Glovento Journal of Integrated Studies (GJIS) | ISSN: 3117-3314
Volume 2 (2026) | Article 80

Classifying Genetic Risk Profiles for Iron Deficiency Anemia in South Asian Population using Simulated Polygenic Risk Scores and Support Vector Machine

[1]Margaret Grace A. Docdoc, [2]Llewelyn S. Dramayo, [3]Christian V. Maderazo

[1,2,3]Dept. of Computer, Info. Sci. & Math., University of San Carlos, Cebu City, Philippines
Email: [1]22103988@usc.edu.ph, [2]22103210@usc.edu.ph, [3]cvmaderazo@usc.edu.ph

Abstract- Iron deficiency anemia disproportionately affects South Asian populations, yet population-specific genetic risk models for this group remain largely absent, as most existing polygenic risk score (PRS) frameworks were derived from European cohorts with limited transferability across ancestries. This study examined the feasibility of classifying genetic risk profiles for iron deficiency anemia in South Asian individuals using simulated PRS and support vector machine (SVM) classification. Genetic data from 489 individuals across five South Asian subpopulations were integrated with iron-related GWAS summary statistics, and each individual was assigned a PRS representing cumulative genetic predisposition. Disease labels were simulated through a Liability Threshold Model (LTM) that assigned case and control status probabilistically. Two SVM classifiers were compared: one using the aggregated PRS as input, and one using individual genetic markers under Leave-One-Chromosome-Out (LOCO) cross-validation to prevent data leakage. The PRS-based classifier achieved AUC-ROC=0.701 and recall=0.727, while the marker-based classifier produced near-random performance with AUC-ROC=0.509. The AUC gap of 0.192 between configurations was the primary finding, demonstrating that PRS aggregation is a necessary preprocessing step where sparse individual markers carry insufficient discriminative signal at this sample size. The study contributed a reproducible, South Asian-focused pipeline built from publicly available data, extensible to real clinical data as genomic resources for this population expand.

Keywords: Polygenic Risk Scores, Support Vector Machine, Iron Deficiency Anemia, South Asian Populations, Liability Threshold Model, Genetic Risk Classification.

I. Introduction

Among the most prevalent nutritional disorders in the world, iron deficiency anemia (IDA) remains a critical concern for women of reproductive age, carrying outsized consequences in resource-limited settings. Estimates from the World Health Organization indicate that approximately 30 percent of women in this demographic group worldwide are affected by anemia (World Health Organization, 2025a), with pregnant women facing an even steeper burden. Beyond individual suffering, IDA contributes to diminished productivity, adverse maternal and perinatal outcomes, and long-term setbacks in human capital development. The scale of the problem is particularly pronounced across South Asia, a region estimated to house around 244 million anemic women (World Health Organization, 2025b). Research comparing hematological profiles across ethnic groups has consistently found that South Asian women tend to carry lower iron reserves and exhibit higher anemia prevalence than women of European or East Asian descent (Kang et al., 2021). While dietary inadequacy and chronic blood loss are recognized as the foremost drivers of iron depletion—and thus IDA (World Health Organization, 2025a)—the inherited genetic factors that modulate individual susceptibility to clinical iron deficiency remain poorly characterized.

The genetic architecture of IDA has received comparatively little systematic investigation. Existing genomic studies have largely concentrated on quantitative iron-related biomarkers—such as circulating ferritin, serum iron, or transferrin saturation—or on classical hereditary iron overload syndromes rather than on IDA as a discrete clinical outcome. Large-scale genome-wide association studies (GWAS) of iron biomarkers have catalogued more than 100 associated genomic loci (Moksnes et al., 2022), yet these investigations disproportionately drew on European participant cohorts and did not explicitly target IDA diagnosis or focus on women during their reproductive years. Although specific loci—including those near TMPRSS6, HAMP, and TFRC—are known to regulate iron physiology, how the corresponding variants translate into altered IDA risk across ethnically diverse populations remains an open question (Moksnes et al., 2022). Furthermore, the GWAS Catalog contains relatively sparse entries linking genetic variants to anemia or iron deficiency, and South Asian ancestry groups are substantially underrepresented, creating a meaningful knowledge gap around population-tailored genetic predictors.

Polygenic risk scores (PRS) represent a promising methodological bridge for addressing this limitation. Rather than relying on single large-effect variants, PRS synthesize the cumulative influence of numerous genetic markers into one continuous index of inherited liability, enabling risk differentiation even when individual SNP effects are modest. Pairing PRS with machine learning approaches such as Support Vector Machines (SVM) further extends their utility by providing flexible, computationally tractable classifiers suitable for genomic data structures. Because large-scale IDA phenotype data for South Asian populations are not publicly accessible, the present study adopted a simulation-based framework leveraging established genetic resources as a methodologically sound and reproducible proof-of-concept.

This investigation examined whether genetic risk profiles for IDA could be feasibly classified within South Asian populations by combining publicly available GWAS summary statistics from UK Biobank with genotype data from the 1000 Genomes Project Phase 3 South Asian subpopulations. PRS were derived for 489 individuals, and binary phenotype labels were generated via a probabilistic Liability Threshold Model (LTM). Two SVM configurations were constructed and evaluated: one treating the aggregated PRS as its sole feature, and a second employing top-ranked SNPs filtered by p-value ($p < 1\times10-4$) within a Leave-One-Chromosome-Out (LOCO) cross-validation scheme designed to avoid circularity between label generation and classifier assessment. By focusing on a population historically underserved by genomic research, this work contributes a replicable computational pipeline that can be extended to empirical phenotype data as they become available. For broader public health purposes, it illustrates the potential of genetics-informed risk stratification as a complement to nutritional and clinical approaches, and advances equity goals in genomic science while supporting the WHO ambition of halving global anemia prevalence by 2030 (World Health Organization, 2025a).

II. Literature review/Study site

2.1 Polygenic Risk Scores in Complex Disease Classification

Polygenic risk scores condense the additive contributions of thousands of SNPs—each individually modest—into a single metric capturing an individual's inherited propensity for a complex trait (Choi et al., 2020). These scores are typically derived from GWAS summary statistics through procedures that account for linkage disequilibrium, either via clumping and thresholding heuristics or through Bayesian shrinkage approaches such as LDpred2 and PRS-CS, which more precisely model effect size attenuation (Zhao et al., 2024). In conditions such as type 2 diabetes, coronary heart disease, and various psychiatric disorders, PRS have demonstrated meaningful predictive value even when any single associated variant is insufficient for clinical inference (Slunecka et al., 2021). A fundamental limitation, however, is the poor generalizability of PRS constructed primarily from European GWAS data when applied to non-European individuals; reduced accuracy in these populations reflects differences in allele frequency distributions, linkage disequilibrium block structure, and the representation of causal variants (Sun et al., 2024). This portability shortfall is especially consequential for conditions such as IDA, where the genomic resource base for non-European populations—South Asians in particular—is notably thin (Moksnes et al., 2022), underscoring the necessity of ancestry-specific PRS development.

2.2 Support Vector Machines in Genomic Risk Classification

Support vector machines are discriminative classifiers that determine class boundaries by locating the hyperplane that maximizes the separation between training instances closest to the decision surface, a property that makes them particularly resilient in high-dimensional settings typical of genomic analyses (Sigala et al., 2024). Their application to genetic risk stratification has spanned a range of input representations, including SNP matrices, gene expression profiles, and polygenic summary scores, with kernel-based variants capable of capturing non-linear structure in the feature space often proving competitive with or superior to linear alternatives (Muneeb et al., 2022). Vivian-Griffiths et al. (2018) illustrated this capacity in the context of schizophrenia, where kernel SVMs applied to polygenic feature sets achieved classification performance on par with logistic regression while showing greater tolerance for collinear predictors. Despite these advantages, genomic applications of SVMs must contend with the curse of dimensionality: when the number of SNP features vastly exceeds sample size, overfitting becomes a serious risk, necessitating dimensionality reduction via PRS aggregation or p-value-based SNP filtering (Shi et al., 2016). Class imbalance—endemic to case-control genetic study designs—presents an additional complication, generally requiring balanced weighting schemes or resampling procedures to prevent classifiers from collapsing to the majority class (Saito & Rehmsmeier, 2015).

2.3 Liability Threshold Models for Simulated Phenotype Generation

When empirically observed phenotype data are unavailable, the liability threshold model offers a theoretically grounded mechanism for generating synthetic binary outcomes from continuous polygenic liability distributions

(So et al., 2011). Within this framework, each individual's unobserved liability is modeled as a normally distributed quantity, a portion of whose variance is attributable to additive SNP heritability (h^2); assignment to case status is determined by whether an individual's liability surpasses a threshold calibrated to the assumed population prevalence of the condition (Wray et al., 2010). This simulation strategy has become an established tool in genomic methodology research, permitting rigorous benchmarking of PRS algorithms and classification pipelines in the absence of accessible real phenotype data (Uffelmann et al., 2023). So et al. (2011) confirmed that LTM-derived labels preserve biologically coherent gradients in case rates across PRS strata, validating the approach for evaluating risk stratification pipelines. A recognized caveat is that simulated labels are inherently circular with respect to the PRS from which liability scores are drawn; addressing this requires independent evaluation designs—such as SNP-level classifiers validated through LOCO cross-validation—to furnish a benchmark not confounded by this circularity (Uffelmann et al., 2023).

III. Materials and Methods/ Methodology

3.1 Data Sources

GWAS summary statistics were obtained from the IEU OpenGWAS Project (dataset ID: UKB-e-280_CSA), extracting rsIDs, chromosomal positions, effect alleles, beta coefficients, standard errors, p-values, and allele frequencies in VCF format via bcftools v1.19. Reference genotype data were retrieved from the 1000 Genomes Project Phase 3 via the IGSR portal in VCF format, covering five South Asian subpopulations: GIH (n = 103), STU (n = 102), ITU (n = 102), PJL (n = 96), and BEB (n = 86), totaling 489 individuals. All datasets were aligned to GRCh37/hg19.

3.2 Data Preprocessing and Quality Control

GWAS variants were excluded for missing rsIDs, missing beta coefficients or standard errors, ambiguous strand orientations (A/T or C/G), non-autosomal chromosomes, and MAF < 0.01; a p-value upper bound of $p \leq 0.5$ was retained, yielding 4,016,015 variants from an initial 9,810,691. Genotype QC was performed using PLINK2 v2.0 with filters --mind 0.05, --geno 0.01, --maf 0.01, and --hwe 1e-6 midp, retaining 8,882,943 variants across 489 samples (229 female, 260 male) (Purcell et al., 2007).

3.3 Polygenic Risk Score Computation

PRS were computed using PRSice-2 v2.3.5 following the clumping and thresholding approach of Choi and O'Reilly (2019), with LD clumping at $r^2 < 0.1$ within a 250 kb window and an inclusion threshold of Pt = 0.5, yielding 228,909 independent SNPs. Each individual's score was the weighted sum of effect allele dosages multiplied by GWAS beta coefficients, subsequently z-score normalized to mean = 0 and SD = 1.

3.4 Phenotype Simulation

As real IDA phenotype data were unavailable for the 1000 Genomes cohort, binary case/control labels were simulated using the probabilistic Liability Threshold Model described by Wray et al. (2010) and Privé et al. (2019), with target prevalence = 10%, SNP heritability $h^2 = 0.15$ (Moksnes et al., 2022), liability threshold T = 1.2816, and random seed = 42. Case probabilities were derived from each individual's normalized PRS z-score and labels were drawn via Bernoulli sampling, producing 53 cases (10.8%) and 436 controls (89.2%).

3.5 Classification Model Development

Two SVM configurations were developed in scikit-learn v1.6.1, following the binary classification framework applied in Vivian-Griffiths et al. (2018) and Gola et al. (2020), with class_weight = 'balanced' and StandardScaler normalization applied throughout. Configuration 1 used the normalized PRS z-score as a single input feature, with an 80/20 stratified train-test split and GridSearchCV optimizing AUC-ROC over linear and RBF kernels, $C \in \{0.01, 0.1, 1, 10, 100\}$, and $\gamma \in \{\text{scale}, \text{auto}, 0.01, 0.1\}$ via 5-fold stratified cross-validation within the training set only. Configuration 2 used 793 SNPs selected at $p < 1\times10^{-4}$ from the GWAS summary statistics, extracted via PLINK2 v2.0, and evaluated under Leave-One-Chromosome-Out cross-validation across 21 autosomes to prevent circularity between simulated labels and classifier features, following the approach recommended by Uffelmann et al. (2023).

3.6 Model Evaluation

Both configurations were assessed using accuracy, precision, recall, F1-score, and AUC-ROC (Sokolova & Lapalme, 2009). For Configuration 1, 95% confidence intervals were computed via bootstrap resampling (n = 10,000, random_state = 42) on the test set, with majority-class baseline accuracy reported as a reference. Configuration 2 metrics were aggregated across all 21 LOCO folds. All analyses were executed in Python 3.12.13 on Google Colab with random seed = 42, using pandas v2.2.1, NumPy v1.26.4, matplotlib v3.8.2, and seaborn v0.13.2.

IV. Results and Discussion

4.1 Data Preprocessing and Simulation Validation

GWAS quality control reduced the raw dataset from 9,810,691 to 4,016,015 variants (40.9% retention), and genotype QC retained 8,882,943 variants across all 489 South Asian samples. PRS computation via PRSice-2 v2.3.5 yielded a z-score normalized distribution with mean = 0 and SD = 1.0, spanning [−3.83, 3.63] units, consistent with expected polygenic additive architecture.

To validate the simulation framework prior to classification, individuals were stratified into PRS tertiles and case rates were compared across groups shown Table 1. Case rates increased from 4.3% in the lowest tertile to 8.0% in the middle and 20.2% in the highest, representing a 4.7-fold gradient. This monotonic dose-response pattern confirmed that the probabilistic LTM correctly encoded cumulative genetic liability into the simulated labels, establishing the biological plausibility of the phenotype assignments before classification results are interpreted.

Table 1: Case rates across PRS tertiles (low, mid, high) among 489 simulated South Asian individuals, showing a 4.7-fold gradient from 4.3% in the lowest tertile to 20.2% in the highest.

PRS	n	Cases	Controls	Case Rate
Low PRS (bottom 33%)	163	7	156	4.3%
Mid PRS (middle 33%)	163	13	150	8.0%
High PRS (top 33%)	163	33	130	20.2%

4.2 Configuration 1: PRS-Based SVM

GridSearchCV selected a linear kernel with C = 0.01 as the optimal configuration, with an internal CV AUC of 0.742 during hyperparameter search. On the held-out test set (n = 98), the model achieved AUC-ROC = 0.701 (95% CI: 0.580–0.810) and recall = 0.727, both exceeding the pre-specified targets of ≥ 0.60 and ≥ 0.60–0.65 respectively. Accuracy was 0.663 (95% CI: 0.571–0.755), falling below the majority-class baseline of 0.888, an expected consequence of class_weight = 'balanced' trading overall accuracy for minority-class sensitivity. Full metrics are reported in Table 2, and the ROC curve is shown in Figure.

The confusion matrix (Figure 2) showed TN = 57, FP = 30, FN = 3, TP = 8. Only three high-risk cases were missed, while 30 low-risk individuals were over-flagged, consistent with a screening-oriented classifier prioritizing recall under severe class imbalance. Precision (0.211) and F1-score (0.327) fell below their targets as a direct consequence of the 10:1 case-to-control ratio; at 10.8% prevalence, high recall necessarily produces a high false positive rate that suppresses precision, consistent with expectations for PRS-based classifiers under class imbalance (Lewis & Vassos, 2020). The 5-fold CV AUC of 0.726 closely tracked the test set AUC of 0.701, indicating stable generalization across data partitions. It is acknowledged that the PRS used as classifier input is derived from the same score used to simulate labels, a circularity inherent to the simulation design that bounds the generalizability of Config 1 results; Config 2 was designed specifically to address this.

4.3 Configuration 2: SNP-Based SVM with LOCO Cross-Validation

Under LOCO cross-validation across 21 autosomes, the SNP-based classifier achieved an aggregated AUC-ROC of 0.509 (95% CI: 0.429–0.591), with per-chromosome AUC values ranging from 0.470 to 0.546 (mean = 0.511, SD = 0.022). No chromosome demonstrated meaningfully above-chance discrimination. The confusion matrix (Figure 3) shows TN = 436, FP = 0, FN = 53, TP = 0, with the model assigning every individual to the low-risk class, yielding precision, recall, and F1-score of 0.000. Accuracy of 0.892 equaled the majority-class baseline and represents trivial all-control prediction rather than genuine classification.

This near-random result was the expected and informative outcome of circularity-controlled evaluation. When the aggregation step was removed and 793 sparse GWAS-ranked SNPs were evaluated independently, the discriminative signal was insufficient to classify genetic risk in a cohort of 489 individuals. The polygenic architecture of iron homeostasis distributes genetic signals broadly across the genome; concentrating features in a small set of top-ranked variants captures too little of that signal for a high-dimensional classifier to exploit at this sample size (Choi et al., 2020; Lewis & Vassos, 2020).

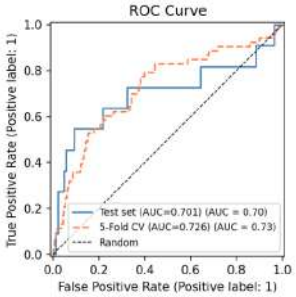

Figure 1: Receiver operating characteristic curve for Configuration 1 (PRS-based SVM) on the held-out test set (AUC = 0.701) and 5-fold cross-validation (AUC = 0.726), with random classifier baseline shown for reference.

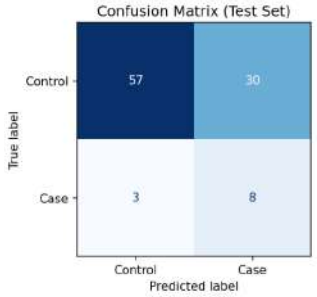

Figure 2: Confusion matrix for Configuration 1 (PRS-based SVM) on the held-out test set (n = 98), showing true negatives, false positives, false negatives, and true positives.

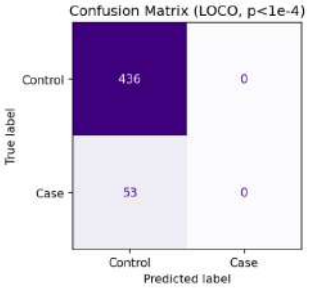

Figure 3: Confusion matrix for Configuration 2 (SNP-based SVM with LOCO cross-validation) aggregated across 21 autosomes (n = 489), reflecting all-control prediction with no minority-class detection

4.4 Comparative Analysis

Table 2 presents the head-to-head comparison between both configurations across all metrics.

Table 2: Classification performance of Configuration 1 (PRS-based SVM) and Configuration 2 (SNP-based SVM with LOCO cross-validation) against pre-specified targets.

Metric	PRS-based SVM	95% CI	SNP-based SVM
Accuracy	0.663	[0.571, 0.755]	0.892†
Precision	0.211	–	0.000
Recall	0.727	–	0.000
F1-Score	0.327	[0.150, 0.490]	0.000
AUC-ROC	0.701	[0.580, 0.810]	0.509

† Config 2 accuracy equals the majority-class baseline (0.892), reflecting trivial all-control prediction with no minority-class detection.

The AUC-ROC gap of 0.192 between configurations is the primary comparative finding of this study. It quantifies the discriminative advantage conferred by PRS aggregation, compressing 228,909 weighted SNP effects into a single score, over 793 individual sparse SNP features evaluated under proper cross-validation. This advantage is consistent with the broader literature documenting that PRS models outperform direct SNP classifiers in small-sample genetic studies due to the signal aggregation property of weighted allele summation (Lewis & Vassos, 2020; Choi et al., 2020; Gola et al., 2020). The results collectively demonstrate that PRS aggregation is a necessary preprocessing step for genetic risk classification in small-sample, ancestrally underrepresented settings, and that direct high-dimensional SNP classification is statistically underpowered at this cohort size.

These findings must be interpreted within the constraints of the simulation design. All classifications reflect simulated genetic liability derived from iron biomarker GWAS rather than clinically confirmed IDA outcomes, and performance estimates cannot be extrapolated directly to real-world diagnostic settings. The cohort of 489 individuals, while representative of South Asian genetic diversity in the 1000 Genomes Project, is insufficient for robust high-dimensional SNP classification and produces wide confidence intervals on all metrics. Population stratification within the South Asian cohort was not explicitly adjusted, which may introduce confounding into classifier performance. Notwithstanding these constraints, the pipeline produced internally consistent, interpretable results and establishes a reproducible framework that can be directly extended when real IDA phenotype data from South Asian cohorts become available.

V. Conclusion

This study demonstrated the feasibility of classifying genetic risk profiles for iron deficiency anemia in South Asian populations using simulated polygenic risk scores and support vector machines. The PRS-based classifier achieved above-chance discrimination, while the SNP-based classifier under circularity-controlled LOCO cross-validation performed at near-random levels, collapsing to all-control prediction. The AUC-ROC gap of

0.192 between configurations establishes that PRS aggregation is a necessary preprocessing step for genetic risk classification when sample sizes are small and genetic signal is distributed broadly across the genome.
The contrast between configurations carries a methodological implication beyond this study: high-dimensional SNP feature matrices are statistically underpowered for classification in cohorts of this scale, and aggregating weighted genetic effects into a continuous score is what makes discrimination tractable. This finding is consistent with the broader literature on PRS performance in underrepresented populations and reinforces the case for ancestry-specific PRS development as genomic resources for South Asian cohorts expand.
These results are a proof of concept grounded in simulated phenotypes rather than clinical IDA outcomes, and performance estimates should not be extrapolated to diagnostic settings without validation on real case-control data. Future work should prioritize replication using clinically confirmed IDA phenotypes from South Asian biobanks, larger cohort sizes to support high-dimensional classification, and South Asian-specific GWAS summary statistics to improve the ancestry-appropriateness of the underlying effect size estimates.

VI. Ethical Approval

This study used only publicly available, anonymized genomic datasets from the 1000 Genomes Project Phase 3 and the IEU OpenGWAS Project. No human participants were recruited, no biological samples were collected, and no identifiable personal data were used. Ethical approval and written informed consent were not required.

VII. Data Availability

The GWAS summary statistics used in this study are publicly available from the IEU OpenGWAS Project (dataset ID: UKB-e-280_CSA) at https://gwas.mrcieu.ac.uk. Reference genotype data are available from the 1000 Genomes Project Phase 3 via the International Genome Sample Resource at https://www.internationalgenome.org. Analysis code is available from the authors upon reasonable request.

Grant /Funding: This research received no external funding.

References

1. Choi, S. W., Mak, T. S. H., & O'Reilly, P. F. (2020). Tutorial: A guide to performing polygenic risk score analyses. Nature Protocols, 15(9), 2759–2772.
2. Choi, S. W., & O'Reilly, P. F. (2019). PRSice-2: Polygenic risk score software for biobank-scale data. GigaScience, 8(7), giz082.
3. Gola, D., Erdmann, J., Muller-Myhsok, B., Schunkert, H., & Konig, I. R. (2020). Polygenic risk scores outperform machine learning methods in predicting coronary artery disease status. Genetic Epidemiology, 44(2), 125–138.
4. Kang, W., et al. (2021). Ethnic differences in iron status. Advances in Nutrition, 12(5), 1838–1853.
5. Lewis, C. M., & Vassos, E. (2020). Polygenic risk scores: From research tools to clinical instruments. Genome Medicine, 12(1), 44.
6. Moksnes, M. R., et al. (2022). Genome-wide meta-analysis of iron status biomarkers. Communications Biology, 5, 591.
7. Muneeb, M., Feng, S. F., & Henschel, A. (2022). Can we convert genotype sequences into images for cases/controls classification? Frontiers in Bioinformatics, 2, 914435.
8. Privé, F., Arbel, J., & Vilhjalmsson, B. J. (2019). LDpred2: Better, faster, stronger. bioRxiv. https://doi.org/10.1101/2020.04.28.066720
9. Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A., Bender, D., & Sham, P. C. (2007). PLINK: A tool set for whole-genome association and population-based linkage analyses. American Journal of Human Genetics, 81(3), 559–575.
10. Saito, T., & Rehmsmeier, M. (2015). The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. PLOS ONE, 10(3), e0118432.
11. Shi, J., et al. (2016). Winner's curse correction and variable thresholding improve performance of polygenic risk modeling based on genome-wide association study summary-level data. PLOS Genetics, 12(12), e1006493.
12. Sigala, R. E., et al. (2024). Machine learning to advance human genome-wide association studies. Genes, 15(1), 34.
13. Slunecka, J. L., et al. (2021). Implementation and implications for polygenic risk scores in healthcare. Human Genomics, 15, 46.
14. So, H.-C., Kwan, J. S. H., Cherny, S. S., & Sham, P. C. (2011). Risk prediction of complex diseases from family history and known susceptibility loci, with applications for cancer screening. American Journal of Human Genetics, 88(5), 548–565.
15. Sokolova, M., & Lapalme, G. (2009). A systematic analysis of performance measures for classification tasks. Information Processing and Management, 45(4), 427–437.
16. Sun, Q., et al. (2024). Improving polygenic risk prediction in admixed populations by explicitly modeling ancestral-differential effects via GAUDI. Nature Communications, 15, 1016.

17. Uffelmann, E., Posthuma, D., & Peyrot, W. J. (2023). Genome-wide association studies of polygenic risk score-derived phenotypes may lead to inflated false positive rates. Scientific Reports, 13, 4219.
18. Vivian-Griffiths, T., et al. (2018). Predictive modeling of schizophrenia from genomic data: Comparison of polygenic risk score with kernel support vector machines approach. American Journal of Medical Genetics. Part B: Neuropsychiatric Genetics, 180(1), 80–85.
19. World Health Organization. (2025a). Anaemia. WHO Fact Sheet, February 2025.
20. World Health Organization. (2025b). Anaemia in women and children. WHO Global Health Observatory, 2025.
21. Wray, N. R., et al. (2010). The genetic interpretation of area under the ROC curve in genomic profiling. PLOS Genetics, 6(2), e1000864.
22. Zhao, Z., et al. (2024). Optimizing and benchmarking polygenic risk scores with GWAS summary statistics. Genome Biology, 25(1), 260.