Polygenic Index Repository User Guide (version 1.0)

Prepared by
Dan Benjamin
David Cesarini
Aysu Okbay
Patrick Turley

January 19, 2021

Contents

In this guide, we summarize the key information regarding the construction of the Repository PGIs, lay out some of the interpretational issues that are likely to arise as researchers begin to use PGIs from the Repository, and outline how we suggest thinking through those issues.

1 Summary Information About Repository PGIs

Here, we provide a brief summary of how the PGIs were constructed (please see Methods for more details). We refer the reader to the relevant tables where more information can be found.

1.1 Phenotype definitions and GWAS for single-trait PGIs

The single-trait PGIs are based on meta-analyses of summary statistics from three sources: GWAS conducted in 23andMe and UKB (some of which are novel), and published GWAS. Supplementary Table 5 lists the phenotype measures used in the UKB GWAS that we conducted ourselves, including information on how repeated measures were handled and the sample size in each of the three UKB partitions. Supplementary Table 6 lists the phenotype definitions and describes the association models for all novel or published 23andMe GWAS, and for published GWAS, it cites the relevant publications.

In order to avoid sample overlap between the GWAS and Repository datasets, we conducted multiple versions of the GWAS meta-analysis for each phenotype (so as to have, for each dataset, a version of the meta-analysis that excludes that dataset). Supplementary Table 8 lists all GWAS meta-analyses used as inputs for the single-trait PGIs. The "Repository Datasets Sumstats are Used for" column shows which meta-analysis the PGI weights come from for each Repository dataset.

1.2 Supplementary phenotypes and MTAG for multi-trait PGIs

The multi-trait PGIs are based on MTAG analyses[1] of genetically correlated (pairwise $r_g > 0.6$) phenotypes. Supplementary Table 9 lists genetic correlations between all pairs of phenotypes considered in the Repository. Based on these genetic correlations, MTAG groups were formed for each phenotype. These groups are listed in Supplementary Table 10. The "Input Files" column lists, for each group, the codes for the single-trait GWAS (see Supplementary Table 8 for the GWAS that the codes refer to) that were included in the multi-trait MTAG analysis. As is the case for the single-trait PGIs, there are multiple versions for each phenotype because of sample overlap with the Repository datasets and the "Repository Datasets Sumstats are Used for" column shows which MTAG analysis the PGI weights for each Repository dataset comes from.

1.3 PGI construction

The PGIs were made using LDpred[2] applied to HapMap3 SNPs. The inclusion criterion was that the "expected" out-of-sample predictive power of a PGI be greater than 1%. The expected predictive power was calculated from the results of the GWAS meta-analysis[3]. The expected predictive power of each single- and multi-trait PGI (including the ones not included in the Repository because they did not pass the cutoff of 1%) are shown in Supplementary Tables 1 and 2, respectively.

1.4 Genotyping, imputation, and phenotype definitions in Repository datasets

Details on genotyping and imputation of the Repository datasets are listed in Supplementary Table 11. Supplementary Table 12 lists the phenotype definitions for the subset of these datasets that we used to validate our PGIs, excluding UK Biobank. The phenotype definitions for UK Biobank can be found in Supplementary Table 5.

1.5 PGIs from publicly available GWAS

In order to assess the gains in predictive power when using the Repository PGIs as opposed to PGIs obtained using publicly available GWAS, we constructed a set of "public PGIs." These "public PGIs" were obtained using the same methodology that we used for our Repository PGIs and weights from the largest GWAS in

the public domain that does not have sample overlap with the validation dataset. Supplementary Table 13 lists these publicly available GWAS. Again, there are multiple versions for each phenotype that were used for different validation datasets in order to avoid sample overlap. The table shows which version was used for which dataset.

1.6 Predictive power of Repository PGIs in validation datasets

Supplementary Table 3 shows the observed predictive power of the single- and multi-trait Repository PGIs in our five validation datasets, together with 95% confidence intervals obtained using a bootstrap with 1000 repetitions. The table also shows the difference between the predictive power of "public PGIs" and single-/multi-trait Repository PGIs, as well as the difference between the predictive power of single- and multi-trait PGIs.

1.7 Estimates of ρ in HRS, WLS, and UKB

In Supplementary Table 4, we provide estimates of the amount of measurement error, ρ, corresponding to single- and multi-trait PGIs for phenotypes available in three of our validation datasets: HRS, WLS, and UKB (third partition). In HRS and WLS, we also provide jackknife standard errors for the ρ estimates. Because producing jackknife standard errors in UKB is very computationally expensive, for UKB we provide standard errors only for three phenotypes: friend satisfaction, educational attainment and height. We chose these three phenotypes so as to have one each corresponding to a single-trait PGI with low (friend satisfaction), medium (educational attainment)and high predictive power (height).

2 Interpretational Considerations

In this section, we lay out some of the interpretational issues that are likely to arise as researchers begin to use PGIs from the Repository, and we outline how we suggest thinking through those issues. The executive summary is as follows:

1. The methodologies used to conduct the GWAS and to construct the PGI weights jointly determine the additive SNP factor that is proxied for by the PGI.

2. These methodologies, together with the PGI phenotype, determine the relative importance of various potential confounds to a causal interpretation of PGI associations. In most applications, researchers should control for PCs (which are available from the datasets, along with the PGIs, as part of the Repository).

3. Whether and which confounds should be highlighted (or can be safely ignored) depends on the application.

4. While a multi-trait PGI generally has higher predictive power than its corresponding single-trait PGI, it is subject to additional potential confounds. This tradeoff should be evaluated when deciding whether to use a single-trait or multi-trait PGI.

5. Currently, the most feasible way to cleanly identify causal effects of a PGI is to conduct a within-family analysis (where the PGI is analyzed in a sibling sample, with sibling fixed effects). In the absence of clean identification of a causal effect, researchers should highlight the potential confounds to a causal interpretation.

6. In interpreting PGI associations (whether causal or not), it is important to keep in mind that genetic effects can operate through environmental mechanisms, and these mechanisms may be modifiable. For this reason, terminology such as "genetic endowment" should be avoided. Researchers should remind readers of the potential role of environmental mechanisms in explaining PGI associations.

The following subsections, numbered 1 through 6, provide more detail on the points above. In addition to attending to these interpretational issues, we urge users of the Repository to conduct power calculations

prior to undertaking analyses; to pursue analyses only if they are adequately powered; and, when feasible, to preregister planned analyses (along with the power calculations).

We note that the GWAS from which the Repository PGIs are constructed were conducted in European-ancestry samples (where "European-ancestry" is operationalized differently depending on the study but almost always involves sample restrictions based on the genetic PCs; e.g., for our UKB GWAS, see the "UKB GWAS" subsection of Section I in Methods). Due to the limited portability of such GWAS results to other ancestries, for the PGIs released to participating datasets, the current version of the Repository is restricted to individuals of European ancestries, as defined by how their genetic PCs cluster together with those classified as having European ancestries in the 1000 Genomes Project (see the "Subject-level QC in Repository Cohorts" subsection of Section II in Methods).

2.1 GWAS and PGI-weight methodologies and the additive SNP factor

In the Supplementary Methods section 6, we showed how the set of control variables used in a GWAS affects the additive SNP factor proxied for by a PGI. The choice of controls, however, is just one of many dimensions of GWAS methodology. A change to any of these dimensions is likely to result in a different additive SNP factor (with a different interpretation). For example, it is increasingly common for datasets to conduct association analyses using mixed-linear models[4,5] rather than OLS. Since mixed-linear models often produce estimates that are more robust to stratification, the additive SNP factor will be akin to that generated by an OLS-based GWAS with some additional controls for stratification. Knowledge of the methodology of the GWAS underlying a particular PGI is therefore often a necessary first step for understanding what additive SNP factor a specific PGI is proxying for. For example, the methodologies underlying the GWASs we conducted in UKB for the PGIs in the Repository are described in the "UKB GWAS" subsection of Section I in Methods. Information about association models in 23andMe GWAS can be found in Supplementary Table 6.

The PGI-weight methodology can matter, as well. For example, our Repository PGI weights are calculated from the GWAS results using the HapMap3 set of SNPs, which primarily captures common genetic variation. If PGI weights were instead calculated based on results from SNPs that capture a different mix of common and rare genetic variation, then the additive SNP factor corresponding to that PGI would have a different interpretation: it would be the best linear predictor based on that set of SNPs.

2.2 Potential confounds to a causal interpretation

It is increasingly understood that standard GWAS approaches with a limited set of controls – for example, sex, age, and up to 10 PCs, as in most of the GWAS underlying the Repository PGIs – generate PGIs that can be subject to a number of confounds to a causal interpretation[6–9]. For example, PGIs for educational attainment derive a substantial share of their overall predictive power from their positive association with rearing environment. In behavior-genetic parlance, this positive correlation arises due to the vertical transmission of the parental phenotypes (parents' phenotypes impact their children's phenotypes). In recent molecular-genetic research, this source of positive gene-environment correlation has been labelled "genetic nurture"[7]. This effect can be further exacerbated by assortative mating at the genetic level.

As another example, when the PCs are estimated in a small sample, they are often not very accurate proxies for ancestry. Failure to adequately control for genetic ancestry gives rise to "population stratification"[10]: because the PGI is correlated with ancestry, which in turn is correlated with ethnicity and regional background, it picks up cultural or environmental factors that are correlated with these factors. In many empirical applications, the goal is to estimate an association that is net of any such cultural and environmental confounds. In such cases, it may be possible to mitigate concerns that the underlying GWAS may have relied on inaccurate ancestry controls by including a richer-than-usual set of environmental controls in the analysis of the PGI (i.e., in the vector $\mathbf{z}_i$ in equations (1) and (2) in the main text).

Indeed, in most applications, researchers should include PCs in the set of environmental controls. When estimating PGI-by-environment interactions, researchers should additionally control for interactions between PCs and the "environment" variable[11]. For these purposes, dataset-specific PCs are made available as part of the Repository. However, it is important to recognize and acknowledge that the PCs are not fully accurate measures of ancestry, so even after controlling for PCs, residual confounding almost surely remains.

The relevance of potential confounds could vary across phenotypes[6,8,9]. For example, genetic nurture effects are much smaller for height than educational attainment. Although the noisiness of PCs as measures of ancestry in a given sample is the same across phenotypes, the noisiness is likely to be substantially more problematic for educational attainment than for height because finer ancestral distinctions (which require more PCs to capture) probably matter for the social and environmental factors that influence educational attainment. More generally, it seems likely that potential confounds to a causal interpretation matter more for PGIs for social and behavioral phenotypes than for PGIs for more biologically proximal phenotypes.

2.3 Importance of confounds depends on the application

The degree to which potential confounds to a causal interpretation matter depends on how the PGI is used. For example, if a PGI is used as a control variable to increase precision for a randomized treatment evaluation[12,13], then the goal is simply to use controls that absorb as much residual variance as possible (and avoid controlling for any variables realized after the randomized intervention). Since the PGI is simply being used as a predictive variable, its interpretation is irrelevant in that case. As a contrasting example, consider the illustrative application in the main text that tests how much parental education mediates the predictive power of the PGI for educational attainment. There, the PGI should be understood as capturing some of the genetic nurture effects and ancestry associations with education. In most applications, the potential confounds do matter and should be highlighted.

2.4 Single- versus multi-trait PGIs

MTAG coefficient estimates are a weighted sum of GWAS coefficient estimates. Relative to GWAS estimates, MTAG coefficients have a lower expected mean-squared error, which means that multi-trait PGIs will in general have greater predictive power.

Multi-trait PGIs, however, do not necessarily have the same interpretation as single-trait PGIs. Because MTAG estimates are a weighted average of GWAS estimates for several traits, the multi-trait PGI based on MTAG estimates is roughly a weighted average of PGIs for the set of included traits. As a result, a multi-trait PGI may be correlated with an outcome variable if that outcome variable is genetically correlated with a supplementary phenotype for the multi-trait PGI. This can even be the case if the outcome variable and the target phenotype are not genetically correlated.

Therefore, **results using the multi-trait PGI have the same interpretation as results using the single-trait PGI in analyses where**

(i) the dependent variable and the PGI correspond to the same phenotype, *and*

(ii) no other covariates are included in the regression that are genetically correlated with any of the supplementary phenotypes used to construct the multi-trait PGI.

However, **results from the multi-trait PGI should be interpreted differently than results from the single-trait PGI—perhaps being driven by a supplementary phenotype rather than the target phenotype—if either (i) or (ii) is violated**. In that case, the risk of spurious results increases when (a) the GWAS sample size for the target GWAS is small relative to the GWAS sample size of the supplementary phenotypes, and (b) the genetic correlation between the target phenotype and the supplementary phenotypes is only moderate. Researchers who use multi-trait PGIs should make clear to readers how large the potential for a confounded interpretation is and how much it matters for the application at hand. To facilitate this, we report the average weight that MTAG assigns to each traits that enter into the multi-trait PGIs. Although these weights may vary by SNP when there is variation in the sample size across SNPs, they are informative about where the predictive power comes from.

As described in Section 3 above, in settings where the PGI is just being used as a covariate (e.g., as a control variable in a randomized controlled trial), the confounds associated with using the multi-trait PGI may be less important. In all settings, however, it is good practice to describe which supplementary phenotypes were included in the multi-trait PGI if they are included in an analysis.

2.5 Identifying causal effects of a PGI

A clean way to identify the causal effects of a PGI is to conduct the analysis of the PGI in a sibling sample and control for family fixed effects (even if the PGI itself is generated from currently-standard (between-family) GWAS, as the Repository PGIs are). The family fixed effects control for all common factors shared by siblings within a family, including the parents that the siblings share. This empirical strategy exploits a natural experiment: conditional on a pair of biological parents, genetic inheritance is random. A robustly estimated non-zero within-family association from a large and attrition-free sample would provide strong evidence of a causal effect of the PGI. The coefficient estimate could be interpreted as a weighted average of treatment effects from hypothetical experiments that randomly modify PGIs at conception[14,15].

The additive SNP factors corresponding to the PGIs in the Repository are not the best linear predictors conditional on a pair of biological parents (because the GWAS underlying the PGI weights do not control for the biological parents). The PGIs proxying for additive SNP factors that would be the best linear predictors for such a "within-family analysis" would be PGIs constructed from GWAS that control for parental genotypes or from GWAS (in sibling samples) that control for family fixed effects. Unfortunately, to date genotyped family-based samples have been too small to produce reliable "within-family PGIs." The Repository does not yet contain any such PGIs. Ultimately, however, when genotyped family-based samples become sufficiently large, the resulting within-family PGIs will be more predictive for within-family analyses than PGIs constructed from currently-standard (between-family) GWAS.

2.6 Genetic effects can operate through environmental mechanisms

We encourage researchers who use PGIs in their research to be mindful of three important issues of interpretation for the causal effects of a PGI. First, a PGI could exert its effects through the environment[16]. Consider a PGI for BMI[12]. Suppose a within-family association analysis yields unambiguous evidence of a within-family association between the PGI and BMI. Even though the within-family design provides strong support for a causal interpretation, this does *not* imply that the SNPs in the PGI must be influencing BMI through some narrowly physiological mechanism. In principle, the sibling differences in BMI could arise because of sibling differences in genes that influence the proneness to eat sweets, exercise habits, or myriads of other behaviors with downstream effects on BMI. PGIs for seemingly "biological" phenotypes can thus have a substantial behavioral component. A PGI for lung health may similarly derive predictive power from SNPs that influence lung health very indirectly, through smoking habits[17,18].

Second and relatedly, it is therefore a fallacy to assume that any genetic sources of heterogeneity captured by a PGI are immutable, or at least harder to modify than environmental sources of heterogeneity. Indeed, the possibility of identifying modifiable mechanisms through which PGIs exert some of their effects motivates some of the research using PGIs[19,20]. To continue the BMI example, the widespread replacement of sugar by low-calorie sweeteners or better behavioral tools for avoiding temptation could eliminate or reduce the effect of the PGI on BMI. Because of these issues, we urge caution in describing PGIs as "genetic endowments," or related terminology that may, however inadvertently, promote the common misunderstanding that genes are a resource that is easily separable from choices made in light of that resource.

Third, because the additive genetic factor depends on the environment, the PGI may be context dependent. That is, the same PGI may have a different predictive power in two different samples if there are differences in the population sampled, the sampling methodology, or the phenotype measure. For example, the research participants from the UKB were recruited through the mail and had a 5.5% response rate. Those that responded to the recruitment mailers were more healthy and more educated than the UK population as a whole[21,22]. Because UKB participants make up a large fraction of the discovery sample for many phenotypes, it may be that the PGI from this Repository does not correspond to a PGI that would be produced from a representative sample or a sample of individuals not from the UK.

3 Citation Instructions

Please include the following citation in any publication based on the Repository PGIs along with the citations for the GWAS included in the single-trait or multi-trait input GWAS for the PGI:

TBA

References

[1] Turley, P., *et al.* Multi-trait analysis of genome-wide association summary statistics using MTAG. *Nature Genetics*, **50**, 229–237 (2018).

[2] Vilhjálmsson, B.J., *et al.* Modeling Linkage Disequilibrium Increases Accuracy of Polygenic Risk Scores. *American Journal of Human Genetics*, **97**, 576–592 (2015).

[3] Daetwyler, H.D., Villanueva, B., and Woolliams, J.A. Accuracy of predicting the genetic risk of disease using a genome-wide approach. *PLoS One*, **3**, e3395 (2008).

[4] Kang, H.M., *et al.* Variance component model to account for sample structure in genome-wide association studies. *Nature Genetics*, **42**, 348–354 (2010).

[5] Loh, P.R., *et al.* Efficient Bayesian mixed-model analysis increases association power in large cohorts. *Nature Genetics*, **47**, 284–290 (2015).

[6] Lee, J.J., *et al.* Gene discovery and polygenic prediction from a genome-wide association study of educational attainment in 1.1 million individuals. *Nature Genetics*, **50**, 1112–1121 (2018).

[7] Kong, A., *et al.* The nature of nurture: Effects of parental genotypes. *Science*, **359**, 424–428 (2018).

[8] Young, A.I., *et al.* Relatedness disequilibrium regression estimates heritability without environmental bias. *Nature Genetics* (2018).

[9] Morris, T.T., *et al.* Population phenomena inflate genetic associations of complex social traits. *Science Advances* (2020).

[10] Hamer, D. and Sirota, L. Beware the chopsticks gene. *Molecular Psychiatry*, **5**, 11–13 (2000).

[11] Keller, M.C. Gene x Environment Interaction Studies Have Not Properly Controlled for Potential Confounders: The Problem and the (Simple) Solution. *Biological Psychiatry*, **75**, 18–24 (2013).

[12] Benjamin, D.J., *et al.* The Promises and Pitfalls of Genoeconomics. *Annual Review of Economics*, **4**, 627–662 (2012).

[13] Rietveld, C.A., *et al.* GWAS of 126,559 individuals identifies genetic variants associated with educational attainment. *Science*, **340**, 1467–1471 (2013).

[14] Angrist, J.D. and Pischke, J.S. *Mostly harmless econometrics: An empiricist's companion* (2008).

[15] Yitzhaki, S. On using linear regressions in welfare economics. *Journal of Business and Economic Statistics* (1996).

[16] Jencks, C. Heredity, environment, and public policy reconsidered. *American Sociological Review*, **45**, 723–736 (1980).

[17] Thorgeirsson, T.E., *et al.* A variant associated with nicotine dependence, lung cancer and peripheral arterial disease. *Nature*, **452**, 638–642 (2008).

[18] Amos, C.I., *et al.* Genome-wide association scan of tag SNPs identifies a susceptibility locus for lung cancer at 15q25.1. *Nature Genetics*, **40**, 616–622 (2008).

[19] Belsky, D.W. and Harden, K.P. Phenotypic Annotation: Using Polygenic Scores to Translate Discoveries From Genome-Wide Association Studies From the Top Down. *Current Directions in Psychological Science*, **28**, 82–90 (2019).

[20] Conley, D. Socio-genomic research using genome-wide molecular data. *Annual Review of Sociology*, **42**, 275–299 (2016).

[21] Fry, A., *et al.* Comparison of Sociodemographic and Health-Related Characteristics of UK Biobank Participants with Those of the General Population. *American Journal of Epidemiology*, **186**, 1026–1034 (2017).

[22] Keyes, K.M. and Westreich, D. UK Biobank, big data, and the consequences of non-representativeness. *The Lancet*, **393**, 1297 (2019).