nature protocols

REVIEW ARTICLE

https://doi.org/10.1038/s41596-020-0353-1

Tutorial: a guide to performing polygenic risk score analyses

Shing Wan Choi[1,2], Timothy Shin-Heng Mak[3] and Paul F. O'Reilly[1,2]✉

A polygenic score (PGS) or polygenic risk score (PRS) is an estimate of an individual's genetic liability to a trait or disease, calculated according to their genotype profile and relevant genome-wide association study (GWAS) data. While present PRSs typically explain only a small fraction of trait variance, their correlation with the single largest contributor to phenotypic variation—genetic liability—has led to the routine application of PRSs across biomedical research. Among a range of applications, PRSs are exploited to assess shared etiology between phenotypes, to evaluate the clinical utility of genetic data for complex disease and as part of experimental studies in which, for example, experiments are performed that compare outcomes (e.g., gene expression and cellular response to treatment) between individuals with low and high PRS values. As GWAS sample sizes increase and PRSs become more powerful, PRSs are set to play a key role in research and stratified medicine. However, despite the importance and growing application of PRSs, there are limited guidelines for performing PRS analyses, which can lead to inconsistency between studies and misinterpretation of results. Here, we provide detailed guidelines for performing and interpreting PRS analyses. We outline standard quality control steps, discuss different methods for the calculation of PRSs, provide an introductory online tutorial, highlight common misconceptions relating to PRS results, offer recommendations for best practice and discuss future challenges.

Introduction

Genome-wide association studies (GWASs) have identified a large number of genetic variants, mostly single nucleotide polymorphisms (SNPs), significantly associated with a wide range of complex traits[1–3]. However, these variants typically have a small effect and correspond to a small fraction of truly associated variants, meaning that they have limited predictive power[4–6]. Using a linear mixed model in the genome-wide complex trait analysis software[7], Yang et al. demonstrated that much of the heritability of height can be explained by evaluating the effects of all SNPs simultaneously[4]. Subsequently, statistical techniques such as linkage disequilibrium (LD) score regression[8,9] and the polygenic risk score (PRS) method[5,10] have also aggregated the effects of variants across the genome to estimate heritability, to infer genetic overlap between traits and to predict phenotypes based on genetic profile[5,6,8–10].

While genome-wide complex trait analysis, LD score regression and PRS can all be exploited to infer heritability and shared etiology among complex traits, PRS is the only approach that provides an estimate of genetic liability to a trait at the individual level. In the **classic PRS method**[5,11–14] (terms in boldface are defined in Box 1), a polygenic risk score is calculated by computing the sum of **risk alleles** that an individual has, weighted by the risk allele effect sizes as estimated by a GWAS on the phenotype. Studies have shown that substantially greater predictive power can usually be achieved by including a large number of SNPs in the PRS rather than restricting to only those reaching genome-wide significance in the GWAS[11,15,16]. As an individual-level proxy of genetic liability to a trait, PRSs are suitable for a range of applications. For example, as well as identifying shared etiology among traits, PRSs have been used to test for genome-wide gene-by-environment and gene-by-gene interactions[15,17], to perform Mendelian randomization studies to infer causal relationships and for patient stratification and sub-phenotyping[15,16,18]. Thus, while polygenic scores represent individual genetic predictions of phenotypes, prediction is often not the end objective: instead, these predictions are commonly aggregated across samples and used for research purposes, interrogating hypotheses via association testing.

Despite the popularity of PRSs, there are minimal guidelines[12] on how best to perform and interpret PRS analyses. Here, we provide a guide to performing PRS analyses, outlining the standard quality control steps required, options for PRS calculation and testing and interpretation of results. We also outline some of the challenges in PRS analyses and highlight common misconceptions in their interpretation. We will not perform a comparison of the power of different PRS methods or provide an overview of PRS applications, since these are available elsewhere[12,14,19,20]. Instead, we focus this article on the issues relevant to PRS analyses irrespective of the method used or the application, so that researchers have a starting point and reference guide for performing polygenic score analyses.

[1]MRC Social, Genetic and Developmental Psychiatry Centre, Institute of Psychiatry, Psychology and Neuroscience, King's College London, London, UK. [2]Department of Genetics and Genomic Sciences, Icahn School of Medicine, Mount Sinai, New York, NY, USA. [3]Centre of Genomic Sciences, University of Hong Kong, Hong Kong, China. ✉e-mail: paul.oreilly@mssm.edu

Box 1 | Key terms and definitions (in order of appearance)

Classic PRS method: the method—commonly known as the C+T method—for calculating PRSs applied in the key early PRS empirical studies, theoretical evaluations and software implementations[5,11,13]. The method involves computing PRSs based on a subset of partially independent (clumped) SNPs exceeding a specific GWAS association *P* value threshold.
Risk allele: the allele of a SNP that increases the risk of disease. An effect allele is simply the allele that was coded for association testing and can either increase or decrease risk.
Effect size: the increase in the trait value (usually reported as a beta) or disease risk (usually reported as an OR) associated with each additional copy of the risk allele.
Summary statistic: a value that summarizes multiple data points with a single number (e.g., a mean or effect size). GWAS data are often made available only as summary statistics.
Minor allele frequency: the frequency of the less frequent allele of a SNP (usually reported as a fraction) in the population.
Base data: the GWAS summary statistics (e.g., effect sizes or *P* values) on which the PRS calculation is based. The base trait is the phenotype of study in the GWAS.
Target data: the genotype-phenotype data, in, for example, PLINK binary format[26], of individuals in whom PRSs are calculated. The PRSs infer genetic liability of the base trait and are tested for association with the target trait.
Shrinkage: a statistical technique applied to reduce estimated effect sizes, inflated due to overfitting (see below), so that they more accurately reflect the true population effect sizes.
SNP heritability: the proportion of phenotypic variance that can be explained by SNPs, often estimated using GWAS data on common SNPs only.
Overfitting: occurs when a prediction model has been over-optimized to sample data due to inclusion of too many parameters, such that it performs relatively poorly when applied to independent data. Closely related to winner's curse, in which predictors most associated with the outcome in sample data have inflated effect size estimates.
Variance explained: typically refers to the variance of a phenotype explained by a set of predictors, or specifically a PRS, in a predictive model assuming linear effects.

Accompanying this article is an online tutorial for guiding users through the steps of a standard PRS analysis, with example data and scripts provided. Definitions of key terms used throughout this article can be found in Box 1.

Introduction to PRSs

We define PRSs, or polygenic scores, as a single value estimate of an individual's genetic liability to a phenotype, calculated as a sum of their genome-wide genotypes, weighted by corresponding genotype **effect size** estimates derived from GWAS **summary statistic** data. The genotypes are typically those of common (**minor allele frequency** > 0.01) biallelic SNPs, since most GWASs to date consist of these, but they could also include rare variants or other forms of polymorphism. The effect size estimates may be scaled or shrunk, as discussed in later sections. The use of summary statistic data for the genotype effect size estimates distinguishes polygenic scores from phenotypic prediction approaches that exploit individual-level data only. In the latter, genotype effect sizes are usually estimated in joint models of multiple variants and prediction performed simultaneously, using approaches such as best linear unbiased prediction[21,22] or least absolute shrinkage and selection operator (LASSO)[23,24]. While such methods may offer great promise in performing powerful prediction within large individual-level data sets[24], we limit our focus to polygenic scores here. Polygenic scores, as defined here by their utilization of GWAS summary statistics, are likely to have enduring application because: (i) data sharing restrictions limit full access to individual-level data; (ii) heterogeneity across cohorts reduces the motivation to pool individual-level data; (iii) the largest sources of individual-level data—population cohorts, such as the UK Biobank[25]—generally have relatively few individuals with specific diseases compared to dedicated case/control studies, for which there is typically only summary statistic data available and (iv) researchers desire to test specific hypotheses within richly phenotyped small-scale local data sets, made feasible by leveraging powerful summary statistics.

Therefore, PRS analyses can be characterized by the two key input data sets that they require: (i) **base data** (GWAS), consisting of summary statistics (e.g., betas and *P* values) of genotype-phenotype associations at genetic variants (hereafter SNPs) genome-wide, typically made available online in text format by the investigators who performed the GWAS; and (ii) **target data**, consisting of genotypes, and usually also phenotype (s), in individuals from a sample to which the researchers performing the PRS analysis have access (often not publicly available), which should be independent of the GWAS sample (discussed below). The target data are typically formatted as PLINK binary files[26]. It is in the target sample that the PRS analyses are performed, which may involve merely computing PRSs in all the target individuals, conducting association testing between the PRSs and phenotypes or outcomes of interest or predicting individuals' risk of disease or medication side effects in clinical settings. Important challenges in the calculation of PRSs are the selection of SNPs for inclusion in the score and what, if any, **shrinkage** to apply to the GWAS effect size estimates. If the parameters of the PRS calculation have not been previously optimized, then the target sample can be used both for this optimization and for the analysis, as long as careful cross-validation or permutation procedures are applied. Ideally, analysis is also performed in an independent validation sample to ensure the generalizability of results. Each of these topics is discussed further in later sections.

If genetic effects could be estimated from GWAS without error, then the PRS would explain variability in the phenotype of target sample individuals equal to the **SNP heritability** (h^2_{SNP}) of the trait[27]. However, due to error in the effect size estimates and inevitable differences in the base and target samples, the predictive power of PRSs are typically substantially

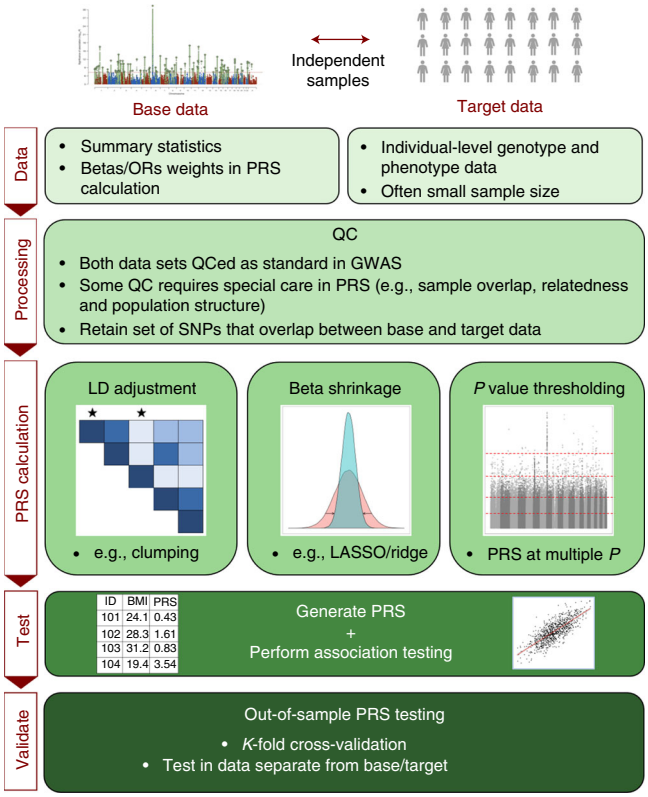

Fig. 1 | The PRS analysis process. PRS analyses can be characterized by their use of base and target data sets. QC of both data sets is described in 'QC of base and target data', while the different approaches to calculating PRSs (e.g., LD adjustment via clumping, beta shrinkage using LASSO regression or *P* value thresholding) are summarized in 'Calculation of PRSs'. Issues relating to utilizing PRSs for association analyses to test hypotheses, including interpretation of results and avoidance of overfitting to the target data, are detailed in 'Interpretation and presentation of results'.

lower than h^2_{SNP}, but will tend towards h^2_{SNP} as GWAS sample sizes increase.

Figure 1 summarizes the fundamental features of a PRS analysis and reflects the structure of this guide. In the next section, we outline recommended quality control (QC) of the base and target data sets.

QC of base and target data

The power and validity of PRS analyses are dependent on the quality of the base and target data. Therefore, both data sets must undergo QC to at least the standards implemented in GWAS studies (see refs. [28–30]), while numerous QC issues specific to PRS analyses need special attention. Below, we outline these QC measures, which should act as a 'QC checklist' for PRS analyses. These QC procedures are intentionally conservative, and particular care should be taken in performing them, because small errors can become inflated when aggregated across SNPs in PRS calculation. Researchers can practice performing these QC steps on example data in our online tutorial: https://choishingwan.github.io/PRS-Tutorial/.

QC relevant to base data only

Heritability check

A critical factor in the accuracy and predictive power of PRSs is the power of the base (GWAS) data[5], and so to avoid reaching misleading conclusions from the application of PRSs, we recommend performing PRS analyses only that use GWAS data with an $h^2_{SNP} > 0.05$. If an h^2_{SNP} estimate has not been reported for these data, then we suggest using software for estimating h^2_{SNP} from GWAS summary statistics, such as LD Score regression[8] or SumHer[31].

Effect allele

Some GWAS results files do not make clear which allele is the **effect allele** and which is the non-effect allele. If the incorrect assumption is made in computing the PRS, then the effect of the PRS in the target data will be in the wrong direction. Therefore, to prevent the generation of spurious results, the identity of the effect allele from the base GWAS data must be obtained from the GWAS investigators if not reported clearly in the GWAS results files.

QC relevant to target data only

We recommend performing PRS analyses that involve association testing on target sample sizes of ≥100 individuals (or effective sample sizes[32] >100 for case/control data) and caution against analyses that utilise base data with low h^2_{SNP} and small target sample size. This is to minimize the generation of misleading results due to the less-stringent QC feasible on small samples, potentially inaccurate adjustments (e.g., from population structure adjustments and LD calculations) and underpowered PRS-trait association tests (see 'Power and accuracy of PRSs: target sample sizes required').

QC relevant to base and target data

File transfer

Since most base GWAS data are downloaded online, and base/target data transferred internally, one should ensure that files have not been corrupted during transfer by using, for example, md5sum[33]). Corrupt files can generate PRS calculation errors.

Genome build

Ensure that the base and target data SNPs have genomic positions assigned on the same genome build[34]. LiftOver[35] is an excellent tool for standardizing genome build across different data sets.

Standard GWAS QC

Researchers should follow established guidelines (e.g., refs. [28–30]) —we recommend ref. [29]—to perform standard GWAS QC on the base and target data. Since the option of performing QC on the base GWAS data will typically be unavailable, researchers should ensure that high-quality QC was performed on the GWAS data that they utilize. We recommend the following QC criteria for standard analyses: genotyping rate >0.99, sample missingness <0.02, Hardy-Weinberg Equilibrium $P > 1 \times 10-^6$, heterozygosity within 3 standard deviations of the mean, minor allele frequency (MAF) >1% (MAF >5% if target sample

N <1000) and imputation 'info score' >0.8. If both the base and target data are large (e.g., N >50,000), then SNPs with MAF <1% may be included, in which case we recommend a minor allele count >100 in the base and target data to ensure the integrity of normality assumptions implicit in association testing and LD calculation. Future work will be required to integrate the effects of extremely rare and common variants and to establish whether their joint effects are typically additive[36]. PLINK is a useful software for performing these, and other, QC procedures[26,37].

Ambiguous SNPs

If the base and target data were generated using different genotyping chips, and the chromosome strand (+/−) that was used for either is unknown, then it is not possible to pair up the alleles of ambiguous SNPs (i.e., those with complementary alleles, either C/G or A/T SNPs) across the data sets, because it will be unknown whether the base and target data are referring to the same allele or not. While allele frequencies could be used to infer which alleles are on the same strand[38], the accuracy of this could be low for SNPs with MAF close to 50% or when the base and target data are from different populations. Therefore, we recommend removing all ambiguous SNPs to avoid introducing this potential source of systematic error.

Mismatching SNPs

SNPs that have mismatching alleles reported in the base and target data are either resolvable by strand-flipping the alleles to their complementary alleles in, for example, the target data, such as for a SNP with A/C in the base data and G/T in the target, or non-resolvable, such as for a SNP with C/G in the base and C/T in the target. Most polygenic score software programs perform strand-flipping automatically for SNPs that are resolvable and remove non-resolvable mismatching SNPs.

Duplicate SNPs

Ensure that there are no duplicated SNPs in either the base or target data (e.g., using *uniq -d* in bash or *duplicated()* in R), since this can cause polygenic score software to crash or produce errors unless the software used specifically checks for duplicated SNPs.

Sex chromosomes

It is standard in GWAS QC to remove individuals for which there is a difference between reported sex and that indicated by the sex chromosomes. While these may be due to differences in sex and gender identity, they could also reflect mislabeling of samples or misreporting and are, thus, considered potentially unreliable data. A sex check can be performed in PLINK[37], in which individuals are called females if their X chromosome homozygosity estimate (F statistic) is <0.2 and males if the estimate is >0.8. In addition to this check, if the aim of an analysis is to model autosomal genetics only, then we recommend that all X and Y chromosome SNPs are removed from the base and target data to eliminate the possibility of non-autosomal sex effects influencing results. However, incorporation of the sex chromosomes has the potential to provide etiological insights and increase the predictive power of PRSs[39] and so may be performed in practice. However, given the different options for modeling of the sex chromosomes[40], reporting of analyses that incorporate the sex chromosomes should highlight how the modeling assumptions may have influenced results.

Sample overlap

Sample overlap between the base and target data can result in substantial inflation of the association between the PRS and the trait tested in the target data[41] and so must be eliminated. The level of inflation is proportional to the fraction of the target sample that overlaps the base sample[41], and so the problem is not resolved by using a large base data set. Ideally, overlapping samples are removed from the base data, and the base GWAS is recalculated. This allows calculation of polygenic scores in all target individuals and, if the base sample is larger than the target, leads to greater power for association testing than removing the overlapping samples from the target data. A practical solution that is often applied in consortium meta-analysis settings is to generate leave-one-out meta-analysis GWAS results[42], whereby each contributing study is excluded from the meta-analysis in turn. This allows each study to be subsequently used as independent target data. Alternatively, leave-one-out meta-analysis results can be calculated analytically by rearranging the meta-analysis formula[43], but this requires availability of the contributing study-level GWAS and the meta-analysis results without subsequent adjustments, such as 'genomic control'[44]. We expect a correction in more complex scenarios of partial or unknown sample overlap, when these strategies would not be appropriate, to be an objective of future methods development; until then, in such settings, we recommend that any risk of overlap is minimized through judicious use of target samples, selecting samples that are unlikely to have also been part of the base sample (e.g., due to age or location of collection). If overlap is still a distinct possibility, then inflation in results cannot be ruled out.

Relatedness

A high degree of relatedness between individuals between the base and target data can also generate inflation of the association between the PRS and target phenotype. While population structure produces a correlation between genetics and environmental risk factors that requires a broad solution, the problem is exacerbated with inclusion of very close relatives, since they may share the same household environment as well (discussed below). Thus, if genetic data from the relevant base data samples can be accessed, then any closely related individuals (e.g., first/second degree relatives) across base and target samples should be removed to eliminate this risk. If this is not an option, then every effort should be made to select base and target data that are unlikely to contain highly related individuals. However, statistical power can be compromised in analyzing base and target samples from different populations, as discussed below, and so ideally base and target samples should be as similar as possible without risking inclusion of overlapping or highly related samples.

Calculation of PRSs

Once QC has been performed on the base and target data, and the data files are formatted appropriately, then the next step is to calculate PRSs for all individuals in the target sample. There are several options in terms of how PRSs are calculated. GWASs are performed on finite samples drawn from particular subsets of the human population, and so the SNP effect size estimates are some combination of true effect and stochastic variation—producing 'winner's curse' (see **overfitting**) among the top-ranking associations—and the estimated effects may not generalize well to different populations (discussed below). The aggregation of SNP effects across the genome is also complicated by the correlation between SNPs—LD. Thus, key factors in the development of methods for calculating PRSs are: (i) the potential adjustment of GWAS estimated effect sizes via, for example, shrinkage, (ii) the tailoring of PRSs to target populations and (iii) the task of accounting for LD. We discuss these issues below, and also those relating to the units that PRS values take, the prediction of traits different from the **base trait** and multi-trait PRS approaches. Each of these issues should be considered when calculating PRSs irrespective of subsequent application. While some of these features of PRS calculation are automated in specific PRS software, it is important to understand the issues underlying PRS calculation to aid study design and interpretation of results.

Shrinkage of GWAS effect size estimates

Given that SNP effects are estimated with uncertainty, and since not all SNPs influence the trait under study, the use of unadjusted effect size estimates of all SNPs could generate poorly estimated PRSs with high standard error. To address this, two broad shrinkage strategies have been adopted: (1) shrinkage of the effect estimates of all SNPs via standard or tailored statistical techniques, and (2) use of *P* value selection thresholds as inclusion criteria for SNPs into the score.

1 PRS methods that perform shrinkage of all SNPs[19,20,45,46] generally exploit commonly used statistical shrinkage/regularization techniques, such as LASSO or ridge regression[19], or Bayesian approaches that perform shrinkage via prior distribution specification[20,45,46]. Under different approaches or parameter settings, varying forms of shrinkage can be achieved: e.g., LASSO regression reduces small effects to zero, while ridge regression shrinks the largest effects more than LASSO but does not reduce any effects to zero. The most appropriate shrinkage to apply is dependent on the underlying mixture of null and true effect size distributions, which are probably a complex mixture of distributions that vary by trait. Since the optimal shrinkage parameters are unknown a priori, PRS prediction is typically optimized across a range of possible parameter values (see below for overfitting issues relating to this), which in the case of LDpred, for example, includes a parameter for the fraction of causal variants[45].

2 In the classic PRS calculation method[5,11,13], only those SNPs with a GWAS association *P* value below a certain threshold (e.g., $P < 1 \times 10^{-5}$) are included in the calculation of the PRS, while all other SNPs are excluded. This approach effectively shrinks all excluded SNPs to an effect size estimate of zero and performs no shrinkage on the effect size estimates of those SNPs included. Since the optimal *P* value threshold is unknown a priori, PRSs are typically calculated over a range of thresholds, association with the target trait is tested for each, and the prediction is optimized accordingly (see Overfitting in PRS-trait association testing). This process is analogous to tuning parameter optimization in the formal shrinkage methods. An alternative way to view this approach is as a parsimonious variable selection method, effectively performing forward selection ordered by GWAS *P* value, involving block-updates of variables (SNPs), with size dependent on the increment between *P* value thresholds. Thus the 'optimal threshold' selected is defined as such only within the context of this forward selection process; a PRS computed from another subset of the SNPs could be more predictive of the target trait, but the number of possible subsets of SNPs is too large to feasibly test given that GWAS are based on millions of SNPs.

Different shrinkage methods offer differences in trait predictive power (varying by trait genetic architecture), parsimony of predictive model and speed of computation, factors that the investigator must weigh in method selection.

Controlling for LD

If genetic association testing is performed using joint models of multiple SNPs[47], then independent genetic effects can be estimated despite the presence of LD. However, association tests in GWASs are typically performed one SNP at a time, which, combined with the strong correlation structure across the genome, makes estimating the independent genetic effects (or best proxies of these if not genotyped/imputed) extremely challenging. If independent effects were estimated in the GWAS or by subsequent fine-mapping, then PRS calculation can be a simple summation of those effects. If, instead, the investigator is using a GWAS based on one-SNP-at-a-time testing, then there are two main options for approximating the PRS that would be obtained from independent effect estimates: (i) SNPs are clumped (i.e., thinned, prioritizing SNPs at the locus with the smallest GWAS *P* value) so that the retained SNPs are largely independent of each other, and, thus, their effects can be summed, assuming additivity; and (ii) all SNPs are included, accounting for the LD between them. In the classic PRS calculation method[5,11,13], option (i) is combined with *P* value thresholding and called the C+T (clumping + thresholding) method, while option (ii) is generally favored in methods that implement traditional shrinkage techniques[19,20,45,46]. The relatively similar performance of the classic approach to more sophisticated methods[14,19,20] may be due to the clumping process capturing conditionally independent effects well; note that clumping does not merely thin SNPs by LD at random (like pruning) but preferentially selects SNPs most associated with the trait under study, and retains multiple SNPs in the same genomic region if there are multiple independent effects there: clumping does not simply retain only the most-associated SNP in a region. A criticism of clumping, however, is that researchers typically select an arbitrarily chosen correlation

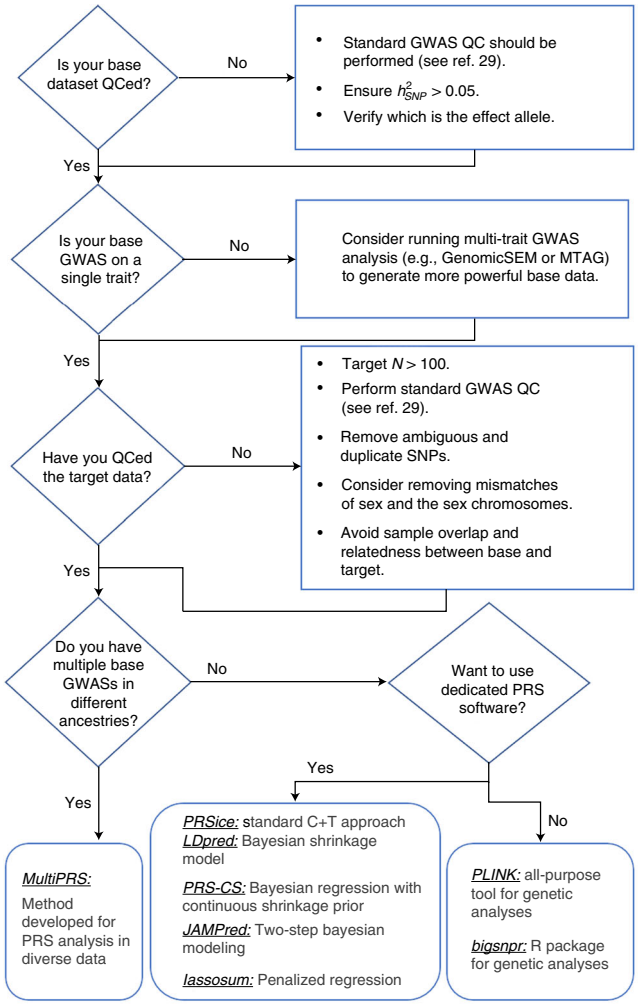

Fig. 2 | Shown is a flow chart of suggested analytical steps that can be followed to perform QC and select software for PRS analyses. GenomicSEM[65] and MTAG[70] are software tools that allow for joint analysis of summary statistics from GWASs of different complex traits and can help to boost power. Common PRS software programs include (but are not limited to) PRSice[13,14], LDpred[45], PRS-CS[20], JAMPred[46] and lassosum[19]. PLINK[26,37] and bigsnpr[49] can be used for the implementation of custom pipelines, and MultiPRS[50] is a method to perform PRS analyses on admixed populations.

threshold[41] for the removal of SNPs in LD, and so while no strategy is without arbitrary features, this may be an area for future development of this approach. The key benefits of the classic PRS method are that it is relatively fast to apply and is more interpretable than present alternatives.

Both clumping and LD modeling require estimation of the LD between SNPs. Assuming that LD values derived from the base data are unavailable, then those from a reference sample of the same ancestry, such as from the 1000 Genomes Project data[48], should be used to approximate these. If there are no reference samples well matched to the population composition of the base data, then the target data can be used to estimate the LD instead. However, if base and target samples are drawn from different populations, then the base data LD may be poorly approximated and PRS accuracy reduced accordingly.

Figure 2 illustrates a PRS analysis pipeline, highlighting QC steps and some of the main software programs presently available to users as options, which may be selected according to scientific question, data, estimated accuracy and speed of PRS computation method[14,19,20,45,46,49,50], and user preference. In our tutorial that accompanies this article (https://choishingwan.github.io/PRS-Tutorial/), readers can perform PRS analyses on example data using several of these programs to become familiar with the process. The tutorial uses summary statistic data from the GIANT consortium[1] and simulated target data, and involves applying PLINK[26,37], PRSice-2[14], LDpred[45] and lassosum[19] to calculate PRSs and illustrate results from standard PRS analyses.

PRS units

When calculating PRSs, the units of the GWAS effect sizes determine the units of the PRS; for example, if calculating a height PRS using effect sizes from a height GWAS that are reported in centimeters, then the resulting PRS will also be in centimeters. The PRS may then be standardized, dividing by the number of SNPs to ensure a similar scale irrespective of the number of SNPs included, or standardized to a standard normal distribution. However, the latter discards information that you may wish to retain, since the absolute values of the PRS may be useful for identifying outliers, detecting problems with the sample or PRS calculation (see 'PRS distribution'), comparing PRSs across different samples or even detecting the effects of natural selection.

If the phenotype values were log-transformed, standardized or inverse normalized before the GWAS, then the reported effect sizes will reflect this. Log-transformed effect sizes can be back-transformed, via exponentiating, to obtain effect sizes in the measured units. The logarithm base used in log-transforming the phenotype data must be known so that the correct exponentiation can be performed. Typically, the data required to back-transform normalized data (in *Z*-score units) are unavailable, so in this case the PRS should be calculated based on the *Z*-score effect size estimates, and the resulting scores will be in *Z*-score (i.e., standard deviation) units. When PRSs are calculated using effect sizes in units of the trait, then an implicit assumption is that the absolute effect of risk alleles is equal in the base and target populations, while when computed in *Z*-score units, the assumption is that the effect sizes are equal in terms of their impact as a fraction of trait variance.

In calculating PRSs on a binary (e.g., case/control) phenotype, the effect sizes used as weights are typically reported as log Odds Ratios (log(ORs)). Assuming that relative risks on a disease accumulate on a multiplicative rather than an additive scale[51], then PRSs should be computed as a summation of log (OR)-weighted genotypes. PRS values are computed in relation to a hypothetical individual with the non-effect allele at every SNP, and, thus, they provide only a relative (compared to other individuals) estimate of risk (or trait effect) rather than an absolute estimate.

Population genetic structure and the generalizability of PRSs

A major concern in GWAS and PRS studies is that their results may be affected by confounding due to population genetic structure. Briefly, the non-random mating of individuals in a population, caused chiefly by the tendency for individuals to find a partner born in a nearby geographic location, generates structure in genetic variation across a population. Since environmental risk factors also tend to be geographically structured, this creates the potential for associations between many genetic variants and the tested trait that are confounded by, for example, location[52,53]. Uncorrected, this can lead to false positive genotype-phenotype associations and consequently inflated estimates of PRS prediction. PRS prediction can also be inflated by a household effect, whereby the genetics of an individual are correlated with their household environment when created by parents (or siblings) with shared genetic tendencies (e.g., of diet, books or exercise)[54,55]. A key difference between these sources of PRS inflation is that the genetic variants leading to inflation due to population genetic structure are typically non-causal of the outcome, being incidentally associated with location and environmental risk factors, whereas those creating the household effect are (indirectly) causal. Stringent adjustment of effects via genetic principal components (PCs)[52] or the use of mixed models[56] should be applied to both the base and target samples to minimize inflation due to population structure, but the possibility of complex structure causing residual confounding cannot be ruled out. However, family data provide a convenient way of testing for the combined impact of population structure and the household effect on PRS prediction. If a unit increase in PRS between 'unrelated' individuals has a larger impact on a trait than a unit increase in PRS between siblings, then population structure and/or the household effect may be inflating PRS prediction in general population samples[55,57,58]. Greater adoption of family designs at the GWAS and PRS stages could be important in the future for disentangling the effects of direct genetics, indirect genetics and population structure on a trait[59].

In contrast, PRS prediction performed in a target sample from a different worldwide population from that of the base sample typically shows significant deflation[58,60–62], due to differences in, for example, genotype effect sizes, allele frequencies and LD. The characteristics of the base sample, such as their age, sex or socioeconomic distribution, influence base trait heritability and, thus, can also affect PRS prediction[58]. Given the potential implications for disparity in healthcare caused by applying PRSs that perform well only in specific subsets of the human population, we expect the issue of the generalizability of PRSs to be an active area of methods development in the coming years[50,59,62].

Figure 3 illustrates some of the major sources of bias in PRS-trait associations, highlighting the potential inflation caused by local correlation between genetics and the environment, and the likely deflation caused by a lack of correlation between the genetics and/or environment of base and target data.

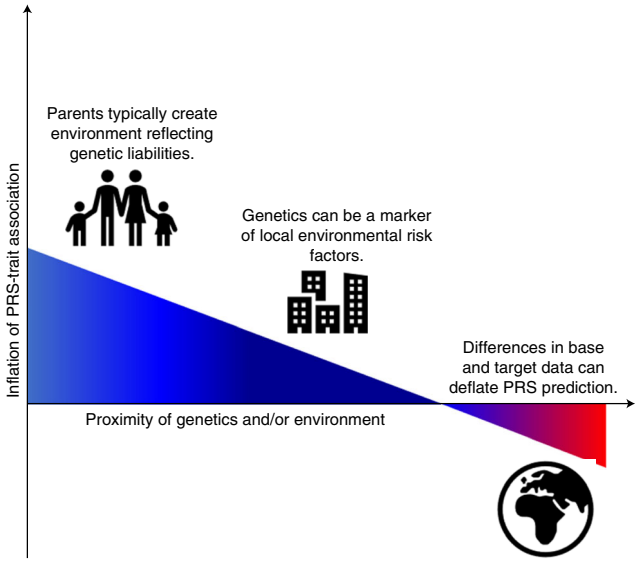

Fig. 3 | Illustration of major sources of inflation/deflation of PRS-trait associations. If the target data differ markedly from the base data in terms of allele frequencies, LD, the environment, selection pressures, etc., then the PRS-trait association will probably be deflated relative to a target sample that is well matched to the base data (note that relative inflation is theoretically possible if the trait has greater heritability in the target sample than the base sample[58]). Correlation between the population structure of genetics and the environment can inflate PRS-trait associations unless they are controlled for fully. This inflation can be exacerbated by a household effect in which parents produce an environment reflecting their genetic tendencies[55], known as passive gene*environment correlation[105]. This figure illustrates in simple form some of the broad major influences on PRS-trait associations and their typical effects; it is not intended to capture the many nuances and exceptions involved or other important effects such as evocative or active genetic-environment correlations or assortative mating[59,105].

Predicting different traits and exploiting multiple PRSs

While PRSs are often analyzed in scenarios in which the base and target phenotype are the same, many published studies involve a target phenotype different from that on which the PRS is based. These analyses fall into three main categories: (i) target trait prediction using a different but similar (or 'proxy') base trait: if there is no large GWAS on the target trait, or it is underpowered compared to a similar trait, then prediction may be improved using a different base trait (e.g., education years to predict cognitive performance[63,64]); (ii) target trait prediction exploiting multiple PRSs based on a range of different traits in a joint model[65–67]; and (iii) testing for shared etiology between base and target trait[68,69]. Applications (i) and (ii) are straightforward in their etiology-agnostic aim of optimizing prediction, achieved by exploiting the fact that a PRS based on one trait is predictive of genetically correlated traits, and that a PRS computed from any base trait is sub-optimal due to the finite size of any GWAS. A common concern in using multiple PRSs as predictors is that the PRSs are computed from the same SNPs and are, thus, inherently correlated. However, this is true of any epidemiological prediction model, since predictors typically comprise multiple shared risk factors. Therefore, when a large number of PRSs (>10) are included as predictors in a joint model, then the risk of overfitting and multicollinearity should be minimized as standard in prediction modeling, such as by applying shrinkage techniques (as in ref. [67]) or using a random effects term to model their correlation (as in ref. [66]).

Alternatively, multi-trait GWAS methods can be used to model the joint effects of genetic variants on multiple phenotypes at the GWAS stage[65,70], before computing PRSs.

Application (iii) is inherently more complex than (i) and (ii) because there are different ways of defining and assessing 'shared etiology'[71]. Shared etiology may be due to so-called horizontal pleiotropy (separate direct effects) or vertical pleiotropy (downstream effect)[71], and there are several quantities that can be estimated—genetic correlation[9], genetic contribution to phenotypic covariance (co-heritability)[72,73] or a trait-specific measure (e.g., where the denominator relates to the genetic variance of only one of the traits).

While there is active method development in these areas[65–67] at present, the majority of PRS studies use the same approach to PRS analysis whether or not the base and target phenotypes differ. However, this is rather unsatisfactory because of the non-uniform genetic sharing between different traits. In PRS analysis, the effect sizes and *P* values are estimated using the base phenotype, independent of the target phenotype. Thus, a SNP with high effect size and significance in the base GWAS may have no effect on the target phenotype. The standard approach could be adapted so that SNPs are prioritized for inclusion in the PRS according to joint effects on the base and target traits, while modifications of other PRS approaches will likely be developed in the future, each tailored to specific scientific questions.

Interpretation and presentation of results

Once PRSs have been calculated, selecting from the options described above, typically a regression is then performed in the target sample, with the PRS as a predictor of the target trait or experimental outcome, and covariates included as appropriate. In this section, we consider how results from PRS analyses are measured and plotted, how to avoid overfitting, the interpretation of results in terms of genetic associations and the potential clinical utility of PRSs and the predictive accuracy and power of PRS analyses.

Association and goodness-of-fit metrics

A typical PRS study involves testing evidence for an association between a PRS and a trait(s) in the target data. The association between PRS and outcome can be measured with standard association or goodness-of-fit metrics, such as the *P* value derived in testing a null hypothesis of no association, phenotypic **variance explained** (R^2) oreffect size estimate (beta or OR) per unit of PRS or between specific strata (e.g., high- versus low-risk individuals), and with measures of discrimination in disease prediction, such as area under the receiver operator curve (AUC) or area under the precision recall curve. The association between the PRS and the target trait is usually tested in a linear (continuous trait) or logistic (binary trait) regression, adjusting for covariates (e.g., genetic PCs, sex and age). When covariates are included in the model, then measures such as the incremental R^2 (increase in R^2 with the addition of the PRS to the model), which isolate the explanatory power of the PRS, should be reported. The incremental R^2 is necessarily greater than zero when testing is performed within a single sample, and so either an adjusted R^2 (accounting for additional parameters) or an out-of-sample R^2 should be reported (see Overfitting in PRS-trait association testing). The inclusion of covariates that are predictors of the outcome should increase statistical power and lead to more accurate estimates of PRS effects in linear regression settings, but in ascertained samples can reduce power in logistic regression settings[74]. Therefore, we recommend reporting results with and without important covariates when testing binary outcomes; confounders, such as genetic PCs, should be included as usual.

While variance explained (R^2) is a well-defined concept for continuous trait outcomes, only conceptual proxies of this measure ('pseudo-R^2') are available for case/control outcomes. A range of pseudo-R^2 metrics is used in epidemiology[75,76], with Nagelkerke R^2 perhaps being the most popular. However, Nagelkerke R^2 and similar metrics produce biased estimates of the phenotypic variance on the liability scale when the case/control ratio is not equal to the disease prevalence[75]. Intuitively, the R^2 on the liability scale here estimates the proportion of variance explained by the PRS of a hypothetical normally distributed latent variable that underlies and causes case/control status[75,77]. Heritability is typically estimated on the liability scale for case/control phenotypes[12,75,77]. Lee et al.[75] developed a pseudo-R^2 metric that accounts for case/control ratio and is measured on the liability scale. Under simulation, we demonstrate that this metric indeed controls for case/control ratios that do not reflect disease prevalence, while Nagelkerke R^2 can be highly biased (Fig. 4). Thus, we recommend use of the Lee R^2 when the disease prevalence can be well approximated, and, if not, the Lee R^2 should be estimated for a range of realistic prevalences to provide a credible interval of R^2 values. Note that if the cases in a study are milder or more severe than typical cases, then the estimated pseudo-R^2 (including the Lee R^2) will be deflated or inflated, respectively.

Graphical representations of results: bar and quantile plots

When the classic C+T method is used, the results of PRS association tests are sometimes displayed as a bar plot, where each bar corresponds to the result from testing a PRS computed from SNPs with a GWAS *P* value exceeding a specific threshold. Typically, a small number of bars are shown, reflecting results at round-figure *P* value thresholds ($5 \times 10^{-}8$, $1 \times 10^{-}5$, $1 \times 10^{-}3$, 0.01, 0.05, 0.1, 0.2, 0.3, etc.). If 'high-resolution' scoring[13] is performed, then a bar representing the most-predictive PRS may be included. Usually, the y-axis corresponds to the phenotypic variance explained by the PRS (R^2 or pseudo-R^2), and the value over each bar (or its color) provides the *P* value of association between the PRS and target trait. See examples of such bar plots in refs. [78–81]. It is important to note that the *P* value threshold of the most predictive PRS is a function of the effect size distribution, the power of the base (GWAS) and target data, the genetic architecture of the trait and the fraction of causal variants, and so should not be interpreted merely as reflecting the fraction of causal variants. For instance, if the GWAS data are relatively underpowered, then the optimal threshold is more likely to be $P = 1$ (all SNPs) even if a small fraction of SNPs are causal (see ref. [5] for details).

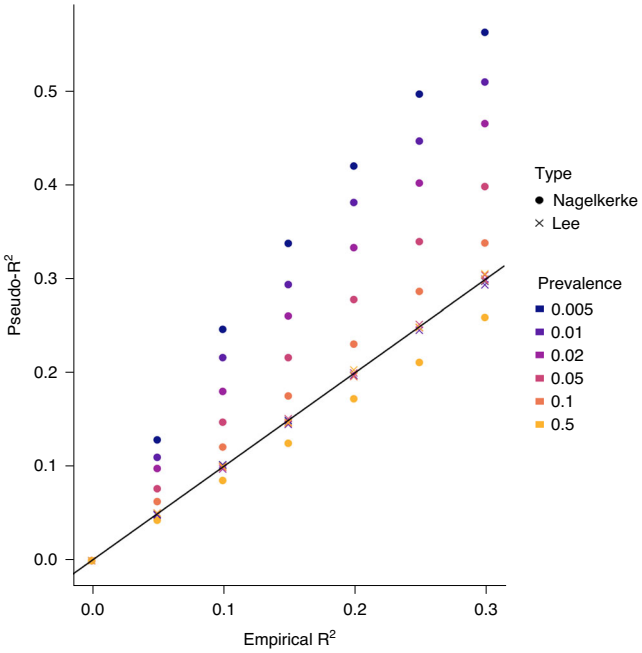

Fig. 4 | Results from a simulation study comparing Nagelkerke pseudo-R^2 with the pseudo-R^2 proposed by Lee et al.[75] that incorporates adjustment for the sample case/control ratio. In the simulation, 2,000,000 samples were simulated (using linear models and *rnorm()* in R) to have a normally distributed phenotype, generated by a normally distributed predictor (e.g., a PRS) explaining a varying fraction of phenotypic variance, with a residual error term to model all other effects. Case/control status was then simulated under the liability threshold model according to a specified prevalence. Cases (5,000) and controls (5,000) were then randomly selected from the population, and the R^2 of the original continuous data (empirical R^2), estimated by linear regression, was compared to both the Nagelkerke R^2 (•) and the Lee R^2 (×) based on the corresponding case/control data by logistic regression.

While metrics such as the AUC and R^2 can provide sample-wide summaries of the predictive power of a PRS, it can be useful to inspect how trait values vary with increasing PRS or to gauge the elevated disease risk among individuals with the highest PRSs. This can be visualized using a quantile plot (Fig. 5a). Quantile or strata plots in PRS studies are usually constructed as described in refs. [18,82–84]. The target sample is first separated into strata of increasing PRS: for instance, 20 equally sized quantiles, each comprising 5% of the PRS sample distribution (Fig. 5a), or unequal strata, usually used to highlight individuals with extreme PRSs (Fig. 5b). The phenotype values of each stratum are then either plotted directly as means or prevalences (as in Fig. 5a,b) or compared to those of a reference stratum (usually the median stratum or the remaining strata combined) one by one, with strata status as a predictor of target phenotype (reference stratum coded 0, test stratum coded 1) in a regression. Performing a regression allows adjustment for covariates and will mean that the y-axis takes values of beta (continuous trait) or OR (binary trait).

Quantile plots corresponding to the effect of a PRS on a normally distributed target trait should reflect the S-shape of the probit function (Fig. 5a). This is because the trait values are more spread out between quantiles at the tails of a normal distribution. Thus, plotting quantiles of PRS versus (absolute) effect on trait shows increasingly larger jumps up/down the y-axis from the median to the extreme upper/lower quantiles. When unequal strata are plotted, with the smallest strata at the tails, then this effect appears stronger. When the target outcome is disease status and prevalence or OR are plotted on the y-axis, then the shape is expected to be different: here, the shape is asymmetrical, showing a marked inflection at the upper end (Fig. 5b), since cases are enriched at the upper end only. Thus, inflections of risk at the tails of the PRS distribution[82,83] should

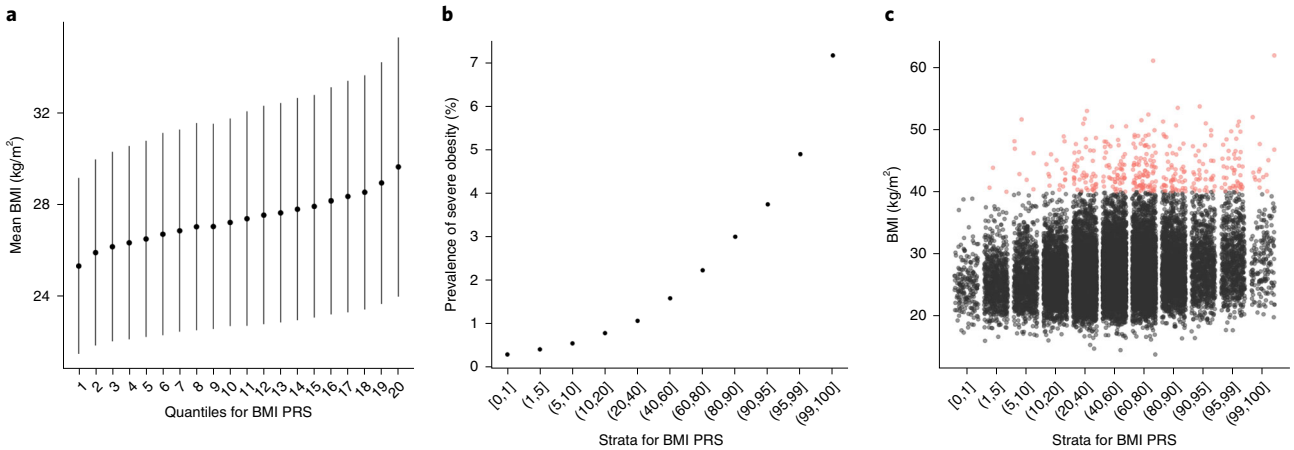

Fig. 5 | Three different ways of representing the same data. The data correspond to body mass index (BMI; in kg/m^2) PRSs calculated in 386,266 individuals in the UK Biobank data, derived using the GIANT BMI GWAS as base data. **a**, Quantile plot with 20 quantiles of increasing BMI PRS versus mean BMI (y-axis). **b**, Strata plot with unequal strata of increasing BMI PRS versus prevalence (%) of severe obesity (BMI > 40). **c**, Strata plot with the same strata as in **b**, but here each individual's BMI value is shown on the y-axis. The sample is randomly thinned to 5% of the total size, and lateral spread within each stratum is applied, to make individual points visible, while red points correspond to individuals with severe obesity. Qualitatively similar patterns as these should be expected for PRSs corresponding to all reasonably heritable continuous or binary traits, with strength of patterns dependent on the predictive power of the PRS (here, the PRS explains ~5% of BMI in these data). BMI here could be considered analogous to the liability underlying a disease in the liability threshold model, and in this way plot **c** may be helpful in imagining the uncertainty in the true liability that underlies a given PRS value for a disease.

be interpreted according to these statistical expectations and not as interesting in themselves.

Interpretation for clinical utility

There is intense interest in the potential clinical utility of PRSs—to improve diagnoses, to select optimal treatment and in particular as part of preventative medicine[85–89]. Preventative medicine typically either seeks to shift entire trait distributions (e.g., to reduce population-wide BMI or salt intake) or to target high-risk individuals (e.g., screening according to age or multiple factors). The efficacy of each strategy in reducing disease burden is dependent on numerous statistical, behavioral and economic factors, discussed elsewhere[90,91]. If targeting high-risk individuals is evaluated as worthwhile for a given disease, then whether PRSs can aid the stratified medicine approach taken should be considered. The PRS has some attractive features as a clinical predictor, including being reasonably inexpensive, non-intrusive, available from birth and requiring only a single measure during a life-time (although effects can vary by age[88]). Also, while PRS must be partially correlated with traditional risk factors given the heritability of almost all risk factors, they probably also offer orthogonal information that cannot be easily measured. One noteworthy example is that of family history as a risk factor: the family history of disease, often a key predictor in disease prediction models, will typically be exactly the same for full siblings despite the substantial variance in genetic liability conferred to them from their parents. Thus, individual-level PRSs have the potential to offer markedly higher predictive power than family history alone. However, present PRSs often have low predictive power, and so claims of their direct clinical utility have drawn scepticism[92–94] and generated much debate.

We use Fig. 5, and BMI as an example, to highlight some of the pertinent issues of the debate. The base data here are the BMI summary statistics generated by the GIANT Consortium[1], while the target data are from the UK Biobank:[25] in these data, the PRS for BMI explains approximately 5% of the variation in BMI in the target data, which is typical predictive accuracy for a BMI PRS using a recent BMI GWAS and for PRSs of most phenotypes with a well-powered GWAS (e.g., with >20 genome-wide significant loci). Figure 5b shows the prevalence of severe obesity across strata of BMI PRSs, and in contrast to the moderate increase in mean BMI across quantiles (Fig. 5a), shows a steep increase in obesity prevalence rates in the upper tail. The comparatively high risk in the most extreme strata, for obesity and other major diseases, has been used as an argument for the clinical utility of PRSs[82,83]. However, Fig. 5c highlights potential limitations. While the upper strata do have an elevated prevalence rate, the uncertainty in the individual predictions is extremely large, such that individuals should avoid interpretation of their PRS value (unless the PRS explains considerably more phenotypic variance than 5%). Furthermore, most obese individuals have a normal or low PRS, highlighting a drawback of focusing on 'high risk' individuals when the prediction model explains a small fraction of phenotypic variation[90]. Finally, the prevalence rates in the top 1% stratum (7.2%) are markedly higher than in the 95%–99% stratum (4.9%), and substantially higher than in the 40%–60% stratum (1.6%), but there are more individuals with severe obesity in the latter strata, and many are close to the obesity threshold. Therefore, focusing on prevalence rates (or risk/ORs) could be misleading in terms of impact on public health, especially if the clinical effects of genetic liability are continuous. It can also be misleading to report, for example, ORs that compare the highest and lowest strata, since these are inflated relative to typical ORs, which compare exposed and unexposed groups.

There is a critical need for rigorous cost-benefit analyses to evaluate how estimated increases in predictive power offered by PRSs are likely to translate into improvements in public health compared to alternatives, such as instead optimizing prediction models based on endophenotypes (e.g., BMI or cholesterol) measured at informative ages or implementing population-wide interventions (e.g., food regulations). Until objective comparisons have been performed, the debate on the topic is likely to remain largely semantic, while huge investments in research and healthcare funds, justified by the promise of the clinical utility of PRSs, could be misguided unless robust evidence is followed.

Interpretation of PRS-trait associations

PRSs for many traits are presently such weak proxies of true genetic liability that the phenotypic variance that they explain is often very small ($R^2 < 0.01$). Association test results of PRS with very small estimated effects should be treated with caution given the possibility that they may have been generated by subtle uncorrected confounding. However, if the results are shown to be robust to confounding (see Population genetic structure and the generalizability of PRSs), then the effect size is not important if the aim is only to establish whether an association exists, which may provide etiological insight.

Pleiotropy is ubiquitous in the genome[71,95], with potentially some shared genetic etiology between the vast majority of phenotypes. This is probably due to the complex, highly interrelated biological and environmental network among human traits. For instance, a genetic predisposition to higher cognitive performance must, on average, lead to greater educational performance and higher socioeconomic position[96]; socioeconomic position is associated with most complex diseases, and thus a component of the genetic etiology of most diseases will be the genetics of cognition. This genetic component is probably extremely small for most diseases, but with sufficient sample size will generate significant (typically negative) genetic correlations between cognition and many diseases (vertical pleiotropy), as well as between diseases (horizontal pleiotropy). Similar examples could be provided for genetic liabilities to addiction, risk-taking, confidence, depression, metabolism, immunity, etc. and the multitude of traits and diseases on which they have downstream effects. While this intimate link between genetics and the complex interrelated network among risk factors and diseases helps to explain both the high levels of pleiotropy and polygenicity observed in genomic data, it also calls for caution in interpretation of genetic overlap between phenotypes: an unconsidered or unknown, shared, small sub-component of genetic risk may

have driven an observed genetic association between two phenotypes, potentially rendering the link between the two phenotypes unimportant. However, despite this complexity, important mechanistic insight can be provided by testing whether the shared etiology between a pair of traits is due to horizontal or vertical pleiotropy[71], which is the focus of Mendelian Randomization methods[97,98]. To this end, PRSs may be useful in establishing the relative strength of genetic associations among a range of traits[43,99] and, in so doing, act as a step toward identifying the causal mechanism[100].

PRS distribution

The central limit theorem dictates that if a PRS is based on a sum of independent variables (here, SNPs) with identical distributions, then the PRS of a sample should approximate the normal (Gaussian) distribution. This is true even if the PRS has extremely low predictive accuracy, since the sum of random numbers is approximately normally distributed, and so a normally distributed PRS in a sample should not be considered as validation of the accuracy of a PRS or of the liability threshold model. However, strong violations of these assumptions, such as the use of many correlated SNPs or a sample of heterogenous ancestry (thus, SNPs with markedly different genotype distributions), can lead to non-normal PRS distributions. Thus, inspection of PRS distributions may highlight calculation errors or problems of population stratification in the target sample for which researchers did not adequately control.

Overfitting in PRS-trait association testing

A common concern in PRS studies that adopt the classic (C+T) approach is whether the use of the most predictive PRS—based on testing at many P value thresholds—overfits to the target data and thus produces inflated results and false conclusions. While such caution is to be encouraged in general, potential overfitting is a normal part of prediction modeling, relevant to the other PRS approaches (Fig. 2), and there are well-established strategies for increasing predictive power while avoiding overfitting[101]. One strategy that we do not recommend is to perform no optimization of parameters—e.g., selecting a single arbitrary P value threshold (such as $P < 1 \times 10^{-8}$ or $P = 1$)—because this may lead to serious underperformance of the PRS prediction, which itself can lead to false conclusions.

The gold-standard strategy for guarding against generating overfit prediction models and results is to perform out-of-sample prediction. First, parameters are optimized using a training sample, and then the optimized model is tested in a test or validation data set to assess performance. In the PRS setting involving base and target data sets, it would be incorrect to believe that out-of-sample prediction has already been performed, because polygenic scoring involves two different data sets; in fact, the training is performed on the target data set, meaning that a third data set is required for out-of-sample prediction. The leave-one-out strategy often adopted in meta-analysis consortia[42] is also at risk of overfitting if parameter optimization and testing are both performed in the data set left out. In the absence of an independent data set, the target sample can be subdivided into training and validation data sets, and this process can be repeated with different partitions of the sample (e.g., performing 10-fold cross-validation[67,102,103]) to obtain more robust model estimates. However, a true out-of-sample, and thus not overfit, assessment of performance can be achieved only via final testing on a sample entirely separate from data used in training.

Without validation data or when the size of the target data makes cross-validation underpowered, an alternative is to generate empirical P values corresponding to the optimized PRS prediction of the target trait, via permutation[14]. While the PRS itself may be overfit, if the objective of the PRS study is association testing of a hypothesis—e.g., H_0: schizophrenia and rheumatoid arthritis have shared genetic etiology—rather than for prediction per se, then generating empirical P values offers a powerful way to achieve this while maintaining appropriate type 1 error[14]. It is also even possible to generate optimized parameters for a PRS when no target data are available[19].

Power and accuracy of PRSs: target sample sizes required

In one of the key PRS papers published to date, Dudbridge 2013[5] investigated the expected power and predictive accuracy of PRSs according to derived formulae based on standard quantitative genetics models[104]. Dudbridge demonstrated that highly significant results observed in PRS association studies were consistent with expectations given the base and target sample sizes used, thus not necessarily due to confounding or bias, and calculated that several published studies with null results were probably underpowered. Dudbridge also showed that the power of PRS association testing is optimized using equal-sized base and target sample sizes, while individual-level predictive accuracy is optimized by maximizing base sample size.

To complement these theoretical expectations, we performed PRS analyses, using the UK Biobank, that may be useful for estimating what target sample sizes are required for PRS-trait association testing. We tested traits with high (height), medium (forced volume capacity; FVC) and low (hand grip strength) SNP heritability. Sampling randomly from the UK Biobank, we generated a base GWAS of size 100,000 individuals, a target sample size of 100,000 for parameter optimization and a range of validation sample sizes from 10 to 2,000. We performed PRS association tests using the classic C+T method, predicting the same trait as used in the base GWAS, and repeating the sampling 200 times to estimate the variability in the results. Figure 6a displays the trait variance explained in the validation data across the range of sample sizes in the three target traits. Figure 6b displays the statistical power from association testing of the PRS and each of the corresponding target traits, showing, for example, that a target sample size of ~200 is required to exceed 80% power for FVC (h^2_{SNP}= 0.23) and ~500 for hand grip strength (h^2_{SNP} = 0.11). While these results correspond to performance in validation data, the statistical power should be reflective of the power in relation to empirical P values estimated in target data (see Overfitting in PRS-trait association testing). We tested continuous traits here, but we would expect these results to reflect those of case/control outcomes with

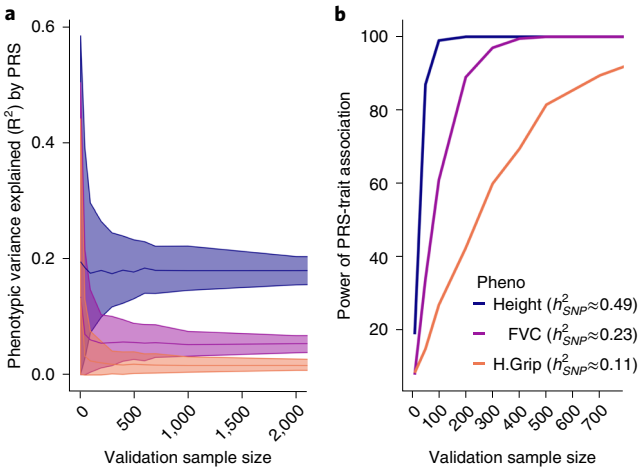

Fig. 6 | Examples of the performance of PRS analyses on real data by validation sample size, according to (a) phenotypic variance explained (R2) and (b) association *P* value. UK Biobank data on height (h^2_{SNP} = 0.49[8]), FVC (h^2_{SNP} = 0.23[8]) and hand grip (h^2_{SNP} = 0.11[8]) were randomly split into two sets of 100,000 individuals and used as base and target data, while the remaining sample was used as validation data of varying sample sizes, from 10 individuals to 2,000 individuals. Each analysis was repeated 200 times with independently selected validation samples. The mean and 95% range of R^2 values across the 200 simulations are depicted in **a**, and statistical power in **b** corresponds to the proportion of simulations that produced a PRS-trait association *P* value < 0.05 in the validation data.

similar heritabilities estimated on the liability scale and equivalent effective sample sizes (see ref. [32]). While these results only approximate the performance of PRS analyses across traits of varying heritability—assuming ancestrally matched base and target samples and without accounting for factors such as trait polygenicity—they may be useful in providing a broad indication of whether researchers' data are sufficiently powered for future analyses or if they should acquire more data.

Conclusions

As GWAS sample sizes increase, polygenic scores are likely to play a central role in the future of biomedical research and personalized medicine. However, the efficacy of their use will depend on the continued development of methods that exploit them, their proper analysis and appropriate interpretation and an understanding of their strengths and limitations.

References

1. Locke, A. E. et al. Genetic studies of body mass index yield new insights for obesity biology. *Nature* **518**, 197–206 (2015).
2. Kunkle, B. W. et al. Genetic meta-analysis of diagnosed Alzheimer's disease identifies new risk loci and implicates Aβ, tau, immunity and lipid processing. *Nat. Genet.* **51**, 414 (2019).
3. Schizophrenia Working Group of the Psychiatric Genomics Consortium. Biological insights from 108 schizophrenia-associated genetic loci. *Nature* **511**, 421–427 (2014).
4. Yang, J. et al. Common SNPs explain a large proportion of heritability for human height. *Nat. Genet.* **42**, 565–569 (2010).
5. Dudbridge, F. Power and predictive accuracy of polygenic risk scores. *PLoS Genet.* **9**, e1003348 (2013).
6. Dudbridge, F. Polygenic epidemiology. *Genet. Epidemiol.* **40**, 268–272 (2016).
7. Yang, J. et al. GCTA: A tool for genome-wide complex trait analysis. *Am. J. Hum. Genet.* **88**, 76–82 (2011).
8. Bulik-Sullivan, B. K. et al. LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nat. Genet.* **47**, 291–295 (2015).
9. Bulik-Sullivan, B. et al. An atlas of genetic correlations across human diseases and traits. *Nat. Genet.* **47**, 1236–1241 (2015).
10. Palla, L. & Dudbridge, F. A fast method that uses polygenic scores to estimate the variance explained by genome-wide marker panels and the proportion of variants affecting a trait. *Am. J. Hum. Genet.* **97**, 250–259 (2015).
11. Purcell, S. M. et al. Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. *Nature* **460**, 748–752 (2009).
12. Wray, N. R. et al. Research review: polygenic methods and their application to psychiatric traits. *J. Child Psychol. Psychiatry* **55**, 1068–1087 (2014).
13. Euesden, J., Lewis, C. M. & O'Reilly, P. F. PRSice: polygenic risk score software. *Bioinformatics* **31**, 1466–1468 (2015).
14. Choi, S. W. & O'Reilly, P. F. PRSice-2: polygenic risk score software for biobank-scale data. *Gigascience* **8**, giz082 (2019).
15. Agerbo, E. et al. Polygenic risk score, parental socioeconomic status, family history of psychiatric disorders, and the risk for schizophrenia: a Danish population-based study and meta-analysis. *JAMA Psychiatry* **72**, 635–641 (2015).
16. Mavaddat, N. et al. Polygenic risk scores for prediction of breast cancer and breast cancer subtypes. *Am. J. Hum. Genet.* **104**, 21–34 (2019).
17. Mullins, N. et al. Polygenic interactions with environmental adversity in the aetiology of major depressive disorder. *Psychol. Med.* **46**, 759–770 (2016).
18. 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* **135**, 2091–2101 (2017).
19. Mak, T. S. H. et al. Polygenic scores via penalized regression on summary statistics. *Genet. Epidemiol.* **41**, 469–480 (2017).
20. Ge, T. et al. Polygenic prediction via Bayesian regression and continuous shrinkage priors. *Nat. Commun.* **10**, 1–10 (2019).
21. Speed, D. & Balding, D. J. MultiBLUP: improved SNP-based prediction for complex traits. *Genome Res.* **24**, 1550–1557 (2014).
22. Zhou, X., Carbonetto, P. & Stephens, M. Polygenic modeling with Bayesian sparse linear mixed models. *PLoS Genet.* **9**, e1003264 (2013).
23. Shi, J. et al. Winner's curse correction and variable thresholding improve performance of polygenic risk modeling based on genome-wide association study summary-level data. *PLoS Genet.* **12**, e1006493 (2016).
24. Lello, L. et al. Accurate genomic prediction of human height. *Genetics* **210**, 477–497 (2018).
25. Sudlow, C. et al. UK Biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS Med.* **12**, e1001779 (2015).
26. Purcell, S. et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* **81**, 559–575 (2007).
27. Evans, L. M. et al. Comparison of methods that use whole genome data to estimate the heritability and genetic architecture of complex traits. *Nat. Genet.* **50**, 737–745 (2018).
28. Coleman, J. R. I. et al. Quality control, imputation and analysis of genome-wide genotyping data from the Illumina HumanCoreExome microarray. *Brief. Funct. Genomics* **15**, 298–304 (2016).
29. Marees, A. T. et al. A tutorial on conducting genome-wide association studies: quality control and statistical analysis. *Int. J. Methods Psychiatr. Res.* **27**, e1608 (2018).
30. Anderson, C. A. et al. Data quality control in genetic case-control association studies. *Nat. Protoc.* **5**, 1564–1573 (2010).

31. Speed, D. & Balding, D. J. SumHer better estimates the SNP heritability of complex traits from summary statistics. *Nat. Genet.* **51**, 277–284 (2019).
32. Han, B. & Eskin, E. Random-effects model aimed at discovering associations in meta-analysis of genome-wide association studies. *Am. J. Hum. Genet.* **88**, 586–598 (2011).
33. Drepper, U., Miller, S. & Madore, D. md5sum(1): compute/check MD5 message digest. Linux man page (accessed 20 October 2018); https://linux.die.net/man/1/md5sum
34. National Center for Biotechnology Information. US National Library of Medicine. Data changes that occur between builds. in *SNP FAQ Archive. NCBI Help Manual*. https://www.ncbi.nlm.nih.gov/books/NBK44467/ (2005).
35. Hinrichs, A. S. et al. The UCSC Genome Browser Database: update 2006. *Nucleic Acids Res.* **34**, D590–D598 (2006).
36. Niemi, M. E. K. et al. Common genetic variants contribute to risk of rare severe neurodevelopmental disorders. *Nature* **562**, 268–271 (2018).
37. Chang, C. C. et al. Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience* **4**, 7 (2015).
38. Chen, L. M. et al. PRS-on-Spark (PRSoS): a novel, efficient and flexible approach for generating polygenic risk scores. *BMC Bioinforma.* **19**, 295 (2018).
39. Accounting for sex in the genome. *Nat. Med.*23, 1243–1243 (2017).
40. König, I. R. et al. How to include chromosome X in your genome-wide association study. *Genet. Epidemiol.* **38**, 97–103 (2014).
41. Wray, N. R. et al. Pitfalls of predicting complex traits from SNPs. *Nat. Rev. Genet.* **14**, 507–515 (2013).
42. Viechtbauer, W. & Cheung, M. W.-L. Outlier and influence diagnostics for meta-analysis. *Res. Synth. Methods* **1**, 112–125 (2010).
43. Socrates, A. et al. Polygenic risk scores applied to a single cohort reveal pleiotropy among hundreds of human phenotypes. Preprint at https://www.biorxiv.org/content/10.1101/203257v1 (2017).
44. Devlin, B. & Roeder, K. Genomic control for association studies. *Biometrics* **55**, 997–1004 (1999).
45. Vilhjálmsson, B. J. et al. Modeling linkage disequilibrium increases accuracy of polygenic risk scores. *Am. J. Hum. Genet.* **97**, 576–592 (2015).
46. Newcombe, P. J. et al. A flexible and parallelizable approach to genome-wide polygenic risk scores. *Genet. Epidemiol.* **43**, 730–741 (2019).
47. Loh, P.-R. et al. Mixed-model association for biobank-scale datasets. *Nat. Genet.* **50**, 906–908 (2018).
48. The 1000 Genomes Project Consortium. A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).
49. Privé, F. et al. Efficient analysis of large-scale genome-wide data with two R packages: bigstatsr and bigsnpr. *Bioinformatics* **34**, 2781–2787 (2018).
50. Márquez-Luna, C., Loh, P.-R. & Price, A. L. Multiethnic polygenic risk scores improve risk prediction in diverse populations. *Genet. Epidemiol.* **41**, 811–823 (2017).
51. Clayton, D. Link functions in multi-locus genetic models: implications for testing, prediction, and interpretation. *Genet. Epidemiol.* **36**, 409–418 (2012).
52. Price, A. L. et al. Principal components analysis corrects for stratification in genome-wide association studies. *Nat. Genet.* **38**, 904–909 (2006).
53. Astle, W. & Balding, D. J. Population structure and cryptic relatedness in genetic association studies. *Stat. Sci.* **24**, 451–471 (2009).
54. Kong, A. et al. The nature of nurture: effects of parental genotypes. *Science* **359**, 424–428 (2018).
55. Selzam, S. et al. Comparing within- and between-family polygenic score prediction. *Am. J. Hum. Genet.* **105**, 351–363 (2019).
56. Price, A. L. et al. New approaches to population stratification in genome-wide association studies. *Nat. Rev. Genet.* **11**, 459–463 (2010).
57. Cheesman, R. et al. Comparison of adopted and nonadopted individuals reveals gene–environment interplay for education in the UK Biobank. *Psychol. Sci.* **31**, 582–591 (2020).
58. Mostafavi, H. et al. Variable prediction accuracy of polygenic scores within an ancestry group. *eLife* **9**, e48376 (2020).
59. Young, A. I. et al. Deconstructing the sources of genotype-phenotype associations in humans. *Science* **365**, 1396–1400 (2019).
60. Kim, M. S., Patel, K. P., Teng, A. K., Berens, A. J. & Lachance, J. Genetic disease risks can be misestimated across global populations. *Genome Biol.* **19**, 179 (2018).
61. Martin, A. R. et al. Human demographic history impacts genetic risk prediction across diverse populations. *Am. J. Hum. Genet.* **100**, 635–649 (2017).
62. Duncan, L. et al. Analysis of polygenic risk score usage and performance in diverse human populations. *Nat. Commun.* **10**, 3328 (2019).
63. Selzam, S. et al. Predicting educational achievement from DNA. *Mol. Psychiatry* **22**, 267–272 (2017).
64. Lee, J. J. et al. Gene discovery and polygenic prediction from a genome-wide association study of educational attainment in 1.1 million individuals. *Nat. Genet.* **50**, 1112–1121 (2018).
65. Grotzinger, A. D. et al. Genomic structural equation modelling provides insights into the multivariate genetic architecture of complex traits. *Nat. Hum. Behav.* **3**, 513–525 (2019).
66. Maier, R. M. et al. Improving genetic prediction by leveraging genetic correlations among human diseases and traits. *Nat. Commun.* **9**, 989 (2018).
67. Krapohl, E. et al. Multi-polygenic score approach to trait prediction. *Mol. Psychiatry* **23**, 1368–1374 (2018).
68. Ruderfer, D. M. et al. Polygenic dissection of diagnosis and clinical dimensions of bipolar disorder and schizophrenia. *Mol. Psychiatry* **19**, 1017–1024 (2014).
69. Bipolar Disorder and Schizophrenia Working Group of the Psychiatric Genomics Consortium. Genomic dissection of bipolar disorder and schizophrenia, including 28 subphenotypes. *Cell* **173**, 1705–1715.e16 (2018).
70. Turley, P. et al. Multi-trait analysis of genome-wide association summary statistics using MTAG. *Nat. Genet.* **50**, 229–237 (2018).
71. Rheenen, W. et al. Genetic correlations of polygenic disease traits: from theory to practice. *Nat. Rev. Genet.* **20**, 567–581 (2019).
72. Visscher, P. M. et al. Statistical power to detect genetic (co)variance of complex traits using SNP data in unrelated samples. *PLoS Genet.* **10**, e1004269 (2014).
73. Janssens, M. J. J. Co-heritability: its relation to correlated response, linkage, and pleiotropy in cases of polygenic inheritance. *Euphytica* **28**, 601–608 (1979).
74. Pirinen, M., Donnelly, P. & Spencer, C. C. A. Including known covariates can reduce power to detect genetic effects in case-control studies. *Nat. Genet.* **44**, 848–851 (2012).
75. Lee, S. H. et al. A better coefficient of determination for genetic profile analysis. *Genet. Epidemiol.* **36**, 214–224 (2012).
76. Heinzl, H., Waldhör, T. & Mittlböck, M. Careful use of pseudo R-squared measures in epidemiological studies. *Stat. Med.* **24**, 2867–2872 (2005).
77. Lee, S. H. et al. Estimating missing heritability for disease from genome-wide association studies. *Am. J. Hum. Genet.* **88**, 294–305 (2011).
78. Won, H.-H. et al. Disproportionate contributions of select genomic compartments and cell types to genetic risk for coronary artery disease. *PLoS Genet.* **11**, e1005622 (2015).

79. Santoro, M. L. et al. Polygenic risk score analyses of symptoms and treatment response in an antipsychotic-naive first episode of psychosis cohort. *Transl. Psychiatry* **8**, 1–8 (2018).
80. Power, R. A. et al. Polygenic risk scores for schizophrenia and bipolar disorder predict creativity. *Nat. Neurosci.* **18**, 953–955 (2015).
81. Mullins, N. et al. GWAS of suicide attempt in psychiatric disorders and association with major depression polygenic risk scores. *Am. J. Psychiatry* **176**, 651–660 (2019).
82. Khera, A. V. et al. Polygenic prediction of weight and obesity trajectories from birth to adulthood. *Cell* **177**, 587–596.e9 (2019).
83. Khera, A. V. et al. Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations. *Nat. Genet.* **50**, 1219–1224 (2018).
84. Du Rietz, E. et al. Association of polygenic risk for attention-deficit/hyperactivity disorder with co-occurring traits and disorders. *Biol. Psychiatry Cogn. Neurosci. Neuroimaging* **3**, 635–643 (2018).
85. Clayton, D. G. Prediction and interaction in complex disease genetics: experience in type 1 diabetes. *PLoS Genet.* **5**, e1000540 (2009).
86. Dudbridge, F., Pashayan, N. & Yang, J. Predictive accuracy of combined genetic and environmental risk scores. *Genet. Epidemiol.* **42**, 4–19 (2018).
87. Torkamani, A., Wineinger, N. E. & Topol, E. J. The personal and clinical utility of polygenic risk scores. *Nat. Rev. Genet.* **19**, 581–590 (2018).
88. Lambert, S. A., Abraham, G. & Inouye, M. Towards clinical utility of polygenic risk scores. *Hum. Mol. Genet.* **28(R2)**, R133–R142 (2019).
89. Gibson, G. On the utilization of polygenic risk scores for therapeutic targeting. *PLoS Genet.* **15**, e1008060 (2019).
90. Rose, G. Sick individuals and sick populations. *Int. J. Epidemiol.* **30**, 427–432 (2001).
91. Wynants, L., Collins, G. S. & Van Calster, B. Key steps and common pitfalls in developing and validating risk models. *BJOG* **124**, 423–432 (2017).
92. Janssens, A. C. J. W. & Joyner, M. J. Polygenic risk scores that predict common diseases using millions of single nucleotide polymorphisms: is more, better? *Clin. Chem.* **65**, 609–611 (2019).
93. Baverstock, K. Polygenic scores: are they a public health hazard? *Prog. Biophys. Mol. Biol.* **149**, 4–8 (2019).
94. Janssens, A. C. J. W. Validity of polygenic risk scores: are we measuring what we think we are? *Hum. Mol. Genet.* **28**, R143–R150 (2019).
95. Boyle, E. A., Li, Y. I. & Pritchard, J. K. An expanded view of complex traits: from polygenic to omnigenic. *Cell* **169**, 1177–1186 (2017).
96. Sniekers, S. et al. Genome-wide association meta-analysis of 78,308 individuals identifies new loci and genes influencing human intelligence. *Nat. Genet.* **49**, 1107–1112 (2017).
97. Bowden, J., Davey Smith, G. & Burgess, S. Mendelian randomization with invalid instruments: effect estimation and bias detection through Egger regression. *Int. J. Epidemiol.* **44**, 512–525 (2015).
98. Hartwig, F. P. et al. Two-sample Mendelian randomization: avoiding the downsides of a powerful, widely applicable but potentially fallible technique. *Int. J. Epidemiol.* **45**, 1717–1726 (2016).
99. Krapohl, E. et al. Phenome-wide analysis of genome-wide polygenic scores. *Mol. Psychiatry* **21**, 1188–1193 (2016).
100. Pingault, J.-B. et al. Using genetic data to strengthen causal inference in observational research. *Nat. Rev. Genet.* **19**, 566–580 (2018).
101. Tibshirani, R. Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. Ser. B Methodol.* **58**, 267–288 (1996).
102. Mak, T. S. H. et al. Polygenic scores for UK Biobank scale data. Preprint at https://www.biorxiv.org/content/10.1101/252270v3 (2018).
103. Machiela, M. J. et al. Evaluation of polygenic risk scores for predicting breast and prostate cancer risk. *Genet. Epidemiol.* **35**, 506–514 (2011).
104. Falconer, D. S. *Introduction to Quantitative Genetics* (Ronald Press, 1960).
105. Jaffee, S. & Price, T. Gene–environment correlations: a review of the evidence and implications for prevention of mental illness. *Mol. Psychiatry* **12**, 432–442 (2007).

Acknowledgements

We thank the participants in the UK Biobank and the scientists involved in the construction of this resource. We thank Jonathan Coleman and Kylie Glanville for help in management of the UK Biobank resource at King's College London, and we thank Jack Euesden, Carla Giner-Delgado, Clive Hoggart, Hei Man Wu, Tom Bond, Gerome Breen, Cathryn Lewis, Cecile Janssens and Pak Sham for helpful discussions. This research has been conducted using the UK Biobank Resource under application 18177 (P.F.O.). P.F.O. receives funding from the UK Medical Research Council (MR/N015746/1). S.W.C. is funded by the UK Medical Research Council (MR/N015746/1). This report represents independent research partially funded by the National Institute for Health Research (NIHR) Biomedical Research Centre at South London and Maudsley NHS Foundation Trust and King's College London. The views expressed are those of the authors and not necessarily those of the NHS, the NIHR or the Department of Health.

Author contributions

P.F.O. conceived, prepared and wrote the manuscript, with feedback from S.W.C. and T.S.-H.M. S.W.C. performed the analyses and produced the online tutorial, with feedback from P.F.O.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to P.F.OR.

Peer review information *Nature Protocols* thanks Dorret Boomsma, Brandon Johnson and Anubha Mahajan for their contribution to the peer review of this work.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Received: 17 December 2018; Accepted: 5 May 2020;
Published online: 24 July 2020

Related links

Key references using this protocol

PRS tutorial: https://choishingwan.github.io/PRS-Tutorial/
GWAS Tutorial: https://github.com/MareesAT/GWA_tutorial
PRSice software: https://www.prsice.info
LDpred software: https://github.com/bvilhjal/ldpred
Lassosum software: https://github.com/tshmak/lassosum/blob/master/README.md
The R project for statistical computing: https://www.r-project.org