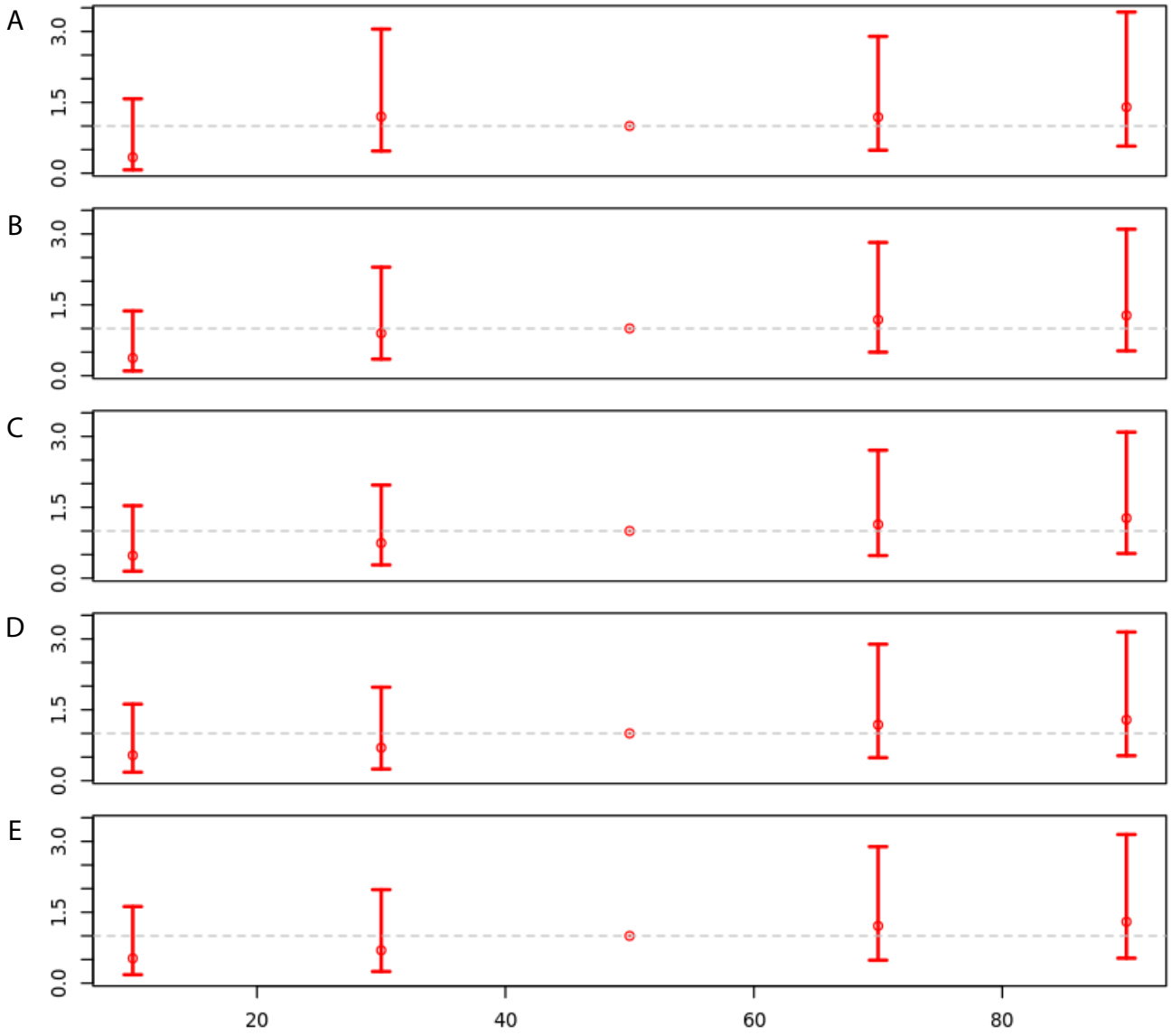
Table 4: Estimated OR (95% C.I.) per s.d. change in PRS, unadjusted (first row) and adjusted for age and sex (second row), by ancestry. AFR=African, AMR=Admixed American, EAS=East Asian, EUR=European, MID=Middle Eastern, OTH=Other, SAS=South Asian. RAW=raw PRS; PC_μ=raw PRS adjusted for first 5 PCs of ancestry; AD_μ=raw PRS adjusted for admixture; PC.AOU_μσ=raw PRS adjusted using a training model developed using random subset of unscreened controls (AOU.REF). AD.AOU_μσ=raw PRS adjusted using a training model developed using AOU.REF and admixture estimates rather than PCs.

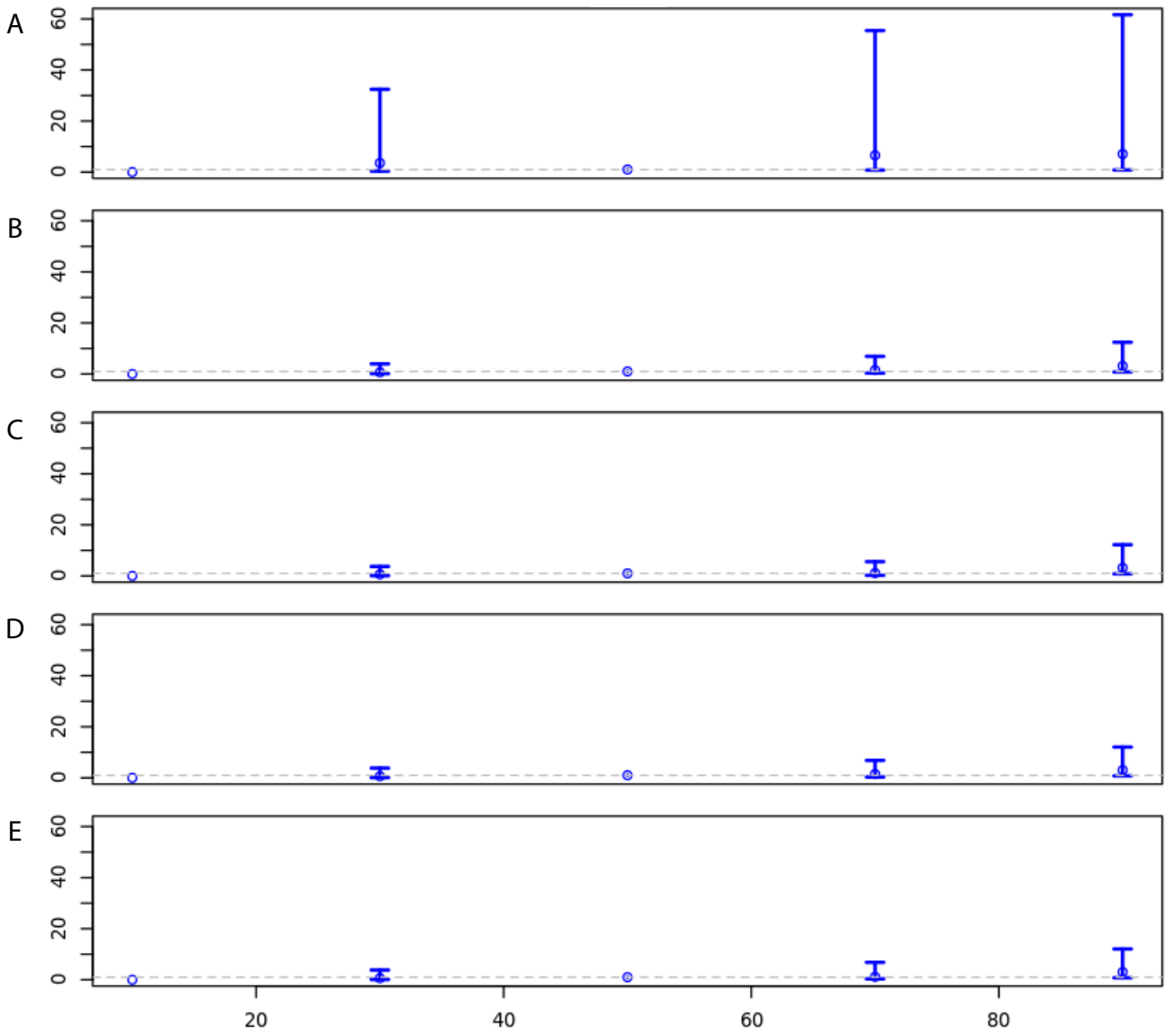
Ancestry	**RAW**	**PC_μ**	**PC.AOU_μσ**	**AD_μ**	**AD.AOU_μσ**
AFR	1.5(1.02,2.08)	1.5(1.02,2.07)	1.4(1.01,1.93)	1.4(1.01,2.04)	1.4(1.01,1.91)
	1.4(0.99,2.02)	1.4(0.99,2.02)	1.4(0.99,1.88)	1.4(0.98,1.99)	1.4(0.98,1.87)
AMR	2.2(1.27,3.85)	2.2(1.26,3.80)	2.2(1.26,3.72)	2.2(1.29,3.84)	2.2(1.28,3.72)
	2.2(1.24,3.75)	2.1(1.23,3.70)	2.1(1.23,3.61)	2.2(1.26,3.74)	2.1(1.25,3.63)
EUR	1.6(1.43,1.89)	1.6(1.43,1.88)	1.7(1.46,1.95)	1.6(1.43,1.87)	1.7(1.45,1.92)
	1.7(1.44,1.90)	1.7(1.44,1.90)	1.7(1.47,1.96)	1.6(1.44,1.88)	1.7(1.46,1.93)
MID	1.1(0.71,1.63)	1.1(0.72,1.63)	1.1(0.7,1.67)	1.1(0.74,1.67)	1.1(0.71,1.77)
	1.1(0.7,1.60)	1.1(0.7,1.59)	1.1(0.68,1.63)	1.1(0.72,1.63)	1.1(0.69,1.73)
OTH	1.5(1.17,1.98)	1.5(1.16,1.97)	1.5(1.17,1.99)	1.5(1.16,1.97)	1.5(1.17,2)
	1.5(1.18,2.01)	1.5(1.17,2.00)	1.6(1.18,2.02)	1.5(1.18,2)	1.5(1.18,2.04)

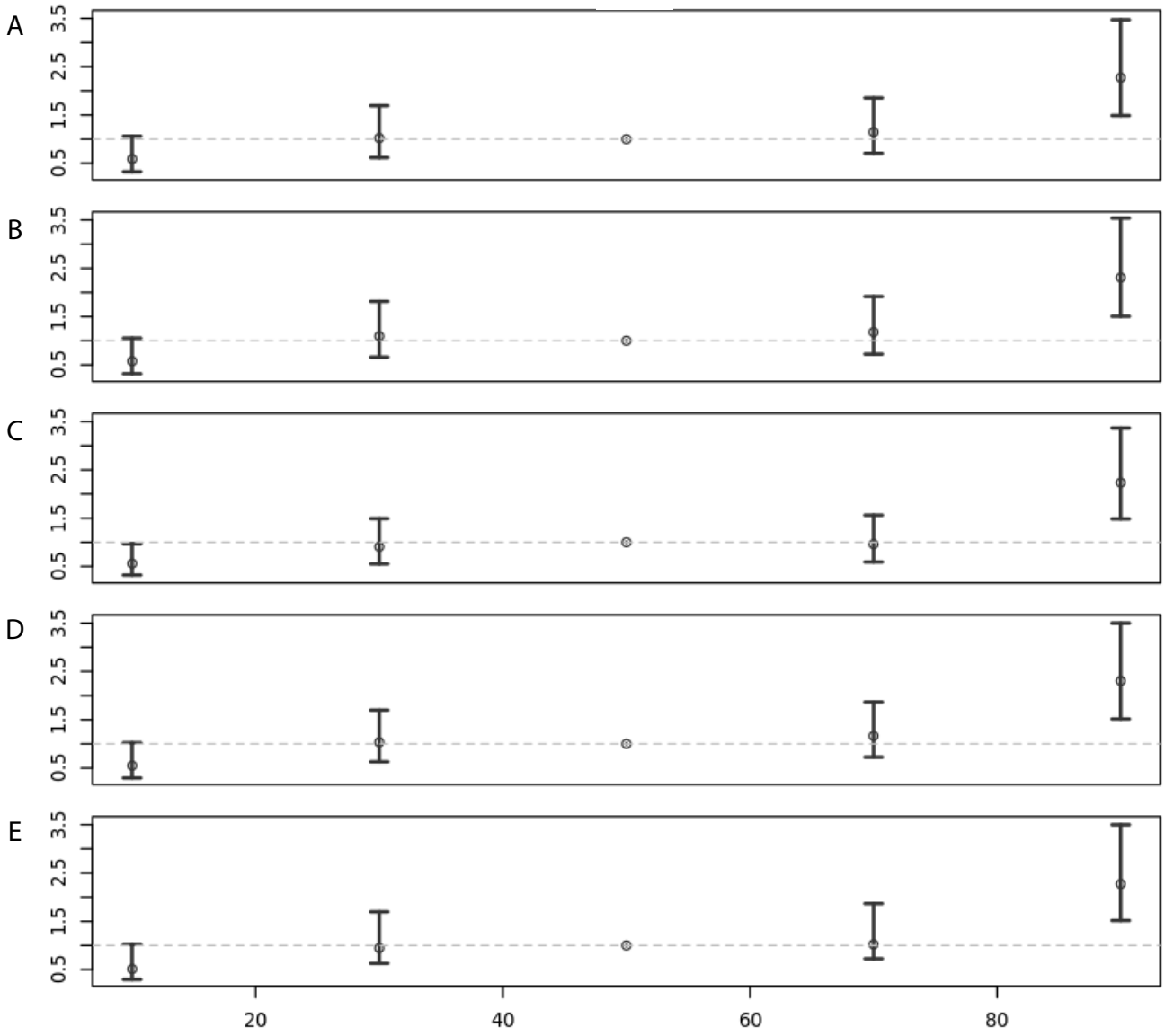
Table 5: Estimated AUC and 95% C.I. for each PRS, by ancestry, comparing cleaned cases to screened controls. For each PRS, the first row contains results for the unadjusted AUC and the second row contains results for the AUC adjusted for age and sex. Although the theoretical lower bound for the C.I. is 0.5, we include the actual lower value from the bootstrap estimation. AFR=African, AMR=Admixed American, EAS=East Asian, EUR=European, MID=Middle Eastern, OTH=Other, SAS=South Asian. RAW=raw PRS; PC_μ=raw PRS adjusted for first 5 PCs of ancestry; AD_μ=raw PRS adjusted for admixture; PC.AOU_μσ=raw PRS adjusted using a training model developed using random subset of unscreened controls (AOU.REF). AD.AOU_μσ=raw PRS adjusted using a training model developed using AOU.REF and admixture estimates rather than PCs.

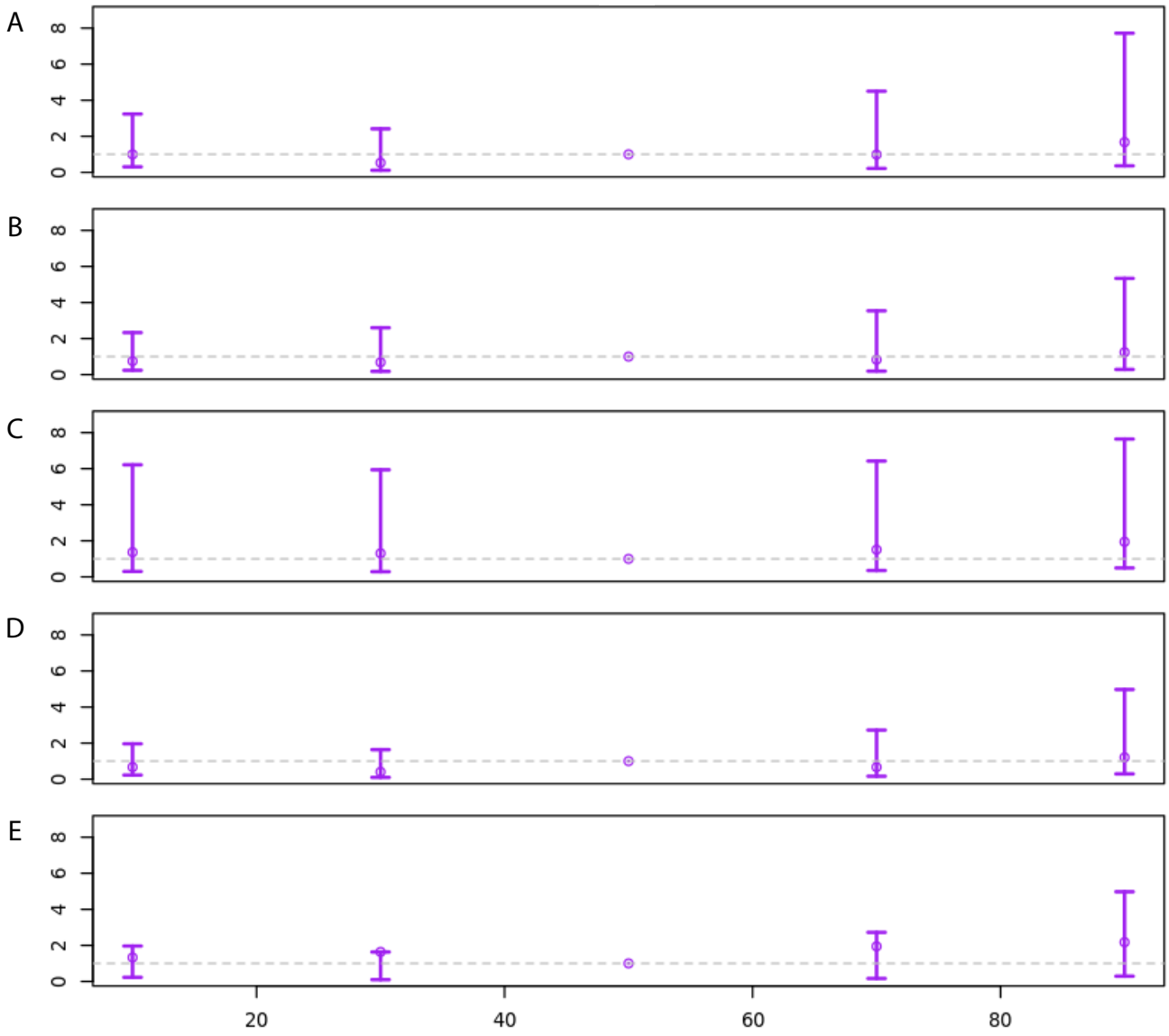
Ancestry	**RAW**	**PC_μ**	**PC.AOU_μσ**	**AD_μ**	**AD.AOU_μσ**
AFR	0.6 (0.51, 0.67)	0.6 (0.51, 0.67)	0.6 (0.51, 0.67)	0.6 (0.50, 0.67)	0.6 (0.50, 0.67)
	0.6 (0.50, 0.69)	0.6 (0.49, 0.68)	0.6 (0.50, 0.69)	0.6 (0.50, 0.69)	0.6 (0.50, 0.69)
AMR	0.7 (0.60, 0.83)	0.7 (0.60, 0.83)	0.7 (0.60, 0.82)	0.7 (0.61, 0.83)	0.7 (0.61, 0.83)
	0.7 (0.58, 0.80)	0.7 (0.57, 0.80)	0.7 (0.57, 0.80)	0.7 (0.58, 0.80)	0.7 (0.59, 0.81)
EUR	0.6 (0.59, 0.67)	0.6 (0.59, 0.67)	0.6 (0.59, 0.67)	0.6 (0.59, 0.67)	0.6 (0.59, 0.67)
	0.6 (0.59, 0.68)	0.6 (0.59, 0.68)	0.6 (0.59, 0.68)	0.6 (0.59, 0.67)	0.6 (0.59, 0.68)
MID	0.5 (0.40, 0.67)	0.5 (0.40, 0.67)	0.5 (0.40, 0.67)	0.5 (0.41, 0.67)	0.5 (0.41, 0.67)
	0.5 (0.39, 0.70)	0.5 (0.39, 0.70)	0.5 (0.38, 0.69)	0.6 (0.40, 0.71)	0.6 (0.40, 0.71)
OTH	0.6 (0.56, 0.70)	0.6 (0.56, 0.70)	0.6 (0.56, 0.70)	0.6 (0.56, 0.70)	0.6 (0.56, 0.70)
	0.7 (0.58, 0.71)	0.6 (0.57, 0.71)	0.6 (0.56, 0.70)	0.6 (0.58, 0.71)	0.6 (0.58, 0.71)

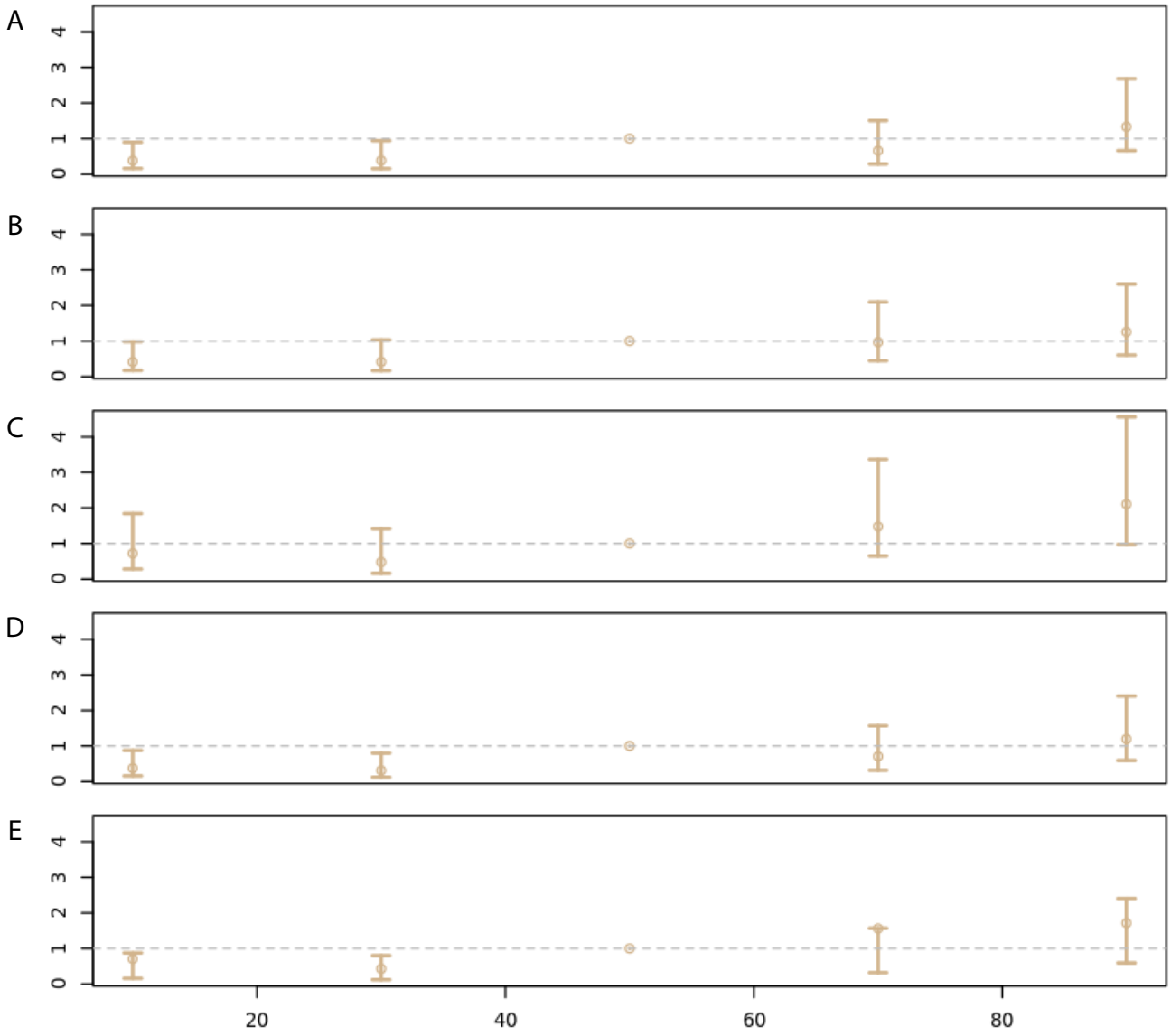

A
B
C
D
E
F
Density
AFR
AMR
EAS
EUR
MID
OTH
SAS

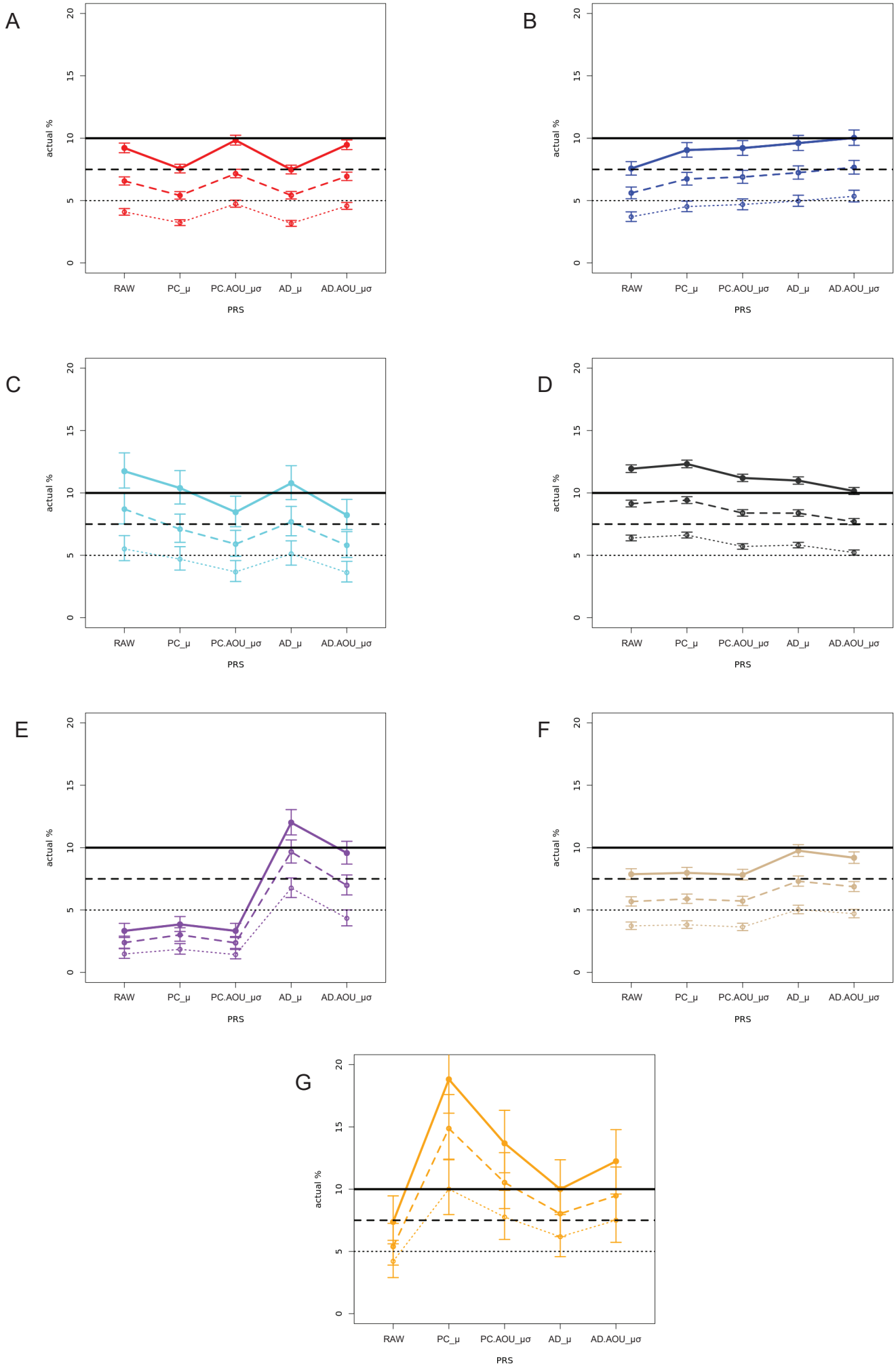

A
B
C
D
E
0.0
1.5
3.0
20
40
60
80


A
B
C
D
E
0
20
40
60
20
40
60
80


A
B
C
D
E
0.5
1.5
2.5
3.5
20
40
60
80


A
B
C
D
E
0
2
4
6
8
20
40
60
80


A
B
C
D
E
0
1
2
3
4
20
40
60
80


A
B
C
D
E
F
G
actual %
0
5
10
15
20
RAW
PC_μ
PC.AOU_μσ
AD_μ
AD.AOU_μσ
PRS