Supplementary Information

April 9, 2021

Includes Supplementary Methods, Supplementary Note, Supplementary Figures and FAQs

Supplementary Methods

April 9, 2021

Contents

1 Setup

Denote individual i's phenotype by $y_i^\star$. For some genetic variant j, denote the allele count for individual i by $x_{ij}^\star$. Without loss of generality in this derivation, we use mean-centered transformation of the phenotype and allele counts, $y_i \equiv y_i^\star - \mathbb{E}\left(y_i^\star\right)$ and $x_{ij} \equiv x_{ij}^\star - \mathbb{E}\left(x_{ij}^\star\right)$. We denote the vector of mean-centered allele counts across J SNPs for individual i by $\mathbf{x}_i = (x_{i1}, x_{i2}, \ldots, x_{iJ})$.

As a benchmark, consider the standardized best linear predictor of y_i given the J SNPs, $\mathbf{x}_i$:

$$g_i \equiv \frac{\mathbf{x}_i\gamma}{\text{sd}\left(\mathbf{x}_i\gamma\right)}, \tag{1}$$

where

$$\gamma \equiv \arg\min_{\tilde{\gamma}} \mathbb{E}\left[\left(y_i - \mathbf{x}_i\tilde{\gamma}\right)^2\right]. \tag{2}$$

Thus, g_i is the weighted sum of genotypes that maximizes the expected power for predicting y_i using a linear combination of genotypes. If the set of J genetic variants is the set of all genetic variants, then g_i is referred to as the standardized additive genetic factor. The variance of y_i explained by the standardized additive genetic factor is called the narrow-sense heritability. When the set of J genetic variants is some set of genotyped SNPs, we refer to g_i as the standardized additive SNP factor and the variance of y_i is referred to as the SNP heritability.

We use h^2_{SNP} to denote the SNP heritability, which is the variation in y_i explained by the additive SNP factor. Because g_i is standardized, the SNP heritability of y_i is also the squared correlation of y_i and g_i:

$$h^2_{SNP} \equiv \frac{\left[\text{Cov}\left(y_i, g_i\right)\right]^2}{\text{Var}\left(y_i\right)\text{Var}\left(g_i\right)} = \frac{\left[\text{Cov}\left(y_i, g_i\right)\right]^2}{\text{Var}\left(y_i\right)}.$$

By basic properties of population regression, we can decompose y_i into two uncorrelated components,

$$\begin{aligned} y_i &= \frac{\text{Cov}\left(y_i, g_i\right)}{\text{Var}\left(y_i\right)} g_i + \varepsilon_{y,i} \\ &= \frac{\text{Cov}\left(y_i, g_i\right)}{\sqrt{\text{Var}\left(y_i\right)}} \frac{1}{\sqrt{\text{Var}\left(y_i\right)}} g_i + \varepsilon_{y,i} \\ &= \frac{h_{SNP}}{\sqrt{\text{Var}\left(y_i\right)}} g_i + \varepsilon_{y,i}, \end{aligned}$$

where ε_i is the component of the phenotype that is uncorrelated with g_i. We have ignored covariates in this definition of the additive SNP factor. For a model that includes covariates, we may define y_i and each x_{ij} as the phenotype and genotypes after having been residualized for the set of covariates.

A PGI for phenotype y_i is also a weighted sum of genotypes,

$$\hat{g}_i \equiv \sum_j x_{ij}\hat{\gamma}_j. \tag{3}$$

The PGI will have maximum predictive power only when $\hat{\gamma}_j = \gamma_j$ for every SNP j. In practice, methods for constructing a PGI calculate the $\hat{\gamma}_j$'s using GWAS summary statistics together with some procedure to account for linkage disequilibrium (e.g., pruning and thresholding, LD-based shrinkage-based methods). The PGI is then usually standardized to have mean zero and variance one. In the theory below, we treat the PGI as if it has been standardized.

Projecting the PGI onto the space spanned by the standardized additive SNP factor, we can express the PGI as

$$\hat{g}_i = \frac{g_i + e_i}{\text{sd}\left(g_i + e_i\right)}, \tag{4}$$

where $\text{Cov}\left(g_i, e_i\right) = 0$. In Section 2.2 below, we assume that e_i is uncorrelated with all other variables when the prediction sample is independent of the sample used to estimate $\hat{\gamma}$. (We highlight there where the assumption is used.) As shown in Section 4, this assumption will be a very good approximation when the PGI is constructed using LDpred-inf, as is the case for all the PGIs in the Repository.

The predictive power of the PGI, R^2, is therefore

$$\begin{aligned}
R^2 &\equiv \frac{\left[\text{Cov}\left(y_i, \hat{g}_i\right)\right]^2}{\text{Var}\left(y_i\right)\text{Var}\left(\hat{g}_i\right)} \\
&= \frac{\left[\text{Cov}\left(y_i, \hat{g}_i\right)\right]^2}{\text{Var}\left(y_i\right)} \\
&= \frac{\left[\text{Cov}\left(y_i, \frac{g_i+e_i}{\text{sd}(g_i+e_i)}\right)\right]^2}{\text{Var}\left(y_i\right)} \\
&= \frac{\left[\text{Cov}\left(y_i, \frac{g_i}{\text{sd}(g_i+e_i)}\right)\right]^2}{\text{Var}\left(y_i\right)} \\
&= \frac{1}{\text{Var}\left(g_i+e_i\right)}\frac{\left[\text{Cov}\left(y_i, g_i\right)\right]^2}{\text{Var}\left(y_i\right)} \\
&= \frac{h^2_{SNP}}{\text{Var}\left(g_i+e_i\right)} \\
&= \frac{h^2_{SNP}}{\text{Var}\left(g_i\right)+\text{Var}\left(e_i\right)} \\
&= \frac{h^2_{SNP}}{1+\text{Var}\left(e_i\right)} < h^2_{SNP}.
\end{aligned}$$

This inequality shows that the predictive power of the PGI is strictly less than the heritability, but the predictive power increases asymptotically towards h^2_{SNP} as the error in the PGI decreases. Also note that this calculation implies

$$\frac{h^2_{SNP}}{R^2} = 1 + \text{Var}\left(e_i\right).$$

We denote this ratio by

$$\rho^2 \equiv \frac{h^2_{SNP}}{R^2}. \tag{5}$$

Using this notation, the PGI can be written as

$$\hat{g}_i = \frac{g_i+e_i}{\text{sd}\left(g_i+e_i\right)} = \frac{g_i+e_i}{\sqrt{1+\text{Var}\left(e_i\right)}} = \frac{g_i+e_i}{\rho}. \tag{6}$$

The error in the PGI will bias any analysis that uses the PGI as a regressor instead of the additive SNP factor. (We use the term *bias* in this case to refer to a difference between expected parameter estimates of a model that includes the additive SNP factor and expected parameter estimates of a model that instead uses the PGI. This is in contrast to whether the PGI itself is a biased predictor of the additive SNP factor or of the phenotype[1].) The "errors-in-variables" bias described here is closely related to what is often referred to as "attenuation bias" because in special cases (such as a univariate regression), measurement error in the regressor attenuates the coefficient on that regressor.

In the following sections, we describe a correction for this bias. We assume that R^2 and h^2_{SNP} are known parameters. In practice, these parameters are not known and would need to be estimated. Using estimates of R^2 and h^2_{SNP}, as opposed to the true value of the parameters, affects the standard errors of the regression estimates. We discuss this issue in section 5 below.

2 Measurement-Error-Corrected Estimator

2.1 The Theoretical Regression and the Feasible Regression

Let ϕ_i denote a mean zero phenotype of interest, and let g_i denote a standardized additive SNP factor. Note that ϕ_i and g_i may correspond to different phenotypes. In addition to g_i, the model may include a vector of

mean-zero covariates, which we denote by $\mathbf{z}_i$, where $\mathbf{\Sigma}_z \equiv \text{Var}(\mathbf{z}_i)$. The model may also include interactions between g_i and some subset of covariates in $\mathbf{z}_i$. Let $\mathbf{z}_{\text{int},i}$ denote the subset of covariates that we would like to interact with g_i where $\mathbf{\Sigma}_{\text{int},z} \equiv \text{Var}(\mathbf{z}_{\text{int},i})$. We denote the vector of interactions by $\mathbf{w}_i \equiv g_i \mathbf{z}_{int,i}$, where $\mathbf{\Sigma}_w \equiv \text{Var}(\mathbf{w}_i)$ and $\mathbf{\Sigma}_{g,w} \equiv \text{Cov}(g_i, \mathbf{w}_i)$. The *theoretical regression* model we would like to estimate is:

$$\phi_i = g_i \beta_g + \mathbf{w}_i \zeta_g + \mathbf{z}_i \delta_g + \varepsilon_{g,i}. \tag{7}$$

We collect the coefficients of this regression into a single vector, denoted $\alpha_g \equiv (\beta_g, \zeta_g, \delta_g)$. We group all genetic variables (which are the variables affected by the measurement error) using the vector $\mathbf{G}_i \equiv (g_i, \mathbf{w}_i)$, where $\mathbf{\Sigma}_G \equiv \text{Var}(\mathbf{G}_i)$ and $\mathbf{\Sigma}_{G,z} \equiv \text{Cov}(\mathbf{G}_i, \mathbf{z}_i)$. The coefficients $\alpha_g \equiv (\beta_g, \zeta_g, \delta_g)$ of the regression (7) are:

$$\begin{aligned}
\alpha_g &= \begin{bmatrix} \text{Var}(\mathbf{G}_i) & \text{Cov}(\mathbf{G}_i, \mathbf{z}_i) \\ & \text{Var}(\mathbf{z}_i) \end{bmatrix}^{-1} \begin{bmatrix} \text{Cov}(\mathbf{G}_i, \phi_i) \\ \text{Cov}(\mathbf{z}_i, \phi_i) \end{bmatrix} \\
&= \begin{bmatrix} \mathbf{\Sigma}_G & \mathbf{\Sigma}_{G,z} \\ & \mathbf{\Sigma}_z \end{bmatrix}^{-1} \begin{bmatrix} \text{Cov}(\mathbf{G}_i, \phi_i) \\ \text{Cov}(\mathbf{z}_i, \phi_i) \end{bmatrix} \\
&= \mathbf{V}_g^{-1} \begin{bmatrix} \text{Cov}(\mathbf{G}_i, \phi_i) \\ \text{Cov}(\mathbf{z}_i, \phi_i) \end{bmatrix},
\end{aligned}$$

where

$$\begin{aligned}
\mathbf{V}_g &\equiv \text{Var} \begin{bmatrix} \mathbf{G}_i \\ \mathbf{z}_i \end{bmatrix} \\
&= \begin{bmatrix} \mathbf{\Sigma}_G & \mathbf{\Sigma}_{G,z} \\ & \mathbf{\Sigma}_z \end{bmatrix}
\end{aligned}$$

is the variance-covariance of all the regressors in the theoretical regression (7).

In practice, however, we do not observe g_i and instead observe the PGI $\hat{g}_i$. Similarly, instead of $\mathbf{w}_i$, we observe $\hat{\mathbf{w}}_i \equiv \hat{g}_i \mathbf{z}_{\text{int},i}$, where $\mathbf{\Sigma}_{\hat{w}} \equiv \text{Var}(\hat{\mathbf{w}}_i)$; and instead of $\mathbf{G}_i$, we observe $\hat{\mathbf{G}}_i = (\hat{g}_i, \hat{\mathbf{w}}_i)$, where $\mathbf{\Sigma}_{\hat{G}} \equiv \text{Var}\left(\hat{\mathbf{G}}_i\right)$. We use $\mathbf{\Sigma}_{\hat{G}z} \equiv \text{Cov}\left(\hat{\mathbf{G}}_i, \mathbf{z}_i\right)$ to denote the covariance between $\hat{\mathbf{G}}_i$ and $\mathbf{z}_i$. So the *feasible regression* in this case is

$$\phi_i = \hat{g}_i \beta_{\hat{g}} + \hat{\mathbf{w}}_i \zeta_{\hat{g}} + \mathbf{z}_i \delta_{\hat{g}} + \varepsilon_{\hat{g},i}, \tag{8}$$

where we now subscript the regressors and error term by $\hat{g}$ instead of g to distinguish them from the regressors and error term in the theoretical regression. As before, we collect the coefficients of this regression into a single vector, denoted $\alpha_{\hat{g}} \equiv (\beta_{\hat{g}}, \zeta_{\hat{g}}, \delta_{\hat{g}})$.

2.2 Bias from Estimating the Feasible Regression

To construct a measurement-error correction, we must first derive the relationship between α_g and $\alpha_{\hat{g}}$. The relationship we derive is closely related to results previously derived by Abel (2017). The primary differences between the relationship we derive and that of Abel (2017) are: (i) in our context, we can describe the amount of measurement error as a function of estimable parameters, h^2_{SNP} and R^2 (whereas Abel's formula treats the amount of measurement error as known), and (ii) our formula accounts for the standardization of PGIs rather than just assuming that the measurement error is an additive component of the observed value of the regressor.

We begin by writing the coefficients of the feasible regression, $\alpha_{\hat{g}} \equiv (\beta_{\hat{g}}, \zeta_{\hat{g}}, \delta_{\hat{g}})$ in terms of the notation defined above. The coefficient vector is equal to

$$\begin{aligned}
\alpha_{\hat{g}} &= \begin{bmatrix} \text{Var}\left(\hat{\mathbf{G}}_i\right) & \text{Cov}\left(\hat{\mathbf{G}}_i, \mathbf{z}_i\right) \\ & \text{Var}(\mathbf{z}_i) \end{bmatrix}^{-1} \begin{bmatrix} \text{Cov}\left(\hat{\mathbf{G}}_i, \phi_i\right) \\ \text{Cov}(\mathbf{z}_i, \phi_i) \end{bmatrix} \\
&= \begin{bmatrix} \mathbf{\Sigma}_{\hat{G}} & \mathbf{\Sigma}_{\hat{G}z} \\ & \mathbf{\Sigma}_z \end{bmatrix}^{-1} \begin{bmatrix} \text{Cov}\left(\hat{\mathbf{G}}_i, \phi_i\right) \\ \text{Cov}(\mathbf{z}_i, \phi_i) \end{bmatrix} \\
&= \mathbf{V}_{\hat{g}}^{-1} \begin{bmatrix} \text{Cov}\left(\hat{\mathbf{G}}_i, \phi_i\right) \\ \text{Cov}(\mathbf{z}_i, \phi_i) \end{bmatrix},
\end{aligned} \tag{9}$$

where

$$\mathbf{V}_{\hat{g}} \equiv \begin{bmatrix} \mathbf{\Sigma}_{\hat{G}} & \mathbf{\Sigma}_{\hat{G}z} \\ & \mathbf{\Sigma}_z \end{bmatrix}.$$

We now take each term that is related to the PGI and derive its relationship with a term related to the additive SNP factor. Considering the first term in the right column vector of (9),

$$\begin{aligned}
\text{Cov}\left(\hat{\mathbf{G}}_i, \phi_i\right) &= \begin{bmatrix} \text{Cov}\left(\hat{g}_i, \phi_i\right) \\ \text{Cov}\left(\hat{\mathbf{w}}_i, \phi_i\right) \end{bmatrix} \\
&= \begin{bmatrix} \text{Cov}\left(\frac{g_i+e_i}{\rho}, \phi_i\right) \\ \text{Cov}\left(\frac{g_i+e_i}{\rho}\mathbf{z}_{int,i}, \phi_i\right) \end{bmatrix} \\
&= \frac{1}{\rho}\begin{bmatrix} \text{Cov}\left(g_i, \phi_i\right) + \text{Cov}\left(e_i, \phi_i\right) \\ \text{Cov}\left(g_i\mathbf{z}_{int,i}, \phi_i\right) + \text{Cov}\left(e_i\mathbf{z}_{int,i}, \phi_i\right) \end{bmatrix} \\
&= \frac{1}{\rho}\begin{bmatrix} \text{Cov}\left(g_i, \phi_i\right) \\ \text{Cov}\left(g_i\mathbf{z}_{int,i}, \phi_i\right) \end{bmatrix} \\
&= \frac{1}{\rho}\begin{bmatrix} \text{Cov}\left(g_i, \phi_i\right) \\ \text{Cov}\left(\mathbf{w}_i, \phi_i\right) \end{bmatrix} \\
&= \frac{1}{\rho}\text{Cov}\left(\mathbf{G}_i, \phi_i\right).
\end{aligned}$$

Note that in the fourth line, we have used the assumption that e_i is independent of all other varables. The above calculation implies that

$$\begin{bmatrix} \text{Cov}\left(\hat{\mathbf{G}}_i, \phi_i\right) \\ \text{Cov}\left(\mathbf{z}_i, \phi_i\right) \end{bmatrix} = \mathbf{P}^{-1}\begin{bmatrix} \text{Cov}\left(\mathbf{G}_i, \phi_i\right) \\ \text{Cov}\left(\mathbf{z}_i, \phi_i\right) \end{bmatrix}, \tag{10}$$

where

$$\mathbf{P} = \begin{bmatrix} \rho\mathbf{I}_{|G|} & \mathbf{0} \\ & \mathbf{I}_{|z|} \end{bmatrix}.$$

Substituting (10) into (9), we get

$$\begin{aligned}
\alpha_{\hat{g}} &= \mathbf{V}_{\hat{g}}^{-1}\begin{bmatrix} \text{Cov}\left(\hat{\mathbf{G}}_i, \phi_i\right) \\ \text{Cov}\left(\mathbf{z}_i, \phi_i\right) \end{bmatrix} \\
&= \mathbf{V}_{\hat{g}}^{-1}\mathbf{P}^{-1}\begin{bmatrix} \text{Cov}\left(\mathbf{G}_i, \phi_i\right) \\ \text{Cov}\left(\mathbf{z}_i, \phi_i\right) \end{bmatrix} \\
&= \mathbf{V}_{\hat{g}}^{-1}\mathbf{P}^{-1}\mathbf{V}_g\mathbf{V}_g^{-1}\begin{bmatrix} \text{Cov}\left(\mathbf{G}_i, \phi_i\right) \\ \text{Cov}\left(\mathbf{z}_i, \phi_i\right) \end{bmatrix} \\
&= \mathbf{V}_{\hat{g}}^{-1}\mathbf{P}^{-1}\mathbf{V}_g\alpha_g.
\end{aligned} \tag{11}$$

We now need to derive the relationship between $\mathbf{V}_{\hat{g}}$ and $\mathbf{V}_g$. We begin by calculating

$$\begin{aligned}
\Sigma_{\hat{G}} &= \begin{bmatrix} \text{Var}\left(\hat{g}_i\right) & \text{Cov}\left(\hat{\mathbf{w}}_i, \hat{g}_i\right) \\ & \text{Var}\left(\hat{\mathbf{w}}_i\right) \end{bmatrix} \\
&= \begin{bmatrix} \text{Var}\left(\hat{g}_i\right) & \text{Cov}\left(\hat{g}_i\mathbf{z}_{\text{int},i}, \hat{g}_i\right) \\ & \text{Var}\left(\hat{g}_i\mathbf{z}_{\text{int},i}\right) \end{bmatrix} \\
&= \begin{bmatrix} \text{Var}\left(\frac{g_i+e_i}{\rho}\right) & \text{Cov}\left(\frac{g_i+e_i}{\rho}\mathbf{z}_{\text{int},i}, \frac{g_i+e_i}{\rho}\right) \\ & \text{Var}\left(\frac{g_i+e_i}{\rho}\mathbf{z}_{\text{int},i}\right) \end{bmatrix} \\
&= \left(\frac{1}{\rho}\mathbf{I}_{|G|}\right)\begin{bmatrix} \text{Var}\left(g_i\right) + \text{Var}\left(e_i\right) & \text{Cov}\left(g_i\mathbf{z}_{\text{int},i}, g_i\right) + \text{Cov}\left(e_i\mathbf{z}_{\text{int},i}, e_i\right) \\ & \text{Var}\left(g_i\mathbf{z}_{\text{int},i}\right) + \text{Var}\left(e_i\mathbf{z}_{\text{int},i}\right) \end{bmatrix}\left(\frac{1}{\rho}\mathbf{I}_{|G|}\right) \\
&= \left(\frac{1}{\rho}\mathbf{I}_{|G|}\right)\left(\mathbf{\Sigma}_G + \mathbf{\Omega}_G\right)\left(\frac{1}{\rho}\mathbf{I}_{|G|}\right),
\end{aligned}$$

where

$$\mathbf{\Omega}_G \equiv \begin{bmatrix} \text{Var}\left(e_i\right) & \text{Cov}\left(e_i \mathbf{z}_{\text{int},i}, e_i\right) \\ & \text{Var}\left(e_i \mathbf{z}_{\text{int},i}\right) \end{bmatrix}$$

is the component of the variance-covariance matrix of $\hat{\mathbf{G}}_i$ that is due to error before standardization. Notice that, in the fourth line above, we have again used the assumption of the independence of e_i. Defining

$$\mathbf{\Omega} \equiv \begin{bmatrix} \mathbf{\Omega}_G & \mathbf{0} \\ & \mathbf{0} \end{bmatrix},$$

we have

$$\mathbf{V}_{\hat{g}} = \mathbf{P}^{-1}\left(\mathbf{V}_g + \mathbf{\Omega}\right)\mathbf{P}^{-1}. \tag{12}$$

Finally, substituting (12) into (11) gives us an equation for how the coefficients from the feasible regression, $\alpha_{\hat{g}}$, are biased relative to the coefficients from the theoretical regression, α_g:

$$\begin{aligned} \alpha_{\hat{g}} &= \mathbf{V}_{\hat{g}}^{-1}\mathbf{P}^{-1}\mathbf{V}_g\alpha_g \\ &= \left[\mathbf{P}^{-1}\left(\mathbf{V}_g + \mathbf{\Omega}\right)\mathbf{P}^{-1}\right]^{-1}\mathbf{P}^{-1}\mathbf{V}_g\alpha_g \\ &= \mathbf{P}\left(\mathbf{V}_g + \mathbf{\Omega}\right)^{-1}\mathbf{V}_g\alpha_g. \end{aligned} \tag{13}$$

2.3 Estimator for Coefficients from the Theoretical Regression

Equation (13) can be rearranged to yield the simple regression-disattenuation estimator mentioned in the main text:

$$\alpha_{corr} \equiv \mathbf{V}_g^{-1}\left(\mathbf{V}_g + \mathbf{\Omega}\right)\mathbf{P}^{-1}\alpha_{\hat{g}}. \tag{14}$$

This estimator, however, is written in terms of $\mathbf{V}_g$ and $\mathbf{V}_g + \mathbf{\Omega}$, which are unobserved. To obtain the estimator we implement, we now derive expressions for $\mathbf{V}_g$ and $\mathbf{V}_g + \mathbf{\Omega}$ in terms of estimable quantities.

Beginning with $\mathbf{V}_g + \mathbf{\Omega}$, equation (12) gives us

$$\begin{aligned} \left(\mathbf{V}_g + \mathbf{\Omega}\right) &= \mathbf{P}\mathbf{P}^{-1}\left(\mathbf{V}_g + \mathbf{\Omega}\right)\mathbf{P}^{-1}\mathbf{P} \\ &= \mathbf{P}\mathbf{V}_{\hat{g}}\mathbf{P}. \end{aligned} \tag{15}$$

Now, turning to $\mathbf{V}_g$, we begin by deriving expressions for $\mathbf{\Sigma}_G$ and $\mathbf{\Sigma}_{G,z}$, which are quadrants of the matrix $\mathbf{V}_g$. We calculate

$$\begin{aligned} \mathbf{\Sigma}_G &= \begin{bmatrix} \text{Var}\left(g_i\right) & \text{Cov}\left(\mathbf{w}_i, g_i\right) \\ & \text{Var}\left(\mathbf{w}_i\right) \end{bmatrix} \\ &= \begin{bmatrix} \text{Var}\left(g_i\right) & \text{Cov}\left(\mathbf{w}_i, g_i\right) - \rho^2\text{Cov}\left(\hat{\mathbf{w}}_i, \hat{g}_i\right) + \rho^2\text{Cov}\left(\hat{\mathbf{w}}_i, \hat{g}_i\right) \\ & \text{Var}\left(\mathbf{w}_i\right) - \rho^2\text{Var}\left(\hat{\mathbf{w}}_i\right) + \rho^2\text{Var}\left(\hat{\mathbf{w}}_i\right) \end{bmatrix} \\ &= \begin{bmatrix} \text{Var}\left(g_i\right) & \text{Cov}\left(\mathbf{w}_i, g_i\right) - \rho^2\text{Cov}\left(\frac{g_i+e_i}{\rho}\mathbf{z}_{\text{int},i}, \frac{g_i+e_i}{\rho}\right) + \rho^2\text{Cov}\left(\hat{\mathbf{w}}_i, \hat{g}_i\right) \\ & \text{Var}\left(\mathbf{w}_i\right) - \rho^2\text{Var}\left(\frac{g_i+e_i}{\rho}\mathbf{z}_{\text{int},i}\right) + \rho^2\text{Var}\left(\hat{\mathbf{w}}_i\right) \end{bmatrix} \\ &= \begin{bmatrix} \text{Var}\left(g_i\right) & \text{Cov}\left(\mathbf{w}_i, g_i\right) - \text{Cov}\left(g_i\mathbf{z}_{\text{int},i}, g_i\right) - \text{Cov}\left(e_i\mathbf{z}_{\text{int},i}, e_i\right) + \rho^2\text{Cov}\left(\hat{\mathbf{w}}_i, \hat{g}_i\right) \\ & \text{Var}\left(\mathbf{w}_i\right) - \text{Var}\left(g_i\mathbf{z}_{\text{int},i}\right) - \text{Var}\left(e_i\mathbf{z}_{\text{int},i}\right) + \rho^2\text{Var}\left(\hat{\mathbf{w}}_i\right) \end{bmatrix} \\ &= \begin{bmatrix} \text{Var}\left(g_i\right) & \text{Cov}\left(\mathbf{w}_i, g_i\right) - \text{Cov}\left(\mathbf{w}_i, g_i\right) - \text{Var}\left(e_i\right)\mathbb{E}\left(\mathbf{z}_{\text{int},i}\right) + \rho^2\text{Cov}\left(\hat{\mathbf{w}}_i, \hat{g}_i\right) \\ & \text{Var}\left(\mathbf{w}_i\right) - \text{Var}\left(\mathbf{w}_i\right) - \text{Var}\left(e_i\right)\text{Var}\left(\mathbf{z}_{\text{int},i}\right) + \rho^2\text{Var}\left(\hat{\mathbf{w}}_i\right) \end{bmatrix} \\ &= \begin{bmatrix} 1 & \rho^2\text{Cov}\left(\hat{\mathbf{w}}_i, \hat{g}_i\right) - \left(\rho^2 - 1\right)\mathbb{E}\left(\mathbf{z}_{\text{int},i}\right) \\ & \rho^2\mathbf{\Sigma}_{\hat{w}} - \left(\rho^2 - 1\right)\mathbf{\Sigma}_{\text{int},z} \end{bmatrix}. \end{aligned} \tag{16}$$

Also,

$$\begin{aligned}
\mathbf{\Sigma}_{\hat{G},z} &= \begin{bmatrix} \text{Cov}\left(\hat{g}_i, \mathbf{z}_i\right) \\ \text{Cov}\left(\hat{\mathbf{w}}_i, \mathbf{z}_i\right) \end{bmatrix} \\
&= \begin{bmatrix} \text{Cov}\left(\hat{g}_i, \mathbf{z}_i\right) \\ \text{Cov}\left(\hat{g}_i \mathbf{z}_{\text{int},i}, \mathbf{z}_i\right) \end{bmatrix} \\
&= \begin{bmatrix} \text{Cov}\left(\frac{g_i+e_i}{\rho}, \mathbf{z}_i\right) \\ \text{Cov}\left(\frac{g_i+e_i}{\rho} \mathbf{z}_{\text{int},i}, \mathbf{z}_i\right) \end{bmatrix} \\
&= \frac{1}{\rho} \begin{bmatrix} \text{Cov}\left(g_i, \mathbf{z}_i\right) \\ \text{Cov}\left(g_i \mathbf{z}_{\text{int},i}, \mathbf{z}_i\right) \end{bmatrix} \\
&= \frac{1}{\rho} \begin{bmatrix} \text{Cov}\left(g_i, \mathbf{z}_i\right) \\ \text{Cov}\left(\mathbf{w}_i, \mathbf{z}_i\right) \end{bmatrix} \\
&= \frac{1}{\rho} \mathbf{\Sigma}_{G,z}.
\end{aligned}$$

Hence,

$$\mathbf{\Sigma}_{G,z} = \rho \mathbf{\Sigma}_{\hat{G},z}. \tag{17}$$

Equations (16) and (17) then give us an expression for $\mathbf{V}_g$ in terms of observables:

$$\begin{aligned}
\mathbf{V}_g &= \begin{bmatrix} \mathbf{\Sigma}_G & \rho \mathbf{\Sigma}_{\hat{G},z} \\ & \mathbf{\Sigma}_z \end{bmatrix} \\
&= \mathbf{P} \begin{bmatrix} \frac{1}{\rho^2} \mathbf{\Sigma}_G & \mathbf{\Sigma}_{\hat{G},z} \\ & \mathbf{\Sigma}_z \end{bmatrix} \mathbf{P},
\end{aligned} \tag{18}$$

where the estimable expression for $\mathbf{\Sigma}_G$ is given by(16) above.

Substituting these expressions for $\mathbf{V}_g$ and $\mathbf{V}_g + \mathbf{\Omega}$, equations (18) and (15), into the estimator, equation (14), gives us our estimator in terms of estimable quantities:

$$\begin{aligned}
\alpha_{\text{corr}} &\equiv \mathbf{V}_g^{-1} \left(\mathbf{V}_g + \mathbf{\Omega}\right) \mathbf{P}^{-1} \alpha_{\hat{g}} \\
&= \left(\mathbf{P} \begin{bmatrix} \frac{1}{\rho^2} \mathbf{\Sigma}_G & \mathbf{\Sigma}_{\hat{G},z} \\ & \mathbf{\Sigma}_z \end{bmatrix} \mathbf{P} \right)^{-1} \left(\mathbf{P} \mathbf{V}_{\hat{g}} \mathbf{P}\right) \mathbf{P}^{-1} \alpha_{\hat{g}} \\
&= \mathbf{P}^{-1} \begin{bmatrix} \frac{1}{\rho^2} \mathbf{\Sigma}_G & \mathbf{\Sigma}_{\hat{G},z} \\ & \mathbf{\Sigma}_z \end{bmatrix}^{-1} \mathbf{V}_{\hat{g}} \alpha_{\hat{g}} \\
&= \mathbf{C} \alpha_{\hat{g}},
\end{aligned} \tag{19}$$

where

$$\mathbf{C} \equiv \mathbf{P}^{-1} \begin{bmatrix} \frac{1}{\rho^2} \mathbf{\Sigma}_G & \mathbf{\Sigma}_{\hat{G},z} \\ & \mathbf{\Sigma}_z \end{bmatrix}^{-1} \mathbf{V}_{\hat{g}}. \tag{20}$$

2.4 Standard Errors

To obtain standard errors for our estimator, note that equation (19) implies

$$\text{Var}\left(\alpha_{\text{corr}}\right) = \mathbf{C} \mathbf{A}_{\hat{g}} \mathbf{C}'. \tag{21}$$

where $\mathbf{A}_{\hat{g}} \equiv \text{Var}\left(\alpha_{\hat{g}}\right)$. We calculate standard errors by taking the square root of the diagonal of this matrix.

2.5 Two Special Cases

To understand the intuition for the estimator, we consider two special cases. First consider a univariate regression of the phenotype on the PGI. In this case, $\mathbf{G}_i = [g_i]$ and $\mathbf{z}_i$ is empty. Thus

$$\begin{aligned}\alpha_{\text{corr}} &= \left[\frac{1}{\rho}\right]\left[\frac{1}{\rho^2}\right]^{-1}[1]\,\alpha_{\hat{g}} \\ &= \rho\alpha_{\hat{g}}.\end{aligned}$$

In this case, our estimator simply re-inflates the coefficient corresponding to the amount of attenuation due to error in the PGI.

Turning to the standard error, by (21), the sampling variance of the corrected estimate is

$$\begin{aligned}\text{Var}\left(\alpha_{\text{corr}}\right) &= \mathbf{C}\mathbf{A}_{\hat{g}}\mathbf{C}' \\ &= \rho\text{Var}\left(\alpha_{\hat{g}}\right)\rho \\ &= \rho^2\text{Var}\left(\alpha_{\hat{g}}\right).\end{aligned}$$

This means the standard error of the corrected estimate is

$$\text{s.e.}\left(\hat{\alpha}_{corr}\right) = \rho\ \text{s.e.}\left(\hat{\alpha}_{\hat{g}}\right).$$

Since the standard error is inflated by exactly the same factor ρ as the regression coefficient, the t-statistic and p-value of the measurement-error-corrected regression coefficient remains the same as without the measurement-error correction.

As a second special case, consider a multivariate regression with a single covariate that is independent of the PGI and an interaction between the covariate and the PGI. In this case, we have

$$\begin{aligned}\alpha_{\text{corr}} &= \begin{bmatrix}\frac{1}{\rho} & 0 & 0\\ & \frac{1}{\rho} & 0\\ & & 1\end{bmatrix}\begin{bmatrix}\frac{1}{\rho^2} & 0 & 0\\ & \frac{1}{\rho^2}\Sigma_z & 0\\ & & \Sigma_z\end{bmatrix}^{-1}\begin{bmatrix}1 & 0 & 0\\ & \Sigma_z & 0\\ & & \Sigma_z\end{bmatrix}\alpha_{\hat{g}} \\ &= \begin{bmatrix}\rho & 0 & 0\\ & \rho & 0\\ & & 1\end{bmatrix}\alpha_{\hat{g}},\end{aligned}$$

where the first element of $\alpha_{\hat{g}}$ is the coefficient associated with the PGI, the second element is the coefficient associated with the interaction, and the third element is the coefficient associated with the covariate. Similar to the univariate special case, the estimator in the simple, independent gene-by-environment interaction case with no other covariates simply inflates the coefficients corresponding to covariates related to the PGI. We conclude from this special that that as a first approximation, we should expect that the estimator will inflate each of the coefficients associated with the PGI or its interactions by ρ. The estimator will deviate from this benchmark to the extent that the PGI is correlated with the interacted environmental factor or any other covariates included in the model.

Using (21), we calculate the variance of the corrected estimates:

$$\begin{aligned}\text{Var}\left(\alpha_{\text{corr}}\right) &= \mathbf{C}\mathbf{A}_{\hat{g}}\mathbf{C}' \\ &= \begin{bmatrix}\rho & 0 & 0\\ & \rho & 0\\ & & 1\end{bmatrix}\text{Var}\left(\alpha_{\hat{g}}\right)\begin{bmatrix}\rho & 0 & 0\\ & \rho & 0\\ & & 1\end{bmatrix} \\ &= \begin{bmatrix}\rho^2\text{Var}\left(\alpha_{\hat{g},1}\right) & \rho^2\text{Cov}\left(\alpha_{\hat{g},1},\alpha_{\hat{g},2}\right) & \rho\text{Cov}\left(\alpha_{\hat{g},1},\alpha_{\hat{g},3}\right)\\ & \rho^2\text{Var}\left(\alpha_{\hat{g},2}\right) & \rho\text{Cov}\left(\alpha_{\hat{g},2},\alpha_{\hat{g},3}\right)\\ & & \text{Var}\left(\alpha_{\hat{g},3}\right)\end{bmatrix},\end{aligned}$$

where $\alpha_{\hat{g},1}$, $\alpha_{\hat{g},2}$, and $\alpha_{\hat{g},3}$ are the three elements of $\alpha_{\hat{g}}$. The standard errors are the square root of the

diagonal of this matrix, giving us

$$\text{s.e.}\left(\hat{\alpha}_{corr}\right) = \begin{bmatrix} \sqrt{\rho^2 \text{Var}\left(\hat{\alpha}_{\hat{g},1}\right)} \\ \sqrt{\rho^2 \text{Var}\left(\hat{\alpha}_{\hat{g},2}\right)} \\ \sqrt{\text{Var}\left(\hat{\alpha}_{\hat{g},3}\right)} \end{bmatrix} = \begin{bmatrix} \rho & 0 & 0 \\ & \rho & 0 \\ & & 1 \end{bmatrix} \text{Var}\left(\hat{\alpha}_{\hat{g}}\right).$$

Thus, each of the standard errors of the corrected estimates is inflated by exactly the same proportion as the inflation of its corresponding corrected estimates. Therefore, in this special case, the t-statistics and p-values of all three measurement-error-corrected regression coefficients remain the same as without the measurement-error correction.

3 Implementation of the Estimator

In the derivation above, we have expressed everything in terms of population parameters. In order to obtain a consistent estimator of α_{corr} and its standard error, we must write them in terms of the data that we observe.

First, consider the parameter ρ. Our estimator $\hat{\rho}$ is

$$\hat{\rho} \equiv \sqrt{\frac{\hat{h}^2_{SNP}}{\hat{R}^2}}.$$

The value $\hat{h}^2_{SNP}$ is an estimate of SNP heritability of the phenotype y_i in the prediction dataset, based on the same set of J SNPs that make up the PGI. The value $\hat{R}^2$ is the estimated predictive power of the PGI for y_i in the prediction dataset.

Note that $\hat{h}^2_{SNP}$ and $\hat{R}^2$ each correspond to the PGI phenotype y_i rather than the (possibly different) phenotype in the regression analysis ϕ_i. If the phenotype y_i is not available in the prediction dataset or if the sample is too small to obtain reliable estimates, $\hat{h}^2_{SNP}$ and $\hat{R}^2$, $\hat{\rho}$ could instead be estimated from a different sample without introducing any bias as long as the genetic correlation of y_i is perfect between the two samples. (The heritability may differ in the samples, but the genetic correlation must be one. This may happen if the individuals are drawn from the same population in the two samples, but the phenotype is measured with greater error in one of them.) Given that a researcher would choose to do this only in the absence of enough data on y_i in the regression sample, perfect genetic correlation cannot be reliably tested and would therefore become an important assumption underlying use of the correction.

Turning to the other parameters besides ρ, in all cases we replace the population variance-covariance matrices with the consistent (sample-analog) estimates of these matrices. For example,

$$\hat{\mathbf{\Sigma}}_z \equiv \frac{1}{N}\mathbf{Z}'\mathbf{Z},$$

were N is the sample size in the regression sample and $\mathbf{Z}$ is the $N \times |z_i|$ matrix of covariates in the regression.

Since $\hat{\rho}$ and each variance-covariance matrix is a consistent estimator of its population counterpart, $\hat{\alpha}_{\text{corr}}$ is a consistent estimator of α_{corr}.

4 Assumption That e_i is Uncorrelated With Other Variables

Recall from equation (4), we have expressed the PGI as

$$\hat{g}_i = \frac{g_i + e_i}{\text{sd}\left(g_i + e_i\right)},$$

where, by construction, $\text{Cov}\left(g_i, e_i\right) = 0$. We assumed that e_i is uncorrelated with all other covariates in the model. In this section, we show that if the SNP weights for the PGI are unbiased estimates of the

SNP weights for the additive SNP factor, then this uncorrelatedness assumption is exactly true. We then show that when SNP weights for the PGI are estimated using LDpred-inf—which is the method we use for the Repository and which does not generate unbiased estimates—given typical parameter values for PGIs in the Repository, the bias in our measurement-error-corrected estimator due to the violation of the uncorrelatedness assumption is negligible.

4.1 Uncorrelatedness Is Implied When Unbiased Estimates of γ_j Are Used

A sufficient condition for our $\text{Cov}(g_i, e_i) = 0$ assumption to hold is that the SNP weights for the PGI are unbiased estimates of the SNP weights for the additive SNP factor. To state it more formally, recall that the standardized additive SNP factor is

$$g_i = \frac{\mathbf{x}_i\gamma}{\text{sd}(\mathbf{x}_i\gamma)},$$

and the PGI is

$$\hat{g}_i = \frac{\mathbf{x}_i\hat{\gamma}}{\text{sd}(\mathbf{x}_i\hat{\gamma})}.$$

The sufficient condition is that $\hat{\gamma}$ is an unbiased estimate of γ. This would be the case, for example, if $\hat{\gamma}$ is estimated by ordinary least squares or logistic regression (rather than a Bayesian approach, such as LDpred, which tends to shrink coefficient estimates relative to those from ordinary least squares). This is roughly equivalent to what is done when PGIs are constructed using "Pruning and Thresholding" methods as long as the PGI weights are estimated in a different sample than the sample used to select the SNPs that are included in the PGI. (Note, however, that because "Pruning and Thresholding" methods construct a PGI using fewer SNPs, the resulting PGI is proxying for an additive SNP factor that is based on fewer SNPs and hence has a lower h^2_{SNP}.) Because $\hat{\gamma}$ is unbiased,

$$\hat{\gamma} = \gamma + \mathbf{e}_\gamma,$$

where $\mathbb{E}(\mathbf{e}_\gamma) = \mathbf{0}$. Since $\mathbf{e}_\gamma$ is sampling error, $\mathbf{e}_\gamma$ is independent of all variables in independent samples. Therefore, the measurement error in the PGI, $e_i = \mathbf{x}_i\mathbf{e}_\gamma$, is also independent of all variables in independent samples.

4.2 Magnitude of the Bias When γ_j Is Estimated Using LDpred-inf

For the Repository, we construct the PGI weights using LDpred-inf. To be precise about this method, it is helpful to express the length-N vector of phenotype values for the N individuals in the discovery sample as

$$\mathbf{Y} = \mathbf{X}\gamma + \mathbf{E},$$

where $\mathbf{X}$ is the $N \times K$ matrix of K genotypes, and $\mathbf{E}$ is a vector of residuals, which is uncorrelated with $\mathbf{X}$. The LDpred-inf estimator for $\hat{\gamma}$ is:

$$\hat{\gamma} = \left(\mathbf{X}'\mathbf{X} + \frac{1}{\sigma^2_\gamma}\mathbf{I}\right)^{-1}\mathbf{X}'\mathbf{Y},$$

where $\sigma^2_\gamma \equiv \text{Var}(\gamma) = \frac{h^2_{SNP}}{K}$ is the prior variance of the additive SNP factor weights. This is equivalent to the ridge regression estimator, with a particular choice of the regularization parameter. This estimator reduces the problem of multicollinearity (due to LD) at the cost of some bias in the estimates. We can calculate the

relationship between $\hat{\gamma}$ and γ as

$$\begin{aligned}
\hat{\gamma} &= \left(\mathbf{X}'\mathbf{X} + \frac{1}{\sigma_\gamma^2}\mathbf{I}\right)^{-1}\mathbf{X}'\mathbf{Y} \\
&= \left(\mathbf{X}'\mathbf{X} + \frac{1}{\sigma_\gamma^2}\mathbf{I}\right)^{-1}\mathbf{X}'\left(\mathbf{X}\gamma + \mathbf{E}\right) \\
&= \left(N\hat{\Sigma}_X + \frac{1}{\sigma_\gamma^2}\mathbf{I}\right)^{-1}\left(N\hat{\Sigma}_X\right)\gamma + e_{\hat{\gamma}} \\
&= \left(\hat{\Sigma}_X + \frac{1}{N\sigma_\gamma^2}\mathbf{I}\right)^{-1}\hat{\Sigma}_X\gamma + e_{\hat{\gamma}}, \qquad (22)
\end{aligned}$$

where $\hat{\Sigma}_X \equiv \frac{1}{N}\mathbf{X}'\mathbf{X}$ is the sample variance-covariance matrix of $\mathbf{X}$ and $e_{\hat{\gamma}} \equiv \left(\mathbf{X}'\mathbf{X} + \frac{1}{\sigma_\gamma^2}\mathbf{I}\right)^{-1}\mathbf{X}'\mathbf{E}$ is the estimation error of $\hat{\gamma}$.

To evaluate the magnitude of the bias in finite samples, we quantify it in a simple case where we regress the phenotype ϕ_i on the standardized additive SNP factor g_i and a single (scalar) covariate z_i. Without loss of generality, we orient z_i such that g_i and z_i have positive covariance. As in the main text, we use α_g to denote the coefficients of this theoretical regression:

$$\phi_i = \left[\begin{array}{cc} g_i & z_i \end{array}\right]\alpha_g + \varepsilon_i.$$

The coefficients from the feasible regression are

$$\alpha_{\hat{g}} = \left[\begin{array}{cc} \mathrm{Var}\left(\hat{g}_i\right) & \mathrm{Cov}\left(\hat{g}_i, z_i\right) \\ & \mathrm{Var}\left(z_i\right) \end{array}\right]^{-1}\left[\begin{array}{c} \mathrm{Cov}\left(\hat{g}_i, \phi_i\right) \\ \mathrm{Cov}\left(z_i, \phi_i\right) \end{array}\right].$$

Our measurement-error-corrected estimator is

$$\begin{aligned}
\alpha_{\mathrm{corr}} &= \left[\begin{array}{cc} \frac{1}{\rho} & 0 \\ & 1 \end{array}\right]\left[\begin{array}{cc} \frac{1}{\rho^2} & \mathrm{Cov}\left(\hat{g}_i, z_i\right) \\ & \mathrm{Var}\left(z_i\right) \end{array}\right]^{-1}\left[\begin{array}{cc} \mathrm{Var}\left(\hat{g}_i\right) & \mathrm{Cov}\left(\hat{g}_i, z_i\right) \\ & \mathrm{Var}\left(z_i\right) \end{array}\right]\alpha_{\hat{g}} \\
&= \left[\begin{array}{cc} \frac{1}{\rho} & 0 \\ & 1 \end{array}\right]\left[\begin{array}{cc} \frac{1}{\rho^2} & \mathrm{Cov}\left(\hat{g}_i, z_i\right) \\ & \mathrm{Var}\left(z_i\right) \end{array}\right]^{-1}\left[\begin{array}{c} \mathrm{Cov}\left(\hat{g}_i, \phi_i\right) \\ \mathrm{Cov}\left(z_i, \phi_i\right) \end{array}\right] \\
&= \left[\begin{array}{cc} \frac{1}{\rho} & 0 \\ & 1 \end{array}\right]\left[\begin{array}{cc} \frac{1}{\rho^2} & \mathrm{Cov}\left(\hat{g}_i, z_i\right) \\ & \mathrm{Var}\left(z_i\right) \end{array}\right]^{-1}\left[\begin{array}{c} \mathrm{Cov}\left(\hat{g}_i, \left[\begin{array}{cc} g_i & z_i \end{array}\right]\alpha + \varepsilon_i\right) \\ \mathrm{Cov}\left(z_i, \left[\begin{array}{cc} g_i & z_i \end{array}\right]\alpha + \varepsilon_i\right) \end{array}\right] \\
&= \left[\begin{array}{cc} \frac{1}{\rho} & 0 \\ & 1 \end{array}\right]\left[\begin{array}{cc} \frac{1}{\rho^2} & \mathrm{Cov}\left(\hat{g}_i, z_i\right) \\ & \mathrm{Var}\left(z_i\right) \end{array}\right]^{-1}\left[\begin{array}{cc} \mathrm{Cov}\left(\hat{g}_i, g_i\right) & \mathrm{Cov}\left(\hat{g}_i, z_i\right) \\ \mathrm{Cov}\left(g_i, z_i\right) & \mathrm{Var}\left(z_i\right) \end{array}\right]\alpha \\
&= \frac{1}{\frac{\mathrm{Var}(z_i)}{\rho^2} - \mathrm{Cov}\left(\hat{g}_i, z_i\right)^2}\left[\begin{array}{cc} \frac{1}{\rho} & 0 \\ & 1 \end{array}\right]\left[\begin{array}{cc} \mathrm{Var}\left(z_i\right) & -\mathrm{Cov}\left(\hat{g}_i, z_i\right) \\ & \frac{1}{\rho^2} \end{array}\right]\left[\begin{array}{cc} \mathrm{Cov}\left(\hat{g}_i, g_i\right) & \mathrm{Cov}\left(\hat{g}_i, z_i\right) \\ \mathrm{Cov}\left(g_i, z_i\right) & \mathrm{Var}\left(z_i\right) \end{array}\right]\alpha \\
&= \frac{1}{\frac{\mathrm{Var}(z_i)}{\rho^2} - \mathrm{Cov}\left(\hat{g}_i, z_i\right)^2}\left[\begin{array}{cc} \frac{1}{\rho}\mathrm{Var}\left(z_i\right) & -\frac{1}{\rho}\mathrm{Cov}\left(\hat{g}_i, z_i\right) \\ -\mathrm{Cov}\left(\hat{g}_i, z_i\right) & \frac{1}{\rho^2} \end{array}\right]\left[\begin{array}{cc} \mathrm{Cov}\left(\hat{g}_i, g_i\right) & \mathrm{Cov}\left(\hat{g}_i, z_i\right) \\ \mathrm{Cov}\left(g_i, z_i\right) & \mathrm{Var}\left(z_i\right) \end{array}\right]\alpha \\
&= \left[\begin{array}{cc} \frac{\frac{1}{\rho}\mathrm{Var}(z_i)\mathrm{Cov}(\hat{g}_i,g_i) - \frac{1}{\rho}\mathrm{Cov}(g_i,z_i)\mathrm{Cov}(\hat{g}_i,z_i)}{\frac{\mathrm{Var}(z_i)}{\rho^2} - \mathrm{Cov}(\hat{g}_i,z_i)^2} & 0 \\ \frac{\frac{1}{\rho^2}\mathrm{Cov}(g_i,z_i) - \mathrm{Cov}(\hat{g}_i,z_i)\mathrm{Cov}(\hat{g}_i,g_i)}{\frac{\mathrm{Var}(z_i)}{\rho^2} - \mathrm{Cov}(\hat{g}_i,z_i)^2} & 1 \end{array}\right]\alpha \\
&= \left[\begin{array}{cc} \frac{\frac{1}{\rho}\mathrm{Var}(z_i)\mathrm{Cov}(\hat{g}_i,g_i) - \frac{1}{\rho}\mathrm{Cov}(g_i,z_i)\mathrm{Cov}(\hat{g}_i,z_i)}{\frac{\mathrm{Var}(z_i)}{\rho^2} - \mathrm{Cov}(\hat{g}_i,z_i)^2} & 0 \\ \frac{\frac{1}{\rho^2}\mathrm{Cov}(g_i,z_i) - \mathrm{Cov}(\hat{g}_i,z_i)\mathrm{Cov}(\hat{g}_i,g_i)}{\frac{\mathrm{Var}(z_i)}{\rho^2} - \mathrm{Cov}(\hat{g}_i,z_i)^2} & 1 \end{array}\right]\alpha.
\end{aligned}$$

Using $\text{Cov}\left(\hat{g}_i, g_i\right) = \text{Cov}\left(\frac{g_i+e_i}{\rho}, g_i\right) = \frac{1}{\rho}$, we have

$$\begin{aligned}
\alpha_{\text{corr}} &= \begin{bmatrix} \frac{\frac{1}{\rho^2}\text{Var}(z_i) - \frac{1}{\rho}\text{Cov}(g_i,z_i)\text{Cov}(\hat{g}_i,z_i)}{\frac{\text{Var}(z_i)}{\rho^2} - \text{Cov}(\hat{g}_i,z_i)^2} & 0 \\ \frac{\frac{1}{\rho^2}\text{Cov}(g_i,z_i) - \frac{1}{\rho}\text{Cov}(\hat{g}_i,z_i)}{\frac{\text{Var}(z_i)}{\rho^2} - \text{Cov}(\hat{g}_i,z_i)^2} & 1 \end{bmatrix} \alpha \\
&= \begin{bmatrix} \frac{\frac{1}{\rho^2}\text{Var}(z_i) - \text{Cov}(\hat{g}_i,z_i)^2 + \text{Cov}(\hat{g}_i,z_i)^2 - \frac{1}{\rho}\text{Cov}(g_i,z_i)\text{Cov}(\hat{g}_i,z_i)}{\frac{\text{Var}(z_i)}{\rho^2} - \text{Cov}(\hat{g}_i,z_i)^2} & 0 \\ \frac{\frac{1}{\rho^2}\text{Cov}(g_i,z_i) - \frac{1}{\rho}\text{Cov}(\hat{g}_i,z_i)}{\frac{\text{Var}(z_i)}{\rho^2} - \text{Cov}(\hat{g}_i,z_i)^2} & 1 \end{bmatrix} \alpha \\
&= \begin{bmatrix} 1 + \frac{\text{Cov}(\hat{g}_i,z_i)\left[\text{Cov}(\hat{g}_i,z_i) - \frac{1}{\rho}\text{Cov}(g_i,z_i)\right]}{\frac{\text{Var}(z_i)}{\rho^2} - \text{Cov}(\hat{g}_i,z_i)^2} & 0 \\ \frac{-\frac{1}{\rho}\left[\text{Cov}(\hat{g}_i,z_i) - \frac{1}{\rho}\text{Cov}(g_i,z_i)\right]}{\frac{\text{Var}(z_i)}{\rho^2} - \text{Cov}(\hat{g}_i,z_i)^2} & 1 \end{bmatrix} \alpha.
\end{aligned}$$

This means that the bias is

$$\begin{aligned}
\alpha_{\text{corr}} - \alpha &= \begin{bmatrix} 1 + \frac{\text{Cov}(\hat{g}_i,z_i)\left[\text{Cov}(\hat{g}_i,z_i) - \frac{1}{\rho}\text{Cov}(g_i,z_i)\right]}{\frac{\text{Var}(z_i)}{\rho^2} - \text{Cov}(\hat{g}_i,z_i)^2} & 0 \\ \frac{-\frac{1}{\rho}\left[\text{Cov}(\hat{g}_i,z_i) - \frac{1}{\rho}\text{Cov}(g_i,z_i)\right]}{\frac{\text{Var}(z_i)}{\rho^2} - \text{Cov}(\hat{g}_i,z_i)^2} & 1 \end{bmatrix} \alpha - \alpha \\
&= \begin{bmatrix} \frac{\text{Cov}(\hat{g}_i,z_i)\left[\text{Cov}(\hat{g}_i,z_i) - \frac{1}{\rho}\text{Cov}(g_i,z_i)\right]}{\frac{\text{Var}(z_i)}{\rho^2} - \text{Cov}(\hat{g}_i,z_i)^2} & 0 \\ \frac{-\frac{1}{\rho}\left[\text{Cov}(\hat{g}_i,z_i) - \frac{1}{\rho}\text{Cov}(g_i,z_i)\right]}{\frac{\text{Var}(z_i)}{\rho^2} - \text{Cov}(\hat{g}_i,z_i)^2} & 0 \end{bmatrix} \alpha \\
&= \left[\text{Cov}\left(\hat{g}_i, z_i\right) - \frac{1}{\rho}\text{Cov}\left(g_i, z_i\right)\right] \begin{bmatrix} \frac{\text{Cov}(\hat{g}_i,z_i)}{\frac{\text{Var}(z_i)}{\rho^2} - \text{Cov}(\hat{g}_i,z_i)^2} \\ -\frac{1}{\rho} \\ \frac{\text{Var}(z_i)}{\rho^2} - \text{Cov}(\hat{g}_i,z_i)^2 \end{bmatrix} \alpha_1, \qquad (23)
\end{aligned}$$

where α_1 is the first element of α.

We next express the first factor, $\left[\text{Cov}\left(\hat{g}_i, z_i\right) - \frac{1}{\rho}\text{Cov}\left(g_i, z_i\right)\right]$, in terms of observable or estimable quantities. To do this, consider the best linear predictor of z_i using the same SNPs that make up g_i. That predictor would have weights

$$\xi \equiv \arg\min_{\tilde{\xi}} \mathbb{E}\left[\left(y_i - \mathbf{x}_i\tilde{\xi}\right)^2\right]$$

(so $\mathbf{x}_i\xi$ is the additive SNP factor for z_i). Therefore,

$$\begin{aligned}
\left[\text{Cov}\left(\hat{g}_i, z_i\right) - \frac{1}{\rho}\text{Cov}\left(g_i, z_i\right)\right] &= \left[\text{Cov}\left(\frac{x_i\hat{\gamma}}{\rho}, x_i\xi\right) - \frac{1}{\rho}\text{Cov}\left(x_i\gamma, x_i\xi\right)\right] \\
&= \frac{1}{\rho}\left[\text{Cov}\left(x_i\left[\left(\hat{\Sigma}_X + \frac{1}{N\sigma_\gamma^2}\mathbf{I}\right)^{-1}\hat{\Sigma}_X\gamma + e_{\hat{\gamma}}\right], x_i\xi\right) - \text{Cov}\left(x_i\gamma, x_i\xi\right)\right] \\
&= \frac{1}{\rho}\left[\text{Cov}\left(x_i\left(\hat{\Sigma}_X + \frac{1}{N\sigma_\gamma^2}\mathbf{I}\right)^{-1}\hat{\Sigma}_X\gamma, x_i\xi\right) - \text{Cov}\left(x_i\gamma, x_i\xi\right)\right] \\
&= \frac{1}{\rho}\left[\gamma'\hat{\Sigma}_X\left(\hat{\Sigma}_X + \frac{1}{N\sigma_\gamma^2}\mathbf{I}\right)^{-1}\text{Var}\left(x_i\right) - \gamma'\text{Var}\left(x_i\right)\right]\xi \\
&= \frac{1}{\rho}\gamma'\left[\hat{\Sigma}_X\left(\hat{\Sigma}_X + \frac{1}{N\sigma_\gamma^2}\mathbf{I}\right)^{-1}\Sigma_X - \Sigma_X\right]\xi.
\end{aligned}$$

The second line follows from substituting equation (22) into the first line. By Woodbury's Identity, we have

$$\begin{aligned}\left[\text{Cov}\left(\hat{g}_i, z_i\right) - \frac{1}{\rho}\text{Cov}\left(g_i, z_i\right)\right] &= \frac{1}{\rho}\gamma'\left[\hat{\Sigma}_X\left(\hat{\Sigma}_X^{-1} - \hat{\Sigma}_X^{-1}\left(\hat{\Sigma}_X^{-1} + n\sigma_\gamma^2\mathbf{I}\right)^{-1}\hat{\Sigma}_X^{-1}\right)\Sigma_X - \Sigma_X\right]\xi \\ &= \frac{1}{\rho}\gamma'\left[\hat{\Sigma}_X\hat{\Sigma}_X^{-1}\Sigma_X - \hat{\Sigma}_X\hat{\Sigma}_X^{-1}\left(\hat{\Sigma}_X^{-1} + n\sigma_\gamma^2\mathbf{I}\right)^{-1}\hat{\Sigma}_X^{-1}\Sigma_X - \Sigma_X\right]\xi \\ &= \frac{1}{\rho}\gamma'\left[\Sigma_X - \left(\hat{\Sigma}_X^{-1} + n\sigma_\gamma^2\mathbf{I}\right)^{-1}\hat{\Sigma}_X^{-1}\Sigma_X - \Sigma_X\right]\xi \\ &= -\frac{1}{\rho}\gamma'\left(\hat{\Sigma}_X^{-1} + n\sigma_\gamma^2\mathbf{I}\right)^{-1}\hat{\Sigma}_X^{-1}\Sigma_X\xi. \end{aligned} \tag{24}$$

Next imagine a weighted regression of ξ onto γ with weights $\left(\Sigma_X^{-1} + n\sigma_\gamma^2\mathbf{I}\right)^{-1}$. This produces

$$\xi = \frac{\sigma_{\xi\gamma}}{\sigma_\gamma^2}\gamma + \mu,$$

with $\sigma_{\gamma\xi} \equiv \text{Cov}\left(\gamma, \xi\right)$ and with the residual μ having the property $\gamma'\left(\Sigma_X^{-1} + n\sigma_\gamma^2\mathbf{I}\right)^{-1}\hat{\Sigma}_X^{-1}\Sigma_X\mu = 0$. Since we have oriented g_i and z_i to have positive covariance, $\sigma_{\gamma\xi} > 0$. Substituting both of these equations into (24) gives us

$$\begin{aligned}\left[\text{Cov}\left(\hat{g}_i, z_i\right) - \frac{1}{\rho}\text{Cov}\left(g_i, z_i\right)\right] &= -\frac{1}{\rho}\gamma'\left(\hat{\Sigma}_X^{-1} + n\sigma_\gamma^2\mathbf{I}\right)^{-1}\hat{\Sigma}_X^{-1}\Sigma_X\left(\frac{\sigma_{\xi\gamma}}{\sigma_\gamma^2}\gamma + \mu\right) \\ &= -\frac{1}{\rho}\left[\frac{\sigma_{\xi\gamma}}{\sigma_\gamma^2}\gamma'\left(\Sigma_X^{-1} + n\sigma_\gamma^2\mathbf{I}\right)^{-1}\hat{\Sigma}_X^{-1}\Sigma_X\gamma + \gamma'\left(\Sigma_X^{-1} + n\sigma_\gamma^2\mathbf{I}\right)^{-1}\hat{\Sigma}_X^{-1}\Sigma_X\mu\right] \\ &= -\frac{1}{\rho}\frac{\sigma_{\xi\gamma}}{\sigma_\gamma^2}\gamma'\left(\Sigma_X^{-1} + n\sigma_\gamma^2\mathbf{I}\right)^{-1}\hat{\Sigma}_X^{-1}\Sigma_X\gamma. \end{aligned}$$

Next, in order to put an upper bound on the magnitude of bias, we will show that $\gamma'\left(\Sigma_X^{-1} + n\sigma_\gamma^2\mathbf{I}\right)^{-1}\hat{\Sigma}_X^{-1}\Sigma_X\gamma < \frac{1}{n\sigma_\gamma^2}\gamma'\gamma$. Again using Woodbury's Identity, we have

$$\begin{aligned}\gamma'\left(\Sigma_X^{-1} + n\sigma_\gamma^2\mathbf{I}\right)^{-1}\hat{\Sigma}_X^{-1}\Sigma_X\gamma - \frac{1}{n\sigma_\gamma^2}\gamma'\gamma &= \gamma'\left[\left(\Sigma_X^{-1} + n\sigma_\gamma^2\mathbf{I}\right)^{-1}\hat{\Sigma}_X^{-1}\Sigma_X - \frac{1}{n\sigma_\gamma^2}\mathbf{I}\right]\gamma \\ &= \gamma'\left[\frac{1}{n\sigma_\gamma^2}\mathbf{I} - \left(\frac{1}{n\sigma_\gamma^2}\mathbf{I}\right)\left(\Sigma_X + \frac{1}{n\sigma_\gamma^2}\mathbf{I}\right)^{-1}\left(\frac{1}{n\sigma_\gamma^2}\mathbf{I}\right)\hat{\Sigma}_X^{-1}\Sigma_X - \frac{1}{n\sigma_\gamma^2}\mathbf{I}\right]\gamma \\ &= -\frac{1}{n^2\sigma_\gamma^4}\gamma'\left(\Sigma_X + \frac{1}{n\sigma_\gamma^2}\mathbf{I}\right)^{-1}\hat{\Sigma}_X^{-1}\Sigma_X\gamma \\ &< 0. \end{aligned}$$

The last step here follows because Σ_X and $\frac{1}{n\sigma_\gamma^2}\mathbf{I}$ are positive definite matrices. So this implies that

$\gamma'\left(\Sigma_X^{-1}+n\sigma_\gamma^2\mathbf{I}\right)^{-1}\hat{\Sigma}_X^{-1}\Sigma_X\gamma < \frac{1}{n\sigma_\gamma^2}\gamma'\gamma$. Since both sides of this inequality are positive, this implies that

$$\begin{aligned}\left[\text{Cov}\left(\hat{g}_i,z_i\right)-\frac{1}{\rho}\text{Cov}\left(g_i,z_i\right)\right] &= -\frac{1}{\rho}\frac{\sigma_{\xi\gamma}}{\sigma_\gamma^2}\gamma'\left(\Sigma_X^{-1}+n\sigma_\gamma^2\mathbf{I}\right)^{-1}\hat{\Sigma}_X^{-1}\Sigma_X\gamma \\ &> -\frac{1}{\rho}\frac{\sigma_{\xi\gamma}}{\sigma_\gamma^2}\frac{1}{n\sigma_\gamma^2}\gamma'\gamma \\ &= -\frac{1}{\rho}\frac{\sigma_{\xi\gamma}}{\sigma_\gamma^2}\frac{1}{n\sigma_\gamma^2}M_e\sigma_\gamma^2 \\ &= -\frac{1}{\rho}\frac{\sigma_{\xi\gamma}}{\sigma_\gamma^2}\frac{M_e}{n} \\ &= -\frac{1}{\rho}r_{\xi\gamma}\frac{\sigma_\xi}{\sigma_\gamma}\frac{M_e}{n} \\ &= -\frac{1}{\rho}r_{\xi\gamma}\sqrt{\frac{h_z^2}{h_{SNP}^2}}\frac{M_e}{n}.\end{aligned}$$

where M_e is the effective population size, $r_{\xi\gamma}$ is the correlation of ξ and γ, and h_z^2 is the SNP heritability of z_i. The parameters $r_{\xi\gamma}$ and h_z^2 are unknown, but the magnitude of this term will be largest when $r_{\xi\gamma}=h_z^2=1$. So we have

$$\left[\text{Cov}\left(\hat{g}_i,z_i\right)-\frac{1}{\rho}\text{Cov}\left(g_i,z_i\right)\right] > -\frac{1}{\rho}\frac{M_e}{nh_{SNP}}. \tag{25}$$

Substituting (25) into (23) therefore gives us an upper bound on the magnitude of the bias of the corrected estimates:

$$b_{upper} = -\frac{1}{\rho}\frac{1}{nh_{SNP}}\begin{bmatrix} \frac{\text{Cov}(\hat{g}_i,z_i)}{\frac{\text{Var}(z_i)}{\rho^2}-\text{Cov}(\hat{g}_i,z_i)^2} \\ -\frac{1}{\rho} \\ \frac{\text{Var}(z_i)}{\rho^2}-\text{Cov}(\hat{g}_i,z_i)^2 \end{bmatrix}\alpha_1. \tag{26}$$

Each of the values in this expression is observed or estimable. This means we can replace each of these parameters with their corresponding estimates to approximate the magnitude of the bias. Calculations using equation (26) imply that when Repository PGIs are used (for which the weights are calculated using LDpred-inf), the bias due to the violation of the $\text{Cov}\left(e_i,z_i\right)=0$ assumption will typically be small.

For example, using values from the Papageorge and Thom application used in this paper, if z_i represents mother's educational attainment, we estimate $\hat{\rho}=1.51$, $\hat{h}_{SNP}^2=0.25$, $n=293,723$, $\text{Var}\left(z_i\right)=9.02$, $\text{Cov}\left(\hat{g}_i,z_i\right)=0.53$ and $\hat{\alpha}_1=1.16$. (Note that this value of α_1 is actually based on controlling for mother and father's education, but since this exercise is just meant to get a approximation of the order or magnitude of the bias, we have not re-evaluated the model with only one covariate.) Substituting these values into (26) gives us

$$b_{upper} = \begin{bmatrix} -6.50\times10^{-7} \\ 8.13\times10^{-7} \end{bmatrix}.$$

This is several orders of magnitude smaller than the measurement-error correction.

5 Potential Bias in the Standard Errors

The standard errors from (21) ignore the estimation error introduced by using $\hat{\rho}$ rather than ρ. We argue here that ignoring this source of uncertainty induces little bias to our standard errors for $\hat{\alpha}_{\text{corr}}$ if ρ is estimated in the same sample as $\alpha_{\hat{g}}$.

Consider the univariate case: regressing the phenotype ϕ_i on only the PGI $\hat{g}_i$. In that case,

$$\begin{aligned}\hat{\alpha}_{\text{corr}} &= \hat{\rho}\hat{\alpha}_{\hat{g}} \\ &= \frac{\hat{h}_{SNP}}{\hat{R}}\hat{\alpha}_{\hat{g}}.\end{aligned}$$

Note that $\hat{R}$ and $\hat{\alpha}_{\hat{g}}$ correspond to how well the PGI predicts y_i and ϕ_i, respectively, in sample. For this reason, the error in $\hat{R}$ and $\hat{\alpha}_{\hat{g}}$ will be positively correlated, which will reduce the standard error of $\hat{\alpha}_{\text{corr}}$. In contrast, the error in $\hat{h}_{SNP}$ will also be correlated with the error in $\hat{\alpha}_{\hat{g}}$, which will increase the standard error. In the further simple setting where $\phi_i = y_i$, note that $\hat{\alpha}_{\hat{g}} = \hat{R}$, which implies that

$$\begin{aligned}\hat{\alpha}_{\text{corr}} &= \hat{\rho}\hat{\alpha}_{\hat{g}} \\ &= \frac{\hat{h}_{SNP}}{\hat{R}}\hat{R} \\ &= \hat{h}_{SNP}.\end{aligned}$$

So the true standard error is equal to the standard error of $\hat{h}_{SNP}$.

If $\hat{\rho}$ is calculated in a different dataset than the dataset used in the regression, the error in $\hat{h}_{SNP}$ and $\hat{R}$ will be uncorrelated with the error in $\hat{\alpha}_{\hat{g}}$. This means that the standard errors reported by our measurement-error software will be anti-conservative. However, since the error in $\hat{h}_{SNP}$ and $\hat{R}$ will be positively correlated, the sampling variance in $\hat{\rho}$ will likely be small, suggesting that the bias in the standard errors will also likely be small relative to the magnitude of the reported standard error.

6 Theoretical Framework with GWAS Controls

The theoretical framework in the main text is derived for PGI weights estimated in a GWAS conducted using ordinary least squares (OLS), without any control variables. In practice, PGI weights are virtually always derived from a GWAS that includes at least some basic set of control variables (typically sex, age, and at least four principal components (PCs) of the genotype data). We omit the covariates in the main text because doing so simplifies the exposition without altering any of the theoretical properties of the true additive SNP factor that we focus on in the main text. However, the choice of covariates is one of many dimensions of GWAS methodology that may matter in important ways in practical applications where a researcher is trying to interpret a PGI from a specific GWAS. To illustrate, we show below that the theoretical framework can be extended to account for the vector of control variables, C, included in the GWAS. The theoretical regression equation that defines the vector needs to be modified to include the control variables:

$$(\gamma^C, \kappa) = \arg\min_{(\tilde{\gamma}^C, \tilde{\kappa})} E[(y_i - x_i'\tilde{\gamma}^C - C_i'\tilde{\kappa})^2],$$

where we use the C superscript to highlight the fact that the optimal weight vector with controls, γ^c, generally differs from the optimal weight vector without controls, which we denoted γ in the main text. Although the additive SNP factor $g_i^C \equiv x_i'\gamma^C$ is in general different from $g_i \equiv x_i'\gamma$, as it is derived from γ^C rather than γ, everything proceeds from here onward like in the main text. Since g_i^c is a best linear predictor, it can be understood as a standardized, noisy measure of an unobserved, latent variable, and the error-in-variables bias and the measurement-error-corrected estimator formulas remain the same, with coefficients from the conditional analyses replacing the univariate coefficients. Compared to the main text, the only difference is that g_i^c is now the best linear predictor of the phenotype *conditional* on the controls.

7 Polygenic Index Repository User Guide

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.

7.1 Summary information about Repository PGIs

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

7.1.1 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.

7.1.2 Supplementary phenotypes and MTAG for multi-trait PGIs

The multi-trait PGIs are based on MTAG analyses[2] 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.

7.1.3 PGI construction

The PGIs were made using LDpred[3] 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[4]. 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. Notably, even though the *expected* predictive power of each PGI is greater than 1%, in many instances, the *actual* predictive power of the PGI in a particular dataset may be less than 1%.

7.1.4 PC construction

As part of the Repository, we also release 20 principal components (PCs) based on the genome-wide data in each of the praticipating cohorts. The primary purpose of the release is to make them available for users who wish to use them as controls for population stratification. In order to make the PCs, we first restricted the samples to European-ancestry individuals and removed markers with imputation accuracy less than 70% or minor allele frequency less than 1%, as well as markers in long-range LD blocks (chr5:44mb-51.5mb, chr6:25mb-33.5mb, chr8:8mb-12mb, chr11:45mb-57mb). We then pruned all SNPs that survived these filters using a 1Mb rolling window (incremented in steps of 5 variants) and an r^2 threshold of 0.1 Next, we calculated the pairwise relatedness between all individuals in our full sample and generated a sample of conventionally unrelated individuals by dropping one individual from each pair of individuals with an estimated relatedness greater than 0.05. We then estimated SNP loadings for the top 20 PCs in this sample of approximately unrelated individuals. Finally, we used the estimated SNP loadings to compute 20 PCs for *all* individuals in the full sample (including both members from all pairs whose estimated relatedness exceeded our 0.05 threshold).

In HRS, we re-labeled the PCs in sets of five in order to address identifiability concerns. Therefore, it is only possible to infer from the variable name of a PC if it is one of the first five PCs (PC 1-5), one of the next five PC (PCs 6-10), etc.

7.1.5 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.

7.1.6 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.

7.1.7 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. Note that the HRS sample used in our validation exercise (2006-2010) is smaller than the HRS sample for which we are releasing PGIs (2006-2012) because we only had access to phenotype data in the former.

7.1.8 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).

7.2 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).

7.2.1 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 researchers to conduct association analyses using mixed-linear models[5,6] 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 the association models in the 23andMe GWASs 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.

7.2.2 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 GWASs underlying the Repository PGIs – generate PGIs that can be subject to a number of confounds to a causal interpretation[7–10]. 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"[8]. 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"[11]: 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 [12]. 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 [7,9,10]. 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.

7.2.3 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 [13,14], 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.

7.2.4 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 in Supplementary Table 10. 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 7.5 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 whenever an analysis employs the multi-trait PGI.

7.2.5 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, at conception, the genotypes of the causal SNPs responsible for the predictive power of the PGI [15,16].

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.

7.2.6 Genetic Effects Can Operate Through Environmental Mechanisms

We urge 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 [17]. Consider a PGI for BMI [13]. 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 myriad 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 [18,19].

Second and relatedly, it is therefore a fallacy to assume that any genetic sources of heterogeneity captured by a PGI are immutable—or even 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 [20,21]. 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 researchers to avoid describing PGIs as "genetic endowments" or other terms 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 is defined conditional on the GWAS phenotype, population, and environment, the same PGI may have different predictive power in different samples if there are differences in the phenotype measure, population sampled, the sampling methodology, or the environmental context. 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 [22,23]. 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.

References

[1] Goddard, M.E., *et al.* Estimating effects and making predictions from genome-wide marker data. *Statistical Science*, **24**, 517–529 (2009).

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

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

[4] 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).

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

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

[7] 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).

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

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

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

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

[12] 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).

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

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

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

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

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

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

[19] 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).

[20] 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).

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

[22] 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).

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

Supplementary Note

April 9, 2021

Contents

1 Data Access Procedures

23andMe

Upon publication of this paper, investigators at non-profit institutions can obtain access to the genome-wide summary statistics from 23andMe used in this paper by completing the 23andMe Publication Dataset Access Request Form. The information provided on this form will be used to generate a Statement of Work (SOW) that will allow 23andMe to transfer data for use in the described research project. The SOW and a Data Transfer Agreement will need to be signed by the institution and 23andMe before data can be shared. The form, as well as additional information and requirements, are available at https://research.23andme.com/dataset-access/.

Add Health

Access to the polygenic indexes and full phenotype data in Add Health is publicly available via a restricted data use contract with the University of North Carolina at Chapel Hill. Obtain access by visiting the CPC Data Portal at data.cpc.unc.edu/projects/2/view or see the Add Health contracts page at www.cpc.unc.edu/projects/addhealth/contracts. Add Health genotype data can be accessed via the database of Genotypes and Phenotypes (dbGaP, www.ncbi.nlm.nih.gov/gap, accession number phs001367.v1.p1).

Dunedin Multidisciplinary Health and Development Study

The datasets reported in the current article are available on request by qualified scientists. Requests require a concept paper describing the purpose of data access, ethical approval at the applicant's university, and provision for secure data access. We offer secure access on the Duke, Otago and King's College campuses. All data analysis scripts and results files are available for review. For more information, see moffittcaspi.trinity.duke.edu/research-topics/dunedin.

ELSA

Polygenic indexes and genotype data are publicly available and are available here: https://www.elsa-project.ac.uk/genetics. Phenotype and other publicly available data can be downloaded from the UK Data Service: https://beta.ukdataservice.ac.uk/datacatalogue/studies/study?id=5050. Use is limited to non-profit research use only. For more information regarding the data please contact o.ajnakna@ucl.ac.uk.

E-Risk

The datasets reported in the current article are available on request by qualified scientists. Requests require a concept paper describing the purpose of data access, ethical approval at the applicant's university, and provision for secure data access. We offer secure access on the Duke and King's College campuses. All data analysis scripts and results files are available for review. For more information, see moffittcaspi.trinity.duke.edu/research-topics/erisk.

EGCUT

Estonian Biobank data is available for academic research. To request phenotype, polygenic index, and/or genotype data, researchers need to fill out a preliminary request form (available at genomics.ut.ee/en/biobank.ee/data-access) and submit it via e-mail to releases@ut.ee. The preliminary request will be evaluated by the Estonian Committee on Bioethics and Human Research. Upon positive review, researchers need to fill out a request form (also available at genomics.ut.ee/en/biobank.ee/data-access) and submit it via e-mail to releases@ut.ee. The data will then be shared pursuant to a Data Use Agreement. For further details, see genomics.ut.ee/en/biobank.ee/data-access.

HRS

Polygenic scores are publicly available and can be downloaded here: https://hrsdata.isr.umich.edu/data-products/ssgac-polygenic-index-pgi-repository. Phenotype and other publicly available data can be downloaded here: hrs.isr.umich.edu/data-products. Genotype data can be accessed via the database of Genotypes and Phenotypes (dbGaP, www.ncbi.nlm.nih.gov/gap, accession number phs000428.v1.p1 and phs000428.v2.p2) with the most recent version forthcoming via NIAGADS (www.niagads.org/). Use is limited to non-profit research use only.

MCTFR

Access to the MCTFR PGIs is available by contacting Matt McGue (mcgue001@umn.edu), who will provide access authorization. Access to MCTFR phenotypic data will require a research proposal the structure of which can be provided by Matt McGue. Please note that the MCTFR is a complex, longitudinal study with thousands of relevant phenotypes assessed at multiple points in time. An overview of the range of phenotypes and developmental periods can be found in Wilson et al. (2019). Use of phenotypic data requires an approved proposal that is approved by the MCTFR Principal Investigator Committee; access to the MCTFR PGIs does not require an approved proposal. Because of the complexities involved, developing a proposal typically involves multiple iterations with MCTFR staff and are dealt with on a case-by-case basis.

STR

Researchers interested in using STR data must obtain approval from the Swedish Ethical Review Authority and from the Steering Committee of the Swedish Twin Registry. Researchers using STR data are required to follow the terms of a number of clauses designed to ensure protection of privacy and compliance with relevant laws. For further information please visit ki.se/en/research/the-swedish-twin-registry.

TTP

Access to the polygenic indexes and phenotype data from the Texas Twin Project is available via a restricted data use contract with the University of Texas at Austin. Restricted data users must develop an IRB-approved research proposal and security plan that ensures secure use of the data to minimize deductive disclosure risks. To apply for restricted-use data, please visit https://redcap.prc.utexas.edu/redcap/surveys/?s=FHJW9KCW8K.

UKB

All bona fide researchers can apply to use the UK Biobank resource for health related research that is in the public interest. Researchers can register and apply for data access at https://www.ukbiobank.ac.uk/register-apply/. Prior to publication of this paper, we will return the Repository PGIs to the UKB in accordance with their "returning results" procedure: https://biobank.ndph.ox.ac.uk/showcase/exinfo.cgi?src=returning_results. UKB will subsequently make the PGIs available to researchers as "Derived data-fields."

WLS

In addition to phenotype data, the polygenic index data is publicly available. As of February 2019, researchers who wish to use these polygenic indexes should email a brief research proposal and a copy or link to their CV to wls@ssc.wisc.edu. Given the need to preserve participant confidentiality, to access the complete genotyped data, researchers will additionally need to receive IRB approval from their home institution and enter into a Data Use Agreement between the researcher's home institution and the University of Wisconsin-Madison. For the most up-do-date instructions, see www.ssc.wisc.edu/wlsresearch/documentation/GWAS/.

2 Dataset Profiles

23andMe

Eriksson, N. *et al.* Web-Based, Participant-Driven Studies Yield Novel Genetic Associations for Common Traits. *PLOS Genetics* 6(6), 1–20 (2010).

Add Health

Harris, K. M. *et al.* Cohort Profile: The National Longitudinal Study of Adolescent to Adult Health (Add Health). *International Journal of Epidemiology* 48(5), 1415–1425 (2019).

Dunedin Multidisciplinary Health and Development Study

Poulton, R, *et at.*. The Dunedin Multidisciplinary Health and Development Study: Overview of the first 40 years, with an eye to the future. *Social Psychiatry and Psychiatric Epidemiology* 50, 679–693 (2015).

ELSA

Steptoe, A. *et al.* "Cohort Profile: The English Longitudinal Study of Ageing (ELSA)." *International Journal of Epidemiology* 42(6), 1640–1648 (2013).

E-Risk

None.

EGCUT

Leitsalu, L. *et al.* Cohort Profile: Estonian Biobank of the Estonian Genome Center (EGCUT), *International Journal of Epidemiology* 44, 1137–1147 (2015).

HRS

Sonnega, A. *et al.* Cohort Profile: the Health and Retirement Study (HRS), *International Journal of Epidemiology* 43(2), 576–85 (2014).

MCTFR

Wilson, S. *et al.* Minnesota Center for Twin and Family Research (MCTFR). *Twin Research and Human Genetics* 22(6), 746–752 (2019).

STR

Zagai, U. *et al.* The Swedish Twin Registry (STR): Content and Management as a Research Infrastructure. *Twin Research and Human Genetics* 22(6), 672–680 (2019).

TTP

Harden, K.P. *et al.* The Texas Twin Project (TTP). *Twin Research and Human Genetics* 16(1), 385–90 (2013).

UKB

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(3) (2015).

WLS

Herd, P. *et al.* Cohort profile: Wisconsin Longitudinal Study (WLS). *International Journal of Epidemiology* 43, 34–41 (2014).

3 Dataset-Specific Acknowledgments

We gratefully acknowledge research participants from all cohorts.

23andMe

We gratefully acknowledge the contributions of members of 23andMe's Research Team, whose names are listed below: Michelle Agee, Babak Alipanahi, Adam Auton, Robert K. Bell, Katarzyna, Bryc, Sarah L. Elson, Pierre Fontanillas, Nicholas A. Furlotte, Karen E. Huber, Nadia K. Litterman, Jennifer C. McCreight, Matthew H. McIntyre, Joanna L. Mountain, Carrie A.M. Northover, Steven J. Pitts, J. Fah Sathirapongsasuti, Olga V. Sazonova, Janie F. Shelton, Suyash Shringarpure, Chao Tian, Joyce Y. Tung, Vladimir Vacic, and Catherine H. Wilson.

Add Health

The National Longitudinal Study of Adolescent to Adult Health (Add Health) is supported by grant P01 HD031921 to Kathleen Mullan Harris from the Eunice Kennedy Shriver National Institute of Child Health and Human Development (NICHD), with cooperative funding from 23 other federal agencies and foundations. Add Health GWAS data were funded by NICHD grants to Harris (R01 HD073342) and to Harris, Boardman, and McQueen (R01 HD060726). For information about access to the data from this study, contact addhealth@unc.edu.

Dunedin Multidisciplinary Health and Development Study

Dunedin Multidisciplinary Health and Development Study research is supported by National Institute on Aging grants R01AG032282, R01AG049789, UK Medical Research Council grant MR/P005918, the New Zealand Health Research Council and New Zealand Ministry of Business, Innovation, and Employment.

ELSA

The English Longitudinal Study of Ageing is jointly run by University College London, Institute for Fiscal Studies, University of Manchester and National Centre for Social Research. Genetic analyses have been carried out by UCL Genomics and funded by the Economic and Social Research Council (ES/K005774/1) and the National Institute on Aging (R01 AG017644). All GWAS data has been deposited in the European Genome-phenome Archive. For more information please refer to www.elsa-project.ac.uk/genetics, or contact o.ajnakina@ucl.ac.uk.

E-Risk

The E-Risk study is funded by grant G1002190 from the UK Medical Research Council and grant HD077482 from the National Institute of Child Health and Development.

EGCUT

EGCUT received funding from the Estonian Research Council Grant PUT1660 and PRG184, Mobilitas Pluss ERA-NET grant SP1GI18045T, Horizon 2020 program grants MMVCM18418R, and European Union through the European Regional Development Fund SLTMR16142T. For more information, please contact Tõnu Esko (tonu.esko@ut.ee).

MCTFR

This project was led by William G. Iacono, PhD. and Matt McGue, PhD (Co-Principal Investigators) at the University of Minnesota, Minneapolis, MN, USA. Co-investigators from the same institution included: Irene J. Elkins, Margaret A. Keyes, James J. Lee, Lisa N. Legrand, Stephen M. Malone, William S. Oetting, Michael B. Miller, Saonli Basu and Scott Vrieze. Funding support for this project was provided through NIDA (U01DA024417). Other support for sample ascertainment and data collection came from several grants: R37DA05147, R01AA09367, R01AA11886, R01DA13240, R01MH66140.

STR

The Swedish Twin Registry (STR) is managed by Karolinska Institutet and receives additional funding through the Swedish Research Council under the grant no 2017-00641. Other funding for the project come from the Ragnar Söderberg Foundation (E9/11), the Swedish Research Council (421-2013-1061).

TTP

The Texas Twin Project is supported by grants R01HD083613 and R01HD092548 from NIH/NICHD and Jacobs Foundation Research Fellowships.

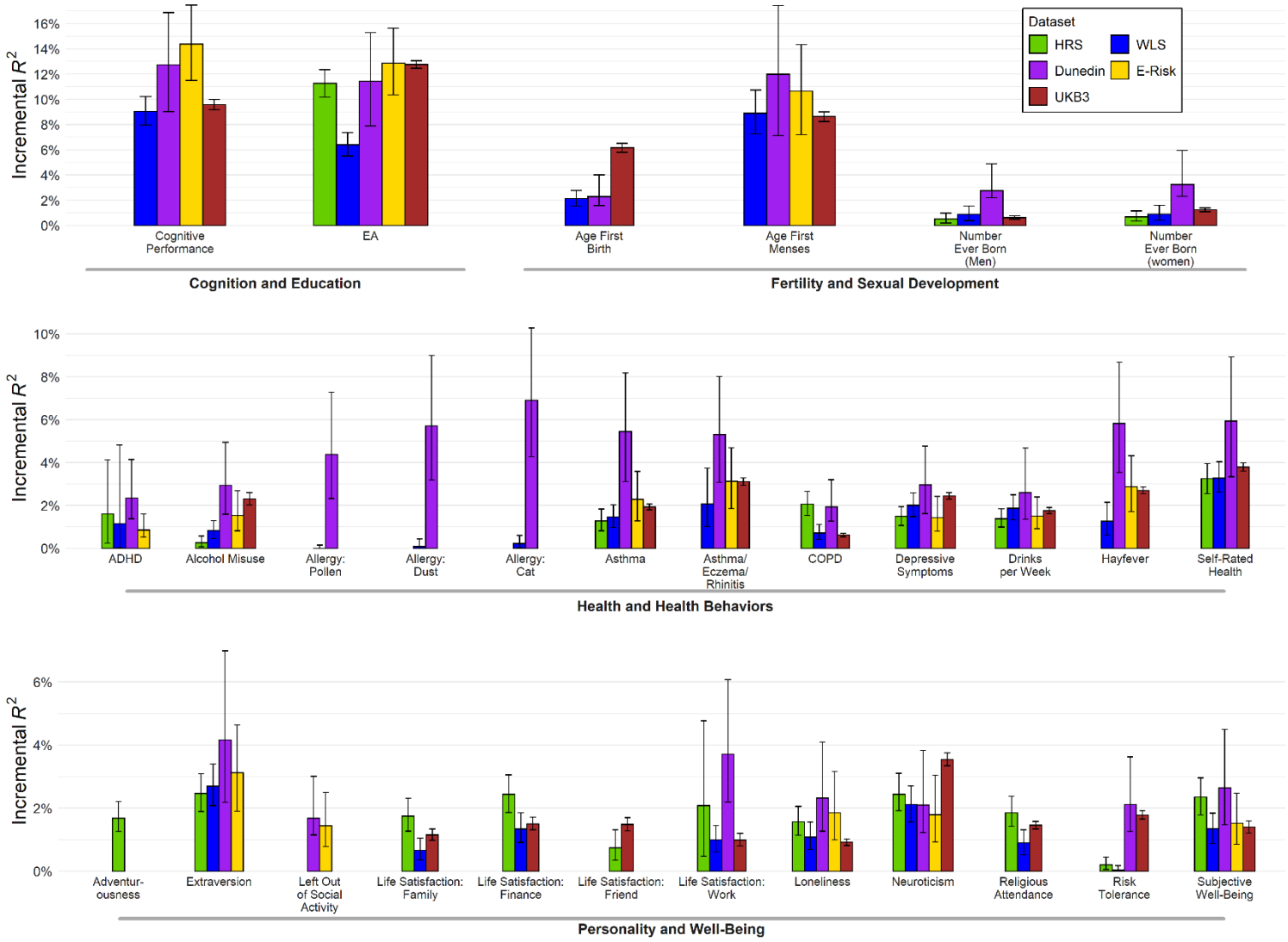
WLS

This research uses data from the Wisconsin Longitudinal Study (WLS) of the University of Wisconsin-Madison. Since 1991, the WLS has been supported principally by the National Institute on Aging (AG-9775, AG-21079, AG-033285, and AG-041868, R01 AG041868-01A1), with additional support from the Vilas Estate Trust, the National Science Foundation, the Spencer Foundation, and the Graduate School of the University of Wisconsin-Madison. Since 1992, data have been collected by the University of Wisconsin Survey Center. The opinions expressed herein are those of the authors. A public use file of data from the Wisconsin Longitudinal Study is available from the Wisconsin Longitudinal Study, University of Wisconsin-Madison, 1180 Observatory Drive, Madison, Wisconsin 53706 and at www.ssc.wisc.edu/WLSresearch/data/.

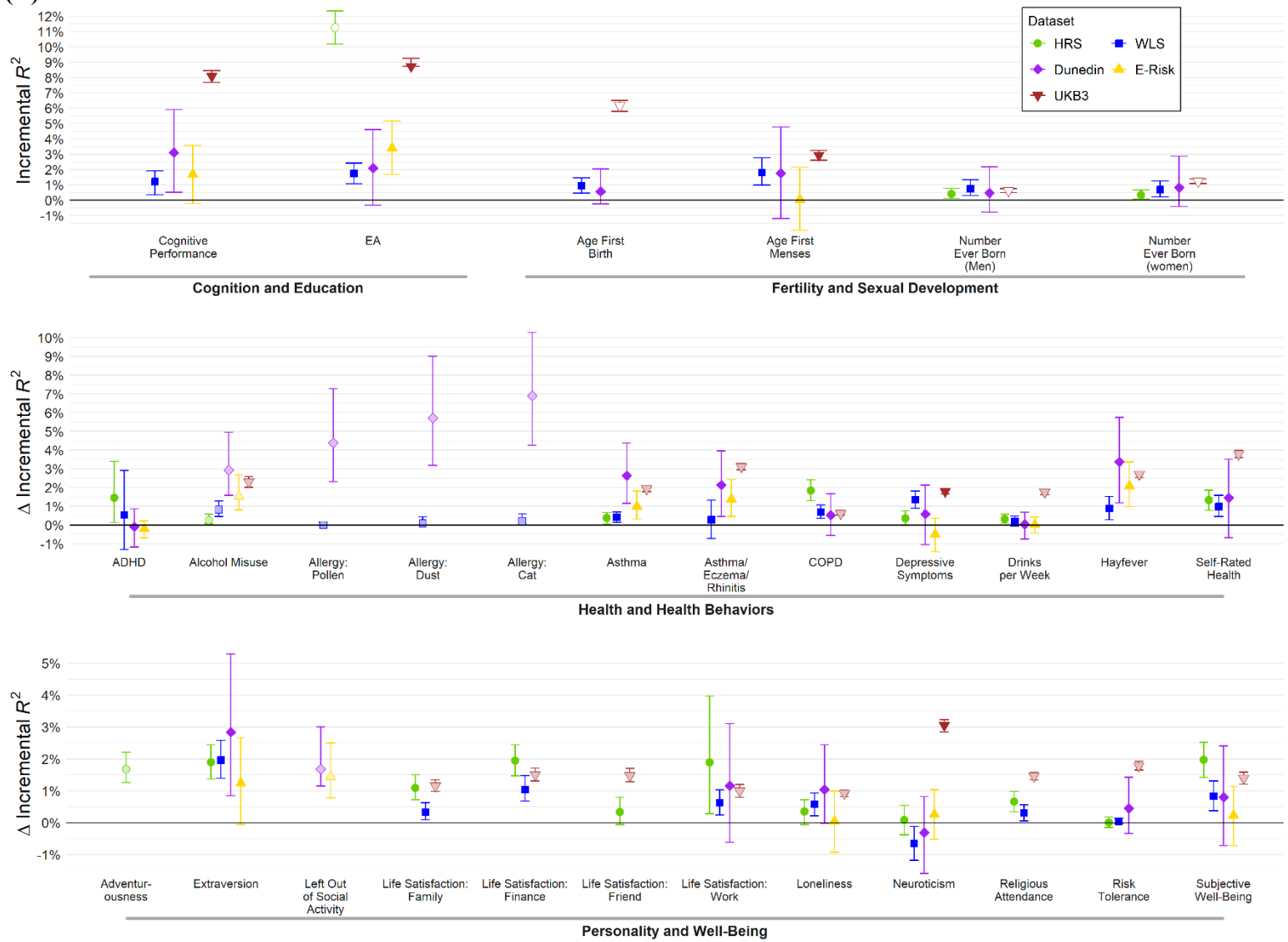
4 Dataset Authorship Contributions

Dataset	Author	Study design & mgmt.	Data collection	Genotyping	Genotype prep.	Phenotype prep.	Data analysis	Writing
23andMe	Aaron Kleinman					X	X	X
23andMe	David A. Hinds	X				X		X
23andMe	23andMe Research Group		X	X		X	X	
Add Health	Kathleen Mullan Harris	X	X	X	X	X		
Dunedin Study	Daniel W. Belsky					X	X	X
Dunedin Study	Avshalom Caspi	X	X				X	X
Dunedin Study	David L. Corcoran			X	X		X	
Dunedin Study	Terrie E. Moffitt	X	X			X		X
Dunedin Study	Richie Poulton	X	X			X		X
Dunedin Study	Karen Sugden			X	X	X	X	X
Dunedin Study	Benjamin S. Williams			X	X	X		
ELSA	Andrew Steptoe	X	X					
ELSA	Olesya Ajnakina				X			
E-Risk	Daniel W. Belsky					X	X	X
E-Risk	Avshalom Caspi	X	X				X	X
E-Risk	David L. Corcoran			X	X		X	
E-Risk	Terrie E. Moffitt	X	X			X		X
E-Risk	Karen Sugden			X	X	X	X	X
E-Risk	Benjamin S. Williams			X	X	X		
EGCUT	Lili Milani	X	X	X	X	X	X	
EGCUT	Tõnu Esko	X	X	X	X	X	X	
MCTFR	William G. Iacono	X	X					
MCTFR	Matt McGue	X	X			X		
STR	Rafael Ahlskog	X						
STR	Patrik K.E. Magnusson	X	X			X		
TTP	Travis T. Mallard				X			
TTP	K. Paige Harden	X						
TTP	Elliot M. Tucker-Drob	X						
WLS	Pamela Herd	X	X	X				
WLS	Jeremy Freese	X	X	X				

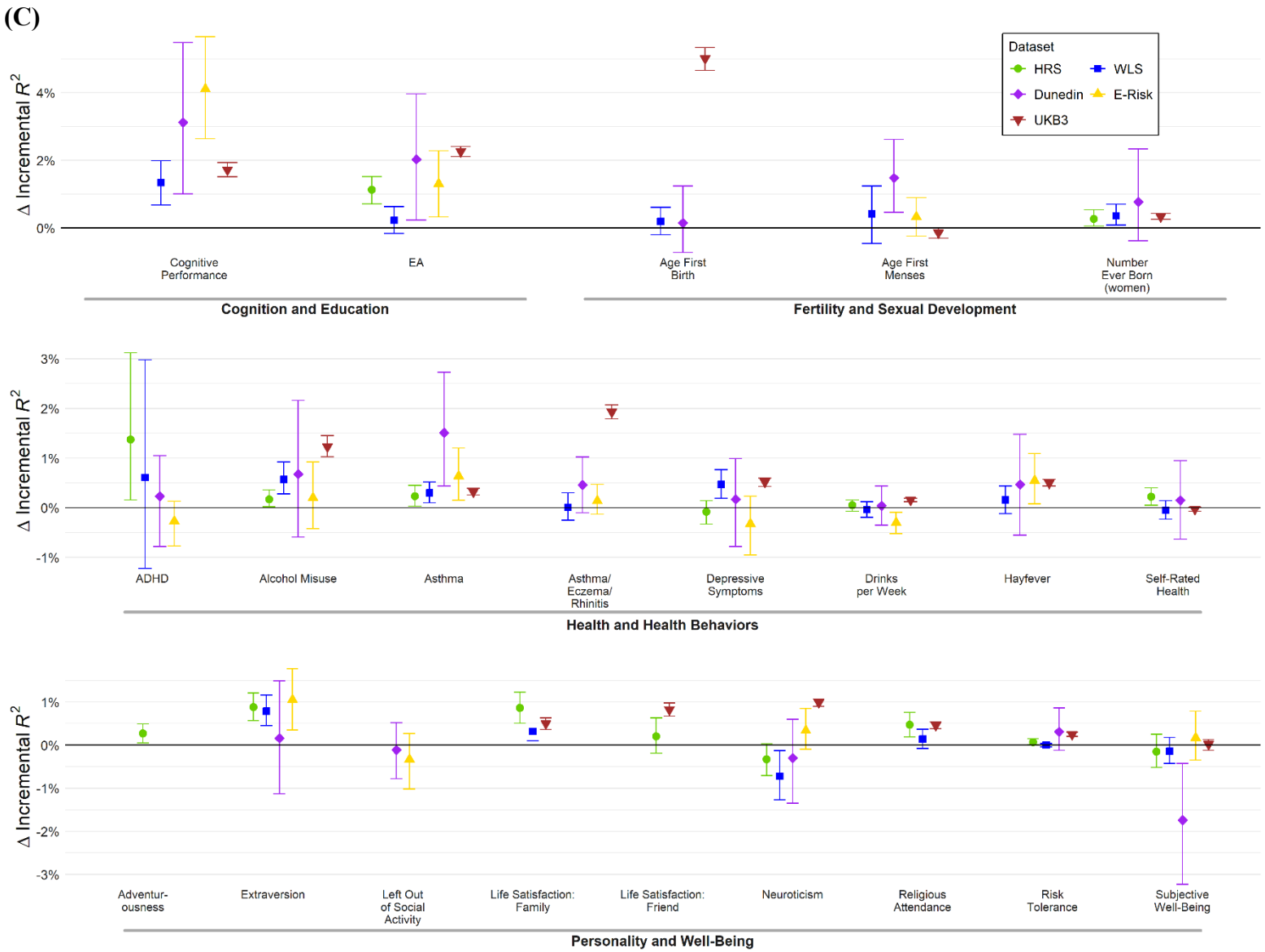
Supplementary Figure 1. Predictive power of Repository multi-trait PGIs
(A)



(B)


Incremental R^2
12%
11%
10%
9%
8%
7%
6%
5%
4%
3%
2%
1%
0%
-1%
Dataset
HRS
WLS
Dunedin
E-Risk
UKB3
Cognitive Performance
EA
Age First Birth
Age First Menses
Number Ever Born (Men)
Number Ever Born (women)
Cognition and Education
Fertility and Sexual Development
Δ Incremental R^2
10%
9%
8%
7%
6%
5%
4%
3%
2%
1%
0%
-1%
ADHD
Alcohol Misuse
Allergy: Pollen
Allergy: Dust
Allergy: Cat
Asthma
Asthma/ Eczema/ Rhinitis
COPD
Depressive Symptoms
Drinks per Week
Hayfever
Self-Rated Health
Health and Health Behaviors
Δ Incremental R^2
5%
4%
3%
2%
1%
0%
-1%
Adventur-ousness
Extraversion
Left Out of Social Activity
Life Satisfaction: Family
Life Satisfaction: Finance
Life Satisfaction: Friend
Life Satisfaction: Work
Loneliness
Neuroticism
Religious Attendance
Risk Tolerance
Subjective Well-Being
Personality and Well-Being

(C)


Dataset
HRS
WLS
Dunedin
E-Risk
UKB3
Δ Incremental R^2
4%
2%
0%
Cognitive Performance
EA
Age First Birth
Age First Menses
Number Ever Born (women)
Cognition and Education
Fertility and Sexual Development
3%
2%
1%
0%
-1%
ADHD
Alcohol Misuse
Asthma
Asthma/ Eczema/ Rhinitis
Depressive Symptoms
Drinks per Week
Hayfever
Self-Rated Health
Health and Health Behaviors
1%
0%
-1%
-2%
-3%
Adventur- ousness
Extraversion
Left Out of Social Activity
Life Satisfaction: Family
Life Satisfaction: Friend
Neuroticism
Religious Attendance
Risk Tolerance
Subjective Well-Being
Personality and Well-Being

Notes: Error bars show 95% confidence intervals from bootstrapping with 1,000 repetitions. Panel (A): Incremental R^2 from adding Repository's multi-trait PGI to a regression of the phenotype on 10 principal components of the genetic relatedness matrix for HRS, WLS, Dunedin, and E-Risk, and on 20 principal components and 106 genotyping batch dummies for UKB. Prior to the regression, phenotypes are residualized on a second-degree polynomial for age or birth year, sex, and their interactions (see Supplementary Tables 5 and 12). For the GWAS-equivalent sample sizes of the summary statistics that the PGIs are based on, see Supplementary Table 10. Panel (B): Difference in incremental R^2 between Repository multi-trait PGI and PGI constructed from publicly available summary statistics using our Repository pipeline. (Note that the latter do not include PGI directly available from datasets, such as the ones accessible from the HRS website.) If no publicly available summary statistics are available for a phenotype, then the difference in incremental R^2 is equal to the incremental R^2 of the single-trait PGI and is represented by an open circle. For the GWAS sample sizes of the PGIs based on publicly available summary statistics, see Supplementary Table 13. Panel (C): Difference in incremental R^2 between Repository multi-trait PGI and the single-trait PGI corresponding to the same phenotype.

Frequently Asked Questions (FAQs)

This document provides information about the study:

Becker *et al.* (2021) "Resource Profile and User Guide of the Polygenic Index Repository" *Nature Human Behaviour*, in press.

The document was prepared by Daniel Benjamin, David Laibson, Michelle N. Meyer, and Patrick Turley. It draws from and builds on the FAQs for earlier SSGAC papers. It has the following sections:

1. **Background**
2. **Study design and results**
3. **Social and ethical implications of the study**
4. **Appendices**

For clarifications or additional questions, please contact Daniel Benjamin (daniel.benjamin@gmail.com).

Table of Contents

1. Background

1.1. Who conducted this study? What are the group's overarching goals?

The authors of the study are researchers affiliated with the Social Science Genetic Association Consortium (SSGAC) as well as data providers (i.e., individuals who act as stewards for datasets and provide other researchers with access to these data for research purposes). The SSGAC is a multi-institutional, international research group that aims to identify statistically robust associations between variation in DNA and variation in social-science-relevant outcomes.

We study the most common sources of genetic variation—single-nucleotide polymorphisms (SNPs). SNPs are sites in the genome where single DNA base pairs commonly differ across individuals. Each SNP usually has two different possible base pairs, which are called alleles. Although there are tens of millions of sites where SNPs are located in the human genome, our work (like most genetic research today that aims to link variation in DNA to variation in disease and other outcomes) investigates only SNPs that can be easily measured with a high level of accuracy. These days, we can easily and accurately measure millions of SNPs, which together capture most of the common genetic variation across people.

The social-science-relevant outcomes that we analyze include differences across people in behavior, preferences, and personality that are traditionally studied by social and behavioral scientists (e.g., anthropologists, economists, political scientists, psychologists, and sociologists). These traits are often also of interest to health and other researchers.

The SSGAC was formed in 2011 to address a specific set of scientific challenges. Most outcomes and behaviors are weakly associated with a very large number of SNPs. Although their collective effect can be meaningful (see FAQs 1.2 & 2.3), we now know that almost every one of these SNPs has an extremely weak association on its own. To identify specific SNPs with such small effects, scientists must study at least hundreds of thousands of people (to separate weak signals from sampling noise). One promising strategy for doing this is for many investigators to pool their data into one large study. This approach has borne considerable fruit when used by medical geneticists interested in a range of medical conditions (Visscher et al., 2017). Most of these advances would not have been possible without large research collaborations between multiple research groups interested in similar questions. The SSGAC was formed in an attempt by social scientists to adopt this research model.

The SSGAC is organized as a working group of the Cohorts for Heart and Aging Research in Genomic Epidemiology (CHARGE), a successful medical consortium. (In genetics research, "cohort" is a term that means "dataset.") The SSGAC was founded by three social scientists—Daniel Benjamin (University of California – Los Angeles), David Cesarini (New York University), and Philipp Koellinger (University of Wisconsin and Vrije Universiteit Amsterdam)—who believe that studying SNPs associated with social scientific outcomes can have substantial positive impacts across many research fields. This includes research that aims to better understand the effects of the environment (e.g., research on policy interventions) and interactions between genetic and environmental effects. The

potential benefits also span a diverse set of research questions in the biomedical sciences, such as why and how educational attainment is linked to longevity and better overall health outcomes.

To conduct such research, the SSGAC implements genome-wide association studies (GWAS) of social-scientific outcomes. For example, to conduct a GWAS of educational attainment (e.g., Lee et al., 2018) every participating cohort calculates the cross-sectional (i.e., within-cohort) correlation between educational attainment and DNA-base-pair variation at a single location on the genome: a SNP. As first discussed above, a SNP is a base-pair of the genome where there is common variation in the human population. This statistical analysis is repeated for each SNP on the genome. The cohort-level results do not contain individual-level data—just summary statistics about these within-cohort statistical associations. The SSGAC then combines these cohort results to produce the overall GWAS results. By using existing datasets and combining cohort results, we can study the genetics of large numbers of individuals (for example, ~1.1 million people in Lee et al. (2018)) at very low cost. The SSGAC publicly shares overall, aggregated results (subject to some Terms of Service; see FAQ 3.7) so that other scientists can build on this work. These publicly available data have already catalyzed many research projects and analyses across the social and biomedical sciences. Among the most useful products of these GWASs for other research are the polygenic indexes that are based on GWAS associations. Polygenic indexes are variables that aggregate the predictive power of many SNPs for predicting the outcome of the GWAS (see FAQ 1.2), and they are the focus on the current paper.

The Advisory Board for the SSGAC is composed of prominent researchers representing various disciplines: Dalton Conley (Sociology, Princeton University), George Davey Smith (Epidemiology, University of Bristol), Tõnu Esko (Molecular Biology and Human Genetics, University of Tartu and Estonian Genome Center), Albert Hofman (Epidemiology, Harvard University), Robert Krueger (Psychology, University of Minnesota), David Laibson (Economics, Harvard University), James Lee (Psychology, University of Minnesota), Sarah Medland (Genetic Epidemiology, QIMR Berghofer Medical Research Institute), Michelle Meyer (Bioethics and Law, Geisinger Health System), and Peter Visscher (Statistical Genetics, University of Queensland).

The SSGAC is committed to the principles of reproducibility and transparency. Major SSGAC publications are usually accompanied by a FAQ document (such as this one). The FAQ document is written to communicate what was found less tersely and technically than in the paper, as well as what can and cannot be concluded from the research findings more broadly. FAQ documents produced for SSGAC publications are available on the SSGAC website.

To date, SSGAC-affiliated papers have studied educational attainment, cognitive performance, subjective well-being, reproductive behavior, risk tolerance, and dietary intake. The SSGAC website contains a list of our major publications, which have been published in journals such as *Science*, *Nature*, *Nature Genetics*, *Proceedings of the National Academy of Sciences*, *Psychological Science*, and *Molecular Psychiatry*.

1.2. What is a polygenic index (PGI)? Why this terminology?

A polygenic index (we use the acronym PGI throughout the paper) is an index composed of a large number of SNPs from across the genome. Each polygenic index is associated with a particular outcome

(for details, see FAQ 1.3). Because a polygenic index aggregates the information from many SNPs, it can "predict" (see FAQ 1.6) far more of the variation among individuals than any single SNP. (Note that even polygenic indexes are not good predictors of outcomes for one person; see FAQ 3.3). Often, the polygenic indexes with the most predictive power are those created using *all* the (millions of) SNPs measured in a SNP array. A SNP array is the currently standard way of measuring common genetic differences across individuals. A SNP array data does not measure the entire genetic sequence of each individual, but it does measure most of the places on the genome where individuals differ.

Our terminology of *polygenic index* is currently non-standard, but most of the authors of the paper prefer it to current terms and hope that this paper, and the Polygenic Index Repository introduced in this paper, make polygenic index a standard term. The traditional terms include polygenic risk score and polygenic score. The word *risk* makes little sense when the polygenic index is for a non-disease outcome (such as height). The word *score* was intended to echo statistical nomenclature but can instead convey an unintended value judgment or valence (i.e., "a higher score must be better"). The word *index* is at least as accurate statistically and does not convey a value judgment.

1.3. How is a polygenic index constructed?

A polygenic index is constructed in three steps. First, a genome-wide association study (GWAS) is conducted, looking at SNPs measured across the entire human genome to see which of them are associated with higher or lower levels of some outcome. As explained above, SNPs are sites in the genome where single DNA base pairs commonly differ across individuals. SNPs usually have two different possible base pairs, or alleles. Although there are tens of millions of sites where SNPs are located in the human genome, GWASs typically investigate only SNPs that can be easily measured (or imputed) with a high level of accuracy. These days, we can easily and accurately measure millions of SNPs, which together capture most of the common genetic variation across people. For each of these millions of SNPs, the GWAS generates an "effect size" corresponding to the (typically miniscule) magnitude of the association between that SNP and the outcome. (We use the term "effect size" because it is a common scientific shorthand for "magnitude of association," but we emphasize that use of the term is not intended to imply that the SNP, or polygenic index, *causes* the outcome; see FAQ 1.5.)

Second, the effect sizes are used to determine the "weight" each SNP will get in the polygenic index. The simplest scheme is to weight each SNP by its effect size as estimated in the GWAS. This simple weighting scheme has one main problem: because SNPs tend to be correlated with nearby SNPs on the genome (a phenomenon called linkage disequilibrium), if one SNP is associated with the outcome, nearby SNPs will also be associated with the outcome. State-of-the-art approaches to determining the weights for a polygenic index are designed to address this problem. We use a common approach called LDpred (Vilhjálmsson et al., 2015). Using the results of a GWAS, LDpred generates a weight for each SNP. These weights are not equal to the SNPs' effect sizes as estimated in the GWAS, mostly because the weights take into account each SNP's correlation with other SNPs. (Even though LDpred addresses the issue of linkage disequilibrium, it does so only for the purpose of generating weights for optimal prediction. LDpred will not necessarily assign more weight to the SNP whose association with the outcome is responsible for nearby SNPs' associations with the outcome. Thus, LDpred is a tool to address the issue of linkage disequilibrium for the purpose of prediction—which is the purpose of a polygenic index—but not for the purpose of unbiased estimation of SNPs' effect sizes. See FAQ 1.5.)

Third, the set of weights for the SNPs are used in a formula for calculating a polygenic index for any particular individual. The formula is a weighted sum of alleles at each SNP (using the weights from the second step). The formula is used to calculate a numerical value of the polygenic index for each individual in some dataset (that was not included in the GWAS).

The sample used for the GWAS in the first step is the training sample for the polygenic index. The larger the GWAS sample size, the greater the predictive power of a polygenic index constructed in the third step. However, this predictive power of a polygenic index has a maximum for each outcome that the polygenic index can approach as the sample size gets bigger, but it can never exceed.

1.4. How might polygenic indexes be useful?

A polygenic index for an outcome provides one measure of the genetic influence on that outcome that can be used in research in a variety of ways. For example, polygenic indexes have been used to:

- partially control for genetic influences in order to generate less noisy estimates of how changes in school policy influence health outcomes (Davies et al., 2018);
- examine how the effect of school policy on health outcomes depends in part on genetic influences (Barcellos, Carvalho and Turley, 2018a);
- study why SNPs predict educational attainment – for example, it appears that some genetic effects on educational attainment operate through associations with cognitive function and traits such as self-control (Belsky et al., 2016), which in turn affect educational attainment;
- investigate how genetic influences on educational attainment differ across environmental contexts (Schmitz and Conley, 2017; Barcellos, Carvalho and Turley, 2018b);
- investigate how genetic influences on BMI vary over the lifecycle (Khera et al., 2019);
- infer the degree of assortative mating (Robinson et al., 2017; Yengo et al., 2018);
- trace recent migration patterns (Domingue et al., 2018; Abdellaoui et al., 2019);
- examine whether polygenic indexes for disease risk are sufficiently predictive to be incorporated into clinical practice for preventative medicine (Khera et al., 2018); and
- develop new statistical tools that may advance our understanding of how parenting and other features of a child's rearing environment influence his or her developmental outcomes (Koellinger and Harden, 2018; Kong et al., 2018).

The idea of using GWAS results to create a polygenic index was initially proposed in 2007 (Wray, Goddard and Visscher, 2007), and the first polygenic index was created in 2009 in a GWAS of schizophrenia and bipolar disorder (Purcell et al., 2009). Since then, polygenic indexes have become a significant part of research that builds on genetics in the medical and social sciences. For example, in the current paper we analyze presentations at the annual meeting of the Behavior Genetics Association. We report that the fraction of presentations that used polygenic indexes increased from 0% in 2009 to 20% in 2019. The list above represents a few illustrative examples of research that uses polygenic indexes.

As discussed in FAQ 1.9 below, one goal of this paper, and the Polygenic Index Repository it introduces, is to facilitate further work using polygenic indexes by making a much wider range of more predictive polygenic indexes available to researchers.

1.5. Does a polygenic index "cause" the outcome of interest?

Polygenic indexes available today, including those we construct in this paper, should not be interpreted as a measure of causal mechanisms.

The genome-wide association studies (GWASs) used as the training data for the polygenic indexes (see FAQ 1.3) identify SNPs that are associated with the outcome, but an observed empirical correlation with a specific SNP need not imply that the SNP *causes* the outcome, for a variety of reasons. First, SNPs are often highly correlated with other, nearby SNPs on the same chromosome. As a result, when one or more SNPs in a region causally influence an outcome (in that particular environment), many non-causal SNPs in that region may also be identified as associated with the outcome (in FAQ 1.3, see the parenthetical "Even though LDpred…" for why LDpred does not solve this problem for the purpose of identifying the causal SNP). In fact, the causal SNP may not have even been measured directly. For example, GWAS that focus on common SNPs would not be able to identify rare or structural types of genetic variation (e.g., deletions or insertions of an entire genetic region) that are causal, but they may identify SNPs that are correlated with these unobserved variants. For these and other reasons, polygenic indexes are likely to be composed of a mix of causal and non-causal SNPs, and the weights used in the formula for constructing the polygenic index (see FAQ 1.3) should not be interpreted as estimates of the causal effects of the SNPs. As a very rough estimate, for social and behavioral outcomes, no more than about one-third of the predictive power of a polygenic index (i.e., the percentage of the variance in the outcome among individuals that the polygenic index explains) is explained by causal genetic effects (Howe et al., 2021). For instance, the most predictive polygenic index for educational attainment currently available explains about 12% of the variance between people, but only one-third of that—about 4%—is causal. (These causal SNPs may be among the SNPs included in the polygenic index or may be physically close to, and therefore correlated with, SNPs that are included.) In contrast, for anthropometric outcomes such as height, it is possible that nearly all of the predictive power of a polygenic index is explained by causal SNPs.

Second, at a particular SNP the frequency of different alleles might vary systematically across environments. If those environmental factors are not accounted for in the association analyses, some of the measured SNP associations with social-science outcomes may be spurious. To use a well-known example often used to explain this idea (Lander and Schork, 1994), any genetic variants common in people of Asian ancestries will be associated statistically with more frequent than average chopstick use, but these variants would not *cause* greater chopstick use; rather, these genetic variants and the outcome of chopstick use are both distributed unevenly among people with different ancestries. This is called the problem of "population stratification." The GWAS underlying the polygenic indexes in this paper employ standard strategies to try to minimize this problem, but the issues raised by population stratification cannot be ruled out entirely. As a result, the polygenic indexes likely reflect population stratification to some extent. In the User Guide that accompanies the Polygenic Index Repository (reproduced in the Supplementary Methods of the paper), we discuss this problem in more detail and discuss strategies for addressing the population stratification in the polygenic indexes

Even in GWAS (such as those we rely on or conduct ourselves) that attempt to address and correct for heterogeneity in genetic ancestry, allele frequencies may nonetheless vary systematically with environmental factors even *within* a group of people of similar genetic ancestry. For example, a SNP that is associated with improved educational outcomes in the parental generation may have downstream effects on parental income and other factors known to influence children's educational outcomes (such

as neighborhood characteristics). This same SNP is likely to be inherited by the children of these parents, creating a correlation between the presence of the SNP in a child's genome and the extent to which the child was reared in an environment with specific characteristics. A recent study of Icelandic families showed that a parental allele associated with higher educational attainment of the parent that is *not* passed on to the parent's offspring is still associated with the child's educational attainment, suggesting that GWAS results for educational attainment partly represent these intergenerational environmental pathways (Kong et al., 2018).

Third, a SNP's effects on an outcome may be indirect, so a SNP that may be "causal" in one environment may have a diminished effect or no effect at all in other environments. For example, variation in a particular SNP on chromosome 15 is associated with lung cancer (Amos et al., 2008; Hung et al., 2008; Thorgeirsson et al., 2008). From this observation alone we cannot conclude that variation in this SNP can cause lung cancer through some direct *biological* mechanism. In fact, it is likely that variation in this SNP, which is part of the nicotinic acetylcholine receptor gene cluster that affects nicotine metabolism, increases lung cancer risk through effects on smoking behavior. In a tobacco-free environment, it is plausible that this association with lung cancer would be substantially weaker and perhaps disappear altogether. Thus, even *if* we have credible evidence that a specific association is not spurious, it is entirely possible that the SNP in question influences the outcome through channels that we, in common parlance, would label environmental (e.g., smoking). Nearly forty years ago, the sociologist Christopher Jencks criticized the widespread tendency to mistakenly treat environmental and genetic sources of variation as mutually exclusive (see also Turkheimer, 2000). As the example of smoking illustrates, it is often overly simplistic to assume that "genetic explanations of behavior are likely to be exclusively physical explanations while environmental explanations are likely to be social" (Jencks, 1980, p723).

1.6. In what sense does a polygenic index "predict" the outcome of interest?

When we and other scientists say that polygenic indexes (and other variables, such as demographics or other environmental factors) "predict" certain outcomes, our use of the word differs in several important ways from how "predict" is used in standard language (e.g., outside of social science research papers). First, we do not mean that the polygenic index guarantees an outcome with 100% probability, or even with a high degree of likelihood. Rather, we mean that the polygenic index is, on average across people, statistically associated with an outcome. In other words, on average, people with a higher numerical value of the polygenic index have a higher likelihood of the outcome compared to people with a lower numerical value. A polygenic index is said to be statistically "predictive" of an outcome even if the polygenic index has only a *weak* association with the outcome—as is the case, for instance, with almost all of the polygenic indexes in this paper. In such cases, the polygenic index is only weakly predictive of the outcome.

Second, in standard language, "prediction" usually refers to the future. In contrast, when scientists say that a polygenic index "predicts" an outcome, they mean that they expect to see the association in *new data*. "New data" means data that haven't been analyzed yet—regardless of whether those data will be collected in the future or have already been collected. In other words, in social science, it makes perfect sense to ask how well a polygenic index predicts outcomes that have already occurred, like how many years of education were attained by older adults.

Finally, in standard language, a "prediction" is often an unconditional guess about what will happen. Instead of meaning it unconditionally, scientists mean that they expect to see an association in new data under certain conditions, for example, that the environment for the new data is the same as the environment in which the GWAS that underlies the polygenic index (see FAQ 1.3) was conducted. In the example given in FAQ 1.5, in which a SNP is associated with lung cancer due to an effect on smoking, we would *not* expect the SNP to be as strongly predictive of lung cancer, or predictive at all, in an environment where tobacco-based products are hard to obtain or absent entirely.

1.7. What polygenic indexes were available to researchers prior to this project?

Prior to this project, only a few datasets had constructed polygenic indexes that researchers could download and use. Notable examples of data providers that did make polygenic indexes directly available to researchers —all of which recognized early on the value of doing so—are the Health and Retirement Study, the Wisconsin Longitudinal Study, and the National Longitudinal Adolescent to Adult Health Study. The UK Biobank does not construct polygenic indexes for its users, but it provides a mechanism by which researchers who use the data and construct polygenic indexes can "return" them to the UK Biobank for use by other researchers. Through this mechanism, polygenic indexes constructed from several GWASs have been made available for researchers to download from the UK Biobank.

To study polygenic indexes in other datasets or for other outcomes, prior to this paper, researchers would need to construct the polygenic indexes themselves, following the steps described in FAQ 1.3. For the first step, most researchers would need to rely on publicly available GWAS results, which include less data and are therefore less predictive than some polygenic indexes in published work that rely on non-public GWAS results (see FAQ 2.3). Recently, to make it easier for researchers to construct polygenic indexes themselves, the Polygenic Score Catalog (Lambert *et al.*, 2020) collected together weights for a range of polygenic indexes (also based on publicly available GWAS results).

As we discuss in more detail in FAQ 2.1, for the Polygenic Index Repository, we constructed a large number of polygenic indexes in each of 11 datasets (including the four mentioned above) and have made the polygenic indexes directly available for researchers to download. The polygenic indexes are often based on more data than is publicly available, and the polygenic indexes are constructed according to a uniform methodology across both outcomes and datasets. For examples of Repository polygenic indexes that were previously not available at all or that were less accurate (i.e., predictive), see FAQ 2.3.

1.8. How do different polygenic indexes for the same outcome differ? How comparable are results across studies that use different polygenic indexes for the same outcome?

There are several reasons why polygenic indexes for the same outcome can differ from each other. As described in FAQ 1.3, there are three steps to creating a polygenic index, and differences can arise at each of these steps. For example, in the first step, researchers could base the polygenic index on different GWAS studies of the same outcome. Different GWAS studies may be based on samples who live under different environmental conditions, may have different measures of the outcome, and/or may have measured different SNPs. As another example, in the second step, researchers could use a different

method of determining polygenic-index weights from the results of a GWAS. For these and other reasons, it has been common for different studies to use different polygenic indexes, even when the polygenic indexes are for the same outcome and are being studied in the same dataset.

The results are typically difficult to compare across such studies for three main reasons:

1. If the polygenic indexes are constructed using different methods, then even though they are both measuring genetic influences on the outcome, the precise definition of these "genetic influences" may differ (see FAQs 3.1 and 3.2).
2. The units for measuring the strength of associations between the polygenic index and other variables generally differ across studies. Researchers usually report results in terms of standard deviations (a statistical unit) of the polygenic index, but if the polygenic index in one study is a more powerful predictor than that in the other study, then one standard deviation of one polygenic index means something different than one standard deviation of the other.
3. If one of the polygenic indexes is a more powerful predictor than the other, then they differ in their signal-to-noise ratio for capturing genetic influences on the outcome. Whenever an explanatory variable is measured with noise, results based on that variable will be distorted, sometimes in unanticipated ways. Since the signal-to-noise ratio differs across the polygenic indexes, results based on them are distorted differentially, further making the results difficult to compare.

1.9. Why create the Polygenic Index Repository?

In brief, the Polygenic Index Repository introduced in this paper has three main goals: (i) to make polygenic indexes for a large number of outcomes more accessible to a wider range of researchers from many fields and disciplines, including early career researchers, researchers without access to the data and/or training required to create the most state-of-the-art polygenic indexes, and researchers who wish to probe the limitations of polygenic indexes; (ii) to increase the use of polygenic indexes that are more accurate (i.e., predictive) than polygenic indexes researchers could construct from publicly available GWAS results; and (iii) to facilitate the comparability of results across studies that use these polygenic indexes.

In more detail, the Polygenic Index Repository addresses several practical obstacles that researchers interested in using polygenic indexes must often confront, including:

1. Constructing a polygenic index from genotype data requires special expertise. Even for researchers with that expertise, it can be a time-consuming process.
2. It is generally desirable to generate polygenic-index weights from the GWAS with the largest sample size because the predictive accuracy of a polygenic index is expected to be largest in that case. However, there are administrative hurdles for accessing some GWAS results, such as those from 23andMe. In practice, researchers often end up constructing polygenic indexes using only publicly available GWAS results. Such polygenic indexes tend to have less predictive power.
3. Publicly available GWAS results are sometimes based on a sample that includes the dataset (or close relatives of dataset members) in which the researcher wants to analyze the polygenic index. Such "sample overlap" spuriously inflates the predictive power of the polygenic index, which can lead to highly misleading results.

4. Because different researchers construct polygenic indexes in different ways, it is hard to compare and interpret results from different studies (see FAQ 1.8)

As we explain in the paper:

> We overcome #1 by constructing the [polygenic indexes] ourselves and releasing them to the data providers, who in turn will make them available to researchers. This simultaneously addresses #2 because we use all the data available to us that may not be easily available to other researchers or to the data providers, including genome-wide summary statistics from 23andMe. Using these genome-wide summary statistics from 23andMe is what primarily distinguishes our Repository from existing efforts by data providers to construct PGIs and make them available…It also distinguishes our Repository from efforts to make publicly available [polygenic index] weights directly available for download (although we also do that, for weights constructed without 23andMe data). To deal with #3, for each [outcome] and each dataset, we construct a [polygenic index] from GWAS summary statistics that excludes that dataset. We overcome #4 by using a uniform methodology across the [outcomes].

In addition to providing polygenic indexes constructed using a uniform methodology (which deals with problem #1 listed in FAQ 1.8), we aim to improve comparability of results based on polygenic indexes in another way (which deals with problems #2 and #3 listed in FAQ 1.8): we derive a "measurement-error-corrected estimator" and provide software for calculating it. This estimator deals with the fact that polygenic indexes can differ from each other in their signal-to-noise ratios. It estimates what the results of an analysis would be if the polygenic index had no noise. It thereby avoids the distortions in results that arise from having a noisy measure. Because it puts results about the polygenic index in the units of the "noiseless" polygenic index, the results from polygenic indexes with different signal-to-noise ratios are expressed in the same units. For more details, see FAQ 2.4.

2. Study Design and Results

2.1. What outcomes are included in the Polygenic Index Repository? How did you choose the outcomes?

We constructed polygenic indexes for 47 outcomes in 11 datasets, using a consistent methodology. The outcomes (listed in Table 1 in the paper) can be categorized into five groups:

- anthropometric (height and body mass index);
- cognition and education (including number of years of formal schooling and performance on cognitive tests);
- fertility and sexual development (including number of children separately for men and women, and age at first menses);
- health and health behaviors (the largest category, which includes self-rated overall health, several alcohol and smoking-related behaviors, and depressive symptoms); and
- personality and well-being (the next largest category, which includes self-rated risk tolerance, subjective well-being, and adventurousness).

The set of 47 outcomes we studied was selected from a larger set of 53 outcomes; we did not create polygenic indexes for the 6 outcomes for which statistical calculations indicated that, based on the GWAS results we had available, a polygenic index was predicted to explain less than 1% of the variation across individuals. Although the specific threshold of 1% is somewhat arbitrary (but see further discussion in FAQ 2.3 below), polygenic indexes with low predictive power are less useful and more likely to generate misleading results (such as false positives) if used.

2.2. How did you create these polygenic indexes?

In order to construct the polygenic indexes, we combined GWAS results from three sources. First, for the 34 outcomes where we could find previously published GWAS, we obtained the publicly available results. Second, we collaborated with the personal genomics company 23andMe. 23andMe contributes to academic research by analyzing the data of customers who consent to participate in research. For this paper, 23andMe provided GWAS results for 37 outcomes, 9 of which had not previously been published. Third, for 25 outcomes, we conducted a GWAS ourselves in the UK Biobank, a large-scale biomedical database accessible to researchers. When more than one of these sources of GWAS results was available for an outcome, we combined the GWAS results together using a statistical method called meta-analysis. In some cases, we constructed "multi-trait polygenic indexes" using GWAS results for multiple outcomes (Turley et al., 2018); these polygenic indexes are often more predictive than a standard "single-trait polygenic index" constructed from GWAS results from a single outcome (FAQ 1.3), but the results from analyzing multi-trait polygenic indexes are sometimes more difficult to interpret (FAQ 2.5).

2.3. How predictive are the polygenic indexes in the Repository?

To assess the predictive power of the polygenic indexes, we used data from 5 of the 11 participating datasets (those for which we had access to both the outcome and genotype data we needed to construct the polygenic indexes). In each of these 5 datasets, we calculated the predictive power of every polygenic index for which the dataset contained data on the relevant outcome (see FAQ 2.1).

The predictive power of the polygenic indexes varies substantially across the outcomes and validation datasets. The polygenic index for height has the greatest predictive power. It predicts 26% to 34% of the variation across individuals, depending on the validation dataset. Next is the polygenic index for body mass index (BMI), whose predictive power ranges from 13% to 15% in our validation datasets. Several outcomes—cognitive performance, age at first menses, and educational attainment—have a polygenic index with predictive power in the range of 6% to 12%. Among the least predictive are the polygenic indexes for satisfaction with family and satisfaction with friendships, whose predictive powers in our validation datasets range from 0.3% to 0.7% (they were included because their predictive power was statistically expected to exceed 1%; see FAQ 2.1). The predictive powers for the other polygenic indexes in the Repository lie somewhere between 1% and 6%.

Although the effects explained by these polygenic indexes are small-to-modest, they can nevertheless be useful in research. For instance, the environmental factors studied in economics research typically have predictive power smaller than 5%, often 1% or smaller. Among the strongest predictors of educational attainment is family socioeconomic status, which has predictive power of roughly 15%. In

a standard categorization used in psychology (Cohen, 1992; percentages here are squared *r* values) predictive power less than 9% is "small" while predictive power greater than 25% (rarely attained in psychological research) is "large." We caution, however, that these comparisons of the effect sizes of polygenic indexes and environmental influences aren't apples-to-apples because researchers usually study one particular environmental factor or many on an outcome, whereas a polygenic index summarizes the predictive power of SNPs across the genome. As discussed further in FAQ 3.3, for social and behavioral outcomes, the sum of all environmental (i.e., non-genetic) influences substantially outweigh the sum of all genetic influences that a polygenic index aims to capture.

As we discuss in FAQ 3.3, an individual's polygenic indexes (even for height) do *not* very accurately predict that individual's outcomes. However, polygenic indexes are useful for *scientific studies* (including social science, health research, etc.). Such studies are concerned with aggregate population trends and averages rather than with individual outcomes. For example, for a polygenic index that predicts 1% of the variation across individuals, studies of its association with other variables can be well powered in sample sizes as small as 785 individuals; 10 out of the 11 datasets participating in the Repository have sample sizes larger than that.

A major goal of the Polygenic Index Repository is to enable other research that is valuable to social scientists and health researchers. Such studies are already being conducted with some polygenic indexes (see FAQ 1.9). For some outcomes, the polygenic indexes in the Repository are more predictive than those that were previously possible to construct; examples include having asthma/eczema/rhinitis, number of cigarettes smoked per day, having migraines, nearsightedness, self-reported physical activity, self-rated overall health, extraversion (i.e., being outgoing), and subjective well-being (i.e., self-reported happiness or life satisfaction). For other outcomes, polygenic indexes were not available prior to this paper because there had been no large published GWASs for those outcomes; examples include childhood reading, self-rated math ability, and self-reported narcissism, and several allergies including to pollen.

2.4. What is the "measurement-error-corrected estimator"? How will it and the Repository improve comparability of results across future studies?

To understand this tool, it's helpful to imagine the theoretically ideal polygenic index that could result from an infinitely large GWAS. In the paper, we call the predictor that would result from this ideal GWAS the "additive SNP factor." The actual polygenic indexes that exist in the world are "noisy" measures of, and therefore only proxies for, this additive SNP factor. The signal-to-noise ratio of a polygenic index—i.e., the extent to which it reflects the additive SNP factor—is determined by the sample size of the GWAS from which the polygenic index is constructed (a larger GWAS leads to less noise and therefore a higher signal-to-noise ratio). The fact that the polygenic index is noisy distorts the results of most analyses that use the polygenic index (relative to what the results would be with the ideal predictor). These distortions can lead researchers to reach incorrect conclusions. For example, in an analysis of how genes and environments interact in influencing some outcome, the noise in the polygenic index will usually cause a researcher to underestimate how strongly genes and environments interact.

Moreover, as discussed in FAQ 1.8, there are many reasons why two polygenic indexes for the same outcome could differ from each other, including differences in the GWAS that the polygenic index is

based on and different methods for constructing the polygenic index. Many of these differences among GWASs produce differences in the signal-to-noise ratios of their resulting polygenic indexes. Two studies using polygenic indexes with different signal-to-noise ratios will, in turn, have results that are distorted to differing degrees, reducing comparability of results across studies that use the polygenic indexes.

The "measurement-error-corrected estimator" we derive in the paper enables researchers to conduct analyses *without* the distortion that comes from the noise. It works because we (often) have a good estimate of how much noise a given polygenic index has. We can use that information to calculate what the results of an analysis would have been if the polygenic index had no noise. The estimator improves comparability of results across papers because it avoids the distortions in results that arise from having a noisy polygenic index. Rather than being distorted to different degrees, two studies using polygenic indexes with different signal-to-noise ratios that use our estimator will both have undistorted results. We have made available the software for this estimator. We will maintain and provide user support for this software.

Moreover, across all the polygenic indexes and across all the datasets participating in the Repository, we constructed the polygenic indexes in a uniform way. To the extent that future studies use the polygenic indexes from the Repository, their results will therefore be more comparable.

2.5. What is in the User Guide that accompanies the Repository?

Along with the polygenic indexes, we have distributed to the participating datasets a User Guide. Data providers will distribute this User Guide to researchers as part of the Repository. The User Guide contains technical details about the construction of the polygenic indexes, as well as details about data and software availability. It also describes a set of key interpretational considerations that researchers should keep in mind when analyzing polygenic indexes. These include when to use a single-trait versus multi-trait polygenic index (see FAQ 2.1) and reasons why associations between a polygenic index and an outcome generally cannot be interpreted as causal (see FAQ 1.5). Finally, the User Guide contains a discussion of six "interpretational considerations" that we urge researchers who use polygenic indexes to consider as part of the responsible conduct and communication of their research (see FAQ 3.7).

2.6. Who can access the Repository polygenic indexes, and how?

Researchers can access the Repository polygenic indexes through the data access procedures for each of the datasets participating in the Repository. These are summarized in the Supplementary Note of the paper. Typically, data providers require researchers to submit a brief a description of the planned research and to sign a Data Use Agreement. The Data Use Agreement usually requires researchers to agree to protect the confidentiality of individuals in the dataset and, to that end, to analyze the data on computers that satisfy certain security protocols.

We provided the polygenic indexes we created to the 11 datasets participating in the Repository, so that the data providers can distribute them to users of the datasets. We designed the Repository this way for three reasons (corresponding to problems #1, #2, and #3 in FAQ 1.9; problem #4 is addressed by using

a consistent methodology for constructing the polygenic indexes). First, because we are making available the polygenic indexes (rather than the GWAS results from which they are constructed), researchers do not need to spend time constructing the polygenic indexes from GWAS results. Second, for many outcomes, the polygenic indexes we construct are based on more data than are in the largest previously published GWAS. Because the Repository polygenic indexes for those outcomes are based on more data, they are more accurate (i.e., predictive) than polygenic indexes that could be constructed based only on publicly available GWAS results. Third, we tailored the polygenic indexes we constructed to each of the 11 datasets. Specifically, we ensured that for a given dataset, its polygenic indexes were *not* based on GWAS results that included that dataset (which would have led to "sample overlap" that would make it problematic to use the polygenic index with that dataset).

2.7. How will the Repository be updated?

We plan to update the Repository regularly as new GWAS are published or new data become available in which we can conduct our own GWAS. The updates will increase the predictive power of polygenic indexes already in the Repository, as well as expand the set of outcomes for which polygenic indexes are available. We also expect to include additional datasets whose stewards want to participate in the Repository and make their data broadly available to the research community.

3. Ethical and social implications of the study

3.1. Do GWAS or the polygenic indexes they produce identify the gene—or genes—"for" a particular outcome?

No. GWAS of complex outcomes identify *many* SNPs that are associated with an outcome like height or educational attainment. Although it was once believed that scientists would discover numerous strong one-to-one associations between specific genes and outcomes, we have known for a number of years that the vast majority of human traits and other outcomes are complex and are influenced by thousands of genes, each of which alone tends to have a small influence on the relevant outcome.

Furthermore, many complex outcomes are also influenced by parts of the genome that are not genes at all but instead serve to regulate genes (e.g., influencing when a gene is turned on or off). Genes typically contain many SNPs (often dozens or hundreds, in some cases thousands), and there are even more SNPs outside of genes than inside genes. Complex outcomes are often influenced by millions of SNPs.

Although the GWAS that produced the polygenic indexes included in the Repository did find several SNPs that are associated with particular outcomes, we believe that characterizing these as "genes for X"—or, more accurately—"SNPs for X" (e.g., educational attainment, height) is still likely to mislead, for many reasons, and we urge researchers and reporters to avoid this usage.

As an example, consider the outcome of educational attainment. First, most of the variation in people's educational attainment is accounted for by social and other environmental factors, not by additive

genetic effects (See FAQ 3.3). "Genes for educational attainment" might be read to imply, incorrectly, that genes are the strongest predictor of variation in educational attainment.

Second, the SNPs that are associated with educational attainment are also associated with many other things. These SNPs are no more "for" educational attainment than for the other outcomes with which they are associated.

Third, the "predictive" power (see FAQ 1.6) of each individual SNP that we identify is very small. Our previous work (Lee et al., 2018) has shown that genetic associations with educational attainment are comprised of thousands, or even millions, of SNPs, each of which has a tiny effect size. Each SNP is therefore weakly associated with, rather than a strong influence on, educational attainment. "Genes for educational attainment" might misleadingly imply the latter.

Fourth, environmental factors can increase or decrease the impact of specific SNPs (see FAQ 3.3). Put differently, even if a SNP is associated with higher or lower levels of educational attainment *on average*, it may have a much larger or smaller effect depending on environmental conditions. Indeed, in our most recent GWAS of educational attainment (Lee et al., 2018) and elsewhere, we report exploratory analyses that provide evidence of such gene-environment interactions. Educational attainment couldn't even exist as a meaningful object of measurement if we didn't have schools, and having schools introduces societal mechanisms that influence who goes to them. Accordingly, genetic associations with educational attainment necessarily will be mediated by societal systems and therefore genetic variation should often be expected to interact with environmental factors when it influences social phenomena, such as educational attainment. "Genes for educational attainment" suggests a stability in the relationship between these genes and the outcome of educational attainment that does not exist.

Finally, SNPs do not affect educational attainment directly. As described in our previous work (Lee et al., 2018), the genes identified as associated with educational attainment tend to be especially active in the brain and involved in neural development and neuron-to-neuron communication. The "predictive" power (see FAQ 1.6) of SNPs on educational attainment may therefore be the result of a long process starting with brain development, followed by the emergence of particular psychological traits (e.g., cognitive abilities and personality). These traits may then lead to behavioral tendencies as well as experiences and treatment by parents, peers, and teachers. All of these factors may additionally interact with the environment in which a person lives. Eventually these traits, behaviors, and experiences may influence (but not completely determine) educational attainment.

3.2. Do polygenic indexes show that these outcomes are determined, or fixed, at conception?

Absolutely not. Social and other environmental factors account for most variation in most of the outcomes for which the Repository contains polygenic indexes. But even if it were true that genetic factors accounted for *all* of the differences among individuals in an outcome, it would *still* not follow that an individual's outcome is "determined" at conception. There are at least three reasons for this.

First, some genetic effects may operate through environmental channels (Jencks, 1980). Again, consider educational attainment as an example. Suppose—hypothetically— that some of the SNPs in the index help students to memorize and, as a result, to become better at taking tests that rely on memorization. In this example, changes to the intermediate environmental channels—the type of tests administered in schools—could have large effects on individuals' educational attainment, even though individuals' genome would not have changed. Certain SNPs may not be associated with educational attainment *at all* if schools did not use tests that rely on memorization. More generally, the polygenic index for educational attainment in the Repository might be less predictive if the education system were organized differently than it is at present (see also FAQ 3.3).

Second, even if the genetic associations with educational attainment operated entirely through non-environmental mechanisms that are difficult to modify (such as direct influences on the formation of neurons in the brain and the biochemical interactions among them), there could still exist powerful environmental interventions that could change the genetic relationships. In a famous example suggested by the economist Arthur Goldberger, even if all variation in unaided eyesight were due to genes, there would still be enormous benefits from introducing eyeglasses (Goldberger, 1979). Similarly, policies such as a required minimum number of years of education and dedicated resources for individuals with learning disabilities can increase educational attainment in the entire population and/or reduce differences among individuals.

Third, even if the genetic effects on an outcome were not influenced by changes in the environment, those environmental changes themselves could still have a major impact on the outcome in the population as a whole. For example, if young children were given more nutritious diets, then everyone's school performance might improve, and college graduation rates might increase. Or consider the outcome of height: 80%-90% of the variation across individuals in height is due to genetic factors. Yet the current generation of people is much taller than past generations due to changes in the environment such as improved nutrition.

3.3. Can the polygenic indexes from the Repository be used to accurately predict a particular person's outcomes?

No. While the "predictive" power (see FAQ 1.6) of our polygenic indexes makes most of them useful in research for some purposes (see FAQ 2.3), these polygenic indexes *fail to predict* the majority of variation across individuals. Even for height—the outcome for which our polygenic index has the greatest predictive power—the index fails to predict 70% of the variation.

Indeed, an important message of a number of our earlier papers is that DNA does *not* "determine" an individual's behavioral and social outcomes, for at least four reasons: First, in the environments in which the outcomes have been measured, other studies have estimated that the additive effects of SNPs will only ever account (even with arbitrarily large samples used to construct polygenic indexes) for a minority of the variation across individuals in the outcomes we study. For example, we estimate that the theoretical upper bound for additive effects of SNPs would account for 46% of the variation in height, 24% in body mass index, 20% in age at first menses, and less than 10% for most of the social/behavioral outcomes we study. So even a hypothetical polygenic index that perfectly reflects the additive SNP factor (see FAQ 2.4) could only explain a small fraction of the variation across

individuals. Second, *today's* polygenic indexes are *not* perfect; they are only able to predict a fraction of that already small fraction of cross-sectional predictive power. Third, since SNPs matter more or less depending on environmental context (see FAQ 3.2), a polygenic index might be less (or more) predictive for individuals in some environments than for individuals in others. Finally, and similarly, polygenic predictions only hold for as long as the environment in which they were developed remains substantially the same.

To illustrate these final two reasons, consider the example of educational attainment (for which we have included a polygenic index in the Repository and on which we have done previous research): if the pedagogy underlying the educational system in which the GWAS that produced the polygenic index was conducted is substantially different than the pedagogy of the *different population* to which that polygenic index is being applied, the polygenic index may be less (or, conceivably, more) predictive in this second population (for an example, see FAQ 3.2). The same is true if the polygenic index is applied to the same population, but at a *later time* when the pedagogy has changed substantially. Just as eyeglasses allow those genetically predisposed to poor vision to have nearly perfect vision, innovations in education (say, an innovation that makes education irresistibly engaging, thus mitigating the risk to those with SNPs associated with lower ability to pay attention or maintain self-control) might result in those with lower polygenic indexes now achieving just as much education, on average, as those with higher polygenic indexes.

As sample sizes for GWAS continue to grow, it will likely be possible to construct polygenic indexes for many outcomes whose predictive power comes closer to the total amount of variation that is theoretically predictable from additive effects of common SNPs for those outcomes (the upper bounds given above). Even these levels of predictive power would pale in comparison to some other scientific predictors. For example, professional weather forecasts correctly predict about 95% of the variation in day-to-day temperatures. Weather forecasters are therefore vastly more accurate forecasters than social science geneticists will ever be.

Note: Polygenic indexes created by GWASs are increasingly used by commercial and research direct-to-consumer platforms to predict individual outcomes. We recognize that returning individual genomic "results" can be a fun way to engage people in research and other projects and has at least the theoretical potential to stoke their interest in, and educate them about, genomics and how genes and environments interact. But it is important that participants/users understand that, at present, most of these individual results, including all social and behavioral outcomes, are not *meaningful* predictions (in the sense that they generally have very little predictive power at the individual level). Failure to make this point clear risks sowing confusion and undermining trust in genetics research.

3.4. Can the polygenic indexes accurately be used for research studies in non-European-ancestry populations?

No. We constructed polygenic indexes only for individuals classified as "European ancestry." (The precise definition of "European ancestry" differs in different datasets, but it usually means that a person's pattern of genetic variation across the genome is statistically close to the average pattern from a "reference sample" for some European country. The reference samples used by geneticists are based on samples of people who live in the European country today and whose recent ancestors also lived in

that country.) Therefore, the Polygenic Index Repository only includes polygenic indexes for these individuals.

The main reason we only constructed polygenic indexes for these individuals is that the polygenic indexes are likely to be much less predictive—and hence much less useful—in a sample of people of non-European ancestries. That is because our original GWAS data was obtained from samples of people with European-ancestry, and GWAS results have been found to have only limited portability across ancestries (Belsky et al., 2013; Domingue et al., 2015, 2017; Martin et al., 2017; Vassos et al., 2017). There are a number of reasons for the limited portability. For one thing, the set of SNPs that are associated with an outcome in people of European ancestries is unlikely to overlap closely with the set of SNPs associated with the outcome in people of non-European ancestries. And even if a given SNP is associated in both ancestry groups, the effect size—in other words, the strength of the association—will almost surely differ. This is primarily because linkage disequilibrium (LD) patterns (i.e., the correlation structure of the genome) vary by ancestry. This means that some SNP may be associated with the outcome because the SNP is in LD (i.e., correlated) with a SNP elsewhere in the genome that causally affects education (see FAQ 1.5). If the strength of the correlation is greater in one ancestry group than in another, then the size of the association will be larger in that ancestry group. Moreover, even if LD patterns were similar in each ancestry group, the association may differ in different groups because environmental conditions differ (see FAQ 1.6). The fact that there are differences across ancestry groups in the set of associated SNPs and their effect sizes means that the weights for constructing polygenic indexes in European-ancestry individuals (FAQ 1.3) would be the "wrong" weights for non-European-ancestry individuals. For a more extensive, excellent discussion of these and related issues, see Graham Coop's blog post, "Polygenic scores and tea drinking."

Unfortunately, this attenuation of predictive power means that for non-European-ancestry populations, many of the benefits of having a polygenic index available will have to wait until large GWAS studies are conducted using samples from these populations. (Currently, most large genotyped samples are of European ancestries.) We intend that future versions of the Polygenic Index Repository will include polygenic indexes for non-European-ancestry populations, once it becomes possible to produce polygenic indexes with adequate predictive power. We believe that the relative scarcity of polygenic indexes that can be used for research that focuses on non-European ancestry groups is a disparity that should be rapidly eliminated by prioritizing GWAS studies that focus on non-European populations.

3.5. Would it be appropriate to use the Repository social and behavioral polygenic indexes in policy or practice?

No. We reiterate that polygenic indexes are poor predictors of social and behavioral outcomes (see FAQs 2.3 and 3.3). Their *incremental* predictive power over and above other, non-genetic predictors that are already used is even smaller than a polygenic index's predictive power on its own. Moreover, the predictive power of the polygenic indexes for social and behavioral outcomes depends on the environment in which the GWAS participants live (FAQ 3.3). Thus, enshrining polygenic indexes in policy risks basing policy (which can be difficult to change) on weak predictions that could become even weaker or nonexistent as the environment changes. Furthermore, the polygenic indexes can operate through environmental channels (FAQ 3.2). Allocating resources based on polygenic indexes could therefore exacerbate inequalities that were originally due to environmental disparities (a similar

risk to that of other biased algorithms that bake in pre-existing discrimination). Using polygenic indexes in order to prioritize giving resources to individuals who are already advantaged would further limit the opportunities of individuals who are disadvantaged, which would be ethically inappropriate. Finally, even if polygenic indexes were used to offer additional resources to disadvantaged individuals, any small potential benefits of using such weak individual predictors would almost certainly be offset by the risk of stigmatization and by the fact that this technology is currently only accessible to people of European ancestries (FAQ 3.4). For all these reasons, we are deeply skeptical that the Repository social and behavioral polygenic indexes have any appropriate role to play in policy now or in the foreseeable future.

3.6. Could research on polygenic indexes lead to discrimination against, or stigmatization of, people with higher or lower polygenic indexes for certain outcomes? If so, why facilitate the spread of polygenic indexes?

Unfortunately, like a great deal of research—including, for instance, research identifying genomic variation associated with increased cancer risk—the results can be misunderstood and misapplied. This includes being used to discriminate against those with higher or lower polygenic indexes for certain outcomes (e.g., in insurance markets). Nevertheless, for a variety of reasons, in this instance, we do not think that the best response to the possibility that useful knowledge could be misused is to refrain from producing the knowledge. Moreover, many researchers already have access to and use polygenic indexes; against this background, the Repository helps ensure that a much wider array of researchers have the same opportunity to access and probe these research tools, and also that the polygenic indexes themselves will be more accurate. Here, we briefly discuss some of the broad potential benefits of this research. We then describe what we see as our ethical duty as researchers conducting this work.

First, one benefit of conducting social-science genetics research in ever larger samples is that doing so allows us to correct the scientific record. An important theme in our earlier work has been to point out that most existing studies in social-science genetics that report genetic associations with behavioral outcomes have serious methodological limitations, fail to replicate, and are likely to be false-positive findings (Benjamin et al., 2012; Chabris et al., 2012, 2015). This same point was made in an editorial in *Behavior Genetics* (the leading journal for the genetics of behavioral outcomes), which stated that "it now seems likely that many of the published [behavior genetics] findings of the last decade are wrong or misleading and have not contributed to real advances in knowledge" (Hewitt, 2012). One of the most important reasons why earlier work has generated unreliable results is that the sample sizes were far too small, given that the true effects of individual SNPs on behavioral outcomes are tiny. Pre-existing claims of genetic associations with complex social-science outcomes have reported widely varying effect sizes, many of them purporting to "predict" as much of the variation across individuals as do the polygenic indexes we construct in this paper that aggregate the effects of millions of SNPs.

Second, behavioral genetics research also has the potential to correct the *social* record and thereby to help *combat* discrimination and stigmatization. For instance, overestimating the role of genetics can be damaging, and the present work can help debunk the myth of genetic determinism. By quantifying how various outcomes are predicted by genetic data, we show that for all of the outcomes we study, the genetic data can explain a very small fraction of the variation across individuals (see FAQ 2.3). By clarifying the *limits* of deterministic views of complex outcomes, recent behavioral genetics research—

if communicated responsibly—could make appeals to genetic justifications for discrimination and stigmatization *less* persuasive to the public in the future.

Third, behavioral genetics research has the potential to yield many other benefits, especially as sample sizes continue to increase—as briefly summarized in FAQ 1.9. Foregoing this research necessarily entails foregoing these and any other possible benefits, some of which will likely be the result of serendipity. Indeed, very few of the uses of polygenic indexes were anticipated when they were first proposed (Wray, Goddard and Visscher, 2007).

In sum, we agree with the U.K. Nuffield Council on Bioethics, which concluded in a report (Nuffield Council on Bioethics, 2002, p114) that "research in behavioural genetics has the potential to advance our understanding of human behaviour and that the research can therefore be justified," but that "researchers and those who report research have a duty to communicate findings in a responsible manner" (see FAQ 3.7).

3.7. What have you done to mitigate the risks of research using Repository polygenic indexes?

In our view, the responsible behavioral genetics research called for by the Nuffield Council on Bioethics (see FAQ 3.6) includes sound methodology and analysis of data (e.g., only conducting analyses that are adequately powered and, when feasible, preregistering power calculations and planned analyses); a commitment to publish all results, including any negative results; and transparent, complete reporting of methodology and findings in publications, presentations, and communications with the media and the public. A critical aspect of the latter is particular vigilance regarding what research results do—and do not—show, and how polygenic indexes can—and cannot—be appropriately used. In an effort to reduce the risk that its results might be misinterpreted by readers, misreported by the media, or misused, the SSGAC has developed and publicly posted FAQs like this document with every major paper it has published since its first paper in 2013.

In addition, the SSGAC will require researchers who download the SNP weights for constructing polygenic indexes to agree to Terms of Service. Among the many terms that we require researchers to agree to, we highlight two here:

> I agree to conduct research that strictly adheres to the principles articulated by the American Society of Human Genetics (ASHG) position statement: "ASHG Denounces Attempts to Link Genetics and Racial Supremacy." (See also International Genetic Epidemiological Society Statement on Racism and Genetic Epidemiology.) **In particular, I will not use these data to make comparisons across ancestral groups.** Such comparisons could animate biological conceptualizations of racial superiority. In addition, such comparisons are usually scientifically confounded due to the effects of linkage disequilibrium, gene-environment correlation, gene-environment interactions, and other methodological problems.
>
> I have read the principles articulated by the ASHG with respect to "Advancing Diverse Participation in Research with Special Consideration for Vulnerable Populations". I agree to adhere to the principles articulated in the final two sections of this statement, "In the Conduct

of Research with Vulnerable Populations, Researchers Must Address Concerns that Participation May Lead to Group Harm" and "The Benefits of Research Participation Are Profound, Yet the Potential Danger that Unethical Application of Genetics Might Stigmatize, Discriminate against, or Persecute Vulnerable Populations Persists."

These Terms of Service stem from the observation that SNP associations are not necessarily causal (see FAQ 1.5) and depend on the environment of the individuals included in the GWAS (see FAQ 1.6). Different ancestry groups arise in the population because they became partially separated from each other many generations ago, for example, due to geographic factors or social forces. When two groups are geographically or socially separated, they also face different environments, which not only may have direct effects on certain outcomes (such as disease risk) but may also change the strength of the association between the outcomes and certain SNPs. Therefore, when individuals from two ancestry groups have different average outcomes, it is extremely difficult to identify whether the difference is due to average genetic differences between the groups or to the different environments faced by the groups. For this reason, it is scientifically invalid to make general statements about ancestry group differences based on SNP associations identified in a GWAS. (Also see FAQ 3.2.) The Terms of Service also require users to securely store the data and to immediately report any breach of the Terms.

Finally, we have developed and provided to participating data providers a User Guide to be distributed to researchers who use Repository polygenic indexes (see FAQ 2.5). We will also provide the User Guide to researchers who download the SNP weights. One section of the User Guide discusses six "interpretational considerations" that are likely to arise when conducting research with polygenic indexes and which we urge researchers to seriously consider as a critical part of responsibly conducting and communicating their research. One recurring ethical concern about genetic research is the tendency for its predictive power to become exaggerated in the media and in the public's minds, at the expense of a more nuanced understanding of how genes and environment interact, the importance of environmental influences, and the ability of interventions to improve outcomes. Many of the interpretational considerations we discuss in the User Guide involve how to anticipate and address potential confounds and how to navigate complex questions about causality and ensure responsible communication of causality.

For instance, the User Guide cautions researchers to appreciate and communicate that associations between a polygenic index and an outcome may operate through *environmental* (rather than biological) mechanisms (see FAQs 3.2 and 3.3).

4. References

Abdellaoui, A. et al. (2019). Genetic correlates of social stratification in Great Britain. *Nature Human Behaviour*, 3 (12), 1332–1342. Available from https://doi.org/10.1038/s41562-019-0757-5.

Amos, C.I. et al. (2008). Genome-wide association scan of tag SNPs identifies a susceptibility locus for lung cancer at 15q25.1. *Nature Genetics*, 40, 616–622. Available from https://doi.org/10.1038/ng.109.

Barcellos, S.H., Carvalho, L.S. and Turley, P. (2018a). Education can reduce health differences related to genetic risk of obesity. *Proceedings of the National Academy of Sciences*, 115 (42), E9765. Available from https://doi.org/10.1073/pnas.1802909115.

Barcellos, S.H., Carvalho, L.S. and Turley, P. (2018b). Education can Reduce Health Disparities Related to Genetic Risk of Obesity: Evidence from a British Reform. *bioRxiv [https://doi.org/10.1101/260463]*. Available from https://doi.org/10.1101/260463.

Belsky, D.W. et al. (2013). Development and evaluation of a genetic risk score for obesity. *Biodemography and Social Biology*, 59 (1), 85–100. Available from https://doi.org/10.1080/19485565.2013.774628.

Belsky, D.W. et al. (2016). The Genetics of Success. *Psychological Science*, 27 (7), 957–972. Available from https://doi.org/10.1177/0956797616643070.

Benjamin, D.J. et al. (2012). The Promises and Pitfalls of Genoeconomics. *Annual Review of Economics*, 4 (1), 627–662. Available from https://doi.org/10.1146/annurev-economics-080511-110939.

Chabris, C.F. et al. (2012). Most reported genetic associations with general intelligence are probably false positives. *Psychological Science*, 23 (11), 1314–1323. Available from https://doi.org/10.1177/0956797611435528.

Chabris, C.F. et al. (2015). The Fourth Law of Behavior Genetics. *Current Directions in Psychological Science*, 24 (4), 304–312. Available from https://doi.org/10.1177/0963721415580430.

Cohen, J. (1992). Statistical Power Analysis. *Current Directions in Psychological Science*, 1 (3), 98–101. Available from https://doi.org/10.1111/1467-8721.ep10768783.

Davies, N.M. et al. (2018). The causal effects of education on health outcomes in the UK Biobank. *Nature Human Behaviour*. Available from https://doi.org/10.1038/s41562-017-0279-y.

Domingue, B.W. et al. (2015). Polygenic Influence on Educational Attainment: New evidence from The National Longitudinal Study of Adolescent to Adult Health. *AERA Open*, 1 (3), 1–13. Available from https://doi.org/10.1177/2332858415599972.

Domingue, B.W. et al. (2017). Mortality selection in a genetic sample and implications for association studies. *International Journal of Epidemiology*, 46 (4), 1285–1294. Available from

https://doi.org/10.1093/ije/dyx041.

Domingue, B.W. et al. (2018). Geographic Clustering of Polygenic Scores at Different Stages of the Life Course. *RSF: The Russell Sage Foundation Journal of the Social Sciences*, 4 (4), 137 LP – 149. Available from https://doi.org/10.7758/RSF.2018.4.4.08.

Goldberger, A.S.A. (1979). Heritability. *Economica*, 46 (184), 327–347.

Hewitt, J.K. (2012). Editorial policy on candidate gene association and candidate gene-by-environment interaction studies of complex traits. *Behavior Genetics*, 42 (1), 1–2. Available from https://doi.org/10.1007/s10519-011-9504-z.

Howe, L.J. et al. (2021). Within-sibship GWAS improve estimates of direct genetic effects. *bioRxiv*, 2021.03.05.433935. Available from https://doi.org/10.1101/2021.03.05.433935.

Hung, R.J. et al. (2008). A susceptibility locus for lung cancer maps to nicotinic acetylcholine receptor subunit genes on 15q25. *Nature*. Available from https://doi.org/10.1038/nature06885.

Jencks, C. (1980). Heredity, environment, and public policy reconsidered. *American Sociological Review*, 45 (5), 723–736.

Khera, A. V. et al. (2018). Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations. *Nature Genetics*, 50 (9), 1219–1224. Available from https://doi.org/10.1038/s41588-018-0183-z.

Khera, A. V et al. (2019). Polygenic Prediction of Weight and Obesity Trajectories from Birth to Adulthood. *Cell*, 177 (3), 587-596.e9. Available from https://doi.org/10.1016/j.cell.2019.03.028.

Koellinger, P.D. and Harden, K.P. (2018). Using nature to understand nurture: Genetic associations show how parenting matters for children's education. *Science*, 359 (6374), 386–387. Available from https://doi.org/10.1126/science.aar6429.

Kong, A. et al. (2018). The nature of nurture: Effects of parental genotypes. *Science*, 359 (6374), 424–428. Available from https://doi.org/10.1126/science.aan6877.

Lambert, S.A. et al. (2020). The Polygenic Score Catalog: an open database for reproducibility and systematic evaluation. *medRxiv*, 2020.05.20.20108217. Available from https://doi.org/10.1101/2020.05.20.20108217.

Lander, E.S. and Schork, N.J. (1994). Genetic dissection of complex traits. *Science*, 265, 2037–48.

Lee, J.J. et al. (2018). Gene discovery and polygenic prediction from a genome-wide association study of educational attainment in 1.1 million individuals. *Nature Genetics*, 50 (8), 1112–1121. Available from https://doi.org/10.1038/s41588-018-0147-3.

Martin, A.R. et al. (2017). Human Demographic History Impacts Genetic Risk Prediction across Diverse Populations. *American Journal of Human Genetics*, 100 (4), 635–649. Available from

https://doi.org/10.1016/j.ajhg.2017.03.004.

Nuffield Council on Bioethics. (2002). *Genetics and human behaviour: the ethical context*. London: Nuffield Council on Bioethics [http://nuffieldbioethics.org/wp-content/uploads/2014/07/Genetics-and-human-behaviour.pdf].

Purcell, S.M. et al. (2009). Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. *Nature*, 460 (7256), 748–752. Available from https://doi.org/10.1038/nature08185.

Robinson, M.R. et al. (2017). Genetic evidence of assortative mating in humans. *Nature Human Behaviour*. Available from https://doi.org/10.1038/s41562-016-0016.

Schmitz, L.L. and Conley, D. (2017). The effect of Vietnam-era conscription and genetic potential for educational attainment on schooling outcomes. *Economics of Education Review*, 61, 85–97. Available from https://doi.org/https://doi.org/10.1016/j.econedurev.2017.10.001.

Thorgeirsson, T.E. et al. (2008). A variant associated with nicotine dependence, lung cancer and peripheral arterial disease. *Nature*, 452 (7187), 638–642. Available from https://doi.org/10.1038/nature06846.

Turkheimer, E. (2000). Three laws of behavior genetics and what they mean. *Current Directions in Psychological Science*, 9 (5), 160–164.

Turley, P. et al. (2018). Multi-trait analysis of genome-wide association summary statistics using MTAG. *Nature Genetics*, 50 (2), 229–237. Available from https://doi.org/10.1101/118810.

Vassos, E. et al. (2017). An Examination of Polygenic Score Risk Prediction in Individuals With First-Episode Psychosis. *Biological Psychiatry*, 81 (6), 470–477. Available from https://doi.org/10.1016/j.biopsych.2016.06.028.

Vilhjálmsson, B.J. et al. (2015). Modeling linkage disequilibrium increases accuracy of polygenicrisk scores. *The American Journal of Human Genetics*, 97 (4), 576–592.

Visscher, P.M. et al. (2017). 10 Years of GWAS Discovery: Biology, Function, and Translation. *American Journal of Human Genetics*, 101 (1), 5–22. Available from https://doi.org/10.1016/j.ajhg.2017.06.005.

Wray, N.R., Goddard, M.E. and Visscher, P.M. (2007). Prediction of individual genetic risk to disease from genome-wide association studies. *Genome research*, 17 (10), 1520–1528. Available from https://doi.org/10.1101/gr.6665407.

Yengo, L. et al. (2018). Imprint of Assortative Mating on the Human Genome. *Nature Human Behaviour*, 2 (12), 2, 948–954. Available from https://doi.org/10.1038/s41562-018-0476-3.