Published in final edited form as:

Trends Genet. 2019 May ; 35(5): 371–382. doi:10.1016/j.tig.2019.02.005.

Progress in Polygenic Composite Scores in Alzheimer's and other Complex Diseases

Danai Chasioti[1,2,3], **Jingwen Yan**[1,2,3], **Kwangsik Nho**[1,2,3], and **Andrew J. Saykin**[2,3,4]

[1]Department of BioHealth Informatics, Indiana University-Purdue University Indianapolis, IN, 46202, USA

[2]Indiana Alzheimer Disease Center, Indiana University School of Medicine, Indianapolis, IN, 46202, USA

[3]Department of Radiology and Imaging Sciences, Indiana University School of Medicine, Indianapolis, IN, 46202, USA

[4]Department of Medical and Molecular Genetics, Indiana University School of Medicine, Indianapolis, IN, 46202, USA

Abstract

Advances in high-throughput genotyping and next generation sequencing coupled with larger sample sizes brings the realization of precision medicine closer than ever. Polygenic approaches incorporating the aggregate influence of multiple genetic variants can contribute to a better understanding of the genetic architecture of many complex diseases and facilitate patient stratification. This review addresses polygenic concepts, methodological developments, hypotheses, and key issues in study design. Polygenic risk scores (PRS) have been applied to many complex diseases and here we focus on Alzheimer's disease (AD) as a primary exemplar. This review was designed to serve as a starting point for investigators wishing to employ PRS in their research and those interested in enhancing clinical study designs through enrichment strategies.

Keywords

polygenic risk score; polygenic hazard score; statistical power; linkage disequilibrium; heritability; Alzheimer's disease

Introduction

Polygenic Landscape of Complex Diseases

The hypothesis of multifactorial etiology of complex diseases originated in Fisher's 1918 quantitative demonstration that human variability in traits such as height and other biometric

characteristics can be explained by the additive effect of multiple genetic factors [9]. Unlike the single-gene etiology of Mendelian diseases, complex diseases are influenced by multiple gene variants and environmental factors [10]. The individual effects of these variants are usually very small [11] making determination of the genetic architecture of complex diseases challenging. Combinatorial genetic metrics such as the PRS and its variations are designed to address these challenges. A variation of PRS using a different type of single nucleotide polymorphism weights (SNP, see Glossary) is the polygenic hazard score (PHS) [3], with the latter utilizing hazard ratios (HR) instead of odds ratios (OR) as SNP weights in the score. Whereas the focus of this review is disease specific, the combinatorial genetic metrics described here are also generalizable to all types of quantitative traits.

The PRS expresses the cumulative genetic risk for an individual as an additive function of the effect of each genetic marker. Polygenic methods have been widely utilized to investigate many diseases, e.g., congenital malformations [12], breast cancer [4, 13], type 2 diabetes (T2D) [14], schizophrenia and other psychiatric disorders [15, 16], and Alzheimer's disease (AD) [3, 17]. Use of PRS for risk stratification and classification is contributing toward the goals of precision medicine. This is enabled by advances in high-throughput genotyping and next generation sequencing (NGS) and the availability of large-scale genome-wide association studies (GWAS), which continuously expand the list of disease-related genetic markers [18]. Additional PRS applications include patient stratification [3, 14, 19, 20], exploration of genetic architecture [13, 21, 22], and studies of genetic overlap between traits [4, 15, 23].

Several review articles have been dedicated to facets of research on PRS [23– 28]. Some of the methodological aspects that influence PRS in the context of psychiatric disorders were discussed in [27]. In [25], the authors systematically reviewed the association of schizophrenia-related PRS with different phenotypes; others mainly focus on disease-specific findings (e.g., [23, 24, 26]) or do not examine methodological factors related to the development and application of PRS.

Here, we review key methodological issues to assist researchers interested in employing PRS for studies of complex diseases and clinicians interested in potential future clinical applications in precision medicine. We overview the state-of-the-art methods for PRS construction and discuss study design and disease characteristics related to performance. Finally, we provide an overview of the contributions of PRS to a wide spectrum of diseases and a detailed overview of applications to Alzheimer's disease.

Calculation of Polygenic Composite Scores

By combining the influence of each SNP into a single measure, the PRS represents the aggregate influence of the genetic variation. There are two approaches for PRS calculation: 1) simple sum of SNPs, and 2) weighted sum of SNPs (Figure 1 and Box 1). The first approach [4, 14, 29, 30] assumes that the disease risk is equally influenced by each SNP. That is rarely realistic as some variants carry a much larger contribution to disease heritability (e.g., the *APOE* ε4 allele in AD [31]). In the weighted sum approach, each SNP is weighted by its estimated disease effect size, therefore accounting for its unique contribution to disease risk or outcome [1–8, 13–17, 19–21, 32–53] (See Table S1 for

examples of methods with publicly available software). Next, we discuss extensively two critical methodological aspects for the PRS development: SNP selection and weight estimation.

SNP selection

The candidate SNP selection is critical because they constitute the PRS's building blocks. A simple strategy is to retain all the SNPs without filtering. This may be effective for genetically underexplored diseases with many small to moderate SNP effects. However, the PRS's performance may suffer by incorporating many non-informative or very weakly associated SNPs. Alternatively, one can retain a subset of SNPs based on predefined criteria (e.g., those passing an arbitrary p-value threshold in the GWAS results). This ad-hoc cut-off selection, however, may omit some informative markers with small effect size. Thus, the PRS-disease association may significantly vary under different thresholds [15, 35, 51]. Another challenge is redundancy of informativeness of variants, especially in the presence of linkage disequilibrium (LD) where nearby SNPs have highly similar associations. This can be addressed by SNP filtering techniques such as LD pruning followed by p-value thresholding. The majority of the SNPs in a LD block are removed by random pruning or clumping (see Box 2). The remaining SNPs are further filtered by thresholding their p-values. PRSice is an example of a software approach employing LD pruning for automated calculation of the PRS [54]. It allows SNP selection under a range of p-value thresholds offering a more precise cut-off choice. One caution is that overfitting issues may arise based on threshold selection criteria [55, 56].

Stepwise regression can also be used for SNP selection [3, 5, 7, 8]. It retains a SNP depending on whether it significantly improves the model's predictive ability. This purely statistical approach has the disadvantage of ignoring prior knowledge of LD structure and possible disease-variant relations.

SNP-weight calculation

Another key factor for PRS performance is the choice of SNP weights. GWAS-derived statistics or risk estimations (e.g., ORs) on an independent sample are commonly used as PRS weights [3, 5, 8, 13, 57]. An extension of this approach that has been promising in AD research is PHS [3, 5, 8, 13, 57]. The PHS is also derived as a weighted sum of SNPs but in this case each SNP's weight is expressed by an HR estimated using a survival model where SNPs are entered as predictors.

GWAS genotypes in a PRS discovery sample may not be representative of those in the validation or application set leading to attenuated performance of the PRS. Other factors that influence performance are LD and regression to the mean or "winner's curse". Adjusting SNP weights may help address these concerns. Next, we consider the two main approaches to optimized SNP re-weighting: 1) those based on Bayesian inference and 2) those based on frequentist inference.

LDpred [58] (Supplementary materials) uses known LD structure as a prior to derive new SNP weights, without requiring raw genotype data or p-value thresholds. When applied on simulation data, LDpred demonstrated improved trait prediction accuracy compared to

traditional methods without LD information [58]. AnnoPred [59] (Supplementary materials) further improved LDpred, by assuming that each SNP's biological identity contributes to the SNP-specific heritability. With this additional assumption and tested on 5 diseases, AnnoPred achieved higher precision in weight estimation (using functional annotation as a prior), better prediction accuracy of disease status, and better risk stratification ability, compared to LDpred [59]. Another Bayesian based method [44] is the *doubly-weighted* PRS (Supplementary materials), which addresses the "winner's curse". It weights each SNP by both its estimated effect on the trait and the probability that its p-value is less than a cut-off. In a study of prevalent T2D, inclusion of the *doubly-weighted* PRS in a logistic model showed significantly better fit than the model with the conventional GWAS-based weighted PRS. Although evidence was not presented in their study, the authors propose that, their method reduces "winner's curse" bias compared to the conventional GWAS-based weighted PRS. The efficiency of the aforementioned methods is highly dependent on parameter tuning. An alternative Bayesian method that requires no parameter tuning [56] (Supplementary materials), corrects a SNP effect by utilizing GWAS *z-statistics* and by assigning a probability for the SNP being not causal (Supplementary materials).

Frequentist approaches, including shrinkage regression (e.g., Least Absolute Shrinkage and Selection Operation (Lasso) [60]) and linear mixed models (LMM) (e.g., GeRSI [61]), have also been utilized for PRS calculation. Shrinkage methods, which penalize the SNP effect estimates to avoid overfitting, show higher precision and power, compared to univariate tests [62]. They successfully handle LD, SNP interactions, and non-genetic covariates [63]. Lasso estimates minimize the sum of squared residuals and assign a penalty on the absolute sum of the predictors' coefficients. Hence, less informative predictors are assigned smaller weights or removed from the model. Lassosum [64] is an example that applies a Lasso-type formula for SNP effect estimation. Despite the need for parameter tuning, it is computationally appealing and outperforms both pruning-thresholding and LDpred methods [64]. LMM, by contrast, treats the most significant SNPs as fixed effects with regard to disease status, and less significant SNPs as having random effects [61]. Here, the fixed effect SNPs are treated as parameters that need to be individually estimated, whereas the random effect SNPs do not require individual estimation since they are considered to be random variables with a common distribution. Both methods, however, are based on distributional assumptions of the genetic effects. Specifically, the shrinkage methods assume a skewed effect distribution, where the majority of the SNPs have small effects and only few have large effects; LMM assumes a normal distribution of effects. If these assumptions are violated, the PRS performance may suffer. To overcome this issue, "hybrid" methods such as Bayesian sparse linear mixed model (BSLMM) [65, 66] and LMM-Lasso [65, 66] were developed that combine the LMM and regularization methodologies.

Both Bayesian and frequentist methods can be further improved by embedding non-genetic information. For example, age-specific OR was used in [67] as *APOE* weight in the PRS and showed significantly improved discriminative power compared to a simple weighted score. In [3] PHS approach employed age-specific PRS weights. Although SNP selection and weighting are key elements in PRS performance, other factors also play an important role. We next consider factors influencing power and accuracy.

Power and accuracy of polygenic composite score

In addition to the methodological factors discussed above, PRS performance is also influenced by study design. As a screening tool, PRS performance should be assessed by the appropriate metrics [28]. More specifically, polygenic composites should not be considered diagnostic tests and the metrics for evaluation are those for a susceptibility screening rather than diagnostic instrument. Thus, the area under the curve (AUC) might be for example more appropriate for assessing the PRS stratification performance on diagnostic markers, and not on disease status directly [28]. Here, we describe some of the factors that could influence analysis results (Fig 2).

Although heritability, as a population metric, is not directly relevant to personal prediction, several aspects of PRS performance are strongly influenced by heritability, necessitating application of PRS in an appropriate population context. For example, while power and accuracy are positively correlated with sample size [59, 68], they are also influenced by the disease heritability. Thus, sample size requirements for achieving the maximum possible AUC, vary based on the heritability [18]. However, heterogeneity problems may arise as the sample size increases [18]. An alternative strategy for power improvement addresses variation of the p-value thresholds for SNP selection. The optimal p-value threshold is determined by the underlying genetic architecture of the disease and the sample size [18, 69]. For example, loosely defined traits (e.g., heterogeneous psychiatric disorders) will benefit more by a relaxed p-value threshold, compared to strictly defined traits (e.g., diseases with a small number of informative SNPs, such as myocardial infarction or stroke) [68]. The heritability of loosely defined traits spreads among a larger number of genetic markers and a relaxed cut-off allows more heritability to be explained. However, threshold increment should be made cautiously as it is usually accompanied by Type I error increase and power reduction [69], which may lead to biased effect estimates with high levels of LD. In contrast, a strict p-value cut-off will be more beneficial for strictly defined traits by eliminating non-informative SNPs [18]. In some cases, the desired performance for a given trait cannot be achieved using only genetic data and incorporation of additional information (e.g., functional annotation of the PRS markers [59] and pathway specific PRS [70]) may be beneficial.

As in all research, study goals should be clearly and operationally defined. Since PRS is used either for association analysis or for individual prediction, the sample requirements vary in each case. In [18] it is suggested that sample sizes are adequate to ensure a well powered association study when independent datasets for training and testing are available. If the latter is not possible, 1:1 splitting ratio between the two sets is advised [18]. In contrast, individual prediction requires a significantly larger training set compared to the testing set [18]. PRS may be unable to successfully discriminate risk groups when there are limited training sample sizes, which attenuates precision in the PRS-explained variation [18].

Additionally, false positive results can occur from the presence of population stratification, due to systematic genetic differences among populations [15, 71]. Mainly implemented using European populations due to greater availability of samples, polygenic scores have

ancestry-specific characteristics that limit application across populations. Thus, in multi-ethnic samples, population structure should be controlled to avoid such bias.

Polygenic risk score applications

Existing research using PRS mainly focuses on two problems: 1) association analysis and 2) outcome prediction. Although use of PRS has not achieved clinical accuracy levels yet, its use has led to some interesting discoveries and shown potential in diseases like cancer [2, 4, 7, 13, 47], psoriasis [19], rheumatoid arthritis (RA) [19], multiple sclerosis (MS) [32], mental disorders [15, 16], atherosclerosis [46], T2D [2, 14, 30, 44], asthma [29], Parkinson disease (PD) [21, 41], and cardiovascular diseases (CVD) [20] including coronary heart disease (CHD) [1].

Association analysis quantifies the relation between two sets of features such as phenotype and genotype (e.g., SNPs). In this context, PRS is used to assess the differential biology between disease types or stages [13, 16, 48], to identify risk strata [19, 72], to assess treatment response [46, 73] and to identify genetic overlap between diseases [4, 15]. Association of a simple sum PRS with T2D risk indicated that, men and women in the highest PRS quantile had ~2.8 and ~2.2 times higher risk of developing T2D respectively, compared to those in the lowest PRS quantile [14]. Similar findings were reported with a GWAS weighted PRS [14]. Another study showed that adopting a healthy lifestyle can reduce the CVD risk, regardless of the individual's genetic background [20]. High genetic risk participants with healthy lifestyle had 46% lower risk of CVD, compared to those with unhealthy lifestyle [20].

For breast cancer patients, significant PRS differences were observed between screen-detected and interval breast cancer cases, indicating the possibility of differential biology underlying the two breast cancer subtypes [13].

PRS also can help with therapy selection for disease prevention. In [46], statin therapy significantly reduced the relative CHD risk in high genetic risk patients (>80th percentile) as compared to patients with low genetic risk.

The PRS has also been employed to explore genetic overlap between different diseases (e.g., application of schizophrenia-specific PRS to bipolar disorder [15]), where the PRS derived from one disease is evaluated in another disease. Motivated by this, the recently proposed *multi-polygenic score* (MPS) [74], combines multiple PRSs from different GWASs, for outcome prediction. Compared to a single PRS, this method explained more variability when applied to three traits (i.e., BMI, educational achievement and cognitive ability). The increased predictive power that MPS achieves should be useful in situations of modest sample size [74].

As an individual prediction tool, PRS has also shown potential in screening studies. For example, in a study on aggressive prostate cancer (PCa), using PHS it was observed that, males in high genetic risk (>98th centile) have almost triple PCa hazard, compared to those in average genetic risk [7]. For PCa patients who had undergone radical prostatectomy, PCa recurrence was predicted with AUC= 88.8% [47]. Moreover, the 10-year recurrence-free rate

for those in high genetic risk is almost half (46.3%), compared to people in the lowest genetic risk group (81.8%).

Although PRS approaches are still experimental, future application in public health and preventative and therapeutic medicine holds significant potential including quantitating the overall burden of genetic risk factors in various subpopulations (primary prevention), identifying high risk individuals who warrant screening for disease (secondary prevention) or serving as a stratification biomarker for treatment optimization (tertiary prevention).

Polygenic risk score in Alzheimer's disease

Late onset AD (LOAD) is a highly prevalent neurodegenerative dementia characterized pathologically by brain accumulation of amyloid beta (Aβ) plaques and neurofibrillary tangles composed of hyperphosphorylated tau. These classic pathological hallmarks of AD are only the most obvious manifestation and belie a broad array of pathophysiological changes affecting numerous systems within the brain and periphery. A small percent of AD cases, typically with an early onset (EOAD) and aggressive course, are monogenic with an autosomal dominant inheritance pattern. Over 95% of AD is genetically complex, highly heritable, and therefore well-suited to polygenic investigation including analysis of heterogeneity and subgroups to support development of a precision medicine approach. Since the mechanistic drivers of LOAD remain unclear, substantial effort is being dedicated to genetic risk score modelling for individual risk prediction and to a systems approach to understanding disease pathogenesis.

APOE ε4, the strongest genetic variant associated with increased risk and earlier onset of LOAD, only partially accounts for the estimated heritability [31]. The contribution of other genetic markers has frequently been highlighted by PRS [49, 67, 75–77] (for a list of SNPs included in published AD PRS, see Table S2 in Supplementary materials). One PRS study including 19 non-*APOE* SNPs successfully stratified *APOE* ε4 carriers into risk subgroups where those with the highest scores exceeded the risk of those with the lowest score by 62% [57]. Another PRS study using 31 non-*APOE* SNPs found that age at onset (AAO) of AD is modulated by the genetic score [3]. *APOE* ε3/ε3 carriers in the highest AD risk stratum, could progress to AD as many as 10 years faster than those in the lowest group [3]. In [77] it was shown that, the PRS predictive accuracy in a neuropathologically confirmed sample does not change significantly after removing *APOE* ε4 and ε2 carriers, indicating similar genetic architecture among the *APOE* genotypes. Non-*APOE* PRS has also been associated with disease stage and progression (e.g., MCI-converters [34] and cognitively normal individuals [3, 36]), suggesting that genetic contributions to AD manifest in a stage-specific manner [36]. Furthermore, non-*APOE* PRS have been used for AD-patient classification [3, 8, 49, 57, 67, 75, 76, 78–80] and AD-subtype discrimination [36], which has helped to reveal diverse mechanisms underlying various AD subtypes.

In addition to clinical indicators of disease status, endophenotypes such as cerebrospinal fluid (CSF) and MRI and PET imaging measures are important AD biomarkers. In most studies, their relation to the genetic composite score was either driven by the *APOE* [38] or could not be established [35, 38, 70, 81] (possibly due to low statistical power and a small number of SNPs in the PRS [39, 74, 81]). One study [17] observed that relaxing the SNP

inclusion threshold from the conventional GWAS-based $p<5\times10^{-8}$ to a nominal $p<0.01$ led to several associations becoming significant, even after excluding *APOE*. This result, however, was not replicated in other studies [36, 74]. The optimal threshold remains an open question and may be related to multiple factors as discussed above.

Accepted CSF biomarkers for AD include Aβ1–42, total tau (t-tau), and phosphorylated tau (p-tau). However, the relation between genetic scores and these CSF biomarkers has not been consistent. Genetic association studies of $A\beta_{1-42}$ with non-*APOE* PRS were not successful in the past [67, 70]. The evidence for the PRS's relation to p-tau [67], t-tau [3, 67] and p-tau/ $A\beta_{1-42}$ ratio [76] remains limited. Recently, it was observed [82] that PHS is associated with increased intracranial Aβ plaque accumulation over time (p-value = 1.28×10^{-7}). In another study [37], the variability explained for $A\beta_{1-42}$ was increased by 1.8%, when in addition to *APOE* other markers were included in PRS.

For neuroimaging measures, many studies have failed to detect a significant association of PRS with baseline AD imaging phenotypes (e.g., hippocampal volume) in cognitively normal older adults [81], young adults and older individuals with MCI [80]. However, when older adults from 4 cohorts where combined into one large sample (>1,600 individuals), the same analysis revealed significant association of the PRS with the mean hippocampal volume at the baseline [80]. In a more recent study [3], PRS was associated with longitudinal volume loss, in both hippocampal and entorhinal cortex areas. In cognitively normal adults, a PRS was marginally associated with annual cortical thinning rates [53] and significantly associated with bi-annual hippocampal complex thinning rates [81].

Currently, PRS seems to be a useful tool for predicting the AAO of AD [3, 5, 67, 75, 76] for both sporadic late and early onset [76], even after excluding *APOE*. However, the degree of prediction varies across studies. One unit increase in the non-*APOE* PRS is estimated to accelerate the AAO by 8 months to a year [75, 76]. Another study with >1,300 AD patients suggested that, a unit increase in PRS (22 IGAP SNPs, including *APOE*) decreases the AAO by up to 2.4 years [67]. As above, *APOE* e3/e3 homozygotes showed PRS strata differences in AAO can reach 10 years [3].

Other important PRS applications include subtype stratification and prediction of disease trajectory. Prediction analysis requires larger sample sizes compared to association analysis [18] but the goal of prognostic prediction may be within range. The AD heritability explained by additive genetic effects as captured by GWAS is estimated to be 24%−33% [31, 83] with the majority attributed to *APOE* [82]. The sample size required to observe reliable PRS effect for prediction is a function of disease heritability [18]. The largest AD GWAS [84] included 25,580 AD cases and 48,466 controls. As sample sizes continue to increase rapidly, PRS performance is expected to soon reach levels acceptable for clinical application in a susceptibility screening framework. Ongoing efforts to improve the accuracy and interpretability of PRS can also be expected to advance our knowledge about AD pathogenesis and help to identify new combinatorial diagnostic/biomarker strategy for the early intervention (for a hypothesis-based list of AD-PRS studies, see Table S3 in Supplementary materials).

Concluding Remarks

Polygenic composite score approaches have been used to identify optimized sets of SNPs whose cumulative genetic effect can better identify susceptibility and predict AAO and phenotypic features that characterize complex diseases. With applications in a wide range of diseases, PRS, the most common genetic composite score, has promise for patient screening and genetic enrichment for therapeutic intervention trials. As sample sizes continue to increase rapidly, PRS performance is expected to soon reach levels sufficient for clinical application in susceptibility screening and stratification for clinical trials within appropriate populations. Although PRS is neither designed to be a diagnostic test nor sufficiently accurate for clinical diagnosis, important applications of PRS in addition to risk stratification include subtype stratification and prediction of disease trajectory. PRS used in this matter are consistent with FDA draft guidance on enrichment strategies [85] and could be used to improve clinical trials by decreasing heterogeneity, increased prognostic accuracy, and enhanced prediction of treatment response. In AD research, PRS have contributed to risk stratification for early detection and helped to elucidate the genetic contribution to disease endophenotypes.

Despite the advances in PRS methodologies discussed above, current polygenic composite score approaches have limitations, including extent of ability to account for disease heritability and insufficient development for full clinical deployment in precision medicine. A number of strategies may lead to better PRS performance (see Outstanding Questions). While current methods focus on additive effects and common variants, future approaches may incorporate the potential role of epistasis and gene-environment interactions, transcriptomic and epigenetic variation, and other patient information through combinatorial strategies. Recent advances in machine learning can be expected to improve PRS models. Another limitation is interpretability. PRS reflect enriched pathways but the downstream mechanisms through which they influence disease is not identified. New computational biology tools and databases can be expected to enhance interpretation of polygenic effects. Future polygenic models developed in relation to quantitative endophenotype data from disease specific biomarkers hold promise for clinically and mechanistically useful prediction. We can look forward to further development of these methods to support the evolving precision medicine of complex disease.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

This work was supported in part by the following NIH grants: P30 AG010133, R01 AG019771, R01 LM011360, R01 CA129769, U01 AG024904, R01 LM012535 and R03 AG054936. Additional support was provided by the Indiana University Network Science Institute (IUNI), ADNI and ADNI DoD.

Glossary

Apolipoprotein E (*APOE*)

APOE is the gene which codes for the synthesis of the protein apolipoprotein E. Specific mutation in this gene has been found to increase the risk of Alzheimer's disease as much as 12 times.

Area under the curve (AUC)

AUC expresses the predictive accuracy of a test on a binary trait. A value of 1 represents a perfect test, while 0.5 shows a test with no better accuracy than chance. In a clinical setting, AUC ≥0.75 is required for screening patients at risk, while AUC ≥0.99 for screening the general population.

Cerebrospinal Fluid (CSF)

Fluid found in and around the brain and spinal cord that reflects the biochemical changes in the brain and is an important biomarker for AD and other brain disorders. Three most commonly studied CSF biomarkers are for AD include: total tau (t-tau), phospho-tau (p-tau), and the 42 amino acid form of β-amyloid (Aβ42).

Late Onset Alzheimer's Disease (LOAD)

Late onset AD, usually defined as onset after age 65, is the most common form of AD. LOAD is genetically complex and highly heritable. Although no deterministic genetic variants have been found, *APOE* e4 allele is currently the strongest genetic risk factor.

Linkage Disequilibrium (LD)

The non-random association between alleles at different loci on the same chromosome. Alleles in LD appear together more (or less) often than expected by chance.

Single nucleotide polymorphism (SNP)

The most common DNA variation. It occurs when a nucleotide in the genome (Adenine: A, Guanine: G, Cytocine: C, Thymine: T) is replaced by another nucleotide. These variations are commonly used in the gene-trait association studies.

Winner's curse

Inflated estimation of the effect of genetic variants selected based on a specific threshold. SNPs that pass the threshold in any given study are typically overestimated compared to the true effect size. This overestimated effect is sample-specific and sample size dependent and frequently leads to difficulty replicating association studies.

References

1. Ripatti S, et al., A multilocus genetic risk score for coronary heart disease: case-control and prospective cohort analyses. Lancet, 2010 376(9750): p. 1393–400. [PubMed: 20971364]
2. Raynor LA, et al., Pleiotropy and pathway analyses of genetic variants associated with both type 2 diabetes and prostate cancer. Int J Mol Epidemiol Genet, 2013 4(1): p. 49–60. [PubMed: 23565322]
3. Desikan RS, et al., Genetic assessment of age-associated Alzheimer disease risk: Development and validation of a polygenic hazard score. PLoS Med, 2017 14(3): p. e1002258. [PubMed: 28323831]
4. Kuchenbaecker KB, et al., Evaluation of Polygenic Risk Scores for Breast and Ovarian Cancer Risk Prediction in BRCA1 and BRCA2 Mutation Carriers. J Natl Cancer Inst, 2017 109(7).
5. Tan CH, et al., Polygenic hazard scores in preclinical Alzheimer disease. Ann Neurol, 2017 82(3): p. 484–488. [PubMed: 28940650]

6. Ten Broeke SW, et al., SNP association study in PMS2-associated Lynch syndrome. Fam Cancer, 2017.
7. Seibert TM, et al., Polygenic hazard score to guide screening for aggressive prostate cancer: development and validation in large scale cohorts. BMJ, 2018 360: p. j5757. [PubMed: 29321194]
8. Tan CH, et al., Polygenic hazard score: an enrichment marker for Alzheimer's associated amyloid and tau deposition. Acta Neuropathol, 2018 135(1): p. 85–93. [PubMed: 29177679]
9. Fisher RA, XV.—The Correlation between Relatives on the Supposition of Mendelian Inheritance. Transactions of the Royal Society of Edinburgh, 1918 52(2): p. 399–433.
10. Craig J, Complex diseases: research and applications. Nature Education, 2008 1(1): p. 184.
11. Price AL, Spencer CC, and Donnelly P, Progress and promise in understanding the genetic basis of common diseases. Proc Biol Sci, 2015 282(1821): p. 20151684. [PubMed: 26702037]
12. Carter CO, Genetics of common disorders. Br Med Bull, 1969 25(1): p. 52–7. [PubMed: 5782759]
13. Li J, et al., Breast cancer genetic risk profile is differentially associated with interval and screen-detected breast cancers. Ann Oncol, 2015 26(3): p. 517–22. [PubMed: 25488685]
14. Cornelis MC, et al., Joint effects of common genetic variants on the risk for type 2 diabetes in U.S. men and women of European ancestry. Ann Intern Med, 2009 150(8): p. 541–50. [PubMed: 19380854]
15. International Schizophrenia, C., et al., Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. Nature, 2009 460(7256): p. 748–52. [PubMed: 19571811]
16. Vassos E, et al., An Examination of Polygenic Score Risk Prediction in Individuals With First-Episode Psychosis. Biol Psychiatry, 2017 81(6): p. 470–477. [PubMed: 27765268]
17. Mormino EC, et al., Polygenic risk of Alzheimer disease is associated with early- and late-life processes. Neurology, 2016 87(5): p. 481–8. [PubMed: 27385740]
18. Dudbridge F, Power and predictive accuracy of polygenic risk scores. PLoS Genet, 2013 9(3): p. e1003348. [PubMed: 23555274]
19. Karlson EW, et al., Cumulative association of 22 genetic variants with seropositive rheumatoid arthritis risk. Ann Rheum Dis, 2010 69(6): p. 1077–85. [PubMed: 20233754]
20. Khera AV, et al., Genetic Risk, Adherence to a Healthy Lifestyle, and Coronary Disease. N Engl J Med, 2016 375(24): p. 2349–2358. [PubMed: 27959714]
21. Nalls MA, et al., Genetic risk and age in Parkinson's disease: Continuum not stratum. Mov Disord, 2015 30(6): p. 850–4. [PubMed: 25778492]
22. Broce IJ, et al., Dissecting the genetic relationship between cardiovascular risk factors and Alzheimer's disease. Acta Neuropathol, 2018.
23. Dima D and Breen G, Polygenic risk scores in imaging genetics: Usefulness and applications. J Psychopharmacol, 2015 29(8): p. 867–71. [PubMed: 25944849]
24. Chatterjee N, Shi J, and Garcia-Closas M, Developing and evaluating polygenic risk prediction models for stratified disease prevention. Nat Rev Genet, 2016 17(7): p. 392–406. [PubMed: 27140283]
25. Mistry S, et al., The use of polygenic risk scores to identify phenotypes associated with genetic risk of schizophrenia: Systematic review. Schizophr Res, 2017.
26. Santoro ML, et al., A current snapshot of common genomic variants contribution in psychiatric disorders. Am J Med Genet B Neuropsychiatr Genet, 2016 171(8): p. 997–1005. [PubMed: 27486013]
27. Wray NR, et al., Research review: Polygenic methods and their application to psychiatric traits. J Child Psychol Psychiatry, 2014 55(10): p. 1068–87. [PubMed: 25132410]
28. Torkamani A, Wineinger NE, and Topol EJ, The personal and clinical utility of polygenic risk scores. Nat Rev Genet, 2018 19(9): p. 581–590. [PubMed: 29789686]
29. Belsky DW, et al., Polygenic risk and the development and course of asthma: an analysis of data from a four-decade longitudinal study. The Lancet Respiratory Medicine, 2013 1(6): p. 453–461. [PubMed: 24429243]
30. Layton J, et al., Type 2 Diabetes Genetic Risk Scores Are Associated With Increased Type 2 Diabetes Risk Among African Americans by Cardiometabolic Status. Clin Med Insights Endocrinol Diabetes, 2018 11: p. 1179551417748942. [PubMed: 29326538]

31. Ridge PG, et al., Alzheimer's disease: analyzing the missing heritability. PLoS One, 2013 8(11): p. e79771. [PubMed: 24244562]
32. De Jager PL, et al., Integration of genetic risk factors into a clinical algorithm for multiple sclerosis susceptibility: a weighted genetic risk score. Lancet Neurol, 2009 8(12): p. 1111–9. [PubMed: 19879194]
33. Reeves GK, et al., Incidence of breast cancer and its subtypes in relation to individual and multiple low-penetrance genetic susceptibility loci. JAMA, 2010 304(4): p. 426–34. [PubMed: 20664043]
34. Rodriguez-Rodriguez E, et al., Genetic risk score predicting accelerated progression from mild cognitive impairment to Alzheimer's disease. J Neural Transm (Vienna), 2013 120(5): p. 807–12. [PubMed: 23180304]
35. Harris SE, et al., Polygenic risk for Alzheimer's disease is not associated with cognitive ability or cognitive aging in non-demented older people. J Alzheimers Dis, 2014 39(3): p. 565–74. [PubMed: 24246418]
36. Adams HH, et al., Genetic risk of neurodegenerative diseases is associated with mild cognitive impairment and conversion to dementia. Alzheimers Dement, 2015 11(11): p. 1277–85. [PubMed: 25916564]
37. Martiskainen H, et al., Effects of Alzheimer's disease-associated risk loci on cerebrospinal fluid biomarkers and disease progression: a polygenic risk score approach. J Alzheimers Dis, 2015 43(2): p. 565–73. [PubMed: 25096612]
38. Vivot A, et al., Association of Alzheimer's related genotypes with cognitive decline in multiple domains: results from the Three-City Dijon study. Mol Psychiatry, 2015 20(10): p. 1173–8. [PubMed: 26033242]
39. Yin X, et al., A weighted polygenic risk score using 14 known susceptibility variants to estimate risk and age onset of psoriasis in Han Chinese. PLoS One, 2015 10(5): p. e0125369. [PubMed: 25933357]
40. Holm J, et al., Associations of Breast Cancer Risk Prediction Tools With Tumor Characteristics and Metastasis. J Clin Oncol, 2016 34(3): p. 251–8. [PubMed: 26628467]
41. Pihlstrom L, et al., A cumulative genetic risk score predicts progression in Parkinson's disease. Mov Disord, 2016 31(4): p. 487–90. [PubMed: 26853697]
42. Gibson J, et al., Assessing the presence of shared genetic architecture between Alzheimer's disease and major depressive disorder using genome-wide association data. Transl Psychiatry, 2017 7(4): p. e1094. [PubMed: 28418403]
43. Lacour A, et al., Genome-wide significant risk factors for Alzheimer's disease: role in progression to dementia due to Alzheimer's disease among subjects with mild cognitive impairment. Mol Psychiatry, 2017 22(1): p. 153–160. [PubMed: 26976043]
44. Lall K, et al., Personalized risk prediction for type 2 diabetes: the potential of genetic risk scores. Genet Med, 2017 19(3): p. 322–329. [PubMed: 27513194]
45. Li H, et al., Breast cancer risk prediction using a polygenic risk score in the familial setting: a prospective study from the Breast Cancer Family Registry and kConFab. Genet Med, 2017 19(1): p. 30–35. [PubMed: 27171545]
46. Natarajan P, et al., Polygenic Risk Score Identifies Subgroup With Higher Burden of Atherosclerosis and Greater Relative Benefit From Statin Therapy in the Primary Prevention Setting. Circulation, 2017 135(22): p. 2091–2101. [PubMed: 28223407]
47. Oh JJ, et al., Genetic risk score to predict biochemical recurrence after radical prostatectomy in prostate cancer: prospective cohort study. Oncotarget, 2017 8(44): p. 75979–75988. [PubMed: 29100285]
48. Sengupta SM, et al., Polygenic Risk Score associated with specific symptom dimensions in first-episode psychosis. Schizophr Res, 2017 184: p. 116–121. [PubMed: 27916287]
49. Chaudhury S, et al., Polygenic risk score in postmortem diagnosed sporadic early-onset Alzheimer's disease. Neurobiol Aging, 2018 62: p. 244 e1–244 e8.
50. Hindy G, et al., Polygenic Risk Score for Coronary Heart Disease Modifies the Elevated Risk by Cigarette Smoking for Disease Incidence. Circ Genom Precis Med, 2018 11(1).
51. Logue MW, et al., Use of an Alzheimer's disease polygenic risk score to identify mild cognitive impairment in adults in their 50s. Mol Psychiatry, 2018.

52. Paul KC, et al., Association of Polygenic Risk Score With Cognitive Decline and Motor Progression in Parkinson Disease. JAMA Neurol, 2018.
53. Sabuncu MR, et al., The association between a polygenic Alzheimer score and cortical thickness in clinically normal subjects. Cereb Cortex, 2012 22(11): p. 2653–61. [PubMed: 22169231]
54. Euesden J, Lewis CM, and O'Reilly PF, PRSice: Polygenic Risk Score software. Bioinformatics, 2015 31(9): p. 1466–8. [PubMed: 25550326]
55. Mak TS, et al., Local True Discovery Rate Weighted Polygenic Scores Using GWAS Summary Data. Behav Genet, 2016 46(4): p. 573–82. [PubMed: 26747043]
56. So HC and Sham PC, Improving polygenic risk prediction from summary statistics by an empirical Bayes approach. Sci Rep, 2017 7: p. 41262. [PubMed: 28145530]
57. Chouraki V, et al., Evaluation of a Genetic Risk Score to Improve Risk Prediction for Alzheimer's Disease. J Alzheimers Dis, 2016 53(3): p. 921–32. [PubMed: 27340842]
58. Vilhjalmsson BJ, et al., Modeling Linkage Disequilibrium Increases Accuracy of Polygenic Risk Scores. Am J Hum Genet, 2015 97(4): p. 576–92. [PubMed: 26430803]
59. Hu Y, et al., Leveraging functional annotations in genetic risk prediction for human complex diseases. PLoS Comput Biol, 2017 13(6): p. e1005589. [PubMed: 28594818]
60. Tibshirani R, Regression Shrinkage and Selection via the Lasso. Journal of the Royal Statistical Society, 1996 58(1): p. 22.
61. Golan D and Rosset S, Effective genetic-risk prediction using mixed models. Am J Hum Genet, 2014 95(4): p. 383–93. [PubMed: 25279982]
62. Abraham GK,A; Zobel J; Inouye M; , Performance and Robustness of Penalized and Unpenalized Methods for Genetic Prediction of Complex Human Disease. Genetic Epidemiology, 2013 37(2): p. 184–195. [PubMed: 23203348]
63. de Vlaming R and Groenen PJ, The Current and Future Use of Ridge Regression for Prediction in Quantitative Genetics. Biomed Res Int, 2015 2015: p. 143712. [PubMed: 26273586]
64. Mak TSH, et al., Polygenic scores via penalized regression on summary statistics. Genet Epidemiol, 2017 41(6): p. 469–480. [PubMed: 28480976]
65. Zhou X, Carbonetto P, and Stephens M, Polygenic Modeling with Bayesian Sparse Linear Mixed Models. PLoS Genet, 2013 9(2).
66. Rakitsch B, et al., A Lasso multi-marker mixed model for association mapping with population structure correction. Bioinformatics, 2013 29(2): p. 206–14. [PubMed: 23175758]
67. Sleegers K, et al., A 22-single nucleotide polymorphism Alzheimer's disease risk score correlates with family history, onset age, and cerebrospinal fluid Abeta42. Alzheimers Dement, 2015 11(12): p. 1452–1460. [PubMed: 26086184]
68. So HC and Sham PC, Exploring the predictive power of polygenic scores derived from genome-wide association studies: a study of 10 complex traits. Bioinformatics, 2017 33(6): p. 886–892. [PubMed: 28065900]
69. Chatterjee N, et al., Projecting the performance of risk prediction based on polygenic analyses of genome-wide association studies. Nat Genet, 2013 45(4): p. 400–5, 405e1–3. [PubMed: 23455638]
70. Darst BF, et al., Pathway-Specific Polygenic Risk Scores as Predictors of Amyloid-beta Deposition and Cognitive Function in a Sample at Increased Risk for Alzheimer's Disease. J Alzheimers Dis, 2017 55(2): p. 473–484. [PubMed: 27662287]
71. Reitz C, et al., Variants in the ATP-binding cassette transporter (ABCA7), apolipoprotein E 4,and the risk of late-onset Alzheimer disease in African Americans. JAMA, 2013 309(14): p. 1483–92. [PubMed: 23571587]
72. Khera AV, et al., Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations. Nat Genet, 2018 50(9): p. 1219–1224. [PubMed: 30104762]
73. Inouye M, et al., Genomic Risk Prediction of Coronary Artery Disease in 480,000 Adults: Implications for Primary Prevention. J Am Coll Cardiol, 2018 72(16): p. 1883–1893. [PubMed: 30309464]
74. Krapohl E, et al., Multi-polygenic score approach to trait prediction. Mol Psychiatry, 2017.

75. Tosto G, et al., Polygenic risk scores in familial Alzheimer disease. Neurology, 2017 88(12): p. 1180–1186. [PubMed: 28213371]

76. Cruchaga C, et al., Polygenic risk score of sporadic late-onset Alzheimer's disease reveals a shared architecture with the familial and early-onset forms. Alzheimers Dement, 2018 14(2): p. 205–214. [PubMed: 28943286]

77. Escott-Price V, et al., Polygenic Risk Score Analysis of Alzheimer's Disease in Cases without APOE4 or APOE2 Alleles. J Prev Alzheimers Dis, 2019 6(1): p. 16–19. [PubMed: 30569081]

78. Escott-Price V, et al., Common polygenic variation enhances risk prediction for Alzheimer's disease. Brain, 2015 138(Pt 12): p. 3673–84. [PubMed: 26490334]

79. Escott-Price V, M., A.J; Huentelman M; Hardy J, Polygenic Risk Score Analysis of Pathologically Confirmed Alzheimer Disease. Annals of Neurology, 2017 82(2): p. 311–314. [PubMed: 28727176]

80. Lupton MK, et al., The effect of increased genetic risk for Alzheimer's disease on hippocampal and amygdala volume. Neurobiol Aging, 2016 40: p. 68–77. [PubMed: 26973105]

81. Harrison TM, et al., An Alzheimer's Disease Genetic Risk Score Predicts Longitudinal Thinning of Hippocampal Complex Subregions in Healthy Older Adults. eNeuro, 2016 3(3).

82. Lambert JC, et al., Meta-analysis of 74,046 individuals identifies 11 new susceptibility loci for Alzheimer's disease. Nat Genet, 2013 45(12): p. 1452–8. [PubMed: 24162737]

83. Lee SH, et al., Estimation and partitioning of polygenic variation captured by common SNPs for Alzheimer's disease, multiple sclerosis and endometriosis. Hum Mol Genet, 2013 22(4): p. 832–41. [PubMed: 23193196]

84. Marioni RE, et al., GWAS on family history of Alzheimer's disease. Transl Psychiatry, 2018 8(1): p. 99. [PubMed: 29777097]

85. Administration, F.a.D. Guidance for Industry Enrichment Strategies for Clinical Trials to Support Approval of Human Drugs and Biological Products 2012 [cited 2019 10 Feb]; Available from: https://www.fda.gov/downloads/drugs/guidancecomplianceregulatoryinformation/guidances/ucm332181.pdf

Highlights

- Combinatorial metrics including the polygenic risk score (PRS) summarize the aggregate influence of multiple common genetic variants.
- Recent methodological advances include optimized variant selection and weighing algorithms.
- Despite considerable progress, current polygenic approaches have limitations in their ability to account for heritability and in readiness for clinical implementation.
- Late onset Alzheimer's disease is a highly heritable complex disease that is particularly well-suited for polygenic analysis of heterogeneity and subtypes to support development of a precision medicine approach.
- Polygenic models and metrics based on disease specific biomarkers or endophenotypes hold promise for prognostic prediction and enhanced mechanistic understanding. Eventually sets of PRS employed in combination may help prioritize therapeutic targets on a personalized basis.

Outstanding Questions

- How can the SNP selection process for PRS and other polygenic composite scores be improved?

  Potential strategies include enhanced algorithms including machine learning and incorporation of prior biological knowledge. However, inclusion of non-informative markers will add noise, increase variability and decrease performance accuracy. Penalization approaches may help optimize the signal to noise in PRS development. Methods to optimally incorporate longitudinal disease trajectory in SNP weight estimation also warrant investigation.

- Would strategies for incorporating non-additive genetic (and other) effects improve PRS performance?

  Most current composite models are based on additive genetic effects of common variants. Genetic interaction, both epistasis and gene by environment influences, are neglected, as are rare variants.

- Is there a need for development of sex and ancestry-specific composite scores?

  Individual characteristics such as sex and racial ancestry significantly modify PRS performance. Research is needed to determine whether one model or separate composite scores are needed.

- Can performance be improved by including other omics data such as transcriptomics and epigenetic markings?

  Similarly, incorporating other endophenotypes such as medical imaging or biomarker results might improve the precision and utility of polygenic scores, perhaps in the context of a clinical decision support system.

- How can we enhance the interpretability of genetic composite scores?

  Interpretability of genetic composite scores remains a challenge with current models as they are not constructed to reveal how selected markers interact mechanistically to affect disease outcomes. Precision medicine requires identification of actionable test results that indicate specific therapeutic targets and are clinically meaningful at the individual level. Enhanced genetic counseling approaches addressing the results of composite risk scores vs. single markers are needed to help explain test implications to patients and families.

Box 1

Main PRS calculation categories

There are multiple mathematical formulas for PRS calculation. The simplest way to derive a PRS for an individual i is by calculating the sum over the risk-allele frequencies (d_{ij}) of each SNP j.

$$PRS_i = \sum_{j \in SNPs} d_{ij}$$

Most PRS models assume that SNPs have an additive effect on the disease risk. In this case, the frequency (d_{ij}) takes values 0, 1 or 2, depending on the number of risk alleles present in the gene. Since one can't assume SNP influences are equal, a weighted version of this formula has been proposed.

$$PRS_i = \sum_{j \in SNPs} \beta_j d_{ij}$$

Here, the PRS is expressed as the sum over the weighted number of alleles per SNPs. Depending on the type and the goal of the study, different weights can be utilized. The most commonly used weight is the GWAS odds ratio (OR), or the univariate linear regression coefficient. Recent studies [1–8] have introduced the Cox-derived hazard ratio (HR) as alternative weight, to account for the time to event, which is otherwise ignored when using the GWAS-OR.

Box 2

Pruning and clumping

LD pruning is the process of genetic marker selection based on their LD. The aim is the final set markers to contain those that are nearly uncorrelated.

While clumping retains one SNP per LD block, pruning can end up with multiple SNPs or no SNPs at all for a region.

Specifically, for LD pruning, the pairwise correlation between the markers in a specific range of the genome (window) is calculated. This region is then scanned and if for any pair of markers, the correlation is greater than the specified threshold, the marker with the smallest minor allele frequency (MAF) is discarded, otherwise both markers are retained. In case both markers have the same MAF, the one in the latter position is pruned. The process continues until the whole genome has been scanned. LD clumping, in contrast, identifies all SNPs with GWAS p values meeting a prespecified value ($p1$; default 0.0001). For each of these index SNPs, clumps are generated. The clumps are constituted by those SNPs that have an LD ($r2$; default 0.5) that is at least equal to a prespecified value, lie within a prespecified physical distance from the index SNP, and their GWAS p value is less than a second significance threshold ($p2$; default 0.01). Each index SNP is used as representative of the corresponding LD region.

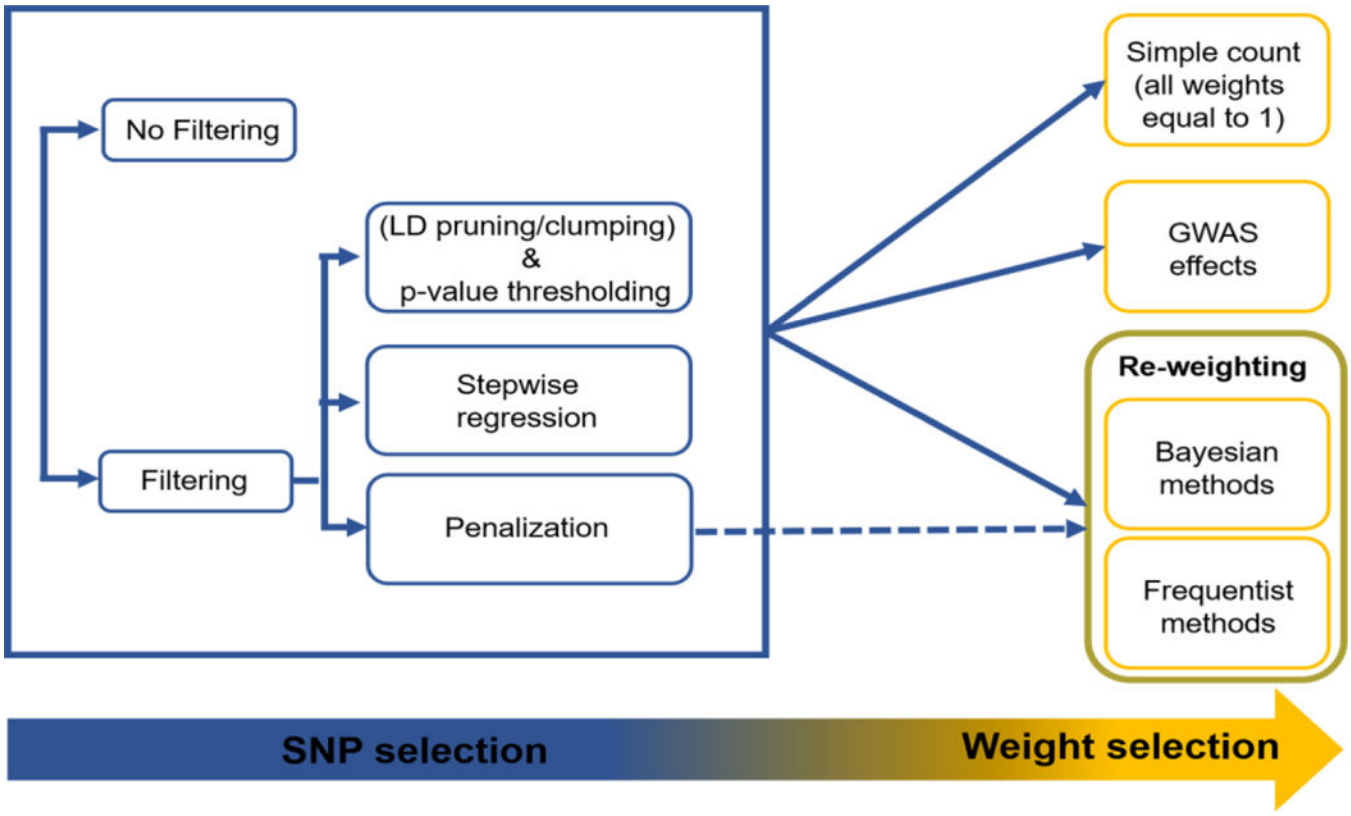

Figure 1.

Key Figure. Polygenic risk score calculation. Step1) SNP selection (with or without filtering), Step 2) weight calculation: Candidate SNPs can be assigned a weight of 1 (PRS is a simple sum of SNP alleles) or weighed using existing GWAS-derived effect sizes. Alternatively, one can re-calculate the SNP weights (re-weighting), that is, estimate new weights by including the SNPs in a regression model (e.g., Cox). Penalization techniques (either frequentist e.g., Lasso or Bayesian e.g., LDpred) can also be used for re-weighting. These methods can achieve SNP selection and weight estimation simultaneously, by setting some of the SNP weights to zero. Penalization methods can be either applied on the filtered or on the original SNP list.

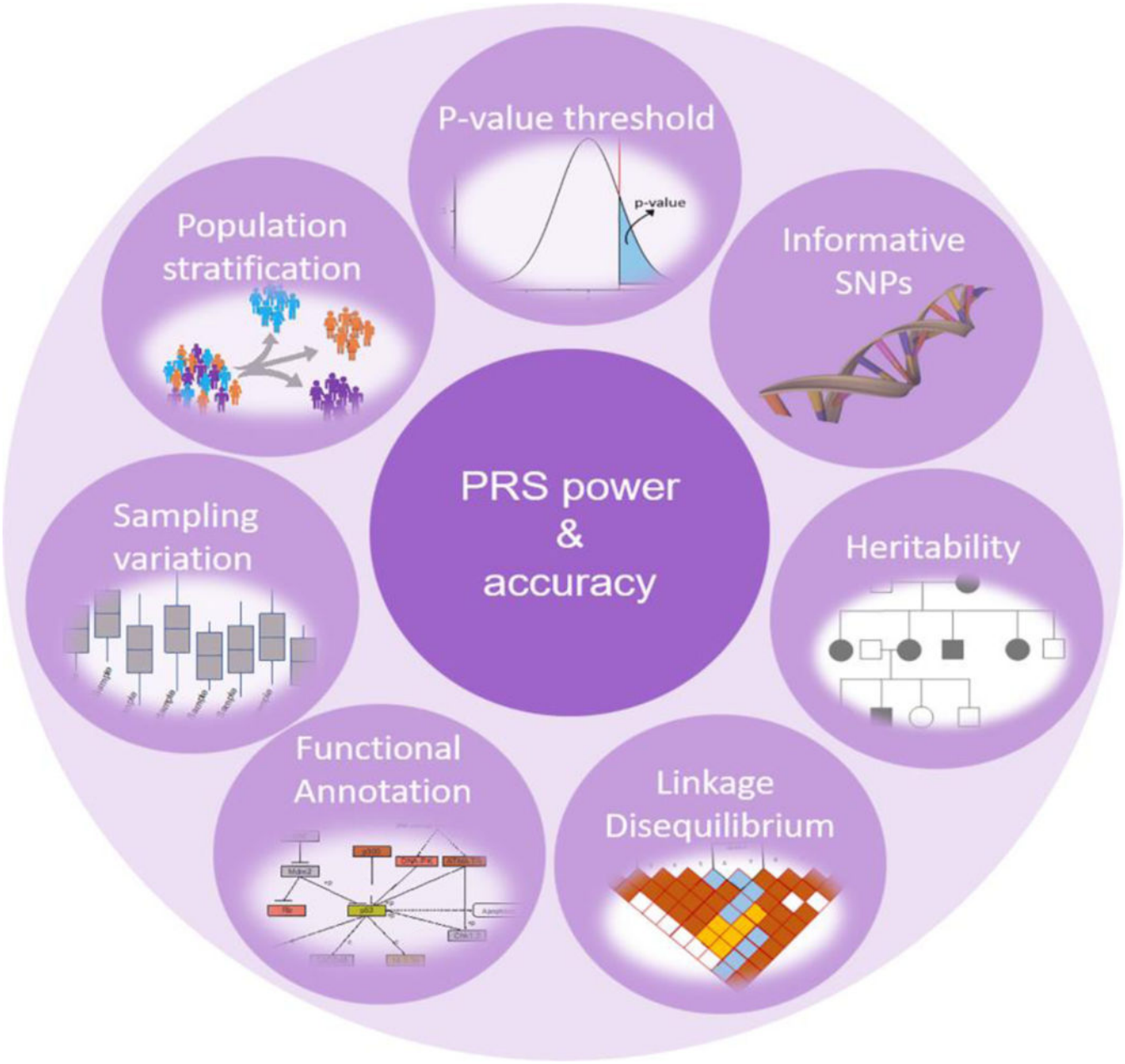

Figure 2.

Factors affecting PRS accuracy. Disease related factors (e.g., heritability, functional annotation, LD structure, and number of informative SNPs) as well as study design aspects (e.g., sample size, p-value threshold for SNP selection, and sampling variability), affect the power and performance of PRS. Depending on the hypothesis tested and the disease characteristics, improved PRS performance is possible via the appropriate sample size, SNP selection threshold and LD control.