Supplementary Note

Model fitting for ratios of PRSs between LDpred2 and PRS-CT

In the context of GENESIS, which is intended for sample size projections via the CT method, we established a model to describe the relationship between the effective sample size (N_eff) and the phenotypic variance ratios elucidated by CT and LDpred2 PRS. The effective sample size was denoted as $N_{eff} = N_{case} * N_{control}/(N_{case} + N_{control})$, where N_{case} and $N_{control}$ represent the counts of cases and controls, respectively. Our training dataset was systematically downsized to seven distinct sample sizes, derived through combinations of the three EAS studies included in our training data, enabling us to evaluate the performance of PRS-CT and LDpred2 PRS at each data point. This analysis yielded seven data points depicting the variance ratio between LDpred2 PRS and PRS-CT across a spectrum of sample sizes. For accurate characterization, we sought models satisfying two conditions: firstly, as N_{eff} increases, the ratio of phenotypic variance explained by the two methods will converge to 1, given that both PRSs approach the heritability of the genetic effects. Secondly, this ratio should inversely correlate with the increase in sample size. Additionally, we assumed that beyond an N_{eff} of 50,000, the variance ratio between PRS-CT and LDpred2 PRS would converge to 1.

We examined five mathematical models to encapsulate this relationship: exponential decay, power law, logistic, Gompertz, and Weibull functions. Defining y as the ratio of phenotypic variance between LDpred2 and CT PRS, and n as the EAS population's effective sample size, the models are formulated as follows:

Exponential decay:

$$y = ae^{-bn} + c,$$

Power law:

$$y = an^{-b} + 1,$$

Logistic:

$$y = \frac{a}{1 + e^{-b(n-c)}} + (1 - a),$$

Geompertz:

$$y = \frac{a}{e^{-be^{n-c}}} + 1,$$

Weibull:

$$y = ae^{-bn^c} + 1,$$

Here, a, b, and c represent the parameters to estimate for each model. To determine the best fit, we calculated the coefficient of determination (R^2) for each model:

$$R^2 = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \overline{y})^2},$$

Here $\overline{y}$ is the mean of the observed values y_i. A model with a higher R^2 indicates a more accurate representation of the data.

Each model was fitted to the data points using nonlinear least squares regression, utilizing the Levenberg-Marquardt algorithm as implemented in the 'minpack.lm' package[1] within R. We selected initial parameter estimates based on the observed range of effective sample sizes and phenotypic variance ratios. The performance of each model was gauged by its R-squared value—the proportion of variance in the observed data that is predictable from the independent variables.

Of the models tested, The Weibull model was identified as the best fit (**Supplementary Figure 1**). The finalized Weibull model was:

$$y = 3.511e^{-0.049n^{0.429}} + 1.$$

This model was used to extrapolate the phenotypic variance of CT PRS to predict the variance for LDpred2. Finally, the projected phenotypic variance was translated into AUC values[2] to estimate the AUC for LDpred2 PRS at different sample sizes.

Proportion of genetic variance explained by LDpred2 PRS across different sample sizes

As described in the last section, we estimated the genetic variance explained by LDpred2 PRS ($\sigma^2_{LDpred2}$) under different sample sizes. This genetic variance is equated with heritability on a frailty scale, premised on the polygenic log-additive model as the fundamental genetic architecture. Specifically, the genetic variance on this scale for all GWAS variants is formulated as $\sigma^2_{GWAS} = var(\sum_{m=1}^{M} \beta_m G_m)$, where G_m is the standardized genotype for the mth SNP, β_m is the true log OR for the mth SNP and M is the total number of causal SNPs within the GWAS variants. To estimate the frailty scale heritability for lung cancer in EAS never-smokers, we used the linkage-disequilibrium score regression[3] to estimate σ^2_{GWAS} using summary statistics from the training dataset along with the provided EAS-specific LDscore derived from the 1000 Genomes Project data. Consequently, the fraction of genetic variance attributable to all GWAS variants that is explained by the LDpred2 PRS is denoted by the ratio $\sigma^2_{LDpred2}/\sigma^2_{GWAS}$.

The conversion of familial risk to genetic variance employed the expression $\lambda_s^2 = exp(\sigma^2)$, where λ_s is the familial risk when a first-order sibling has the disease, and σ^2 is the genetic

variance on frailty-scale[4]. The familial risk of lung cancer in EAS never-smokers was reported as a 1.84-fold increase[5], correspond to a genetic variance σ^2 of 1.22. Thus, the LDpred2 PRS's elucidation of familial risk is quantified by the ratio $\sigma^2_{LDpred2}/\sigma^2$.

References

1. CRAN - Package minpack.lm. https://cran.r-project.org/web/packages/minpack.lm/index.html.
2. Pharoah, P. D. P. *et al.* Polygenic susceptibility to breast cancer and implications for prevention. *Nat Genet* **31**, 33–36 (2002).
3. Bulik-Sullivan, B. *et al.* LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nature Genetics 2015 47:3* **47**, 291–295 (2015).
4. Pharoah, P. D. P. *et al.* Polygenic susceptibility to breast cancer and implications for prevention. *Nat Genet* **31**, 33–36 (2002).
5. Matakidou, A., Eisen, T. & Houlston, R. S. Systematic review of the relationship between family history and lung cancer risk. *Br J Cancer* **93**, 825 (2005).