Multimodal Prediction of Primary Open-Angle Glaucoma Using Polygenic Risk Scores and Clinical Features in a High-Risk African Ancestry Cohort

Yan Zhu[1], Aude Benigne Ikuzwe Sindikubwabo[2], Yuki Bradford[3], Rebecca Salowe[1], Lannawill Caruth[2], Kenneth Pham[1], Moksha Laxmi[1], Vrathasha Vrathasha[1], Marine-Ayan Ibrahim Aibo[1], Roy Lee[1], Mina Halimitabrizi[1], Isabel Di Rosa[1], Leila Ghaffari[1], Jie He[1], Marylyn D. Ritchie[3], Joan M. O'Brien[1]*, Shefali Setia-Verma[2]*

[1]Center for Genetics of Complex Disease, Department of Ophthalmology, University of Pennsylvania, Philadelphia, PA 19104, USA
[2]Department of Pathology and Laboratory Medicine, University of Pennsylvania, Philadelphia, PA 19104, USA
[3]Department of Genetics, University of Pennsylvania, Philadelphia, PA 19104, USA
[5]Present address: 225 N 38th Street, Philadelphia, PA 19104, USA
*Lead contact
*Corresponding Author: joan.o'brien@pennmedicine.upenn.edu, Shefali.SetiaVerma@pennmedicine.upenn.edu

SUMMARY

Primary open-angle glaucoma (POAG) disproportionately affects individuals of African ancestry, yet early detection tools remain limited. Using the largest African ancestry POAG cohort (N = 7,352), we developed machine learning models that integrate clinical features with four ancestry-matched polygenic risk scores (PRS): two genome-wide (PRS-CS POAAGG, PRS-CS MEGA) and two curated (MTAG PRS, MEGA PRS). Models trained on a clean cohort (N = 196) and tested in 1,013 suspects showed improved prediction with PRS, especially curated scores. MEGA PRS-based predictions correlated with IOP, CDR, and RNFL, validating biological relevance. Combining PRS with inter-eye asymmetry features further enhanced performance. These results support curated PRS such as MEGA PRS and MTAG PRS as interpretable, equitable tools for early glaucoma risk stratification.

KEYWORDS

primary open-angle glaucoma, machine learning, multimodal modeling, polygenic risk score, genetic risk score, phenotype integration, African American cohort, risk prediction, ocular phenotypes, precision medicine

INTRODUCTION

Glaucoma is one of the leading causes of irreversible blindness worldwide, affecting over 70 million individuals and projected to rise dramatically with population aging and urbanization trends.[1,2] Primary open-angle glaucoma (POAG), the most prevalent form of glaucoma, accounts for approximately 74% of all diagnosed cases.[3] POAG is characterized by progressive degeneration of the optic nerve, often accompanied by elevated intraocular pressure (IOP), thinning of the retinal nerve fiber layer (RNFL) and structural changes in the optic disc such as increased cup-to-disc (CDR).[4-6] Importantly, individuals of African ancestry are disproportionately affected, experiencing earlier onset, more rapid progression, and up to five-fold higher risk of blindness compared to those of European ancestry.[7,8] Despite these known disparities, early detection and risk stratification in this high-risk population remain inadequate, underscoring an urgent need for precision tools that can support earlier and more accurate diagnosis.[9]

In recent years, advances in genomics have enabled the development of polygenic risk scores (PRS), which aggregate the effects of multiple genetic variants to estimate an individual's inherited susceptibility to complex diseases like POAG.[10-12] These tools have shown promise in various conditions, including type 2 diabetes, [13,14] coronary heart disease [15,16] and breast cancer,[17-19] and are beginning to gain traction in ophthalmology.[20,21] Most existing PRS are trained in European-centric cohorts and often lack transferability to other ancestries.[22,23] By comparison, curated PRS constructed from biologically validated, ancestry-matched loci offer a more interpretable and potentially more reliable approach for

glaucoma risk prediction.[24] Moreover, while genetic scores alone often lack diagnostic resolution, clinical phenotypes such as IOP, RNFL, and CDR—though routinely measured—capture disease only after structural or functional damage has begun.[25] Thus, a compelling opportunity lies in the integrative use of both genetic and phenotypic data to develop more accurate and equitable risk prediction models for glaucoma, especially in populations with historically limited representation in genomic studies.[22]

Artificial intelligence, particularly supervised machine learning (ML), offers a powerful framework to harness the complexity of multidimensional health data.[26,27] In contrast to traditional regression-based risk models, machine learning algorithms can capture non-linear interactions among diverse inputs and automatically learn relevant feature representations.[28,29] In ophthalmology, machine learning techniques have been successfully employed to analyze fundus photographs, segment optic nerve head structures, and quantify RNFL thickness using OCT scans.[30] These models can detect glaucoma-related changes that are often imperceptible to the human eye, providing a critical advantage in screening and disease monitoring.[31] Previous studies have successfully applied machine learning to ophthalmic imaging modalities—including segmentation of optic nerve head structures from fundus images and automated quantification of RNFL thickness from OCT scans.[32,33] Yet, relatively few studies have examined the utility of integrating genetic risk—particularly PRS—with ocular phenotypes in large, ancestrally diverse POAG cohorts.[34,35] Moreover, no prior work has systematically compared multiple ML classifiers across different combinations of phenotypic and genetic features in a rigorously curated African American cohort. This gap highlights the need to evaluate whether ancestry-matched genetic scores, especially curated PRS, can provide added predictive value beyond standard phenotypes.

In this study, we leverage the Primary Open-Angle African American Glaucoma Genetics (POAAGG) cohort—the largest and most deeply phenotyped African ancestry glaucoma dataset—to develop and evaluate machine learning models for POAG risk prediction.[8,36,37] We systematically apply three widely used classifiers— multilayer perceptron (MLP),[38] random forest (RF),[39-41] and support vector machine (SVM) [42-44]— to a clinically "clean" cohort (N=196) defined by stringent phenotype cutoffs. The larger full cohort (N=7352) was used to perform genome-wide association studies (GWAS) and generate four PRS tailored to African ancestry individuals. These include two genome-wide PRS derived using PRS-CS (PRS-CS POAAGG and PRS-CS MEGA), and two interpretable, curated PRS constructed from selected high-confidence loci (MTAG PRS and MEGA PRS). We hypothesize that curated PRS, derived from ancestry-matched and biologically informed GWAS loci, can serve as robust and interpretable predictors of POAG risk in African ancestry individuals. When integrated into machine learning models with demographic and phenotypic features, these PRS will enhance predictive accuracy across classifiers (RF, SVM, MLP), with the strongest effects observed in clean, phenotype-defined cohorts—highlighting their potential for early screening and precision risk stratification. Our results demonstrate that curated PRS (MEGA PRS and MTAG PRS) consistently outperform genome-wide PRS in clean cohorts, suggesting complementary roles for each genetic strategy in advancing precision glaucoma care for individuals of African ancestry.

RESULTS

Study Design and Data Integration Framework

To develop a robust and interpretable machine learning framework for glaucoma risk prediction, we designed a structured pipeline that integrates phenotypic filtering, genomic risk scores, and model interpretability through feature importance and clinical validation (Figure 1A). Starting from a large-scale genotyped cohort (N=7,352), we applied stringent phenotype-based criteria to curate a biologically extreme clean cohort (N=196) used for model training, and an independent test cohort (N=1,013)—composed entirely of suspects—for model evaluation and enrichment analysis.

We generated four PRS: two genome-wide scores using PRS-CS (PRS-CS POAAGG and PRS-CS MEGA), and two curated PRS based on selected, high-confidence glaucoma-associated loci (MTAG PRS and MEGA PRS). All scores were calculated using ancestry-matched GWAS summary statistics and applied uniformly to both training and test sets.

Three machine learning models—RF, MLP, and SVM—were trained across five different feature sets and evaluated using standard metrics (AUC, accuracy, sensitivity, specificity, and F1-score). Feature

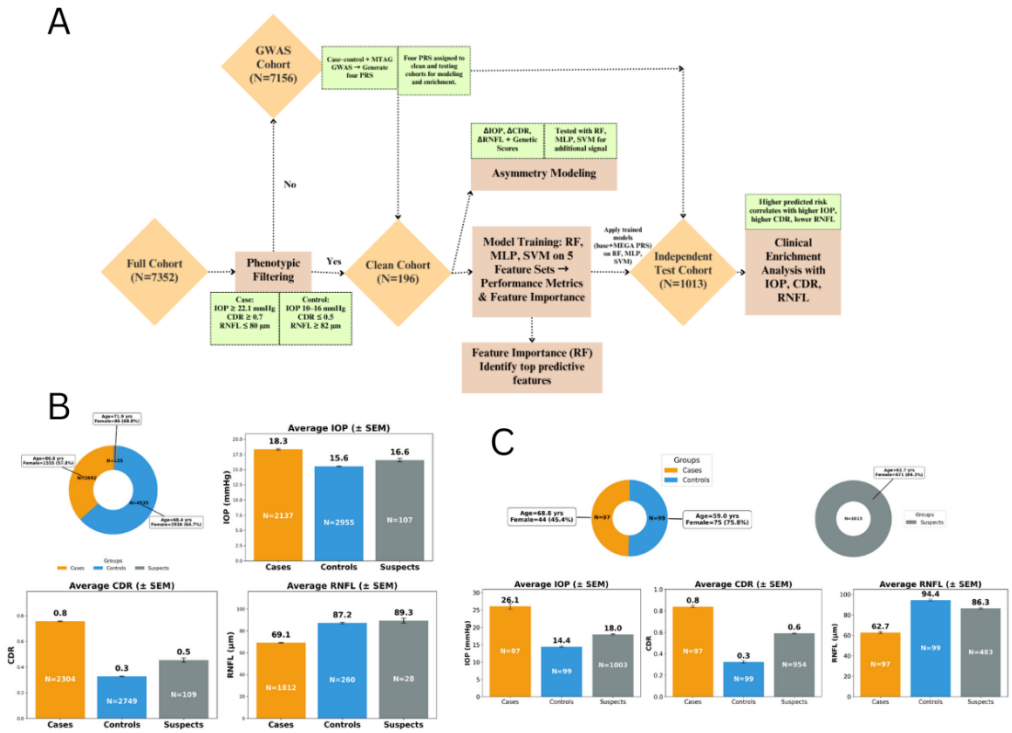
importance was extracted using RF, and predicted risk scores were correlated with clinical parameters (IOP, CDR, RNFL) in the test set.

The full cohort consisted of 7,352 genotyped individuals enrolled in the POAAGG study, of which 2,692 were POAG cases, 4,535 were controls, and 125 were suspects (Figure 1B). The average age was 80.8 years in cases, 68.4 years in controls, and 71.9 years in suspects. Females accounted for 57.8%, 64.7%, and 68.8% of these subgroups, respectively. Clinical averages showed distinct group separations: cases had significantly higher intraocular pressure (IOP = 18.3 mmHg), larger cup-to-disc ratios (CDR = 0.8), and thinner retinal nerve fiber layers (RNFL = 69.1 µm), while controls showed lower IOP (15.6 mmHg), CDR (0.3), and thicker RNFL (87.2 µm). These phenotype-based differences justified our use of threshold-based filtering to define a clean training subset and a biologically relevant suspect test cohort.

For model training, we curated a clean cohort of 196 individuals by applying stringent phenotype thresholds to isolate biologically extreme cases and definitively healthy controls (Figure 1A and 1C). POAG cases were selected if they met at least one of the following in either eye: IOP ≥ 22.1 mmHg, CDR ≥ 0.7, or RNFL ≤ 80 µm. Controls were required to meet all of the following: IOP between 10–16 mmHg, CDR ≤ 0.5, and RNFL ≥ 82 µm. The final training set included 97 cases and 99 controls. This clean cohort exhibited a pronounced phenotype gap: average IOP of 26.1 mmHg in cases vs. 14.4 mmHg in controls, CDR of 0.8 vs. 0.3, and RNFL of 62.7 µm vs. 94.4 µm. This separation helps mitigate label leakage and ensures the biological validity of the training labels. Gender imbalance was also evident (female: 45.4% in cases vs. 75.8% in controls), which was retained in the modeling to reflect real-world bias.

To assess the generalizability of our trained models, we applied them to an independent test cohort composed of 1,013 suspect individuals—defined as subjects who did not meet criteria for definite POAG or healthy control but showed clinical ambiguity (Figure 1C). This cohort represents a biologically heterogeneous group with early or subclinical disease, making it an ideal external validation set for enrichment analysis. The average IOP was 18.0 mmHg, CDR was 0.6, and RNFL was 86.3 µm—intermediate between clean cases and controls. Notably, the suspect cohort was enriched for female participants (66.2%) with a mean age of 62.7 years. By comparing predicted risk scores against continuous phenotype severity, we further demonstrated that higher predicted glaucoma risk correlates with higher IOP, larger CDR, and thinner RNFL—confirming the biological relevance of the trained models.

Figure 1. Overview of Study Design, Cohorts, and Phenotypic Characteristics



(A) Study Design. Schematic illustration of the study workflow. The full cohort (N=7352) was phenotypically filtered using stringent thresholds for IOP, CDR, and retinal nerve fiber layer thickness (RNFL) to generate the clean training cohort (N=196). Case-control status was determined using biologically extreme definitions (e.g., IOP ≥ 22.1 mmHg, CDR ≥ 0.7, RNFL ≤ 80 µm for cases; IOP 10–16 mmHg, CDR ≤ 0.5, RNFL ≥ 82 µm for controls). Four PRS were computed and applied to the clean cohort and testing cohort. Three models (RF, MLP, SVM) were trained on the clean cohort and evaluated on the independent test cohort of suspects (N=1013).

(B) Full Cohort Characteristics. Demographic and phenotypic distributions for the full cohort (N=7352). Top left panel: donut chart showing case, control, and suspect proportions with age and sex distribution annotated. Top right and bottom panels: mean ± SEM of IOP, CDR, and RNFL across groups, with sample size (N) indicated inside each bar.

(C) Training and Testing Cohort Characteristics. Demographic and phenotypic distributions for the clean training cohort (N=196) and the independent suspect-only test cohort (N=1013). Top left panel: case/control composition of the clean cohort with age and sex annotated. Top right panel: suspect cohort distribution with demographics. Bottom: comparison of IOP, CDR, and RNFL across case, control, and suspect groups, with mean ± SEM and available sample counts shown.

PRS-Based Prediction Models Reveal Limited but Interpretable Signals

To evaluate the predictive utility of genetic features alone, we constructed and benchmarked four PRS developed from African ancestry GWAS data. These include two genome-wide PRS computed using PRS-CS (PRS-CS POAAGG and PRS-CS MEGA) and two curated PRS derived from selected, high-confidence loci (MTAG PRS and MEGA PRS). Each score was designed to capture complementary aspects of glaucoma genetic architecture—balancing genome-wide signal breadth with locus-level interpretability—and was evaluated in both the clean training cohort (N = 196) and the independent suspect testing cohort (N = 1,013).[12,14,45 46]

PRS-CS POAAGG and PRS-CS MEGA were generated using a Bayesian polygenic modeling framework that integrates genome-wide SNP effect sizes with African ancestry linkage disequilibrium reference panels. MTAG PRS and MEGA PRS were built using distinct sets of curated variants: the MTAG PRS

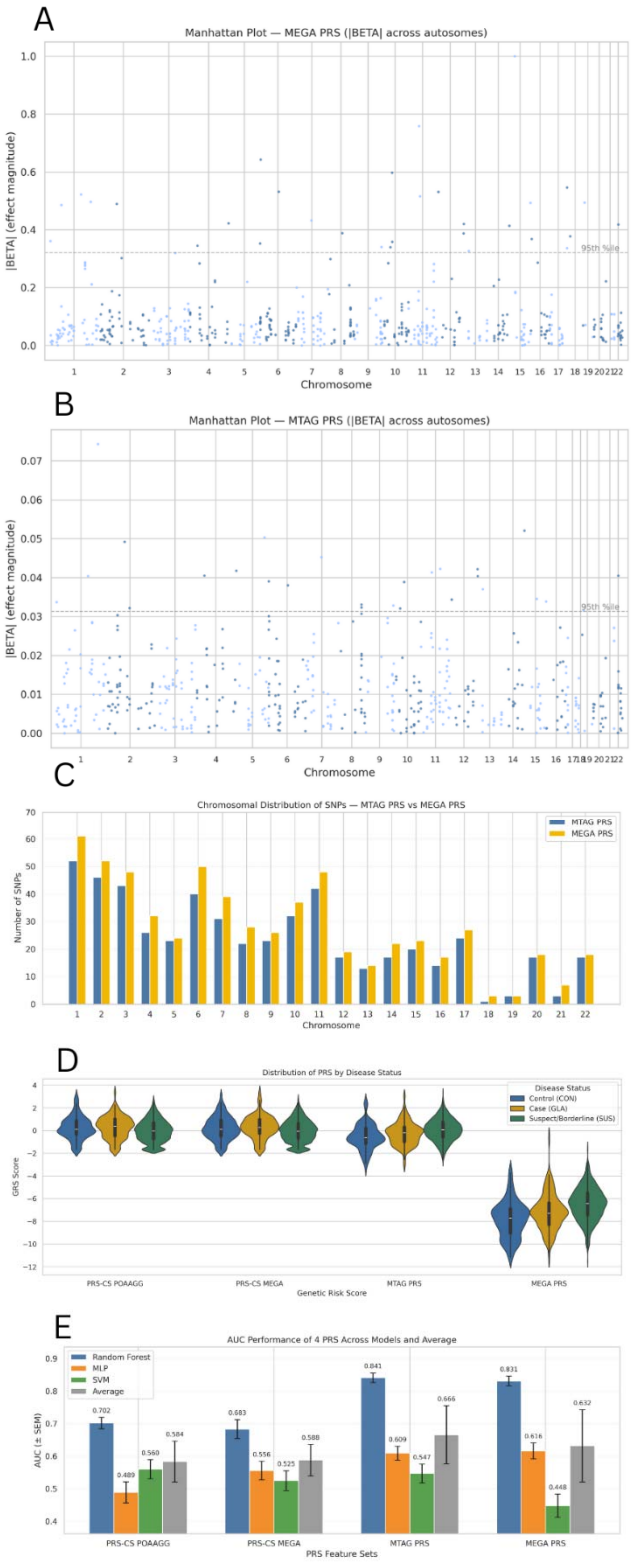
incorporated 526 unique loci from multi-trait quantitative GWAS emphasizing IOP and endophenotypic traits, while the MEGA PRS included 616 loci from the MEGA case-control GWAS focused on POAG status.[11] All variants were mapped to hg38 and scored using ancestry-specific effect weights (see STAR Methods).[47]

We visualized the structure and genomic distribution of these four PRS (Figure 2). Panels A and B show Manhattan-style plots illustrating the variant effect magnitudes across chromosomes for MEGA PRS and MTAG PRS, respectively. The chromosomal distribution plot (Figure 2C) highlights the broader genomic coverage of MEGA PRS compared to MTAG PRS. Violin plots (Figure 2D) reveal a consistent pattern across all PRS: higher scores among POAG cases, lower scores among controls, and intermediate values among suspects—supporting the biological relevance of each PRS in capturing subclinical risk.

When used as standalone features, all four PRS demonstrated moderate but interpretable predictive performance in the clean cohort (Figure 2E). Average AUC values across the three classifiers (RF, MLP, SVM) showed a consistent ranking among the PRS. MTAG PRS achieved the highest mean AUC (0.666 ± 0.067), followed by MEGA PRS (0.632 ± 0.092), PRS-CS POAAGG (0.584 ± 0.048), and PRS-CS MEGA (0.588 ± 0.063). Although the absolute values were moderate, the relative consistency across models indicates that each PRS captures reproducible genetic signal.

Together, these results suggest that while genome-wide PRS capture broad heritable risk, curated PRS such as MEGA PRS and MTAG PRS provide clearer separation between affected and unaffected individuals. However, the overall discrimination remains moderate—highlighting that genetic data alone only partially explains POAG risk. Their value is best realized when integrated with clinical and demographic features, as explored in the following sections.

Figure 2. Performance and Characteristics of Four Polygenic Risk Scores (PRS) in African Ancestry Glaucoma Cohorts


A
Manhattan Plot — MEGA PRS (|BETA| across autosomes)
|BETA| (effect magnitude)
95th %ile
Chromosome
B
Manhattan Plot — MTAG PRS (|BETA| across autosomes)
|BETA| (effect magnitude)
95th %ile
Chromosome
C
Chromosomal Distribution of SNPs — MTAG PRS vs MEGA PRS
MTAG PRS
MEGA PRS
Number of SNPs
Chromosome
D
Distribution of PRS by Disease Status
Disease Status
Control (CON)
Case (GLA)
Suspect/Borderline (SUS)
GRS Score
PRS-CS POAAGG
PRS-CS MEGA
MTAG PRS
MEGA PRS
Genetic Risk Score
E
AUC Performance of 4 PRS Across Models and Average
Random Forest
MLP
SVM
Average
AUC (± SEM)
0.702
0.489
0.560
0.584
0.683
0.556
0.525
0.588
0.841
0.609
0.547
0.666
0.831
0.616
0.448
0.632
PRS-CS POAAGG
PRS-CS MEGA
MTAG PRS
MEGA PRS
PRS Feature Sets

(A–B) Manhattan-style plots of absolute effect sizes (|BETA|) across autosomes for the MEGA PRS (A) and MTAG PRS (B), showing variant distribution and strength. SNPs included in each PRS are plotted by chromosomal location with dotted lines indicating the 90th percentile of effect magnitude.
(C) Chromosomal distribution of SNPs across the MEGA PRS (yellow) and MTAG PRS (blue), highlighting broad but distinct genomic coverage of each curated PRS.
(D) Violin plots showing standardized distributions of all four PRS (PRS-CS POAAGG, PRS-CS MEGA, MTAG PRS, MEGA PRS) across clean cases (orange), clean controls (blue), and suspect individuals (green). Suspects consistently fall between cases and controls, validating biological plausibility.
(E) Predictive performance of each PRS across three machine learning classifiers (RF, SVM, MLP) in the clean cohort (N = 196). Bars show mean AUC ± SEM, with gray bars representing the average across all three classifiers.

Clean Cohort Analysis Reveals Robust and Consistent Gains from Polygenic Risk Integration

In the clean training cohort (97 POAG cases, 99 controls), we evaluated the predictive performance of three classifiers—RF, SVM, and MLP—using feature sets that incrementally incorporated polygenic risk information. Each classifier was first trained on a baseline model consisting of age and gender, and then retrained with one of four finalized PRS: PRS-CS POAAGG, PRS-CS MEGA, MTAG PRS, and MEGA PRS.

As shown in Figure 3A–C, adding any PRS to the base model improved AUC in all three classifiers. For RF (Figure 3B), AUC increased from 0.81 (base) to 0.91 when incorporating either MTAG PRS or MEGA PRS, with slightly lower gains observed for PRS-CS scores. SVM (Figure 3A) and MLP (Figure 3C) showed similar improvement trends, with performance gains ranging from +0.04 to +0.16 depending on the PRS used. These consistent improvements across models demonstrate that PRS—especially those based on curated loci—provide reliable predictive value even in a phenotype-limited setting.

Although all four PRS were derived from African ancestry GWAS data, they reflect different genetic strategies: PRS-CS POAAGG and PRS-CS MEGA aggregate genome-wide polygenic signal, while MTAG PRS and MEGA PRS are based on selected loci with higher effect sizes and biological relevance. This diversity in design likely explains the complementary performance patterns, with curated PRS (especially MEGA PRS) showing stronger and more stable gains across classifiers.

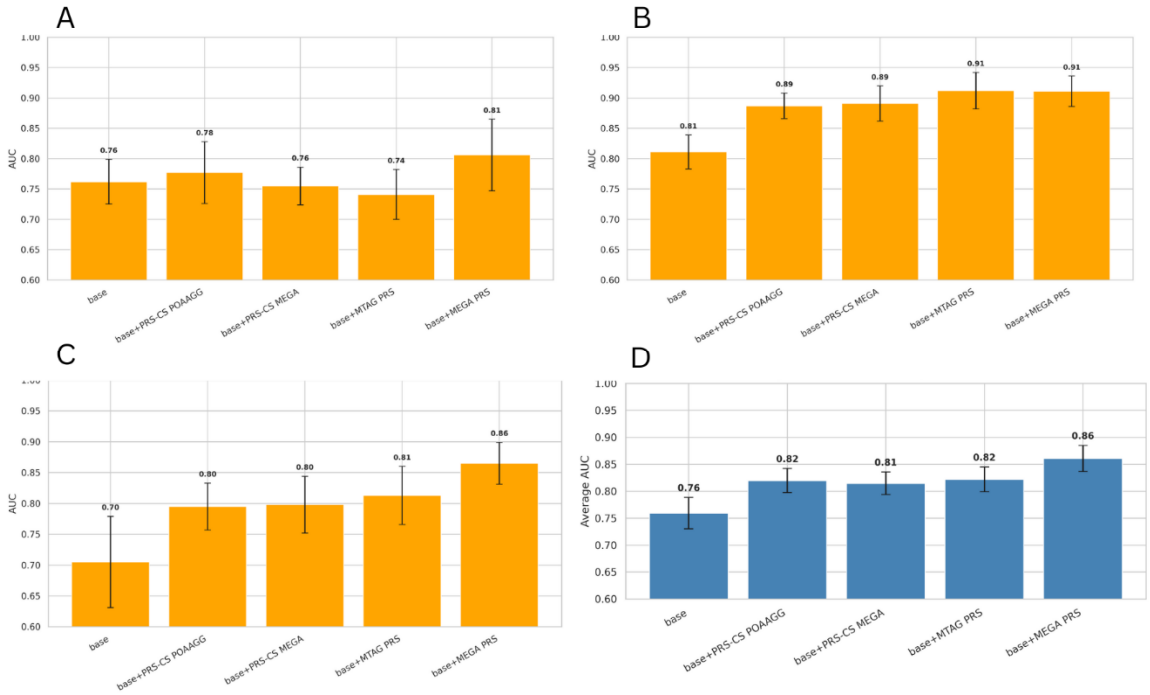
To evaluate model-agnostic trends, we computed the average AUC across RF, MLP, and SVM for each PRS-enhanced feature set (Figure 3D). The MEGA PRS achieved the highest mean AUC (0.86), followed closely by MTAG PRS (0.82). PRS-CS POAAGG and PRS-CS MEGA also showed improvement over the baseline model (0.76), achieving average AUCs of 0.82 and 0.81, respectively. These results suggest that both genome-wide and curated PRS provide additive value, with curated scores showing slightly stronger alignment with case-control labels in this clean African ancestry cohort.

These gains are particularly notable given the limited sample size and absence of direct disease biomarkers in the models. The consistency across classifiers also reinforces the robustness of PRS as generalizable predictors, rather than artifacts of any specific algorithm.

From a methodological perspective, RF and MLP achieved the highest peak AUCs, while SVM tended to underperform slightly, especially with MEGA PRS. This trend may reflect differences in how each model handles small, nonlinear datasets: RF excels in low-sample settings with strong feature importance, whereas MLP benefits from richer input spaces. Still, all three classifiers improved meaningfully with PRS inclusion, underscoring their collective utility.

In summary, integrating PRS into simple demographic models meaningfully improves POAG risk prediction in a well-defined African ancestry cohort. These results support the use of polygenic scores for early-stage screening and justify further validation in larger or less curated populations—such as the suspect cohort examined later in the study.

Figure 3. Model performance across feature sets in the clean cohort (N = 196).



(A) SVM models trained on the base feature set (age + gender) and versions that include each of the four PRS. Adding any PRS improved performance, with AUC increasing from 0.76 (base) to 0.78–0.81 across PRS-enhanced models.
(B) RF models showed the largest gains, rising from 0.81 (base) to 0.89–0.91 after adding PRS, indicating strong and consistent effects across all genetic inputs.
(C) MLP models also improved with PRS integration, from 0.70 (base) to 0.80–0.86, demonstrating that genetic information adds predictive value even in smaller training sets.
(D) Average AUC across all three classifiers (RF, SVM, MLP) summarizes overall model performance. The MEGA PRS achieved the highest average AUC (0.86), followed by MTAG PRS (0.82), PRS-CS POAAGG (0.82), and PRS-CS MEGA (0.81). These results confirm that adding PRS—especially curated scores like MEGA PRS—consistently boosts glaucoma risk prediction in the clean cohort.

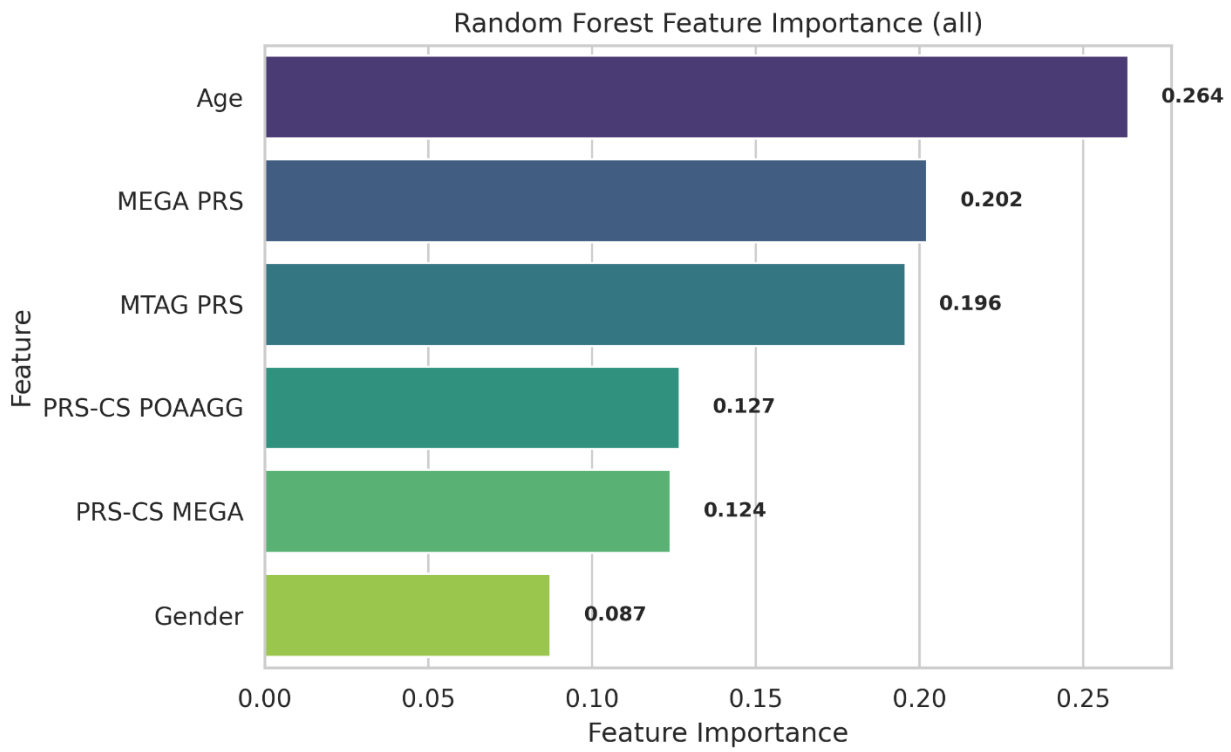
Feature Importance Analysis Highlights the Complementary Role of Genetic Scores

To understand which features contributed most to prediction accuracy, we analyzed feature importance from the Random Forest model trained on the full set of variables in the clean cohort (Figure 4). As expected, age remained the top predictor, with the highest importance score (0.264), consistent with its known role as a major risk factor for glaucoma.[48]

Genetic scores also ranked highly, reinforcing their value beyond demographics. Among the four PRS, MEGA PRS and MTAG PRS showed the strongest contributions, with importance scores of 0.202 and 0.196, respectively. These curated scores outperformed both genome-wide PRS, suggesting that carefully selected variants may offer more targeted risk information. PRS-CS POAAGG (0.127) and PRS-CS MEGA (0.124) also contributed meaningful signal, though at slightly lower levels. This indicates that broad polygenic risk still adds value, especially when combined with curated inputs. Gender contributed the least (0.087), which aligns with prior findings that sex differences, while present, are often overshadowed by age and genetic risk in POAG models.

Taken together, these results show that genetic features—especially curated PRS like MEGA and MTAG—can rival traditional clinical predictors in importance. Their consistent contribution supports their use as interpretable and transferable biomarkers. The ranking also illustrates how Random Forest models adaptively balance overlapping predictors, assigning weight based on unique information contributed by each PRS.

Figure 4. Feature Importance Rankings in Random Forest Models Using Clean Cohort



Feature importance values were derived from a Random Forest model trained on the full feature set (age, gender, and all four PRS). Age remained the strongest predictor (importance□=□0.264), followed by MEGA□PRS (0.202) and MTAG□PRS (0.196). PRS-CS□POAAGG (0.127) and PRS-CS□MEGA (0.124) also contributed meaningful signal, while gender had the smallest impact (0.087). These results highlight that genetic features—especially curated PRS such as MEGA□PRS and MTAG□PRS—carry strong and complementary predictive value alongside traditional demographic factors.

Testing Cohort Analysis: Clinical Enrichment Validates Biological Signal in Suspect Population

To assess the biological plausibility and real-world utility of our trained models, we applied the MLP, RF, and SVM classifiers—originally trained on the clean 196-case/control cohort—to an independent test cohort of 1,013 POAAGG glaucoma suspects. These individuals do not have definitive diagnostic labels but have comprehensive phenotyping (IOP, CDR, RNFL), enabling us to evaluate predictions via a clinical enrichment framework. For this analysis, we focused on the base+MEGA PRS feature set, which achieved the highest AUCs across all three models in the clean cohort, thereby providing the most robust and generalizable predictor. Each of the three classifiers was applied using this feature set to generate individual risk scores for the suspect cohort. Specifically, we tested whether higher predicted POAG risk scores correlated with known clinical indicators of glaucomatous damage: elevated IOP, enlarged vertical CDR, and thinning of the RNFL. Because no binary labels are available, traditional classification metrics (e.g., AUC) cannot be computed; instead, we assessed Pearson correlations between predicted risk (binned to deciles) and the three quantitative phenotypes.

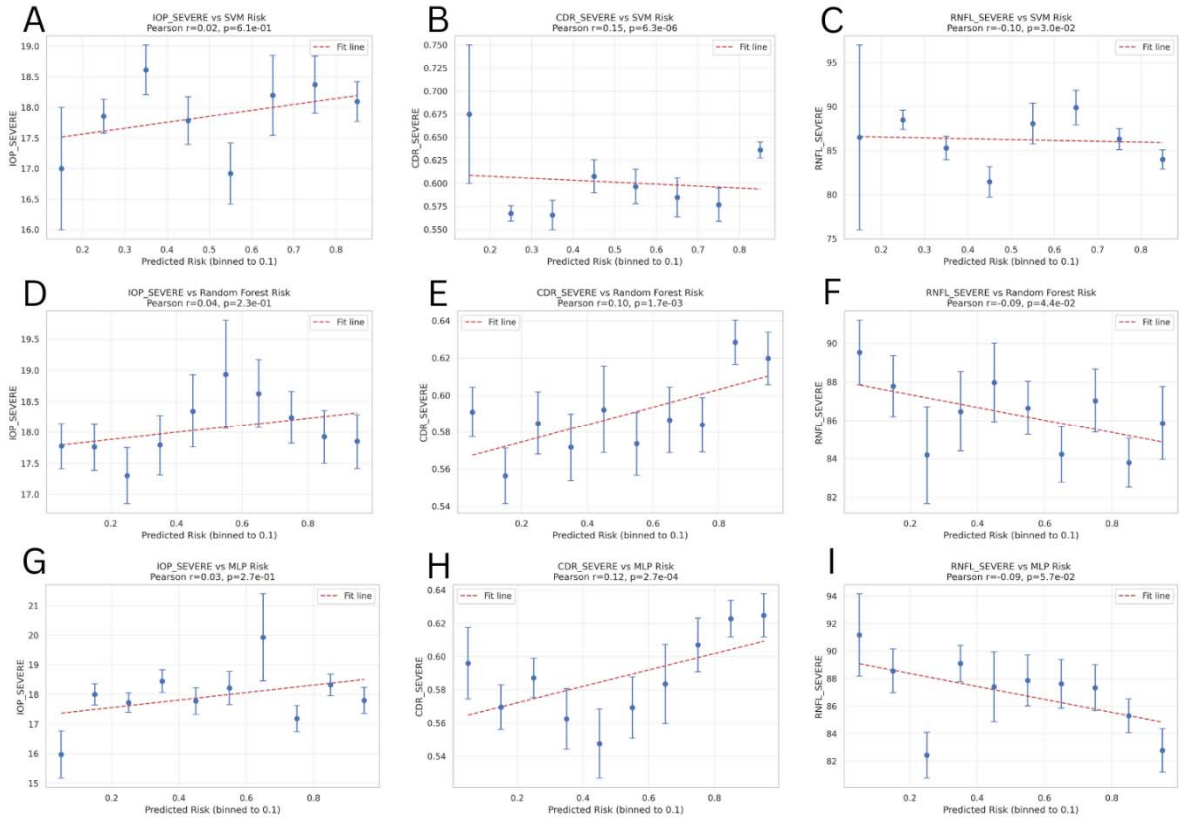
Across all three models, predicted risk was consistently associated with disease severity, particularly for IOP and RNFL. The SVM model (Figure 5A–C) showed a significant correlation between predicted risk and CDR ($r = 0.15$, $p = 6.3\times10^{-}$□), and modest associations with IOP ($r = 0.02$, $p = 0.61$) and RNFL thickness ($r = -0.10$, $p = 0.03$). These results suggest that higher risk scores from the SVM model were associated with increased IOP level and thinner RNFL, even though the CDR signal was weaker.

The Random Forest model (Figure 5D–F) demonstrated the strongest and most consistent enrichment across clinical measures. Predicted risk was positively correlated with CDR ($r = 0.10$, $p = 1.7\times10^{-3}$) and IOP ($r = 0.04$, $p = 0.23$), and negatively correlated with RNFL ($r = -0.09$, $p = 0.044$). Although the IOP correlation was not statistically significant, the trend was in the expected direction. These results suggest that RF predictions effectively captured biologically meaningful differences in disease severity.

The MLP model (Figure 5G–I) also showed encouraging trends. CDR again had the strongest correlation ($r = 0.12$, $p = 2.7\times10^{-}$□), with more modest, non-significant associations observed for IOP ($r = 0.03$, $p = 0.27$) and RNFL ($r = -0.09$, $p = 0.057$). While not all associations reached statistical significance, the directional consistency across all three models reinforces the biological validity of the MEGA PRS-informed predictions.

Overall, this enrichment analysis provides compelling evidence that models trained on a clean case–control cohort can generalize to an unlabeled, real-world population. The consistent correlations between predicted risk and structural markers of glaucoma—particularly CDR and RNFL—indicate that the risk scores are not arbitrary, but instead reflect underlying disease processes. This supports the use of PRS-based models, and the MEGA PRS in particular, as potential tools for early risk stratification in high-risk populations that may not yet meet diagnostic thresholds for POAG.

Figure 5. Predicted glaucoma risk scores from the MEGA PRS model correlate with clinical severity in the suspect cohort.



Risk scores were generated using models trained on the clean cohort (base + MEGA PRS) and applied to 1,013 glaucoma suspects. Panels A–C show results from the SVM model, panels D–F from the Random Forest model, and panels G–I from the MLP model. Each panel plots predicted risk (binned by decile) against a key glaucoma phenotype: IOP, CDR, and RNFL thickness. Predicted risk scores showed consistent positive correlations with CDR (B, E, H) and IOP (A, D, G), and inverse correlations with RNFL (C, F, I). The strongest enrichment was observed for CDR for RF and MLP. These results validate the biological relevance of PRS-informed predictions and highlight the potential of MEGA PRS models for early risk stratification in clinically ambiguous populations.

Phenotypic Asymmetry as a Clinically Meaningful Marker, Enhanced by Genetic Risk Scores

In glaucoma diagnosis and risk stratification, one of the earliest red flags clinicians look for is asymmetry between the two eyes. It is well recognized that in the early stages of POAG, one eye may begin to deteriorate while the fellow eye remains within normal limits. This asymmetry can manifest in IOP, CDR, or RNFL thinning—often before the disease becomes symmetric or clinically definitive. Importantly, using absolute IOP, CDR, or RNFL values directly in our models would introduce circularity, since these same thresholds were applied to define the clean case/control cohort. To avoid this issue, we instead derived

inter-eye differences (ΔIOP, ΔCDR, and ΔRNFL). These asymmetry features are statistically independent of the clean cohort selection criteria and carry strong clinical meaning, as they reflect the earliest, eye-specific manifestations of glaucomatous damage observed in practice.

When we used asymmetry features alone, ΔIOP emerged as the strongest single predictor, achieving an AUC of 0.85–0.92 across all three models (Figure 6A–C). This aligns with clinical intuition—an unexplained pressure difference between eyes often prompts closer monitoring. ΔCDR and ΔRNFL also showed predictive value, though with slightly lower performance when used alone.

What made the results even more compelling was the impact of adding PRS to these asymmetry features. Across nearly all combinations, performance improved. For instance, in the Random Forest model (Figure 6B), AUC increased from 0.92 with ΔIOP alone to 0.95–0.97 when combined with any PRS. ΔCDR improved from 0.83 to as high as 0.90 when paired with PRS. Even ΔRNFL, the weakest standalone asymmetry feature, jumped from 0.78 to 0.93 with MEGA PRS included.

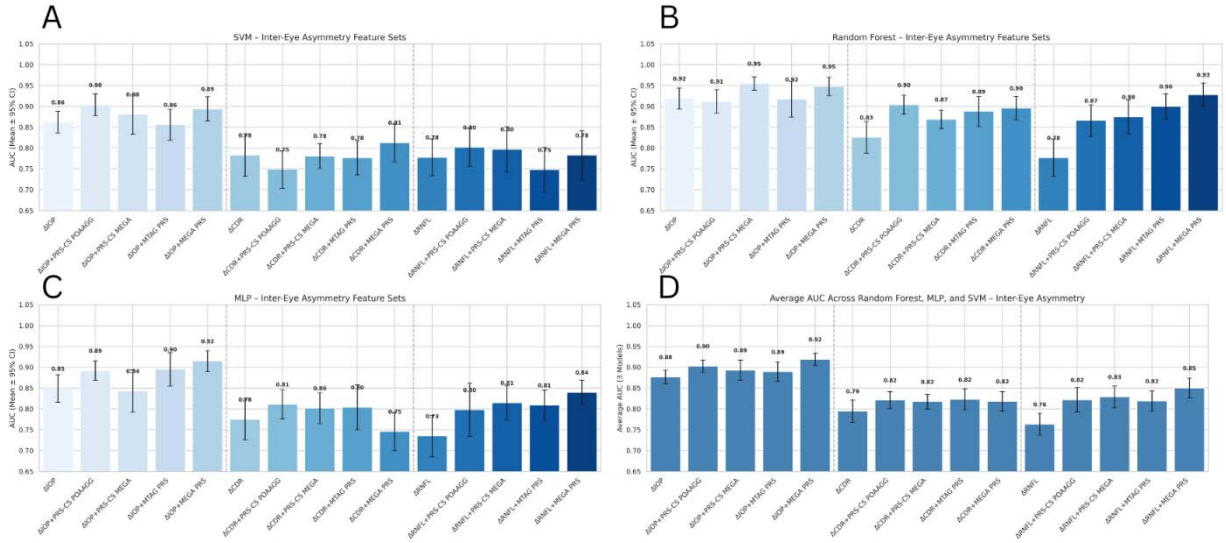
This pattern held across models. To compare overall trends, we calculated the average AUC across all three models (Figure 6D). ΔIOP + MEGA PRS delivered the highest average AUC (0.92), followed closely by ΔIOP with other PRS. Improvements were also notable for ΔCDR and ΔRNFL when combined with any of the four PRS.

These results suggest that PRS are especially powerful when layered onto subtle clinical features. Asymmetry features alone capture early signals of disease, but pairing them with genetic risk greatly boosts prediction—especially in patients who may not yet meet diagnostic criteria.

From a clinical perspective, this could help flag borderline patients who might otherwise be missed or dismissed in routine screenings. This is especially valuable in populations with limited access to advanced imaging or long-term follow-up. The ability to combine simple measurements like inter-eye difference with genetic data opens the door to more proactive, personalized glaucoma care.

In summary, phenotypic asymmetry and genetic risk work better together. Their integration produced some of the highest predictive performance in our study and represents a promising strategy for identifying at-risk individuals early—before irreversible damage occurs.

Figure 6. Inter-Eye Asymmetry Enhances Genetic Risk Modeling Across Modalities



(A) SVM models trained on inter-eye asymmetry features (ΔIOP, ΔCDR, ΔRNFL) showed improved performance when combined with PRS. The best-performing SVM model used ΔIOP + MEGA PRS, achieving an AUC of 0.89 (up from 0.86 with ΔIOP alone). Gains for ΔCDR and ΔRNFL were more modest, with AUCs increasing from 0.78 to 0.81 and 0.78 to 0.80, respectively.
(B) RF models achieved the highest overall performance. ΔIOP alone yielded an AUC of 0.92, which increased slightly to 0.95 with genetic score added. ΔCDR models improved from 0.83 to 0.90, while ΔRNFL models rose from 0.78 to 0.93 when paired with MEGA PRS.
(C) MLP models also benefited from PRS integration. AUC improved from 0.85 (ΔIOP alone) to 0.92 with MEGA PRS. ΔCDR models improved from 0.78 to 0.81, and ΔRNFL from 0.73 to 0.84 with MEGA PRS.
(D) Average AUC across all models showed that ΔIOP + MEGA PRS achieved the top overall performance (AUC = 0.92), followed by ΔIOP with other PRS. PRS also boosted prediction for ΔCDR and ΔRNFL features across the board. These results suggest that combining subtle inter-eye asymmetry with genetic risk scores—especially MEGA PRS—can improve early glaucoma risk detection across modeling approaches.

DISCUSSION

First, our results underscore the robust predictive utility of the four finalized polygenic risk scores (PRS-CS POAAGG, PRS-CS MEGA, MTAG PRS, and MEGA PRS), which consistently improved performance across all three machine learning models. In the clean cohort of 196 stringently selected cases and controls, the base model—comprising only age and gender—already achieved a surprisingly high mean AUC of 0.76. This performance can be attributed to two key factors: (1) the rigorous selection criteria used to define the clean cohort, which amplified phenotypic separation and reduced noise, and (2) the well-established role of age as a dominant risk factor for POAG. When genetic features were added, all four PRS enhanced model performance, with curated scores such as MEGA PRS and MTAG PRS showing the largest gains. Together, MEGA PRS and MTAG PRS captured complementary aspects of glaucoma genetics—MEGA PRS derived from African-ancestry case-control GWAS of POAG (AUC=0.63), and MTAG PRS from quantitative, IOP-based meta-analysis—linking disease susceptibility with underlying endophenotypes (AUC=0.67). Furthermore, when all four PRS were integrated (base + PRS-CS POAAGG, PRS-CS MEGA, MTAG PRS, and MEGA PRS), the average AUC increased to 0.86 on average, outperforming the base model and demonstrating consistent predictive value across classifiers. These results highlight that combining both genome-wide and curated PRS can offer a more complete view of genetic risk in African-ancestry populations. Overall, these findings provide compelling evidence that ancestry-matched PRS—especially curated scores like MEGA PRS and MTAG PRS—are not only statistically robust but also clinically meaningful, supporting their potential role in early glaucoma screening and precision risk assessment.

Second, this work highlights the importance of genetic equity and the need to develop population-specific tools for precision medicine. A key challenge in PRS development is the limited transferability of models trained in European ancestry cohorts to individuals of non-European descent. This discrepancy is especially acute for individuals of African ancestry, who carry greater genomic diversity and often exhibit distinct linkage disequilibrium structures that are not captured by Eurocentric GWAS. In our study, we found that PRS models only performed well when they were built using African ancestry data. For example, the genome-wide scores—PRS-CS POAAGG and PRS-CS MEGA—performed reasonably well, but only because they were generated using GWAS data from African ancestry cohorts. If built from European datasets, their accuracy would likely have been much lower. Even more notably, the curated scores—MTAG PRS and MEGA PRS—which were constructed from African-specific glaucoma loci, delivered the strongest and most consistent results. This highlights the urgent need for diverse representation in genetic studies and the development of ancestry-aware risk models. Our findings are not merely technical—they have real implications for reducing health disparities in glaucoma care, which disproportionately affects African American populations.

Third, our comparison of RF, SVM, and MLP models illustrates the balance between predictive power and interpretability in small-cohort applications. All three models demonstrated strong, stable performance using relatively simple feature sets, making them well-suited for scenarios where data is limited. Notably, RF offered a distinct advantage by enabling feature importance analysis—revealing age, MTAG PRS and MEGA PRS as consistently top-ranked predictors—which enhances transparency and supports clinical trust and regulatory adoption. These findings suggest that model selection for clinical translation should not only consider accuracy but also the need for interpretability, scalability, and robustness in real-world settings.

Fourth, we demonstrate that models trained exclusively on the clean cohort retain clinical validity when applied to a real-world test cohort of 1,013 POAAGG glaucoma suspects. Applying our best-performing models—RF, SVM, and MLP trained on the base + MEGA PRS feature set—we generated individual risk scores for each suspect and validated these scores against known clinical indicators (IOP, CDR, RNFL) using a clinical enrichment approach. Despite the absence of case/control labels, predicted risk showed strong and consistent correlations with disease-relevant phenotypes: individuals in the top decile of

predicted risk had higher IOP, bigger CDR, and thinner RNFL than those in lower deciles. This demonstrates that our models capture meaningful subclinical variation and can stratify individuals along a biologically plausible disease continuum. These results have important clinical implications. In real-world settings—especially in underserved or resource-limited environments—many patients fall into a diagnostic gray zone and remain untreated despite showing early signs of damage. Our framework offers a scalable, genetics-informed risk stratification approach that could flag such individuals for closer monitoring, earlier intervention, or targeted genetic counseling. Importantly, this strategy does not rely on costly imaging or expert annotations, making it adaptable to a range of care settings. The ability to move from model predictions to actionable clinical triage represents a critical step toward operationalizing AI tools in preventive glaucoma care.

Fifth, we emphasize the biological rationale and predictive power of inter-eye asymmetry as a glaucoma-relevant feature. Glaucoma often manifests asymmetrically in its early stages—one eye begins to show signs of damage while the other appears normal. Our clean cohort selection relied on strict thresholds for absolute IOP, CDR, and RNFL values, which could inadvertently introduce circularity if the same features were reused in modeling. To address this, we introduced inter-eye differences (ΔIOP, ΔCDR, ΔRNFL) as independent and biologically meaningful features. Across all three AI models, these asymmetry metrics provided strong predictive signals, with ΔIOP consistently yielding the highest AUCs. The improvement was most pronounced when genetic risk scores were added, suggesting that phenotypic asymmetry and inherited susceptibility act as complementary early disease signals. This approach not only avoids feature leakage from the cohort definition process but also aligns closely with real-world clinical heuristics used by glaucoma specialists.

Sixth, our findings offer a clinically grounded framework for glaucoma risk prediction that is both adaptable and powerful across real-world scenarios. In phenotype-rich situations, where data such as IOP, CDR, or RNFL are readily available, integrating curated PRS with these features consistently produced additive and sometimes super-additive gains across models. This synergy reflects how genetic predisposition and clinical expression capture distinct yet complementary aspects of glaucoma risk, allowing for sharper stratification. For instance, a patient with modest optic nerve asymmetry but a high PRS may warrant earlier referral or intensified monitoring, even if they do not yet meet conventional diagnostic thresholds. These interactions likely mirror complex biological processes, including gene–environment interplay and polygenic modifiers, that drive early disease heterogeneity. At the same time, the system remains powerful in phenotype-limited settings, which often occur in real-world practice. Many individuals—especially those in early or preclinical stages, or those lacking access to advanced imaging—cannot be stratified reliably using phenotypic markers alone. Here, blood-derived PRS offers a low-barrier, ancestry-matched genomic predictor that can forecast future risk independent of immediate clinical measures. This is particularly impactful for African ancestry populations, where underdiagnosis remains common and health disparities persist. By enabling both data-rich integration and data-scarce prediction, our framework provides a flexible pathway toward equitable clinical deployment. With prospective validation, embedding compact and interpretable PRS into electronic health records and genetic screening platforms could support scalable, cost-effective, and personalized glaucoma care in diverse healthcare environments.

In conclusion, our study demonstrates that integrating curated PRS with ML models offers a powerful and equitable framework for early glaucoma risk prediction, particularly in underrepresented African ancestry populations. Through rigorous evaluation across clean and heterogeneous cohorts, and with validation using real-world suspect data, we establish both the scientific validity and translational potential of our approach. These findings lay the groundwork for population-specific, AI-enabled precision screening tools that can be deployed in diverse clinical settings to reduce diagnostic delay and promote health equity in glaucoma care.

Limitations of the study

Despite the promising performance of our machine learning models and the rigorously curated nature of the POAAGG cohort, several limitations should be acknowledged. First, our study focused exclusively on

African American individuals from a single high-risk glaucoma cohort. While this enabled ancestry-matched PRS development and promoted equity in precision medicine, it limits generalizability to other populations. Given known differences in genetic architecture, PRS derived here may not transfer well to non-African ancestries. Future efforts should include multi-ethnic, multi-center cohorts to assess the portability and equity of genetic risk prediction.

Second, although we used cleanly defined cases and controls to train models and validated them in a large suspect cohort, the binary classification framework may oversimplify POAG's continuum. Suspect individuals—used for enrichment analysis—may represent transitional or early-stage disease. Future models should incorporate probabilistic or multi-class labeling to better reflect real-world diagnostic uncertainty and support early detection.

Third, while we integrated key clinical features such as age, IOP, CDR, and RNFL, other glaucoma-relevant phenotypes (e.g., visual field loss, optic disc morphology, or corneal hysteresis) were excluded due to data sparsity or inconsistency. Their absence may limit the comprehensiveness of clinical modeling, especially for atypical presentations or advanced diseases.

Fourth, the integration of complex machine learning models such as MLP and random forest enhances predictive performance but introduces interpretability challenges that may hinder clinical translation. While our random forest model provided feature importance rankings, other models like MLP operate as black boxes, limiting their appeal in settings where decision transparency is critical.[49,50] Hybrid approaches that incorporate interpretable algorithms (e.g., logistic regression or decision trees) or techniques such as SHAP[51] or LIME[52] may help bridge this gap.[53] In addition, although we demonstrated consistent trends across full and clean cohorts, external validation using truly independent, prospective datasets remains necessary to assess robustness, generalizability, and potential confounders related to data collection protocols, healthcare access, or socioeconomic bias.

Finally, our PRS were built from fixed SNP sets derived from African ancestry GWAS, including MEGA PRS and MTAG PRS, which performed strongly. However, we did not model fine-grained SNP weighting or interactions, and some included loci lack functional annotation. Future work should explore dynamic or pathway-informed PRS construction, integrate functional genomics, and validate biological mechanisms to enhance interpretability and clinical actionability.

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources, data, and code should be directed to and will be fulfilled by the lead contact, Joan M. O'Brien (joan.o'brien@pennmedicine.upenn.edu).

Materials availability

This study did not generate new unique reagents.

Data and code availability

- The de-identified genotype, phenotype, and PRS data used in this study are derived from the POAAGG study. Due to participant privacy concerns and institutional regulations, these data cannot be deposited in a public repository. However, de-identified datasets used for model training and evaluation are available from the lead contact upon request and pending data use agreement.
- Summary statistics and curated SNP lists used to compute PRS are available in the supplemental table.
- All original code for data processing, feature engineering, and model training (RF, MLP, SVM) has been deposited at GitHub and is available by request.

ACKNOWLEDGMENTS

This study was supported by the National Eye Institute, National Institutes of Health, through a Research Project Grant (R01EY023557) and a Vision Research Core Grant (P30EY001583). Additional funding was provided by the F.M. Kirby Foundation, Research to Prevent Blindness, the University of Pennsylvania Hospital Board of Women Visitors, and the Paul and Evanina Bell Mackall Foundation Trust. Institutional support was received from the Department of Ophthalmology at the Perelman School of Medicine and the VA Hospital in Philadelphia, Pennsylvania.

The funding sources had no role in the design and conduct of the study; collection, management, analysis, and interpretation of the data; preparation, review, or approval of the manuscript; or the decision to submit the manuscript for publication.

AUTHOR CONTRIBUTIONS

Conceptualization: Yan Zhu, Shefali Setia-Verma, Joan M. O'Brien
Methodology: Yan Zhu, Aude Benigne Ikuzwe, Yuki Bradford, Lannawill Caruth
Software and Formal Analysis: Yan Zhu, Aude Benigne Ikuzwe, Yuki Bradford, Lannawill Caruth
Investigation: Yan Zhu, Rebecca Salowe, Moksha Laxmi, Vrathasha Vrathasha, Kenneth Pham, Marine-Ayan Ibrahim Aibo, Roy Lee, Mina Halimitabrizi, Isabel Di Rosa, Leila Ghaffari, Jie He, Colleen McDermott
Resources: Marylyn D. Ritchie, Shefali Setia-Verma, Joan M. O'Brien
Data Curation: Yan Zhu, Aude Benigne Ikuzwe
Writing – Original Draft: Yan Zhu, Rebecca Salowe
Writing – Review & Editing: Moksha Laxmi, Vrathasha Vrathasha, Kenneth Pham, Marine-Ayan Ibrahim Aibo, Roy Lee, Mina Halimitabrizi, Isabel Di Rosa, Leila Ghaffari, Jie He, Colleen McDermott
Supervision: Marylyn D. Ritchie, Shefali Setia-Verma, Joan M. O'Brien
Funding Acquisition: Joan M. O'Brien

All authors have read and agreed to the published version of the manuscript.

DECLARATION OF INTERESTS

The authors declare no competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

SUPPLEMENTAL INFORMATION

Will be available when paper published.

STAR METHODS

KEY RESOURCES TABLE

REAGENT OR RESOURCE	SOURCE	IDENTIFIER
Deposited data		
POAAGG study cohort		Controlled Access
MEGA cohort summary statistics		Controlled Access
RGC genotype imputed data		Controlled Access
Software and algorithms		
Python 3.11	Python Software Foundation	RRID:SCR_008394
R 4.2.2	https://www.r-project.org	
Scikit-learn v1.3	scikit-learn developers	RRID:SCR_002577
TensorFlow 2.13 (Keras API)	Google / Meta	RRID:SCR_016345 / RRID:SCR_018194
Matplotlib / Seaborn	Python	N/A
PRS-CS	https://github.com/getian107/PRScs	N/A
PLINK v1.9 / v2.0	https://www.cog-genomics.org/plink/	RRID:SCR_001757
Other		
Google Colab environment	https://colab.research.google.com	
GitHub repository (to be deposited)		

EXPERIMENTAL MODEL AND STUDY PARTICIPANT DETAILS

Human Study Participants

This study was conducted using human data from the Primary Open-Angle African American Glaucoma Genetics (POAAGG) study, an ongoing, IRB-approved initiative based at the University of Pennsylvania. The POAAGG study aims to identify genetic, phenotypic, and environmental risk factors for POAG in individuals of African ancestry.

Participants were recruited from ophthalmology clinics affiliated with the University of Pennsylvania Health System and the Philadelphia Veterans Affairs Medical Center. All individuals self-identified as African American or of African ancestry and provided written informed consent prior to participation. The study was approved by the Institutional Review Board at the University of Pennsylvania and was conducted in accordance with the tenets of the Declaration of Helsinki.

For the current analysis, a total of 7,352 participants with high-quality genomic and clinical data were included. This comprised:

- a clean case/control cohort (N = 196), selected using stringent thresholds on IOP, CDR, and RNFL thickness to define biologically extreme POAG cases and healthy controls;
- a suspect cohort (N = 1,013), used to assess model generalizability and perform clinical enrichment analyses;

- and the broader full cohort (N = 7,352), used to generate four PRS from African ancestry-specific glaucoma GWAS.

Collected variables included demographic information (age, gender), comprehensive ocular phenotypes (IOP, CDR, RNFL), and genome-wide genotyping data. These data supported the construction of machine learning models and the evaluation of genetic risk prediction tools tailored to African ancestry populations.

Gender was included as a covariate in all models, although stratified analyses were not performed due to limited sample size in the clean cohort. We acknowledge that gender-specific genetic and clinical factors may influence glaucoma risk and deserve further investigation in future studies. No animal models, cell lines, or microbial samples were used.

METHOD DETAILS

1. Data Preprocessing and Cohort Selection

We utilized data from the Primary Open-Angle African American Glaucoma Genetics (POAAGG) study, the largest known genetic and phenotypic dataset for POAG in African ancestry individuals. The full analysis cohort consisted of 7,352 participants with available genotype data, demographic information (age, gender) and ocular phenotypes—including IOP, vertical CDR, and RNFL thickness.

To enable rigorous training and benchmarking, we curated a clean cohort (N = 196) comprising 97 POAG cases and 99 controls with well-defined, extreme phenotypes. POAG cases were required to meet at least one of the following in either eye: IOP ≥ 22.1 mmHg, CDR ≥ 0.7, or RNFL ≤ 80 µm. Controls were required to meet all three of the following in both eyes: IOP between 10–16 mmHg, CDR ≤ 0.5, and RNFL ≥ 82 µm. Individuals with intermediate or conflicting values were excluded to avoid diagnostic ambiguity and ensure reliable binary classification labels.

To evaluate the real-world translational validity of our models, we applied trained classifiers to a test cohort of 1,013 individuals labeled as glaucoma “suspects”, who did not meet definitive case or control criteria but had full phenotype data. These suspect cases were not included in model training. We assessed prediction performance in this group using a clinical enrichment analysis, correlating predicted risk probabilities with continuous measures of disease severity (IOP, CDR, RNFL) to infer subclinical glaucoma risk.

All individuals self-identified as African American, and genetic ancestry was further confirmed using principal component analysis. Demographic features were numerically encoded (e.g., male = 1, female = 0), and all continuous clinical traits were standardized via z-score transformation. Missing numerical values were imputed using the cohort-wise mean. Preprocessing steps—including imputation, scaling, and feature selection—were implemented within scikit-learn pipelines to ensure reproducibility during cross-validation and model evaluation.

We constructed and evaluated 5 feature sets:

- (1) base = age + gender
- (2) base + PRS-CS POAAGG
- (3) base + PRS-CS MEGA
- (4) base + MTAG PRS
- (5) base + MEGA PRS

This design enabled us to systematically quantify the individual and synergistic contributions of demographic, clinical, and genetic information in glaucoma risk prediction. The full feature definitions and modeling pipelines are available in our repository by request.

2. Polygenic Risk Score (PRS) Construction

To incorporate genetic susceptibility into our glaucoma prediction models, we computed four ancestry-matched polygenic risk scores: two curated PRS derived from high-confidence SNPs (MTAG PRS and MEGA PRS). [54,55] and two genome-wide PRS generated using PRS-CS (PRS-CS POAAGG and PRS-CS

MEGA). [8,56] These scores were based on SNP-level genotype data from African ancestry participants in the POAAGG and MEGA studies, with rigorous quality control and harmonization.[56]

The MEGA PRS was constructed using 616 SNPs identified from a binary case-control GWAS meta-analysis (MEGA) focused on POAG in African ancestry populations. Loci were selected based on genome-wide or suggestive significance ($p < 5\times10^{-}$) and biological relevance to glaucoma. The MTAG PRS included 526 SNPs drawn from a multi-trait GWAS using IOP as a quantitative endophenotype, capturing subclinical risk pathways relevant to glaucoma. Both are as reported in our previous *Cell* publication. [11] These curated PRS were manually filtered for quality, mapped to hg38 coordinates, and scored using ancestry-specific effect sizes.

Effect sizes for each SNP were extracted from the original GWAS summary statistics. Both unweighted (allele-count based) and weighted scores were computed using PLINK v1.9. For the weighted version, each SNP was scored as:

$$\mathrm{PRS}_{\mathrm{weighted}} = \sum_{i=1}^{n} \beta_i G_i$$

where β_i isthe GWAS-derived log-odds ratio or beta coefficient, and G_i is the number of risk alleles (0, 1, or 2) at SNP *i*. Scores were standardized using rank-based z-transformation prior to modeling. Genomic distributions of the curated PRS were visualized in Manhattan-style plots and chromosome-wise bar plots (see Figure 2A–C), and score distributions by phenotype group were shown using violin plots (Figure 2D).

In parallel, we generated two genome-wide PRS using the PRS-CS algorithm, a Bayesian regression framework with continuous shrinkage priors.

- PRS-CS POAAGG was derived from GWAS summary statistics of 7,500 African American participants in the internal POAAGG study.
- PRS-CS MEGA was based on the MEGA African ancestry meta-analysis (14,000 participants), which included multiple cohorts.

PRS-CS was run using the African ancestry LD reference panel from the 1000 Genomes Project.[46] Default global shrinkage parameters were used. SNP weights from each model were merged with individual-level genotype data using PLINK 2.0 (--score) and custom Python scripts. All genome-wide PRS were also standardized using z-score normalization to ensure comparability across models.[47,57]

3. Machine Learning Models: Random Forest, MLP, and SVM

We applied three machine learning classifiers to each feature set: RF, Multi-layer Perceptron (MLP), and SVM, implemented using the scikit-learn library (v1.3.0).[58] Each classifier was trained to distinguish POAG cases from controls based on different combinations of demographic, clinical, and genetic inputs.

- RF [40,59] was configured with 100 decision trees (n_estimators=100), using the default max_depth=None and class_weight='balanced' to correct for label imbalance.
- MLP [38] employed a single hidden layer with 32 neurons (hidden_layer_sizes=(32,)), ReLU activation, and stochastic gradient descent with early stopping (max_iter=300).
- SVM [44] was used with a radial basis function (RBF) kernel (kernel='rbf'), and probability calibration was enabled (probability=True) to support AUC estimation.

Each model was trained and evaluated using 10-fold bootstrap resampling. In each bootstrap iteration:

1. The dataset was resampled with replacement and stratified by label (stratify=y),
2. An 80/20 train/test split was applied,
3. The training pipeline included mean imputation (SimpleImputer), z-score standardization (StandardScaler), and the classifier.

Predictions from each iteration were aggregated to compute mean performance metrics, including AUC, accuracy, precision, sensitivity, specificity, and F1-score. Confidence intervals were estimated as ±1.96

standard deviations across bootstraps. This approach was used consistently across all feature sets and classifiers to ensure fair model comparisons and robust estimation.

4. Feature Importance via Random Forest

To improve model interpretability and identify which variables contributed most to POAG prediction, we analyzed feature importances derived from RF classifiers. We focused this analysis on the base with all four PRS feature set, which comprised demographic variables (age, gender), and genetic predictors (PRS-CS POAAGG, PRS-CS MEGA, MTAG PRS and MEGA PRS). The RF algorithm inherently calculates feature importance by measuring the average decrease in Gini impurity across all decision trees when a particular feature is used for splitting.

We trained RF classifiers on 10 bootstrap-resampled splits of the clean cohort (N = 196), each time extracting the feature_importances_ attribute from the trained model. These raw values were then averaged across bootstrap iterations to yield a robust estimate of variable influence, and standard deviations were computed to quantify variability. The resulting ranked feature list provided a direct comparison of the relative contribution of each variable to the model's predictive power.

Taken together, this feature importance analysis enhances transparency of the ensemble model and helps bridge the gap between black-box machine learning predictions and clinically meaningful variables. This interpretability is critical for establishing trust in AI-assisted diagnostic tools in ophthalmology.

6. Software and Reproducibility

All data preprocessing, feature engineering, and machine learning modeling were performed in Python (version 3.11)[57] using open-source libraries, including scikit-learn (version 1.3.0)[58] for model implementation, pandas and numpy for data manipulation, matplotlib and seaborn for visualization, and scipy for statistical analysis. Genetic data processing, including risk score calculation and dataset merging, was conducted using PLINK v1.9[47] and PRS-CS[46], with ancestry-specific LD panels used to infer SNP weights for polygenic risk scores.

All scripts were developed and executed within Google Colaboratory (Colab), ensuring an accessible, cloud-based computational environment that supports consistent hardware and software dependencies across users. The project followed strict reproducibility protocols. Each model training and evaluation task was encapsulated in modular code blocks with clearly defined input/output interfaces. Outputs for each model-feature set combination were stored as labeled CSV and Excel files, with naming conventions that reflect the model type (e.g., RF, MLP, SVM), feature set used, and performance metrics obtained.

To promote transparency and facilitate replication, all Google Colab notebooks, including those used for bootstrapped model evaluation, feature importance extraction, and visualization generation, have been deposited in an accessible GitHub repository (linked in the Data and Code Availability section). Each notebook includes narrative explanations and step-by-step annotations for ease of use. Additional files—such as input phenotype dictionaries, SNP annotation tables, and processed feature matrices—are included either as supplemental materials (e.g., Table S1–S5) or within the GitHub release archive.

All visualizations shown in the main text (e.g., Figures 2–5) were generated directly from the modeling outputs using standardized plotting functions, which support error bar overlays and stratified grouping. These functions are also available in the repository and can be reused or adapted by other researchers. Overall, the combination of modular scripting, labeled outputs, and cloud-based execution ensures the full reproducibility of our AI modeling pipeline and supports the broader adoption of these tools for glaucoma risk prediction.

QUANTIFICATION AND STATISTICAL ANALYSIS

All statistical analyses and machine learning experiments were conducted using Python (version 3.11) with the scikit-learn library (version 1.3.0), along with supporting packages such as NumPy (1.24.4), pandas (2.0.3), and matplotlib (3.7.2) for visualization. Custom scripts for preprocessing, model training, and evaluation were developed in Google Colab and are available upon request.

We trained and evaluated three machine learning algorithms—RF, MLP, and SVM—on the clean cohort (N = 196) using stratified 10× bootstrap resampling. In each iteration, the dataset was randomly resampled with replacement, and split into training (80%) and testing (20%) subsets while maintaining the

original case–control ratio through stratification. The exact value of *n* for each experiment, representing the number of unique individuals included in the respective cohort, is specified in the main text and figure legends.

Model performance was assessed using five metrics: AUC, accuracy, F1-score, sensitivity (recall for the positive class), and specificity (recall for the negative class). For each metric, we report the mean and standard deviation (SD) across the stratified 10× bootstrap resampling. These summary statistics are displayed in Figures 2–4 and Supplemental Tables S1–S3.

Feature importance scores were derived from the Random Forest model using the mean decrease in Gini impurity across trees. Rankings of feature importance were visualized for the base with all four PRS, with corresponding summary metrics reported in the supplemental tables.

Assumptions of normality or homoscedasticity were not required, given the nonparametric nature of the machine learning models used. Data preprocessing steps (e.g., z-score normalization for PRS scores) are described in the Methods section and were uniformly applied to all models.

All scripts used in model training and evaluation can be shared by the lead contact upon reasonable request.

ADDITIONAL RESOURCES

This study is not a clinical trial, there is no clinical registry number. There is no external sites that have been generated to support discussion or use of the information/data/material created by the manuscript.

REFERENCES

1. Weinreb, R.N., and Khaw, P.T. (2004). Primary open-angle glaucoma. The lancet *363*, 1711-1720.
2. Zhang, N., Wang, J., Li, Y., and Jiang, B. (2021). Prevalence of primary open angle glaucoma in the last 20 years: a meta-analysis and systematic review. Scientific reports *11*, 13762.
3. Quigley, H.A., and Broman, A.T. (2006). The number of people with glaucoma worldwide in 2010 and 2020. British journal of ophthalmology *90*, 262-267.
4. Weinreb, R.N., Leung, C.K., Crowston, J.G., Medeiros, F.A., Friedman, D.S., Wiggs, J.L., and Martin, K.R. (2016). Primary open-angle glaucoma. Nature reviews Disease primers *2*, 1-19.
5. Sommer, A., Tielsch, J.M., Katz, J., Quigley, H.A., Gottsch, J.D., Javitt, J., and Singh, K. (1991). Relationship between intraocular pressure and primary open angle glaucoma among white and black Americans: the Baltimore Eye Survey. Archives of ophthalmology *109*, 1090-1095.
6. Hollands, H., Johnson, D., Hollands, S., Simel, D.L., Jinapriya, D., and Sharma, S. (2013). Do findings on routine examination identify patients at risk for primary open-angle glaucoma?: The rational clinical examination systematic review. Jama *309*, 2035-2042.
7. Tielsch, J.M., Sommer, A., Katz, J., Royall, R.M., Quigley, H.A., and Javitt, J. (1991). Racial variations in the prevalence of primary open-angle glaucoma: the Baltimore Eye Survey. Jama *266*, 369-374.
8. Khachatryan, N., Pistilli, M., Maguire, M.G., Salowe, R.J., Fertig, R.M., Moore, T., Gudiseva, H.V., Chavali, V.R., Collins, D.W., and Daniel, E. (2019). Primary open-angle African American glaucoma genetics (POAAGG) study: gender and risk of POAG in African Americans. PloS one *14*, e0218804.
9. Bettin, P., and Di Matteo, F. (2013). Glaucoma: present challenges and future trends. Ophthalmic research *50*, 197-208.
10. Lewis, C.M., and Vassos, E. (2020). Polygenic risk scores: from research tools to clinical instruments. Genome medicine *12*, 44.
11. Verma, S.S., Gudiseva, H.V., Chavali, V.R., Salowe, R.J., Bradford, Y., Guare, L., Lucas, A., Collins, D.W., Vrathasha, V., and Nair, R.M. (2024). A multi-cohort genome-wide association study in African ancestry individuals reveals risk loci for primary open-angle glaucoma. Cell *187*, 464-480. e410.
12. Igo Jr, R.P., Kinzy, T.G., and Cooke Bailey, J.N. (2019). Genetic risk scores. Current protocols in human genetics *104*, e95.
13. Läll, K., Mägi, R., Morris, A., Metspalu, A., and Fischer, K. (2017). Personalized risk prediction for type 2 diabetes: the potential of genetic risk scores. Genetics in Medicine *19*, 322-329.
14. Chasman, D.I., Giulianini, F., Demler, O.V., and Udler, M.S. (2020). Pleiotropy-based decomposition of genetic risk scores: association and interaction analysis for type 2 diabetes and CAD. The American Journal of Human Genetics *106*, 646-658.
15. Ripatti, S., Tikkanen, E., Orho-Melander, M., Havulinna, A.S., Silander, K., Sharma, A., Guiducci, C., Perola, M., Jula, A., and Sinisalo, J. (2010). A multilocus genetic risk score for coronary heart disease: case-control and prospective cohort analyses. The Lancet *376*, 1393-1400.
16. Morrison, A.C., Bare, L.A., Chambless, L.E., Ellis, S.G., Malloy, M., Kane, J.P., Pankow, J.S., Devlin, J.J., Willerson, J.T., and Boerwinkle, E. (2007). Prediction of coronary heart disease risk using a genetic risk score: the Atherosclerosis Risk in Communities Study. American journal of epidemiology *166*, 28-35.
17. Shieh, Y., Hu, D., Ma, L., Huntsman, S., Gard, C.C., Leung, J.W., Tice, J.A., Vachon, C.M., Cummings, S.R., and Kerlikowske, K. (2016). Breast cancer risk prediction using a clinical risk model and polygenic risk score. Breast cancer research and treatment *159*, 513-525.

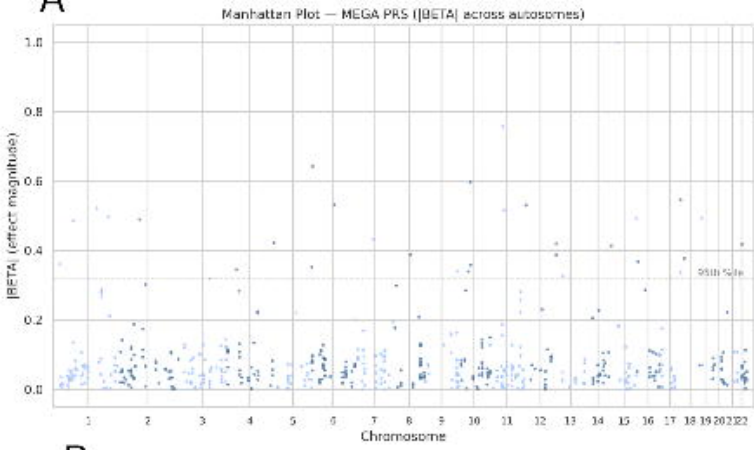
757 18. Roberts, E., Howell, S., and Evans, D.G. (2023). Polygenic risk scores and breast cancer risk
758 prediction. The Breast *67*, 71-77.
759 19. Mavaddat, N., Michailidou, K., Dennis, J., Lush, M., Fachal, L., Lee, A., Tyrer, J.P., Chen, T.-H.,
760 Wang, Q., and Bolla, M.K. (2019). Polygenic risk scores for prediction of breast cancer and breast
761 cancer subtypes. The American Journal of Human Genetics *104*, 21-34.
762 20. Qassim, A., Souzeau, E., Siggs, O.M., Hassall, M.M., Han, X., Griffiths, H.L., Frost, N.A., Vallabh,
763 N.A., Kirwan, J.F., and Menon, G. (2020). An intraocular pressure polygenic risk score stratifies
764 multiple primary open-angle glaucoma parameters including treatment intensity.
765 Ophthalmology *127*, 901-907.
766 21. Gao, X.R., Huang, H., and Kim, H. (2019). Polygenic risk score is associated with intraocular
767 pressure and improves glaucoma prediction in the UK Biobank cohort. Translational vision
768 science & technology *8*, 10-10.
769 22. Han, X., Hewitt, A.W., and MacGregor, S. (2021). Predicting the future of genetic risk profiling of
770 glaucoma: a narrative review. JAMA ophthalmology *139*, 224-231.
771 23. Wang, Y., Kanai, M., Tan, T., Kamariza, M., Tsuo, K., Yuan, K., Zhou, W., Okada, Y., Huang, H., and
772 Turley, P. (2023). Polygenic prediction across populations is influenced by ancestry, genetic
773 architecture, and methodology. Cell Genomics *3*.
774 24. Lambert, S.A., Abraham, G., and Inouye, M. (2019). Towards clinical utility of polygenic risk
775 scores. Human molecular genetics *28*, R133-R142.
776 25. Sekimitsu, S., Xiang, D., Smith, S.L., Curran, K., Elze, T., Friedman, D.S., Foster, P.J., Luo, Y.,
777 Pasquale, L.R., and Peto, T. (2023). Deep ocular phenotyping across primary open-angle
778 glaucoma genetic burden. JAMA ophthalmology *141*, 891-899.
779 26. Rajpurkar, P., Chen, E., Banerjee, O., and Topol, E.J. (2022). AI in health and medicine. Nature
780 medicine *28*, 31-38.
781 27. Wang, F., and Preininger, A. (2019). AI in health: state of the art, challenges, and future
782 directions. Yearbook of medical informatics *28*, 016-026.
783 28. Ngiam, K.Y., and Khor, W. (2019). Big data and machine learning algorithms for health-care
784 delivery. The Lancet Oncology *20*, e262-e273.
785 29. Beam, A.L., and Kohane, I.S. (2018). Big data and machine learning in health care. Jama *319*,
786 1317-1318.
787 30. Zhu, Y., Salowe, R., Chow, C., Li, S., Bastani, O., and O'Brien, J.M. (2024). Advancing glaucoma
788 care: integrating artificial intelligence in diagnosis, management, and progression detection.
789 Bioengineering *11*, 122.
790 31. Zeppieri, M., Gardini, L., Culiersi, C., Fontana, L., Musa, M., D'Esposito, F., Surico, P.L., Gagliano,
791 C., and Sorrentino, F.S. (2024). Novel Approaches for the Early Detection of Glaucoma Using
792 Artificial Intelligence. Life *14*, 1386.
793 32. Alipanahi, B., Hormozdiari, F., Behsaz, B., Cosentino, J., McCaw, Z.R., Schorsch, E., Sculley, D.,
794 Dorfman, E.H., Foster, P.J., and Peng, L.H. (2021). Large-scale machine-learning-based
795 phenotyping significantly improves genomic discovery for optic nerve head morphology. The
796 American Journal of Human Genetics *108*, 1217-1230.
797 33. Han, X., Steven, K., Qassim, A., Marshall, H.N., Bean, C., Tremeer, M., An, J., Siggs, O.M.,
798 Gharahkhani, P., and Craig, J.E. (2021). Automated AI labeling of optic nerve head enables
799 insights into cross-ancestry glaucoma risk and genetic discovery in> 280,000 images from UKB
800 and CLSA. The American Journal of Human Genetics *108*, 1204-1216.
801 34. Truong, B., Hull, L.E., Ruan, Y., Huang, Q.Q., Hornsby, W., Martin, H., Van Heel, D.A., Wang, Y.,
802 Martin, A.R., and Lee, S.H. (2024). Integrative polygenic risk score improves the prediction
803 accuracy of complex traits and diseases. Cell Genomics *4*.

35. Zhai, S., Mehrotra, D.V., and Shen, J. (2024). Applying polygenic risk score methods to pharmacogenomics GWAS: challenges and opportunities. Briefings in Bioinformatics *25*, bbad470.
36. Mamidipaka, A., Di Rosa, I., Lee, R., Zhu, Y., Chen, Y., Salowe, R., Addis, V., Sankar, P., Daniel, E., and Ying, G.-S. (2023). Factors associated with large cup-to-disc ratio and blindness in the primary open-angle African American glaucoma genetics (POAAGG) study. Genes *14*, 1809.
37. Gudiseva, H.V., Verma, S.S., Chavali, V.R., Salowe, R.J., Lucas, A., Collins, D.W., Rathi, S., He, J., Lee, R., and Merriam, S. (2020). Genome wide-association study identifies novel loci in the Primary Open-Angle African American Glaucoma Genetics (POAAGG) study. BioRxiv, 2020.2002. 2027.968156.
38. Popescu, M.-C., Balas, V.E., Perescu-Popescu, L., and Mastorakis, N. (2009). Multilayer perceptron and neural networks. WSEAS Transactions on Circuits and Systems *8*, 579-588.
39. Qi, Y. (2012). Random forest for bioinformatics. In Ensemble machine learning, (Springer), pp. 307-323.
40. Belgiu, M., and Drăguţ, L. (2016). Random forest in remote sensing: A review of applications and future directions. ISPRS journal of photogrammetry and remote sensing *114*, 24-31.
41. Breiman, L. (2001). Random forests. Machine learning *45*, 5-32.
42. Suthaharan, S. (2016). Support vector machine. In Machine learning models and algorithms for big data classification: thinking with examples for effective learning, (Springer), pp. 207-235.
43. Hearst, M.A., Dumais, S.T., Osuna, E., Platt, J., and Scholkopf, B. (1998). Support vector machines. IEEE Intelligent Systems and their applications *13*, 18-28.
44. Cortes, C., and Vapnik, V. (1995). Support-vector networks. Machine learning *20*, 273-297.
45. Chen, L., Wang, Y.-F., Liu, L., Bielowka, A., Ahmed, R., Zhang, H., Tombleson, P., Roberts, A.L., Odhams, C.A., and Cunninghame Graham, D.S. (2020). Genome-wide assessment of genetic risk for systemic lupus erythematosus and disease severity. Human molecular genetics *29*, 1745-1756.
46. Ge, T., Chen, C.-Y., Ni, Y., Feng, Y.-C.A., and Smoller, J.W. (2019). Polygenic prediction via Bayesian regression and continuous shrinkage priors. Nature communications *10*, 1776.
47. Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M.A., Bender, D., Maller, J., Sklar, P., De Bakker, P.I., and Daly, M.J. (2007). PLINK: a tool set for whole-genome association and population-based linkage analyses. The American journal of human genetics *81*, 559-575.
48. Coleman, A.L., and Miglior, S. (2008). Risk factors for glaucoma onset and progression. Survey of ophthalmology *53*, S3-S10.
49. Doshi-Velez, F., and Kim, B. (2017). Towards a rigorous science of interpretable machine learning. arXiv preprint arXiv:1702.08608.
50. Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nature machine intelligence *1*, 206-215.
51. Lundberg, S.M., and Lee, S.-I. (2017). A unified approach to interpreting model predictions. Advances in neural information processing systems *30*.
52. Ribeiro, M.T., Singh, S., and Guestrin, C. (2016). " Why should i trust you?" Explaining the predictions of any classifier. pp. 1135-1144.
53. Molnar, C., Casalicchio, G., and Bischl, B. (2020). Interpretable machine learning–a brief history, state-of-the-art and challenges. (Springer), pp. 417-431.
54. Choi, S.W., Mak, T.S.-H., and O'Reilly, P.F. (2020). Tutorial: a guide to performing polygenic risk score analyses. Nature protocols *15*, 2759-2772.
55. Khera, A.V., Chaffin, M., Aragam, K.G., Haas, M.E., Roselli, C., Choi, S.H., Natarajan, P., Lander, E.S., Lubitz, S.A., and Ellinor, P.T. (2018). Genome-wide polygenic scores for common diseases

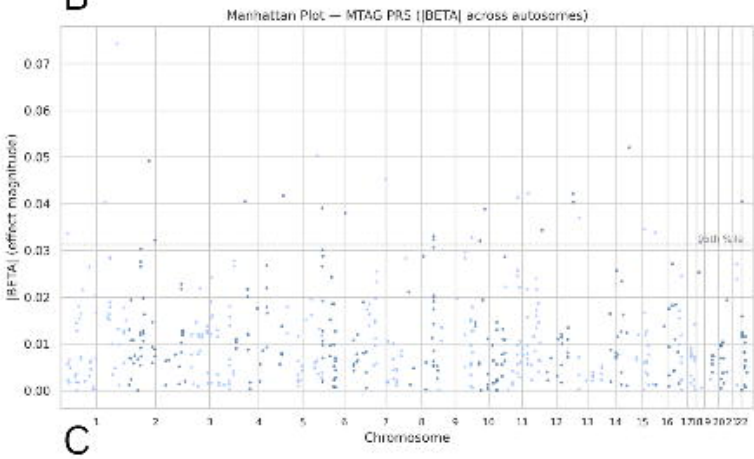
851 identify individuals with risk equivalent to monogenic mutations. Nature genetics *50*, 1219-
852 1224.
853 56. Sirugo, G., Williams, S.M., and Tishkoff, S.A. (2019). The missing diversity in human genetic
854 studies. Cell *177*, 26-31.
855 57. Van Rossum, G., and Drake Jr, F.L. (1995). Python tutorial (Centrum voor Wiskunde en
856 Informatica Amsterdam, The Netherlands).
857 58. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M.,
858 Prettenhofer, P., Weiss, R., and Dubourg, V. (2011). Scikit-learn: Machine learning in Python. the
859 Journal of machine Learning research *12*, 2825-2830.
860 59. Efron, B., and Tibshirani, R.J. (1994). An introduction to the bootstrap (Chapman and Hall/CRC).

861

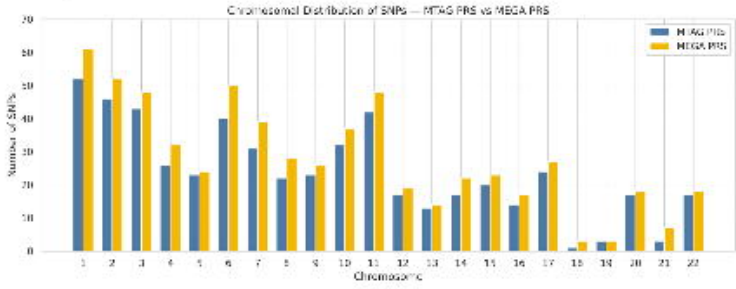
A


Manhattan Plot — MEGA PRS (|BETA| across autosomes)
|BETA| (effect magnitude)
95th %ile
Chromosome

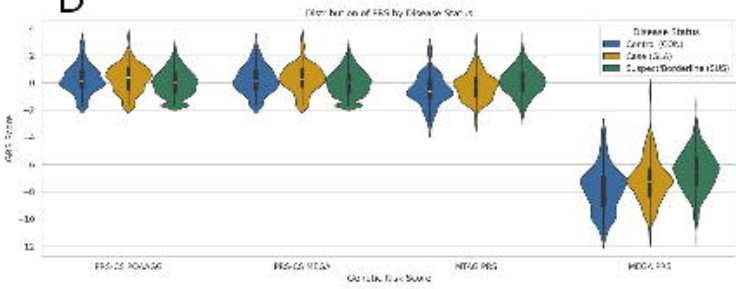
B


Manhattan Plot — MTAG PRS (|BETA| across autosomes)
|BETA| (effect magnitude)
95th %ile
Chromosome

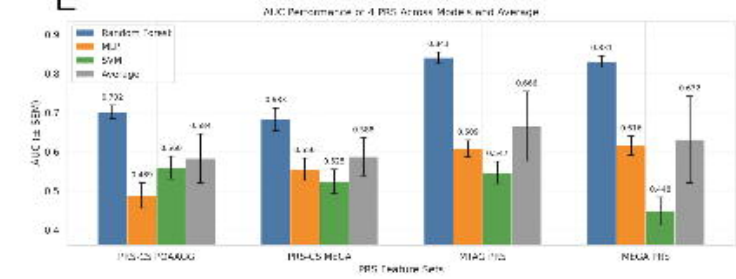
C

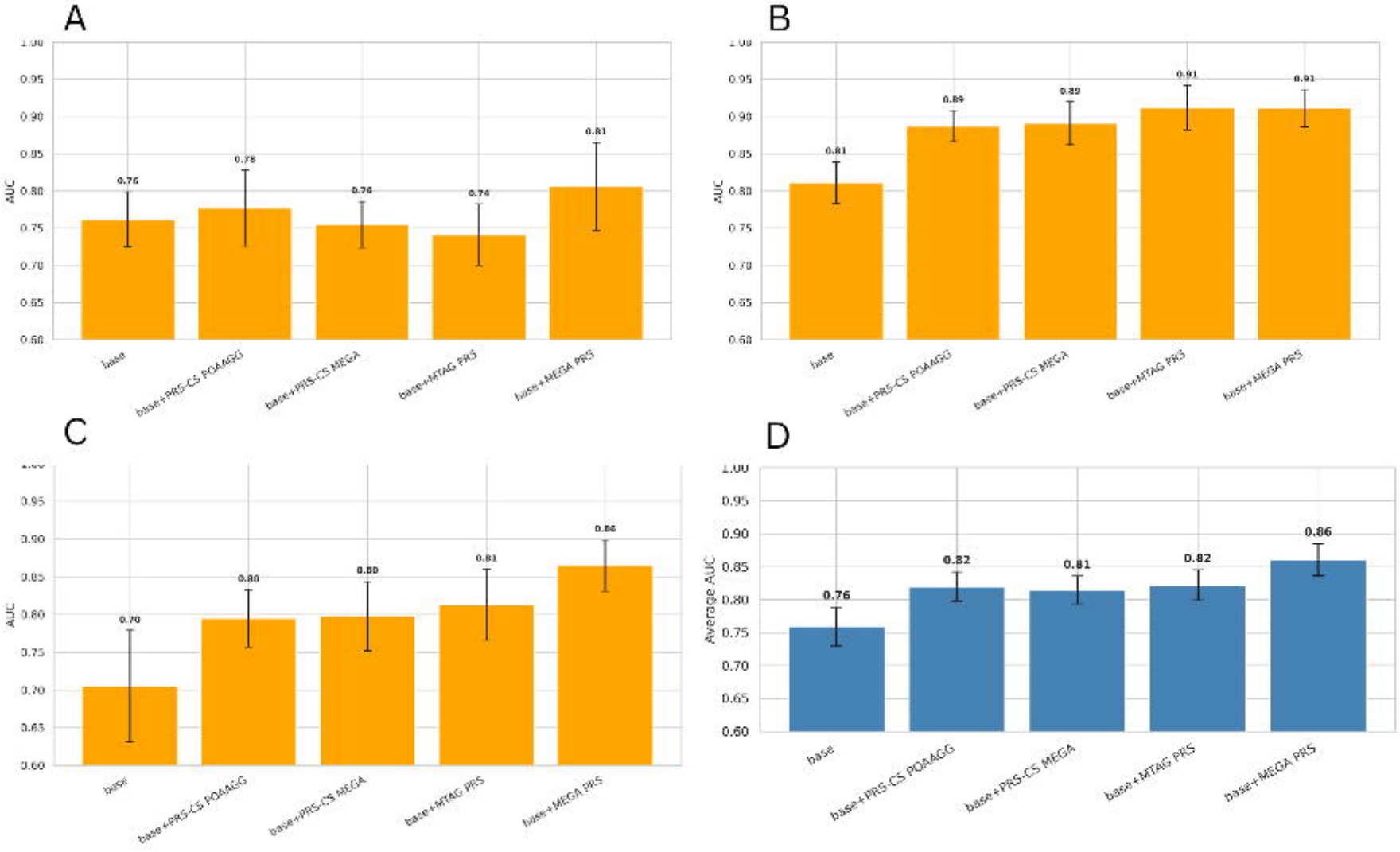

Chromosomal Distribution of SNPs — MTAG PRS vs MEGA PRS
MTAG PRS
MEGA PRS
Number of SNPs
Chromosome

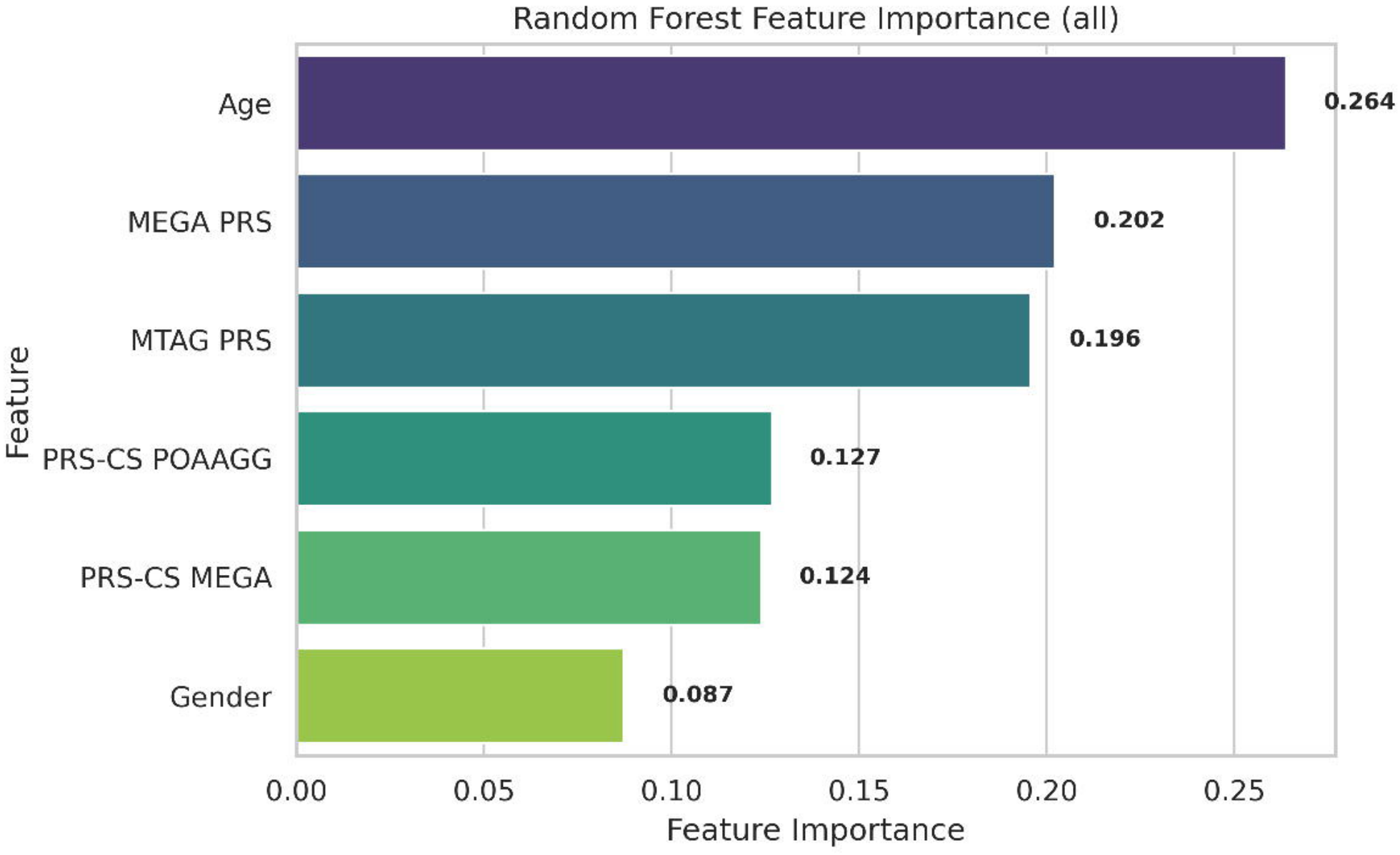
D

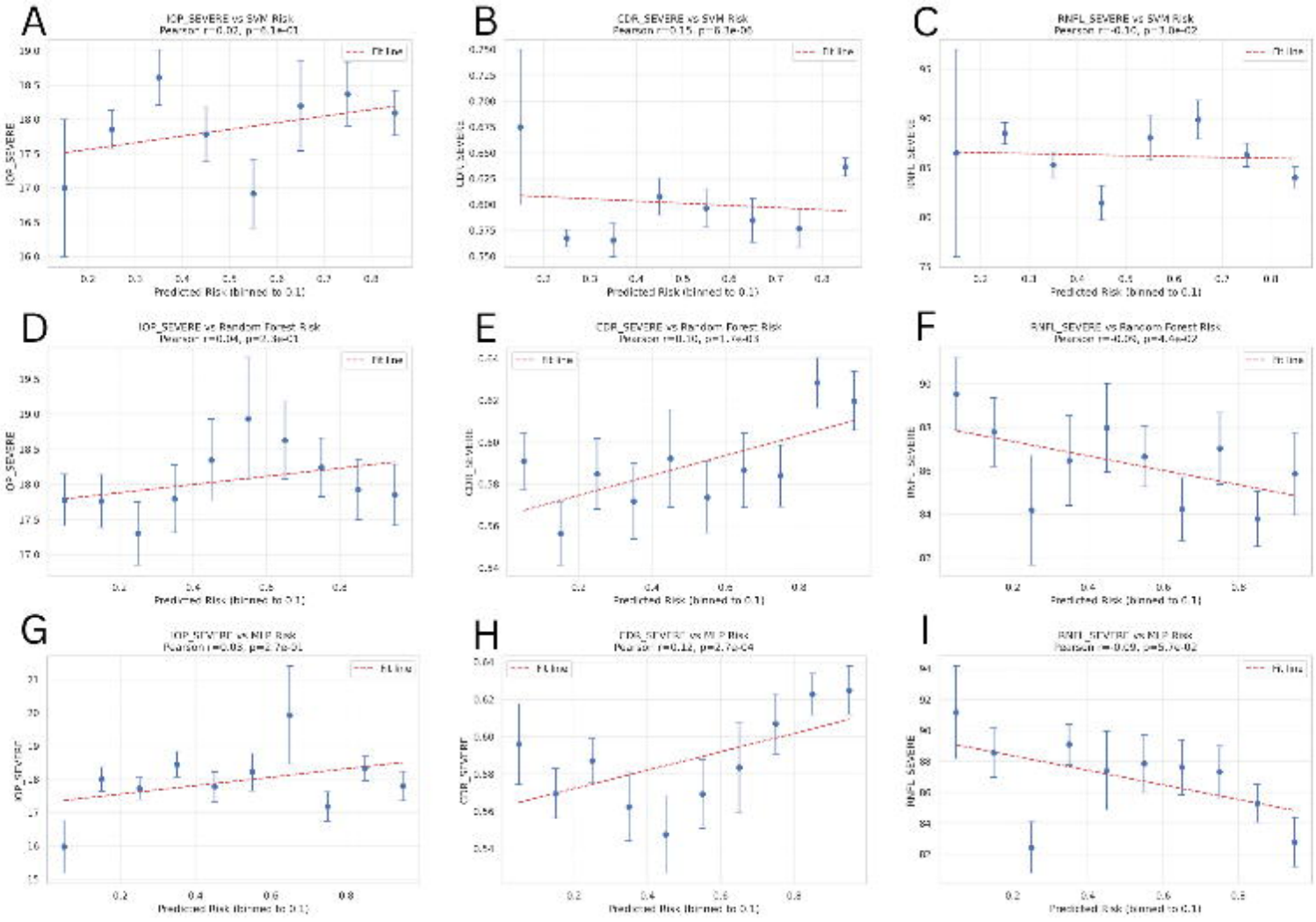

Distribution of PRS by Disease Status
Disease Status
Controls (CON)
Cases (CAD)
Suspect/Borderline (SUS)
PRS Score
MTAG PRS
MEGA PRS
Genetic Risk Score

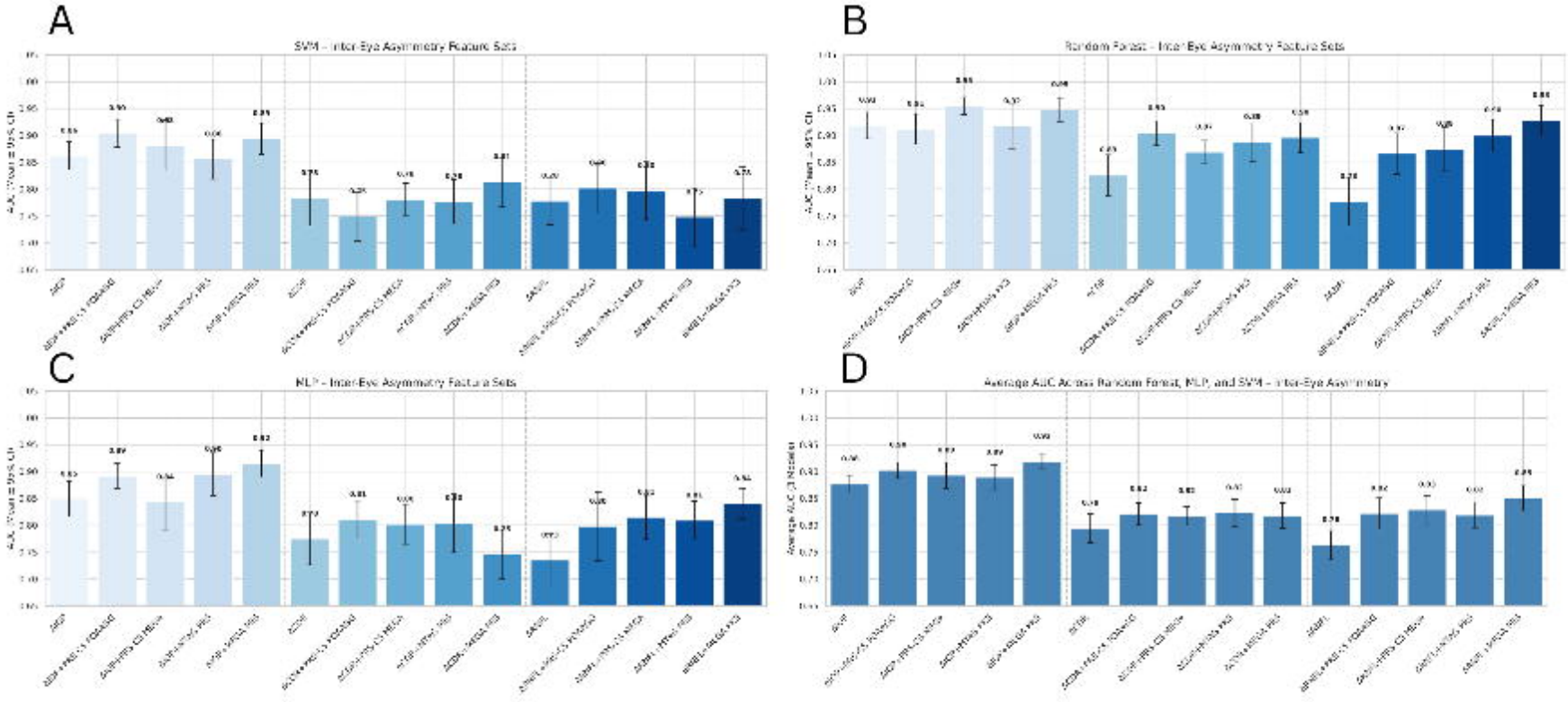
E


AUC Performance of 4 PRS Across Models and Average
Random Forest
MLP
SVM
Average
AUC (± SEM)
PRS Feature Sets

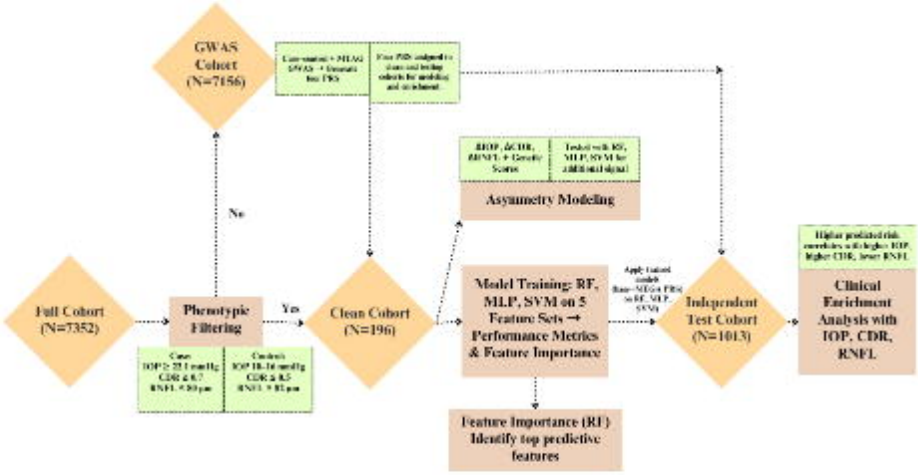

A
AUC
0.76
0.78
0.76
0.74
0.81
base
base+PRS-CS POAAGG
base+PRS-CS MEGA
base+MTAG PRS
base+MEGA PRS
B
AUC
0.81
0.89
0.89
0.91
0.91
base
base+PRS-CS POAAGG
base+PRS-CS MEGA
base+MTAG PRS
base+MEGA PRS
C
AUC
0.70
0.80
0.80
0.81
0.86
base
base+PRS-CS POAAGG
base+PRS-CS MEGA
base+MTAG PRS
base+MEGA PRS
D
Average AUC
0.76
0.82
0.81
0.82
0.86
base
base+PRS-CS POAAGG
base+PRS-CS MEGA
base+MTAG PRS
base+MEGA PRS


Random Forest Feature Importance (all)
Age
MEGA PRS
MTAG PRS
PRS-CS POAAGG
PRS-CS MEGA
Gender
0.264
0.202
0.196
0.127
0.124
0.087
Feature
0.00
0.05
0.10
0.15
0.20
0.25
Feature Importance

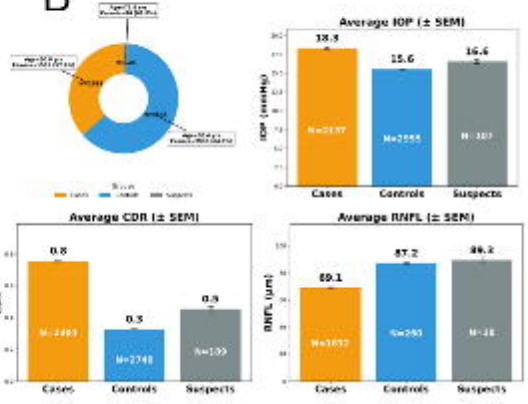

A
IOP_SEVERE vs SVM Risk
Fit line
IOP_SEVERE
Predicted Risk (binned to 0.1)
B
CDR_SEVERE vs SVM Risk
Fit line
CDR_SEVERE
Predicted Risk (binned to 0.1)
C
RNFL_SEVERE vs SVM Risk
Fit line
RNFL_SEVERE
Predicted Risk (binned to 0.1)
D
IOP_SEVERE vs Random Forest Risk
Fit line
IOP_SEVERE
Predicted Risk (binned to 0.1)
E
CDR_SEVERE vs Random Forest Risk
Fit line
CDR_SEVERE
Predicted Risk (binned to 0.1)
F
RNFL_SEVERE vs Random Forest Risk
Fit line
RNFL_SEVERE
Predicted Risk (binned to 0.1)
G
IOP_SEVERE vs MLP Risk
Fit line
IOP_SEVERE
Predicted Risk (binned to 0.1)
H
CDR_SEVERE vs MLP Risk
Fit line
CDR_SEVERE
Predicted Risk (binned to 0.1)
I
RNFL_SEVERE vs MLP Risk
Fit line
RNFL_SEVERE
Predicted Risk (binned to 0.1)


A
SVM – Inter-Eye Asymmetry Feature Sets
B
Random Forest – Inter-Eye Asymmetry Feature Sets
C
MLP – Inter-Eye Asymmetry Feature Sets
D
Average AUC Across Random Forest, MLP, and SVM – Inter-Eye Asymmetry

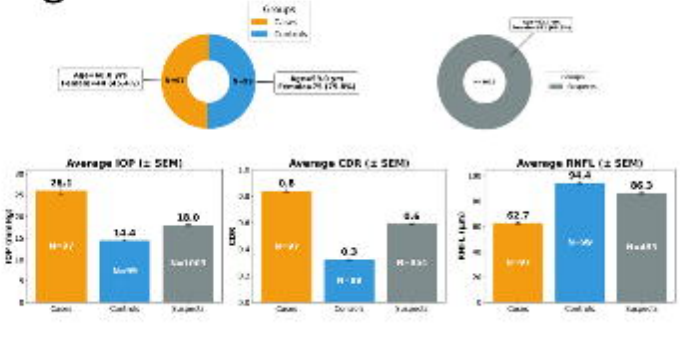
A


GWAS Cohort (N=7156)
No
Full Cohort (N=7352)
Phenotypic Filtering
Yes
Clean Cohort (N=196)
Asymmetry Modeling
Model Training: RF, MLP, SVM on 5 Feature Sets → Performance Metrics & Feature Importance
Feature Importance (RF) Identify top predictive features
Independent Test Cohort (N=1013)
Clinical Enrichment Analysis with IOP, CDR, RNFL

B


Average IOP (± SEM)
18.8
15.6
16.6
Cases
Controls
Suspects
Average CDR (± SEM)
0.8
0.3
0.3
Average RNFL (± SEM)
89.1
87.2
86.2

C


Average IOP (± SEM)
Average CDR (± SEM)
0.8
0.3
0.6
Average RNFL (± SEM)
62.7
94.4
86.3