illumina®

Considerations for polygenic risk score development

Explore how polygenic risk scores are developed, from design to validation, and learn which additional reporting and quality aspects to consider for your research.

Introduction

In recent years, polygenic scores have become powerful tools for genetic disease research, genetic testing, and clinical trials, as evidenced by the increasing number of related research articles (Figure 1). In principle, a polygenic score is a weighted sum of allele counts across variants, where the weight reflects the magnitude of association between variant alleles and a trait or disease (as defined in a discovery genome-wide association study (GWAS) for the trait of interest). While the term polygenic score can apply to any phenotype, it is more commonly referred to as polygenic risk score (PRS) or genetic or genomic risk score (GRS) when describing disease risk.[1]

Historically, the number of alleles included in overall risk assessment have been limited, due to the relatively small size of GWAS. Genome-wide risk scores are becoming relatively common, and the number of identified single nucleotide polymorphisms (SNPs) associated with some diseases has increased into the millions.[1] PRS clinical validity, utility, and sophistication is likely to improve with increased interest from government initiatives, the pharmaceutical industry, and health systems, and as the scientific community continues to improve methods and gains access to larger scale GWAS data, carried out in populations of different ethnic backgrounds.

There is great potential for PRSs to play a key role in personalized medicine in the future. Many common diseases with a need for preventive and personalized medicine have some level of heritability. PRSs have the potential to explain this heritability. PRSs have been shown to outperform conventional methods, such as family history analysis, for predicting risk for many diseases.[2] Furthermore, it has been shown that for some diseases, eg, cancer and cardiovascular disease, a polygenic component can have a stronger effect on clinical risk, relative to monogenic variants. A PRS can explain a greater portion of the disease/trait heritability than monogenic mutations and can also modulate the penetrance of monogenic mutations.[3, 4]

This application note provides an overview of available tools and methods for PRS development and calculation.

PRS development process

Development of a PRS is a multistage process that begins with testing in a risk model predicting the phenotype of interest, which may be combined with other nonclinical variables. Collectively, all variables included in the risk model are referred to as the risk model parameters. After fitting procedures to select the best risk model, this model is validated in an independent sample. The performance of a model is demonstrated through risk score distribution, discrimination, predictive ability, and calibration (Figure 2).

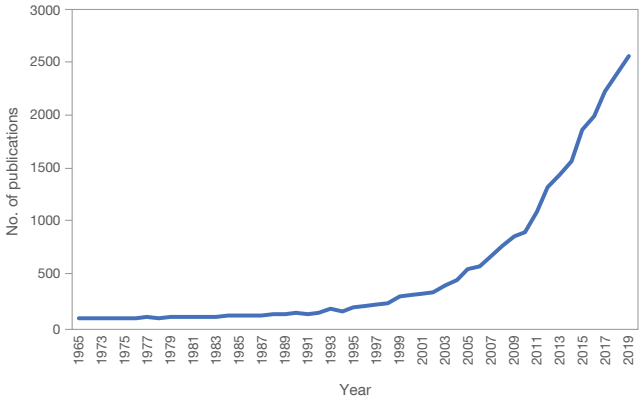

Figure 1: Number of publications with PRS or GRS—A keyword search for "PRS" and "GRS" in publications in the PubMed database shows the dramatic increase in recent years.

PRS construction

Variants included in PRS construction are selected from a discovery cohort. Either whole-genome sequencing (WGS) or genotyping data can be used. Common variants (and their effect sizes) are typically selected from GWAS summary statistics, which are often made available publicly (eg, via the NHGRI-EBI GWAS Catalog). Because only a small portion of the genome is analyzed during genotyping, uncalled variants need to be imputed based on known haplotypes. Imputation can include several reference panels, eg, the 1000 Genomes Project (1KGP) and the Haplotype Reference Consortium (HRC). There are several commonly used imputation tools, such as Beagle 5.0 (or an adapted version) and Minimac. Free imputation servers, such as the TopMed imputation server or the Michigan University imputation server (General Data Protection Regulation (GDPR)-compliant), have imputation panels available.

The Summary Statistics of a GWAS need to undergo a variant selection and/or variant effects modelling to improve the predictive power of the PRS. Four published methods include:

1. **Pruning and thresholding (P+T)**: This "standard" method will set a p-value as a threshold for SNP selection and uses informed linkage disequilibrium (LD) pruning to discard SNPs in LD at a set threshold. The Pruning method is also referred to as "Clumping" (or C+T) when the variant with the highest effect size is selected in each LD block. Both steps are arbitrary. The optimum p-value threshold to impose on the discovery sample depends on its sample size and the genetic architecture of the trait.[5]

2. **Bayesian PGS**: This approach, also referred to as "LDPred," can outperform P+T, particularly with large sample sizes. The method requires the definition of a tuning parameter (ρ), which is an estimate of genetic variants assumed to be causal. P-value thresholds are varied and multiple "LDPred risk scores" are calculated with the use of priors with varying fractions of markers with nonzero effects. A number of PGSs are calculated For C+T with varying ρ.[6]
3. **Stacked clumping and thresholding (SCT)**: This recently developed method involves machine learning that combines C+T, least absolute shrinkage and selection operator (LASSO), and ridge statistical procedure for regression analysis. SCT uses per-SNP effect sizes and p-values to perform repeated P+T/C+T over a four-dimensional grid of parameters (LD squared correlation, p-value threshold, clumping window size, and imputation quality).[7]
4. **Meta-scoring**: This approach combines multiple polygenic scores for a trait/disease into a meta-score (a metaGRS or metaPGS). The method assumes that each individual polygenic score will suffer, to an extent, from regression dilution bias and, therefore, combining them into a more powerful meta-score will reduce this bias. The meta-score itself is constructed as a linear (or non-linear) combination of individual polygenic scores, and typically includes those for both the target trait and related traits.[2]

For each method, many PRSs are built using different parameter values. The PRSs must then be validated. Automated quality control (QC) is possible with several software tools: PLINK, PRSice, LDPred.[8]

PRS validation

Currently, the gold standard for validation of the PRSs, developed from the discovery cohort, is to validate against an independent data set, ideally biobank-scale and inclusive of diverse ancestries.

PRS testing

The output of the validation phase is the selection of an optimal PRS, with the highest performance in accurately predicting the trait of interest (Figure 3). Finally, the external validation phase involves computing the PRS in an independent set of individuals or population, unseen by the PRS construction process, and assessing its predictive power to confirm its predictive performance and to minimize the possibility of overfitting during PRS construction.

Due to its size, the UK BioBank allows for the building of large independent data sets, with the caveat that the ~500K individuals in the UK BioBank are mainly from European descent. Moreover, most GWASs have been performed in European data sets, potentially limiting the utility of the GWAS summary statistics in other ancestries that are characterized by different LD patterns. This overrepresentation of participants of European ancestry results in PGSs having less optimal predictive power for other ethnic groups.[9,10] Several published articles provide more details regarding PRS development.[8–12]

After a PRS for a specific trait or disease outcome has been established, tested, and validated, it can be used to calculate relative genetic risk for traits/diseases in new samples and cohorts. One promising application for PRSs is integration as an additional risk factor in absolute risk models, eg, the Breast and Ovarian Analysis of Disease Incidence and Carrier Estimation Algorithm (BOADICEA).

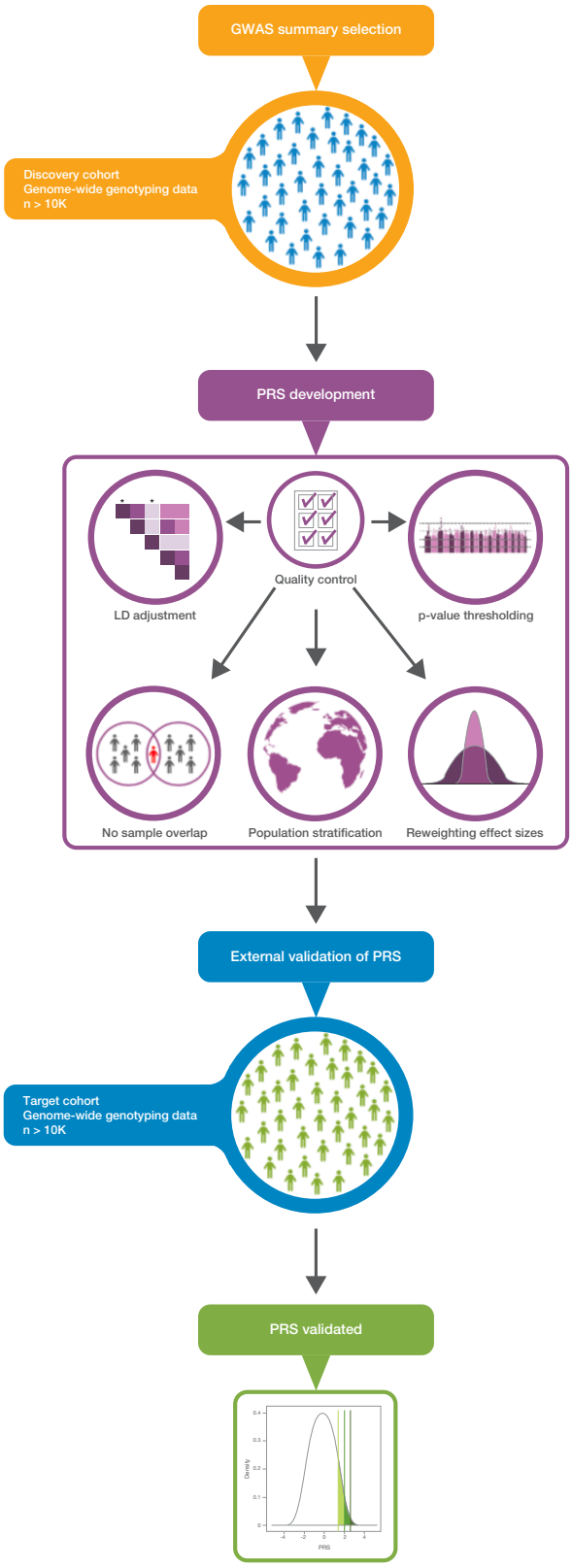

Figure 2: PRS development overview—Development of a PRS is a multistage process that proceeds from initial variant selection to validation and testing of various possible PRSs to determine the most successful model.

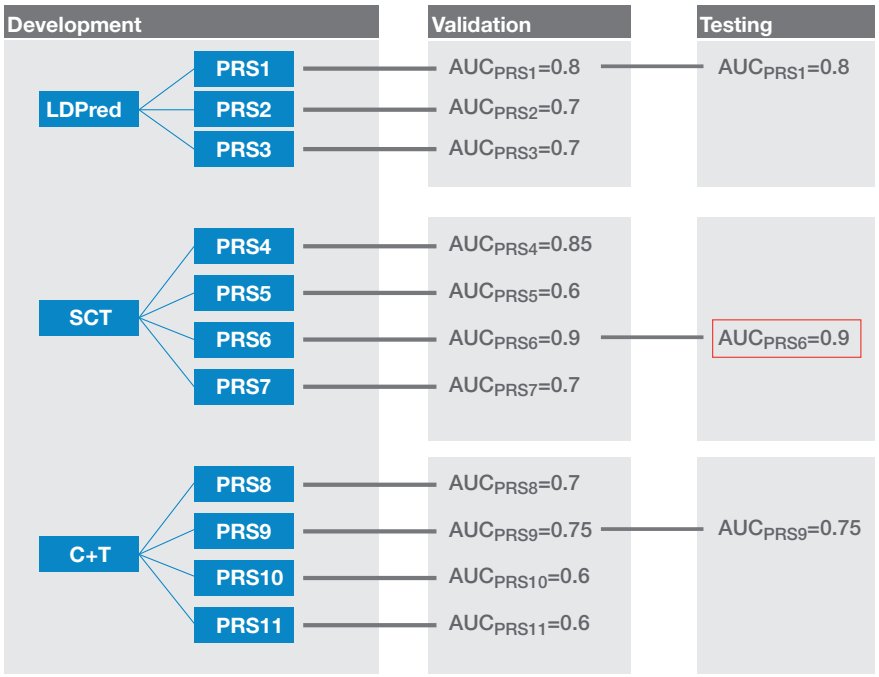

Figure 3: Schematic representation of PRS selection process—Multiple PRSs are created during development using different methods that are then validated and tested to identify the most successful PRS (red box).

The actual score for a specific trait can be communicated in several ways, depending on its intended application:

- A conventional PRS reports absolute risk (eg, lifetime risk) or relative risk (relative to a reference population number)
- PRS results are often normalized by scaling to a mean of 0 and a standard deviation (SD) of 1 to facilitate interpretation (Figure 4)
- PRS interpretation for screening (eg, prostate cancer/breast cancer) can be done with "age-equivalent risk" (patient's risk is equivalent to a person of x years old)

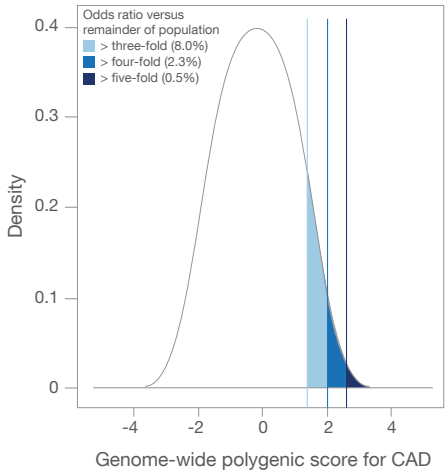

Figure 4: PRS results normalized by scaling—A representative plot of PRS results with values distributed on the X-axis scaled to a mean of 0 with a standard deviation of 1. Shaded portions of the histogram represent proportions of the population with indicated increased risk.[13]

Performance reporting

In addition to a PRS result, performance metrics of the PRS should be reported that include:

- The hazard ratio (HR) for a trait, calculated per SD increment (one SD increment is associated with x HR with a p-value of y)
- The area under the receiver operating curve (AUC), also known as the C-statistic or C-index, which ranges from 0.5 (no discriminative ability) to 1 (perfect discriminative ability). C-statistics quantify the predictive accuracy of the PRS, relative to other PRSs or nongenetic predictors. A PRS has the potential to improve the AUC of existing risk models (eg, ASCVD, Framingham, QRISK3, SCORE, etc.), allowing for reclassification of average/high risk cases and optimizing clinical care[14]
- Fold increased risk in the tail (usually selected from the 80th to the 99th percentile) of the PRS distribution compared to the remainder. For PRS application in common diseases, an increased risk of at least 3-fold can have relevant clinical applications
- Goodness-of-fit measures, such as R2 (defined as the squared correlation between a phenotype and a predictor of the phenotype)
- The effect of a PRS on a patient cohort can also be evaluated by calculating the Net Reclassification Index (NRI) to verify whether the addition of the new predictor to an absolute risk model results in differential classification of individuals across thresholds[15–16]

Standardized reporting

Currently, implementation of PRSs in the clinical setting is hampered by the lack of regulated, standardized reporting. This challenge limits the use, interpretation, and comparison of PRSs. The Polygenic Score (PGS) Catalog, a central resource of published PGSs, allows for methodological and data transparency. Depositing PRSs/PGSs in a resource such as the PGS Catalog provides a valuable resource for widespread adoption, standardizatrion, and improvement.

Summary

PRSs have become powerful tools for genetic disease research, genetic testing, and clinical trials. Development of a PRS begins with initial construction, followed by validation and testing to determine the optimal PRS. It can then be used to calculate relative genetic risk for diseases in new samples and cohorts.

Learn more

To learn more about polygenic risk scores, visit www.illumina.com/areas-of-interest/complex-disease-genomics.html

References

1. Wand H, Lambert SA, Tamburro C, et al. Improving reporting standards for polygenic scores in risk prediction studies. *Nature*. 2021;591:211–219.
2. Inouye M, Abraham G, Nelson CP, et al. Genomic Risk Prediction of Coronary Artery Disease in 480,000 Adults: Implications for Primary Prevention. *J Am Coll Cardiol*. 2108;72(16):1883-1893.
3. Fahed AC, Wang M, Homburger JR, et al. Polygenic background modifies penetrance of monogenic variants conferring risk for coronary artery disease, breast cancer, or colorectal cancer. *medRxiv*. 2109;19013086; doi: https://doi.org/10.1101/19013086.
4. Mars N, Koskela JT, Ripatti P, et al. Polygenic and clinical risk scores and their impact on age at onset and prediction of cardiometabolic diseases and common cancers. *Nat Med*. 2020;26(4):549-557.
5. Wray NR, Kemper KE, Hayes BJ, Goddard ME, Visscher PM. Complex Trait Prediction from Genome Data: Contrasting EBV in Livestock to PRS in Humans. *Genetics*. 2019;211(4):1131–1141.
6. Vilhjálmsson BJ, Yang J, Finucane HK, et al. Modeling Linkage Disequilibrium Increases Accuracy of Polygenic Risk Scores. *Am J Hum Genet*. 2015;97(4):576–592.
7. Privé F, Vilhjálmsson BJ, Aschard H, Blum MGB. Making the most of Clumping and Thresholding for polygenic scores. bioRxiv. 2019;105(6):1213-1221.
8. Choi SW, Mak TSH, O'Reilly PF. A guide to performing Polygenic Risk Score analyses. *bioRxiv*. 2018; 416545;doi: https://doi.org/10.1101/416545.
9. Martin AR, Kanai M, Kamatani Y, Okada Y, Neale BM, Daly MJ. Clinical use of current polygenic risk scores may exacerbate health disparities. *Nat Genet*. 2019;51(4):584-591.
10. Martin AR, Gignoux CR, Walters RK, et al. Human demographic history impacts genetic risk prediction across diverse populations. *Am J Hum Genet*. 2017;100(4):635-649.
11. Lambert SA, Abraham G, Inouye M. Towards clinical utility of polygenic risk scores. *Hum Mol Genet*. 2019;28(R2):R133-R142.
12. Chatterjee N, Shi J, Garcia-Closas M. Developing and evaluating polygenic risk prediction models for stratified disease prevention. *Nat Rev Genet*. 2019;(7):392–406.
13. Khera AV, Chaffin M, Aragam KG, et al. Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations. *Nat Genet*. 2018;50(9):1219-1224.
14. Elliott J, Bodinier B, Bond TA, et al. Predictive Accuracy of a Polygenic Risk Score–Enhanced Prediction Model vs a Clinical Risk Score for Coronary Artery Disease. *JAMA*. 2020;323(7):636-645.
15. Cook NR. Use and misuse of the receiver operating characteristic curve in risk prediction. Circulation. 2007;115(7): 928-935
16. Janssens AC, Martens FK. Prediction Research - An Introduction. 2018; Version 2.2.

Illumina • 1.800.809.4566 toll-free (US) • +1.858.202.4566 tel • techsupport@illumina.com • www.illumina.com