Multimodal Prediction of Primary Open-Angle Glaucoma Using Genetic 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], Laxmi Moksha[1], Vrathasha Vrathasha[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 an integrative machine learning framework combining curated genetic risk scores (GRS 1–4), polygenic risk scores (PRS), and demographic features. A clean, biologically extreme training cohort (N=196) enabled robust model development with Random Forest, Support Vector Machine, and Multi-Layer Perceptron classifiers, followed by validation in 1,013 unlabeled glaucoma suspects. Curated GRS—particularly GRS 2 and GRS 4—consistently outperformed PRS and significantly boosted predictive accuracy (AUC up to 0.95 when combined with base features). Risk predictions correlated with intraocular pressure, cup-to-disc ratio, and retinal nerve fiber layer thickness in suspects, confirming biological relevance. These results establish curated, ancestry-matched GRS as interpretable and clinically translatable predictors, paving the way for equitable, genetics-driven precision screening strategies in high-risk populations.

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) and genetic risk scores (GRS), 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 GRS 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, 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 GRS and 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 GRS, can provide added predictive value beyond standard phenotypes.

In this study, we leverage the 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 ancestry-matched PRS and GRS, which were then evaluated within the clean cohort. Our models incorporate two ancestry-matched PRS (from the MEGA and POAAGG GWAS) and four curated GRS derived from glaucoma-associated loci from African ancestry. We hypothesize that curated GRS, 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, GRS 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. From our results, we find that curated GRS indeed outperform PRS in clean cohorts, while PRS remain useful in broader 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

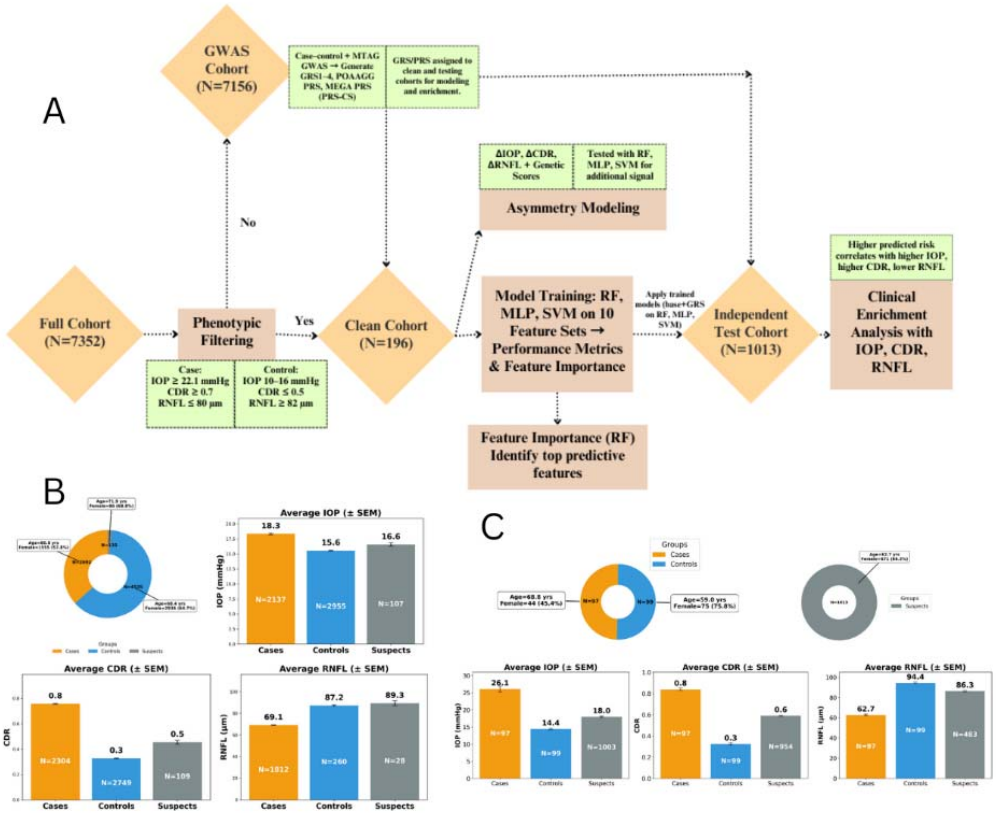
To develop a robust and interpretable machine learning framework for glaucoma risk prediction, we designed a structured pipeline that integrates phenotypic filtering, genomic risk scores, and model interpretability through feature importance and clinical validation (Figure 1A). Starting from a large-scale 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. Genetic risk scores (GRS and PRS) were generated using both manually curated loci and polygenic scoring methods (PRS-CS), and applied uniformly to both training and test sets. Three machine learning models—RF, MLP, and SVM—were trained across ten 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). Genetic risk scores (GRS1–4) and PRS were computed and applied to the clean cohort and testing cohort. 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.

GRS-Based Prediction Models Reveal Limited but Interpretable Signals

To evaluate the predictive utility of curated genetic features alone, we constructed and benchmarked four genetic risk scores (GRS 1–4) derived from GWASconducted within African ancestry populations. These

scores were designed to capture biologically relevant POAG loci while maintaining interpretability, and were applied to our internal training cohort (N=196) and external testing cohort (N=1013). [12,14,45 46]

The four GRS models were derived from loci reported in our previous paper and reflect different combinations of binary and quantitative GWAS sources.[11] GRS 1 and GRS 2 were based on a binary case-control GWAS from the MEGA African ancestry dataset, as published in our previous work.[11] Specifically, GRS 1 included 46 SNPs reaching genome-wide significance ($p < 5\times10^{-8}$), while GRS 2 expanded this set to 174 SNPs by incorporating loci with suggestive association signals ($p < 5\times10^{-5}$). GRS 3 and GRS 4 were derived from a multi-trait quantitative GWAS (MTAG) meta-analysis focused on IOP as a continuous trait—thereby capturing endophenotypic variation linked to subclinical disease progression. GRS 3 included 42 genome-wide significant loci, while GRS 4 incorporated 190 loci based on more inclusive p-value thresholds. All GRS variants were mapped to hg38, manually curated, and scored using PLINK with ancestry-specific effect weights (see STAR Methods).[47]

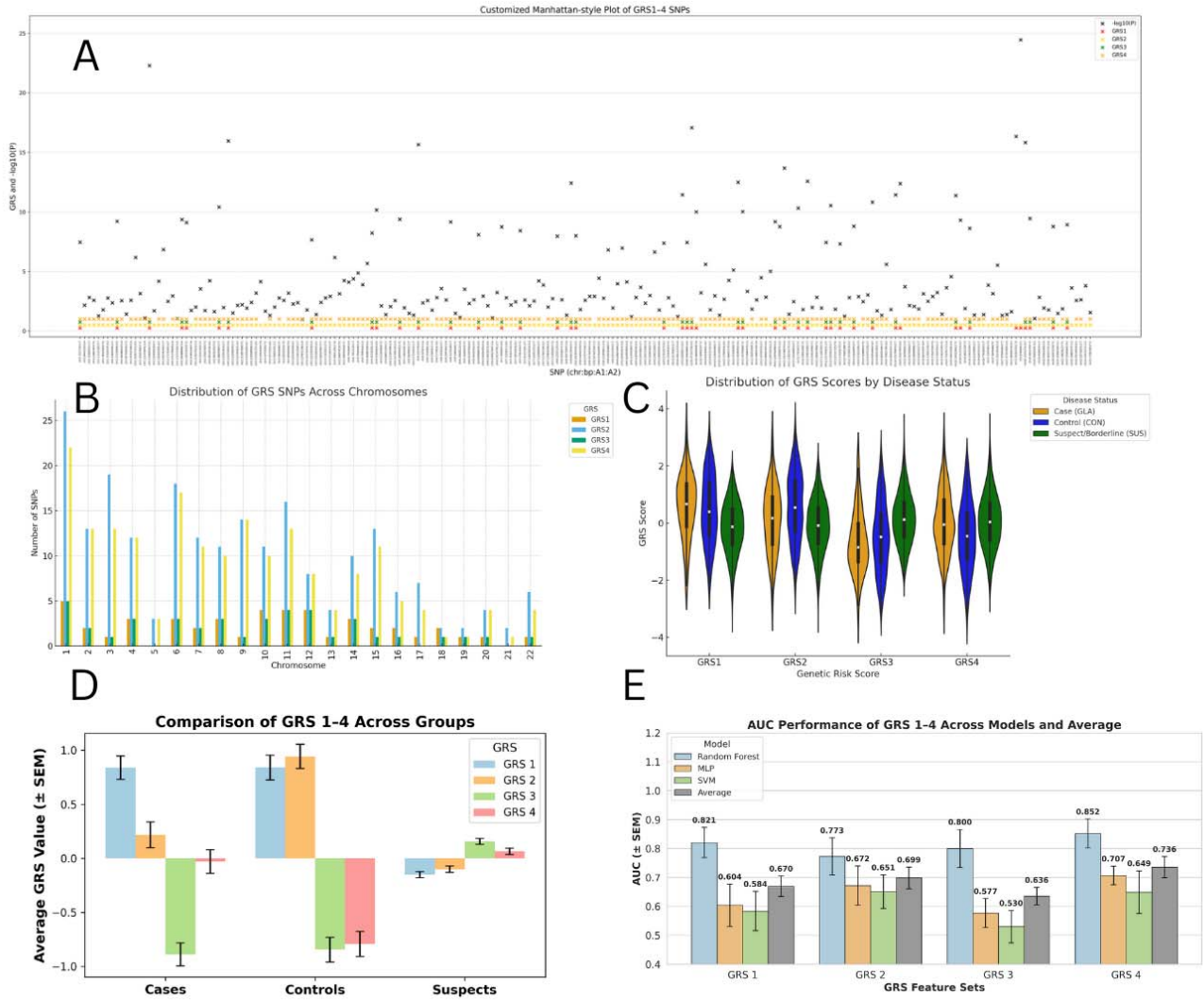
We visualized the composition and genomic distribution of these scores in multiple ways (Figure 2). A customized Manhattan-style plot (Figure 2A) shows the overlap of SNPs across GRS 1–4 and highlights the original GWAS signal strength for each locus. Chromosomal breakdown (Figure 2B) indicates wide genomic coverage for GRS 2 and GRS 4, whereas GRS 1 and GRS 3 were more concentrated. Distribution plots of GRS values across cases, controls, and suspects (Figure 2C) reveal a modest but consistent trend: higher GRS values among cases and lower values among controls, with suspects spanning an intermediate range.

In both the 196-sample clean cohort and the 1,013-sample independent test cohort (Figure 2D), curated GRS demonstrated distinct distributional patterns between cases and controls. While the direction of effect varied across scores—cases were not consistently higher than controls—GRS 2 and GRS 4 showed the clearest and most reproducible separation between the two groups. Suspects generally exhibited intermediate values, reinforcing their role as a biologically plausible group on the disease spectrum. These findings suggest that specific curated GRS, particularly GRS 2 and GRS 4, capture meaningful genetic differences between affected and unaffected individuals.

We next assessed the predictive performance of GRS 1–4 using three machine-learning classifiers (RF, MLP, SVM) trained on the 196 clean cohort (Figure 2E). RF consistently outperformed MLP and SVM across all GRS, with GRS 4 achieving the highest RF performance (AUC = 0.852 ± 0.052), followed by GRS 1 (AUC = 0.821 ± 0.057). MLP and SVM models demonstrated lower and more variable AUCs, with several values below 0.60, indicating weaker generalization in this small, well-curated cohort. To provide a model-agnostic summary, we computed the mean AUC across RF, MLP, and SVM for each GRS. The average AUCs confirmed GRS 4 as the strongest overall predictor (mean AUC = 0.736), followed by GRS 2 (0.699), GRS 1 (0.670), and GRS 3 (0.636). This cross-model aggregation highlights that GRS 4 performs robustly regardless of algorithm choice, reinforcing its potential as the most informative genetic score for POAG risk prediction in our dataset.

Despite their conceptual appeal and ancestry-specific design, the GRS scores demonstrated only moderate discriminatory power when used alone. These findings reinforce a key limitation: sparse genetic models—even when biologically informed—only partially reflect the complex architecture of POAG. Their utility may be enhanced by integration with phenotypic and endophenotypic features, which we explore in subsequent sections.

Figure 2. Performance and Characteristics of Genetic Risk Scores (GRS 1–4)



(A). Customized Manhattan-style representation of SNPs used in GRS 1–4.
Each SNP is plotted on the x-axis using its full genomic position (chr:bp:A1:A2), sorted by chromosomal order. The -log10(p-value) from the original GWAS is plotted on the y-axis (black crosses). Colored markers indicate SNP inclusion across four GRSs: red (GRS 1), yellow (GRS 2), green (GRS 3), and orange (GRS 4), allowing visual comparison of overlap and signal strength across loci.

(B). Distribution of GRS SNPs across chromosomes.
This bar plot shows the number of SNPs from each GRS (1–4) on chromosomes 1 to 22. GRS 2 and GRS 4 span broader genomic regions with SNPs distributed more widely than GRS 1 and GRS 3.

(C). Distribution of GRS scores by disease status.
Violin plots display the standardized GRS 1–4 values across three groups: cases (orange), controls (blue), and suspects (green). Cases tend to have higher GRS values, while controls and suspects show overlapping but generally lower distributions, suggesting moderate separation power.

(D). Comparison of GRS 1–4 across cases, controls, and suspects.
Bar plots show mean ± SEM of four curated genetic risk scores (GRS 1–4) in the clean training cohort (cases and controls) and independent testing cohort (suspects). Colors indicate the four GRS models, and x-axis groups represent disease status categories. GRS 1 and GRS 2, based on genome-wide significant and sub-threshold loci from binary POAG GWAS, show higher average values in both cases and controls relative to suspects, whereas GRS 3 and GRS 4 (derived from quantitative MTAG IOP traits) exhibit negative mean values for cases and controls but modestly positive values in suspects, highlighting distinct genetic risk distributions across groups.

(E). AUC performance of GRS-based models across three machine learning methods and their average.
Bar plots show mean AUC (± SEM) for GRS 1–4 when used as single feature sets to predict POAG status in the clean training cohort (N = 196). Results are reported for Random Forest, MLP, and SVM models, with a fourth bar representing the average performance across all three models. Random Forest consistently achieved the highest predictive power, with AUCs ranging from 0.77 to 0.85 across GRS sets, whereas MLP and SVM showed moderate but comparable performance. The inclusion of the

averaged bar highlights overall model stability and allows comparison of genetic risk score signal strength independent of model choice.

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

In the 196-participant clean training cohort (97 POAG cases, 99 controls), we systematically evaluated the predictive performance of three classifiers—RF, SVM, and MLP—across ten feature sets that incrementally incorporated genetic risk information. Each classifier was evaluated across ten predefined feature sets, starting with a baseline model (age and gender) and extending to combinations of curated genetic risk scores (GRS1–4), polygenic risk scores (PRS from POAAGG and MEGA), and a comprehensive "all features" set that integrated both.

As shown in Figure 3A–C, all three classifiers exhibited substantial improvements in AUC upon inclusion of genetic risk scores. For the RF model (Figure 3B), performance increased from an AUC of 0.80 using the base feature set to 0.93 when incorporating all four GRS (base+GRS), with a similarly high AUC of 0.95 when all genetic scores (GRS+PRS) were included. SVM followed a similar trend (Figure 3A), improving from 0.75 (base) to 0.90 (base+GRS), while the MLP classifier (Figure 3C) increased from 0.76 (base) to 0.93 (base+GRS), matching the peak performance of RF. Notably, the improvements were robust across multiple GRS and PRS inputs, suggesting that genetic information provides generalizable predictive value across diverse algorithmic frameworks.

Although GRS1–4 were derived from overlapping loci, they captured complementary signals due to differences in their derivation: GRS1 and GRS2 were built from binary case-control GWAS results in African ancestry individuals, while GRS3 and GRS4 were derived from MTAG-based quantitative trait GWAS focusing on IOP. These distinctions reflect differences in underlying effect-size estimates, trait definitions, and locus selection strategies. The combination of all four GRS scores thus allows the model to integrate orthogonal dimensions of genetic risk while mitigating noise from any single score. This is supported empirically by the consistent superiority of the base+GRS feature set over models using any individual GRS alone.

To evaluate the consistency of genetic contribution across models, we computed the average AUC of each feature set across the three classifiers (Figure 3D). This aggregate view revealed that the base+GRS feature set achieved the highest average AUC (0.93), followed by the full model (0.91) and base+GRS2 (0.87). In contrast, the base model alone averaged only 0.77, highlighting the added predictive value of polygenic features even in the absence of phenotypic data. The inclusion of PRS from either the POAAGG or MEGA datasets also improved performance (average AUCs of 0.82), but the gain was smaller than that observed with the curated GRS sets. These findings suggest that while genome-wide PRS captures diffuse risk, curated GRS based on known POAG loci may better align with pathophysiologic mechanisms in this African American population.

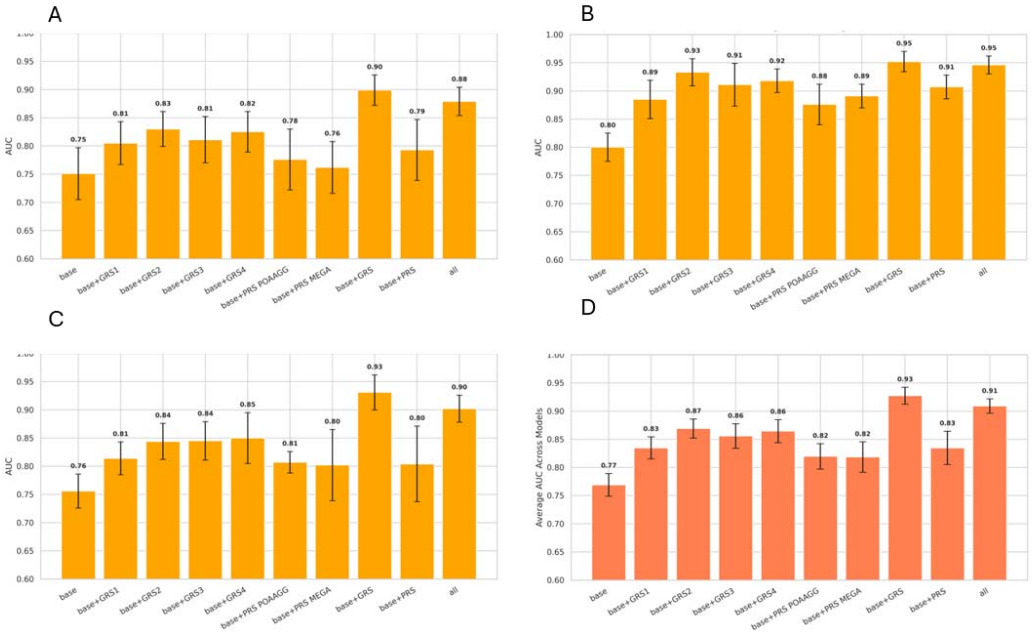
Adding curated GRS to the base model (age + gender) consistently improved predictive performance across all classifiers. On average, the increase in AUC from base to base+GRS ranged from +0.15 to +0.17 (Figure 3A–C). These improvements are notable given the limited sample size and the absence of direct disease markers such as CDR or RNFL thickness in the feature sets. Importantly, the similar gains observed across RF, SVM, and MLP indicate that the added predictive value of GRS is robust and not dependent on a specific modeling approach, but rather reflects underlying genetic risk patterns (Figure 3D).

These findings also carry methodological implications. Ensemble-based (RF) and margin-based (SVM) classifiers yielded strong and stable gains with GRS integration, while the neural network model (MLP) achieved comparable peak performance but required the inclusion of more comprehensive genetic inputs. This observation is consistent with known differences in sample efficiency: RF and SVM models are typically more resilient in small-cohort settings, while MLP performance may degrade with sparse training data unless feature richness compensates for limited sample size. Nonetheless, the overall directionality of improvement across all three models affirms that polygenic risk information is a reliable and transferable predictor.

Taken together, these results provide compelling evidence that GRS can significantly enhance glaucoma risk prediction even in the absence of detailed phenotyping. In the clean cohort, all models demonstrated

clear and reproducible gains from GRS integration, particularly when combining multiple scores that capture complementary dimensions of genetic architecture. These results validate the use of polygenic scores in early-stage screening models and lay the groundwork for clinical enrichment analyses in the larger external and suspect cohorts.

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



(A) SVM models trained on ten feature sets, ranging from demographics only ("base") to combinations with individual GRS, PRS, and their union. Adding GRS notably improved predictive performance, with base+GRS achieving an AUC of 0.90.

(B) RF models trained on the same feature sets. RF achieved the strongest overall results, with base+GRS reaching an AUC of 0.95, the highest among all models and feature sets.

(C) MLP models trained on the ten feature sets. Despite being more sensitive to sample size, MLP also showed clear gains from GRS integration, with base+GRS achieving an AUC of 0.93.

(D) Average AUC across SVM, RF, and MLP for each feature set, summarizing cross-model performance. Base+GRS consistently outperformed other single-feature configurations (mean AUC = 0.93), while the full model with all genetic scores also performed strongly (mean AUC = 0.91). These results highlight the robustness of curated GRS as predictive features across classifiers.

Feature Importance Analysis Highlights the Complementary Role of Genetic Scores

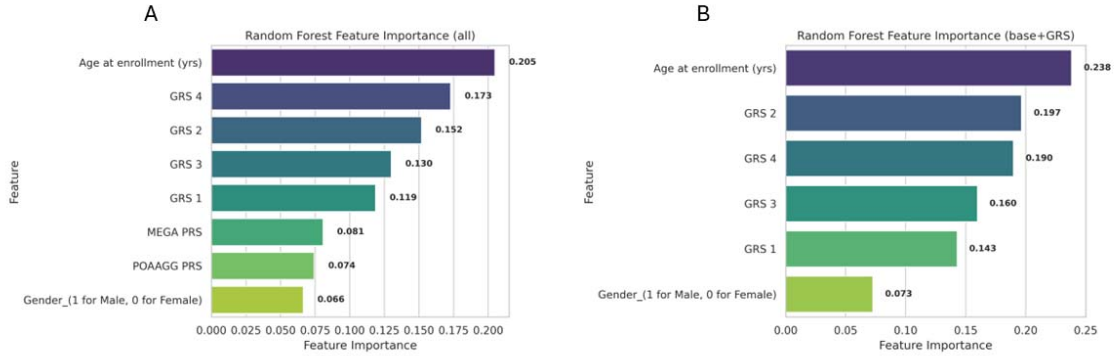
To better understand the internal logic of our predictive models, we performed a feature importance analysis using the Random Forest classifier trained on the clean case–control cohort. As shown in Figure 4, we examined two configurations: one using the full feature set (base + GRS + PRS) and another restricted to base + GRS only. In both settings, age at enrollment emerged as the most important predictor, consistent with its well-established role as a major clinical risk factor for POAG.[48]

Importantly, the four curated genetic risk scores (GRS1–4) consistently ranked among the top contributors, reinforcing their value beyond demographic information. In the base+GRS model (Figure 4B), GRS2 and GRS4 closely followed age, contributing importance scores of 0.197 and 0.190 respectively. However, in the full model (Figure 4A), where PRS and additional features were included, GRS4 surpassed GRS2 in importance (0.173 vs. 0.152). This shift suggests potential feature interaction effects or redundancy between GRS and PRS, which may influence how the Random Forest distributes weight when more features are present.

Across both models, GRS3 and GRS1 contributed meaningful but slightly lower importance, while gender consistently showed minimal influence. Notably, the MEGA PRS and POAAGG PRS, though less impactful than the curated GRSs, still contributed non-negligible signal in the full model, underscoring that genome-wide polygenic burden offers complementary risk information.

Together, these findings highlight the robustness and interpretability of curated genetic scores in POAG risk modeling. Their consistent importance—second only to age—supports their role as transferable and clinically relevant biomarkers. Furthermore, the changes in ranking across models illustrate the adaptive nature of Random Forest in handling correlated features, and reinforce the idea that combining multiple genetic signals enhances predictive accuracy through complementary effects rather than redundancy.

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



(A) Feature importances derived from the Random Forest model using the full feature set (base + all GRS and PRS).
(B) Feature importances from the same model using a restricted feature set (base + GRS only).
In both models, age at enrollment emerged as the most important predictor. Notably, GRS 4 surpassed GRS 2 in the full feature model, whereas GRS 2 ranked higher than GRS 4 in the base+GRS model. This reversal likely reflects feature interaction dynamics and redundancy among genetic scores when combined with PRSs. PRS features (MEGA and POAAGG PRS) contributed moderately but did not exceed curated GRSs in importance.

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+GRS 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.

Among the three classifiers, SVM (Figure 5A-5C) and RF (Figure 5D-5F) demonstrated the strongest and most consistent clinical enrichment. The SVM model showed robust correlations with all three phenotypic markers: CDR ($r = 0.82$, $p = 3.7\times10^{-3}$), IOP ($r = 0.62$, $p = 5.5\times10^{-2}$), and RNFL ($r = -0.51$, $p = 0.13$). While the RNFL correlation was not statistically significant, the trend direction was as expected—higher predicted risk associated with thinner RNFL. The Random Forest model also performed strongly, with significant positive correlations for CDR ($r = 0.81$, $p = 4.9\times10^{-3}$) and IOP ($r = 0.65$, $p = 3.4\times10^{-2}$), and a modest negative correlation with RNFL ($r = -0.61$, $p = 5.9\times10^{-2}$). These patterns demonstrate that the ensemble and margin-based models successfully capture biologically meaningful patterns associated with glaucoma pathophysiology, even in an unlabeled suspect population.

In contrast, the MLP model (Figure 5G-5I) showed weaker and more variable enrichment. While CDR showed a modest positive correlation ($r = 0.64$, $p = 4.6\times10^{-2}$), the IOP and RNFL trends were attenuated and non-significant ($r = 0.07$, $p = 0.84$; $r = -0.49$, $p = 0.15$, respectively). These results suggest that while MLP captured some relevant signal, especially with respect to optic nerve cupping, its clinical interpretability and generalizability were more limited—likely reflecting its sensitivity to sample size and model complexity in smaller training sets.

Taken together, these findings provide compelling evidence that models trained in a carefully curated POAG case–control cohort can generalize to an unlabeled, real-world suspect population. The strong enrichment for structural markers (CDR and RNFL) confirms that the predicted risks are not arbitrary but aligned with disease-relevant anatomy and physiology. This not only underscores the biological validity of our genetic classifiers but also highlights their translational potential for early risk stratification and screening in high-risk populations lacking definitive diagnoses.

Figure 5. Clinical Enrichment of Predicted Risk Scores in External Test Cohort (N = 1013)

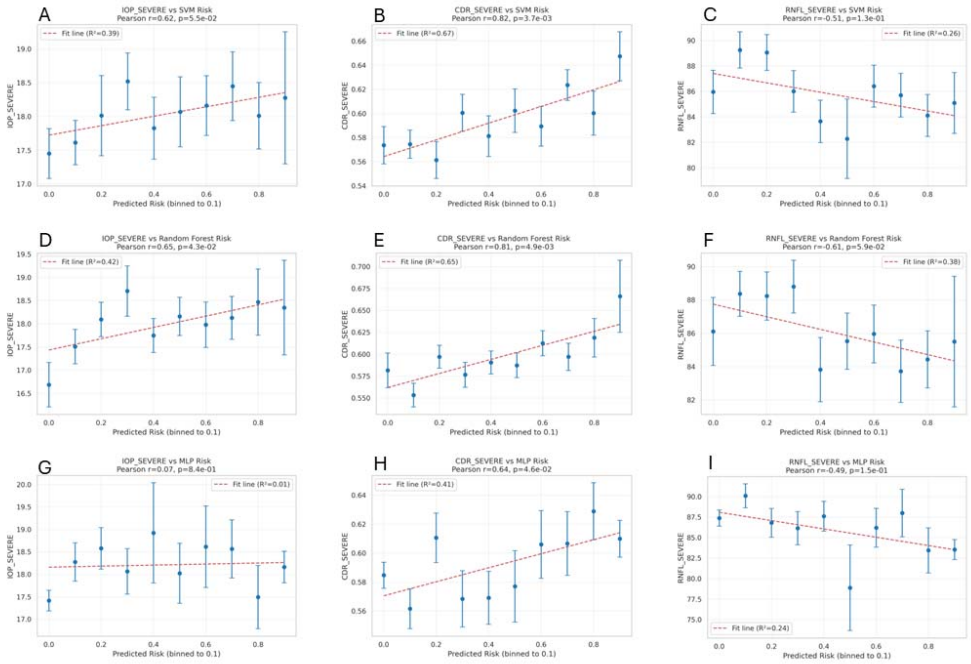


Figure 5 Caption: Predicted risk scores from SVM (A–C), Random Forest (D–F), and MLP (G–I) were positively correlated with IOP, CDR, and inversely with RNFL. Strongest enrichment was seen for CDR in RF ($r = 0.81$) and SVM ($r = 0.82$). MLP showed weaker correlations overall. These results confirm that higher predicted risk aligns with greater clinical severity, supporting the real-world utility of RF and SVM models.

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.

Across all three model types (SVM, RF, and MLP), ΔIOP stood out as the most powerful asymmetry feature, achieving AUCs of 0.87–0.89 when used alone (Figure 6). This reflects real-world clinical intuition: even a modest pressure difference between eyes can raise concern, especially if one eye remains in the

“normal” range. ΔCDR and ΔRNFL also carried predictive value, though to a lesser degree when used in isolation.

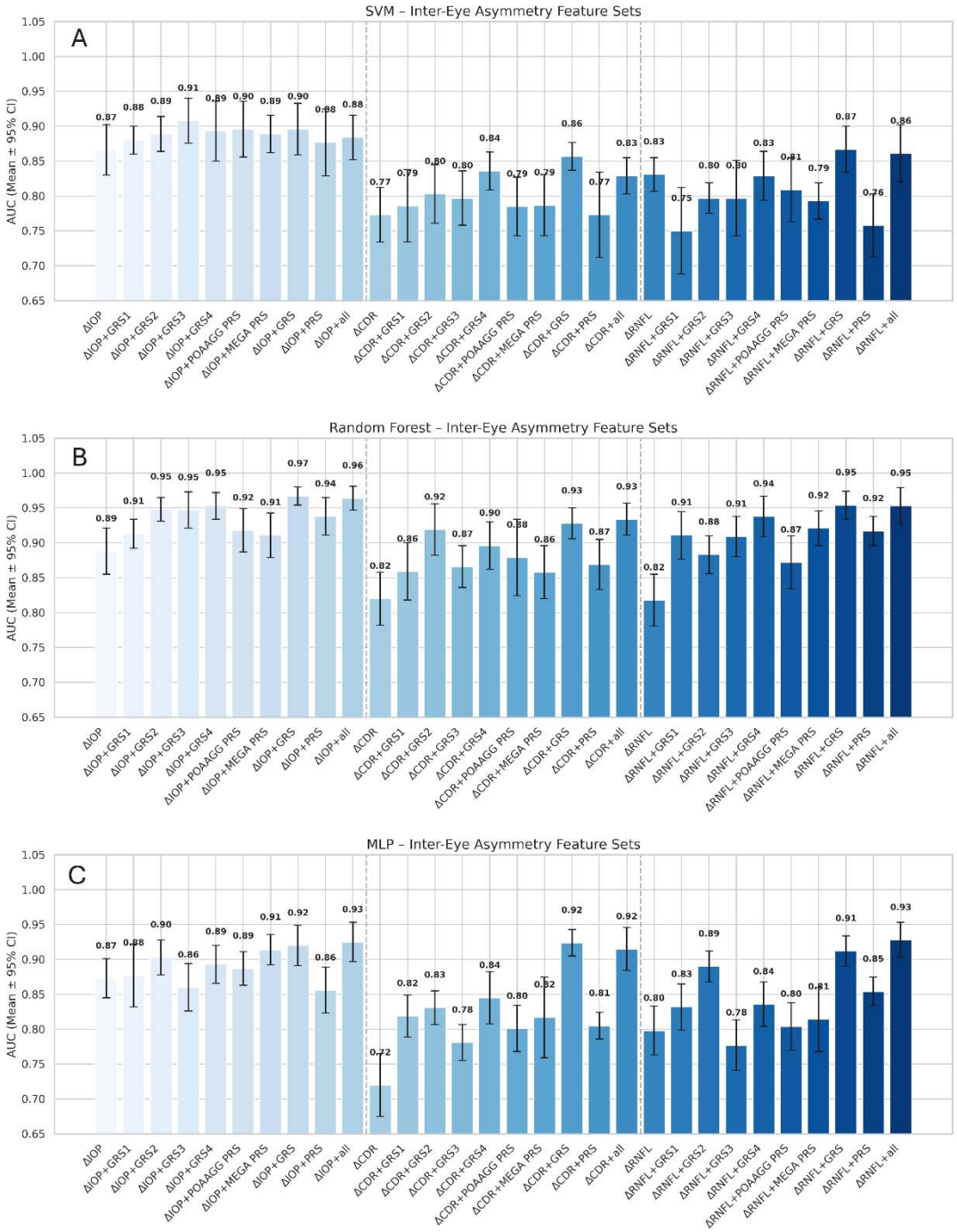
What’s particularly striking—and clinically promising—is that adding polygenic risk scores (GRS1–4 and PRS) substantially improved prediction accuracy across nearly all feature combinations. For example, combining ΔIOP with genetic scores increased the AUC of Random Forest models (Figure 6B) from 0.89 to 0.97, and that of MLP (Figure 6C) from 0.87 to 0.93. Even for weaker features like ΔRNFL, GRS inclusion lifted RF AUC from 0.82 to 0.95, and MLP from 0.80 to 0.91. These results show that GRS provide robust complementary value, especially in subtle or borderline cases.

Integration of curated GRS improved asymmetry-based prediction across all three classifiers, but the magnitude and consistency of the improvement varied by algorithm and feature set (Figure 6A–C). For SVM (Figure 6A), adding GRS to ΔIOP increased the AUC from 0.87 (ΔIOP alone) to as high as 0.91 (ΔIOP+GRS3), while ΔCDR features showed more modest improvements (0.77 → 0.86 with ΔCDR+GRS). In Random Forest (Figure 6B), the effect was more consistent: ΔIOP alone yielded AUC = 0.89, which rose to 0.97 with GRS, and ΔCDR improved from 0.82 to 0.93 when combined with GRS. For MLP (Figure 6C), ΔIOP also benefited from GRS integration (0.87 → 0.92), while ΔCDR improved from 0.72 to 0.92 (ΔCDR+GRS). Across models, the largest and most reproducible gains were observed in RF, followed by MLP, with SVM showing smaller but still meaningful improvements. These results suggest that nonlinear models—particularly RF—are better suited to capture the interaction between phenotypic asymmetry and distributed genetic risk.

Clinically, these findings are highly relevant. Many patients present with asymmetric findings that may not cross diagnostic thresholds but still raise suspicion. Our results suggest that integrating genetic scores with asymmetry features can push these borderline cases into clearer diagnostic territory, offering a more confident risk stratification tool for clinicians. This is especially useful in populations with limited access to longitudinal follow-up or advanced imaging.

In summary, asymmetry features reflect the earliest clinical signs of glaucoma, and when paired with polygenic risk scores, they unlock some of the highest predictive performance observed in our study. This highlights a promising pathway toward early detection, especially in individuals who may not yet meet traditional diagnostic criteria but are genetically predisposed. Combining these two perspectives—clinical observation and genetic liability—brings us closer to personalized, precision-based glaucoma care.

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



(A) SVM models trained on inter-eye asymmetry features (ΔIOP, ΔCDR, ΔRNFL) with or without genetic risk scores (GRS1–4, POAAGG PRS, MEGA PRS). ΔIOP-based models performed best (AUC up to 0.91), with smaller but consistent gains from GRS integration.

(B) RF models using the same asymmetry feature sets. RF achieved the strongest overall performance, with ΔIOP models reaching AUC = 0.95–0.97 and ΔCDR models improving from 0.82 to 0.93 after adding GRS.

(C) MLP models with asymmetry features and genetic predictors. ΔIOP-based models improved from AUC = 0.87 (ΔIOP alone) to 0.92 with GRS, while ΔCDR increased from 0.72 to 0.92 when combined with GRS.

DISCUSSION

First, our results underscore the robust predictive utility of *curated GRS*, particularly GRS 2 and GRS 4, which consistently outperformed both PRS and the baseline demographic model across all three classifiers. 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 0.77. 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. In this context, the ability of GRS 4 alone to reach an AUC of 0.74—nearly matching the base model—underscores the biological strength and potential of GRS as an interpretable and actionable genomic feature.. GRS 2 and GRS 4 were designed to capture complementary facets of glaucoma genetic architecture: GRS 2 incorporated sub-threshold loci from an African-ancestry case-control GWAS, while GRS 4 drew from a quantitative IOP-based MTAG meta-analysis. The high discriminative power of these two scores likely reflects their capacity to capture both disease susceptibility and subclinical endophenotypes. Furthermore, when all four GRS were integrated (base+GRS 1–4), the average AUC climbed to 0.93—surpassing the performance of both ancestry-matched PRS (base+POAAGG PRS or base+MEGA PRS, AUC=0.82) and combined PRS models (base+PRS, AUC=0.83). This reinforces the idea that a small number of biologically curated, ancestry-specific loci may carry disproportionate predictive value in well-characterized cohorts. Overall, these findings provide compelling evidence that curated GRS are not only statistically robust but also clinically meaningful. Their performance rivals and even exceeds demographic baselines, affirming their potential role in early-stage screening pipelines and justifying their further development for translational glaucoma risk assessment.

Second, this work underscores 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 harbor greater genomic diversity and often exhibit distinct linkage disequilibrium structures that are not captured by Eurocentric GWAS. In our study, standard PRS models underperformed unless recalibrated using ancestry-matched GWAS summary statistics. By integrating POAAGG-specific PRS and curating GRS from African-ancestry glaucoma loci, we demonstrated substantial improvements in predictive accuracy and fairness. 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. Both RF and SVM 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, GRS 2, and GRS 4 as consistently top-ranked predictors—which enhances transparency and supports clinical trust and regulatory adoption. In contrast, MLP required richer inputs, such as combined genetic scores, to achieve comparable AUCs, reflecting its greater data dependency. 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 + GRS 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 GRS 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 GRS 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 GRS offers a low-barrier, ancestry-matched genomic predictor that can forecast future risk independent of immediate clinical measures. The fact that GRS 2 and GRS 4 approach or exceed the predictive strength of traditional risk factors like age demonstrates their utility as standalone markers. 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 GRS 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 GRS 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 GRS/PRS development and promoted equity in precision medicine, it limits generalizability to other populations. Given known differences in genetic architecture, GRS/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 disease.

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 GRS were built from fixed SNP sets derived from African ancestry GWAS, including GRS 2 and GRS 4, 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 GRS 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 GRS 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 publicly available.

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

Supplemental tables and figures will be added in a future version.

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 ancestry-matched PRS and curate GRS from African ancestry-specific glaucoma GWAS.

Collected variables included demographic information (age, sex, family history), 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.

Sex 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 sex-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.

METHODS

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, sex) 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; family history: 1 = yes, 0 = no), 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 10 feature sets, ranging from isolated domains (e.g., base features or GRS alone) to fully integrated inputs (e.g., base + GRS + PRS). Key feature sets included:

- (1) base = age + gender
- (2–5) base + GRS 1–4
- (6) base + POAAGG PRS
- (7) base + MEGA PRS
- (8) base + GRS 1-4
- (9) base + POAAGG PRS + MEGA PRS
- (10) base + all GRS + all 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 public repository.

2. Genetic Risk Score (GRS) and Polygenic Risk Score (PRS) Construction

To incorporate genetic susceptibility into our glaucoma prediction models, we computed both curated Genetic Risk Scores (GRS 1–4)[54,55] and genome-wide Polygenic Risk Scores (POAAGG PRS and MEGA PRS).[8,56] These scores were derived using SNP-level genotype data obtained from the POAAGG and MEGA arrays[56], following rigorous quality control and standardization procedures.

We constructed four Genetic Risk Scores (GRS 1–4) using hand-selected single nucleotide polymorphisms (SNPs) derived from African ancestry-specific GWAS datasets, as reported in our previous publication. [11] All SNPs were mapped to hg38 coordinates.

- GRS 1 was derived from the 46 genome-wide significant loci ($p < 5\times10^{-}$□) identified in the MEGA-based binary case-control GWAS of POAG among African ancestry individuals.
- GRS 2 expanded this to 174 loci by including suggestive associations ($p < 5\times10^{-}$□) from the same study, incorporating additional variants that may reflect moderate or polygenic effects.
- GRS 3 and GRS 4 were based on a multi-trait meta-analysis (MTAG) using IOP as a quantitative trait, capturing endophenotypic variation related to glaucoma pathophysiology. GRS 3 included 42 genome-wide significant loci, while GRS 4 incorporated a broader set of 190 SNPs using a relaxed p-value threshold.

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:

$$GRS_{weighted} = \sum_{i=1}^{n} \beta i G i$$

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. SNP inclusion across GRS 1–4 was visualized using customized Manhattan-style plots and chromosome-wise bar plots (see Figure 3A–B), and their distributions across disease status groups were compared using violin plots (Figure 3C).

In parallel with curated GRS, we generated two genome-wide PRS:

- MEGA PRS was constructed using GWAS summary statistics from the MEGA African ancestry meta-analysis (N ≈ 14,005).

- POAAGG PRS was derived from the internal POAAGG GWAS dataset (N ≈ 7,508 African American participants), providing a discovery cohort closely matched to the target population.

Both scores were calculated using the PRS-CS (Bayesian regression with continuous shrinkage priors) framework. Input summary statistics were preprocessed to match the LD reference panel [46] (African-ancestry LD from the 1000 Genomes Project), and default global shrinkage parameters were used.

All GRS and PRS scores were merged with individual-level phenotype data using PLINK 2.0 (--score) and custom Python scripts.[47,57] To mitigate distributional skewness and enable comparability across models, all genetic scores were standardized using rank-based inverse normal transformation (z-score normalization) prior to model input.

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 “all” feature set, which comprised demographic variables (age, gender), and genetic predictors (GRS 1–4, POAAGG PRS, 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 “all” feature set and the “base+GRS” model (Figure 4), 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 GRS and 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.

811 18. Roberts, E., Howell, S., and Evans, D.G. (2023). Polygenic risk scores and breast cancer risk
812 prediction. The Breast *67*, 71-77.
813 19. Mavaddat, N., Michailidou, K., Dennis, J., Lush, M., Fachal, L., Lee, A., Tyrer, J.P., Chen, T.-H.,
814 Wang, Q., and Bolla, M.K. (2019). Polygenic risk scores for prediction of breast cancer and breast
815 cancer subtypes. The American Journal of Human Genetics *104*, 21-34.
816 20. Qassim, A., Souzeau, E., Siggs, O.M., Hassall, M.M., Han, X., Griffiths, H.L., Frost, N.A., Vallabh,
817 N.A., Kirwan, J.F., and Menon, G. (2020). An intraocular pressure polygenic risk score stratifies
818 multiple primary open-angle glaucoma parameters including treatment intensity.
819 Ophthalmology *127*, 901-907.
820 21. Gao, X.R., Huang, H., and Kim, H. (2019). Polygenic risk score is associated with intraocular
821 pressure and improves glaucoma prediction in the UK Biobank cohort. Translational vision
822 science & technology *8*, 10-10.
823 22. Han, X., Hewitt, A.W., and MacGregor, S. (2021). Predicting the future of genetic risk profiling of
824 glaucoma: a narrative review. JAMA ophthalmology *139*, 224-231.
825 23. Wang, Y., Kanai, M., Tan, T., Kamariza, M., Tsuo, K., Yuan, K., Zhou, W., Okada, Y., Huang, H., and
826 Turley, P. (2023). Polygenic prediction across populations is influenced by ancestry, genetic
827 architecture, and methodology. Cell Genomics *3*.
828 24. Lambert, S.A., Abraham, G., and Inouye, M. (2019). Towards clinical utility of polygenic risk
829 scores. Human molecular genetics *28*, R133-R142.
830 25. Sekimitsu, S., Xiang, D., Smith, S.L., Curran, K., Elze, T., Friedman, D.S., Foster, P.J., Luo, Y.,
831 Pasquale, L.R., and Peto, T. (2023). Deep ocular phenotyping across primary open-angle
832 glaucoma genetic burden. JAMA ophthalmology *141*, 891-899.
833 26. Rajpurkar, P., Chen, E., Banerjee, O., and Topol, E.J. (2022). AI in health and medicine. Nature
834 medicine *28*, 31-38.
835 27. Wang, F., and Preininger, A. (2019). AI in health: state of the art, challenges, and future
836 directions. Yearbook of medical informatics *28*, 016-026.
837 28. Ngiam, K.Y., and Khor, W. (2019). Big data and machine learning algorithms for health-care
838 delivery. The Lancet Oncology *20*, e262-e273.
839 29. Beam, A.L., and Kohane, I.S. (2018). Big data and machine learning in health care. Jama *319*,
840 1317-1318.
841 30. Zhu, Y., Salowe, R., Chow, C., Li, S., Bastani, O., and O'Brien, J.M. (2024). Advancing glaucoma
842 care: integrating artificial intelligence in diagnosis, management, and progression detection.
843 Bioengineering *11*, 122.
844 31. Zeppieri, M., Gardini, L., Culiersi, C., Fontana, L., Musa, M., D'Esposito, F., Surico, P.L., Gagliano,
845 C., and Sorrentino, F.S. (2024). Novel Approaches for the Early Detection of Glaucoma Using
846 Artificial Intelligence. Life *14*, 1386.
847 32. Alipanahi, B., Hormozdiari, F., Behsaz, B., Cosentino, J., McCaw, Z.R., Schorsch, E., Sculley, D.,
848 Dorfman, E.H., Foster, P.J., and Peng, L.H. (2021). Large-scale machine-learning-based
849 phenotyping significantly improves genomic discovery for optic nerve head morphology. The
850 American Journal of Human Genetics *108*, 1217-1230.
851 33. Han, X., Steven, K., Qassim, A., Marshall, H.N., Bean, C., Tremeer, M., An, J., Siggs, O.M.,
852 Gharahkhani, P., and Craig, J.E. (2021). Automated AI labeling of optic nerve head enables
853 insights into cross-ancestry glaucoma risk and genetic discovery in> 280,000 images from UKB
854 and CLSA. The American Journal of Human Genetics *108*, 1204-1216.
855 34. Truong, B., Hull, L.E., Ruan, Y., Huang, Q.Q., Hornsby, W., Martin, H., Van Heel, D.A., Wang, Y.,
856 Martin, A.R., and Lee, S.H. (2024). Integrative polygenic risk score improves the prediction
857 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

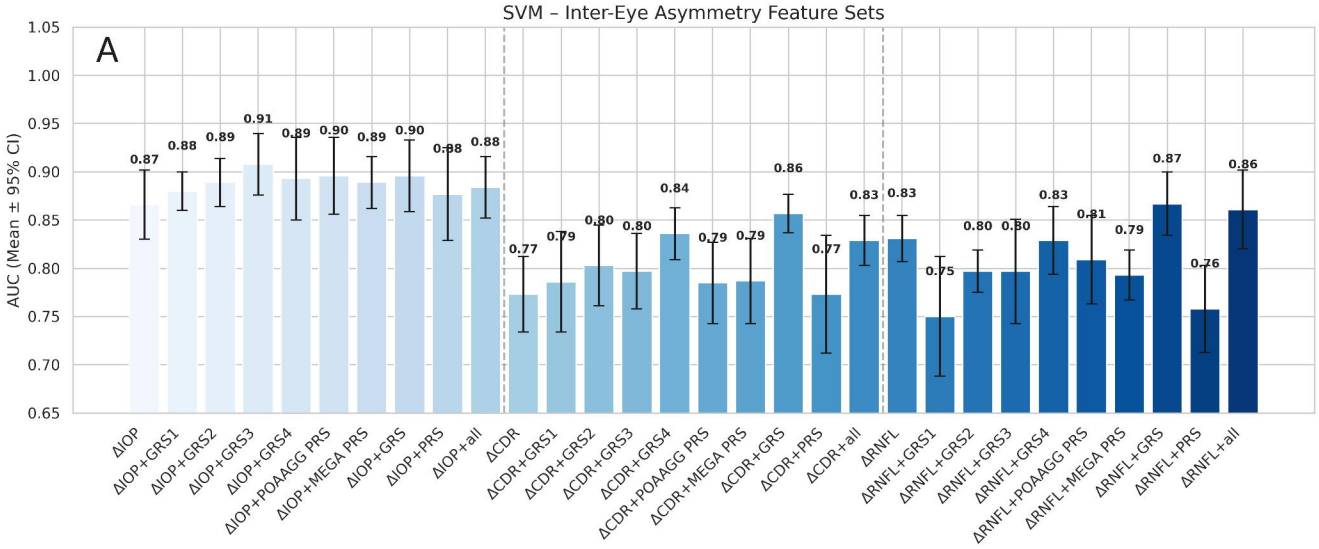
identify individuals with risk equivalent to monogenic mutations. Nature genetics *50*, 1219-1224.

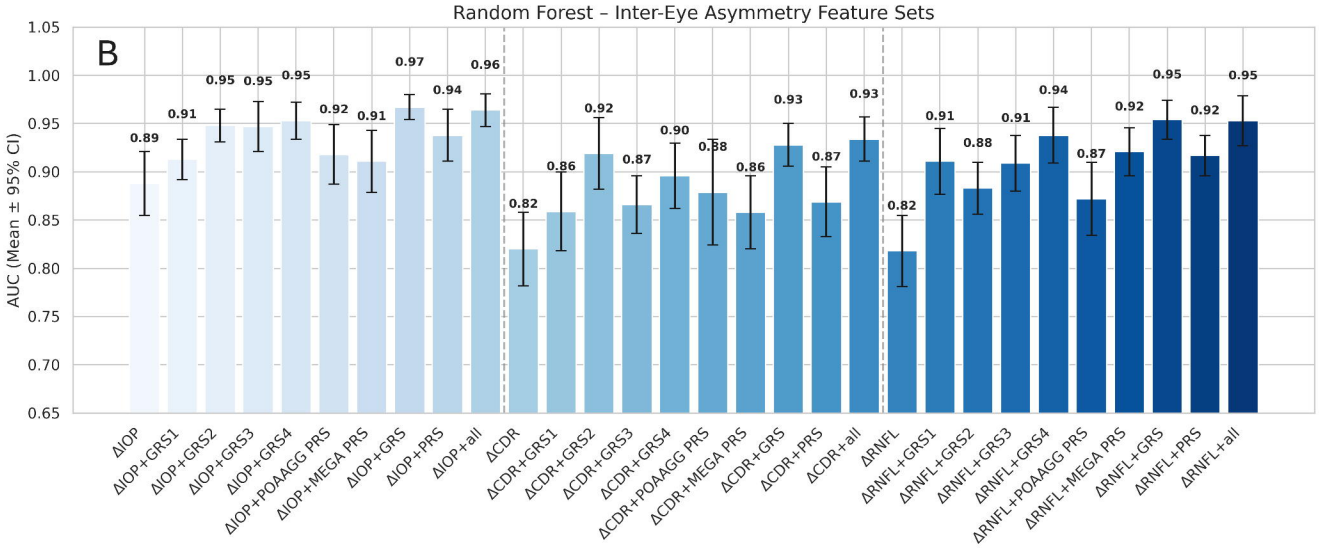
56. Sirugo, G., Williams, S.M., and Tishkoff, S.A. (2019). The missing diversity in human genetic studies. Cell *177*, 26-31.

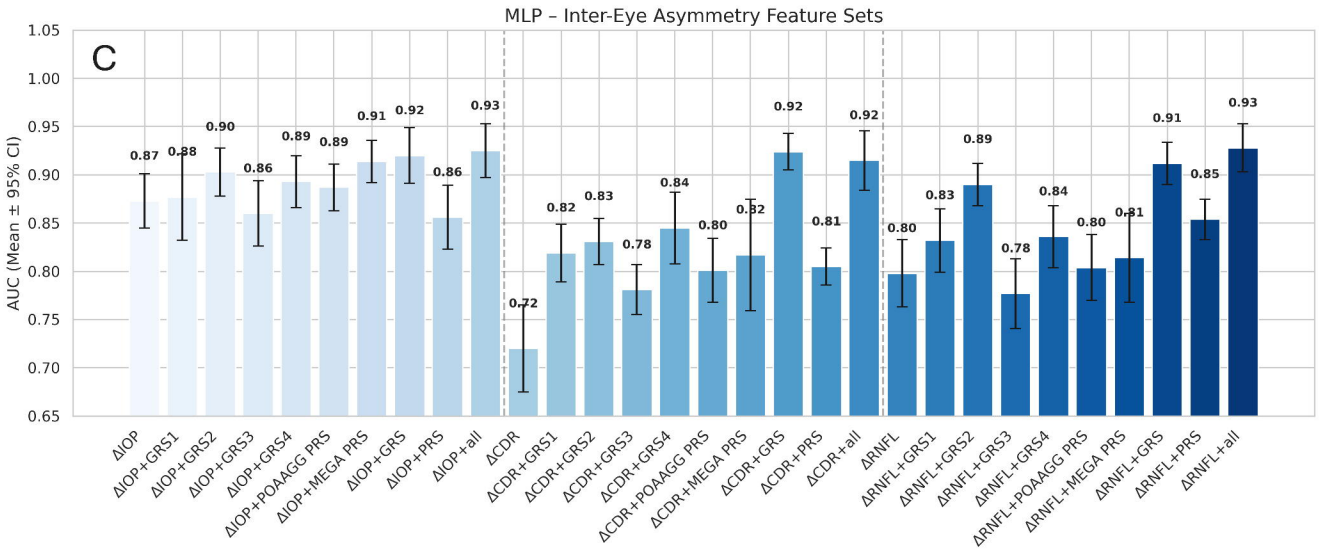
57. Van Rossum, G., and Drake Jr, F.L. (1995). Python tutorial (Centrum voor Wiskunde en Informatica Amsterdam, The Netherlands).

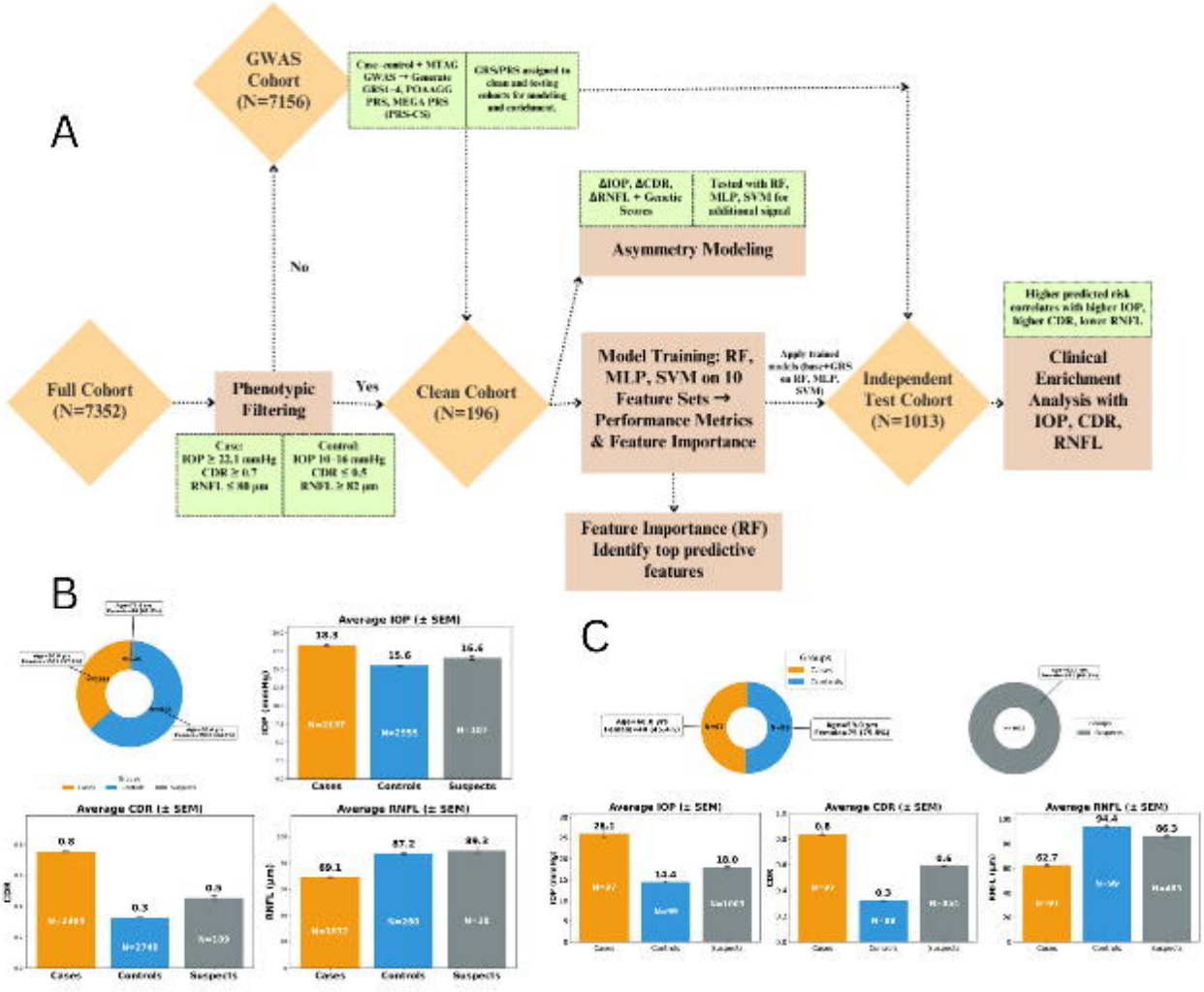
58. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., and Dubourg, V. (2011). Scikit-learn: Machine learning in Python. the Journal of machine Learning research *12*, 2825-2830.

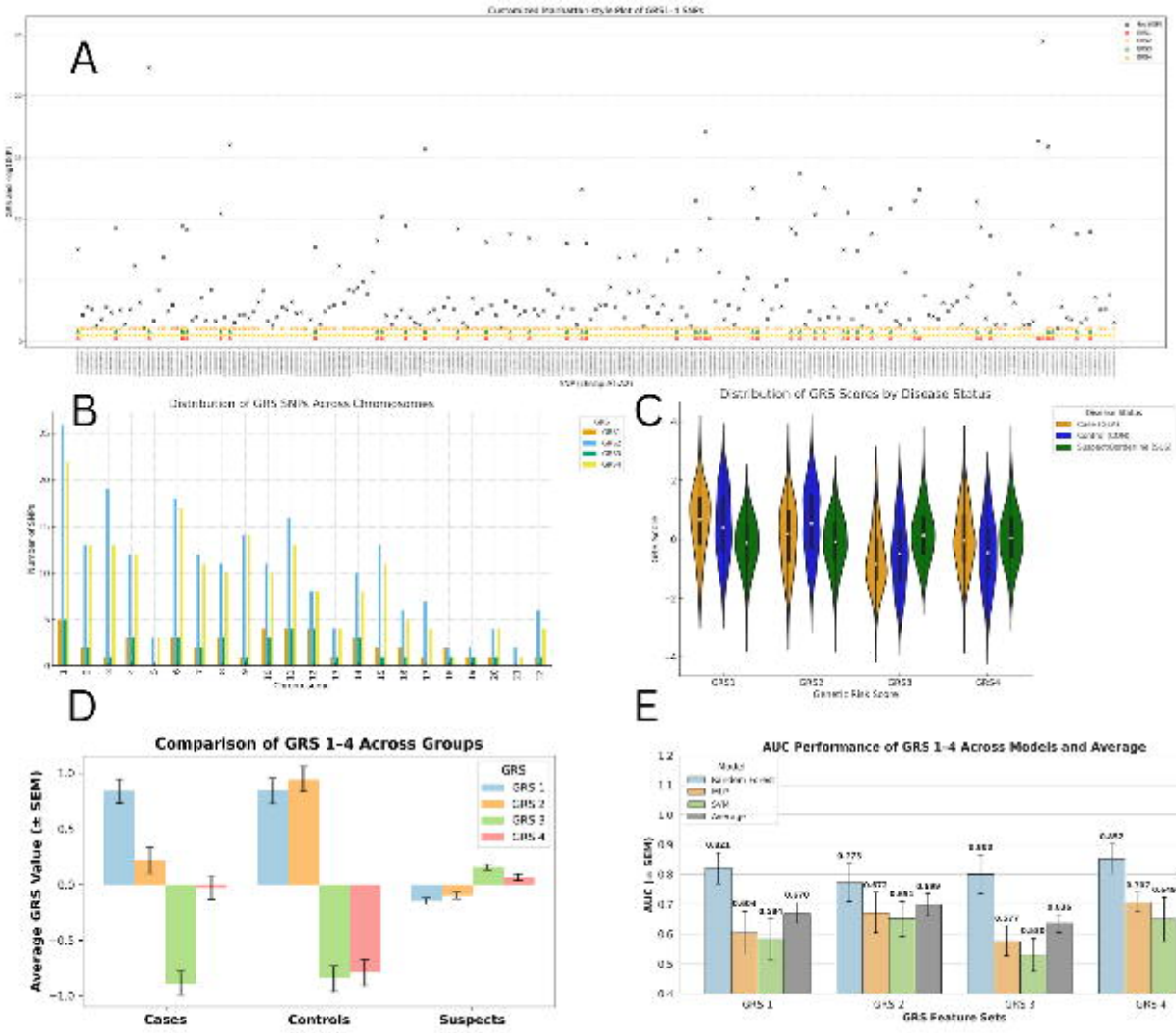
59. Efron, B., and Tibshirani, R.J. (1994). An introduction to the bootstrap (Chapman and Hall/CRC).

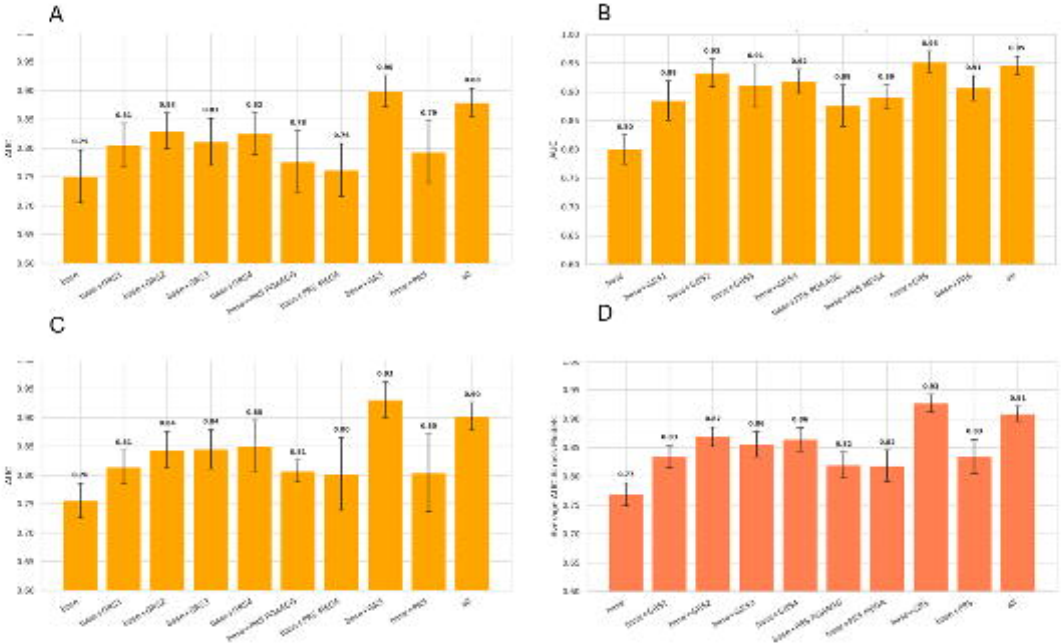

SVM – Inter-Eye Asymmetry Feature Sets
A
AUC (Mean ± 95% CI)
1.05
1.00
0.95
0.90
0.85
0.80
0.75
0.70
0.65
0.87
0.88
0.89
0.91
0.89
0.90
0.89
0.90
0.88
0.88
0.77
0.79
0.80
0.80
0.84
0.79
0.79
0.86
0.77
0.83
0.83
0.75
0.80
0.80
0.83
0.81
0.79
0.87
0.76
0.86
ΔIOP
ΔIOP+GRS1
ΔIOP+GRS2
ΔIOP+GRS3
ΔIOP+GRS4
ΔIOP+POAAGG PRS
ΔIOP+MEGA PRS
ΔIOP+GRS
ΔIOP+PRS
ΔIOP+all
ΔCDR
ΔCDR+GRS1
ΔCDR+GRS2
ΔCDR+GRS3
ΔCDR+GRS4
ΔCDR+POAAGG PRS
ΔCDR+MEGA PRS
ΔCDR+GRS
ΔCDR+PRS
ΔCDR+all
ΔRNFL
ΔRNFL+GRS1
ΔRNFL+GRS2
ΔRNFL+GRS3
ΔRNFL+GRS4
ΔRNFL+POAAGG PRS
ΔRNFL+MEGA PRS
ΔRNFL+GRS
ΔRNFL+PRS
ΔRNFL+all


Random Forest – Inter-Eye Asymmetry Feature Sets
B
AUC (Mean ± 95% CI)
1.05
1.00
0.95
0.90
0.85
0.80
0.75
0.70
0.65
0.89
0.91
0.95
0.95
0.95
0.92
0.91
0.97
0.94
0.96
0.82
0.86
0.92
0.87
0.90
0.88
0.86
0.93
0.87
0.93
0.82
0.91
0.88
0.91
0.94
0.87
0.92
0.95
0.92
0.95
ΔIOP
ΔIOP+GRS1
ΔIOP+GRS2
ΔIOP+GRS3
ΔIOP+GRS4
ΔIOP+POAAGG PRS
ΔIOP+MEGA PRS
ΔIOP+GRS
ΔIOP+PRS
ΔIOP+all
ΔCDR
ΔCDR+GRS1
ΔCDR+GRS2
ΔCDR+GRS3
ΔCDR+GRS4
ΔCDR+POAAGG PRS
ΔCDR+MEGA PRS
ΔCDR+GRS
ΔCDR+PRS
ΔCDR+all
ΔRNFL
ΔRNFL+GRS1
ΔRNFL+GRS2
ΔRNFL+GRS3
ΔRNFL+GRS4
ΔRNFL+POAAGG PRS
ΔRNFL+MEGA PRS
ΔRNFL+GRS
ΔRNFL+PRS
ΔRNFL+all

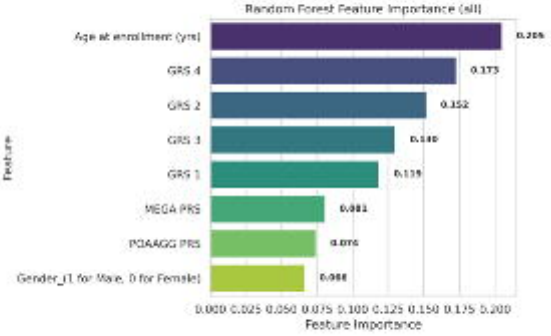

MLP – Inter-Eye Asymmetry Feature Sets
C
AUC (Mean ± 95% CI)
1.05
1.00
0.95
0.90
0.85
0.80
0.75
0.70
0.65
0.87
0.88
0.90
0.86
0.89
0.89
0.91
0.92
0.86
0.93
0.72
0.82
0.83
0.78
0.84
0.80
0.82
0.92
0.81
0.92
0.80
0.83
0.89
0.78
0.84
0.80
0.81
0.91
0.85
0.93
ΔIOP
ΔIOP+GRS1
ΔIOP+GRS2
ΔIOP+GRS3
ΔIOP+GRS4
ΔIOP+POAAGG PRS
ΔIOP+MEGA PRS
ΔIOP+GRS
ΔIOP+PRS
ΔIOP+all
ΔCDR
ΔCDR+GRS1
ΔCDR+GRS2
ΔCDR+GRS3
ΔCDR+GRS4
ΔCDR+POAAGG PRS
ΔCDR+MEGA PRS
ΔCDR+GRS
ΔCDR+PRS
ΔCDR+all
ΔRNFL
ΔRNFL+GRS1
ΔRNFL+GRS2
ΔRNFL+GRS3
ΔRNFL+GRS4
ΔRNFL+POAAGG PRS
ΔRNFL+MEGA PRS
ΔRNFL+GRS
ΔRNFL+PRS
ΔRNFL+all


A
GWAS Cohort (N=7156)
Case–control + MTAG GWAS → Generate GRS1–4, POAAGG PRS, MEGA PRS (PRS-CS)
GRS/PRS assigned to clean and testing cohorts for modeling and enrichment.
ΔIOP, ΔCDR, ΔRNFL + Genetic Scores
Tested with RF, MLP, SVM for additional signal
Asymmetry Modeling
No
Full Cohort (N=7352)
Phenotypic Filtering
Yes
Clean Cohort (N=196)
Case: IOP ≥ 22.1 mmHg CDR ≥ 0.7 RNFL ≤ 80 µm
Control: IOP 10–16 mmHg CDR ≤ 0.5 RNFL ≥ 82 µm
Model Training: RF, MLP, SVM on 10 Feature Sets → Performance Metrics & Feature Importance
Apply trained models (base+GRS on RF, MLP, SVM)
Independent Test Cohort (N=1013)
Higher predicted risk correlates with higher IOP, higher CDR, lower RNFL
Clinical Enrichment Analysis with IOP, CDR, RNFL
Feature Importance (RF) Identify top predictive features
B
Average IOP (± SEM)
18.5
15.6
16.6
Cases
Controls
Suspects
Average CDR (± SEM)
0.8
0.3
0.5
Average RNFL (± SEM)
89.1
87.2
86.2
C
Groups
Average IOP (± SEM)
28.1
13.4
16.0
Average CDR (± SEM)
0.8
0.3
0.6
Average RNFL (± SEM)
62.7
94.4
86.5

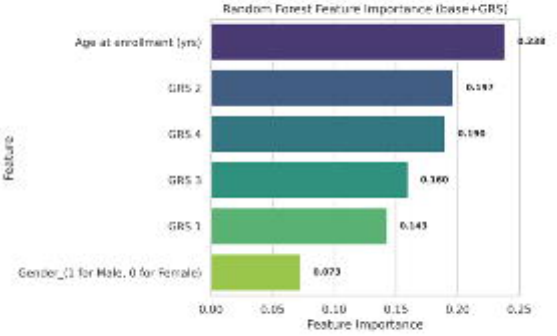

A
B
Distribution of GRS SNPs Across Chromosomes
C
Distribution of GRS Scores by Disease Status
GRS1
GRS2
GRS3
GRS4
Genetic Risk Score
D
Comparison of GRS 1-4 Across Groups
Average GRS Value (± SEM)
GRS
GRS 1
GRS 2
GRS 3
GRS 4
Cases
Controls
Suspects
E
AUC Performance of GRS 1-4 across Models and Average
AUC (± SEM)
GRS 1
GRS 2
GRS 3
GRS 4
GRS Feature Sets

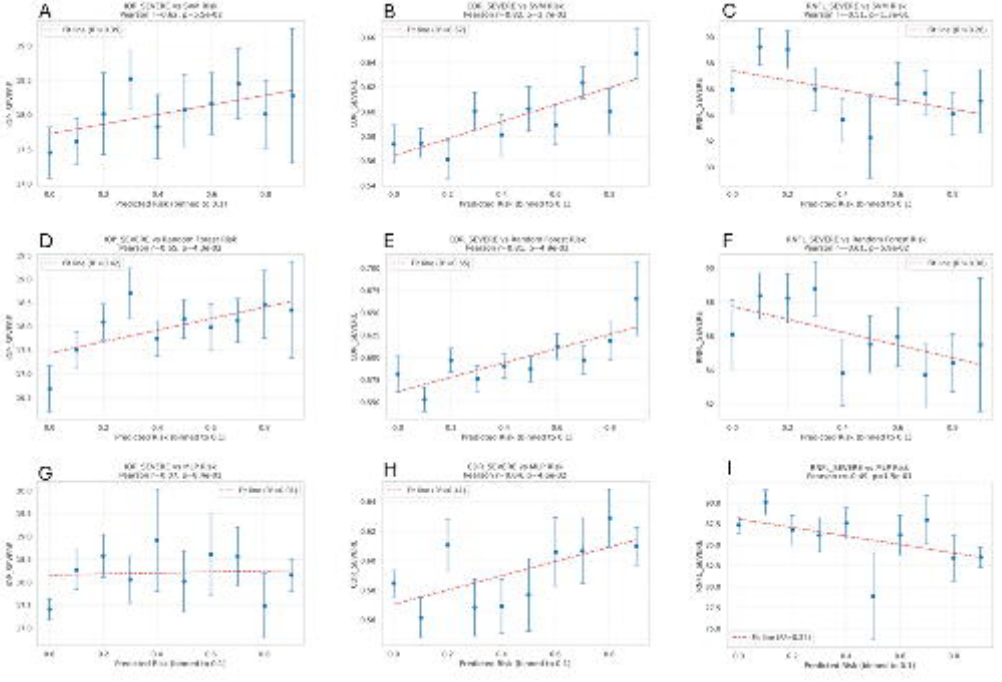

A
B
C
D

A


Random Forest Feature Importance (all)
Age at enrollment (yrs)
GRS 4
GRS 2
GRS 3
GRS 1
MEGA PRS
PGAAGG PRS
Gender_(1 for Male, 0 for Female)
0.206
0.175
0.152
0.140
0.119
0.081
0.076
0.068
Feature
0.000 0.025 0.050 0.075 0.100 0.125 0.150 0.175 0.200
Feature Importance

B


Random Forest Feature Importance (base+GRS)
Age at enrollment (yrs)
GRS 2
GRS 4
GRS 3
GRS 1
Gender_(1 for Male, 0 for Female)
0.240
0.197
0.190
0.160
0.143
0.073
Feature
0.00 0.05 0.10 0.15 0.20 0.25
Feature Importance


A
B
C
D
E
F
G
H
I