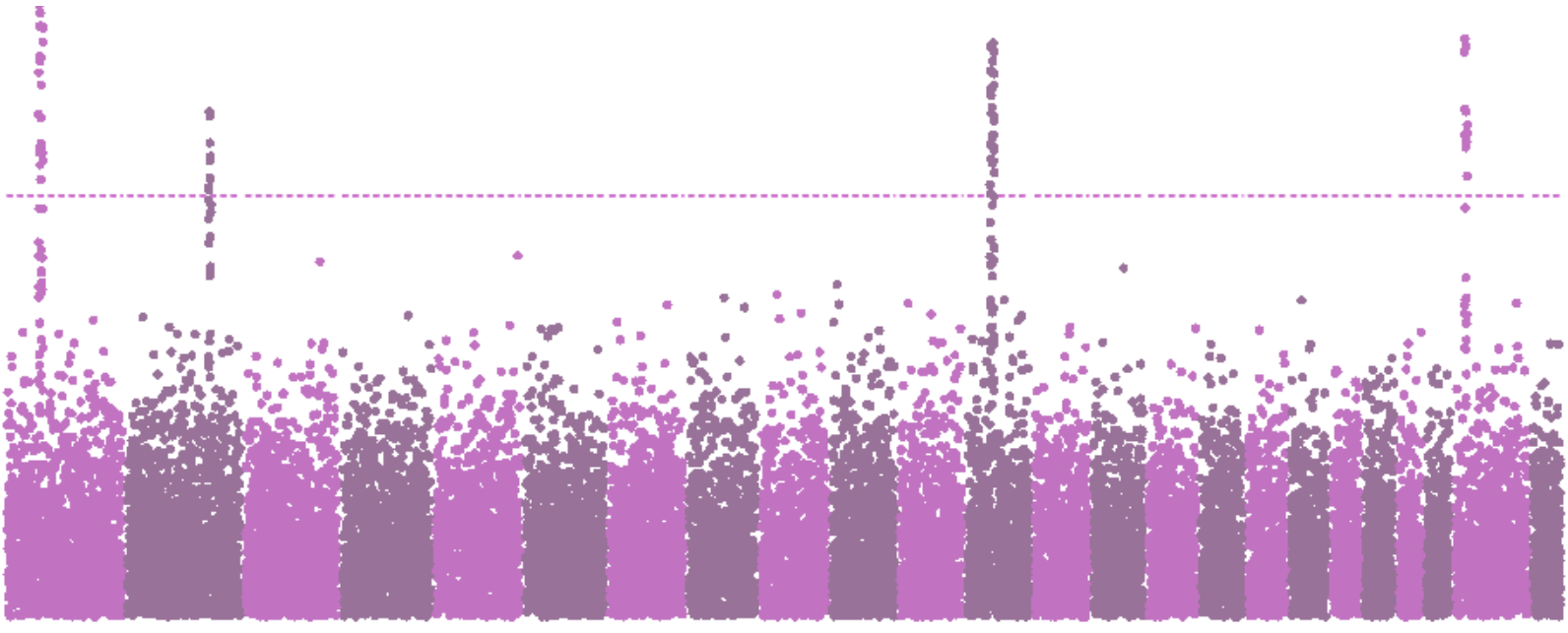
Polygenic Risk Scores validation

March, 2022

Puya Yazdi, Chief Science Officer, Omics Edge



Contents

Polygenic Risk Scores

Risk assessments of clinical disease consider demographics, medical history, lifestyle and lab results. Heredity plays a significant role in many common human diseases. Yet, the genetic contribution to disease risk is not well-integrated into clinical practice.

Polygenic Risk Scores (PRS) sum up the impacts of many inherited variants on the risk of developing a specific disease. They can identify high-risk individuals to target for early therapeutic intervention (Choi, Mak, and O'Reilly 2020; Torkamani, Wineinger, and Topol 2018). Most PRS models include thousands of variants with small effects on disease risk. These variants are identified by Genome-Wide Association Studies (GWAS). GWASs look for genetic differences between disease cases and healthy controls. Each GWAS examines millions of variants. If variants are much more or less common in the case group they may be associated with the disease. Each variant's effect size is an estimate of its contribution to disease risk. Aggregating the effects of the disease-associated variants yields PRSs (Choi, Mak, and O'Reilly 2020; Uffelmann et al. 2021).

PRS Validation

PRS models built from larger or even multiple GWASs tend to be more predictive. We generate and test a great number of PRS models. Once we build a PRS model, we check the risk predictions in another validation dataset. These datasets include genetic information and case-control status in a new independent group. In a robust PRS model, we expect cases in the validation dataset to receive higher risk scores more often than controls.

In Figure 1, we examine a stroke and a type 2 diabetes (T2D) PRS model. These models predict an individual's relative genetic risk of stroke or developing T2D. The stroke PRS model was built from a large multi-ancestry GWAS of approximately 67,000 stroke cases and 450,000 controls (Malik et al. 2018). The T2D PRS model was created from a GWAS that included approximately 63,000 Europeans with T2D and 600,000 controls (Xue et al. 2018).

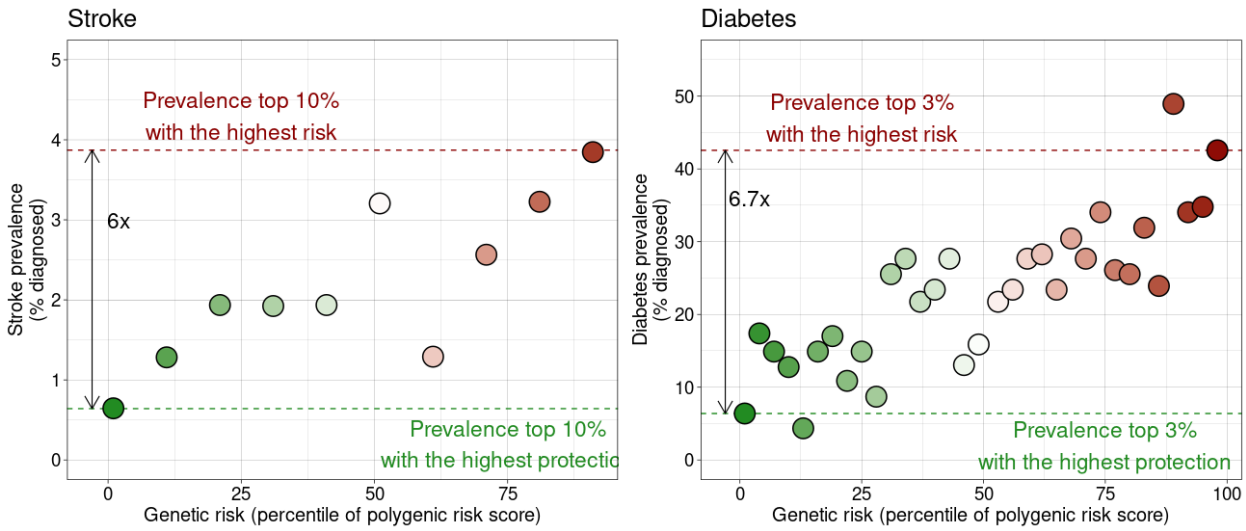


Fig 1: Performance of Stroke and type 2 Diabetes models. The polygenic risk score (PRS) is plotted against the actual prevalence of the disease. Each dot represents multiple individuals with that PRScore. Here we see the expected trend of greater disease prevalence cases at higher genetic risk levels. There are fewer points in the stroke PRS model as we have fewer individuals affected by stroke in our validation datasets.

Portability of PRS Models

When new GWAS are performed in study groups with a different ancestry or demographic composition, they identify new variants and also find some previously-tagged variants to be unassociated in these new populations. Repeated GWAS show that PRS models that predict risk well in one group may not be accurate in another group. These differences exist because patterns of linkage disequilibrium (LD) vary between populations. LD is the association between stretches of variants due to shared inheritance. GWAS use tag variants to mark each stretch and the associated variants. Differences in LD by population mean that the tagged stretch can change. So, a tag variant might tag a disease-associated variant in one population but not in another. This leads to differences in variants that affect a disease, for different populations. This also affects effect sizes, and ultimately, PRS accuracy.

Still, the majority of GWAS participants are European. As a result, PRS models drawn from these studies are difficult to generalize to other populations. Studies show that for many common human diseases, PRS accuracy can drop as low as 30-40% when applied to non-GWAS populations. The lowest predictability is observed for Asians, Africans, and those with multiple ancestries. One study found PRS accuracy decreased with increasing African ancestry in mixed Afro-Europeans (Cavazos and Witte 2021) (Fig 2). Demographic differences, such as socioeconomic status or age, can also impact PRS performance. This restricts PRS use in real

patient populations (Sirugo, Williams, and Tishkoff 2019; Martin et al. 2019; Mostafavi et al. 2020; Weissbrod et al. 2021; Cavazos and Witte 2021).

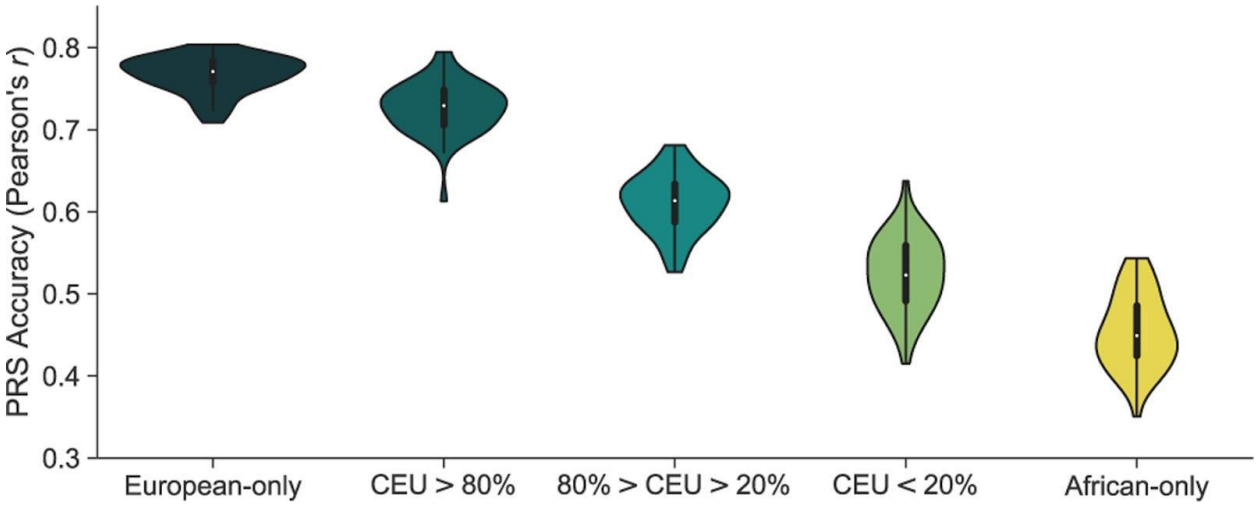


Fig 2: (From Cavazos and Witte, 2020). Accuracy of PRS models derived from a European-ancestry GWAS. CEU refers to the proportion of European ancestry, as opposed to African ancestry. As CEU decreases, the proportion of African ancestry of the individuals increases. Accuracy drops with increasing African Ancestry, and is lowest for African-only.

To check that our PRS are accurate in diverse patient populations, we create models for different ancestries. We then test their performance in multiple populations. We also apply techniques to decrease the impact of biased associations. These practices increase the likelihood that PRS will be usable in clinical settings. We have made further improvements to PRS performance by introducing these new approaches. These changes have increased our risk prediction accuracy beyond existing methods (Fig 3).

One metric we use to assess model accuracy is the 'area under the curve' (AUC) (Choi, Mak, and O'Reilly 2020) (Fig 3). Higher AUC values mean a PRS model is better at predicting risk. An AUC value near 50% indicates near-random predictive performance. Models that do not initially perform well are refined further until their accuracy increases to acceptable levels. If we are unable to improve them, they are discarded.

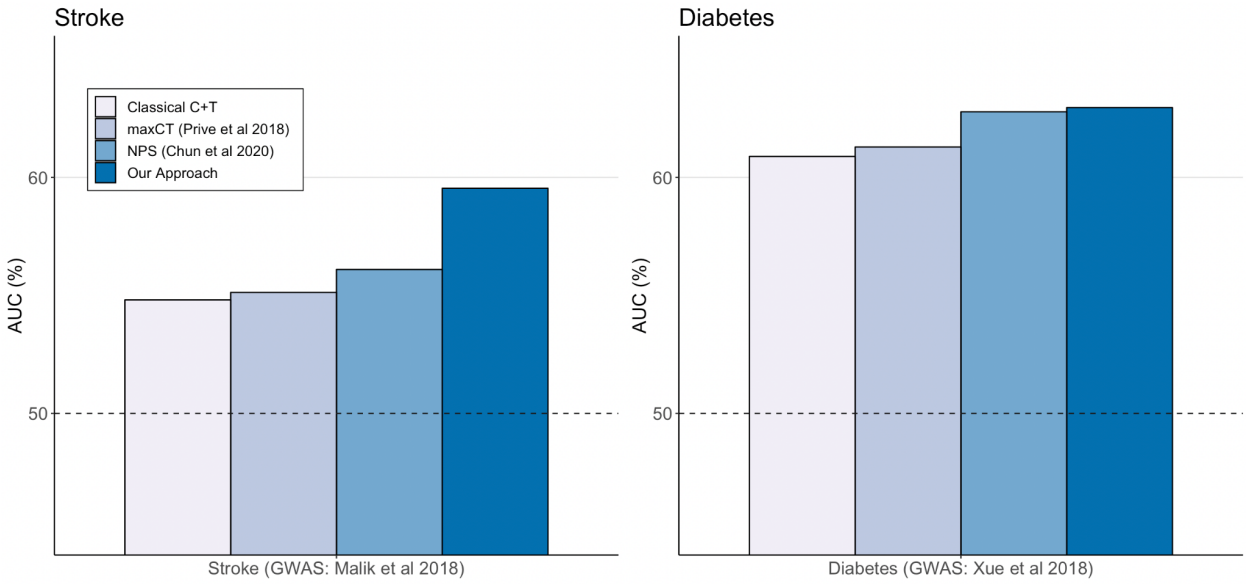


Fig 3: Comparison of PRS stratification ability, as measured by AUC in a validation set of 1,500 Asians using PRS models optimized for this ancestry. These PRS models are for stoke and type 2 diabetes risk. The comparison is between industry-leading approaches and our in-house approach.

Many clinicians already use risk calculators to identify high-risk patients and guide interventions. Screening tools often incorporate ethnicity to improve accuracy. This is because some risk factors, like high BMI, do not impact each ethnicity in the same way (Chan et al. 2009; Ma and Chan 2013). Differences in disease risk between populations can have genetic or environmental causes. Our goal is to help clinicians understand the genetic contribution to patient disease risk. For this to be possible, PRS models must be accurate in all patient populations. A clear picture of risk origins can assist with early detection and intervention, and ultimately improve patient outcomes.

References

Cavazos, Taylor B., and John S. Witte. 2021. “Inclusion of Variants Discovered from Diverse Populations Improves Polygenic Risk Score Transferability.” *Human Genetics and Genomics Advances* 2 (1): 100017. https://doi.org/10.1016/j.xhgg.2020.100017.

Chan, Juliana C. N., Vasanti Malik, Weiping Jia, Takashi Kadowaki, Chittaranjan S. Yajnik, Kun-Ho Yoon, and Frank B. Hu. 2009. “Diabetes in Asia: Epidemiology, Risk Factors, and Pathophysiology.” *JAMA* 301 (20): 2129. https://doi.org/10.1001/jama.2009.726.

Choi, Shing Wan, Timothy Shin-Heng Mak, and Paul F. O’Reilly. 2020. “Tutorial: A Guide to Performing Polygenic Risk Score Analyses.” *Nature Protocols* 15 (9): 2759–72. https://doi.org/10.1038/s41596-020-0353-1.

Ma, Ronald C.W., and Juliana C.N. Chan. 2013. “Type 2 Diabetes in East Asians: Similarities and Differences with Populations in Europe and the United States.” *Annals of the New York Academy of Sciences* 1281 (1): 64–91. https://doi.org/10.1111/nyas.12098.

Malik, Rainer, AFGen Consortium, Cohorts for Heart and Aging Research in Genomic Epidemiology (CHARGE) Consortium, International Genomics of Blood Pressure (iGEN-BP) Consortium, INVENT Consortium, STARNET, BioBank Japan Cooperative Hospital Group, et al. 2018. “Multiancestry Genome-Wide Association Study of 520,000 Subjects Identifies 32 Loci Associated with Stroke and Stroke Subtypes.” *Nature Genetics* 50 (4): 524–37. https://doi.org/10.1038/s41588-018-0058-3.

Martin, Alicia R., Masahiro Kanai, Yoichiro Kamatani, Yukinori Okada, Benjamin M. Neale, and Mark J. Daly. 2019. “Clinical Use of Current Polygenic Risk Scores May Exacerbate Health Disparities.” *Nature Genetics* 51 (4): 584–91. https://doi.org/10.1038/s41588-019-0379-x.

Mostafavi, Hakhamanesh, Arbel Harpak, Ipsita Agarwal, Dalton Conley, Jonathan K Pritchard, and Molly Przeworski. 2020. “Variable Prediction Accuracy of Polygenic Scores within an Ancestry Group.” *ELife* 9 (January): e48376. https://doi.org/10.7554/eLife.48376.

Sirugo, Giorgio, Scott M. Williams, and Sarah A. Tishkoff. 2019. “The Missing Diversity in Human Genetic Studies.” *Cell* 177 (1): 26–31. https://doi.org/10.1016/j.cell.2019.02.048.

Torkamani, Ali, Nathan E. Wineinger, and Eric J. Topol. 2018. “The Personal and Clinical Utility of Polygenic Risk Scores.” *Nature Reviews Genetics* 19 (9): 581–90. https://doi.org/10.1038/s41576-018-0018-x.

Uffelmann, Emil, Qin Qin Huang, Nchangwi Syntia Munung, Jantina de Vries, Yukinori Okada, Alicia R. Martin, Hilary C. Martin, Tuuli Lappalainen, and Danielle Posthuma. 2021. “Genome-Wide Association Studies.” *Nature Reviews Methods Primers* 1 (1): 59. https://doi.org/10.1038/s43586-021-00056-9.

Weissbrod, Omer, Masahiro Kanai, Huwenbo Shi, Steven Gazal, Wouter J. Peyrot, Amit V. Khera, Yukinori Okada, et al. 2021. “Leveraging Fine-Mapping and Non-European Training Data to Improve Cross-Population Polygenic Risk Scores.” Preprint. Genetic and Genomic Medicine. https://doi.org/10.1101/2021.01.19.21249483.

Xue, Angli, eQTLGen Consortium, Yang Wu, Zhihong Zhu, Futao Zhang, Kathryn E. Kemper, Zhili Zheng, et al. 2018. “Genome-Wide Association Analyses Identify 143 Risk Variants and Putative Regulatory Mechanisms for Type 2 Diabetes.” *Nature Communications* 9 (1): 2941. https://doi.org/10.1038/s41467-018-04951-w.