A Novel PGS for Pediatric Sleep Disordered Breathing Trained on ABCD Suggestive SNPs Outperforms PGS Catalog

Parisa Boodaghi Malidarreh[1], Mohammad Sadegh Nasr[1], Dongdong Li[2], Paul Yi[3], Vishwa Parekh[3], Amal Isaiah[3,4,5], Jacob M Luber[1]

[1] Department of Computer Science and Engineering, University of Texas at Arlington, Arlington, TX, USA

[2]Department of Population Medicine, Harvard Medical School and Harvard Pilgrim Health Care Institute

[3]Department of Diagnostic Radiology and Nuclear Medicine, University of Maryland School of Medicine, Baltimore, MD, USA

[4]Department of Otorhinolaryngology—Head and Neck Surgery, University of Maryland School of Medicine, Baltimore, MD, USA

[5]Department of Pediatrics, University of Maryland School of Medicine, Baltimore, MD, USA

Introduction

Pediatric Sleep Disordered Breathing (SDB) encompasses a spectrum of respiratory disorders occurring during sleep, ranging from primary snoring to obstructive sleep apnea (OSA). These conditions are of significant concern due to their potential impact on children's physical health, cognitive development, and quality of life. The etiology of pediatric SDB is multifaceted, involving anatomical, neuromuscular, and genetic factors, highlighting the need for comprehensive diagnostic and predictive tools [2].

The Polygenic Score (PGS) Catalog offers a comprehensive repository of PGS across various health conditions and traits, including sleep disorders. These scores represent the cumulative effect of multiple genetic variants on an individual's trait or disease risk. The catalog facilitates comparison and validation of PGS across different studies and populations. Many of these scores are based on the UK Biobank. The UK Biobank is a large-scale biomedical database and research resource containing in-depth genetic and health information from half a million UK participants. This resource has been instrumental in numerous genetic studies, including those related to sleep disorders. However, its representation is predominantly of European ancestry, which may limit its applicability to diverse populations [3].

In this study, we propose to investigate the effectiveness of Polygenic Scores (PGS) for predicting Sleep Disordered Breathing (SDB) based on Adolescent Brain Cognitive Development (ABCD) study. Our preliminary analysis suggests that Single Nucleotide Polymorphisms (SNPs) data,

derived from the NIDA NIH ABCD cohort, exhibit a superior predictive capability compared to those derived from the UK Biobank.

Methods

The Adolescent Brain Cognitive Development (ABCD) study, funded by the National Institute on Drug Abuse (NIDA) and the National Institutes of Health (NIH), represents the largest long-term study of brain development and child health in the United States [4]. This cohort provides invaluable data on genetic, neurobiological, behavioral, environmental, and social factors that influence health and disease outcomes, including SDB. The ABCD cohort's genetic data, especially SNPs suggestive of SDB risk, present a unique resource for developing predictive models [4]. The baseline performance of PGS Catalog scoring for SDB, primarily derived from the UK Biobank data, provides a foundational understanding of genetic predisposition to SDB. Leveraging the smokescreen panel for imputation of SNPs in diverse ancestries, including European and African, led to enhanced performance in custom PGS creation, surpassing baseline metrics provided by the PGS catalog. These improvements were quantitatively validated using the Mann Whitney U statistic to compare Receiver Operating Characteristic (ROC) curve Area Under the Curve (AUC) metrics.

Results

Our analysis reveals that custom PGS for SDB, developed using the ABCD cohort's suggestive SNPs and enhanced through imputation panels like the smokescreen and TopMed for various ancestries (top SNPs listed in Table 1) [5], significantly outperforms existing scores from the PGS catalog (Figure 1, European ancestries AUC = .54 vs .44, P = 9.7e-105, African Ancestries AUC = .56 vs .52, p = 0.00008, Other ancestries AUC = 0.64 vs .49, p value = 0.0009). This superiority was evident across European, African, and other ancestries, underscoring the importance of a diverse genetic database. The Mann Whitney U statistic's application in comparing ROC curve AUCs underscored the predictive accuracy of our custom PGS, highlighting the potential limitations of relying solely on data from populations of European descent.

Discussion

The observed predictive performance improvement suggests that pediatric SDB may be more significantly influenced by environmental than genetic factors, a hypothesis warranting further investigation. Future research should focus on integrating environmental and lifestyle variables with genetic data to enhance predictive models for SDB. This holistic approach could lead to more effective screening, prevention, and management strategies for pediatric SDB, ultimately improving health outcomes for affected children worldwide.

Acknowledgments

we extend our deepest gratitude to Biraaj Rout, Yike Shen, and Helen H Shang for their invaluable contributions to this study. Biraaj Rout offered significant support in software domain. Yike Shen and Helen H Shang provided critical insights that shaped our research. This work was supported by the Cancer Prevention and Research Institute of Texas (CPRIT) Recruitment of First-Time, Tenure-Track Faculty Members Grant (RR220015) (JML) and University of Texas System Rising STARs award (JML).

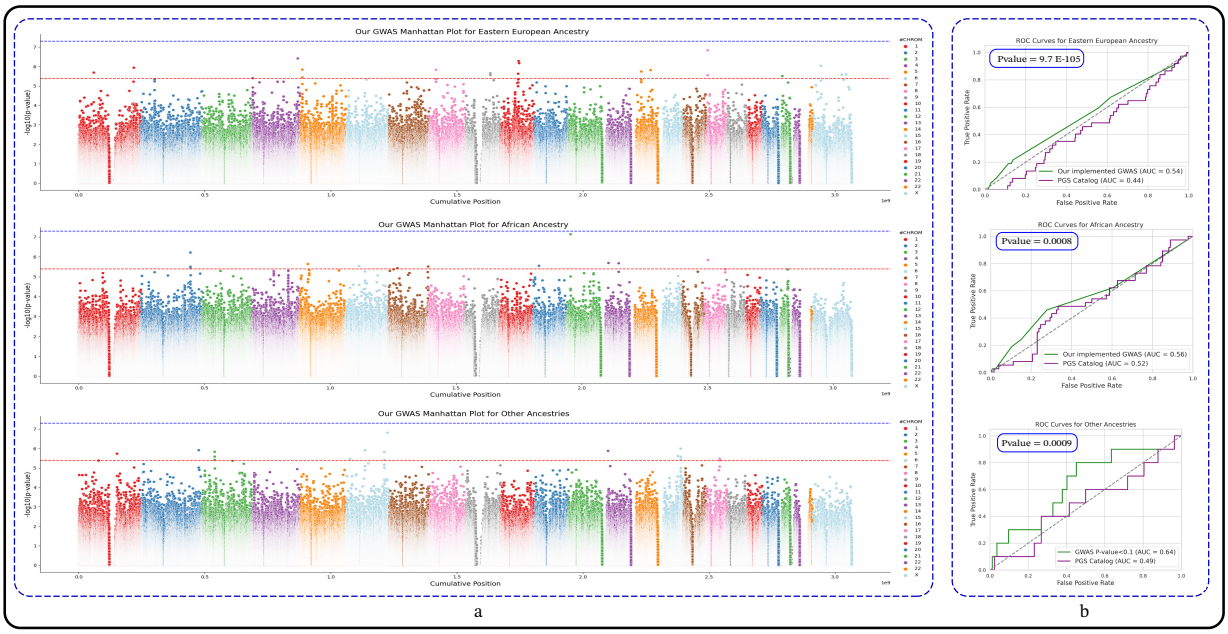

Figure 1. Manhattan plot and Receiver Operating Characteristic (ROC) Curves illustration. Panel a presents the Manhattan plot for Eastern European Ancestry, African Ancestry, and Other Ancestries, where each point represents a single nucleotide polymorphism (SNP). To distinguish between SNPs of varying significance, two thresholds are introduced: a stringent 'significant' threshold (p-value = 5e-8, blue dashed line) for highly relevant SNPs, and a relaxed 'suggestive' threshold (p-value = 4e-6, red dashed line) to identify SNPs that may have a potential association with the trait under investigation. Panel b depicts the ROC curves for a logistic regression model trained on each ancestry group within the validation dataset, alongside a comparison to existing Polygenic Score (PGS) catalog results for snoring. The legend details the differences in the Area Under Curve (AUC) between our GWAS results and the PGS catalog, highlighting the predictive performance for each group.

Race	Chromosome	Position	SNP ID	P-value	FDR-BH	Gene name
White	10	70709952	rs117268431	5.19549E-07	0.357889	ADAMTS14
	8	26605828	rs13249543	1.46729E-06	0.357889	DPYSL2
	14	78204338	rs6574425	1.52505E-06	0.357889	NRXN3
	14	41116442	rs12894660	1.80973E-06	0.357889	LINC02315
	9	97177781	rs17323053	2.21004E-06	0.357889	ANKRD18CP
	10	68850958	rs11815653	2.35406E-06	0.357889	STOX1
	9	97282823	rs75432338	2.73314E-06	0.357889	SUGT1P4-STRA6LP
	17	5510947	rs78233376	2.80797E-06	0.357889	NLRP1
Black	2	195262759	rs112707587	6.06144E-07	0.992139	LOC105376755
	5	31900352	rs116517295	2.29484E-06	0.992139	PDZD2
	7	158025866	rs735806	3.07193E-06	0.992139	PTPRN2
	2	195243225	rs113654421	3.1513E-06	0.992139	LOC105376755
	2	195224736	rs113215629	3.60489E-06	0.992139	LOC105376755
other	6	75145535	rs75280438	1.21373E-06	0.489263	COL12A1
	2	227196599	rs10174487	1.21711E-06	0.489263	COL4A3
	13	23289150	rs75038170	1.32385E-06	0.489263	SGCG
	1	151899720	rs113828754	1.85022E-06	0.489263	7THEM4
	15	79234855	rs145755639	2.4308E-06	0.489263	ANKRD34C-AS1
	3	48045219	rs116786562	2.588E-06	0.489263	MAP4
	15	90395419	rs28372225	3.72282E-06	0.489263	IQGAP1
	3	47601616	rs115329200	3.82528E-06	0.489263	SMARCC1
	3	47684466	rs116026478	3.82528E-06	0.489263	SMARCC1
	3	47906985	rs13072917	3.82528E-06	0.489263	MAP4

Table 1: Dataset of different races with P-value and FDR correction

References

Baurley, James W., et al. "Smokescreen: a targeted genotyping array for addiction research." *BMC genomics* 17 (2016): 1-12.

Isaiah, A., Ernst, T.M., Liang, H. *et al.* Associations between socioeconomic gradients and racial disparities in preadolescent brain outcomes. *Pediatr Res* **94**, 356–364 (2023). https://doi.org/10.1038/s41390-022-02399-9

Lambert, Samuel A., et al. "The Polygenic Score Catalog as an open database for reproducibility and systematic evaluation." *Nature Genetics* 53.4 (2021): 420-425.

Garavan, H., et al. "Recruiting the ABCD sample: Design considerations and procedures." *Developmental cognitive neuroscience* 32 (2018): 16-22.

Taliun, Daniel, et al. "Sequencing of 53,831 diverse genomes from the NHLBI TOPMed Program." *Nature* 590.7845 (2021): 290-299.

Tanigawa Y, Qian J, et al. “Significant sparse polygenic risk scores across 813 traits in UK Biobank.” PLoS Genet.” 18.3 (2022): e1010105

Supplementary Materials

GWAS of snoring trait in ABCD (Adolescent Brain Cognitive Development) dataset: We conducted a Genome-Wide Association Study (GWAS) on 5,720 participants before initiating genotype data quality control. This participant group was composed of 4,021 individuals identifying as White, 1,087 as Black, and 612 as belonging to other races. The ABCD study itself is a large-scale, forward-looking longitudinal study that is observing the developmental trajectory of approximately 12,000 young individuals, beginning at the ages of 9-10, over a span of ten years at 21 different research sites across the United States. The study has an extensive range of developmental aspects under its purview, which includes brain structure and function, social and emotional development, cognitive progression, mental health, substance use and perceptions, gender identity, sexual health, as well as a host of physical health and environmental factors.

To ensure the integrity of our GWAS, stringent quality control measures were applied. We removed genetic variants such as single nucleotide polymorphisms (SNPs) with a minor allele frequency (MAF) below 1% and excluded any participant with more than 1% missing genotype data. Moreover, to mitigate the influence of linkage disequilibrium (LD) and to maintain the independence of the genetic variants under study, we utilized the PLINK software to exclude SNPs in high linkage disequilibrium (pairwise $r^2 > 0.8$). This step was critical to refine our dataset by eliminating closely correlated SNPs, thus reducing the possible confounding impact of LD on our genetic association findings. Subsequently, we performed GWAS focusing on the binary trait of snoring, categorizing the participants into non-snoring (control) and habitual snoring (case) groups. We employed generalized logistic regression analysis in Plinkv2 to achieve this. Our model was represented as $y = G\beta_G + X\beta_X + e$ where y denoted the phenotype of interest, which in this study was the snoring status; G represented the matrix of genotypic data; X indicated the matrix of covariates; and e symbolized the error term. To account for population stratification, the first ten principal components (PC1-PC10) were included as covariates in our analysis.

Polygenic Risk Score for Our Implemented GWAS: Our approach to computing the Polygenic Risk Scores (PRS) entailed utilizing the Beta coefficients associated with each single nucleotide polymorphism (SNP) derived from the GWAS for each participant in the study cohort. The dataset was stratified by ethnic categories — Black, White, and Other — and further subdivided into training, validation, and testing subsets. We employed a logistic regression algorithm, training it on the designated training subset, and subsequently evaluating its performance on the validation and testing subsets.

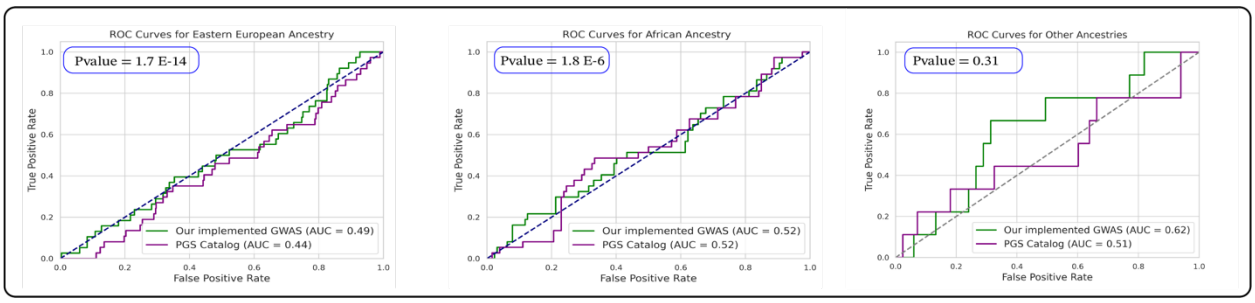

Figure 2. ROC curves for three groups in test dataset