Polygenic Prediction of Depression Liability Reveals Broad Genetic Associations Across Development, Families, and Populations

Tobias Wolfram[†,*], Spencer Moore[†], Jonathan Anomaly[†]
Justin Schleede[†], Mohammed Ahangari[†]

[†] *Herasight Research* [*] *Corresponding author*

Abstract

Polygenic prediction of depression is limited by the depth of depression phenotyping available at biobank scale as well as sample size. Here we develop a depression score combining information from depression and genetically correlated psychiatric traits using phenotype reconstruction, multi-trait genetic integration and Bayesian effect estimation. Compared with the strongest publicly available depression-only score (Psychiatric Genomics Consortium 2025, excluding 23andMe), the liability-R^2 for depression in UK Biobank rose by 49.7% and within-family standardized associations with latent internalizing were up to 54% larger. Across 2,328 childhood and adolescent outcomes in the ABCD study, the association profile of the integrated score was almost identical to that of the depression-only score (r = 0.99), while being uniformly stronger. 1,494 outcomes related to mood, behavior, social functioning, hobbies and sleep were associated with the score — 423 of them within-family. In ABCD, adolescents in the highest score decile were 3.5 times as likely as those in the lowest decile to report suicidal thoughts by age 14 (23% versus 7%) and 1.9 times as likely to report nicotine or cannabis use by age 17 (47% versus 25%). In adults, the score predicted worse general health, chronic pain, and life dissatisfaction. Associations with psychiatric symptoms were substantially stronger in girls than in boys (boys' coefficients 0.68 times girls'), particularly for girl-typical symptoms, and showed no moderation by household resources, hardship or deprivation that remained significant after correction for multiple testing. Prediction was attenuated in line with genetic distance from training data for Hispanic and African American children. Depression prediction can thus be strengthened substantially by integrating the internalizing spectrum, without detectable loss of specificity or within-family predictive ability.

1 Introduction

Major depression (MD)[1] is a common psychiatric disorder and a leading cause of disability worldwide. In 2021, depressive disorders affected approximately 332 million people and accounted for 56 million disability-adjusted life years lost (Rong et al., 2025). Twin-study estimates place the heritability of depression liability at approximately 35–40% (Sullivan et al., 2000) and genome-wide association studies (GWAS) have identified almost 700 associations at more than 600 loci throughout the genome linked to it (Adams et al., 2025).

Polygenic scores (PGSs) are constructed from GWAS summary statistics and can be used to predict individual depression risk, explaining up to 5.8% of MD liability variance in Europeans (Adams et al., 2025). PGSs have been used for the prediction of recurrence and psychiatric comorbidity (Als et al., 2023) and to characterize distinct depression trajectories (Grimes et al., 2024), while phenome-wide studies identified broad adult correlates of depression PGSs and adolescent correlates of cross-disorder genetic liability (Paul et al., 2024; Shen et al., 2020). These applications motivate the development of stronger predictors for research on genetic liability and its broader phenotypic expression.

However, depression is heterogeneous in its clinical presentation. Individuals meeting the same diagnostic criteria can differ substantially in their symptom profiles (Fried and Nesse, 2015). Genetic differences have also been identified between early- and late-onset depression (Shorter et al., 2025), while individual depressive symptoms show different associations with environmental stress (Fried et al., 2015). At biobank scale, detailed depression assessment is often available only for a subset of participants, creating a trade-off between phenotyping depth and sample size (Dahl et al., 2023). Alternative definitions of depression can capture partly distinct genetic architectures, with minimal phenotyping, resulting in genome-wide association signals of low specificity (Cai et al., 2020). Combining the available information can result in higher-quality measures, but complete information is available for only a subset of participants. By combining related measurements, phenotype imputation can improve prediction while preserving specificity to major depressive disorder (MDD) (Dahl et al., 2023).

A related issue is the shared genetic architecture of depression and other psychiatric traits. Depression and anxiety are highly genetically correlated (Thorp et al., 2021). Depression, PTSD and anxiety are associated with an internalizing factor in genomic analyses (Grotzinger et al., 2026). Polygenic scores for depression are associated with the general factor of childhood psychopathology (Waszczuk et al., 2023). Such overlaps provide opportunities for multi-trait analysis of GWAS, which together with modern Bayesian methods and functional genomic annotations can be used to improve polygenic prediction (Turley et al., 2018; Zheng et al., 2024). In this way, depression PGSs can also capture genetic signals that are not specific to MDD (Dahl et al., 2023), which necessitates careful characterisation of shared and trait-specific genetic associations.

Measurement reliability introduces a further consideration. Repeated assessment in female twins demonstrated imperfect reliability of lifetime depression diagnoses (Kendler et al., 1993). Error in the measured phenotype can attenuate PGS associations, motivating comparisons of observed outcomes with estimates that account for measurement reliability (Wolfram et al., 2026). Such estimates depend on the measurement model and should be distinguished from prediction of an observed diagnosis, with both metrics being potentially fit-for-purpose, depending on application.

The interpretation of PGS associations also depends on their persistence within families. Predictions among unrelated individuals can also derive from gene-environment correlations induced by population stratification, indirect genetic effects, and correlations with other genetic factors as a result of non-random mating, including assortative mating. Within-family designs separate out

[1]Following common usage, we use major depressive disorder (MDD) for the clinical diagnosis and major depression (MD) for the broader definitions available at biobank scale, which approximate it (McIntosh et al., 2019).

the component due to direct genetic effects alone (Benjamin et al., forthcoming; Chegeni et al., 2026; Huider et al., 2026; Tan et al., 2024; Veller and Coop, 2024; Young et al., 2022), enabling quantification of the degree to which polygenic prediction reflects causal effects that vary within families.

Finally, prediction may vary across populations and environmental contexts. Studies have examined sex differences in the genetic architecture of depression and the interplay of depression PGSs and socioeconomic conditions in adolescent outcomes (Campbell et al., 2026; Thomas et al., 2025). Prediction accuracy also varies across the genetic ancestry continuum, with limited transferability of European-derived depression scores to African-ancestry samples (Ding et al., 2023; Kanjira et al., 2025). These findings motivate comparisons not only of individual outcomes, but also of whether the broader pattern of associations differs across groups.

Here, we develop an integrated depression PGS using phenotype information from multiple large-scale biobanks together with external GWAS summary statistics. The score combines information from depression and genetically correlated psychiatric traits using phenotype reconstruction, multi-trait genetic integration and Bayesian effect estimation, while retaining depression as the focal prediction target (Methods). On a common scale, the integrated summary statistics carry about twice the genome-wide association signal of the PGC depression GWAS (Results).

We evaluate prediction in UK Biobank and the Adolescent Brain Cognitive Development (ABCD) Study and contrast it with a recent best-in-class literature PGS developed by the PGC 2025 European-ancestry meta-analysis excluding 23andMe, the largest publicly available depression summary data (Adams et al., 2025). We first compare our summary data with that of the PGC, then examine prediction of depression and the role of outcome reliability, followed by associations with latent internalizing psychopathology, extending these analyses to physical and mental health, along with behavioral, cognitive, and socioeconomic outcomes. Finally, we examine sex differences, interactions with measured family environments and cross-ancestry portability. Together, these analyses investigate how well the score predicts MD and the overall depression-adjacent phenome as well as the context-dependency of these predictions.

2 Results

2.1 Integration strengthens the genome-wide depression signal

We compared the integrated discovery summary statistics with those of the published PGC GWAS (without 23andMe) on HapMap sites. The integrated GWAS carried 2.10 times (95% CI 2.04–2.17) the genome-wide association signal of the PGC GWAS, measured as the ratio of LD-score regression slopes, which scale with sample size times SNP heritability (Bulik-Sullivan et al., 2015). The increase was uniform across the LD-score distribution (2.07–2.16 across quintiles), and the two sets of statistics were almost perfectly genetically correlated ($r_g = 0.98$).

2.2 The MD PGS predicts depression, internalizing and adult life and health outcomes in population and family analyses

We defined an MD reference phenotype in our UK Biobank validation sample using the published LifetimeMDD definition (Cai et al., 2020; Dahl et al., 2023) and compared our integrated predictor's performance with the PGC benchmark (Methods). Classifications often changed on reassessment. Of 11,567 UK Biobank participants initially classified as cases and reassessed after a median of 6.2 years, only 47.7% remained cases, whereas 38.9% moved to a symptom-positive unclassified category and 13.5% were classified as controls. Treating these changes as measurement error, we report approximate liability-scale estimates before and after a model-based

repeat-stability correction (Methods, Supplementary Table S3C).

The approximate population liability-R^2 was 7.0% for the integrated PGS and 4.6% for the PGC score before correction, and 9.1% and 6.1% after correction, a relative gain of 49.7% (95% CI 27.6–78.0%, Table 1, Supplementary Table S3A). To assess within-family prediction, we fitted SNIPar models with parental scores imputed from genotyped siblings (Guan et al., 2025; Young et al., 2022) in 7,054 participants from 6,085 sibships[2]. The direct/population coefficient ratio was 1.03 (95% CI 0.83–1.23) for the integrated score and 1.02 for the PGC score (Table 1), so there was no detectable within-family attenuation, although reductions of up to about 17% cannot be excluded. Applying these ratios to the population estimates gave model-implied within-family liability-R^2 of 9.6% and 6.3%, respectively (Figure 1A). Matched observed-scale results confirmed this picture, and the integrated PGS also explained more observed variance in anxiety symptoms, neuroticism and internalizing than the PGC score, with relative gains (83–94%) larger than for depression itself (51%), although with a wide interval (Supplementary Table S4). In ABCD, which contributed no discovery data, the relative gain for depression diagnosed with the computerized K-SADS in childhood and adolescence was similar (liability-R^2 4.9% versus 3.5%, a 39% gain, 95% CI 17–66%), although this construct differs from adult lifetime depression (Supplementary Note 2).

Table 1: UK Biobank depression: population and within-family prediction by the integrated and PGC scores.

	Integrated	PGC
Population liability-R^2, uncorrected (%)	7.0 (5.5–8.7)	4.6 (3.4–6.1)
Population liability-R^2, repeat-corrected (%)	9.1 (7.2–11.3)	6.1 (4.5–7.9)
Direct/population ratio	1.03 (0.83–1.23)	1.02 (0.78–1.26)
Model-implied within-family liability-R^2, repeat-corrected (%)	9.6 (6.0–14.5)	6.3 (3.5–10.3)

Note. 7,054 participants (1,679 cases) from 6,085 sibships of the sibling validation sample. Liability-R^2 assumes a population prevalence of 15%, the repeat correction uses the polychoric reliability (Supplementary Table S3D). Direct/population ratio is derived from linear mixed models of standardized case status on the parental-compatible SNIPar score, without (population) and with (direct) adjustment for parental scores imputed from siblings. Intervals are 95% intervals from 4,000 paired whole-family bootstrap draws.

To study the broader mood–anxiety spectrum, we modelled latent internalizing, a hierarchical factor capturing variance shared across depression, anxiety and neuroticism items in UK Biobank (28 PHQ-9, GAD-7 and EPQ-R items) and across anxious/depressed, withdrawn/depressed and somatic-complaint items in ABCD (30 Youth Self-Report items), and estimated within-family PGS associations from sibling differences in the score (Methods). The integrated PGS coefficients were 27.8% (95% CI 17.4–39.2%) higher than those for PGC in UK Biobank ($\beta = 0.272$ versus 0.213) and 54.2% (8.5–119.2%) higher in ABCD ($\beta = 0.269$ versus 0.175, Figure 1B and Supplementary Table S6), and the gains were essentially unchanged under alternative factor models. Within-family attenuation was similar for the two scores in UK Biobank, whereas in ABCD the integrated score retained a larger share of its population association with internalizing (by 37.3 percentage points, 95% CI 4.2–70.3, Supplementary Table S7).

We also observed broad associations with poorer general health, tiredness, life dissatisfaction and chronic pain in the UK Biobank sibling validation sample (10,387–36,257 participants depending on outcome, Figure 1C). Most of these associations remained within families, with direct/population ratios of 0.86 (95% CI 0.76–0.97) for general health burden and 0.87 (0.72–1.02)

[2]Each belonged to a sibship with at least two genotyped siblings, although most sibships contained only one participant with a depression classification (951 of 6,085 contained two or more). SNIPar imputes parental genotypes from all genotyped siblings, so every participant contributed, as with the imputed parental score as a covariate, the direct coefficient reflects only the random inherited component of each participant's score.

for chronic pain. The socioeconomic and cognitive associations, by contrast, were substantially attenuated, with ratios of −0.07 for lower education, −0.04 for lower occupational status, 0.15 for lower fluid intelligence and 0.31 for lower household income. Patterns were similar for the PGC score (Supplementary Note 4).

Associations with 23 diagnoses across psychiatric, cardiometabolic, immune and inflammatory, cancer and other disease categories (Table 2) were more limited: direct within-family associations after Benjamini–Hochberg correction remained significant for just bipolar disorder, coronary artery disease, hypertension, type 2 diabetes, venous thromboembolism and asthma, with direct/population ratios significantly below 1 for hypertension (0.64) and type 2 diabetes (0.60). Results were directionally similar for the PGC score (Supplementary Data).

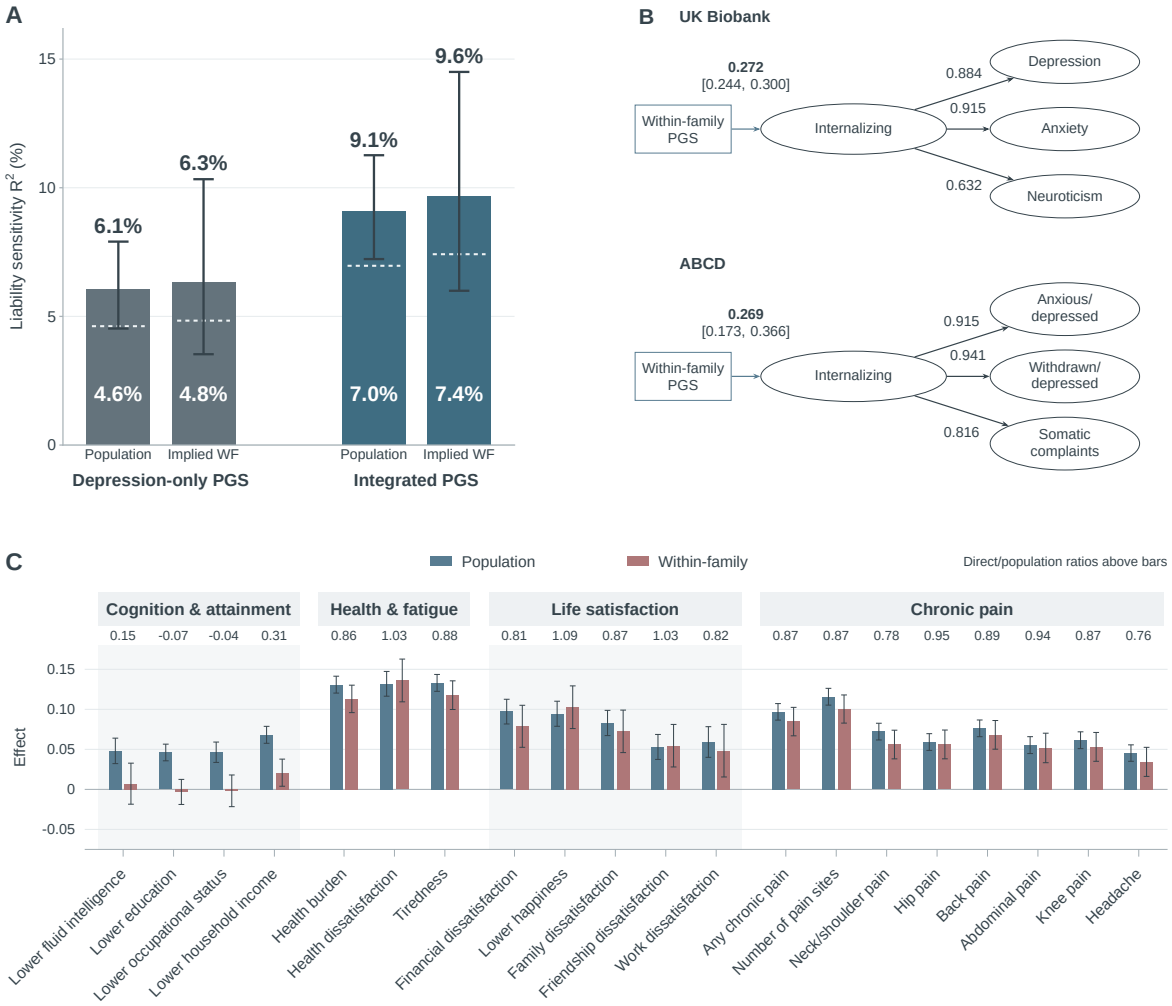


Figure 1: **Depression, internalizing, and adult family associations. A,** Bars and labels above the whiskers show approximate liability-R^2 while correcting for measurement error, with 95% CIs. White labels/dashed marks give the estimates before correction. Within-family liability-R^2 values are model-implied (population value × squared SNIPar direct/population ratio), their intervals include uncertainty in the ratio from 4,000 whole-family bootstrap draws. **B,** Standardised within-family associations with latent internalizing (sibling difference estimate) in UK Biobank and ABCD. **C,** Standardised population and direct (parental-score-adjusted) SNIPar associations of the integrated score with adult outcomes in the UK Biobank sibling validation sample. Numbers above the bars give direct/population ratios. Intervals are 95% CIs.

Table 2: Population and within-family associations of the integrated depression score with disease outcomes in UK Biobank.

Disease	Population	Within-family
	β (95% CI below)	β (95% CI below)
Psychiatric diagnoses		
Bipolar disorder	**0.0183***** (0.0080, 0.0286)	**0.0250**** (0.0065, 0.0434)
Schizophrenia	**0.0169**** (0.0055, 0.0283)	0.0159 (−0.0031, 0.0349)
Psychosis spectrum (F20–F29)	0.0125* (0.0014, 0.0235)	0.0173 (−0.0018, 0.0364)
Cardiometabolic disease		
Coronary artery disease	**0.0323***** (0.0220, 0.0427)	**0.0251**** (0.0074, 0.0428)
Gout	**0.0191***** (0.0087, 0.0294)	0.0142 (−0.0034, 0.0319)
Hypertension	**0.0590***** (0.0490, 0.0691)	**0.0378***** (0.0211, 0.0544)
Type 2 diabetes	**0.0443***** (0.0337, 0.0550)	**0.0266**** (0.0090, 0.0441)
Venous thromboembolism	**0.0241***** (0.0139, 0.0344)	**0.0275**** (0.0096, 0.0455)
Immune and inflammatory disease		
Asthma	**0.0445***** (0.0335, 0.0554)	**0.0475***** (0.0293, 0.0657)
Atopic dermatitis	0.0088 (−0.0033, 0.0210)	0.0058 (−0.0132, 0.0247)
Inflammatory bowel disease	**0.0194***** (0.0092, 0.0295)	0.0089 (−0.0084, 0.0262)
Multiple sclerosis	−0.0020 (−0.0125, 0.0086)	0.0076 (−0.0109, 0.0262)
Psoriasis	0.0099 (−0.0003, 0.0202)	0.0126 (−0.0052, 0.0305)
Vitiligo	0.0036 (−0.0073, 0.0145)	0.0116 (−0.0067, 0.0299)
Cancer		
Basal cell carcinoma	−0.0036 (−0.0140, 0.0068)	−0.0103 (−0.0280, 0.0074)
Breast cancer (female)	0.0053 (−0.0082, 0.0187)	0.0120 (−0.0114, 0.0354)
Colorectal cancer	0.0004 (−0.0098, 0.0106)	−0.0078 (−0.0261, 0.0106)
Melanoma	0.0004 (−0.0097, 0.0106)	0.0152 (−0.0024, 0.0329)
Prostate cancer (male)	−0.0036 (−0.0195, 0.0123)	−0.0011 (−0.0287, 0.0265)
Testicular cancer (male)	−0.0047 (−0.0223, 0.0128)	−0.0327* (−0.0605, −0.0048)
Other diagnosed disease		
Alzheimer's disease	0.0076 (−0.0023, 0.0175)	0.0002 (−0.0170, 0.0173)
Glaucoma	**0.0129*** (0.0022, 0.0235)	0.0109 (−0.0066, 0.0284)
Osteoporosis	**0.0217***** (0.0112, 0.0322)	0.0208* (0.0031, 0.0386)

Note. Coefficients are standardised and binary diagnoses were fitted as linear outcomes. Intervals are 95% confidence intervals based on family-clustered standard errors. * P<0.05, ** P<0.01, *** P<0.001 denote two-sided unadjusted Wald tests, while bold indicates multiple testing correction BH q<0.05, across all 23 diagnoses for each model. Breast cancer is restricted to women, prostate and testicular cancer to men. Within-family coefficients are SNIPar direct effects, adjusted for parental scores imputed from siblings.

2.3 Depression liability reveals broad associations across the childhood phenome

The rich data available in ABCD enabled us to study early-life correlates of the integrated PGS across 2,328 outcomes from 28 instruments concerning social functioning, behavior and self-regulation, mood, stress and psychiatric symptoms, and activity, recreation and sleep. Of these, 1,494 were associated with the PGS after Benjamini–Hochberg correction across the 2,328 outcomes ($q<0.05$), including 423 with nominally significant and directionally similar within-family associations ($P<0.05$, Figure 2A). The profile extended to emotional dysregulation, bullying and peer difficulties, attention and behavioral problems, and sleep disturbance. Supplementary Data gives all outcomes and estimates.

For the 1,854 outcomes measurable in the subset of siblings and dizygotic twins within ABCD, the errors-in-variables slope of within-family on population coefficients across outcomes was 1.01 (95% CI 0.66–1.35, Figure 2B). On average, the profile was therefore carried by direct effects, although the interval allows substantial attenuation or amplification.

Against the PGC score, the mean incremental observed-scale R^2 was 24.7% higher (95% CI 10.8–41.4%, Supplementary Table S4), though the within-family R^2 gain was uncertain (12.3%, Supplementary Data). Nevertheless, the two scores produced closely aligned population association profiles across all 2,328 outcomes in the childhood phenome ($r = 0.988$, Supplementary Table S10). The PGC score retained somewhat less of its population association profile within families (slope 0.90 versus 1.01), but the paired difference was not significant (Supplementary Table S7).

For six illustrative outcomes, we estimated PGS associations across assessments at approximately ages 10–17 (Figure 2C). By the last assessment (age 14 for suicidal thoughts and self-injury, 16 for depression and social anxiety, 17 for the others), children in the highest PGS decile were more likely than those in the lowest to report suicidal thoughts (23.3% versus 6.7%, RR 3.48), nonsuicidal self-injury (20.5% versus 6.9%, RR 2.98), current depression (5.0% versus 1.0%, RR 5.07), current social anxiety (7.3% versus 2.4%, RR 3.01), nicotine or cannabis use (46.9% versus 25.1%, RR 1.87) and psychiatric inpatient treatment since the preceding assessment (3.8% versus 0.8%, RR 4.82). Larger absolute probabilities did not always imply larger relative associations: risk ratios for social anxiety and inpatient treatment were larger, though less precise, at earlier assessments.

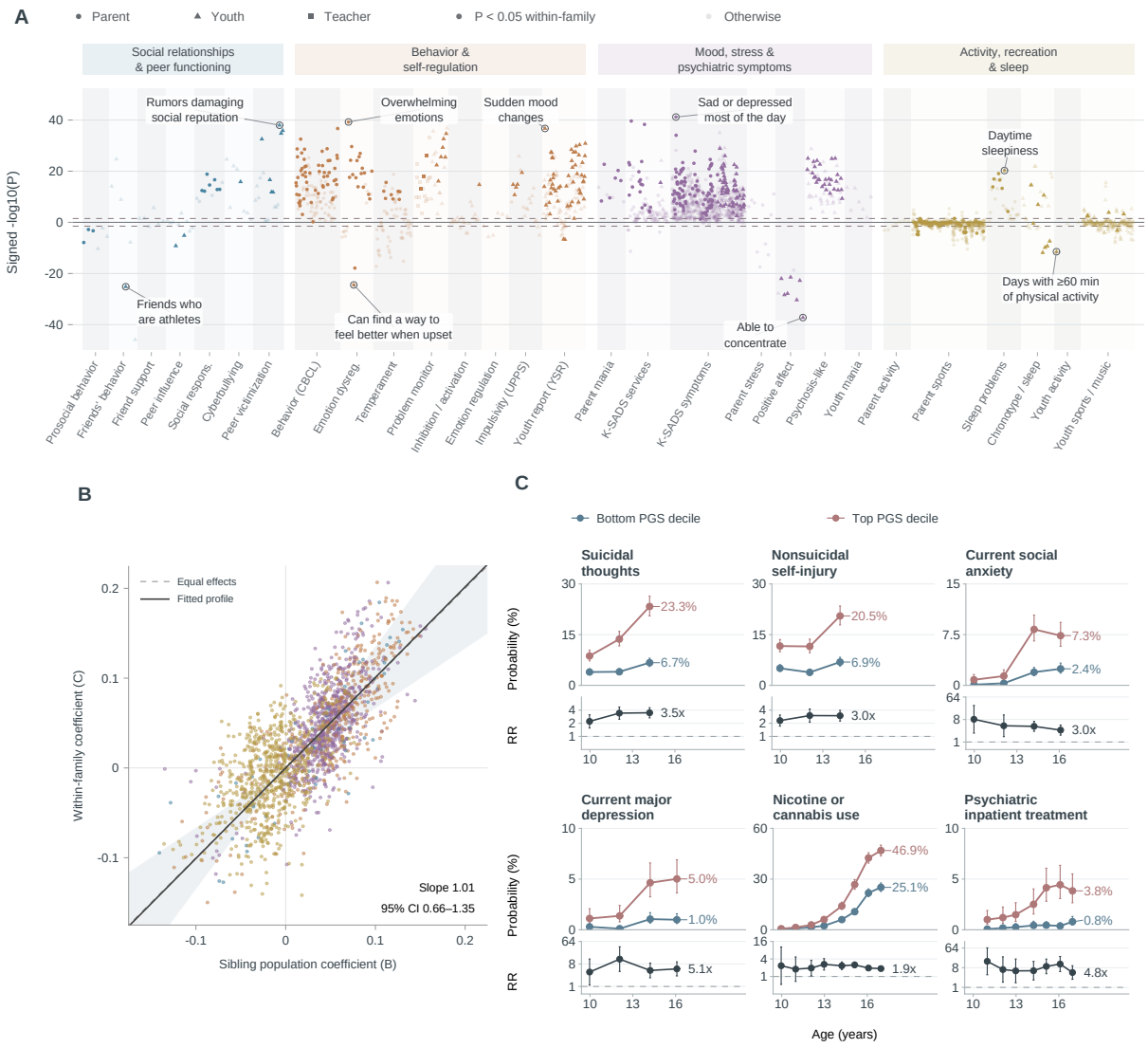


Figure 2: **Childhood phenome and developmental trajectories. A,** Signed population association strength across 2,328 outcomes. Colors identify domains and symbols distinguish reporters. More opaque points have same-direction nominally significant within-family associations (P<0.05). Dashed lines show the BH FDR 0.05 cutoff across the displayed outcomes. **B,** Matched within-family and population coefficients, with the covariance-corrected profile slope and 95% interval. The dashed line denotes equal effects. **C,** Model-standardized probabilities at the top and bottom PGS deciles across assessments, with corresponding risk ratios and pointwise 95% intervals. Suicidal thoughts and self-injury refer to reported past/lifetime history and nicotine or cannabis use to lifetime use, inpatient treatment to the period since the preceding assessment. Supplementary Note 5 describes the models. Supplementary Data provide the PheWAS associations.

2.4 Sex differences in symptom associations and household moderation

We next asked whether the strength and expression of depression-related genetic liability varied by sex and household context in ABCD.

Among the 726 outcomes in the selected psychiatric symptom batteries, 107 showed sex differences after Benjamini–Hochberg correction across these 726 tests. Associations were stronger in girls for 101 outcomes and in boys for six (Figure 3A). The contrasts more positive in girls included fatigue, guilt, low self-esteem, indecision, panic and uncontrollable worry, and self-injury. Those more positive in boys primarily concerned conduct problems (physical fights, bullying, breaking and entering, vandalism and cruelty to animals). Supplementary Data provides all sex-specific coefficients and contrasts.

Given these differences, we examined in an exploratory analysis whether sex-specific PGS associations varied with the cultural gender typicality of symptom content. A language model rated each item from strongly boy-typical to strongly girl-typical using only its wording or description and response options. Among the 716 items with usable ratings, associations in girls increased with girl-typicality, whereas boys showed little linear trend (Figure 3B). Per one-point increase in typicality, the PGS coefficient increased by 0.0074 pooled outcome SD in girls and changed little in boys (between-sex difference 0.0082, 95% CI 0.0050–0.0114). Because these coefficients combine differences in association and symptom dispersion, within-sex outcome standardization shrank the difference (Supplementary Data).

Across the symptom profile, boys' coefficients were 0.68 times those in girls (95% CI 0.53–0.82), corresponding to approximately 32% smaller associations on average (Figure 3C). After within-sex outcome standardization, the difference shrank to 21%. Results were similar for the whole phenome analysis and for the PGC benchmark score (Supplementary Note 6). In UK Biobank adults, the score's liability-R^2 for the reference depression definition was likewise higher in women than in men (repeat-corrected 9.9% versus 8.2%), but the difference was not statistically significant (1.7 percentage points, 95% CI −2.7 to 5.8).

In contrast, household resources, hardship and residential deprivation showed no clear moderation of latent internalizing or either the symptom or full PheWAS profile (Table 3). For the symptom profile, point estimates suggested somewhat stronger associations under disadvantage, but none remained significant after Holm correction across the three exposures (smallest adjusted $P = 0.08$). Latent internalizing interactions were close to zero (at most 0.016 latent SD per exposure SD) and did not differ detectably between the two scores (Supplementary Note 6).

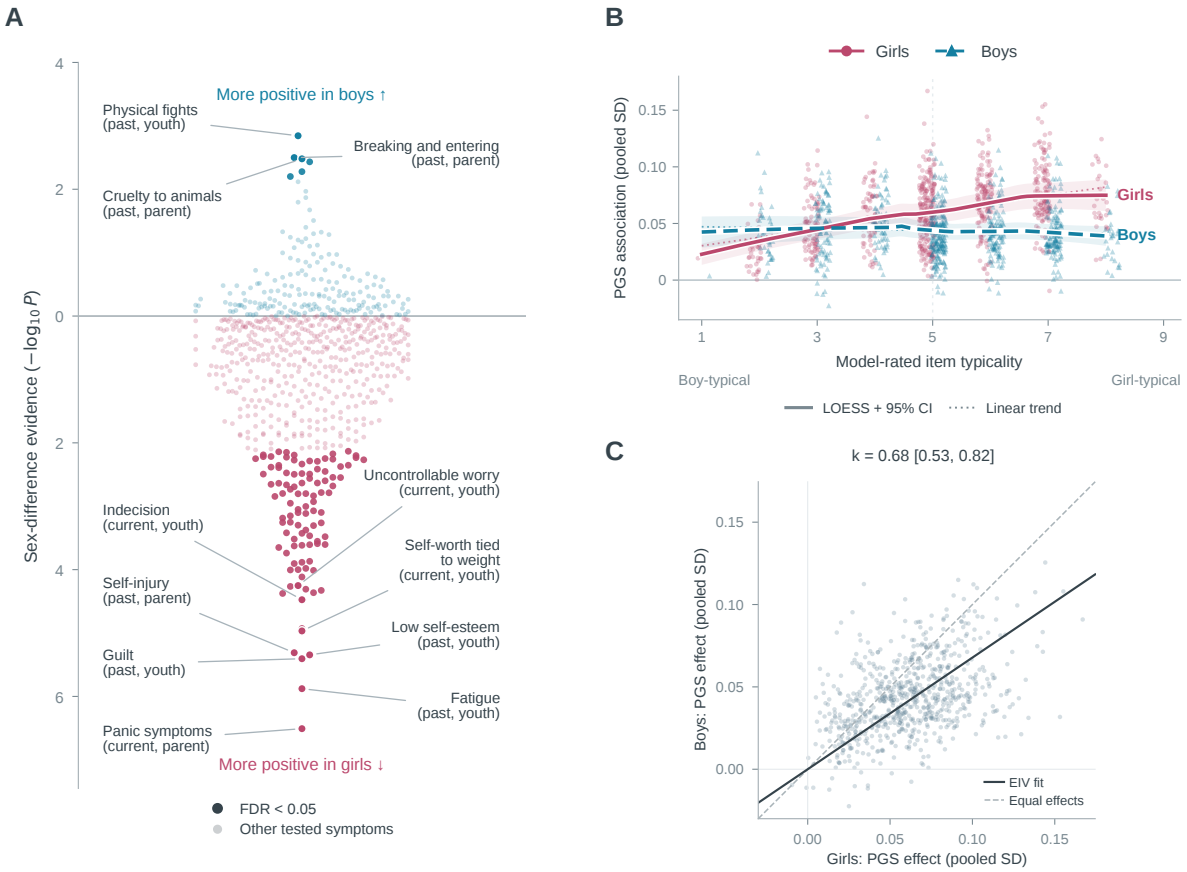


Figure 3: **Sex-specific PGS associations across psychiatric symptoms in ABCD. A,** Sex contrasts in the selected psychiatric symptom batteries, with signed vertical position given by $\text{sign}(\beta_{\text{male}} - \beta_{\text{female}}) \times [-\log_{10}(P)]$ for a two-sided sex-difference test. Blue denotes a more positive male coefficient and rose a more positive female coefficient. This contrast does not necessarily indicate a larger absolute association. Darker, larger points pass Benjamini–Hochberg FDR <0.05 across the 726 displayed outcomes. **B,** Relationship between cultural gender typicality of symptom content and sex-specific PGS associations. A language model rated items from boy-typical to girl-typical using their descriptions, reporter and response options, without access to participant data or association results. Rose circles and the solid curve denote girls, blue triangles and the dashed curve denote boys. Curves are local linear LOESS fits with pointwise 95% confidence bands. Dotted lines show equal-item-weighted linear trends. Coefficients are standardised. The panel includes 716 rated items. Ten outcomes without usable ratings remain in A and C. **C,** Girls' and boys' coefficients for the same outcomes as A. The dashed line indicates equality and the solid line the corrected-moment errors-in-variables slope, k, with its 95% interval. See Supplementary Note 6 for selection, ratings, scaling and uncertainty.

Table 3: Sex and household moderation of latent internalizing, psychiatric symptoms and the full PheWAS profile in ABCD.

Moderator	Internalizing Interaction (latent SD)	Symptom profile Profile moderation	Full PheWAS profile Profile moderation
Sex (boys minus girls)	−0.099 [−0.151, −0.047]	−0.322 [−0.473, −0.177]	−0.292 [−0.430, −0.158]
Resources	−0.007 [−0.039, 0.025]	−0.145 [−0.306, 0.012]	−0.102 [−0.248, 0.045]
Hardship	0.007 [−0.021, 0.035]	0.163 [0.017, 0.307]	0.108 [−0.018, 0.229]
Residential deprivation	0.016 [−0.026, 0.059]	0.108 [−0.080, 0.295]	0.107 [−0.063, 0.274]

Estimates with 95% CIs below. Household exposures are standardized. Profile estimates are boys-to-girls multipliers minus one (sex) or slopes of interaction coefficients on the covariate-averaged main PGS coefficients (household exposures). No household profile slope was significant after Holm correction across the three exposures. The symptom profile covers 726 outcomes (Figure 3). The full profile includes all eligible PheWAS outcomes.

2.5 Core associations and phenotype profiles across population groups

Lastly, we compared PGS performance across population groups, first for depression and latent internalizing and then across the childhood phenome.

In UK Biobank, the integrated PGS association with depression was lower than the European reference by 61.9% in the African-ancestry (AFR) group, 71.2% in East Asians (EAS) and 15.7% in South Asians (SAS, Figure 4A). Intervals were wide because of small samples and included published expected effect reductions of 57.6%, 30.3% and 19.6%, respectively (Privé et al., 2022) (Supplementary Note 7 and Table S12A).

In ABCD, the integrated-score association with latent internalizing was reduced by 27.0% in the Hispanic group and 67.2% (95% CI 35.1–98.1%) in the African American group, compared with published expected reductions of 14.0% and 34.7%, respectively (Moore et al., 2025; Wolfram et al., 2026) (Supplementary Table S12A).

Across the ABCD phenome, the slope relative to White children was 0.78 (95% CI 0.61–0.95) in Hispanic children and 0.65 (0.38–0.91) in African American children (Figure 4B), indicating approximately 22% and 35% attenuation in the overall coefficient profiles, compared with the same expected reductions of 14.0% and 34.7%. Supplementary Figure S1 shows the full coefficient range. The PGC score showed a similar pattern, with slopes of 0.79 and 0.58 (Supplementary Table S12B).

3 Discussion

We developed a depression polygenic score that outperforms the strongest publicly available depression score, based on the European-ancestry PGC 2025 summary statistics excluding 23andMe (Adams et al., 2025). By integrating observed and reconstructed depression phenotypes from multiple biobanks with external broader psychopathology spectrum GWAS, we obtained summary statistics carrying about twice the genome-wide association signal of the PGC GWAS. In UK Biobank, the approximate liability R^2 for depression rose from 4.6% to 7.0% (a relative increase of 51%) and, after correction for repeat-measurement error, from 6.1% to 9.1% (49.7%). Within families, standardized coefficients with latent internalizing were 27.8% larger than those of the benchmark score in UK Biobank and 54.2% stronger in ABCD, with gains extending to neuroticism and anxiety.

A common concern with integrating broader phenotypes and genetically correlated traits is that power is gained at the cost of specificity, so that the score increasingly reflects anxiety, neuroticism or similar traits rather than depression (Cai et al., 2020; Dahl et al., 2023). A score that had drifted toward the internalizing spectrum would gain least on strict depression and most on internalizing-type outcomes. The integrated score did gain on depression itself, in population and within-family analyses alike. In the population analyses, however, the gains for anxiety, neuroticism and the internalizing composite were larger than for depression (83–94% versus 51% in incremental R^2, corresponding to roughly 35–39% versus 23% on the coefficient scale, Supplementary Table S4). Within families, by contrast, the gains were of the same order for all four outcomes (20–33% for anxiety, internalizing and neuroticism, Supplementary Table S6, and about 24% for depression, with direct coefficients of 7.74 versus 6.25 percentage points per SD, Supplementary Note 2). Neuroticism showed the largest within-family gain (33%), so some reweighting toward neuroticism cannot be excluded, although the depression estimate is too imprecise to distinguish this from a uniform gain. Across the childhood phenome the gain was proportional: the association profiles of the two scores were closely aligned across all 2,328 ABCD outcomes ($r = 0.988$), and the integrated score's coefficients were on average 11% larger and its mean incremental R^2 24.7% higher in the 1,854-outcome common inventory (Supplementary Tables S4 and S10). The

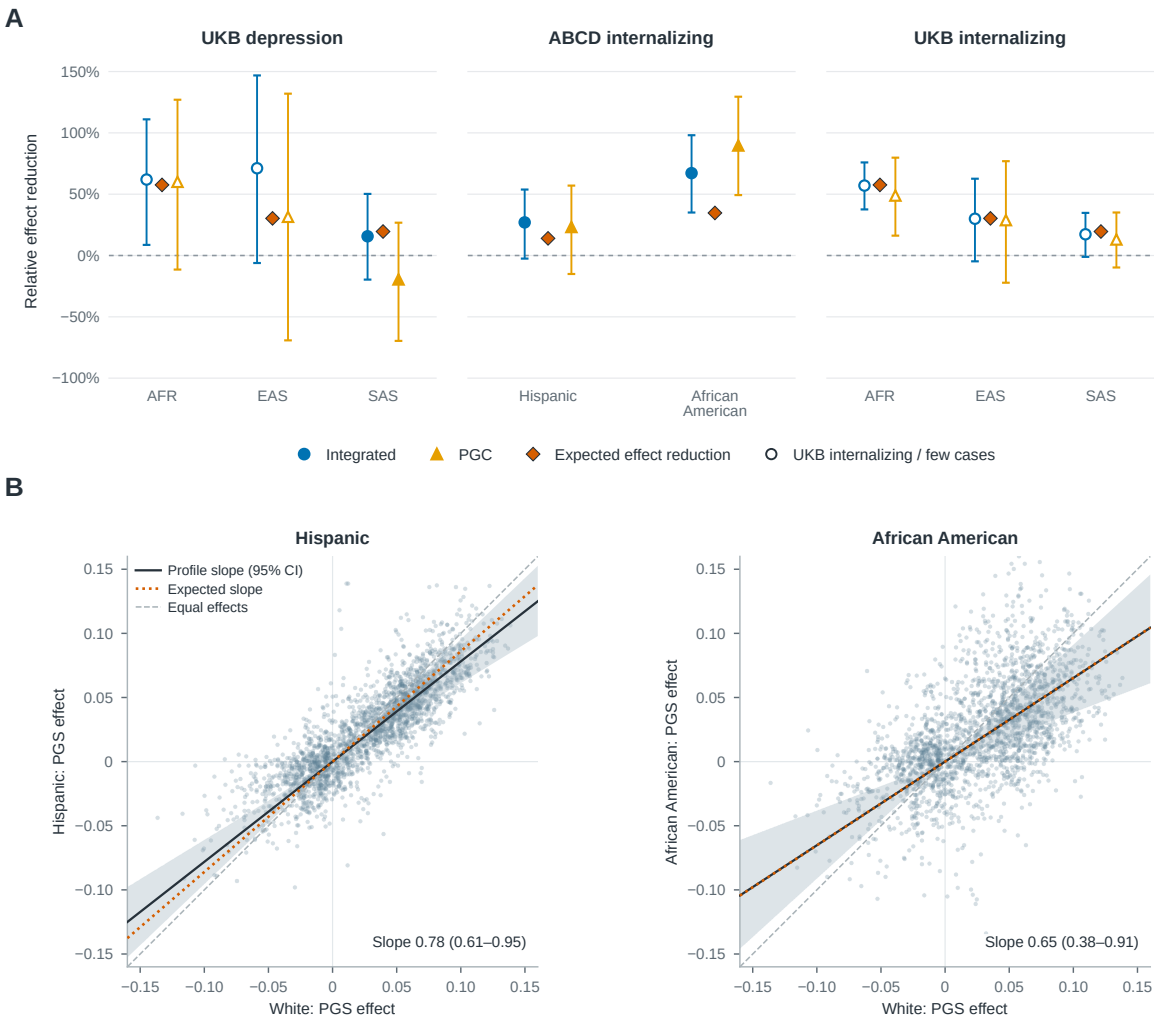


Figure 4: **Core and phenotype-wide portability. A,** Relative reduction, $1 - \beta_g/\beta_{\text{ref}}$, for UK Biobank depression and ABCD and UK Biobank internalizing. Positive values indicate smaller associations and negative values larger associations than the reference group (EUR in UK Biobank, White children in ABCD). Circles denote the integrated PGS and triangles the PGC benchmark. Depression uses log-odds coefficients. Internalizing coefficients are standardised. Bars show 95% intervals. Open markers identify the UK Biobank internalizing analysis and AFR/EAS diagnosis cells with few cases. **B,** ABCD phenotype-wide associations. Solid lines show errors-in-variables slopes. Shaded bands describe pointwise 95% CIs. Dashed lines denote equal effects. The view is cropped. The complete plotting range is shown in Supplementary Figure S1. Diamonds in A show expected effect reductions and dotted lines in B expected profile slopes, derived from published relative R^2 estimates for UK Biobank (Privé et al., 2022) and ABCD (Moore et al., 2025; Wolfram et al., 2026). See Supplementary Note 7 for sample, model, measurement-comparability and benchmark limitations.

integrated score thus strengthened the associations the depression-only score already had rather than adding new ones. Together, these results limit the concern about specificity but do not rule it out.

Beyond the comparison with PGC, the score predicts not only depression but a broad liability to mood and anxiety problems, both in the population as well as within-family. Symptom dimensions of this kind capture liability that a lifetime diagnosis, which compresses it into two imperfectly reliable categories (Kendler et al., 1993), does not.

The score also predicted a wide range of adult outcomes. Associations with poor health, tiredness, chronic pain and life dissatisfaction were retained within families. By contrast, associations with fluid intelligence, education and occupational status were reduced to (almost) zero within families, and the association with household income became barely significant. For socioeconomic outcomes, population stratification, assortative mating, and/or indirect genetic effects are therefore likely to account for much of the population association. Because ascertainment in UK Biobank also shifts genetic correlations between depression and cognition (Nguyen et al., 2022; Tyrrell et al., 2021), the cognitive associations in particular should be interpreted with caution. Disease associations were smaller on the standardized scale, and their within-family point estimates were often lower. The absence of detectable within-family attenuation for depression (direct/population ratio 1.03, 95% CI 0.83–1.23) agrees with twin, trio and imputed-parent studies of depression polygenic scores (Huider et al., 2026; Trindade Pons et al., 2025), although our interval does not exclude attenuation of up to about 17%. Genome-wide, family-based GWAS estimate a direct–population genetic correlation of only 0.68 for depression, indicating substantial confounding of population effects, although population effects were on average smaller rather than larger than direct effects (Tan et al., 2024), in contrast to the within-sibship shrinkage reported for strongly associated variants for depressive symptoms (Howe et al., 2022).

In a phenome-wide scan of 2,328 outcomes in ABCD, 1,494 were associated with the score after correction for multiple testing. Higher scores were associated with greater peer victimization, including reputation-damaging rumors, greater difficulty regulating emotions, more sleep problems and less physical activity. These associations retained their strength within families on average. Children in the top score decile had substantially elevated probabilities for suicidal thoughts, self-injury and were about five times as likely as children in the bottom decile to have received psychiatric inpatient treatment since the preceding assessment. These results extend earlier work on adolescent depression trajectories and on the childhood correlates of internalizing liability (Grimes et al., 2024; Paul et al., 2024).

Household resources, material hardship and residential deprivation did not detectably moderate the score's association with latent internalizing, in line with the minimal interaction between depression PGS and household income reported in ABCD (Campbell et al., 2026). Across the psychiatric symptom profile, point estimates suggested somewhat stronger associations under disadvantage, but none remained significant after correction across the three exposures. Sex, by contrast, mattered substantially. Across 726 psychiatric symptom outcomes, a corrected-moment profile slope placed boys' coefficients at around 32% less than girls', a pattern also evident for the PGC score on the same psychiatric outcomes and across the broader phenome. Symptoms with significant sex differences almost always had more positive coefficients in girls, including panic symptoms, fatigue, guilt, low self-esteem and self-injury.

The sexes may also differ in which symptoms express the liability. In girls, associations increased with the rated girl-typicality of symptoms, whereas in boys they were largely independent of it. With sex-specific standardization, however, the difference between the two trends attenuated, so the gradient depends substantially on sex differences in symptom variance, which for binary and near-binary items largely reflects sex differences in symptom prevalence. The overall pattern aligns with the adolescent onset of female excess in depression (Salk et al., 2017), though, as male and female depression share most of their genetic architecture (Thomas et al., 2025), the weaker

association in boys does not require different causal variants and can be the result of differences in symptom expression, reporting or prevalence, including the hormonal changes of puberty (Angold et al., 1999).

Cross-ancestry effect reductions were in line with expectations for a score trained on European-ancestry data (Ding et al., 2023; Kanjira et al., 2025; Wang et al., 2020). Although intervals were wide, almost every estimate was compatible with published portability expectations for other traits (Moore et al., 2025; Privé et al., 2022). Closing the portability gap will require more diverse discovery data.

Taken together, the integrated score predicts depression more strongly and captures more of the internalizing spectrum, including neuroticism, while in its general phenome-wide association profile remaining closely aligned with that of the depression-only benchmark score and showing minimal within-family effect reduction. For depression, broad phenotypic integration across the internalizing spectrum could therefore roughly double the association signal of a depression GWAS without evidence of the loss of specificity that has been a concern for minimally phenotyped GWAS (Cai et al., 2020).

Several limitations apply. First, the PGC benchmark excluded 23andMe, which contributed about 113,000 European-ancestry cases and 1.77 million controls, an effective sample size of about 425,000 against about 1.15 million for the European meta-analysis without it (Adams et al., 2025). A score including these data would be a stronger benchmark, but the same data could be added to our framework, and phenotype reconstruction would plausibly benefit from 23andMe's survey data as well. Second, the integrated score's effects were estimated with SBayesRC, which uses functional annotations, whereas the benchmarks used SBayesR, as in the PGC study. Part of the score-level gain may therefore reflect the estimation method rather than phenotype reconstruction and multi-trait integration. The comparison of the summary statistics themselves, which precedes effect estimation, is unaffected. Third, the measurement-error correction treats disagreement between two questionnaires as error, although it may also reflect new episodes or changed recall (Kendler et al., 1993; Moffitt et al., 2010), though under a liability framing, that's still an imperfect indication of a stable liability. Fourth, participation bias can distort genetic associations in UK Biobank (Schoeler et al., 2023), and the ABCD family sample and the non-European samples were small.

4 Methods

4.1 Participants and study design

For validation, we used UK Biobank and ABCD data (ABCD Study, 2026; Sudlow et al., 2015). We confirmed no overlap between local discovery GWAS and UK Biobank validation samples. External GWAS inputs to the integrated score excluded UK Biobank, while ABCD contributed no discovery data.

UK Biobank sample. UK Biobank validation used genotyped sibships, with family relationships identified using KING. Family validation analyses used participants in the United Kingdom genetic-ancestry group (Privé et al., 2022), excluding heterozygosity/missingness outliers and sex-chromosome aneuploidies.

In line with the reference phenotype construction (Supplementary Note 1), participants in the analysis in Figure 1A additionally required British self-reported ethnicity, the UK Biobank genetic ethnic grouping labelled "Caucasian," and agreement between reported and genetic sex. The sample comprised 7,054 participants. Repeat-assessment corrections used respondents with

interpretable classifications at both assessments and were not restricted to the family cohort. The latent internalizing analysis required at least two eligible siblings per family, each with at least one observed symptom item, leading to 36,296 participants from 17,656 families. Adult SNIPar analyses used 10,387–36,316 participants from 7,776–17,668 sibships of the sibling validation sample, depending on outcome. Supplementary Note 2 specifies the depression case/control rules, cross-population samples are described under Population portability.

ABCD sample. We used phenotypes from Release 7, with consortium-processed genotypes and genetic/demographic metadata. Primary family analyses used parent-reported White siblings and dizygotic twins, monozygotic twins and twins with missing or conflicting zygosity were excluded. The latent-family sample contained 1,172 participants in 593 families. Broader population and portability analyses used separate eligible samples, with the broader White population including singletons, siblings and twins (Supplementary Notes 1 and 5–7). Score comparisons matched participants and covariates within each analysis. Reference score units remained fixed after outcome restriction and resampling (Supplementary Note 1).

4.2 Polygenic score construction and benchmarks

Scope of disclosure. This manuscript describes the evaluation of an integrated depression predictor and comparisons with literature-based GWAS benchmarks. Detailed predictor construction methods and input specifications are withheld in this version and will be disclosed in the published paper. The validation methods and benchmark specifications are described here and in the Supplement.

The integrated depression PGS combined GWAS of observed and reconstructed depression phenotypes from multiple biobanks with external GWAS summary statistics using multi-trait analysis of GWAS (MTAG) (Turley et al., 2018). Posterior SNP effects were subsequently estimated using SBayesRC in GCTB 2.5.2 (Zheng et al., 2024). To compare the integrated GWAS with the published PGC GWAS without 23andMe, we used ratios of LD-score regression slopes on identical variants (Bulik-Sullivan et al., 2015). These slopes are proportional to sample size times SNP heritability and are estimated from the association statistics alone (Supplementary Note 1).

PGC benchmarks. Depression-only comparative analyses used the European PGC MDD2025 release without 23andMe (Adams et al., 2025) and, in line with the PGC study, SBayesR (Lloyd-Jones et al., 2019) for score construction. ABCD used the published UK Biobank-inclusive GWAS. All UK Biobank analyses used PGC excluding UK Biobank combined with a local depression GWAS excluding evaluation participants and their relatives. The local phenotype was designed to approximate PGC depression ascertainment, using observed diagnostic and questionnaire evidence without phenotype reconstruction or auxiliary-trait integration. GWAS estimates were combined by inverse-variance meta-analysis. To assess alignment with the PGC depression target, we used LD-score regression (Bulik-Sullivan et al., 2015) and found a genetic correlation close to unity (unconstrained $\hat{r}_g = 1.011$). Supplementary Note 1 specifies phenotype, harmonization and effective-sample-size rules.

4.3 Depression measurement and family validation

The MD reference phenotype used the LifetimeMDD definition (Cai et al., 2020; Dahl et al., 2023), which is based on UK Biobank's mental-health questionnaire (Supplementary Note 2). Measured PGS performance depends on the depression definition and on the cohort in which it is evaluated: shallow definitions have shown lower SNP heritability and less depression-specific

genetic signal than detailed diagnostic phenotypes (Cai et al., 2020; Dahl et al., 2023), and in the PGC's own leave-one-cohort-out evaluation, the full PGC 2025 score (including 23andMe) explained 4.1% of liability variance in community-recruited cohorts but 7.0% in cohorts recruited from inpatient, outpatient or register settings (Adams et al., 2025, Table S6). Absolute estimates are therefore not directly comparable across definitions and cohorts, whereas the two scores were always compared on the same phenotype and participants.

We estimated population associations with the reference depression phenotype using logistic regression of case/control status on the full PGS, adjusting for sex, birth year, its square, assessment center, array and genetic principal components (PCs) 1–20 and obtained liability-scale R^2 using established conversions assuming a population depression prevalence of 15% (see "Repeat measurement and liability" and Supplementary Note 2).

All SNIPar analyses (depression and adult outcomes) used sibships of the sibling validation sample, which excludes families with a genotyped parent, so both parental scores were imputed by SNIPar from the siblings' genotypes (Young et al., 2022). Participant and parental scores were calculated using identical variants and SNP weights. We fitted two linear mixed models to the same participants, using the same covariates and score scaling:

$$Y^*_{if} = \alpha_P + \beta_P G_{if} + \gamma_P^\top C_{if} + u_{Pf} + \varepsilon_{Pif},$$
$$Y^*_{if} = \alpha_D + \beta_D G_{if} + \eta \widehat{P}_f + \gamma_D^\top C_{if} + u_{Df} + \varepsilon_{Dif}.$$

Y^* denotes case/control status scaled to unit SD, G is the participant's PGS, $\widehat{P}$ is the (imputed) parental-score sum in the same units, C contains the covariates, and u denotes a family random intercept. The population coefficient β_P estimates the association without parental-score adjustment, whereas β_D estimates the association conditional on parental scores. The ratio β_D/β_P denotes the effect retained within families. Coefficient standard errors are SNIPar's model-based estimates, and the interval for the ratio came from 4,000 whole-family bootstrap samples that refitted both models. As a robustness check, we also ran probit models, which gave nearly identical depression associations to the linear models on the probability scale (Supplementary Note 2).

Latent internalizing used 28 ordinal PHQ/GAD/neuroticism items in UK Biobank and 30 Youth Self-Report (YSR) items averaged across all available ABCD assessments (4.5-, 5.5-, 6- and 7-year follow-ups). Neuroticism used the 12-item Eysenck Personality Questionnaire–Revised Short Form. The hierarchies represented shared internalizing variation across depression, anxiety and neuroticism in UK Biobank, and anxious/depressed, withdrawn/depressed and somatic symptoms in ABCD. UK Biobank used weighted least squares (WLSMV) due to its ordinal item responses, with pairwise ordinal information. ABCD treated averaged item responses as continuous and used robust maximum likelihood with full-information handling of missing items (MLR/FIML) and family clustering. Sibling models included the family mean and the child's deviation for each score, $G_{if} = \bar{G}_f + (G_{if} - \bar{G}_f)$, in the latent regression, where the deviation coefficient estimated the within-family association. Within-family childhood phenome models included family fixed effects, with a separate intercept for each family. For UK Biobank anxiety, the estimated within-family advantage remained positive, but its confidence interval included zero when complete responses to all 28 questionnaire items were required (Supplementary Table S6B).

Repeat measurement and liability. Because case/control status indicates whether continuous liability exceeds a threshold, conversion to the liability scale depends on assumptions concerning prevalence and ascertainment. We transformed log-odds coefficients per score SD using standard conversions to the liability scale (Lee et al., 2012; Moore et al., 2025).

Recall or reporting error can further weaken the association between a case/control label and a PGS (Kendler et al., 1993). To correct for measurement error, we reconstructed the depression

algorithm in two questionnaires administered in UK Biobank in 2016–2017 and 2022–2023. Of 79,072 participants for whom the reference phenotype could be constructed at both assessments, 25,022 (31.6%) changed classification after a median of 6.2 years (Supplementary Table S3C). We compared binary Cohen's kappa, a three-state polychoric correlation and a two-component Gaussian-copula mixture as alternative measures of reliability s. The working correction divided the population log-odds coefficient by $\sqrt{s}$ before applying the coefficient-derived liability mapping at assumed prevalence $K = 0.15$, following the PGC depression analysis (Adams et al., 2025), the uncorrected specification used $s = 1$. The polychoric specification is used in Figure 1A, and Table S3 compares alternatives.

The implied within-family liability-R^2 in Figure 1A was the population summary multiplied by the squared same-score SNIPar direct/population ratio, assuming transfer to the full-score association in population variance units. Supplementary Note 2 gives the derivation, calibration and resampling details.

4.4 Adult health and disease associations

Twenty health, pain, cognitive and socioeconomic outcomes and 23 diagnoses were analyzed with and without parental PGS using SNIPar, with binary diagnoses treated as linear outcomes. Models adjusted for sex, birth year, its square and PCs 1–20. For sex-restricted cancers, covariates were birth year and PCs.

For adult outcomes, coefficient intervals used family-clustered robust (HC0) covariance, conditional on the fitted variance components, and direct/population ratio intervals used Fieller's method (Fieller, 1954) with the same covariance. Fieller's method inverts the test that the direct coefficient minus the ratio times the population coefficient equals zero. Unlike the delta method, it does not rely on a linear approximation and gives asymmetric or unbounded intervals when the denominator is imprecise. For each score, BH correction was applied to all 23 diagnoses, separately for population and direct-effect models. We used Fieller intervals also for portability reductions and observed-scale R^2 gains, with jackknife or influence-function covariances, and log-ratio intervals for gains in latent coefficients (Supplementary Notes 2, 3 and 7). The core depression ratio instead came from the 4,000-sample whole-family bootstrap used for its liability-R^2. The ratio and the model-implied within-family R^2, which multiplies the population R^2 by the squared ratio, therefore share one set of resampled families, which also re-estimates the reliability correction and the variance components. Throughout, confidence intervals are pointwise 95% intervals unless stated otherwise. In multiple-testing families with a fixed set of tests, a test that could not be computed was counted as $P = 1$, whereas the BH correction of household-moderation tests across outcomes and exposures included only the tests that could be computed.

Supplementary Note 4 gives the outcome definitions, measurement occasions, disease codes, control exclusions and model specifications.

4.5 Childhood phenome and developmental trajectories

We tested PGS associations with 2,328 ABCD outcomes using repeated observations across eligible assessment waves. We fitted linear models to continuous outcomes and to binary outcomes coded 0 or 1, adjusting for sex, baseline age, PCs 1–10, site and assessment wave. Standard errors accounted for correlations among observations from the same family.

We compared population and within-family associations in siblings and dizygotic twins. These models used identical participants and assessments, with the same PGS and outcome scaling. The within-family model included a separate intercept for each family, estimating associations from differences between siblings. For the 1,854 outcomes with estimates from the full population

sample and both sibling models, we used regression slopes to summarize differences in association strength between models.

Slopes were constrained to pass through zero and corrected for sampling error. Confidence intervals used 10,000 draws, applying each family's random weight consistently across outcomes and models to preserve their correlations (Supplementary Note 5).

Comparison models between our score and the PGC benchmarks used the same design and outcomes. We compared coefficient profiles and mean incremental observed-scale R^2, defined relative to original outcome total variance, including for family fixed-effect models (Supplementary Note 5 and Tables S4 and S10). The population-profile correlation of 0.99 was estimated across all 2,328 outcomes. Family-model comparisons retained the 1,854-outcome common inventory.

Developmental trajectories. Six binary outcomes, selected as illustrative examples, used logistic models with assessment indicators and PGS-by-assessment interactions, adjusting for sex, baseline age, PCs 1–10, site and within-assessment age deviation. Probabilities were averaged over the empirical top and bottom score deciles using a common covariate distribution. Supplementary Note 5 and Table S11 give outcome definitions, recall periods and minimum participant/event counts.

4.6 Sex differences and gene-environment interaction

We fitted separate models in boys and girls, allowing covariate effects to differ by sex. Both sexes used the same reference PGS SDs and outcome SDs calculated across sexes. We tested differences between boys' and girls' coefficients across 726 psychiatric outcomes, accounting for relatedness between participants, including brothers and sisters. Positive differences indicate more positive coefficients in boys, which need not mean stronger associations.

We summarized the overall sex difference using a regression slope relating boys' coefficients to girls' coefficients, with the line constrained to pass through zero. Each outcome received equal weight, and the slope was corrected for sampling error and correlated estimation errors between sexes. Confidence intervals used 10,000 draws that reweighted families' contributions to the estimates (Supplementary Note 6). This slope describes an average pattern, while individual outcomes can differ.

A sensitivity analysis standardized outcomes separately in boys and girls while keeping PGS scaling unchanged. Comparisons between the integrated and PGC scores used identical outcome sets, both within the psychiatric subset and across the full phenome.

Symptom content. For the exploratory content analysis, separate `gpt-5.6-sol` requests rated items from 1 (strongly boy-typical) to 9 (strongly girl-typical), using item wording or source descriptions, reporter and response options. The ratings describe whether symptoms are culturally perceived as more typical of boys or girls. They do not measure participants' gender or how often symptoms actually occur. Requests contained no participant data, association results, previous ratings or conversation history. Ratings describe perceived cultural typicality among US adolescents aged 12–15. For 716 usable ratings, equally weighted item-level trends tested their relationship to coefficients in each sex and the between-sex trend difference.

Confidence intervals accounted for correlations across items and between relatives, including brothers and sisters. After standardizing outcomes separately in boys and girls, PGS associations in girls still increased with how girl-typical the symptoms were rated, however, the difference between girls' and boys' trends was attenuated. Supplementary Note 6 provides the methods, the ratings are part of the Supplementary Data.

Gene-environment interaction. We examined baseline family resources (income and parental education), material hardship measured with seven items, and residential deprivation measured by the Area Deprivation Index (Kind and Buckingham, 2018). Higher resources indicate greater advantage, whereas higher hardship and deprivation indicate greater disadvantage. Exposures were assigned to families and standardized before selecting participants with outcome data. Families with missing required information or conflicting reports were excluded from that exposure's analysis.

We tested whether PGS associations varied with each exposure using separate population models. Each model included the PGS, exposure and their interaction. Models also included a squared-exposure term and interactions of both the PGS and exposure with covariates (Keller, 2014). Latent internalizing models adjusted for sex, average age at YSR assessments, PCs 1–10 and site. Phenome models adjusted for sex, baseline age, PCs 1–10, site and assessment wave.

Analyses covered latent internalizing, the symptom subset and the full eligible phenome. Across outcomes, we summarized moderation using regression slopes relating interaction coefficients to PGS coefficients. The PGS coefficients represented associations at average exposure, averaged over the covariate values of the analysed observations. Slopes were corrected for sampling error and correlated errors between estimates.

For latent internalizing, sex differences were calculated as boys' minus girls' PGS coefficients, using the same latent scale for both sexes. For phenome summaries, we subtracted one from the boys-to-girls multiplier, so zero indicated no overall difference. Statistical tests and confidence intervals accounted for relatedness within families. Supplementary Note 6 gives the exposure definitions and multiple-testing corrections.

4.7 Population portability

UK Biobank used precomputed genetic-PC groups labeled United Kingdom, Nigeria, China and India. ABCD used parent-reported race/ethnicity groups: White (reference), Hispanic and Black (labelled African American). Multiple-group assignments were excluded except for the Hispanic group (Supplementary Note 1). The UK Biobank reference used the discovery-excluded family sample (United Kingdom PC group), the Nigeria, China and India groups included all eligible scored participants, without a sibling requirement. Matched-score core comparisons estimated $100(1-\beta_g/\beta_{\mathrm{ref}})$ using group-specific PGS SDs and each score's own reference-group denominator. Details are given in Supplementary Note 7.

Acknowledgements

This research has been conducted using the UK Biobank Resource under Application Number #103244. This work uses data provided by patients and collected by the NHS as part of their care and support.

Furthermore, data used in the preparation of this article were obtained from the Adolescent Brain Cognitive Development (ABCD) Study (https://abcdstudy.org), held in the NIMH Data Archive (NDA) under application #21063. This is a multisite, longitudinal study designed to recruit more than 10,000 children aged 9–10 and follow them over 10 years into early adulthood. The ABCD Study® is supported by the National Institutes of Health and additional federal partners under award numbers U01DA041048, U01DA050989, U01DA051016, U01DA041022, U01DA051018, U01DA051037, U01DA050987, U01DA041174, U01DA041106, U01DA041117, U01DA041028, U01DA041134, U01DA050988, U01DA051039, U01DA041156, U01DA041025, U01DA041120, U01DA051038, U01DA041148, U01DA041093, U01DA041089, U24DA041123, U24DA041147. A full list of supporters is available at https://abcdstudy.org/federal-partners.html. A listing of participating sites and a complete listing of the study investigators can be found at https://abcdstudy.org/consortium_members/. ABCD consortium investigators designed and implemented the study and/or provided data but did not necessarily participate in the analysis or writing of this report. This manuscript reflects the views of the authors and may not reflect the opinions or views of the NIH or ABCD consortium investigators.

We thank the participants and investigators of the All of Us Research Program and the NIH for making the research data available. We also acknowledge the participants and investigators of FinnGen.

Ethics and consent. UK Biobank obtained approval from the North West Multi-centre Research Ethics Committee and informed consent from participants. ABCD obtained central or local institutional review board approval, parental consent and child assent. The All of Us Research Program obtained institutional review board approval and informed consent from participants. This study analyzed existing cohort data.

Data access. This study used the All of Us Research Program's Controlled Tier v9 dataset, available to authorized researchers through the Researcher Workbench (https://www.researchallofus.org/data-tools/workbench/). Access requires an institutional Data Use and Registration Agreement covering the Controlled Tier, researcher registration and identity verification, completion of Responsible Conduct of Research training, and agreement to the Data User Code of Conduct.

Competing interests. All authors are affiliated with Herasight, which develops commercial polygenic predictors.

References

ABCD Study. Abcd data release 7.0: Release notes and assessment schedule. https://docs.abcdstudy.org/v/7_0_2/documentation/release_notes/7_0.html, 2026. Release 13 May 2026; accessed 21 September 2026.

Thomas M. Achenbach and Leslie A. Rescorla. *Manual for the ASEBA School-Age Forms & Profiles*. University of Vermont, Research Center for Children, Youth, & Families, Burlington, VT, 2001. ISBN 978-0-938565-73-4. URL https://store.aseba.org/MANUAL-FOR-THE-ASEBA-SCHOOL-AGE-FORMS-PROFILES/productinfo/505/.

Mark J. Adams, Fabian Streit, Xiangrui Meng, Swapnil Awasthi, Brett N. Adey, Karmel W. Choi, et al. Trans-ancestry genome-wide study of depression identifies 697 associations implicating cell types and pharmacotherapies. *Cell*, 188(3):640–652.e9, 2025. doi: 10.1016/j.cell.2024.12.002. URL https://doi.org/10.1016/j.cell.2024.12.002.

Thomas D. Als, Mitja I. Kurki, Jakob Grove, Georgios Voloudakis, Karen Therrien, Elisa Tasanko, et al. Depression pathophysiology, risk prediction of recurrence and comorbid psychiatric disorders using genome-wide analyses. *Nature Medicine*, 29(7):1832–1844, 2023. doi: 10.1038/s41591-023-02352-1. URL https://doi.org/10.1038/s41591-023-02352-1.

A. Angold, E. J. Costello, A. Erkanli, and C. M. Worthman. Pubertal changes in hormone levels and depression in girls. *Psychological Medicine*, 29(5):1043–1053, 1999. doi: 10.1017/S0033291799008946. URL https://doi.org/10.1017/S0033291799008946.

Daniel J. Benjamin, David Cesarini, Patrick Turley, and Alexander Strudwick Young. Social-Science Genomics: Progress, Challenges, and Future Directions. *Journal of Economic Literature*, forthcoming. URL https://www.aeaweb.org/articles?id=10.1257/jel.20261604. Working-paper version: NBER Working Paper 32404, 2024. doi: 10.3386/w32404. URL https://doi.org/10.3386/w32404.

Brendan Bulik-Sullivan, Hilary K. Finucane, Verneri Anttila, Alexander Gusev, Felix R. Day, Po-Ru Loh, et al. An atlas of genetic correlations across human diseases and traits. *Nature Genetics*, 47(11):1236–1241, 2015. doi: 10.1038/ng.3406.

Na Cai, Joana A. Revez, Mark J. Adams, Till F. M. Andlauer, Gerome Breen, Enda M. Byrne, et al. Minimal phenotyping yields genome-wide association signals of low specificity for major depression. *Nature Genetics*, 52(4):437–447, 2020. doi: 10.1038/s41588-020-0594-5. URL https://doi.org/10.1038/s41588-020-0594-5.

Claire E. Campbell, Jessica Morrel, W. James Gauderman, and Megan M. Herting. Effects of genetic risk for depression and household socioeconomic status on emotional behavior and brain development in early adolescence. *Developmental Cognitive Neuroscience*, 79:101692, 2026. doi: 10.1016/j.dcn.2026.101692. URL https://doi.org/10.1016/j.dcn.2026.101692.

Razieh Chegeni, Ragnhild Bang Nes, Rosa Cheesman, Ziada Ayorech, Eivind Ystrøm, and Espen Røysamb. Direct and indirect parental genetic effects on offspring susceptibility to internalizing problems across development. *Nature Mental Health*, 4(5):754–765, 2026. doi: 10.1038/s44220-026-00619-3. URL https://doi.org/10.1038/s44220-026-00619-3.

Jacob Cohen. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46, 1960. doi: 10.1177/001316446002000104. URL https://doi.org/10.1177/001316446002000104.

Andrew Dahl, Michael Thompson, Ulzee An, Morten Krebs, Vivek Appadurai, Richard Border, et al. Phenotype integration improves power and preserves specificity in biobank-based genetic

studies of major depressive disorder. *Nature Genetics*, 55(12):2082–2093, 2023. doi: 10.1038/s41588-023-01559-9. URL https://doi.org/10.1038/s41588-023-01559-9.

Yi Ding, Kangcheng Hou, Ziqi Xu, Aditya Pimplaskar, Ella Petter, Kristin Boulier, et al. Polygenic scoring accuracy varies across the genetic ancestry continuum. *Nature*, 618 (7966):774–781, 2023. doi: 10.1038/s41586-023-06079-4. URL https://doi.org/10.1038/s41586-023-06079-4.

E. C. Fieller. Some Problems in Interval Estimation. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 16(2):175–185, 1954. doi: 10.1111/j.2517-6161.1954.tb00159.x. URL https://doi.org/10.1111/j.2517-6161.1954.tb00159.x.

E. I. Fried, R. M. Nesse, C. Guille, and S. Sen. The differential influence of life stress on individual symptoms of depression. *Acta Psychiatrica Scandinavica*, 131(6):465–471, 2015. doi: 10.1111/acps.12395. URL https://doi.org/10.1111/acps.12395.

Eiko I. Fried and Randolph M. Nesse. Depression is not a consistent syndrome: An investigation of unique symptom patterns in the STAR*D study. *Journal of Affective Disorders*, 172:96–102, 2015. doi: 10.1016/j.jad.2014.10.010. URL https://doi.org/10.1016/j.jad.2014.10.010.

Poppy Z. Grimes, Mark J. Adams, Gladi Thng, Amelia J. Edmonson-Stait, Yi Lu, Andrew McIntosh, et al. Genetic Architectures of Adolescent Depression Trajectories in 2 Longitudinal Population Cohorts. *JAMA Psychiatry*, 81(8):807–816, 2024. doi: 10.1001/jamapsychiatry.2024.0983. URL https://doi.org/10.1001/jamapsychiatry.2024.0983.

Andrew D. Grotzinger, Josefin Werme, Wouter J. Peyrot, Oleksandr Frei, Christiaan de Leeuw, Lucy K. Bicks, et al. Mapping the genetic landscape across 14 psychiatric disorders. *Nature*, 649(8096):406–415, 2026. doi: 10.1038/s41586-025-09820-3. URL https://doi.org/10.1038/s41586-025-09820-3.

Junming Guan, Tammy Tan, Seyed Moeen Nehzati, Michael Bennett, Patrick Turley, Daniel J. Benjamin, and Alexander Strudwick Young. Family-based genome-wide association study designs for increased power and robustness. *Nature Genetics*, 57(4):1044–1052, 2025. doi: 10.1038/s41588-025-02118-0. URL https://doi.org/10.1038/s41588-025-02118-0.

Sture Holm. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6(2):65–70, 1979. URL https://www.jstor.org/stable/4615733.

Laurence J. Howe, Michel G. Nivard, Tim T. Morris, Ailin F. Hansen, Humaira Rasheed, Yoonsu Cho, et al. Within-sibship genome-wide association analyses decrease bias in estimates of direct genetic effects. *Nature Genetics*, 54(5):581–592, 2022. doi: 10.1038/s41588-022-01062-7. URL https://doi.org/10.1038/s41588-022-01062-7.

Floris Huider, Yuri Milaneschi, René Pool, Bernardo de A. P. C. Maciel, Scott D. Gordon, Jihee Han, et al. Genetics of major depressive disorder in a homogeneous population with uniform phenotyping. *Molecular Psychiatry*, 31(10):5889–5899, 2026. doi: 10.1038/s41380-026-03666-5. URL https://doi.org/10.1038/s41380-026-03666-5.

S. C. Kanjira, M. J. Adams, Y. Jiang, C. Tian, 23andMe Research Team, C. M. Lewis, et al. Polygenic prediction of major depressive disorder and related traits in African ancestries UK Biobank participants. *Molecular Psychiatry*, 30(1):151–157, 2025. doi: 10.1038/s41380-024-02662-x. URL https://doi.org/10.1038/s41380-024-02662-x.

Matthew C. Keller. Gene × Environment Interaction Studies Have Not Properly Controlled for Potential Confounders: The Problem and the (Simple) Solution. *Biological Psychiatry*, 75(1):18–24, 2014. doi: 10.1016/j.biopsych.2013.09.006. URL https://doi.org/10.1016/j.biopsych.2013.09.006.

Kenneth S. Kendler, Michael C. Neale, Ronald C. Kessler, Andrew C. Heath, and Lindon J. Eaves. The lifetime history of major depression in women: Reliability of diagnosis and heritability. *Archives of General Psychiatry*, 50(11):863–870, 1993. doi: 10.1001/archpsyc.1993.01820230054003. URL https://doi.org/10.1001/archpsyc.1993.01820230054003.

Amy J.H. Kind and William R. Buckingham. Making Neighborhood-Disadvantage Metrics Accessible — The Neighborhood Atlas. *New England Journal of Medicine*, 378(26):2456–2458, 2018. doi: 10.1056/NEJMp1802313. URL https://doi.org/10.1056/NEJMp1802313.

Kurt Kroenke, Robert L. Spitzer, and Janet B. W. Williams. The PHQ-9: Validity of a brief depression severity measure. *Journal of General Internal Medicine*, 16(9):606–613, 2001. doi: 10.1046/j.1525-1497.2001.016009606.x. URL https://doi.org/10.1046/j.1525-1497.2001.016009606.x.

Sang Hong Lee, Michael E. Goddard, Naomi R. Wray, and Peter M. Visscher. A Better Coefficient of Determination for Genetic Profile Analysis. *Genetic Epidemiology*, 36(3):214–224, 2012. doi: 10.1002/gepi.21614. URL https://doi.org/10.1002/gepi.21614.

Luke R. Lloyd-Jones, Jian Zeng, Julia Sidorenko, Loïc Yengo, Gerhard Moser, Kathryn E. Kemper, et al. Improved polygenic prediction by Bayesian multiple regression on summary statistics. *Nature Communications*, 10(1):5086, 2019. doi: 10.1038/s41467-019-12653-0. URL https://doi.org/10.1038/s41467-019-12653-0.

Andrew M. McIntosh, Patrick F. Sullivan, and Cathryn M. Lewis. Uncovering the Genetic Architecture of Major Depression. *Neuron*, 102(1):91–103, 2019. doi: 10.1016/j.neuron.2019.03.022. URL https://doi.org/10.1016/j.neuron.2019.03.022.

T. E. Moffitt, A. Caspi, A. Taylor, J. Kokaua, B. J. Milne, G. Polanczyk, et al. How common are common mental disorders? Evidence that lifetime prevalence rates are doubled by prospective versus retrospective ascertainment. *Psychological Medicine*, 40(6):899–909, 2010. doi: 10.1017/S0033291709991036. URL https://doi.org/10.1017/S0033291709991036.

Spencer Moore, Ivan Davidson, Jonathan Anomaly, Jeremiah H. Li, Mohammad Ahangari, Lauren Moissiy, Michael Christensen, Alexander Strudwick Young, David Stern, and Tobias Wolfram. Development and validation of polygenic scores for within-family prediction of disease risks. *medRxiv*, page 2025.08.06.25333145, 2025. doi: 10.1101/2025.08.06.25333145. URL https://www.medrxiv.org/content/10.1101/2025.08.06.25333145v2. Version 2, posted 14 August 2025.

Thuy-Dung Nguyen, Arvid Harder, Ying Xiong, Kaarina Kowalec, Sara Hägg, Na Cai, et al. Genetic heterogeneity and subtypes of major depression. *Molecular Psychiatry*, 27(3):1667–1675, 2022. doi: 10.1038/s41380-021-01413-6. URL https://doi.org/10.1038/s41380-021-01413-6.

Ulf Olsson. Maximum likelihood estimation of the polychoric correlation coefficient. *Psychometrika*, 44(4):443–460, 1979. doi: 10.1007/BF02296207. URL https://doi.org/10.1007/BF02296207.

Sarah E. Paul, Sarah M. C. Colbert, Aaron J. Gorelik, Emma C. Johnson, Alexander S. Hatoum, David A. A. Baranger, et al. A phenome-wide association study of cross-disorder genetic liability in youth genetically similar to individuals from European reference populations. *Nature Mental Health*, 2(11):1327–1341, 2024. doi: 10.1038/s44220-024-00313-2. URL https://doi.org/10.1038/s44220-024-00313-2.

Florian Privé, Hugues Aschard, Shai Carmi, Lasse Folkersen, Clive Hoggart, Paul F. O'Reilly, et al. Portability of 245 polygenic scores when derived from the UK Biobank and applied to

9 ancestry groups from the same cohort. *The American Journal of Human Genetics*, 109(1): 12–23, 2022. doi: 10.1016/j.ajhg.2021.11.008. URL https://doi.org/10.1016/j.ajhg.2021.11.008.

Jian Rong, Xueqin Wang, Pan Cheng, Dan Li, and Dahai Zhao. Global, regional and national burden of depressive disorders and attributable risk factors, from 1990 to 2021: results from the 2021 Global Burden of Disease study. *The British Journal of Psychiatry*, 227(4):688–697, 2025. doi: 10.1192/bjp.2024.266. URL https://doi.org/10.1192/bjp.2024.266.

Rachel H. Salk, Janet S. Hyde, and Lyn Y. Abramson. Gender differences in depression in representative national samples: Meta-analyses of diagnoses and symptoms. *Psychological Bulletin*, 143(8):783–822, 2017. doi: 10.1037/bul0000102. URL https://doi.org/10.1037/bul0000102.

Tabea Schoeler, Doug Speed, Eleonora Porcu, Nicola Pirastu, Jean-Baptiste Pingault, and Zoltán Kutalik. Participation bias in the UK Biobank distorts genetic associations and downstream analyses. *Nature Human Behaviour*, 7(7):1216–1227, 2023. doi: 10.1038/s41562-023-01579-9. URL https://doi.org/10.1038/s41562-023-01579-9.

Xueyi Shen, David M. Howard, Mark J. Adams, W. David Hill, Toni-Kim Clarke, Major Depressive Disorder Working Group of the Psychiatric Genomics Consortium, et al. A phenome-wide association and Mendelian Randomisation study of polygenic risk for depression in UK Biobank. *Nature Communications*, 11(1):2301, 2020. doi: 10.1038/s41467-020-16022-0. URL https://doi.org/10.1038/s41467-020-16022-0.

John R. Shorter, Joëlle A. Pasman, Siim Kurvits, Andreas Jangmo, Joonas Naamanka, Arvid Harder, et al. Genome-wide association analyses identify distinct genetic architectures for early-onset and late-onset depression. *Nature Genetics*, 57(12):2972–2979, 2025. doi: 10.1038/s41588-025-02396-8. URL https://doi.org/10.1038/s41588-025-02396-8.

Robert L. Spitzer, Kurt Kroenke, Janet B. W. Williams, and Bernd Löwe. A brief measure for assessing generalized anxiety disorder: The GAD-7. *Archives of Internal Medicine*, 166 (10):1092–1097, 2006. doi: 10.1001/archinte.166.10.1092. URL https://doi.org/10.1001/archinte.166.10.1092.

Cathie Sudlow, John Gallacher, Naomi Allen, Valerie Beral, Paul Burton, John Danesh, 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 Medicine*, 12(3):e1001779, 2015. doi: 10.1371/journal.pmed.1001779. URL https://doi.org/10.1371/journal.pmed.1001779.

Patrick F. Sullivan, Michael C. Neale, and Kenneth S. Kendler. Genetic Epidemiology of Major Depression: Review and Meta-Analysis. *American Journal of Psychiatry*, 157(10):1552–1562, 2000. doi: 10.1176/appi.ajp.157.10.1552. URL https://doi.org/10.1176/appi.ajp.157.10.1552.

Tammy Tan, Hariharan Jayashankar, Junming Guan, Seyed Moeen Nehzati, Mahdi Mir, Michael Bennett, et al. Family-GWAS reveals effects of environment and mating on genetic associations. *medRxiv*, 2024. doi: 10.1101/2024.10.01.24314703. URL https://doi.org/10.1101/2024.10.01.24314703. Preprint.

Jodi T. Thomas, Jackson G. Thorp, Floris Huider, Poppy Z. Grimes, Rujia Wang, Pierre Youssef, et al. Sex-stratified genome-wide association meta-analysis of major depressive disorder. *Nature Communications*, 16(1):7960, 2025. doi: 10.1038/s41467-025-63236-1. URL https://doi.org/10.1038/s41467-025-63236-1.

Jackson G. Thorp, Adrian I. Campos, Andrew D. Grotzinger, Zachary F. Gerring, Jiyuan An, Jue-Sheng Ong, et al. Symptom-level modelling unravels the shared genetic architecture of anxiety and depression. *Nature Human Behaviour*, 5(10):1432–1442, 2021. doi: 10.1038/s41562-021-01094-9. URL https://doi.org/10.1038/s41562-021-01094-9.

Lisa Townsend, Kenneth Kobak, Catherine Kearney, Michael Milham, Charissa Andreotti, Jasmine Escalera, et al. Development of Three Web-Based Computerized Versions of the Kiddie Schedule for Affective Disorders and Schizophrenia Child Psychiatric Diagnostic Interview: Preliminary Validity Data. *Journal of the American Academy of Child & Adolescent Psychiatry*, 59(2):309–325, 2020. doi: 10.1016/j.jaac.2019.05.009. URL https://doi.org/10.1016/j.jaac.2019.05.009.

Victória Trindade Pons, Albertine J. Oldehinkel, and Hanna M. van Loo. Genetic nurture effects in depressive and anxiety disorders and symptoms, and in related traits. *Molecular Psychiatry*, 30(12):5694–5700, 2025. doi: 10.1038/s41380-025-03265-w. URL https://doi.org/10.1038/s41380-025-03265-w.

Patrick Turley, Raymond K. Walters, Omeed Maghzian, Aysu Okbay, James J. Lee, Mark Alan Fontana, et al. Multi-trait analysis of genome-wide association summary statistics using MTAG. *Nature Genetics*, 50(2):229–237, 2018. doi: 10.1038/s41588-017-0009-4. URL https://doi.org/10.1038/s41588-017-0009-4.

Jessica Tyrrell, Jie Zheng, Robin Beaumont, Kathryn Hinton, Tom G. Richardson, Andrew R. Wood, et al. Genetic predictors of participation in optional components of UK Biobank. *Nature Communications*, 12(1):886, 2021. doi: 10.1038/s41467-021-21073-y. URL https://doi.org/10.1038/s41467-021-21073-y.

Carl Veller and Graham M. Coop. Interpreting population- and family-based genome-wide association studies in the presence of confounding. *PLOS Biology*, 22(4):e3002511, 2024. doi: 10.1371/journal.pbio.3002511. URL https://doi.org/10.1371/journal.pbio.3002511.

Ying Wang, Jing Guo, Guiyan Ni, Jian Yang, Peter M. Visscher, and Loic Yengo. Theoretical and empirical quantification of the accuracy of polygenic scores in ancestry divergent populations. *Nature Communications*, 11(1):3865, 2020. doi: 10.1038/s41467-020-17719-y. URL https://doi.org/10.1038/s41467-020-17719-y.

Monika A. Waszczuk, Jiaju Miao, Anna R. Docherty, Andrey A. Shabalin, Katherine G. Jonas, Giorgia Michelini, et al. General v. specific vulnerabilities: polygenic risk scores and higher-order psychopathology dimensions in the Adolescent Brain Cognitive Development (ABCD) Study. *Psychological Medicine*, 53(5):1937–1946, 2023. doi: 10.1017/S0033291721003639. URL https://doi.org/10.1017/S0033291721003639.

Tobias Wolfram, Spencer Moore, Jeremiah H. Li, Jonathan Anomaly, Ivan Davidson, and Michael Christensen. Interpreting Polygenic Prediction of Cognitive Ability: Evidence for Direct, Reliable, and Portable Genetic Effects. *Intelligence & Cognitive Abilities*, 2(1):1–19, 2026. doi: 10.65550/001c.158459. URL https://doi.org/10.65550/001c.158459.

Alexander I. Young, Seyed Moeen Nehzati, Stefania Benonisdottir, Aysu Okbay, Hariharan Jayashankar, Chanwook Lee, et al. Mendelian imputation of parental genotypes improves estimates of direct genetic effects. *Nature Genetics*, 54(6):897–905, 2022. doi: 10.1038/s41588-022-01085-0. URL https://doi.org/10.1038/s41588-022-01085-0.

Zhili Zheng, Shouye Liu, Julia Sidorenko, Ying Wang, Tian Lin, Loic Yengo, et al. Leveraging functional genomic annotations and genome coverage to improve polygenic prediction of complex traits within and between ancestries. *Nature Genetics*, 56(5):767–777, 2024. doi: 10.1038/s41588-024-01704-y. URL https://doi.org/10.1038/s41588-024-01704-y.

Supplementary Information

This Supplementary Information describes the samples, measures and statistical models behind each result in the main text and reports additional results. Supplementary Note 1 covers the samples, genotype processing and polygenic scores used throughout. Supplementary Notes 2 to 7 follow the order of the main text. Complete association results for all outcomes are provided in the Supplementary Data workbook.

Throughout, "integrated score" refers to our depression polygenic score and "PGC score" to the benchmark score built from Psychiatric Genomics Consortium summary statistics. The two scores were always compared on identical participants, outcomes and covariates. UK Biobank population groups are labelled EUR (reference), AFR, EAS and SAS, and ABCD groups are labelled White (reference), Hispanic and African American (Supplementary Notes 1 and 7).

Supplementary Note 1: Samples, genotypes and polygenic scores

This note describes the validation samples, genotype processing, score standardization and the construction of the PGC benchmark scores. These elements are shared by all later analyses.

UK Biobank. We set aside for validation all genetically inferred full-sibling and parent–offspring pedigrees among participants in the United Kingdom genetic-ancestry group defined by Privé et al. (2022), and excluded every pedigree member from local discovery. Pedigrees were built using kinship, age and sex. The sibling validation sample comprised the reserve's sibships with no genotyped parent. Participants flagged as heterozygosity or missingness outliers (field 22027) or as having a sex-chromosome aneuploidy (field 22019) were excluded.

The depression analyses (Supplementary Note 2) additionally required British self-reported ethnicity (field 21000), the UK Biobank genetic ethnic grouping "Caucasian" (field 22006) and agreement between reported and genetic sex (fields 31 and 22001). Each participant needed a case or control classification and complete genotype and covariate data, but their siblings did not need to be classified. The latent internalizing analysis (Supplementary Note 3) instead required at least two siblings per family who had each answered at least one symptom item and had complete covariates.

ABCD. We used phenotypes from ABCD Release 7 (ABCD Study, 2026), genotypes imputed by the consortium, and the cohort's family and demographic records (family identifiers, parent-reported race and ethnicity, sex and genetic principal components). Family analyses used non-twin siblings and dizygotic twins, monozygotic twins and twins with missing or conflicting zygosity were excluded. A family entered the reference sample if at least two of its members had complete scores and covariates and enough internalizing data: at least two observed internalizing composites among the parent-reported Child Behavior Checklist (CBCL), the Youth Self-Report (YSR) and the youth Brief Problem Monitor (BPM), or one such composite plus at least three teacher-reported BPM items. The latent YSR analysis then required at least one answered YSR item per child, but not a second YSR-reporting sibling (Table S1). Broader population analyses used all eligible White children, including singletons, siblings and twins. Sex was the parent-reported sex at baseline, or the sex assigned at birth in the Release 7 participant table for the latent family analysis.

ABCD population groups were based on parent-reported race and ethnicity (Wolfram et al., 2026). Children reported as Hispanic formed the Hispanic group, whatever their reported race. Each remaining child was assigned to a group if exactly one of four categories applied: White, Black, Asian Indian, or East Asian (Chinese, Filipino, Japanese or Korean). We analysed the White (reference), Hispanic and Black (labelled African American) groups. The East Asian and Asian Indian groups were below the 150-child minimum for model fitting (Supplementary Note 7). Children reported in more than one of the four categories were excluded, and so were children reported in none of them.

Table S1: ABCD reference and YSR analysis sample flow

Selection	Children	Families
White genotyped children with both scores and complete reference covariates	6,325	5,129
Eligible non-twin siblings or confirmed dizygotic twins	1,398	761
Meeting the multi-instrument availability requirement	1,396	760
Reference sample: at least two eligible relatives per family	1,268	632

Selection	Children	Families
YSR analysis: at least one observed YSR item	1,172	593

Each row applies an additional restriction. Score moments and family means used the reference sample before YSR restriction. The final analysis included 17 families with only one YSR-observed child. Their family means included the other reference-eligible relatives.

Genotypes and scoring. In UK Biobank, imputed genotypes were restricted to biallelic variants with at most 5% missing calls, minor-allele frequency of at least 1% and Hardy–Weinberg $P \geq 10^{-50}$. Participants with more than 3% missing genotypes (10% in the African-ancestry groups) were removed. Full scores were computed with PLINK 2 from imputed dosages, as the average allele-weighted dosage over the scored variants. Weight variants that were absent from the target data or had incompatible alleles were skipped, and missing genotypes were replaced by their expected dosage given the allele frequency.

Scores for the family analyses were computed by SNIPar (Young et al., 2022). SNIPar used hard-called genotypes, converted from the imputed data for the family reserve after restricting to a list of well-imputed variants (imputation INFO > 0.99) and removing variants with Hardy–Weinberg $P < 10^{-6}$. Parental genotypes were imputed from the genotypes of siblings and the genome segments they share. Full-sibling and parent–offspring relationships among UK-ancestry participants were inferred with KING 2.3.2, and SNIPar built pedigrees using year of birth and sex. Identity-by-descent segments were inferred with SNIPar's `ibd.py` from directly genotyped array SNPs (607,402 of 784,256 retained), using the HapMap II genetic map and a genotyping-error probability of 10^{-4}. Child and parental scores were computed from the same variants and weights: those present in the child's genotypes, in the imputed parental genotypes and among the score's non-zero weights. The within-family analyses of depression (Supplementary Note 2) therefore used this parental-compatible version of each score, while the logistic model of depression used the full score.

In ABCD, scores were computed from hard calls derived from the consortium's imputed genotypes.

Score standardization. Scores were standardized to unit variance in the study sample of each analysis before restricting to participants with a particular outcome (for adult UK Biobank outcomes, within each outcome's sample, which changed the score SD by less than 0.3%).

PGC benchmark scores. In ABCD, the benchmark was an SBayesR score (Lloyd-Jones et al., 2019) computed from the published European PGC MDD2025 summary statistics without 23andMe (Adams et al., 2025), which include UK Biobank. Because those statistics include UK Biobank, the UK Biobank benchmark was rebuilt. The PGC statistics without UK Biobank were meta-analysed with a local UK Biobank depression GWAS that excluded every validation participant and their relatives (kinship ≥ 0.0442).

The local GWAS phenotype used observed evidence:

- *Cases* met a detailed CIDI-based lifetime definition, reported depression, reported a professional diagnosis, or had a hospital diagnosis of F32 or F33.
- *Exclusions* took priority: bipolar, psychotic or personality disorders.
- *Controls* required observed negative evidence and no ambiguous evidence. Ambiguous evidence was antidepressant use without a qualifying diagnosis, positive core depressive symptoms, probable depression, or another mood-disorder hospital code.

Unrelated participants were then selected without regard to case status. Associations were tested with GCTA fastGWA, using a linear model of the 0/1 phenotype on hard-called genotypes, with age, sex, genotyping array and principal components 1–20 as covariates. Variants required a minor-allele frequency above 0.01, at least 20 copies of the minor allele and genotype missingness below 0.02.

The local and PGC results were meta-analysed by allele-aligned inverse-variance weighting. Variants were removed if their allele frequencies differed by more than 0.10 between the two sources. Strand-ambiguous variants were also removed if either frequency was within 0.10 of 0.5. The effective sample size was $4/(1/N_{\text{cases}} + 1/N_{\text{controls}})$. Both benchmark scores were estimated with SBayesR using the European banded HapMap3 linkage-disequilibrium reference.

Discovery-signal comparison. The integrated GWAS (final multi-trait statistics, before effect estimation) was compared with the published PGC GWAS without 23andMe on 960,778 autosomal HapMap3 variants that were also present in the UK Biobank focal depression GWAS. The MHC region, variants with

allele frequency outside 0.01–0.99, strand-ambiguous variants and variants with $\chi^2 \geq 80$ in either GWAS were excluded. One cross-trait LD-score regression (Bulik-Sullivan et al., 2015) with 1000 Genomes European LD scores and free intercepts was fitted with the same fixed sample size for both GWASs, so that the ratio of the two heritability estimates equals the ratio of LD-score slopes ($N \times h^2_{\mathrm{SNP}}$). The 95% interval of the ratio was $\exp(\log r \pm 1.96\,\mathrm{SE})$, with the SE from a paired 200-block jackknife of the log ratio. Uniformity was checked with the ratio of mean χ^2 excess over the intercepts within LD-score quintiles, and the genetic correlation was the unconstrained LD-score estimate with its block-jackknife SE.

Similarity of the two benchmark GWASs. Cross-trait LD-score regression (Bulik-Sullivan et al., 2015) compared the published PGC GWAS with our PGC-plus-UK Biobank GWAS. The estimated genetic correlation on common HapMap sites was 1.011 (SE 0.00115, 95% interval 1.009–1.014), showing that the two sets of summary statistics are nearly identical.

Supplementary Note 2: Depression in UK Biobank and ABCD

Case and control definition. We used the LifetimeMDD definition (Cai et al., 2020; Dahl et al., 2023) (Table S2): Cases required at least two weeks of depressed mood or loss of interest, at least five of eight lifetime symptom domains and substantial interference with normal roles, or alternatively at least five of the nine current symptoms present nearly every day, including depressed mood or loss of interest, and the same impairment. As in the original definition (Cai et al., 2020), cases with a self-reported psychotic, bipolar or substance-use disorder were left unclassified, which implements DSM-5 criteria C–E for cases. Controls explicitly denied a period of prolonged depressed mood and did not meet the case rule, while participants who met neither definition were also left unclassified. Cai et al. (2020) describe the controls only as participants without the case criteria. Applied to the same white British quality-controlled sample, our implementation gave 16,189 cases and 49,846 controls (published: 16,301 and 50,870), case counts within 14 of the published counts at every assessment centre, and the published lead association (rs926552, odds ratio 0.902 versus 0.903). In the analysis sample below, 1,637 of the 1,679 cases qualified only through the lifetime route, 41 through both routes and one only through the current-symptom route.

Population and within-family models. The analysis sample comprised 7,054 participants (1,679 cases) from 6,085 families of the sibling validation sample (Supplementary Note 1). All models adjusted for sex, birth year and its square, assessment centre, genotyping array and principal components 1–20. The population association of the full score was estimated by logistic regression. Within-family associations were estimated with SNIPar (Young et al., 2022) using the parental-compatible score. SNIPar fits two linear mixed models with family random intercepts by restricted maximum likelihood: a child-only model, which estimates the population association, and a model that adds the sum of the two imputed parental scores, which estimates the direct (within-family) association. In these linear models the binary outcome was standardized. Chromosome scores were combined before SNIPar's genome-wide adjustment for assortative mating. SNIPar applies this adjustment only when the estimated correlation between the two parents' scores is large relative to its standard error. The adjustment was therefore applied to the integrated score but not to the PGC score, whose estimated correlation was too small ($r = 0.018$, SE 0.008, versus 0.008, SE 0.008), however omitting the adjustment changed the integrated score's direct/population ratio by less than 0.001. Coefficients were estimated by generalized least squares at the REML variance components, and we report SNIPar's model-based standard errors. Because each imputed parental score is the expected score given the siblings' genotypes and their identity-by-descent sharing, its imputation error enters the residual, so estimates and standard errors remain valid without further correction (Young et al., 2022). The interval for the ratio came from the whole-family bootstrap described below. The ratio of the direct to the population coefficient, $r = \widehat{\beta}_D/\widehat{\beta}_P$, summarizes how much of the population association is retained within families.

Stability of the depression classification. Reporting error in a single case–control classification attenuates associations with a polygenic score. To estimate how stable the classification is, we applied the same rules to the matching items of the first (2016–2017) and repeat (2022–2023) mental-health questionnaires. At each assessment a respondent was a control, a case, or unclassified. The unclassified group reported qualifying symptoms without meeting the case or control definition, missing answers alone did not place anyone in it, and excluded cases were removed. We used all 79,072 respondents with an interpretable state at both assessments, including people outside the validation sample, as long as they met the ancestry and quality criteria of the depression sample. Table S3C shows the transitions.

We summarized stability in three ways (Table S3D):

- *Binary kappa.* Cohen's kappa (Cohen, 1960), $\kappa = (p_o - p_e)/(1 - p_e)$, among respondents who were a case or a control at both assessments.
- *Polychoric correlation.* A three-state polychoric correlation (Olsson, 1979), which assumes one bivariate normal liability and was estimated by maximum likelihood with separate thresholds for each questionnaire.
- *Stability mixture.* A two-component Gaussian-copula mixture. It fixes the thresholds at the observed marginal proportions and estimates two ordered correlations and a mixing weight π:

$$p_{ab} = \pi p_{ab}(\rho_{\text{low}}) + (1 - \pi)p_{ab}(\rho_{\text{high}}), \qquad s = \pi\rho_{\text{low}} + (1 - \pi)\rho_{\text{high}}.$$

  The two components describe the response distribution, not clinical subgroups.

Using a stability estimate s as a reliability assumes a stable standardized liability T, measured by two equally reliable questionnaires whose errors are independent of each other, of T and of the score: $U_j = \sqrt{s}\,T + \sqrt{1 - s}\,\varepsilon_j$. Associations are then attenuated by $\sqrt{s}$, and we corrected the log odds ratio as $b/\sqrt{s}$, which is an approximation for covariate-adjusted logistic coefficients. The mixture reproduced the transition table more closely than the polychoric model: the largest difference between observed and fitted cell proportions was 0.62 versus 2.76 percentage points. Neither fit statistic accounts for family structure, and a closer fit does not by itself validate the reliability interpretation. Figure 1A uses the polychoric correlation, whereas the mixture implies a slightly larger correction (Table S3A).

Liability-scale variance explained. We converted the full-score log odds ratio into approximate liability-scale R^2 with the conversion in Supplementary Table 5 of Moore et al. (2025), which includes the ascertainment adjustment of Lee et al. (2012). Let b be the log odds ratio per score SD, $\widetilde{b} = b/\sqrt{s}$ its corrected value ($s = 1$ without correction), P the case fraction of the analysis sample and $K = 0.15$ the assumed population prevalence. Under a binormal model, in which scores are normally distributed with equal variance in cases and controls, the standardized case–control difference in the score, d, satisfies

$$\widetilde{b} = d\sqrt{1 + d^2P(1 - P)}, \qquad R^2_o = \frac{d^2P(1 - P)}{1 + d^2P(1 - P)}.$$

With $a = \widetilde{b}^2P(1 - P)$, the observed-scale value is

$$R^2_o = \frac{4a}{\left(1 + \sqrt{1 + 4a}\right)^2}.$$

With $t = \Phi^{-1}(1 - K)$, $z = \phi(t)$ and $i = z/K$, define

$$C = \frac{[K(1 - K)]^2}{z^2P(1 - P)}, \qquad \theta = i\frac{P - K}{1 - K}\left(i\frac{P - K}{1 - K} - t\right).$$

The population and model-implied within-family liability summaries are then

$$R^2_{\text{pop}}(s) = \frac{CR^2_o}{1 + C\theta R^2_o}, \qquad R^2_{\text{WF,implied}}(s) = R^2_{\text{pop}}(s)\,r^2.$$

These summaries are derived from association coefficients under a binormal approximation and an assumed prevalence. Sex-specific values (Results) refitted the logistic model separately in women and men, with a sex-specific repeat-stability correction and the same $K = 0.15$ in both sexes.

Uncertainty. Confidence intervals came from 4,000 whole-family bootstrap samples, stratified by membership of the validation and calibration samples. Repeat respondents without relatives in the reserve were resampled as single-person units. The same bootstrap weights were used for both scores and all specifications, so paired differences account for the overlap between samples. In each bootstrap sample we refitted the logistic models and the four SNIPar models (two per score), recalculated the transition statistics and the mixture thresholds, and updated the case fraction and the ascertainment terms C and θ. Score references and K were held fixed. Intervals are bootstrap percentiles, and ratios above one and R^2 values were not truncated.

Observed-scale variance explained. The main analyses compare the scores on different scales: liability R^2 for depression, latent coefficients for internalizing and coefficient profiles for the phenome. To compare them on one scale, we computed the incremental variance explained on the observed scale,

$(\mathrm{SSE}_0 - \mathrm{SSE}_1)/\mathrm{TSS}$, where SSE_0 and SSE_1 are the residual sums of squares without and with the score and TSS is the total sum of squares. Binary outcomes were analysed as 0/1 variables without correction for measurement error. Both scores used identical rows and covariates. The UK Biobank internalizing outcome was the mean of the raw PHQ, GAD and neuroticism sums among participants who answered all 28 items, and the anxiety and neuroticism outcomes were the corresponding scale sums. The ABCD outcome was the mean of the three YSR scale sums, averaged over visits with complete answers. The adult panel averages all 43 UK Biobank life, health and disease outcomes with equal weight, without selection on significance. Uncertainty came from 40 whole-family deletions shared by both scores, refitting every model and variance, with paired t_{39} Fieller intervals. A deletion in which an outcome was constant contributed zero for both scores. Table S4 gives the results.

Probit models. The population and within-family models of binary outcomes are linear. As a check, we refitted the child-only and parental-adjusted models with a probit link and a family random intercept,

$$\Pr(Y_{if} = 1 \mid G_{if}, P_f, C_{if}, u_f) = \Phi(\alpha + \beta G_{if} + \eta P_f + \gamma^\top C_{if} + u_f), \qquad u_f \sim N(0, \tau^2),$$

with $\eta = 0$ in the child-only model, for core depression, the 23 diagnoses and the seven binary pain outcomes. Participants, covariates and score units were identical to the linear models. We compared the linear coefficients, expressed as changes in probability, with the probit average marginal effects integrated over the family intercept,

$$\mathrm{AME} = \frac{1}{n} \sum_i \frac{\beta}{\sqrt{1+\tau^2}}\, \phi\!\left(\frac{a_i}{\sqrt{1+\tau^2}}\right),$$

where a_i is the linear predictor without the random intercept. Multiplied by 100, both are percentage points per score SD. For depression, intervals for the estimates, the link differences and the parental-adjustment comparisons came from 200 paired whole-family bootstrap samples that refitted both links, both scores and both parental specifications. For the adult outcomes, probit intervals used the delta method, conditional on the fitted random-intercept variance and the observed covariate distribution, and the linear intervals used the family-clustered covariance. These comparisons are descriptive, without paired tests.

In the depression sample (7,054 participants, 1,679 cases, 6,085 families), the two links agreed closely. For the integrated score, the child-only association was 7.499 percentage points per SD in the linear model (95% CI 6.673 to 8.438) and 7.453 (6.637 to 8.312) in the probit model, a difference of −0.046 (−0.177 to 0.067). After parental adjustment it was 7.737 (5.740 to 9.489) and 7.709 (5.785 to 9.374). For the PGC score, the corresponding values were 6.113 (5.194 to 7.085) and 6.107 (5.261 to 7.065) without, and 6.254 (4.782 to 8.144) and 6.273 (4.744 to 8.026) with, parental adjustment. Across the 30 adult binary outcomes, both scores and both specifications, the probit-minus-linear differences ranged from −0.054 to 0.045 percentage points, and the largest absolute difference was for the integrated score's parental-adjusted association with any chronic pain. For atopic dermatitis (19 cases), vitiligo (37 cases) and testicular cancer (60 cases), the probit family variance converged to its boundary of zero.

Depression diagnosis in ABCD. ABCD contributed no discovery data. The prespecified ABCD outcome counted a child as a case if the youth or a parent reported present, past or partially remitted major depressive disorder on the computerized K-SADS (Townsend et al., 2020) at baseline or at the two-, four- or six-year follow-up. All other children with a diagnostic interview were controls. The analysis used the 6,323 White children of the population sample with K-SADS data (1,855 cases, 5,127 families) in the paper's reference score units. Logistic regressions adjusted for sex, age at the last interview and PCs 1–10. Per reference-score SD, the log-odds coefficient was 0.377 for the integrated score and 0.317 for the PGC score (19.0% higher, 95% CI 8.9–29.9%). Liability-R^2 was 4.9% (3.2–6.5%) and 3.5% (2.3–4.7%), a relative gain of 39.4% (16.7–66.2%, $P = 0.002$). Intervals used the 40 whole-family jackknife groups of the observed-scale R^2 comparisons, with Fieller intervals for ratios. The liability conversion used the sample prevalence (29%) because the cohort was not ascertained on depression, and no repeat correction was applied. Because either informant at any of four interviews could establish a case, and the children had not yet passed through the period of risk, this outcome differs from adult lifetime depression, so only the relative gain is comparable with UK Biobank.

Table S2: UK Biobank depression definition (LifetimeMDD)

Component	Rule (UK Biobank field)
Lifetime route	At least two weeks of depressed mood (20446) or loss of interest (20441), plus at least five of eight symptom domains: depressed mood (20446), loss of interest (20441), weight change (20536), sleep change (20532), tiredness (20449), worthlessness (20450), difficulty concentrating (20435) and thoughts of death (20437). "Yes" counts as present, and weight gain, weight loss or both count as weight change. The questionnaire has no psychomotor item. No additional requirement on the fraction of the day or the frequency of depressed days (20436, 20439) was imposed.
Current route	All nine current-symptom items answered, with at least five present "nearly every day", including depressed mood or loss of interest: depressed mood (20510), loss of interest (20514), appetite (20511), sleep (20517), psychomotor change (20518), tiredness (20519), feelings of inadequacy (20507), concentration (20508) and thoughts of suicide or self-harm (20513). This is an item count, not a PHQ sum-score cutoff.
Impairment	Both routes also require "a lot" of interference with normal roles during the worst lifetime episode (20440), as in the original definition (Cai et al., 2020). There is no separate measure of current impairment.
Case exclusions	Otherwise qualifying cases are left unclassified if they self-reported schizophrenia (20002 code 1289), mania or bipolar disorder (1291), alcohol dependency (1408), opioid dependency (1409) or other substance abuse or dependency (1410), or if field 20126 indicates bipolar I or II disorder. As in the original definition, these exclusions apply only to cases.
Controls	An explicit "No" to prolonged depressed mood (20446) and failure to meet the symptom-plus-impairment rule. A negative answer on loss of interest is not required, and the case exclusions are not applied to controls. The original publication describes controls only as participants without the case criteria. This rule matches the size of the published control group to within 2% (Supplementary Note 2).
Missing answers	Missing or declined answers never count as symptoms. The lifetime route counts the symptoms observed without requiring every follow-up item, whereas the current route requires all nine items. A missing impairment answer cannot make a case, and a missing mood answer cannot make a control. Participants who meet neither definition, including excluded cases, remain unclassified.
Repeat questionnaire	The same rules were applied to the 2022–2023 questionnaire using the matching fields (first-questionnaire field in parentheses): 29011 (20446), 29012 (20441), 29021 (20536), 29022 (20532), 29018 (20449), 29027 (20450), 29026 (20435), 29029 (20437), 29031 (20440), 29003 (20510), 29002 (20514), 29006 (20511), 29004 (20517), 29009 (20518), 29005 (20519), 29007 (20507), 29008 (20508) and 29010 (20513). In the repeat questionnaire, weight-change codes 0, 1 and 2 denote a change (3, no change) and impairment code 0 denotes "a lot". Case exclusions and eligibility were held fixed across the two questionnaires.

Table S3: UK Biobank depression: liability-scale summaries, score comparisons and stability of the classification

A. Approximate liability-R^2, % (95% CI)

	Population		Implied within-family	
Specification	Integrated	PGC	Integrated	PGC
Uncorrected	7.0 (5.5, 8.7)	4.6 (3.4, 6.1)	7.4 (4.6, 11.2)	4.8 (2.7, 7.9)
Binary kappa	8.5 (6.7, 10.5)	5.6 (4.2, 7.4)	9.0 (5.6, 13.5)	5.9 (3.3, 9.6)
Polychoric	9.1 (7.2, 11.3)	6.1 (4.5, 7.9)	9.6 (6.0, 14.5)	6.3 (3.5, 10.3)
Stability mixture	9.3 (7.4, 11.5)	6.2 (4.6, 8.1)	9.9 (6.1, 14.8)	6.5 (3.6, 10.6)

B. Paired integrated-minus-PGC comparisons (95% CI)

	Population		Implied within-family	
Specification	Difference, pp	Relative gain, %	Difference, pp	Relative gain, %
Uncorrected	2.35 (1.44, 3.35)	50.8 (28.2, 79.8)	2.58 (0.73, 4.64)	53.4 (13.4, 114.8)
Binary kappa	2.82 (1.73, 4.02)	50.0 (27.8, 78.5)	3.10 (0.86, 5.58)	52.6 (12.9, 113.5)
Polychoric	3.01 (1.85, 4.28)	49.7 (27.6, 78.0)	3.31 (0.90, 5.96)	52.3 (12.7, 113.0)
Stability mixture	3.07 (1.89, 4.37)	49.6 (27.6, 77.8)	3.38 (0.92, 6.08)	52.1 (12.6, 112.8)

C. Repeat-status transitions

	Repeat assessment, 2022–2023			
Initial assessment, 2016–2017	Control	Unclassified	Case	Total
Control	31,052	4,253	505	35,810
Unclassified	11,605	17,484	2,606	31,695
Case	1,559	4,494	5,514	11,567
Total	44,216	26,231	8,625	79,072

D. Repeat-stability inputs

Correction	Repeat N	Input s (95% CI)	Calibration
Binary kappa	38,630	0.810 (0.802, 0.818)	Case/control agreement
Polychoric	79,072	0.751 (0.747, 0.757)	Single Gaussian copula
Stability mixture	79,072	0.733 (0.727, 0.739)	Two-component Gaussian copula

Notes. Panels A and B use the 7,054 participants (1,679 cases) described in Supplementary Note 2, with identical covariates and score units for both scores and an assumed prevalence of 0.15. "Uncorrected" uses $s = 1$, and the other rows divide the log odds ratio by $\sqrt{s}$ before conversion. Within-family values are model-implied from the SNIPar direct-to-population ratio. Panel B gives differences in percentage points and relative gains, $100(R^2_{\text{Integrated}}/R^2_{\text{PGC}} - 1)$. Intervals come from 4,000 paired whole-family bootstrap samples.

Panel C includes all respondents classified at both assessments. The median interval between assessments was 6.2 years, and 25,022 respondents (31.6%) changed category. Of the initial cases, 47.7% were cases again, 38.9% were unclassified and 13.5% were controls. Kappa (Panel D) excludes respondents who were unclassified at either assessment, whereas the polychoric and mixture estimates use all three states.

Table S4: Incremental observed-scale variance explained on harmonised samples

		Added R^2, %			
Outcome	N	PGC	Integrated	Gain, % (95% CI)	Outcomes
UK Biobank: depression	7,054	2.434	3.685	51.4 (26.7, 90.6)	1
UK Biobank: internalizing	9,895	2.418	4.697	94.2 (68.1, 131.1)	1
ABCD: internalizing	5,696	2.161	3.179	47.1 (22.5, 81.7)	1
UK Biobank: anxiety	11,839	1.540	2.818	82.9 (59.4, 115.1)	1
UK Biobank: neuroticism	29,500	2.356	4.429	88.0 (72.6, 106.9)	1
UK Biobank: 43 adult outcomes	10,387–36,316	0.275	0.432	57.2 (41.7, 76.5)	43

N is the number of participants (a range for the adult panel). Both scores use identical observations for each outcome. Values are uncorrected observed-scale R^2. The adult panel gives equal-weight means across outcomes, and gains compare those means. Intervals are paired 95% Fieller intervals from 40 whole-family deletions. For the 1,854 ABCD phenome outcomes of Table S10, the same jackknife gave mean observed R^2 of 0.237% for the PGC score and 0.295% for the integrated score, a 24.7% gain (95% CI 10.8 to 40.5%), and the corresponding multiplier interval (Supplementary Note 5) is 10.8 to 41.4%.

Supplementary Note 3: Latent internalizing within families

This note concerns the analyses presented in Figure 1B. It describes the symptom items, the factor models, the separation of each score into between-family and within-family components, and the complete-response checks.

Symptom items. UK Biobank used 28 items (Table S5): the nine PHQ-9 (Kroenke et al., 2001) and seven GAD-7 (Spitzer et al., 2006) items from the first online mental-health questionnaire (2016–2017), scored 0–3, and the twelve yes/no items of the EPQ-R-S neuroticism scale from the baseline assessment (2006–2010). ABCD used the 30 items of the anxious/depressed, withdrawn/depressed and somatic-complaints scales of the Youth Self-Report (Achenbach and Rescorla, 2001), which the children completed at the 4.5-, 5.5-, 6- and 7-year follow-ups, the YSR was not administered at baseline. Item 91 (suicidal thoughts) was not administered in ABCD. Each ABCD item was averaged over the visits at which the child answered it, and one answer sufficed.

Factor models. In the primary hierarchical model, each item loaded on its scale factor (PHQ, GAD and neuroticism in UK Biobank, and the three YSR scales in ABCD), and the three scale factors loaded on a general internalizing factor. In the alternative bifactor model, each item loaded on both the general factor and its scale factor, and all factors were uncorrelated. Neither model allowed correlated residuals. UK Biobank models treated the items as ordinal and were estimated by WLSMV with the theta parameterization, using all available pairs of items. ABCD models treated the averaged items as continuous and were estimated by robust maximum likelihood (MLR) with full-information handling of missing items and clustering by family. The factor structures were chosen pragmatically during exploratory model development, to summarize the variance shared across the symptom items in a single, less noisy outcome, and were then held fixed for all score comparisons. We do not interpret them as the definitive structure of internalizing psychopathology. They serve as a common target for comparing the scores, and the comparisons were essentially unchanged under an alternative bifactor specification (Supplementary Table S6). Fit indices (Table S6C) come from the models without the scores. In ABCD the bifactor model fitted better than the hierarchical model, but the score comparisons were nearly identical under both structures (Table S6A).

Within-family associations. Each child's score Z was split into the family mean Z_B and the child's deviation from it, $Z_W = Z - Z_B$. All genetic and covariate paths pointed to the general factor:

$$I_{if} = \beta_W Z_{W,if} + \beta_B Z_{B,f} + \gamma^\top C_{if} + \zeta_{if}.$$

The coefficient β_W is the within-family association. Population associations were estimated on the same participants, with the total score in place of its two components. Covariates were sex, birth year and its square (each standardized) and principal components 1–20 in UK Biobank, and sex, mean YSR age and principal components 1–10 in ABCD. Coefficients were divided by the total SD of the latent factor, $\sqrt{\psi + b^\top \Sigma_X b}$, where ψ is the residual variance, b the vector of path coefficients and Σ_X the covariance of the predictors. The within-family deviation was not rescaled. Factors were oriented so that higher values mean more symptoms. Family means used every scored sibling in the reference sample (Supplementary Note 1), each weighted equally. An ABCD family could therefore contribute a single child with YSR data whose family mean included the other siblings.

Confidence intervals. Due to WLSMV, in UK Biobank, families were divided into 40 groups. Each group was left out in turn and all parameters were re-estimated, with the score, item and covariate transformations held fixed. The covariance matrix was 39/40 times the sum of cross-products of the deviations of the 40 leave-one-group-out estimates, and intervals used the t distribution with 39 degrees of freedom. In ABCD, MLR allowed for intervals using family-level influence functions derived from the likelihood scores. These propagated the uncertainty in the total latent variance, including the empirical covariance of the predictors, with the finite-sample factor $\sqrt{G/(G-1)}$ for G families. Because both scores were analysed on the same families, the same jackknife groups or influences also gave the covariance needed for paired comparisons. Relative gains were computed on the log-ratio scale. The retention comparison (Table S7) contrasts $100\,\beta_W/\beta_P$ between the scores, where β_P is the population coefficient in the same family sample. Every reported latent model, including each jackknife refit, converged, was identified, had admissible covariance matrices and had a projected gradient below 10^{-3}.

Complete-response checks. The primary analyses used every answered item, which assumes that missing items are missing at random. We therefore repeated the UK Biobank analyses among participants who answered all items of the relevant scale, or all 28 items. Completeness was assessed for each participant,

so a family could contribute a single complete responder. The family decomposition and all transformations were unchanged (Table S6B).

Table S5: Symptom items of the latent internalizing models

Cohort	Scale (items)	Items (UK Biobank field or YSR item number)	Coding
UK Biobank	PHQ-9 depression (9)	Depressed mood (20510), loss of interest (20514), sleep (20517), tiredness (20519), appetite (20511), feelings of inadequacy (20507), concentration (20508), slowness or restlessness (20518), thoughts of suicide or self-harm (20513)	0–3 (not at all to nearly every day), 2016–2017 questionnaire
UK Biobank	GAD-7 anxiety (7)	Nervousness or anxiety (20506), inability to stop worrying (20509), worrying too much (20520), trouble relaxing (20515), restlessness (20516), irritability (20505), foreboding (20512)	0–3, 2016–2017 questionnaire
UK Biobank	EPQ-R-S neuroticism (12)	Mood swings (1920), miserableness (1930), irritability (1940), hurt feelings (1950), fed-up feelings (1960), nervous feelings (1970), worrier or anxious feelings (1980), tense or "highly strung" (1990), worrying after embarrassment (2000), suffering from "nerves" (2010), loneliness or isolation (2020), guilty feelings (2030)	Yes/no, baseline assessment 2006–2010
ABCD	YSR anxious/depressed (12)	Items 14, 29, 30, 31, 32, 33, 35, 45, 50, 52, 71 and 112	0–2, averaged over follow-ups
ABCD	YSR withdrawn/depressed (8)	Items 5, 42, 65, 69, 75, 102, 103 and 111	0–2, averaged over follow-ups
ABCD	YSR somatic complaints (10)	Items 47, 51, 54 and 56a–56g	0–2, averaged over follow-ups

All items were scored so that higher values indicate more symptoms. YSR responses (0 not true, 1 somewhat or sometimes true, 2 very or often true) come from the 4.5-, 5.5-, 6- and 7-year follow-ups.

Table S6: Within-family associations with latent internalizing

A. Integrated and PGC scores under two factor structures

Cohort	Factor model	Integrated β (95% CI)	PGC β (95% CI)	Gain, % (95% CI)
UK Biobank	Hierarchical	0.272 (0.244, 0.300)	0.213 (0.183, 0.242)	27.8 (17.4, 39.2)
	Bifactor	0.264 (0.237, 0.290)	0.208 (0.180, 0.236)	26.8 (16.7, 37.8)
ABCD	Hierarchical	0.269 (0.173, 0.366)	0.175 (0.075, 0.274)	54.2 (8.5, 119.2)
	Bifactor	0.266 (0.171, 0.361)	0.172 (0.076, 0.268)	54.4 (9.4, 118.1)

B. UK Biobank: complete-response analyses

Outcome	Items required	N/families	Integrated β (95% CI)	PGC β (95% CI)	Gain, % (95% CI)
Internalizing	All 28 items	9,895/8,055	0.234 (0.192, 0.275)	0.185 (0.141, 0.230)	26.2 (9.0, 46.2)
Neuroticism	Available items	36,292/17,656	0.217 (0.193, 0.240)	0.162 (0.140, 0.185)	33.4 (23.9, 43.6)
	All 12 items	29,500/16,952	0.226 (0.200, 0.252)	0.166 (0.139, 0.193)	36.5 (25.2, 48.7)
	All 28 items	9,895/8,055	0.202 (0.155, 0.249)	0.142 (0.097, 0.186)	42.9 (20.2, 69.9)
Anxiety	Available items	12,000/9,340	0.186 (0.143, 0.229)	0.155 (0.109, 0.200)	20.3 (0.8, 43.5)
	All 7 items	11,839/9,240	0.187 (0.143, 0.231)	0.152 (0.105, 0.199)	23.3 (2.8, 48.0)
	All 28 items	9,895/8,055	0.172 (0.125, 0.218)	0.144 (0.096, 0.192)	19.4 (−3.3, 47.4)

C. Fit of the measurement models without scores

Cohort	Factor model	CFI	TLI	RMSEA	SRMR
UK Biobank	Hierarchical	0.952	0.948	0.038	0.063
	Bifactor	0.945	0.936	0.043	0.064
ABCD	Hierarchical	0.839	0.826	0.064	0.058
	Bifactor	0.910	0.896	0.050	0.040

Notes. Coefficients are standardized within-family associations (Supplementary Note 3), and gains are $100(\beta_{\text{Integrated}}/\beta_{\text{PGC}} - 1)$. Within each cohort, both scores and both factor models use the same participants: in UK Biobank, 36,296 participants from 17,656 families (the available-item sample, whose internalizing row is the hierarchical row of Panel A), and in ABCD, 1,172 children from 593 families. In ABCD, the hierarchical-model difference between the scores was 0.095 (95% CI 0.033 to 0.156, $P = 0.0025$). Panel B requires complete answers to the items of each scale or to all 28 items, with completeness assessed per participant. CFI, TLI and RMSEA are scaled fit indices.

Table S7: Share of the population association retained within families

Outcome	Integrated, %	PGC, %	Difference, percentage points (95% CI below)
UK Biobank: depression	103.2	102.3	0.9 (−13.7, 14.3)
ABCD: internalizing	130.6	93.3	37.3 (4.2, 70.3)
UK Biobank: internalizing	92.1	97.4	−5.2 (−12.5, 2.1)
UK Biobank: neuroticism	94.0	96.2	−2.3 (−7.6, 3.1)
UK Biobank: anxiety	89.0	100.9	−12.0 (−28.2, 4.3)
ABCD: phenome profile	100.8	90.2	10.6 (−10.9, 32.2)

Retention is $100\,\beta_{\text{within}}/\beta_{\text{population}}$ in the same family sample, and values above 100% indicate a larger within-family association. Differences subtract PGC from integrated retention, with paired 95% confidence intervals. For UK Biobank depression, the per-score 95% intervals were 83.1–123.3% (integrated) and 78.1–126.5% (PGC). Depression uses the SNIPar direct and population coefficients with the 4,000-sample family bootstrap (Supplementary Note 2), latent outcomes use the jackknife (UK Biobank) or family influences (ABCD) of Supplementary Note 3, and the phenome row compares the corrected within-family-on-population profile slopes across 1,854 outcomes (Supplementary Note 5).

Supplementary Note 4: Adult health, life and disease outcomes

This note relates to Figure 1C and Table 2. It defines the 20 life and health outcomes and the 23 diagnoses and summarizes the results for the PGC score.

Outcomes. Table S8 defines the life and health outcomes and Table S9 the diagnoses. Every outcome was oriented so that higher values mean greater burden or disadvantage. A participant was a case if any listed hospital diagnosis, self-report or additional rule applied in any available record. Participants without case evidence but with a related condition listed as a control exclusion were set to missing, and all others were controls. The diagnoses therefore describe recorded lifetime history, with no common follow-up window or record-coverage cutoff. The outcome labelled Alzheimer's disease combines hospital diagnoses of Alzheimer's disease (G30), dementia in Alzheimer's disease (F00), vascular dementia (F01), dementia in other diseases (F02) and unspecified organic mental disorder (F09) with the self-reported item "dementia, Alzheimer's disease or cognitive impairment". It therefore describes dementia broadly rather than clinically diagnosed Alzheimer's disease alone.

Models and tests. Each outcome was analysed in the sibling validation sample with the two SNIPar models of Supplementary Note 2: a child-only model for the population association and a model adding the parental-score sum for the direct association. Models adjusted for sex, birth year and its square, and principal components 1–20, except that the sex-restricted cancers adjusted for birth year and principal components only. Scores and outcomes were standardized in each outcome's analysed sample, so coefficients are outcome SDs per score SD, and binary diagnoses were analysed as linear outcomes. Two-sided Wald tests used family-clustered robust (HC0) standard errors. The direct-to-population ratio has paired Fieller intervals, which are unbounded when the population association is not clearly different from zero. Benjamini–Hochberg correction covered all 23 diagnoses together, separately for each score and for the population and direct coefficients.

Results for the PGC score. Complete results for both scores are in the Supplementary Data. For every life and health outcome, the population coefficient of the PGC score was smaller than that of the integrated score, for example 0.107 versus 0.131 for health burden. The share retained within families was similar for the two scores: the direct-to-population ratio was 0.85 (PGC) versus 0.86 (integrated) for health burden and 0.88 versus 0.87 for any chronic pain. For education and occupational status the ratio was close to zero with both scores.

Table S8: UK Biobank life and health outcomes

Outcome	UKB fields	Definition and orientation
Health burden	2178	Poor = 1, fair = 2, good = 3, excellent = 4. Response mean, reversed.
Health dissatisfaction	4548	Satisfaction coding below. Response mean, reversed.
Tiredness	2080	Not at all = 1, several days = 2, more than half the days = 3, nearly every day = 4. Response mean.
Financial dissatisfaction	4581	Satisfaction coding below. Response mean, reversed.
Lower happiness	4526	Satisfaction coding below. Response mean, reversed.
Family dissatisfaction	4559	Satisfaction coding below. Response mean, reversed.
Friendship dissatisfaction	4570	Satisfaction coding below. Response mean, reversed.
Work dissatisfaction	4537	Satisfaction coding below. Response mean, reversed.
Any chronic pain	2956, 3404, 3414, 3571, 3741, 3773, 3799, 4067, 6159	1 if any of the eight chronic-pain indicators is positive, 0 if at least one is observed and none is positive, and missing otherwise.
Number of pain sites	Same eight pain fields, 6159	Number of positive chronic-pain indicators, ignoring missing ones. Missing if none is observed. Includes general (all-over) pain.
Neck/shoulder pain	3404, 6159	Site-specific chronic-pain rule (see notes).
Hip pain	3414, 6159	Site-specific chronic-pain rule (see notes).
Back pain	3571, 6159	Site-specific chronic-pain rule (see notes).
Abdominal pain	3741, 6159	Site-specific chronic-pain rule (see notes).
Knee pain	3773, 6159	Site-specific chronic-pain rule (see notes).
Headache	3799, 6159	Site-specific chronic-pain rule (see notes).

Outcome	UKB fields	Definition and orientation
Lower fluid intelligence	20016, 20128	Raw fluid-intelligence score at the earliest available assessment (the first instance when assessments tied). Participants who attempted at most three questions in every instance were excluded. Values were multiplied by −1 so that higher values indicate lower ability.
Lower education	6138	Highest qualification converted to years: university degree 20, NVQ/HND/HNC 19, other professional qualification 15, A or AS levels 13, O levels/GCSE/CSE 10 and none 7. Missing if declined or unrecognized. Values were multiplied by −1 so that higher values indicate lower educational attainment.
Lower occupational status	20277	SOC2000 occupation code converted to the sex-specific CAMSIS occupational-status scale (male scale for men, female scale for women). Values were multiplied by −1 so that higher values indicate lower occupational status.
Lower household income	738	Household income before tax: less than GBP 18,000 = 1, GBP 18,000–30,999 = 2, GBP 31,000–51,999 = 3, GBP 52,000–100,000 = 4 and greater than GBP 100,000 = 5. Declined answers were missing. Values were multiplied by −1 so that higher values indicate lower household income.

Health, satisfaction and tiredness outcomes are means of the recognized responses over all available assessments, and at least one valid response was required. Satisfaction was coded extremely unhappy = 1, very unhappy = 2, moderately unhappy = 3, moderately happy = 4, very happy = 5 and extremely happy = 6. "Reversed" means multiplied by −1, so that every outcome is oriented towards greater burden or disadvantage.

The eight chronic-pain indicators cover general (all-over), neck or shoulder, hip, back, abdominal, knee, head and facial pain. Each field was combined over its available assessments: any "yes" to pain lasting three months or longer counted as pain, and other observed answers counted as none. A site not reported among the pain types of the last month (field 6159) was coded as no pain at that site. A reported site with a missing answer to the three-month question remained missing.

Fluid intelligence used the earliest available assessment. CAMSIS is the published sex-specific occupational-status scale for SOC2000 codes.

Table S9: UK Biobank diagnoses: sources, case codes and control exclusions

Outcome	UKB fields	Case codes	Additional rules and exceptions
Psychiatric diagnoses			
Bipolar disorder	20002, 41270, 20126, 20544	ICD-10: F31, F31[0-9]. Self-report: 1291.	Probable bipolar I/II (20126) or professional mania, hypomania, bipolar or manic-depression diagnosis (20544) also defines a case.
Schizophrenia	20002, 41270, 20544	ICD-10: F20, F20[0-9]. Self-report: 1289.	Professional schizophrenia diagnosis (20544) also defines a case. F20.4 (post-schizophrenic depression) is included, but F32 depression alone is not a case code.
Psychosis spectrum (F20–F29)	41270, 41202, 41204, 40001, 40002	ICD-10 prefixes: F20–F29.	Any matching diagnosis, main/secondary diagnosis, underlying death cause or contributory death cause after code normalization. No duration or incident-onset restriction.

Outcome	UKB fields	Case codes	Additional rules and exceptions
Cardiometabolic disease			
Coronary artery disease	20002, 41270, 41272, 6150	ICD-10: I21, I21[0-49], I22, I22[0189], I23, I23[0-68], I241, I252. Self-report: 1075. OPCS-4: K40[1-489], K41[1-489], K45[1-5], K49[1-489], K502, K75[1-489].	Field 6150 reporting heart attack also defines a case.
Gout	20002, 41270, 20003	ICD-10: M10, M10[0-49]. Self-report: 1466.	Allopurinol use at any time (20003, code 1140875408) also defines a case. Control exclusions: ICD-10: M11, M11[01289].
Hypertension	20002, 41270	ICD-10: I10. Self-report: 1065, 1072.	I11–I13 and I15 do not qualify as case codes.
Type 2 diabetes	20002, 41270	ICD-10: E11, E11[0-9]. Self-report: 1223.	—
Venous thromboembolism	20002, 41270	ICD-10: I26, I26[09], I80, I80[0-389], I82, I82[0-389], O871, O882. Self-report: 1093, 1094, 1068.	—
Immune and inflammatory disease			
Asthma	20002, 41270, 6152	ICD-10: J45, J45[0189], J46. Self-report: 1111.	Doctor-diagnosed asthma (6152) also defines a case. Control exclusions: ICD-10: J40, J41, J41[018], J42, J43, J43[01289], J44, J44[0189], J47, J980. Self-report: 1113, 1114.
Atopic dermatitis	41270	ICD-10: L20, L20[089].	—
Inflammatory bowel disease	20002, 41270	ICD-10: K50, K50[0189], K51, K51[0-589]. Self-report: 1462, 1463, 1461.	—
Multiple sclerosis	20002, 41270	ICD-10: G35. Self-report: 1261.	—
Psoriasis	20002, 41270	ICD-10: L40, L40[0-589]. Self-report: 1453.	—
Vitiligo	20002, 41270	ICD-10: L80. Self-report: 1661.	—
Cancer			
Basal cell carcinoma	20001, 41270	ICD-10: C44, C44[0-9]. Self-report: 1061.	C44 also includes nonmelanoma skin malignancies beyond histologically confirmed basal-cell carcinoma.
Breast cancer (female)	20001, 41270	ICD-10: C50, C50[0-689]. Self-report: 1002.	Restricted to women.
Colorectal cancer	20001, 41270	ICD-10: C18, C18[0-9], C19, C20. Self-report: 1020, 1022, 1023.	—
Melanoma	20001, 41270	ICD-10: C43, C43[0-9]. Self-report: 1059.	—
Prostate cancer (male)	20001, 41270	ICD-10: C61. Self-report: 1044.	Restricted to men.
Testicular cancer (male)	20001, 41270	ICD-10: C62, C62[019]. Self-report: 1045.	Restricted to men.
Other diagnosed disease			
Alzheimer's disease	20002, 41270	ICD-10: G30 and G30[0189] (Alzheimer's disease), and F00, F01, F02 and F09 with all subcodes (dementia in Alzheimer's disease, vascular dementia, dementia in other diseases, and unspecified organic mental disorder). Self-report: 1263 (dementia, Alzheimer's disease or cognitive impairment).	The outcome covers dementia broadly, including vascular dementia and cognitive impairment, rather than clinically diagnosed Alzheimer's disease alone.

Outcome	UKB fields	Case codes	Additional rules and exceptions
Glaucoma	20002, 41270, 6148, 6119	ICD-10: H40, H401, H402, H408, H409. Self-report: 1277.	Reported glaucoma (6148) or any nonmissing affected-eye response (6119) also defines a case. Control exclusions: ICD-10: G453, H17, H17[0189], H18, H18[0-9], H192, H31, H31[0-489], H32, H32[08], H33, H33[0-5], H34, H34[01289], H35, H35[0-9], H36, H36[08], H44, H44[0-9], H45, H45[018], H40[3-6].
Osteoporosis	20002, 41270	ICD-10: M80, M80[0-589], M81, M81[0-689], M82, M82[018]. Self-report: 1309.	—

A participant was a case if any listed code or additional rule applied in any available record. Case evidence took precedence over control exclusions: participants without case evidence but with a listed control-exclusion code were set to missing, and all others, including those with no recorded evidence, were controls. Sex restrictions apply to the named outcomes.

ICD-10 codes came from hospital diagnoses (41270), and self-report codes from noncancer illness (20002) or cancer (20001), as listed. Coronary procedures used OPCS-4 (41272). Psychosis additionally used main diagnoses (41202), secondary diagnoses (41204), underlying death causes (40001) and contributory causes (40002). Other fields are named in the outcome-specific rules. Bracketed digits list the included fourth characters of a code. The outcomes describe recorded history, without a common follow-up period or record-coverage cutoff.

Supplementary Note 5: Childhood phenome and developmental trajectories

This note relates to Figure 2. It describes how the ABCD outcomes were coded and selected, the association models, the profile slope used to compare sets of associations (also used in Supplementary Notes 6 and 7), the comparison with the PGC score and the trajectory analyses.

Outcome inventory. Each ABCD item was analysed separately for each reporter (youth, parent or teacher). When a question was asked only after a positive screening question, children with a negative screen were coded as not having the outcome, and children with a positive screen but a missing follow-up answer were left missing. We excluded items that record details asked only of a subgroup, unordered categorical items, administrative fields, duplicates and items whose question routing was uncertain. The resulting inventory of 2,328 outcomes defines the Benjamini–Hochberg family of the phenome-wide analysis (Figure 2A).

Association models. Each outcome was regressed on the score using all assessments of each child (linear probability models for binary outcomes), with sex, baseline age, principal components 1–10, study site and assessment as covariates, and family-clustered (CR1) standard errors. An outcome was analysed if at least 200 children contributed, the outcome varied, and, for binary outcomes, at least 20 children were observed in each category. Outcomes with 20–49 children in a category are flagged as sparse. A child whose status changed between assessments could count in both categories. The sibling analyses used non-twin siblings and dizygotic twins from families with at least two children with usable observations, not necessarily from the same assessment. They comprised a population model fitted to these siblings, which required at least 100 families, and a within-family model with family fixed effects, which required at least 20 families in which the children differed in both outcome and score. Both sibling models required at least ten children in each category of a binary outcome. The family fixed effects were centred within the outcome-specific rows of each analysis. Figure 2A highlights outcomes whose within-family association was nominally significant and in the same direction as the population association. Benjamini–Hochberg-adjusted results are in the Supplementary Data.

Profile slopes. To summarize how two sets of association coefficients relate across many outcomes, we used a slope through the origin corrected for sampling error in both sets:

$$\widehat{k} = \frac{\sum_j w_j\{\widehat{x}_j\widehat{y}_j - \widehat{\mathrm{Cov}}(\widehat{x}_j, \widehat{y}_j)\}}{\sum_j w_j\{\widehat{x}_j^2 - \widehat{\mathrm{Var}}(\widehat{x}_j)\}}.$$

Here x_j and y_j are the two association coefficients for outcome j, not individual observations, and the sampling variances and covariances come from the fitted models. The slope compares:

- the PGC score (x) with the integrated score (y),
- sibling population coefficients (x) with within-family coefficients (y),
- girls (x) with boys (y),
- main score effects (x) with score-by-exposure interactions (y),
- White children (x) with another population group (y).

Each outcome received equal weight ($w_j = 1$) in the primary analyses. Weighting each battery equally was a sensitivity analysis. Every comparison used all outcomes that could be estimated in both models, whatever the strength of their score associations. The slope describes the relative size of two association profiles, not prediction accuracy for individuals.

Uncertainty came from 10,000 wild-bootstrap draws that multiplied each family's linearized influence on the slope by a random sign shared by all members of that family. This preserves the dependence between outcomes, models, sexes and scores. The sampling covariances and reference moments were held fixed. The denominator $D = \sum_j w_j\{\widehat{x}_j^2 - \widehat{\mathrm{Var}}(\widehat{x}_j)\}$ had to be positive. For comparisons of scores, of family models and of population groups, a denominator below 1.96 of its standard errors marked a weak reference, and intervals were still reported. Sex profiles required $D > 5\ \mathrm{SE}(D)$. For differences between the two scores' household slopes, we used paired delta-method intervals when both denominators exceeded 1.96 standard errors, and otherwise each score's Fieller confidence set and a conservative common-ratio Anderson–Rubin test. Fieller sets can be unbounded or consist of two disjoint intervals, and we report them as they are.

Comparison with the PGC score. Both scores were standardized in the same White reference sample and compared on the identical outcome rows of the 1,854 outcomes that could be estimated in the full population, sibling population and within-family models (Table S10). We also compared the incremental variance explained on the observed scale,

$$\Delta R^2_{sj} = b^2_{sj}\,\frac{\mathrm{SS}_{G_{sj}|C}}{\mathrm{TSS}_{Y_j}},$$

where b_{sj} is the coefficient of score s for outcome j, $\mathrm{SS}_{G_{sj}|C}$ the sum of squares of the score after adjustment for the covariates, and TSS_{Y_j} the total sum of squares of the outcome. In family models, the adjustment included the fixed effects, and the total sum of squares was computed before family centring. Relative gains compare the equal-weight means of the increments. Their intervals are Fieller intervals from paired family influences that propagate the coefficients and the score and outcome variances, and the tests of equal means use the 10,000 shared-family draws. Population and within-family gains are based on different samples and do not measure attenuation. The similarity of the two population profiles is also described by the Pearson correlation of the two scores' coefficients across all 2,328 outcomes. Results for all 2,328 outcomes and for battery-equal weights are in the Supplementary Data.

Developmental trajectories. Six outcomes were chosen to illustrate developmental patterns (Table S11). Score deciles were defined once, in all scored White children, before the outcomes were chosen. Logistic models included the continuous score, assessment and their interaction, together with sex, baseline age, principal components 1–10, study site and the deviation of age from the assessment mean, with family-clustered (CR1) standard errors. An outcome was modelled if it was measured at three or more assessments spanning at least two years of mean age. Each retained assessment needed at least ten children in each outcome category and 100 children in each extreme decile, and the retained assessments together needed at least 30 children in each category. Predicted probabilities for the lowest and highest deciles averaged the model predictions over the observed scores within each decile and a common covariate distribution: each child's earliest complete assessment, with the age deviation set to zero. Mean ages label the assessments. Intervals for the probabilities and risk ratios are pointwise and come from the joint covariance of the logit probabilities and log risk ratios.

Table S10: ABCD phenome: integrated and PGC scores compared on identical outcomes

Profile slope of integrated-score on PGC-score coefficients			
Model	Outcomes	Slope	95% CI
Population	1,854	1.11	(1.04, 1.18)
Sibling population	1,854	0.98	(0.83, 1.14)
Within family	1,854	1.10	(0.84, 1.37)

The analysis uses the 1,854 outcomes estimable in all three models, weighted equally. Profile slopes are corrected for sampling error, with 95% intervals from 10,000 shared-family multiplier draws. The sibling population and within-family models use the same observations. The correlation between the two scores' population coefficients was $r = 0.988$ across all 2,328 outcomes and $r = 0.990$ across the 1,854 outcomes. Correlated sampling error contributes to these values.

Table S11: Developmental trajectories: outcome definitions and estimates by assessment

Assessment	**Mean age**	**Lowest score decile, %**	**Highest score decile, %**	**Risk ratio (95% CI)**
Current major depression. Youth report of a current K-SADS diagnosis at each assessment.				
Baseline	10.0	0.3 (0.1, 0.7)	1.1 (0.6, 2.1)	3.79 (1.14, 12.64)
Year 2	12.1	0.1 (0.0, 0.3)	1.4 (0.8, 2.4)	12.50 (4.00, 39.01)
Year 4	14.2	1.1 (0.7, 1.7)	4.6 (3.2, 6.6)	4.40 (2.23, 8.68)
Year 6	16.1	1.0 (0.6, 1.6)	5.0 (3.6, 6.9)	5.07 (2.65, 9.71)
Current social anxiety. Youth report of a current K-SADS diagnosis at each assessment.				
Baseline	10.0	0.1 (0.0, 0.3)	0.8 (0.4, 1.6)	8.16 (2.28, 29.27)
Year 2	12.1	0.3 (0.2, 0.6)	1.4 (0.8, 2.3)	4.49 (1.65, 12.24)
Year 4	14.2	1.9 (1.4, 2.7)	8.3 (6.6, 10.4)	4.26 (2.69, 6.73)
Year 6	16.1	2.4 (1.8, 3.3)	7.3 (5.8, 9.3)	3.01 (1.89, 4.79)
Suicidal thoughts. Youth and parent K-SADS history of active suicidal thoughts, reported at each assessment. Both reports required, positive if either is positive.				
Baseline	10.0	3.9 (3.1, 4.9)	8.7 (7.2, 10.4)	2.21 (1.56, 3.13)
Year 2	12.1	4.0 (3.2, 4.9)	13.7 (11.7, 16.0)	3.41 (2.47, 4.71)
Year 4	14.2	6.7 (5.6, 8.0)	23.3 (20.5, 26.4)	3.48 (2.66, 4.54)
Nonsuicidal self-injury. Youth and parent K-SADS history of self-injury without suicidal intent. Both reports required, positive if either is positive.				
Baseline	10.0	5.0 (4.2, 6.1)	11.6 (9.9, 13.5)	2.30 (1.72, 3.09)
Year 2	12.1	3.8 (3.0, 4.8)	11.5 (9.6, 13.7)	3.00 (2.12, 4.26)
Year 4	14.2	6.9 (5.7, 8.3)	20.5 (17.9, 23.5)	2.98 (2.25, 3.94)
Nicotine or cannabis use. Youth report of lifetime nicotine and cannabis use by each assessment. Both items required, positive if either is positive.				
Baseline	10.0	0.3 (0.1, 0.7)	0.6 (0.3, 1.3)	2.34 (0.53, 10.30)
Year 1	11.0	0.8 (0.4, 1.4)	1.4 (0.8, 2.2)	1.78 (0.68, 4.69)
Year 2	12.1	1.5 (1.0, 2.2)	2.9 (2.1, 4.0)	1.95 (1.05, 3.61)
Year 3	13.0	2.3 (1.7, 3.1)	6.1 (4.8, 7.7)	2.62 (1.67, 4.11)
Year 4	14.2	6.0 (5.0, 7.3)	14.1 (12.1, 16.3)	2.34 (1.73, 3.15)
Year 5	15.1	10.7 (9.2, 12.4)	26.7 (24.0, 29.7)	2.50 (2.00, 3.13)
Year 6	16.1	21.8 (19.6, 24.2)	42.5 (39.5, 45.6)	1.95 (1.67, 2.27)
Year 7	17.0	25.1 (22.7, 27.7)	46.9 (43.7, 50.1)	1.87 (1.61, 2.16)
Psychiatric inpatient treatment. Parent report of inpatient psychiatric treatment since the previous assessment (follow-ups 1 to 7). A negative screening answer counts as no treatment, and a positive screen with a missing follow-up answer is missing.				
Year 1	11.0	0.1 (0.0, 0.2)	1.0 (0.5, 1.9)	15.71 (3.79, 65.13)
Year 2	12.1	0.2 (0.1, 0.5)	1.2 (0.6, 2.2)	6.60 (1.74, 25.07)
Year 3	13.0	0.3 (0.1, 0.6)	1.5 (0.8, 2.7)	5.73 (1.63, 20.12)
Year 4	14.2	0.4 (0.2, 0.8)	2.5 (1.6, 4.0)	5.83 (2.18, 15.58)

Assessment	Mean age	Lowest score decile, %	Highest score decile, %	Risk ratio (95% CI)
Year 5	15.1	0.5 (0.3, 0.8)	4.1 (2.8, 6.1)	9.20 (4.24, 19.95)
Year 6	16.1	0.4 (0.2, 0.7)	4.4 (3.1, 6.4)	11.80 (5.32, 26.17)
Year 7	17.0	0.8 (0.5, 1.3)	3.8 (2.7, 5.5)	4.82 (2.35, 9.88)

Model-standardized probabilities (with 95% confidence intervals) for children in the lowest and highest deciles of the integrated score, and the ratio of the two. Intervals are pointwise. The outcomes are illustrations chosen by the authors, and the probabilities are not incidence estimates.

Supplementary Note 6: Sex differences and household moderation

This note supports Figure 3 and Table 3. It describes the sex-specific analyses, the symptom-content ratings, the household exposures and the moderation models, and compares the integrated and PGC scores.

Sex-specific associations. Separate models for girls and boys used the same score and outcome scaling as the pooled analyses (the common score SD and the pooled outcome SD) and let the covariate effects differ by sex. Family influences were computed from the original families, so brothers and sisters remained linked. The psychiatric inventory comprised 726 outcomes from the Kiddie Schedule for Affective Disorders and Schizophrenia (K-SADS), the Prodromal Psychosis Scale, the General Behavior Inventory–Mania and the 7-Up Mania items. Benjamini–Hochberg correction covered the 726 boy-minus-girl tests. The direction of each contrast follows the coding of the outcome, so a positive contrast does not necessarily mean a larger absolute association in boys.

The boys-to-girls multiplier k is the profile slope of Supplementary Note 5, with girls' coefficients as x and boys' coefficients as y. It was computed for the 726 psychiatric outcomes and for all 2,139 outcomes that could be estimated in both sexes. With $D = \sum_j \{x_j^2 - \widehat{\mathrm{Var}}(x_j)\}$, the linearized influence of family f on the slope is

$$U_{kf} = D^{-1} \sum_j \{(y_j - 2\widehat{k}x_j)U_{xfj} + x_j U_{yfj}\},$$

and the resampling draws were $\widehat{k} + \sum_f \xi_f U_{kf}$ with shared-family random signs ξ_f. The same draws gave percentile intervals and centred two-sided tests for the difference between the two scores. In a sensitivity analysis, coefficients and influences were rescaled by the ratio of the pooled to the sex-specific outcome SD, keeping the score units and outcomes unchanged.

Symptom-content ratings. To test whether sex differences followed how strongly a symptom is associated with one gender, a language model (`gpt-5.6-sol`) rated the cultural typicality of each psychiatric item among US adolescents aged 12–15, from 1 (strongly typical of boys) through 5 (neutral or mixed) to 9 (strongly typical of girls). Each item was rated in a separate request containing only its wording, reporter and response options, without participant data or association results. Ten items with insufficient, nonspecific or erroneous descriptions were excluded, leaving 716. For each sex, the equally weighted outcome coefficients were regressed on the rating centred at 5. The slope is the change in coefficient per rating point, and the difference between the sexes used the paired family influences. The descriptive curves are local linear (LOESS) fits with tricube weights, a span of 0.70, no robustness iterations and pointwise bands with fixed weights. The ratings and their covariances were treated as fixed, so the uncertainty of the ratings is not reflected in the intervals. The ratings describe perceived cultural typicality, not the observed prevalence of symptoms or any individual's gender.

Household exposures. Exposures were measured at baseline, from reports that agreed within families.

- *Household resources* combined household income, in its ten ordered categories, with the highest education of the caregiver or partner in five levels (below high school, high school or GED, some college or associate degree, bachelor's degree, postgraduate degree). Both components were standardized across eligible families, averaged with equal weight and standardized again. Both components were required.
- *Material hardship* counted seven unmet needs in the previous year: inability to afford food, lost telephone service, incomplete rent or mortgage payment, eviction, utility shutoff, and forgone medical and dental care. All seven items were required.

- *Residential deprivation* was the scaled weighted sum of the Area Deprivation Index for the baseline address (`le_l_adi__addr1_wsum`, Release 7), not its percentile.

A consistent report from one relative could fill another relative's missing income, education or hardship item. A family was excluded from an exposure if its members disagreed on any required component or if data remained missing. The deprivation index required a value for every family member and agreement across the address measures. Each exposure was standardized across eligible families, with one observation per family, before any outcome restriction. The exposures are therefore family-level baseline measures, not verified shared addresses at each later assessment. Higher values indicate more resources, more hardship and more deprivation.

Moderation models. Following Keller (2014), each model included the score, the exposure E, E^2, the score-by-E interaction, the covariates, and the interactions of both the score and the exposure with every covariate. Latent internalizing models used the expanded White population with MLR estimation, full-information handling of missing items, family clustering and delta-method intervals based on the total latent SD. They included site indicators, centred on the site distribution of the original reference sample, with their interactions with the score and the moderator. The same site assignments were used for both scores and all moderators. Sex was coded 0 (girls) and 1 (boys). Phenome models also included the site-by-exposure and assessment-by-exposure interactions. In household profiles, the profile slope of Supplementary Note 5 relates the interaction coefficients to the main score coefficients, separately for the symptom and full-phenome inventories. For sex, the reported profile estimate is the boys-to-girls multiplier minus one. Holm correction (Holm, 1979) covered the three household exposures, both for the latent interactions tested against zero and for the symptom-profile tests, whereas outcome-by-exposure tests used Benjamini–Hochberg correction. Main score coefficients refer to an exposure of zero and to average covariate values (centred covariates in the latent models and, in the phenome models, the score coefficient averaged over the covariate values of the analysed observations through the score-by-covariate terms). A latent interaction is the change in the score coefficient, in latent SD units, per SD of the exposure, whereas profile moderation is a proportional change across all outcome coefficients.

Comparison with the PGC score. Both scores were compared on identical participants and outcomes. The sex profiles were nearly identical. Across the 726 psychiatric outcomes, k was 0.68 (95% CI 0.53 to 0.82) for the integrated score and 0.68 (0.52 to 0.84) for the PGC score (difference −0.01, −0.10 to 0.09, $P = 0.903$). Across all 2,139 outcomes, the corresponding values were 0.71 (0.57 to 0.84) and 0.70 (0.55 to 0.85), with a difference of 0.00 (−0.09 to 0.09, $P = 0.948$). With within-sex standardization, k was 0.79 for both scores across the psychiatric outcomes and 0.77 and 0.76 across all outcomes (all $P \geq 0.866$). For latent internalizing, the score-by-sex interaction (boys minus girls) was −0.099 for the integrated score and −0.084 for the PGC score (difference −0.015, −0.048 to 0.018, $P = 0.362$). The household interactions were −0.007 and −0.016 for resources, 0.007 and 0.011 for hardship, and 0.016 and −0.007 for deprivation. No difference between the scores was significant (all $P \geq 0.114$, Holm-adjusted $P \geq 0.458$), and none of the household interactions differed from zero (Holm-adjusted $P = 1.000$ for the integrated score and $P \geq 0.943$ for the PGC score). These comparisons used 5,720 children from 4,644 families for sex, and 5,311 (4,366 families), 5,609 (4,581) and 5,215 (4,308) children for resources, hardship and deprivation.

Supplementary Note 7: Portability across population groups

This note is about Figure 4. It defines the population groups, the models for the core outcomes and the phenome, the published expectations shown for context, and the observed-score check.

Population groups. In UK Biobank we used the genetic groups defined by Privé et al. (2022) from principal components, which that study labels by country: United Kingdom, Nigeria, China and India. We refer to them as EUR, AFR, EAS and SAS. These labels describe genetic similarity, not nationality or birthplace. The EUR reference was the family reserve (United Kingdom group), which was excluded from all discovery analyses (Supplementary Note 1). The AFR, EAS and SAS groups included all eligible participants with scores, with no sibling requirement. In ABCD we used the parent-reported groups of Supplementary Note 1: White (reference), Hispanic and African American.

The reduction in association is $1 - \beta_g/\beta_{\text{ref}}$, where β_g is the coefficient in group g and β_{ref} the coefficient of the same score in the reference group. Coefficients are log odds ratios for depression and latent-SD coefficients for internalizing, with the score standardized within each group. Ratios were not constrained, so a reduction can exceed 100% or fall below zero.

Depression in UK Biobank. Depression used the case and control rules of Supplementary Note 2 in all four groups, with the case-specific psychiatric exclusions and without the British or Caucasian restriction. Agreement between reported and genetic sex and the absence of a sex-chromosome aneuploidy flag were required. Models adjusted for sex, birth year and its square, and principal components 1–10. Each group's score moments and analysis rows were fixed, and uncertainty came from 40 whole-family deletion groups shared by both scores, with t_{39} intervals. A group was analysed if it had at least 40 participants, including 20 cases and 20 controls, and flagged as small if it had fewer than 200 participants or fewer than 100 cases or controls. The AFR and EAS depression analyses are small by this rule and are shown with open markers in Figure 4A.

Latent internalizing. In ABCD, the latent models used the 30 averaged YSR items with MLR estimation, full-information handling of missing items and no correlated residuals, with item and covariate transformations fixed and each group's own score moments. Children needed at least one answered item, complete sex, age and principal components, and genotype and family quality control. A group needed at least 150 children for the model to be fitted and 500 to be included in the main analysis. Intervals used family influences that propagated the fitted parameters, the empirical covariance of the predictors and the joint covariance of the scores and groups, including families that spanned groups.

In UK Biobank, the latent models used the hierarchical model of Supplementary Note 3 without correlated residuals, WLSMV estimation with the theta parameterization, all available pairs of items, and sex, birth year and its square (each standardized) and principal components 1–20 as covariates. Score moments were computed in each group before any item or covariate restriction. Sparse response categories were merged without reference to the scores: adjacent positive PHQ and GAD categories were combined until each category had at least ten observations in every group, keeping the zero category and the binary neuroticism items. The final coding combined categories 2 and 3 for depressed feelings (field 20510), appetite (20511), concentration (20508), restlessness (20516), irritability (20505) and foreboding (20512), and all positive categories for slowness or restlessness of movement (20518) and thoughts of suicide or self-harm (20513), whereas the other PHQ and GAD items kept all four categories. The coding and all transformations were the same for both scores and in every jackknife refit. All 40 whole-family deletions were refitted, including the empirical covariance of the predictors in the total latent variance, and the joint deletion covariance gave t_{39} coefficient intervals and Fieller ratio intervals. The same model and fit criteria applied to every group and both scores. In the non-EUR groups, most participants had not completed the online questionnaire, so for most of them the latent factor is informed only by the neuroticism items from the baseline assessment (Table S12C gives the share who answered all 28 items).

Intervals for reductions. Fieller intervals were obtained by inverting

$$(\widehat{\beta}_g - r\widehat{\beta}_{\text{ref}})^2 \le q^2\{V_g + r^2 V_{\text{ref}} - 2rC_{g,\text{ref}}\},$$

where V and C are the sampling variances and covariance, and the reduction interval is $[1 - U_r, 1 - L_r]$. The critical value q came from the t_{39} distribution for the UK Biobank analyses and for the observed-score analyses in both cohorts, and from the normal distribution for the ABCD latent analyses. Unbounded or disjoint Fieller sets are reported as such. Similar standardized coefficients do not show that a measure works in the same way in every group.

Phenome profiles. Phenome-wide portability compared each group's outcome coefficients with those of White children using the profile slope of Supplementary Note 5, weighting outcomes equally and including every outcome eligible in both groups (at least 200 children and, for binary outcomes, 20 in each category) whatever its significance. The models used the covariates of the phenome-wide analysis, and the coefficients of both groups were expressed in the score and outcome SDs of White children. Figure S1 shows all points underlying Figure 4B, with the same outcomes, slopes and intervals.

Published expectations. The expected reductions shown for context are $1 - \sqrt{R^2_{\text{relative}}}$, where R^2_{relative} is the variance explained in a group divided by that in the European reference. For UK Biobank we used the relative partial r^2 that Privé et al. (2022) estimated across 245 traits: 0.180 for AFR, 0.486 for EAS and 0.647 for SAS (expected coefficient retention 0.424, 0.697 and 0.804). For ABCD we used the corresponding estimates of Moore et al. (2025), as applied to ABCD by Wolfram et al. (2026). For the Hispanic group, a linear model of these estimates on genetic distance predicts a relative R^2 of 0.740 at the intermediate distance of admixed American populations (0.32, Ding et al., 2023). For African American children we interpolated between the relative R^2 of 0.283 that Moore et al. (2025) found in Nigerian-ancestry individuals and 1 for Europeans, $0.8 \times 0.283 + 0.2 \times 1 = 0.4264$, assuming 80% African ancestry. That proportion is an assumption of Wolfram et al. (2026), not a measurement in ABCD. The

expected coefficient retention is therefore 0.860 and 0.653. These expectations come from other traits, were not calibrated to depression and carry no uncertainty and we did not test observed against expected values. The relative R^2 compares variance explained within each group, which corresponds to within-group standardization of the score and outcome. It therefore matches the core-outcome reductions in Table S12A, but the phenome profile slopes use the score and outcome SDs of White children, so their comparison with the expected slopes assumes similar SDs in the compared groups.

Observed-score check. To see whether the latent measurement model drove the portability results, we repeated the core comparisons with observed symptom scores. In ABCD, the outcome was the mean of the raw anxious/depressed, withdrawn/depressed and somatic-complaints sums, averaged over the visits with complete answers. In UK Biobank, it was the mean of the raw PHQ, GAD and neuroticism sums among participants who answered all 28 items. A second version prorated each scale when at least 8 of 9 PHQ, 6 of 7 GAD and 10 of 12 neuroticism items were answered. Score units and covariates were unchanged, and each outcome definition had its own reference-group denominator. Observed coefficients use the empirical SD of the composite, whereas latent coefficients use the model-implied SD. Forty whole-family deletion refits re-estimated the outcome SDs, and pointwise 95% Fieller intervals used t_{39} and required every deletion to be estimable (Table S12C and Figure S2). The ABCD composites covered more than 99% of each group and broadly reproduced the latent pattern. The UK Biobank comparisons are more restricted, because most participants in every group lacked complete PHQ and GAD answers. In the SAS group, the change from the latent to the observed-score reduction came mainly from the reference-group denominator. Because the sample, the outcome and the inference all change between the latent and observed analyses, the difference cannot be attributed to the measurement model alone.

Table S12: Portability of the integrated and PGC scores across population groups

A. Reduction of the core association relative to the reference group, % (95% CI)				
Group	Participants	Integrated	PGC	Expected
UK Biobank: depression				
AFR	374	61.9 (8.7, 111.0)	60.1 (−11.5, 127.0)	57.6
EAS	281	71.2 (−6.1, 146.8)	31.5 (−69.2, 132.0)	30.3
SAS	581	15.7 (−19.6, 50.2)	−19.5 (−69.6, 26.8)	19.6
UK Biobank: latent internalizing				
AFR	4,015	57.0 (37.6, 75.9)	48.9 (16.2, 79.8)	57.6
EAS	1,803	30.0 (−4.7, 62.6)	28.8 (−22.2, 76.9)	30.3
SAS	6,507	17.3 (−1.1, 34.8)	13.1 (−9.8, 35.1)	19.6
ABCD: latent internalizing				
Hispanic	2,031	27.0 (−2.5, 53.8)	23.2 (−15.1, 57.0)	14.0
African American	1,491	67.2 (35.1, 98.1)	89.7 (49.3, 129.5)	34.7
B. ABCD phenome: profile slope relative to White children (95% CI)				
Group	Outcomes	Integrated	PGC	Expected
Hispanic	2,147	0.78 (0.61, 0.95)	0.79 (0.60, 0.98)	0.860
African American	2,102	0.65 (0.38, 0.91)	0.58 (0.29, 0.85)	0.653
C. Observed-score check: standardized coefficients (95% CI) and coverage				
Group	Participants (% of eligible)	Integrated	PGC	
ABCD: mean of the three YSR scale sums				
White	5,696 (99.6%)	0.182 (0.154, 0.210)	0.151 (0.123, 0.180)	
Hispanic	2,015 (99.2%)	0.123 (0.081, 0.166)	0.107 (0.055, 0.159)	
African American	1,483 (99.5%)	0.055 (−0.005, 0.115)	0.017 (−0.034, 0.068)	
UK Biobank: mean of the PHQ, GAD and neuroticism sums, all 28 items answered				
EUR	12,028 (26.3%)	0.213 (0.193, 0.233)	0.156 (0.137, 0.175)	
AFR	383 (9.5%)	0.078 (−0.022, 0.179)	0.055 (−0.027, 0.137)	
EAS	312 (17.3%)	0.088 (−0.034, 0.209)	0.101 (−0.040, 0.242)	
SAS	626 (9.6%)	0.244 (0.170, 0.319)	0.232 (0.154, 0.310)	
UK Biobank: as above, at least 80% of each scale's items answered				
EUR	14,057 (30.7%)	0.219 (0.201, 0.237)	0.163 (0.147, 0.178)	
AFR	490 (12.2%)	0.096 (0.011, 0.182)	0.086 (−0.004, 0.177)	

Continued on next page

Portability (continued)

EAS	402 (22.3%)	0.110 (−0.004, 0.224)	0.114 (−0.014, 0.242)
SAS	785 (12.1%)	0.221 (0.143, 0.299)	0.186 (0.104, 0.267)

Panel A gives $100(1-\beta_g/\beta_{\text{ref}})$, where each score is standardized within the group and compared with its own reference-group coefficient (UK Biobank EUR, ABCD White children). Depression uses log odds ratios and internalizing standardized latent coefficients. Intervals are pointwise 95% Fieller intervals. The UK Biobank AFR and EAS depression samples include 49 and 54 cases. Panel B uses the score and outcome SDs of White children for both groups, with profile slopes corrected for sampling error and 95% multiplier intervals. "Expected" values are derived from published relative variance explained for other traits (Supplementary Note 7). They carry no uncertainty and are shown for context. Panel C gives observed-score coefficients per group-specific score SD with 95% jackknife intervals. Coverage is relative to each group's sample before any item-completeness requirement. Figure S2 shows the corresponding reductions.

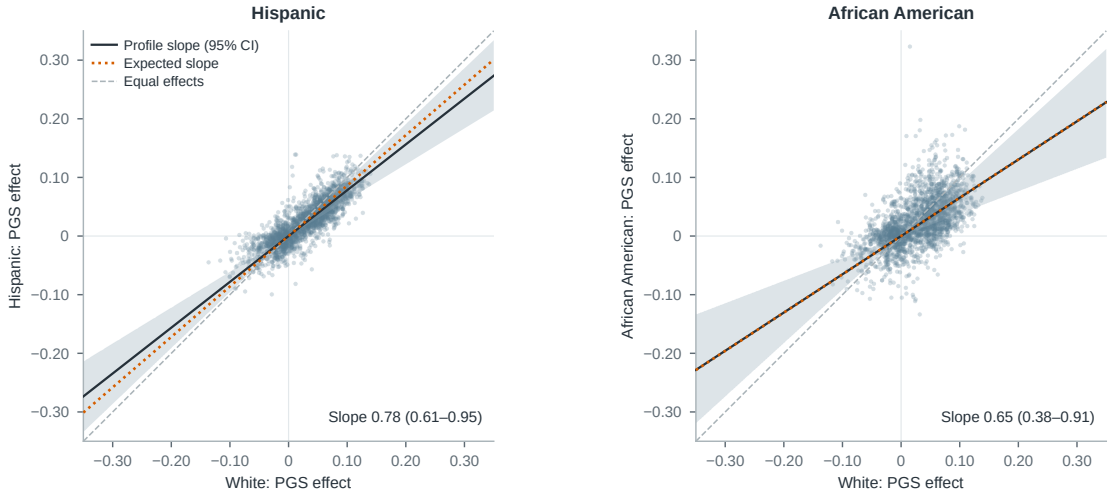


Figure S1: **Full-range phenome portability profiles.** Associations of the integrated score with every eligible ABCD outcome in Hispanic (left) and African American children (right), plotted against the associations in White children. All coefficients are in White-group score and outcome SD units. Figure 4B shows the same outcomes, slopes and intervals on a narrower range. Solid lines show the profile slope with its 95% confidence band, dashed lines equal effects and dotted lines the expected slope (Supplementary Note 7).

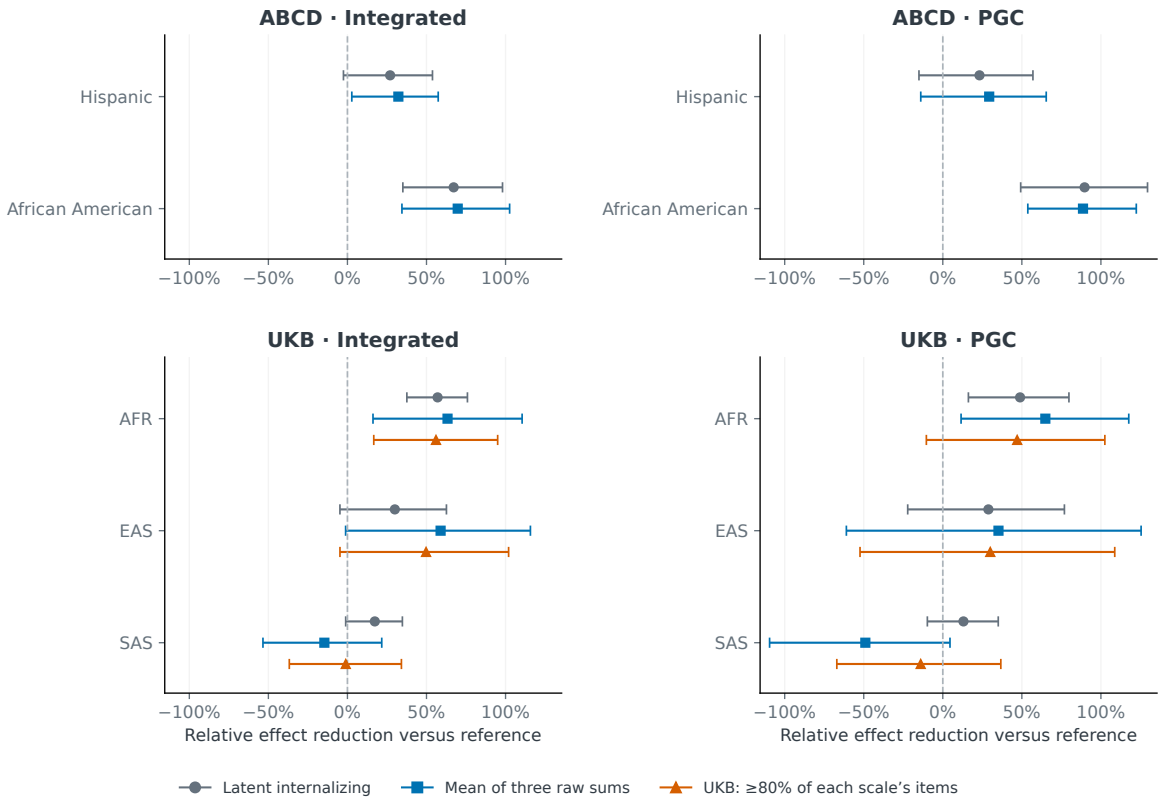


Figure S2: **Observed-score check of portability.** Reduction in association relative to the reference group, $1 - \beta_g/\beta_{\text{ref}}$, for the integrated and PGC scores (reference: White children in ABCD, EUR in UK Biobank). Grey circles show latent internalizing, blue squares the mean of the three observed scale sums, and orange triangles the UK Biobank composite that allows a few unanswered items. Bars are pointwise 95% Fieller intervals. Negative values indicate a larger association than in the reference group. Table S12C gives the coefficients and sample coverage.