Optimization of Multi-Ancestry Polygenic Risk Score Disease Prediction Models

Authors: Jon Lerga-Jaso[1], Alex Osama[1], Charlie Manson[1], Andrew Terpolovsky[1], Biljana Novković[1], Sandra Bohn[1], Adriano De Marino[1], Mark Kunitomi[2], Puya G. Yazdi[1,†]

Affiliation:

[1]Research & Development, Omics Edge, Miami, FL, USA

[2]Almaden Genomics, San Francisco, CA, USA

†Corresponding author. Email: pyazdi@omicsedge.com

Abstract

Polygenic risk scores (PRS) have ushered in a new era in genetic epidemiology, offering insights into individual predispositions to a wide range of diseases. This study aimed to develop and benchmark multi-ancestry PRS models capable of predicting disease risk across diverse populations. Leveraging trans-ethnic GWAS meta-analysis, we generated novel summary statistics for 30 medically-related traits and assessed the predictive performance of four PRS algorithms. Algorithm efficacy across traits varied, with Stacked C+T (SCT) performing better when trait prevalence was higher, while LDpred and Lassosum performed better when trait prevalence was lower. Subsequent integration of PRS algorithm outputs through logistic regression and incorporation of additional medical information enhanced model accuracy. Notably, including ancestry information further improved predictive performance, underscoring its importance in PRS model development. Our findings highlight the potential of multi-ancestry PRS models in clinical settings, demonstrating superior predictive accuracy and broad applicability across different ancestral backgrounds.

Introduction

Polygenic risk scores (PRS) have emerged as a transformative tool in genetic epidemiology, harnessing the wealth of data generated by genome-wide association studies (GWAS) to predict an individual's predisposition to complex diseases[1]. By aggregating the effects of numerous genetic variants, each contributing small to modest effects, PRS offers an understanding of genetic susceptibility across a spectrum of human diseases, including cardiovascular conditions, psychiatric disorders, and cancers[2,3,4,5,6]. The utility of PRS extends beyond mere risk prediction; it holds promise for personalized medicine, where interventions can be tailored based on an individual's genetic risk profile. Moreover, PRS can enhance disease screening strategies, inform clinical decision-making, and potentially guide lifestyle and therapeutic interventions aimed at

mitigating disease risk[1,3,7,8,9]. Despite this potential, the application of PRS in clinical settings is in its nascent stages, grappling with challenges such as improving the accuracy and interpretability of scores, ensuring equitable performance across diverse populations, and integrating genetic risk information with environmental and lifestyle factors for a comprehensive approach to disease prediction and prevention[1,10,11].

Recent advancements in computational methods, alongside the exponential growth in GWAS sample sizes, have markedly enhanced the predictive power of PRS[8,12]. However, developing and validating PRS models that are generalizable across different ancestries remains imperative, as current models often exhibit reduced efficacy in non-European populations[1,8]. Additionally, any PRS used in a clinical setting must demonstrate clinical utility and easy interpretability that will alter patient care decisions based on the results[13].

With those aims in mind, we set about to create and validate multi-ancestry PRS that incorporate diverse clinical characteristics. We treated this task as a binary classification problem and systematically tested different combinations of inputs to make disease predictions that are easily interpretable and would alter physician intervention decisions. We build and evaluate different PRS models for 30 different medically-related traits starting from assessing the benefit of GWAS meta-analysis, before adding different types of inputs and risk factors to our models to maximize their predictive power.

Results

Trans-ethnic GWAS Meta-analysis and PRS Algorithm Benchmarking

We generated novel genome-wide summary statistics for 30 medically-related traits, leveraging trans-ancestry genome-wide association study (GWAS) meta-analysis of diverse cohort data, including the UK Biobank (Table 1)[14,15,16,17,18,19,20,21,22,23,24,25,26,27,28,29,30,31,32,33,34,35,36,37,38,39,40,41,42,43,44,45,46,47,48,49,50,51,52,53,54,55,56,57,58,59,60,61,62,63,64,65,66,67,68,69,70,79]. The following traits were selected: age-related macular degeneration, Alzheimer's disease, asthma, atrial fibrillation, bipolar disorder, body mass index (BMI), bowel cancer, breast cancer, cardiovascular disease, celiac disease, coronary artery disease, Crohn's disease, hypertension, ischemic stroke, melanoma, multiple sclerosis, osteoporosis, Parkinson's disease, primary open angle glaucoma, prostate cancer, psoriasis, rheumatoid arthritis, schizophrenia, systemic lupus erythematosus, total cholesterol, triglycerides, type 1 diabetes, type 2 diabetes, ulcerative colitis, venous thromboembolic disease. For the purpose of PRS training, the three continuous traits, BMI, total cholesterol and triglycerides, were converted into binary outcomes, obesity, hypercholesterolemia, and hypertriglyceridemia, respectively.

GWAS meta-analyses were conducted using METAL[71], incorporating publicly available external GWAS sources, predominantly sourced from the GWAS Catalog[72]. We kept out a set of 104,604

individuals from the UK Biobank (UKB) to serve as a training and testing cohort for PRS prediction and downstream validations. The cohort was composed of 76,998 in the training and 27,606 in the testing set. We adopted the approach described in a recent preprint by Thompson et al. (2022)[8], which aimed to maximize the representation of non-European ancestries in this UKB subset of samples. The remainder of the samples, previously classified as the White British Unrelated (WBU) subset in the UKB project[14], were utilized for the GWAS component of the meta-analysis. This strategy facilitated direct comparisons of our results against the PRS developed by Thompson et al. using the same testing set (see Methods section for more details).

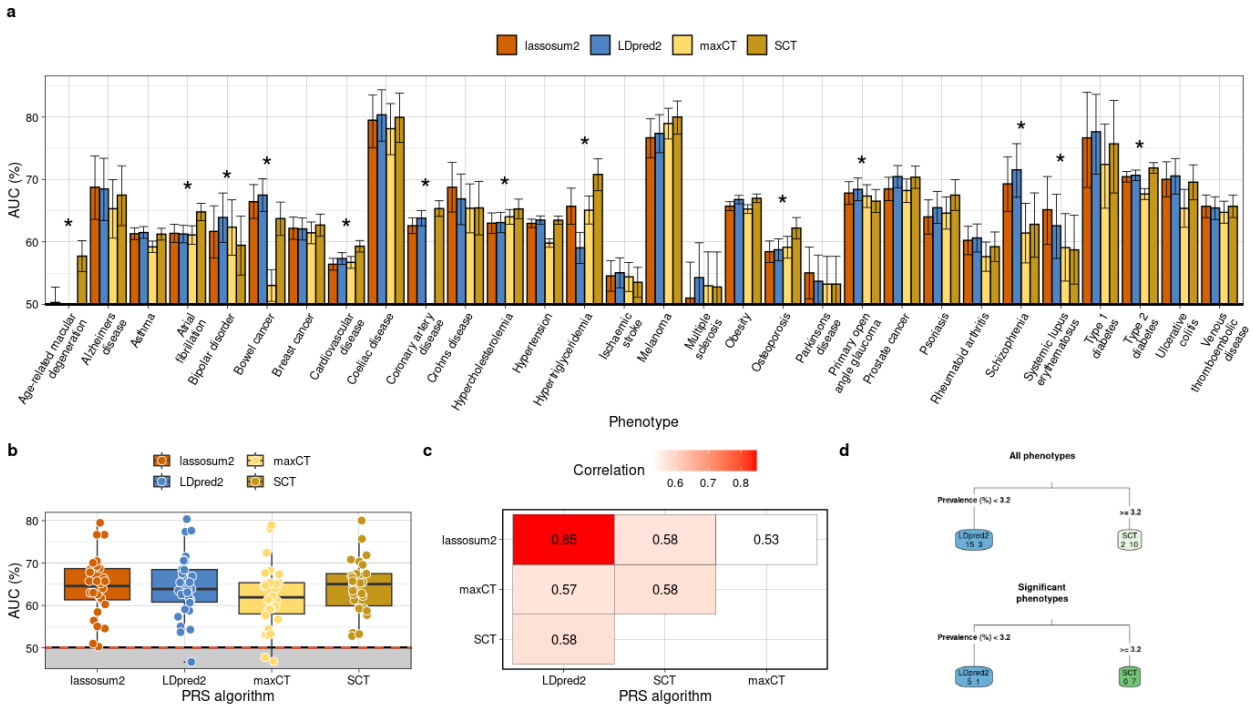

Figure 1. Comparison of four PRS prediction algorithms applied to 30 medically-related traits. a) Prediction accuracy using area under the curve (AUC) as a performance metric in the testing set of UKB. Conditions marked with an asterisk indicate those where LDpred2 or lassosum2 were significantly better at risk prediction than Stacked C+T (SCT) or vice versa. Error bars present the mean and 95% confidence interval of 10,000 non-parametric bootstrap replicates. **b)** Prediction results summarized per algorithm. **c)** Pearson correlation of PRS scores across the 30 phenotypes between pairs of algorithms. **d)** Decision trees indicating disease prevalence as the main factor to distinguish between LDpred2 and SCT top performance among all phenotypes (top) or where accuracy results were significantly better for a specific method (bottom).

To benchmark predictive performance and elucidate each method's strengths and weaknesses, we trained four PRS algorithms: lassosum2[73], LDpred2[74], maxCT and Stacked C+T (SCT)[75]. PRS performance was evaluated using the area under the receiver operating characteristic curve (AUC), with the best results for each phenotype ranging from 80.4% (coeliac disease) to 54.3% (multiple sclerosis), and a median of 65.5%. Results across different phenotypes displayed

correlations for different methods (Figure 1a). Notably, differences were observed for 13 traits: SCT outperformed other models in eight medical traits, LDpred2 excelled in four, and lassosum2 in one. Overall, SCT demonstrated superior performance (Figure 1b). Pairwise correlations of PRS scores generated by each algorithm across all phenotypes revealed high correlations between lassosum2 and LDpred2, whereas correlations with SCT were less pronounced. We hypothesized that integrating insights from different models could lead to enhanced predictive accuracy. Moreover, we identified trait prevalence as a primary factor influencing accuracy, with LDpred2 (or lassosum2) showing better performance in conditions with lower prevalence and SCT performing optimally in higher prevalence settings (Figure 1d).

Predictive Performance of PRS

Building on the idea above of integrating outputs from multiple PRS algorithms, we retrained a PRS model using logistic regression incorporating the scores previously obtained from SCT, LDpred2 and lassosum2. We evaluated the newly constructed PRS scores within the UKB testing cohort, comparing them against the best result obtained from the individual PRS algorithms. Additionally, we compared the results obtained with UKB summary statistics from the WBU set with those obtained through meta-analysis. These findings were benchmarked against the PRS scores by Thompson et al (2022)[8], where the authors demonstrated that their PRS release surpassed a comprehensive array of previously published PRS models.

Our meta-analysis, which combined the three PRS algorithms, generally outperformed the other methods (Figure 2a,b). This combined method exceeded the best individual algorithm in 23 out of 30 phenotypes. Moreover, our results were superior to those reported by Thompson et al. for 23 traits, with significant differences observed in 15 of these phenotypes. PRS analyzes utilizing only summary statistics from the UKB WBU set yielded the poorest performance relative to other strategies. However, results exclusively based on the UKB data showed marginally better outcomes for four traits. This negligible improvement may be attributable to random variation within the analysis, as none of these differences were statistically significant and were closely paralleled by the accuracy reached by our meta-analysis.

Subsequently, we analyzed the performance across different ancestries in UKB individuals, who were stratified by their genetically predominant ancestry using principal component analysis. Our findings were comparable to those of Thompson *et al*. (Figure 2c). As previously noted[76], accuracy for non-European groups dropped relative to Europeans, with Africans experiencing the largest reduction in accuracy (Figure 2d).

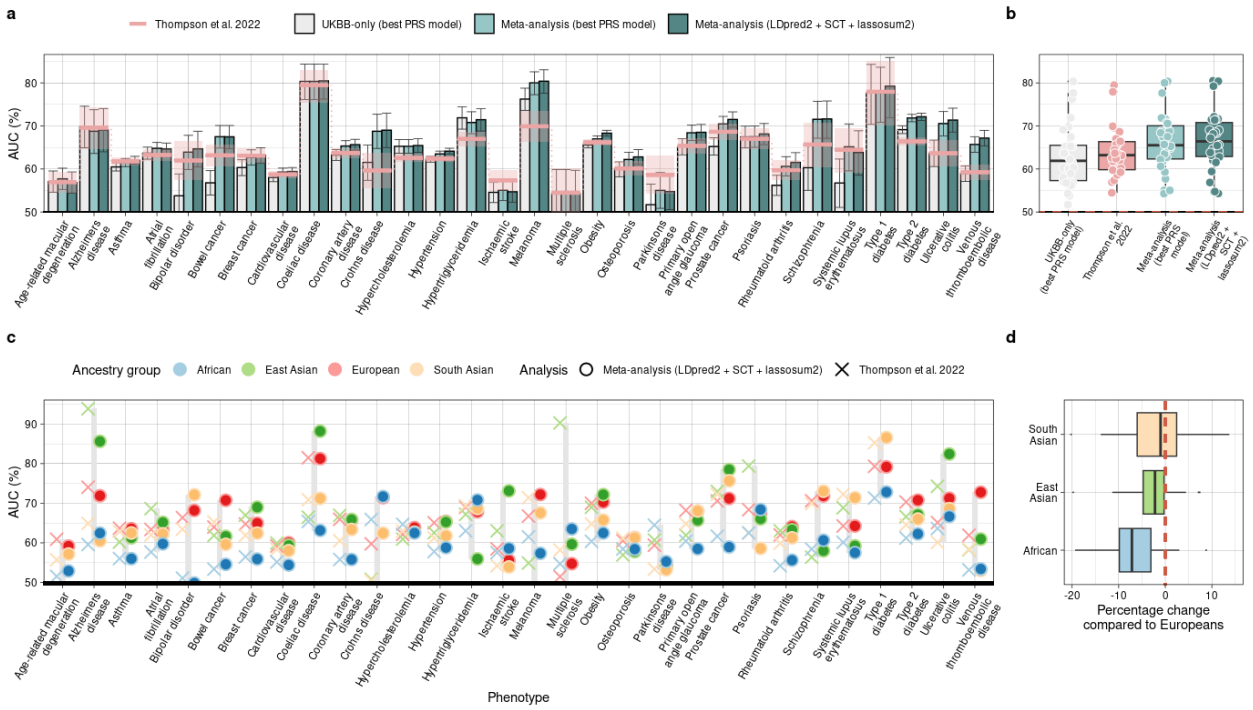

Figure 2. Predictive performance of PRS models built upon UKB-only and meta-analysis GWAS in the UK Biobank testing set. a) Prediction accuracy was measured using area under the curve (AUC) as a performance metric. Error bars indicate 95% confidence intervals of 10,000 non-parametric bootstrap replicates. Results were compared against AUC obtained with PRS scores from Thompson *et al*. (2022) (pink line and shaded area as 95% CI). **b)** Prediction results summarized per PRS study. **c)** Predictive performance by ancestry. **d)** Relative percentage change in performance (AUC) in non-European groups compared to accuracy obtained in Europeans.

Ancestry Incorporation Into PRS Models

Next, we proceeded to add ancestry information to the PRS algorithms. We applied an iterative approach, settling on the use of global ancestry information as deduced by Orchestra[77] in addition to the first four genotyping principal components (PCs) (Fig. 3). Not surprisingly, we found an overall improvement in accuracy with the addition of ancestry information. Adding only the PC components to the baseline (PRS-alone) model improved performance for 26 traits, with a mean average improvement of 1%. Moreover, the addition of global ancestry information deduced by Orchestra improved accuracy for 23 out of 30 traits compared to the addition of PCs, with a total average improvement of 1.22% over PRS-alone models.

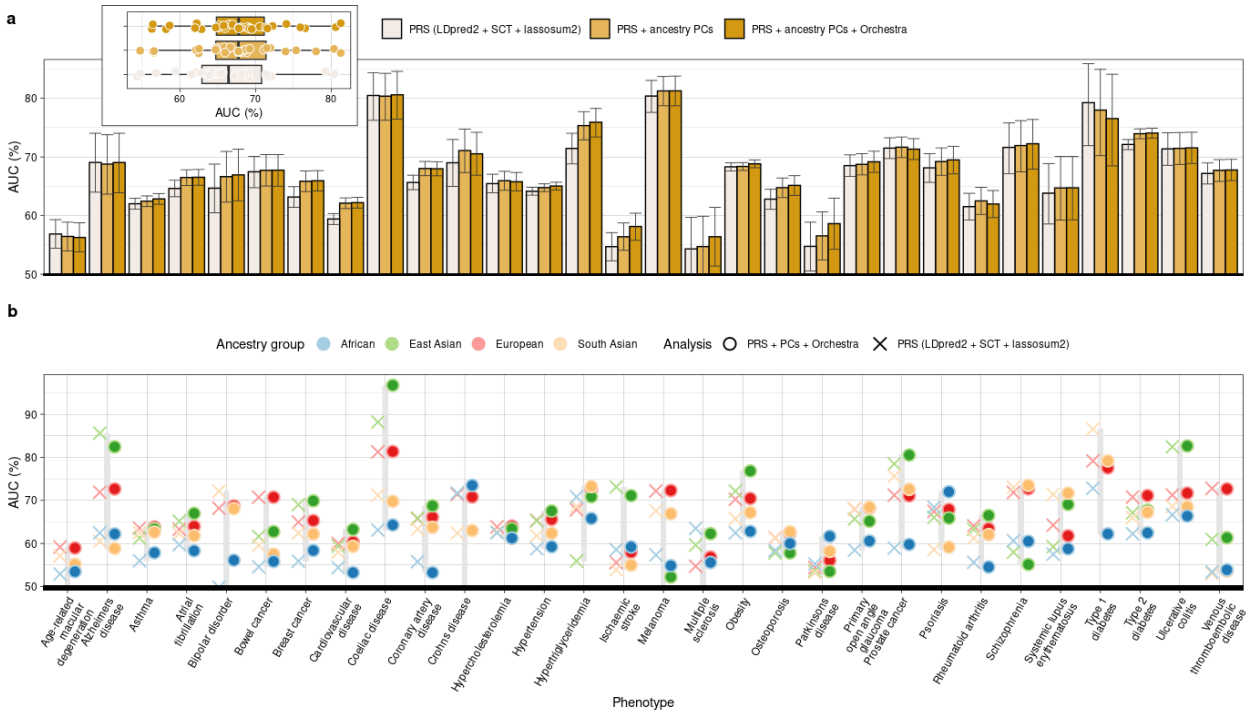

Figure 3. Predictive performance of PRS models incorporating ancestry information. a) Prediction accuracy was measured using AUC. Bars indicate 95% confidence intervals of 10,000 non-parametric bootstrap replicates. Models shown include performance with PRS alone, traditional logistic regression to include genotyping PCs as ancestry information, and a final model including both PCs and global results from the local ancestry deconvolution method Orchestra. Inset, summary of AUC results per model. **b)** Breakdown of results per ancestry group as a comparison between PRS-only and PRS plus full ancestry model.

Addition of Clinical Characteristics to PRS-Based Disease Risk Models

While creating predictive risk models based on genetic information alone is a good way to test and benchmark various PRS models against each other, in a real world clinical setting, the addition of non-genetic information to the final model is needed in order to increase performance. With this in mind, we incorporated other clinical characteristics, easily gatherable in a clinical setting. We added this information to PRS and ancestry to make a final logistic regression model that would serve as a prospective diagnostic test, with an easily interpretable positive or negative value, typical of other diagnostic tests used in a clinical setting. Table 2 contains the various inputs for each trait model which includes additional variables: sex, risk factors and age.

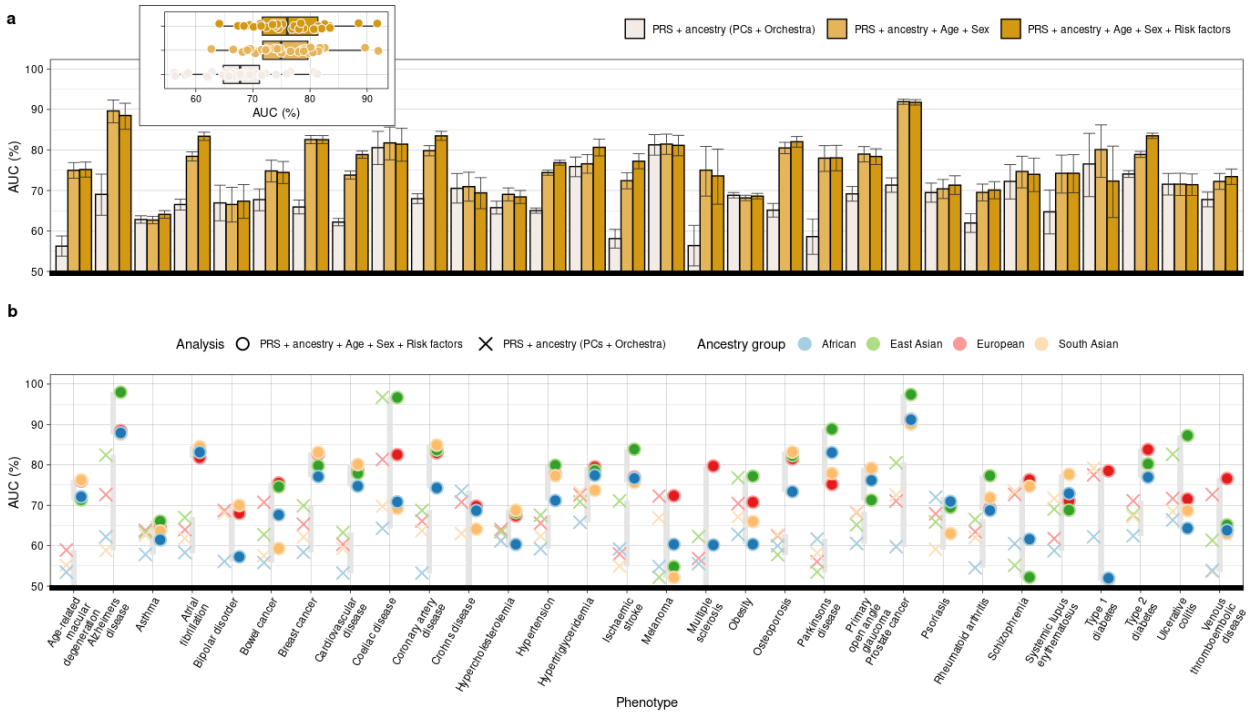

Figure 4. Predictive performance of PRS models incorporating ancestry, age, sex and risk factors information. a) Prediction accuracy was measured using AUC. Bars indicate 95% confidence intervals of 10,000 non-parametric bootstrap replicates. Models shown include performance with the model from Figure 3 (PRS and ancestry information), traditional logistic regression to include genotyping PCs as ancestry information, age and sex; and a final model including this in addition to risk factors. Inset, summary of AUC results per model. **b)** Same results per ancestry group as a comparison between PRS plus ancestry model and full model including genetic and non-genetic variables.

As can be seen in Figure 4, there was improvement in a large subset of the traits tested with some models nearly reaching the 90% accuracy range with the AUC metric. Including age and sex to the previous (PRS and ancestry; see Figure 3) model improved performance for 27 traits, with a mean average improvement of 8.1%. The addition of risk factors added a modest extra 0.72% to the final performance. Many of these models demonstrated sufficient enough overall accuracy to potentially merit being used in a clinical setting. To further elucidate the effectiveness of these models as diagnostic tests, we next calculated the diagnostic odds ratio[78] for each model. As can be seen in Figure 5, diagnostic odds ratios ranged from a low of 1.38 for obesity in Africans to a high of 34.1 for Alzheimer's disease in South Asians.

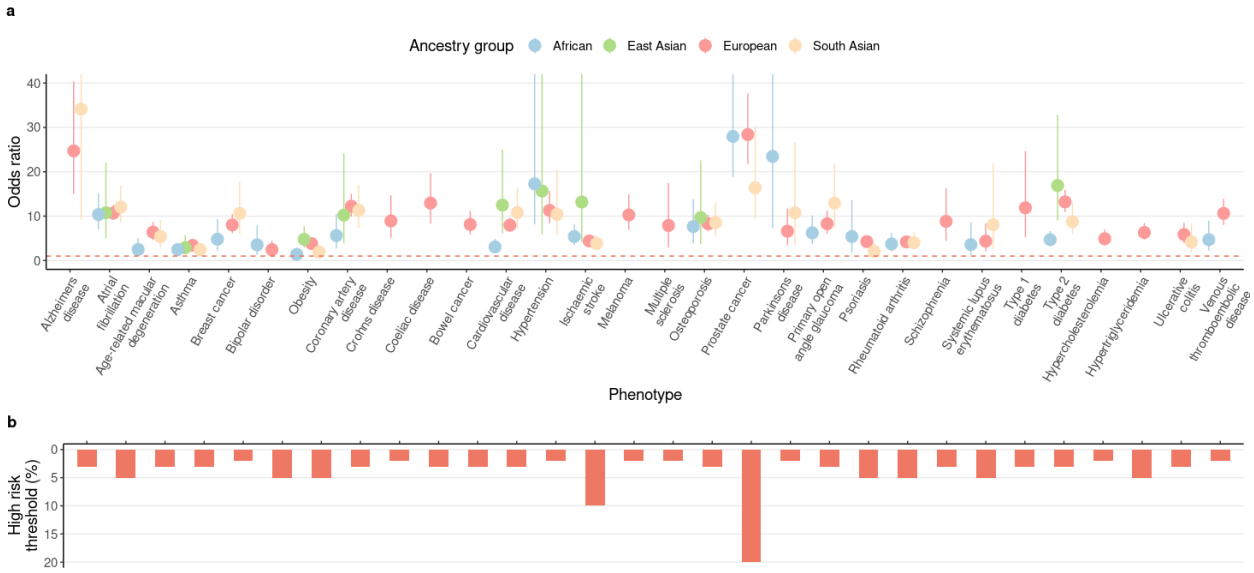

Figure 5. Diagnostic test effectiveness across medical traits. a) Odds ratios are reported per ancestry group, defined as the ratio of the odds that an individual with a clinical condition will score above a specified risk threshold, to the odds of scoring below this threshold without the condition. Error bars represent 95% confidence intervals. Ancestries with fewer than five true positive cases were excluded due to insufficient data. **b)** 'High risk threshold' illustrates the percentile cutoff for a specific condition, above which results are considered high-risk. Odds ratios were optimized using thresholds at the top 2%, 3%, 5%, 10%, and 20% percentiles.

Discussion

Polygenic risk scores (PRS) represent a new frontier in the field of personalized medicine, offering the potential to predict an individual's susceptibility to various diseases by aggregating the effects of numerous genetic variants across the genome. The promise of PRS lies in their ability to harness the wealth of data generated by genome-wide association studies (GWAS), thereby enabling risk stratification at a population level and the possibility of targeted interventions for those at high risk[1,2,3,6,7,9]. This could lead to more precise preventive measures, refined screening protocols, and tailored therapeutic strategies, ultimately improving clinical outcomes. However, the utility of PRS is currently hampered by several limitations. Chief among these is the reduced predictive power in diverse populations due to the majority of GWAS being conducted in individuals of European ancestry, leading to biased risk predictions when applied to other ethnic groups[1,8]. Additionally, the integration of PRS into clinical practice faces challenges, including the need for validated models that have the same level of accuracy and utility of traditional diagnostic tests used in a clinical setting[9,10].

With that in mind we set about to comprehensively optimize and test PRS-based disease prediction models utilizing traditional diagnostic test metrics. First, we performed trans-ancestry GWAS and meta-analysis across 30 medically-related traits. We used the resulting final GWAS

to test the performance of four PRS algorithms against each other on these 30 traits: lassosum2, LDpred2, maxCT and Stacked C+T (SCT). Seeing that certain algorithms performed better with certain traits, we built a new PRS model using logistic regression incorporating scores from SCT, LDpred2, and lassosum2 and directly benchmarked on a sample-by-sample basis against Thomspon et al., demonstrating superior performance in 23 out of 30 models and improvement in portability across different ancestral populations compared to what has been discussed in the literature.

We further refined these models by adding ancestry information, both in the form of principal components and global ancestry inference proportions obtained from Orchestra, demonstrating improvements in accuracy with each addition of these separate inputs. The incorporation of ancestry information emerged as a pivotal factor in improving model accuracy, directly addressing the critical challenge of enhancing PRS performance across non-European populations – a limitation long recognized in the field. Finally, we added a subset of non-genetic clinical characteristics to each model iin a final classification-based logistic regression model. Our results demonstrated that many models had sufficient enough accuracy to warrant consideration of being used in a clinical setting. With ever increasing GWAS sample sizes, refinements, and improvements in PRS algorithms, these testing metrics should only improve with time.

We have optimized and systematically improved diagnostic accuracy of PRS by both increasing GWAS power and using additional inputs to increase final prediction accuracy. A next step would be to fully validate the entire process from instrumental genotype measurements to final classification in individuals that do not belong to cohorts used for training and validation in this paper. Nonetheless, as our current benchmarking demonstrates, we are one step closer to using PRS in a clinical setting across various ancestral populations.

Competing Interests

J.L.J, A.O., C.M., A.T, B.N., S.B., A.D.M. and P.G.Y. are either employed by and/or hold stock or stock options in Omics Edge, a subsidiary of Genius Labs. In addition, P.G.Y. has equity in Systomic Health LLC and Ethobiotics LLC. This does not alter our adherence to journal policies on sharing data and materials. There are no other relevant activities or financial relationships which have influenced this work.

Data Availability

This study used the openly available GWAS catalog, Finngen and deCODE genetics datasets and the UK Biobank, dbGaP and The Million Veterans Program datasets that are available to researchers upon application. UK Biobank research was conducted under application number #84038 and The Million Veterans Program #30221. dbGaP access was obtained for

phs001868.v1.p1. All data produced in the present study are available upon request to the corresponding author.

References

1. Lewis, C.M., Vassos, E. Polygenic risk scores: from research tools to clinical instruments. *Genome Med* **12**, 44 (2020).

2. Fahed, A.C., Wang, M., Homburger, J.R. *et al.* Polygenic background modifies penetrance of monogenic variants for tier 1 genomic conditions. *Nat Commun* **11**, 3635 (2020).

3. Khera, A.V., Chaffin, M., Aragam, K.G. *et al.* Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations. *Nat Genet* **50**, 1219–1224 (2018).

4. Wray, N.R., Ripke, S., Mattheisen, M. *et al*. Genome-wide association analyses identify 44 risk variants and refine the genetic architecture of major depression. *Nat Genet* **50**,668-681 (2018).

5. Mars, N., Widén, E., Kerminen, S. *et al.* The role of polygenic risk and susceptibility genes in breast cancer over the course of life. *Nat Commun* **11**, 6383 (2020).

6. Mavaddat, N., Michailidou, K., Dennis, J. *et al*. Polygenic Risk Scores for Prediction of Breast Cancer and Breast Cancer Subtypes. *Am J Hum Genet* **104**, 21-34 (2019).

7. Riveros-Mckay, F., Weale, M.E., Moore, R. *et al*. Integrated Polygenic Tool Substantially Enhances Coronary Artery Disease Prediction. *Circ Genom Precis Med*. **14**, e003304 (2021).

8. Thompson, D.J., Wells, D., Selzam, S. *et al*. UK Biobank release and systematic evaluation of optimised polygenic risk scores for 53 diseases and quantitative traits. Preprint at medRxiv 2022.06.16.22276246; https://doi.org/10.1101/2022.06.16.22276246 (2022).

9. Torkamani, A., Wineinger, N.E., Topol, E.J. The personal and clinical utility of polygenic risk scores. *Nat Rev Genet* **19**, 581–590 (2018).

10. Sud, A., Horton, R.H., Hingorani, A.D *et al*. Realistic expectations are key to realising the benefits of polygenic scores. BMJ **380**, e073149 (2023).

11. Hingorani, A.D., Gratton, J., Finan, C., *et al*. Performance of polygenic risk scores in screening, prediction, and risk stratification: secondary analysis of data in the Polygenic Score Catalog. BMJ Medicine **2**, e000554 (2023).

12. Albiñana, C., Zhu, Z., Schork, A.J. *et al.* Multi-PGS enhances polygenic prediction by combining 937 polygenic scores. *Nat Commun* **14**, 4702 (2023).

13. Wang, Y., Tsuo, K., Kanai, M. et al. Challenges and Opportunities for Developing More Generalizable Polygenic Risk Scores. *Annu Rev Biomed Data Sci* **10**, 293-320 (2022).

14. Bycroft, C., Freeman, C., Petkova, D. *et al.* The UK Biobank resource with deep phenotyping and genomic data. *Nature* **562**, 203–209 (2018).

15. Kurki, M.I., Karjalainen, J., Palta, P. *et al.* FinnGen provides genetic insights from a well-phenotyped isolated population. *Nature* **613**, 508–518 (2023).

16. Sakaue, S., Kanai, M., Tanigawa, Y. *et al.* A cross-population atlas of genetic associations for 220 human phenotypes. *Nat Genet* **53**, 1415–1424 (2021).

17. Kanai, M., Ulirsch, J.C., Karjalainen, J. *et al*. Insights from complex trait fine-mapping across diverse populations. Preprint at medRxiv 2021.09.03.21262975 https://doi.org/10.1101/2021.09.03.21262975 (2021).

18. Wojcik, G.L., Graff, M., Nishimura, K.K. *et al*. Genetic analyses of diverse populations improves discovery for complex traits. *Nature* **570**, 514-518 (2019).

19. Guindo-Martínez, M., Amela, R., Bonàs-Guarch, S. *et al.* The impact of non-additive genetic associations on age-related complex diseases. *Nat Commun* **12**, 2436 (2021).

20. Moreno-Grau, S., de Rojas, I., Hernández, I. *et al*. Genome-wide association analysis of dementia and its clinical endophenotypes reveal novel loci associated with Alzheimer's disease and three causality networks: The GR@ACE project. *Alzheimers Dement* **15**, 1333-1347 (2019).

21. Lambert, JC., Ibrahim-Verbaas, C., Harold, D. *et al.* Meta-analysis of 74,046 individuals identifies 11 new susceptibility loci for Alzheimer's disease. *Nat Genet* **45**, 1452–1458 (2013).

22. Low, SK., Takahashi, A., Ebana, Y. *et al.* Identification of six new genetic loci associated with atrial fibrillation in the Japanese population. *Nat Genet* **49**, 953–958 (2017).

23. Christophersen, I., Rienstra, M., Roselli, C. *et al.* Large-scale analyses of common and rare variants identify 12 new loci associated with atrial fibrillation. *Nat Genet* **49**, 946–952 (2017).

24. Choi, S.H., Weng, L.C., Roselli, C. *et al*., Association Between Titin Loss-of-Function Variants and Early-Onset Atrial Fibrillation. *JAMA* **320**, 2354-2364 (2018).

25. Demenais, F., Margaritte-Jeannin, P., Barnes, K.C. *et al.* Multiancestry association study identifies new asthma risk loci that colocalize with immune-cell enhancer marks. *Nat Genet* **50**, 42–53 (2018).

26. Chang, X., March, M., Mentch, F. *et al*. Genetic architecture of asthma in African American patients. *J Allergy Clin Immunol* **151**, 1132-1136 (2023).

27. Michailidou, K., Beesley, J., Lindstrom, S. *et al.* Genome-wide association analysis of more than 120,000 individuals identifies 15 new susceptibility loci for breast cancer. *Nat Genet* **47**, 373–380 (2015).

28. Mullins, N., Forstner, A.J., O'Connell, K.S. *et al.* Genome-wide association study of more than 40,000 bipolar disorder cases provides new insights into the underlying biology. *Nat Genet* **53**, 817–829 (2021).

29. Nikpay, M., Goel, A., Won, H.H. *et al*. A comprehensive 1000 Genomes–based genome-wide association meta-analysis of coronary artery disease. *Nat Genet* **47**, 1121–1130 (2015).

30. de Lange, K., Moutsianas, L., Lee, J. *et al.* Genome-wide association study implicates immune activation of multiple integrin genes in inflammatory bowel disease. *Nat Genet* **49**, 256–261 (2017).

31. Garcia-Etxebarria, K., Merino, O., Gaite-Reguero, A. *et al.* Local genetic variation of inflammatory bowel disease in Basque population and its effect in risk prediction. *Sci Rep* **12**, 3386 (2022).

32. Garcia-Etxebarria, K., Etxart, A., Barrero, M. *et al*. Performance of the Use of Genetic Information to Assess the Risk of Colorectal Cancer in the Basque Population. *Cancers (Basel)* **14**, 4193 (2022).

33. Tanikawa, C., Kamatani, Y., Takahashi, A. *et al.* GWAS identifies two novel colorectal cancer loci at 16q24.1 and 20q13.12. *Carcinogenesis* **39**, 652-660 (2018).

34. Fernandez-Rozadilla, C., Timofeeva, M., Chen, Z. *et al.* Deciphering colorectal cancer genetics through multi-omic analysis of 100,204 cases and 154,587 controls of European and east Asian ancestries. *Nat Genet* **55**, 89–99 (2023).

35. Oh, S.W., Lee, J.E., Shin, E. *et al*. Genome-wide association study of metabolic syndrome in Korean populations. *PLoS One* **15**, e0227357 (2020).

36. Malik, R., Chauhan, G., Traylor, M. *et al.* Multiancestry genome-wide association study of 520,000 subjects identifies 32 loci associated with stroke and stroke subtypes. *Nat Genet* **50**, 524–537 (2018).

37. Mishra, A., Malik, R., Hachiya, T. *et al.* Stroke genetics informs drug discovery and risk prediction across ancestries. *Nature* **611**, 115–123 (2022).

38. Landi, M.T., Bishop, D.T., MacGregor, S. *et al.* Genome-wide association meta-analyses combining multiple risk phenotypes provide insights into the genetic architecture of cutaneous melanoma susceptibility. *Nat Genet* **52**, 494–504 (2020).

39. Andlauer, T.F., Buck, D., Antony, G. *et al*. Novel multiple sclerosis susceptibility loci implicated in epigenetic regulation. *Sci Adv*. **2**, e1501678 (2016).

40. The International Multiple Sclerosis Genetics Consortium & The Wellcome Trust Case Control Consortium 2. Genetic risk and a primary role for cell-mediated immune mechanisms in multiple sclerosis. *Nature* **476**, 214–219 (2011).

41. International Multiple Sclerosis Genetics Consortium (IMSGC). Analysis of immune-related loci identifies 48 new susceptibility variants for multiple sclerosis. *Nat Genet* **45**, 1353–1360 (2013).

42. Schumacher, F.R., Al Olama, A.A., Berndt, S.I. *et al.* Association analyses of more than 140,000 men identify 63 new prostate cancer susceptibility loci. *Nat Genet* **50**, 928–936 (2018).

43. Ito, S., Liu, X., Ishikawa, Y. *et al.* Androgen receptor binding sites enabling genetic prediction of mortality due to prostate cancer in cancer-free subjects. *Nat Commun* **14**, 4863 (2023).

44. Rodrigo, L.M., Nyholt, D.R. Imputation and Reanalysis of ExomeChip Data Identifies Novel, Conditional and Joint Genetic Effects on Parkinson's Disease Risk. *Genes (Basel)* **12**, 689 (2021).

45. Nalls, M.A., Blauwendraat, C., Vallerga, C.L. *et al*. Identification of novel risk loci, causal insights, and heritable risk for Parkinson's disease: a meta-analysis of genome-wide association studies. *Lancet Neurol* **18**, 1091-1102 (2019).

46. Le Guen, Y., Napolioni, V., Belloy, M.E. *et al*. Common X-Chromosome Variants Are Associated with Parkinson Disease Risk. *Ann Neurol* **90**, 22-34 (2021).

47. Gharahkhani, P., Jorgenson, E., Hysi, P. *et al.* Genome-wide meta-analysis identifies 127 open-angle glaucoma loci with consistent effect across ancestries. *Nat Commun* **12**, 1258 (2021).

48. Stuart, P.E., Tsoi, L.C., Nair, R.P. *et al*. Transethnic analysis of psoriasis susceptibility in South Asians and Europeans enhances fine-mapping in the MHC and genomewide. *HGG Adv* **3**, 100069 (2022).

49. Ishigaki, K., Sakaue, S., Terao, C. *et al.* Multi-ancestry genome-wide association analyses identify novel genetic mechanisms in rheumatoid arthritis. *Nat Genet* **54**, 1640–1651 (2022).

50. Langefeld, C., Ainsworth, H., Graham, D. *et al.* Transancestral mapping and genetic load in systemic lupus erythematosus. *Nat Commun* **8**, 16021 (2017).

51. Wang, YF., Zhang, Y., Lin, Z. *et al.* Identification of 38 novel loci for systemic lupus erythematosus and genetic heterogeneity between ancestral groups. *Nat Commun* **12**, 772 (2021).

52. Song, Q., Lei, Y., Shao, L. *et al*. Genome-wide association study on Northern Chinese identifies KLF2, DOT1L and STAB2 associated with systemic lupus erythematosus. *Rheumatology (Oxford)* **60**, 4407-4417 (2021).

53. Trubetskoy, V., Pardiñas, A.F., Qi, T. *et al.* Mapping genomic loci implicates genes and synaptic biology in schizophrenia. *Nature* **604**, 502–508 (2022).

54. Inshaw, J.R.J., Sidore, C., Cucca, F. *et al*. Analysis of overlapping genetic association in type 1 and type 2 diabetes. *Diabetologia* **64**,1342-1347 (2021).

55. Robertson, C.C., Inshaw, J.R.J., Onengut-Gumuscu, S. *et al.* Fine-mapping, trans-ancestral and genomic analyses identify causal variants, cells, genes and drug targets for type 1 diabetes. *Nat Genet* **53**, 962–971 (2021).

56. Chen, J., Sun, M., Adeyemo, A. et al. Genome-wide association study of type 2 diabetes in Africa. *Diabetologia* **62**, 1204-1211 (2019).

57. Mansour Aly, D., Dwivedi, O.P., Prasad, R.B. *et al.* Genome-wide association analyses highlight etiological differences underlying newly defined subtypes of diabetes. *Nat Genet* **53**, 1534–1542 (2021).

58. Spracklen, C.N., Horikoshi, M., Kim, Y.J. *et al*. Identification of type 2 diabetes loci in 433,540 East Asian individuals. *Nature* **582**, 240-245 (2020).

59. Gaziano, J.M., Concato, J., Brophy, M. *et al*. Million Veteran Program: A mega-biobank to study genetic influences on health and disease. *J Clin Epidemiol* **70**, 214-23 (2016).

60. Cai, L., Wheeler, E., Kerrison, N.D. *et al.* Genome-wide association analysis of type 2 diabetes in the EPIC-InterAct study. *Sci Data* **7**, 393 (2020).

61. Loh, M., Zhang, W., Ng, H.K. *et al.* Identification of genetic effects underlying type 2 diabetes in South Asian and European populations. *Commun Biol* **5**, 329 (2022).

62. Scott, R.A., Scott, L.J., Mägi, R. *et al*. An Expanded Genome-Wide Association Study of Type 2 Diabetes in Europeans. *Diabetes* **66**, 2888-2902 (2017).

63. Gurdasani, D., Carstensen, T., Fatumo, S. *et al*. Uganda Genome Resource Enables Insights into Population History and Genomic Discovery in Africa. *Cell* **179**, 984-1002.e36 (2019).

64. Hoffmann, T.J., Theusch, E., Haldar, T. *et al.* A large electronic-health-record-based genome-wide study of serum lipids. *Nat Genet* **50**, 401–413 (2018).

65. Thareja, G., Belkadi, A., Arnold, M. Differences and commonalities in the genetic architecture of protein quantitative trait loci in European and Arab populations. *Hum Mol Genet* **32**, 907-916 (2023).

66. Graham, S.E., Clarke, S.L., Wu, KH.H. *et al.* The power of genetic diversity in genome-wide association studies of lipids. *Nature* **600**, 675–679 (2021).

67. Sohail, M., Palma-Martínez, M.J., Chong, A.Y. *et al.* Mexican Biobank advances population and medical genomics of diverse ancestries. *Nature* **622**, 775–783 (2023).

68. Chen, Y., Lu, T., Pettersson-Kymmer, U. *et al.* Genomic atlas of the plasma metabolome prioritizes metabolites implicated in human diseases. *Nat Genet* **55**, 44–53 (2023).

69. Liu, J., van Sommeren, S., Huang, H. *et al.* Association analyses identify 38 susceptibility loci for inflammatory bowel disease and highlight shared genetic risk across populations. *Nat Genet* **47**, 979–986 (2015).

70. Ghouse, J., Tragante, V., Ahlberg, G. *et al.* Genome-wide meta-analysis identifies 93 risk loci and enables risk prediction equivalent to monogenic forms of venous thromboembolism. *Nat Genet* **55**, 399–409 (2023).

71. Willer, C.J., Li, Y., Abecasis, G.R.. METAL: fast and efficient meta-analysis of genomewide association scans. *Bioinformatics* **26**, 2190-1 (2010).

72. Sollis, E., Mosaku, A., Abid, A. *et al*. The NHGRI-EBI GWAS Catalog: knowledgebase and deposition resource. *Nucleic Acids Res* **51**, D977-D985 (2023).

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

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

75. Privé, F., Vilhjálmsson, B.J., Aschard, H. et al. Making the Most of Clumping and Thresholding for Polygenic Scores. *Am J Hum Genet* **105**, 1213-1221 (2019).

76. Ding, Y., Hou, K., Xu, Z. *et al.* Polygenic scoring accuracy varies across the genetic ancestry continuum. *Nature* **618**, 774–781 (2023).

77. Lerga-Jaso, J., Novković, B., Unnikrishnan, D. *et al*. Retracing Human Genetic Histories and Natural Selection Using Precise Local Ancestry Inference. Preprint at bioRxiv 2023.09.11.557177 https://doi.org/10.1101/2023.09.11.557177 (2023).

78. Šimundić, A.M. Measures of Diagnostic Accuracy: Basic Definitions. *EJIFCC***19**, 203-11 (2009).

79. Wightman, D.P., Jansen, I.E., Savage, J.E. *et al.* A genome-wide association study with 1,126,563 individuals identifies new risk loci for Alzheimer's disease. *Nat Genet* **53**, 1276–1282 (2021).

Methods

GWAS Meta-Analysis

We generated novel genome-wide summary statistics for 30 medically-related traits, leveraging trans-ancestry genome-wide association study (GWAS) meta-analysis of diverse cohort data, including the UK Biobank (Table 1)[14,15,16,17,18,19,20,21,22,23,24,25,26,27,28,29,30,31,32,33,34,35,36,37,38,39,40,41,42,43,44,45,46, 47,48,49,50,51,52,53,54,55,56,57,58,59,60,61,62,63,64,65,66,67,68,69,70,79]. For the purpose of PRS training, the three continuous traits, BMI, total cholesterol and triglycerides, were converted into binary outcomes, obesity, hypercholesterolemia, and hypertriglyceridemia, respectively. GWAS meta-analyses were conducted using METAL[71], incorporating publicly available external GWAS sources, predominantly sourced from the GWAS Catalog[72]. We kept out a set of 104,604 individuals from the UK Biobank (UKB) to serve as a training and testing cohort for PRS prediction and downstream validations. The cohort was composed of 76,998 in the training and 27,606 in the testing set. The training set was further divided into two sub-cohorts for PRS algorithm training (TS1 set) and the development of new risk models combining distinct PRS scores from different sources, ancestry, age, sex, and other risk factors. TS1 and TS2 are composed of 30,000 and 46,998 samples, respectively. We adopted the approach described in a recent preprint by Thompson *et al.* (2022)[8], which aimed to maximize the representation of non-European ancestries in this UKB subset of samples. The remainder of the samples, previously classified as the White British Unrelated (WBU) subset in the UKB project[14], were utilized for the GWAS component of the meta-analysis. This strategy facilitated direct comparisons of our results against the PRS developed by Thompson *et al*. using the same testing set.

PRS Model Benchmarking

To benchmark predictive performance and elucidate each method's strengths and weaknesses, we compared four PRS algorithms: lassosum2[73], LDpred2[74], maxCT and Stacked C+T (SCT)[75]. The training set TS1 was used to determine the optimal hyper-parameters for lassosum2, LDpred2, and maxCT, in addition to the stacking weights for SCT.

For maxCT, tuning parameters included a squared correlation (r^2) threshold for clumping and a base size for the clumping window, with default settings specified in Prive *et al*. 2019[75]. Additionally, a series of 50 thresholds on *P* values were applied, ranging from the least to the most significant in the final meta-analyzed summary statistics, and evenly distributed on a log-log scale. LDpred2's tuning involved selecting the proportion of causal SNPs from a sequence of 21 logarithmically spaced values ranging from 10^{-5} to 1, and inferring the per-SNP heritability chosen from {0.3, 0.7, 1, 1.4} times the total heritability -as estimated by LD score regression-. We enabled the 'sparse' option to truncate minor effects to zero, which effectively doubled the number of models considered for training by incorporating both sparse and non-sparse models. lassosum2's tuning parameters included the lasso penalty lambda (L1-regularization), selected from 30 values logarithmically spaced between lambda0 -the maximum value where all

coefficients are zero, derived from the maximum absolute standardized beta coefficient- and 1% of lambda0, along with a delta (L2-regularization) parameter for the linkage disequilibrium (LD) matrix chosen from {0.001,0.01,0.1,1}.

Both LDpred2 and lassosum2 were implemented using the functions *snp_ldpred2_grid* and *snp_lassosum2* from the R package *bigsnpr*, respectively. We adopted an LD radius of 3 cM to approximate local LD patterns, which assumes that variants further away than this distance are not correlated. Moreover, genetic markers were further restricted to the HapMap3 panel, following recommendations by Prive et al. 2020. Conversely, maxCT and SCT were executed using corresponding functions from the same R package (*snp_grid_clumping* and *snp_grid_PRS*, and *snp_grid_stacking*, respectively). For these analyses, we limited the set of genetic variants to those achieving a significance level of $P < 0.1$.

The models with the highest prediction accuracy on the training TS1 set were selected based on the Area under the ROC curve (AUC). The predictive performance of these final models was then evaluated on the independent testing set. Unlike diseases encoded as binary outcomes (case/control), optimal values for the quantitative traits -body mass index, total cholesterol, and triglyceride levels- were initially determined by training models to maximize the predictive R^2 between observed and predicted phenotypes. Subsequently, these phenotypes were converted to binary outcomes to facilitate AUC calculation and comparison in the testing samples.

Multi-algorithm PRS

We developed new PRS models for each disease by integrating scores previously obtained from SCT, LDpred2, and lassosum2. These models were retrained using logistic regression, employing the TS2 set as the training set to optimize the integration of these scores. This approach allowed us to harness the strengths of each individual scoring method to enhance overall predictive accuracy. Furthermore, this strategy proved to be an effective method of leveraging a larger cohort for training without the need to directly retrain the more computationally intensive and time-consuming PRS algorithms described above, which had been previously calibrated within a smaller, yet adequate cohort (TS1), thereby streamlining the overall process.

The Addition of Ancestry Information to PRS

To improve the predictive accuracy of our PRS models, we incorporated ancestry information through an iterative process. We began by incorporating the first four principal components (PCs) identified from a principal component analysis (PCA) on variants that were pruned to ensure linkage equilibrium and had a minor allele frequency (MAF) above 5%, capturing broad ancestry patterns. We then enhanced these models by integrating ancestry estimates from Orchestra, a method for local ancestry deconvolution. We aggregated and averaged the results from all genomics regions into a comprehensive global ancestry measure. Orchestra's results were provided at level 2, which offers a regional granularity intermediate between population-

specific and continental scales of ancestry detail (see reference 77). Both measures were incorporated into a logistic regression model and trained in the TS2 set.

The Addition of Clinical Characteristics to PRS-Based Disease Risk Models

With this in mind, we incorporated age, sex, and other clinical characteristics, easily gatherable in a clinical setting (Table 2). We chose characteristics that are known risk factors for each medically-related trait/condition. We added this information to PRS and ancestry to make a final logistic regression model that would serve as a prospective diagnostic test, with an easily interpretable positive or negative value, typical of other diagnostic tests used in a clinical setting.

Table 1. Studies and database sources for the GWAS meta-analysis for the 30 medically-related traits, including the numbers of cases and controls in each dataset.

Trait	Study/Database	Cases	Controls
Age-related macular degeneration	UKB[14]	7,476	328,367
	Finngen[15]	7,582	348,936
	Guindo-Martínez *et al.* (2021)[19]	3,685	52,952
	∑	**18,743**	**730,255**
Alzheimer's disease	UKB[14]	1,760	335,351
	Finngen[15]	13,393	363,884
	Moreno-Grau *et al.* (2019)[20]	6063	6305
	Lambert *et al.* (2021)[21]	17,008	37,154
	Wightman *et al.* (2021)[79]	39,918	358,140
	∑	**78,142**	**1,100,834**
Asthma	UKB[14]	48,278	288,833
	Finngen[15]	42,163	202,399
	Demenais *et al.* (2017)[25]	5,186	7,660
	Guindo-Martínez *et al.* (2021)[19]	9,209	47,428
	Sakaue *et al.* (2021)/Biobank Japan[16]	13,015	162,933
	Chang *et al.* (2022)[26]	3,876	2,607
	∑	**121,727**	**711,860**
Atrial fibrillation	UKB[14]	22,180	285,774
	Finngen[15]	45,766	191,924
	Sakaue *et al.* (2021)/Biobank Japan[16]	4,150	155,540
	Choi *et al.* (2018)[24]	2,781	4,959
	Low *et al.* (2017)[22]	8,180	28,612
	Christophersen *et al.* (2017)[23]	18,398	91,536
	∑	**101,455**	**758,345**
Bipolar disorder	Mullins et al. (2021)[28]	40,463	313,436
	∑	**40,463**	**313,436**
Bowel cancer	UKB[14]	6,171	302,196
	Finngen[15]	5,458	287,137
	Sakaue *et al.* (2021)/Biobank Japan[16]	8,305	159,386
	Tanikawa *et al.* (2018)[33]	6,692	27,178
	Garcia-Etxebarria *et al.* (2022)[32]	835	940
	Fernandez-Rozadilla *et al.* (2023)[34]	100,204	154,587
	∑	**127,665**	**931,424**
Breast cancer	UKB[14]	12,933	152,004
	Finngen[15]	15,680	167,189
	Sakaue *et al.* (2021)/Biobank Japan[16]	6,325	73,225
	Michailidou *et al.* (2015)[27]	46,785	42,892
	∑	**81,723**	**435,310**
Cardiovascular disease	UKB[14]	37,890	269,427
	Finngen[15]	185,353	191,924
	Guindo-Martínez *et al.* (2021)[19]	15,009	41,628
	Nikpay *et al.* (2015)[29]	60,801	123,504
	Kanai *et al.* (2021)/Biobank Japan[17]	32,512	146,214
	∑	**331,565**	**772,697**
Celiac disease	UKB[14]	2,656	334,455

	∑	**2,656**	**334,455**
Coronary artery disease	UKB	31,451	276,086
	Finngen[15]	47,550	313,400
	Kanai *et al.* (2021)/Biobank Japan[17]	32,512	146,214
	Nikpay *et al.* (2015)[29]	60,801	123,504
	∑	**172,314**	**859,204**
Crohn's disease	UKB[14]	2,064	335,047
	Finngen[15]	1,665	375,445
	Garcia-Etxebarria *et al.* (2022)[31]	284	935
	de Lange *et al.* (2017)[30]	4,474	9,500
	∑	**8,487**	**720,927**
Hypercholesterolemia (Total cholesterol)	UKB[14]	83,838	
	∑	**83,838**	
Hypertension	UKB[14]	121,548	184,542
	Finngen[15]	111,581	265,626
	Guindo-Martínez *et al.* (2021)[19]	28,391	28,246
	Oh *et al.* (2020)[35]	1,864	7,812
	Wojcik *et al.* (2019)[18]	27,123	22,018
	∑	**290,507**	**508,244**
Hypertroglyceridemia (Total triglycerides)	UKB[14]	83,838	
	Graham *et al.* (2021)	1,121,280	
	∑	**1,205,118**	
Ischemic stroke	UKB[14]	5,761	302,890
	Finngen[15]	39,818	271,817
	Malik *et al.* (2018) * only EUR[36]	34,217	406,111
	Mishra *et al.* (2022) * only non-EUR[37]	24,568	269,090
	∑	**104,364**	**1,249,908**
Melanoma	UKB[14]	4,168	331,681
	Finngen[15]	3,960	286,874
	Landi *et al.* (2020)[38]	30,143	81,405
	∑	**38,271**	**699,960**
Multiple sclerosis	UKB[14]	1,632	335,479
	∑	**1,632**	**335,479**
Obesity (BMI)	UKB[14]	336,073	
	Finngen[15]	412,181	
	Sakaue *et al.* (2021)/Biobank Japan[16]	163,835	
	Wojcik *et al.* (2019)[18]	49,335	
	∑	**961,424**	
Osteoporosis	UKB[14]	14,636	321,851
	Finngen[15]	7,300	358,014
	Sakaue *et al.* (2021)/Biobank Japan[16]	9,794	168,932
	Guindo-Martínez *et al.* (2021)[19]	5,399	51,238
	∑	**37,129**	**900,035**
Parkinson's disease	UKB[14]	2,343	334,768
	Finngen[15]	4,235	373,042
	Sakaue *et al.* (2021)/Biobank Japan[16]	340	175,788
	Rodrigo *et al.* (2021)[44]	5,167	5,366
	Nalls *et al.* (2019)[45]	15,056	12,637

	Le Guen *et al.* (2021)[46]	3,500	299,408
	∑	**30,641**	**1,201,009**
Primary open angle glaucoma	UKB[14]	7,674	326,897
	Gharahkhani *et al.* (2021)[47]	6,935	39,588
	∑	**14,609**	**366,485**
Prostate cancer	UKB[14]	9,687	140,038
	Finngen[15]	13,216	119,948
	Sakaue *et al.* (2021)/Biobank Japan[16]	5,672	84,660
	Schumacher *et al.* (2018)[42]	79,148	61,106
	Ito *et al.* (2023)[43]	107,281	197,733
	∑	**215,004**	**603,485**
Psoriasis	UKB[14]	6,358	330,753
	Finngen[15]	9,267	364,071
	Sakaue *et al.* (2021)/Biobank Japan[16]	206	172,289
	Stuart *et al.* (2021)[48]	18,557	29,914
	∑	**34,388**	**897,027**
Rheumatoid arthritis	UKB[14]	7,719	329,392
	Finngen[15]	12,555	240,862
	Ishigaki *et al.* (2022)[49]	35,871	1,698,689
	∑	**56,145**	**2,268,943**
Schizophrenia	UKB[14]	698	336,413
	Finngen[15]	6,515	364,160
	Trubetskoy *et al.* (2022)[53]	67,323	93,456
	Sakaue *et al.* (2021)/Biobank Japan[16]	99	177,794
	∑	**74,635**	**971,823**
Systemic lupus erythematosus	UKB[14]	584	336,527
	Finngen[15]	652	353,088
	Sakaue *et al.* (2021)/Biobank Japan[16]	317	175,937
	Langefeld *et al.* (2017)[50]	6,748	11,516
	Wang *et al.* (2021)[51]	8,798	16,470
	Song *et al.* (2021)[52]	512	994
	∑	**17,611**	**894,532**
Type 1 diabetes	UKB[14]	1,200	333,542
	Finngen[15]	8,967	308,373
	Sakaue *et al.* (2021)/Biobank Japan[16]	1,219	132,032
	Inshaw *et al.* (2021)[54]	7,467	10,218
	Robertson *et al.* (2021)[55]	22,153	37,374
	∑	**41,006**	**821,539**
Type 2 diabetes	UKB[14]	24,800	281,093
	Finngen[15]	57,698	308,252
	Chen *et al.* (2019)[56]	2,633	1,714
	Guindo-Martínez *et al.* (2021)[19]	6,967	49,670
	Mansour *et al.* (2021)[57]	9,486	2,744
	Spracklen *et al.* (2020)[58]	77,418	356,122
	Wojcik *et al.* (2019)[18]	14,042	31,683
	Million Veterans Program[59]	102,683	170,726
	Cai *et al.* (2020)[60]	9,978	12,348
	Loh *et al.* (2022)[61]	16,677	33,856
	Scott *et al.* (2017)[62]	26,676	132,532

		∑	**349,058**	**1,380,740**
Ulcerative colitis	UKB[14]		3,914	333,197
	Finngen[15]		5,034	371,530
	Liu *et al.* (2015)[69]		6,968	20,464
		∑	**15,916**	**725,191**
Venous thromboembolism	UKB[14]		13,934	323,177
	Finngen[15]		19,372	357,905
	Million Veterans Program[59]		11,844	211,753
	Ghouse *et al.* (2023)[70]		81,190	1,419,671
		∑	**126,340**	**2,312,506**

Table 2. Risk factors included in the final PRS-based disease prediction model.

Trait	Additional Risk Factors										
	BMI	CAD	CD	CED	CVD	HT	ISS	RA	T1D	T2D	UC
Age-related macular degeneration	✓	/	/	/	/	✓	/	/	/	/	/
Alzheimer's disease	✓	/	/	/	/	✓	✓	/	✓	✓	/
Asthma	✓	/	/	/	/	/	/	/	/	/	/
Atrial fibrillation	✓	✓	/	/	✓	✓	/	/	/	✓	/
Bipolar disease	/	/	/	/	/	/	/	/	/	/	/
Bowel cancer	✓	/	✓	/	/	/	/	/	✓	✓	✓
Breast cancer	✓	/	/	/	/	/	/	/	/	/	/
Cardiovascular disease	✓	/	/	/	/	✓	/	/	✓	✓	/
Celiac disease	/	/	/	/	/	/	/	/	✓	/	/
Coronary artery disease	✓	/	/	/	/	✓	/	/	✓	✓	/
Crohn's disease	/	/	/	/	/	/	/	/	/	/	/
Hypercholesterolemia	/	/	/	/	/	/	/	/	/	/	/
Hypertension	✓	/	/	/	/	/	/	/	✓	✓	/
Hypertrygliceridemia	/	/	/	/	/	/	/	/	/	/	/
Ischemic stroke	✓	/	/	/	/	✓	/	/	/	/	/
Melanoma	/	/	/	/	/	/	/	/	/	/	/
Multiple sclerosis	/	/	✓	/	/	/	/	/	✓	/	✓
Obesity	/	/	/	/	/	/	/	/	/	/	/
Osteoporosis	/	/	✓	✓	/	/	/	✓	/	/	✓
Parkinson's disease	/	/	/	/	/	/	/	/	/	/	/
Primary open angle glaucoma	/	/	/	/	/	✓	/	/	✓	✓	/
Prostate cancer	✓	/	/	/	/	/	/	/	/	/	/
Psoriasis	✓	/	/	/	/	✓	/	/	/	/	/
Rheumatoid arthritis	✓	/	/	/	/	/	/	/	/	/	/
Schizophrenia	/	/	/	/	/	/	/	/	/	/	/
Systemic lupus erythematosus	/	/	/	/	/	/	/	/	/	/	/
Type 1 diabetes	/	/	/	/	/	/	/	/	/	/	/
Type 2 diabetes	✓	/	/	/	/	/	/	/	/	/	/
Ulcerative colitis	/	/	/	/	/	/	/	/	/	/	/
Venous thromboembolism	✓	✓	/	/	✓	✓	/	/	✓	✓	/

BMI = body mass index; CAD = Coronary artery disease; CD = Crohn's disease; CED = Celiac disease; CVD = Cardiovascular disease; HT = Hypertension; ISS = Ischemic stroke; RA = Rheumatoid arthritis; T1D = Type 1 diabetes; T2D = Type 2 diabetes; UC = Ulcerative colitis.