The American Journal of Human Genetics, Volume *108*

Supplemental information

Leveraging both individual-level genetic data and GWAS summary statistics increases polygenic prediction

Clara Albiñana, Jakob Grove, John J. McGrath, Esben Agerbo, Naomi R. Wray, Cynthia M. Bulik, Merete Nordentoft, David M. Hougaard, Thomas Werge, Anders D. Børglum, Preben Bo Mortensen, Florian Privé, and Bjarni J. Vilhjálmsson

Supplementary Tables and Figures

Traits	Individual-level dataset	Population prevalence	Individual-level sample size	GWAS sample size	Ratio int:ext	Intercept bivariate LDSC internal-external (SE)	r_g internal-external (SE)
BD	iPSYCH 2012	0.01	4091	48609	1:12	3e-04 (0.0067)	1.4051 (0.1994)
SCZ		0.01	6389	48307	1:8	0.0027 (0.0061)	0.56 (0.0662)
AN		0.01	7713	35274	1:5	0 (0.0062)	0.8147 (0.0945)
ASD		0.01	23512	10610	2:1	0.0016 (0.0059)	0.6222 (0.1004)
ADHD		0.05	25658	12214	2:1	0.0019 (0.0058)	1.3366 (0.1399)
MDD iPSYCH		0.08	30222	646483	1:21	6e-04 (0.0068)	0.7729 (0.0483)
BD	iPSYCH 2015	0.01	8436	48609	1:6	6e-04 (0.0063)	0.7855 (0.0804)
SCZ		0.01	15421	48307	1:3	0.0085 (0.0069)	0.6175 (0.0677)
ASD		0.01	39068	10610	4:1	0.0039 (0.0051)	0.6241 (0.0671)
ADHD		0.05	43405	12214	4:1	0.0019 (0.0067)	1.3137 (0.1216)
MDD iPSYCH		0.08	49234	646483	1:13	0.0016 (0.0074)	0.8115 (0.0477)
CAD	UK Biobank	0.03	35457	162973	1:5	0.0183 (0.0055)	0.8644 (0.0672)
BC		0.07	35707	227688	1:6	0.0301 (0.0089)	0.9378 (0.085)
T2D		0.05	57086	88825	1:2	0.019 (0.0081)	0.9567 (0.0595)
MDD UKB		0.15	83900	123796	1:2	0.0093 (0.0071)	0.8156 (0.0632)
BMI		-	269106	339224	1:1	0.0053	0.9536

						(0.0109)	(0.0347)
Height		-	269407	253288	1:1	0.099 (0.0265)	0.9389 (0.0417)

Table S1: Summary of real datasets*. Effective sample sizes of the 12 analyzed complex traits from the individual-level datasets, along with the effective sample sizes of the corresponding external GWAS publication. The table reflects sizes of European, unrelated samples (see Methods).*

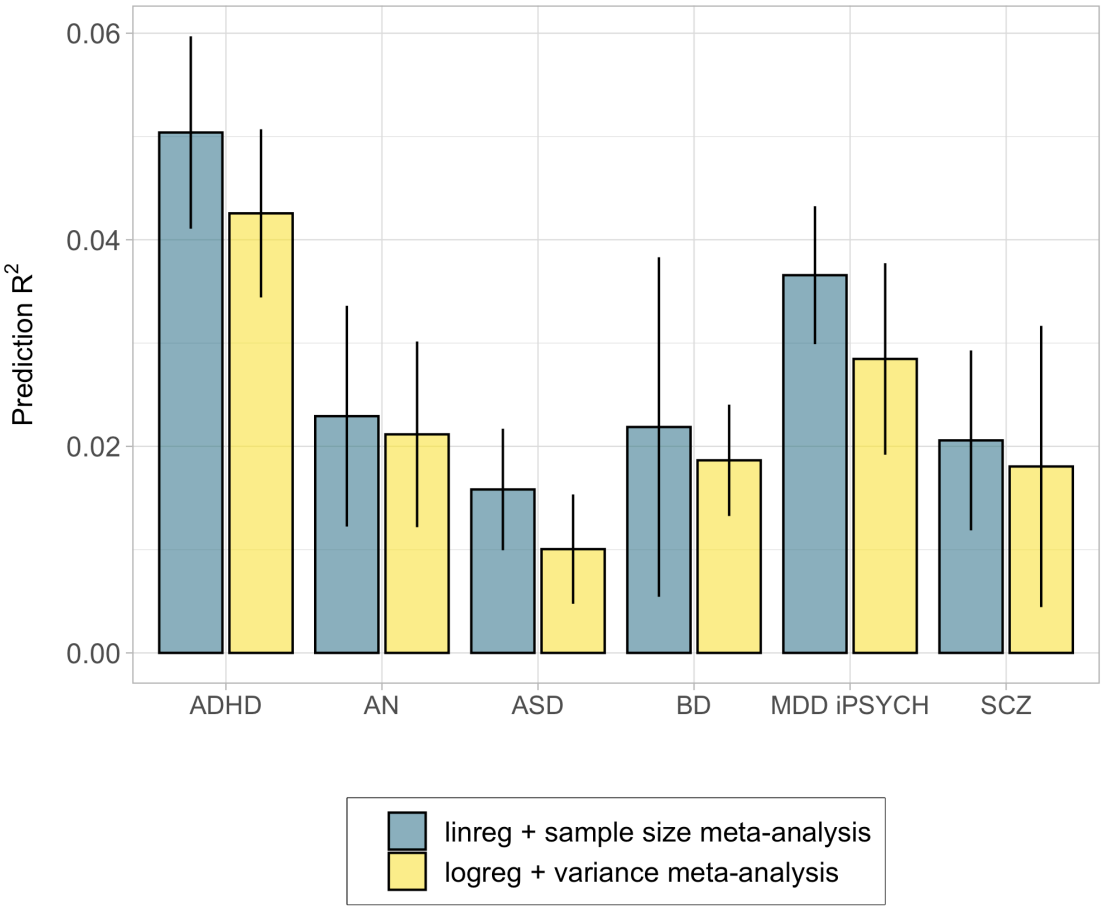

Figure S1 Prediction accuracy of the data-combining approaches based on different GWAS and meta-analysis methods in 6 major psychiatric disorders from iPSYCH 2012. *Each panel displays the mean and 95% CI of the PRS R^2 (y-axis) for each Meta-GWAS PRS (x-axis). Some variation is expected due to randomness in the cross-validation subsets. Mean and 95% CI of the R^2 were obtained from 10k non-parametric bootstrap samples of the 5 cross-validation subsets.*

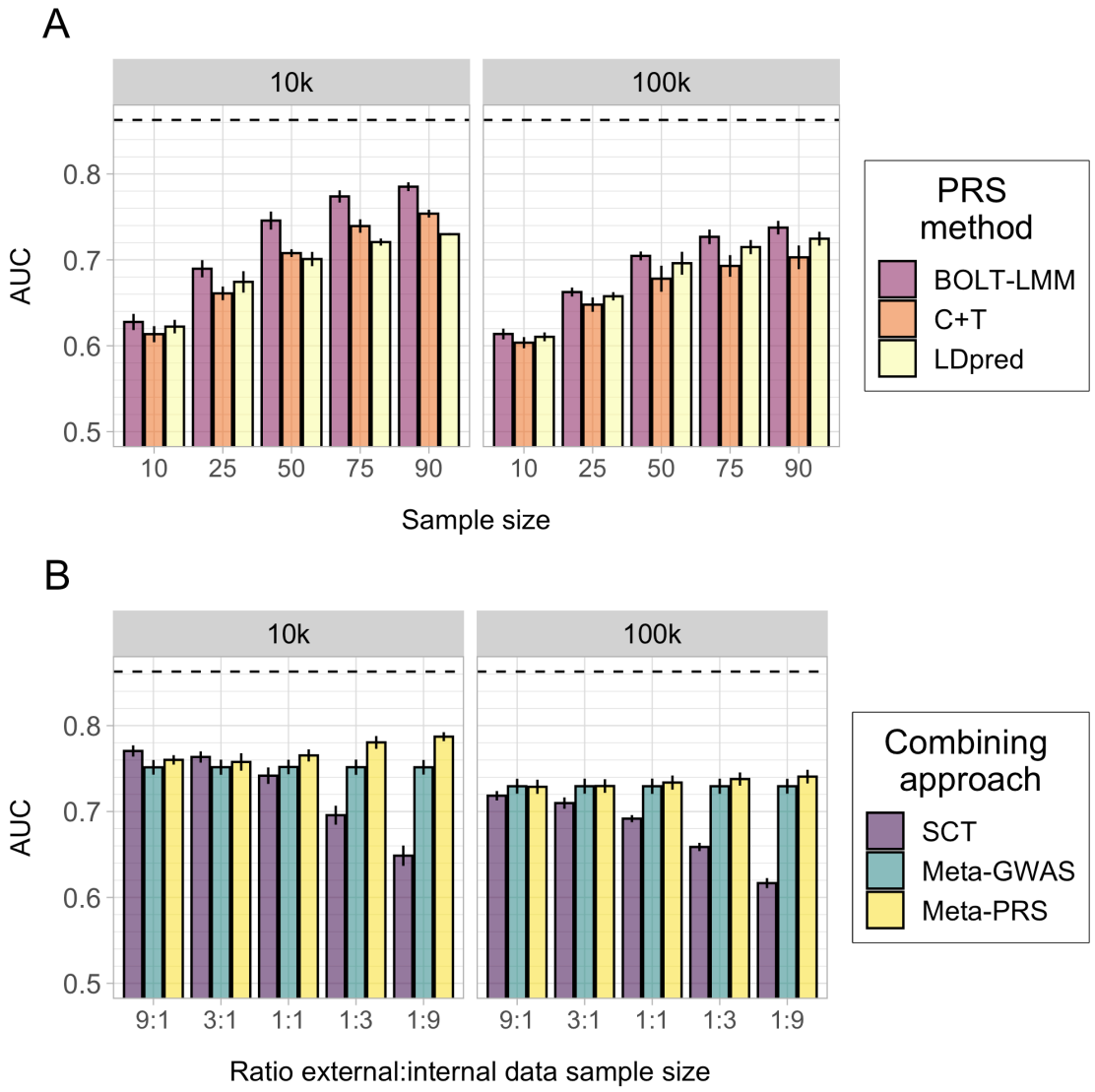

Figure S2 Prediction accuracy of the PRSs in the simulation study in terms of AUC. *Each panel displays the mean and 95% CI of the PRS AUC (y-axis) for each data-combining approach. The traits were simulated from a liability threshold model with 10,000 (10k) and 100,000 (100k) causal SNPs and heritability h^2 of 0.5, and case-control status was inferred from a disease prevalence of 0.2. Mean and 95% CI of AUC were obtained from 10k non-parametric bootstrap samples of 5 independent replicates. The black line represents the AUC_{max} (0.852) for these simulations. A) Effect of training sample size in the PRSs prediction accuracy. The x-axis indicates the percentage of individuals from the total training set (N = 303,728) used as individual-level data for BOLT-LMM or GWAS summary statistics for C+T and LDpred. B) Effect of the ratio between internal and external data in the combining approaches. The x-axis indicates the relative amount of external vs. internal data, e.g. 3:1 indicates a scenario where the external data was 75% and the internal data was 25% of the total sample.*

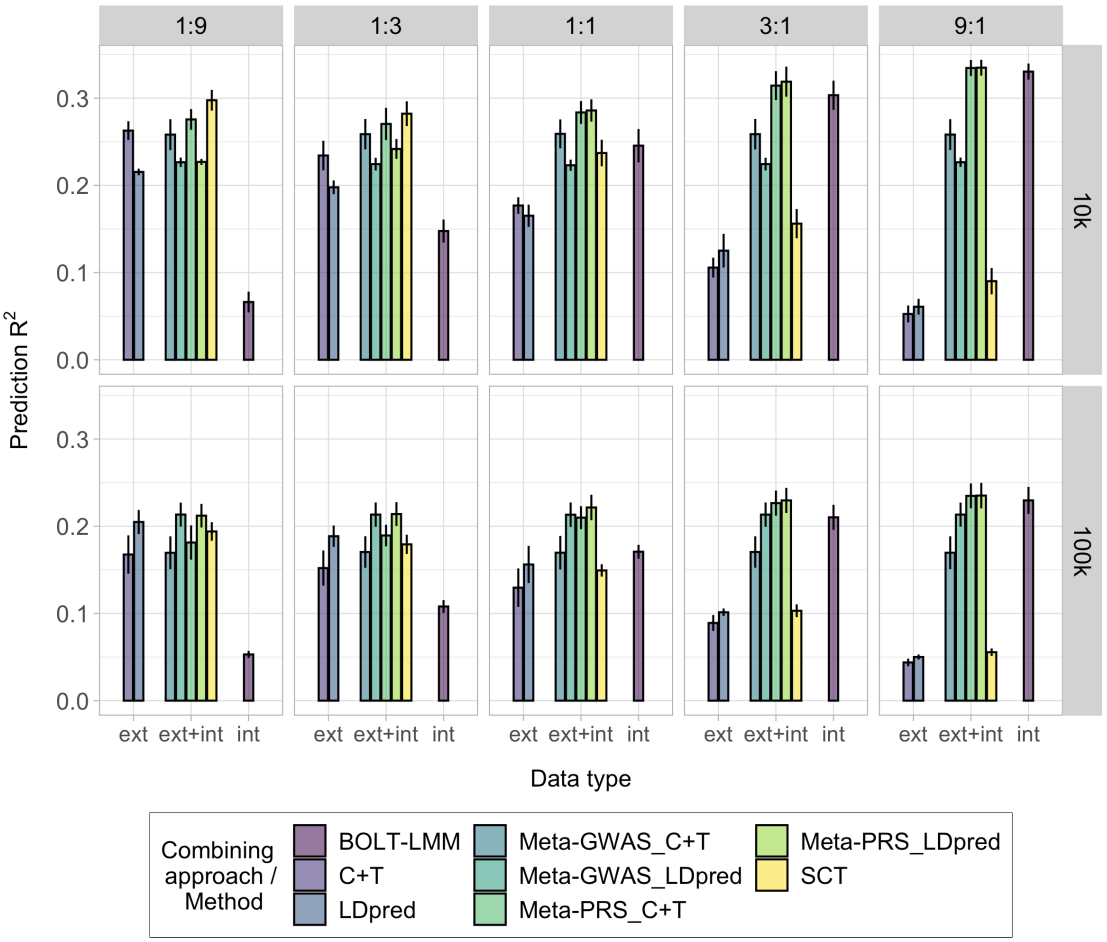

Figure S3 Prediction accuracy of the data-combining approaches using different GWAS summary statistics-based PRS methods in the simulated data. *Each panel displays the mean and 95% CI of the PRS R^2 (y-axis) for each data-combining approach and PRS method, of PRSs trained on individual-level data (int), GWAS summary statistics (ext) or both (ext+int) (x-axis). In the case of Meta-GWAS, C+T and LDpred were used on the meta-analyzed summary statistics and in Meta-PRS, C+T and LDpred were used to compute the external PRS. The traits were simulated from a liability threshold model with 10,000 (10k) and 100,000 (100k) causal SNPs and heritability h^2 of 0.5, and case-control status was inferred from a disease prevalence of 0.2. Mean and 95% CI of prediction R^2 were obtained from 10k non-parametric bootstrap samples of 5 independent replicates.*

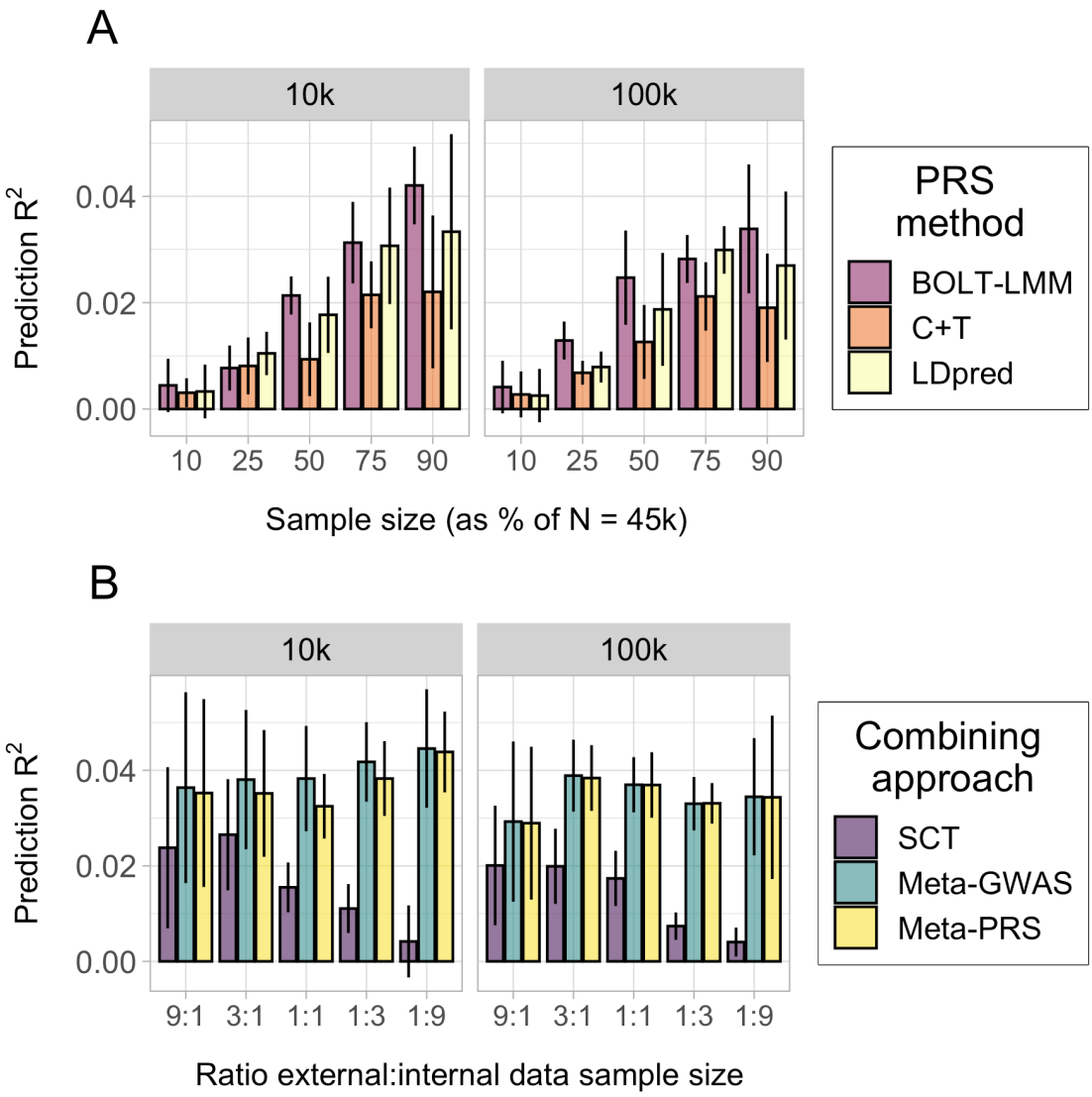

Figure S4 Prediction accuracy of the data-combining approaches in the small (50k individuals) simulation study. *Each panel displays the mean and 95% CI of the PRS prediction R^2 (y-axis) for each data combining approach. The traits were from a liability threshold model with 10,000 (10k) and 100,000 (100k) causal SNPs and heritability h^2 of 0.5, and case-control status was inferred from a disease prevalence of 0.2. Mean and 95% CI of prediction R^2 were obtained from 10k non-parametric bootstrap samples of 5 independent replicates. A) Effect of training sample size in the PRSs prediction accuracy. The x-axis indicates the percentage of individuals from the total training set (N = 45,000) used as individual-level data for BOLT-LMM or GWAS summary statistics for C+T and LDpred. B) Effect of the ratio between internal and external data in the combining approaches.The x-axis indicates the relative amount of external vs. internal data, e.g. 3:1 indicates a scenario where the external data was 25% and the internal data was 75% of the total sample.*

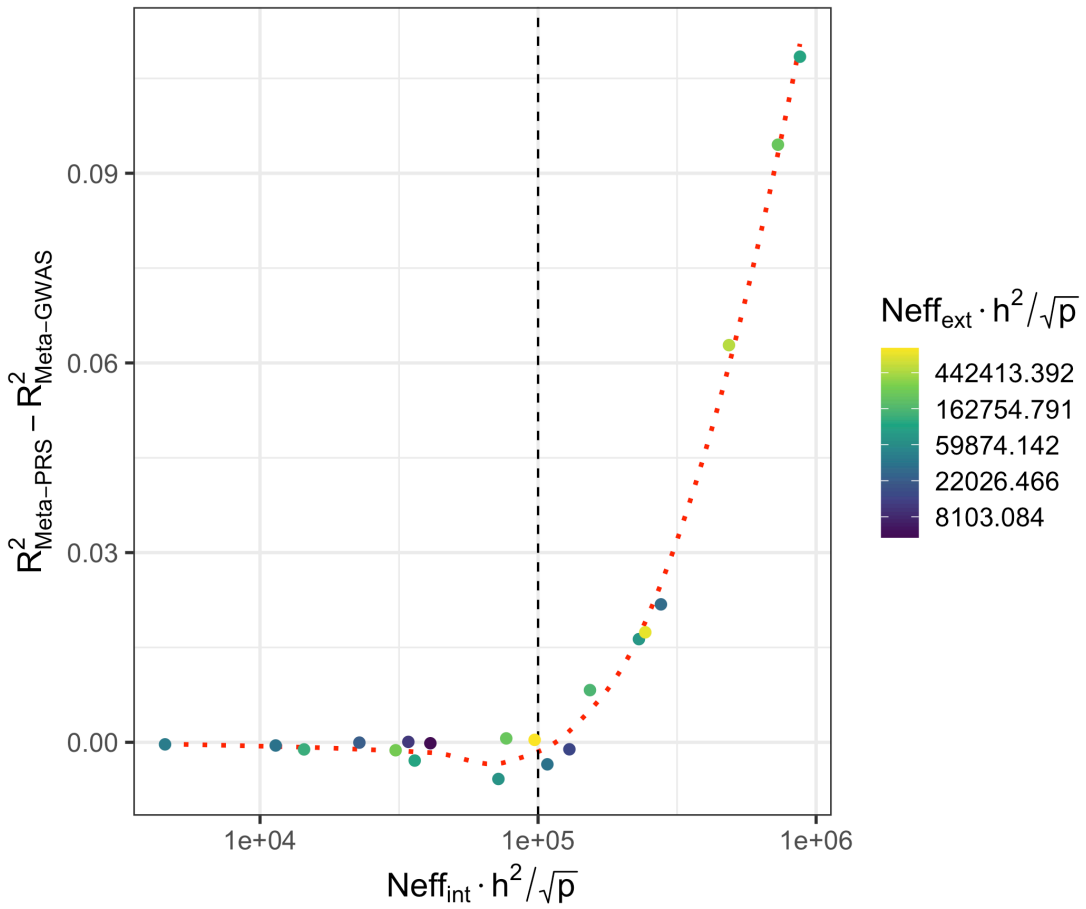

Figure S5 Difference in prediction accuracy between Meta-PRS and Meta-GWAS in the simulation study (both large and small). *The plot displays the difference in mean prediction R^2 between the PRS using Meta-PRS and Meta-GWAS (y-axis) as a function of the internal effective sample size ($Neff_{int}$), the SNP-heritability (h^2) and the proportion of causal variants (p) (x-axis). $Neff_{ext}$ indicates the effective sample size of the external data. The SNP-heritability was 0.5 for all simulations, $Neff_{int}$ and $Neff_{ext}$ can be found in Table 2 and p had value 0.1 (100k causal variants) or 0.01 (10k causal variants). The line at 100k indicates the threshold value of $Neff_{int} \cdot h^2/\sqrt{p}$ where the difference is significant.*

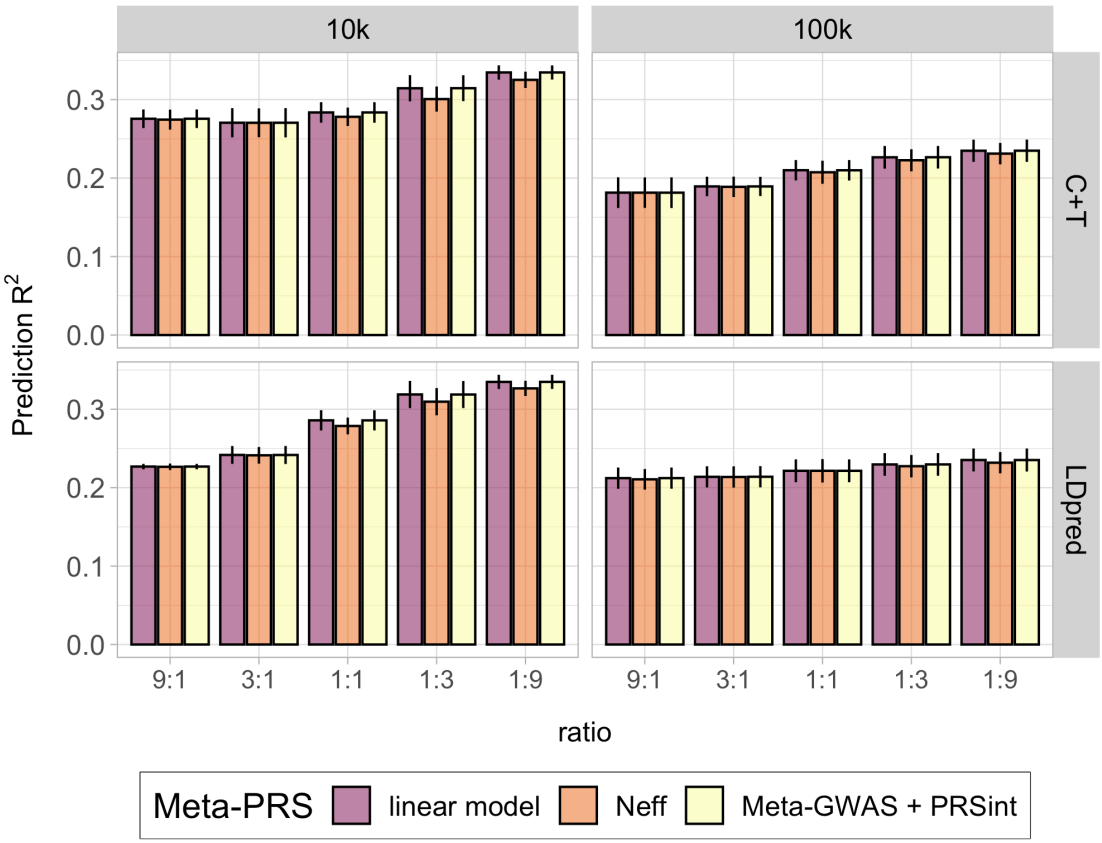

Figure S6 Prediction accuracy of Meta-PRS types in the simulation analysis. *Each panel displays the mean and 95% CI of the PRS prediction R^2 (y-axis) for Meta-PRS in each simulated scenario using either C+T or LDpred to generate the external PRS. The weights were obtained using linear regression (linear model), the square root of the training effective sample size (Neff) or a linear regression between the Meta-GWAS PRS and the internal BOLT-LMM PRS (Meta-GWAS + PRSint). In the cases with a linear regression model, the weights are trained in an independet validation dataset (see Table 1). The traits were simulated from a liability threshold model with 10,000 (10k) and 100,000 (100k) causal SNPs and heritability h^2 of 0.5, and case-control status was inferred from a disease prevalence of 0.2. Mean and 95% CI of prediction R^2 were obtained from 10k non-parametric bootstrap samples of 5 independent replicates.*

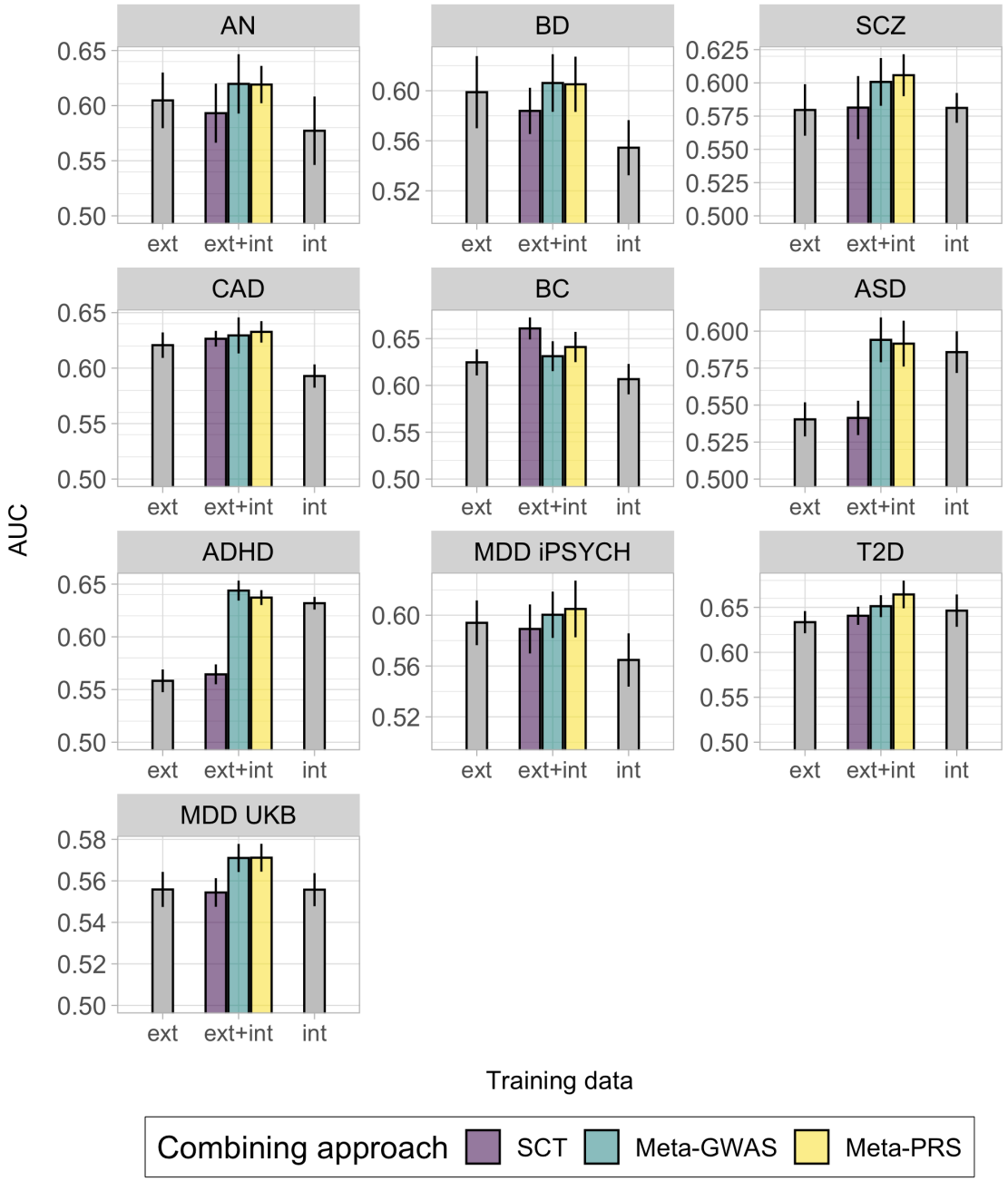

Figure S7 Prediction accuracy of the data-combining approaches in 12 complex traits from iPSYCH 2015 and UK Biobank. *Each panel displays the mean and 95% CI of the PRS AUC (y-axis) for each data-combining approach, of PRS trained on individual-level data (int), GWAS summary statistics (ext) or both (ext+int) (x-axis). The methods noted as int and ext were fitted using BOLT-LMM with individual-level data and LDpred or C+T with GWAS summary statistics, respectively. For simplification, only the ext PRS with larger mean prediction R^2 is shown. Mean and 95% CI of the AUC were obtained from 10k non-parametric bootstrap samples of the 5 cross-validation subsets.*

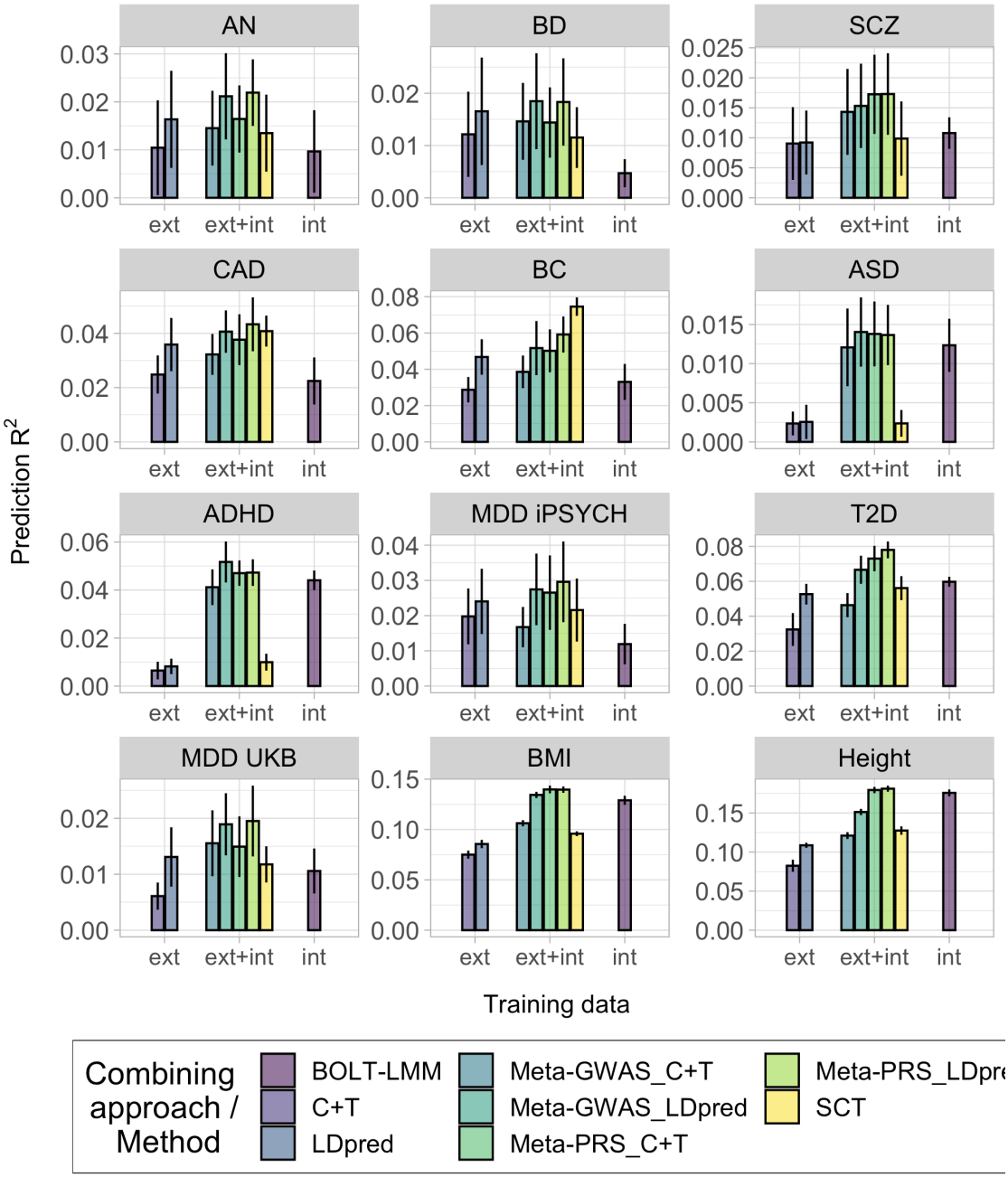

Figure S8 Prediction accuracy of the data-combining approaches using different GWAS summary statistics-based PRS method in 12 complex traits from iPSYCH 2015 and UK Biobank. *Each panel displays the mean and 95% CI of the PRS R^2 (y-axis) for each data-combining approach and PRS method, of PRSs trained on individual-level data (int), GWAS summary statistics (ext) or both (ext+int) (x-axis). In the case of Meta-GWAS, C+T and LDpred were used on the meta-analyzed summary statistics and in Meta-PRS, C+T and LDpred were used to compute the external PRS. The prediction R^2 was transformed to the liability-scale using a population prevalence of 0.01 (ASD), 0.05 (ADHD), 0.15 (MDD UKB), 0.05 (T2D), 0.01 (AN), 0.03 (CAD), 0.01 (SCZ), 0.07 (BC), 0.01 (BD) and 0.08 (MDD iPSYCH). Mean and 95% CI of*

the AUC were obtained from 10k non-parametric bootstrap samples of the 5 cross-validation subsets.

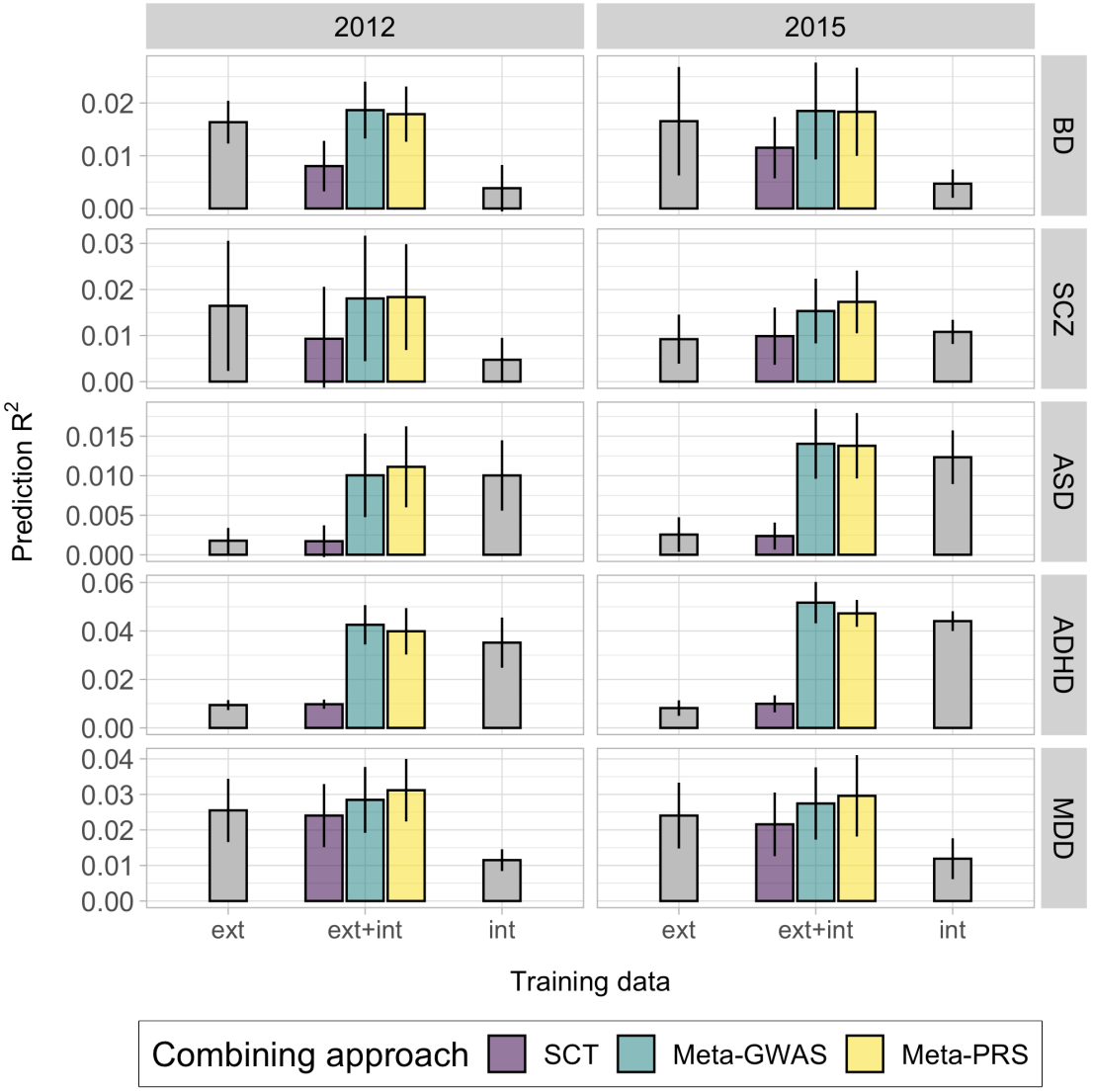

Figure S9 Prediction accuracy of the data-combining approaches in iPSYCH 2012 and iPSYCH 2015 *Each panel displays the mean and 95% CI of the PRS R^2 (y-axis) for each data-combining approach and PRS method, of PRSs trained on individual-level data (int), GWAS summary statistics (ext) or both (ext+int) (x-axis). For simplification, only the ext PRS with larger mean prediction R^2 is shown. The prediction R^2 was transformed to the liability-scale using a population prevalence of 0.01 (ASD), 0.05 (ADHD), 0.01 (AN), 0.01 (SCZ), 0.01 (BD) and 0.08 (MDD). Mean and 95% CI of the AUC were obtained from 10k non-parametric bootstrap samples of the 5 cross-validation subsets.*

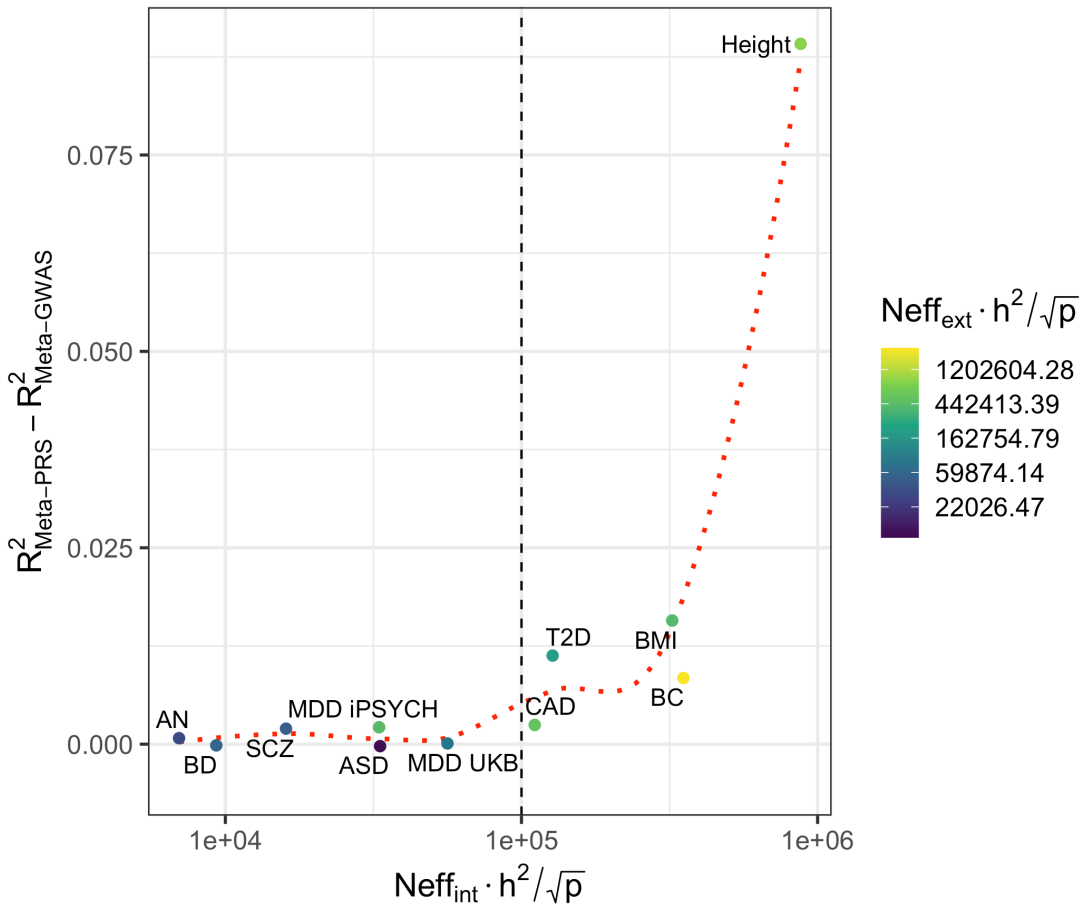

Figure S10 Difference in prediction accuracy between Meta-PRS and Meta-GWAS in 12 complex traits from iPSYCH 2015 and UK Biobank. *The plot displays the difference in mean prediction R^2 between the PRS using Meta-PRS and Meta-GWAS (y-axis) as a function of the internal effective sample size ($Neff_{int}$), the SNP-heritability (h^2) and the proportion of causal variants (p) (x-axis). $Neff_{ext}$ indicates the effective sample size of the external data. The line at 100k (selected from the simulations) indicates the threshold value of $Neff_{int} \cdot h^2/\sqrt{p}$ where the difference is significant. The h^2 estimates were taken from the GWAS publications in Table 2 for each trait. The p estimates were taken from Table S1 in Privé et al. 2021 or set to 0.05 for the psychiatric disorders.*

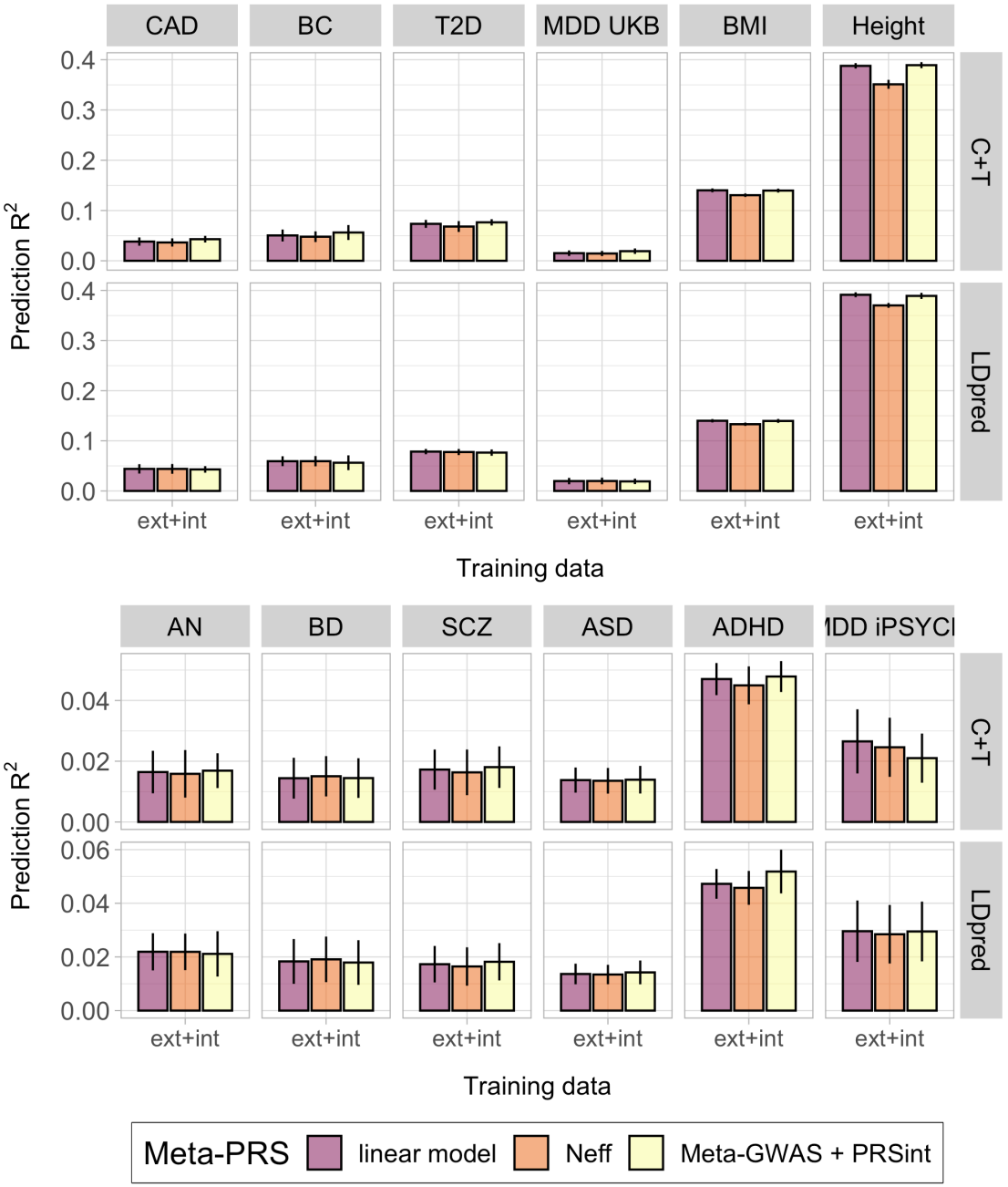

Figure S11 Prediction accuracy of Meta-PRS types in 12 complex traits from iPSYCH 2015 and UK Biobank. *Each panel displays the mean and 95% CI of the PRS prediction R^2 (y-axis) for Meta-PRS in each trait using either C+T or LDpred to generate the external PRS. The weights were obtained using linear regression (linear model), the square root of the training effective sample size (Neff) or a linear regression between the Meta-GWAS PRS and the internal BOLT-LMM PRS (Meta-GWAS + PRSint). In the cases with a linear regression model, the weights are trained in an independet validation dataset (see Table 1). The prediction R^2 was transformed to the liability-scale using a population prevalence of 0.01 (ASD), 0.05 (ADHD), 0.15 (MDD UKB), 0.05 (T2D), 0.01 (AN), 0.03 (CAD), 0.01 (SCZ), 0.07 (BC), 0.01 (BD) and*

0.08 (MDD iPSYCH). Mean and 95% CI of the AUC were obtained from 10k non-parametric bootstrap samples of the 5 cross-validation subsets.