The American Journal of Human Genetics, Volume *109*

Supplemental information

The construction of cross-population polygenic risk scores using transfer learning

Zhangchen Zhao, Lars G. Fritsche, Jennifer A. Smith, Bhramar Mukherjee, and Seunggeun Lee

Supplemental Information

1. Supplemental Figures

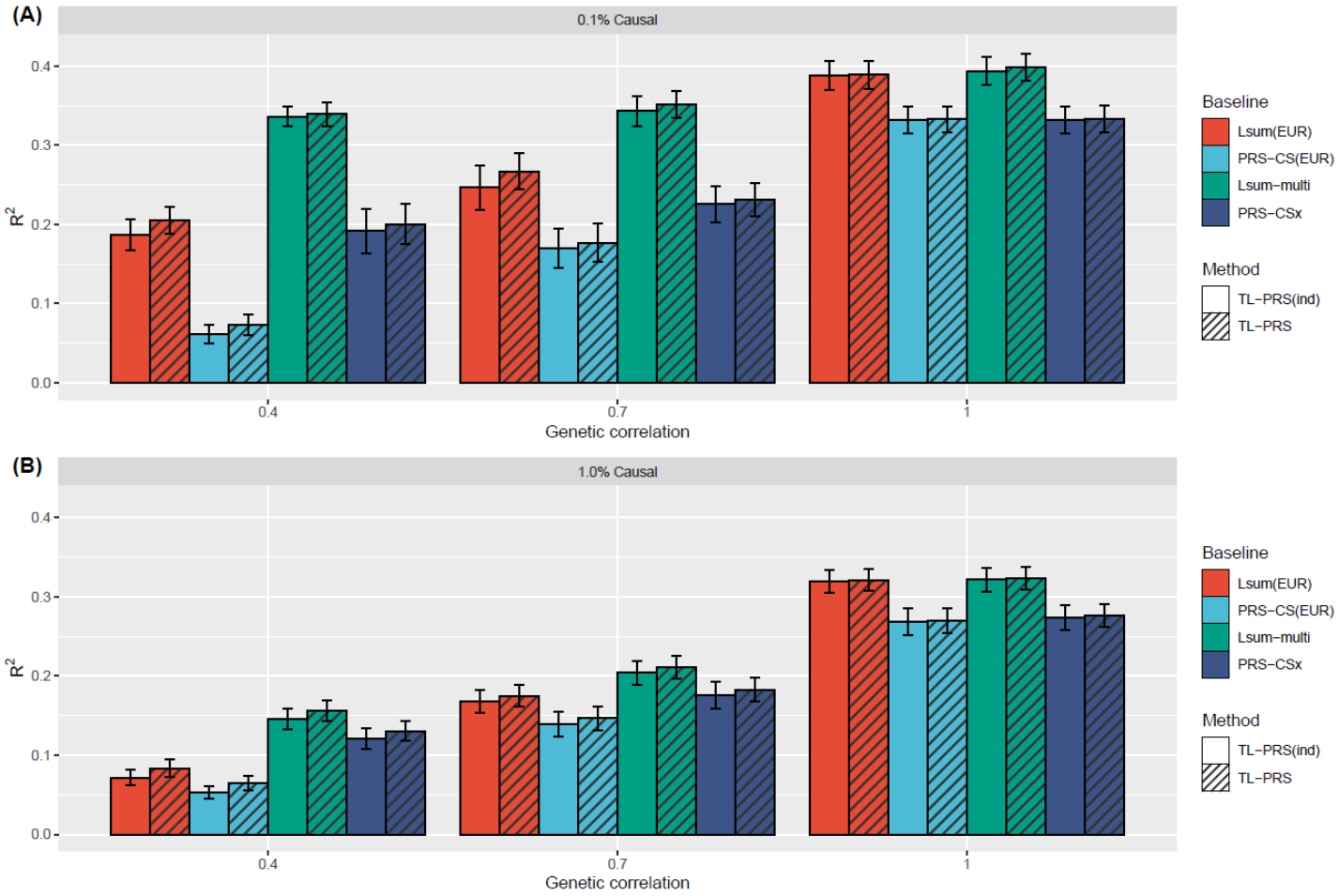
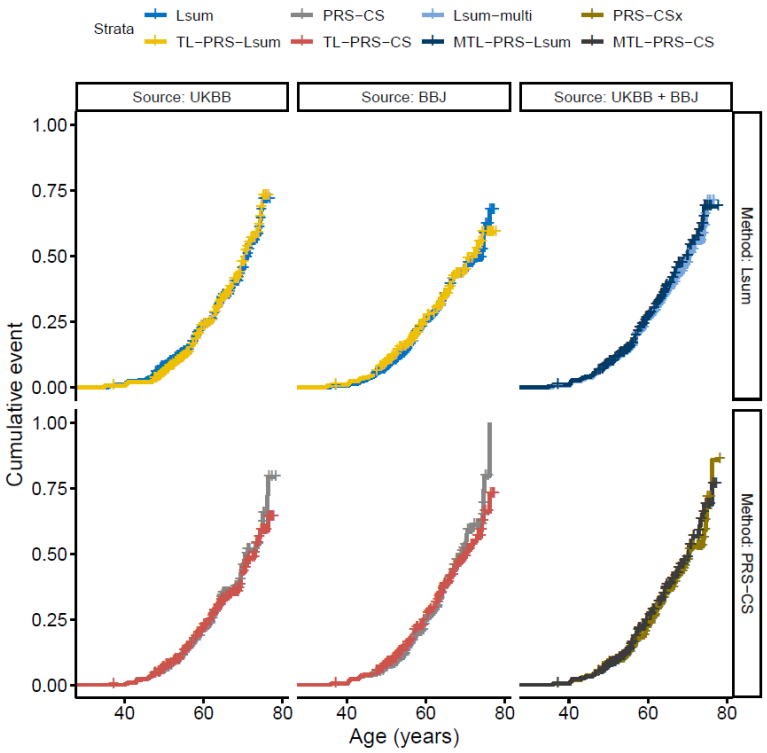
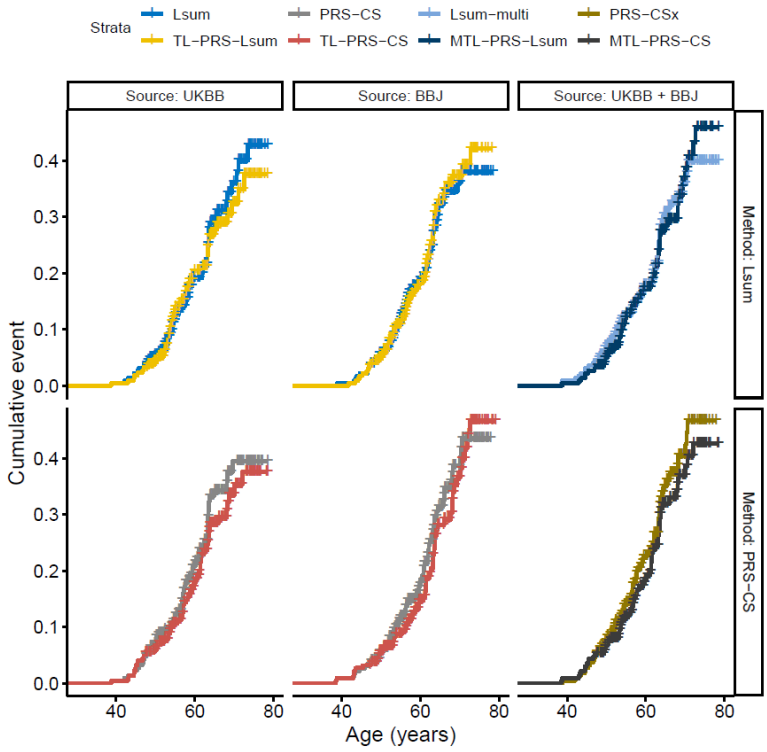


Figure S1. Prediction accuracy of TL-PRS(ind) and TL-PRS methods in simulations. (A) The proportion of causal variants is 0.1%; (B) The proportion of causal variants is 1.0%. In each setting, three different cross-population genetic correlations (0.4, 0.7 and 1.0) were considered. Heritability was fixed at 50%. Prediction accuracy was measured by the squared correlation (R^2) between the simulated and predicted phenotypes in the testing dataset, averaged across 20 simulation replicates. Error bar indicates the standard deviation of R^2 across simulation replicates.

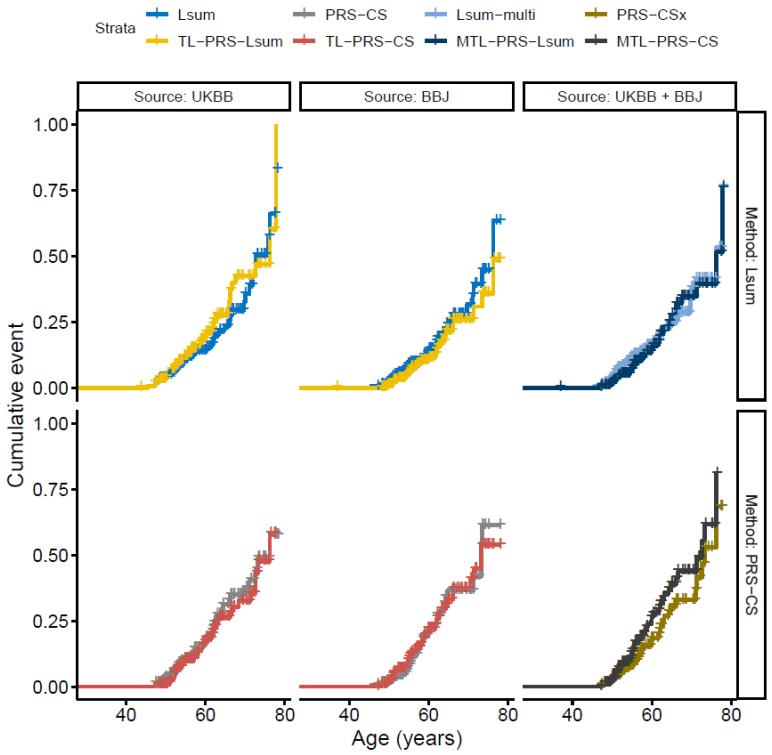
(a)


Strata
Lsum
TL−PRS−Lsum
PRS−CS
TL−PRS−CS
Lsum−multi
MTL−PRS−Lsum
PRS−CSx
MTL−PRS−CS
Source: UKBB
Source: BBJ
Source: UKBB + BBJ
Method: Lsum
Method: PRS−CS
Cumulative event
1.00
0.75
0.50
0.25
0.00
40
60
80
Age (years)

(b)


Strata
Lsum
TL−PRS−Lsum
PRS−CS
TL−PRS−CS
Lsum−multi
MTL−PRS−Lsum
PRS−CSx
MTL−PRS−CS
Source: UKBB
Source: BBJ
Source: UKBB + BBJ
Method: Lsum
Method: PRS−CS
Cumulative event
0.4
0.3
0.2
0.1
0.0
40
60
80
Age (years)

(c)



(d)

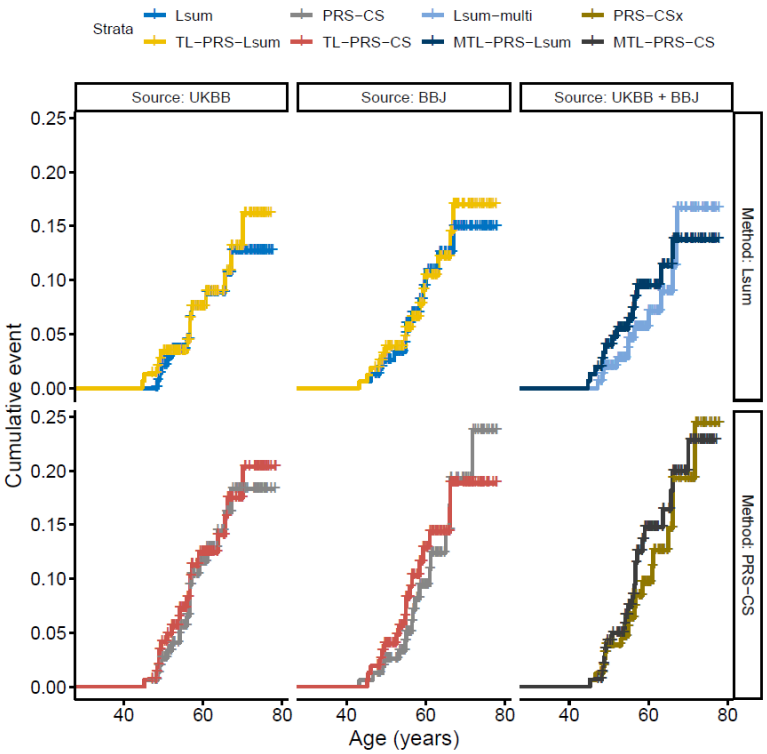


Figure S2. Cumulative event plot in terms of the top 10% PRS constructed by transfer learning methods and their baseline methods. Note that y-axes are on different scales in the different panels. (a) South Asian testing samples, Type 2 diabetes, Case: Control=419:2211; (b)South Asian testing samples, Coronary artery disease, Case: Control=362:2270; (c)African testing samples, Type 2 diabetes, Case: Control=177:1812; (d) African testing samples, Coronary artery disease, Case: Control=94:1902.

2. Supplemental Tables

Methods	Training dataset	Validation dataset	Testing dataset
TL-PRS	Only requires summary statistics	Individual-level data are recommended.	Requires individual-level data to assess prediction performance
TL-PRS (ind)	Requires individual-level data	Requires individual-level data	Requires individual-level data to assess prediction performance

Table S1. The model requirements of TL-PRS and TL-PRS (ind).

Target Population	Trait	Total sample size	Training sample size	Validation sample size	Testing sample size
South Asian (SAS)	Simulation and real phenotypes	10,285	5,000	2,650	2,635
African (AFR)	real phenotypes	8,168	4,000	2,169	1,999

Table S2. List of data sets used in simulations and analyses of real phenotypes.

(a)

Method	Pre-training data (GWAS summary statistics of source ancestry, and 1000 Genome Project Data)	Training data (target ancestry group)	Validation data (target ancestry group)	Testing data (target ancestry group)
PT	Train PT using pre-training data. Individual-level data are required.	Select the hyperparameters using the combination of training and validation data. Individual-level data are required.		Assess prediction performance using individual-level data.
Lsum	Train Lsum using pre-training data. Individual-level data are required.	Select the hyperparameters using the combination of training and validation data. Individual-level data are required.		Assess prediction performance using individual-level data.
PRS-CS	Train PRS-CS using pre-training data. Individual-level data are required.	Select the hyperparameters using the combination of training and validation data. Individual-level data are **not** required.		Assess prediction performance using individual-level data.
TL-PRS-Lsum	Pre-train Lsum using pre-training data. Individual-level data are required.	Validate the pre-trained baseline Lsum model, and use it to train TL-PRS-Lsum. Individual-level data are **not** required.	Select hyper parameters of the TL-PRS-Lsum model using validation data. Individual-level data are recommended.	Assess prediction performance using individual-level data.
TL-PRS-CS	Pre-train PRS-CS using pre-training data. Individual-level data are required.	Validate the pre-trained baseline PRS-CS model, and use it to train TL-PRS-CS. Individual-level data are **not** required.	Select hyper parameters of the TL-PRS-CS model using validation data. Individual-level data are recommended.	Assess prediction performance using individual-level data.

(b)

Method	Training data (target ancestry group)	Validation data (target ancestry group)	Testing data (target ancestry group)
PT-multi Lsum-multi PRS-CSx	Select the weights (hyperparameter) to linearly combine single-source prediction models using combination of training and validation data. Individual-level data are required to fine-tune the weight parameter.		Assess prediction performance using individual-level data
MTL-PRS-Lsum MTL-PRS-CS	Select the weights to construct the baseline Lsum-multi/PRS-CSx model, and then implement TL-PRS. Individual-level data are **not** required.	Select hyper parameters of the MTL-PRS models using validation data. Individual-level data are recommended.	Assess prediction performance using individual-level data

Table S3. The implementation of prediction methods in the simulation and application of UK Biobank (a) The implementation of single-source prediction methods. (b) The implementation of multi-source prediction methods.

		0.1% Causal			1% Causal		
Genetic Correlation		**0.4**	**0.7**	**1**	**0.4**	**0.7**	**1**
TL-PRS-Lsum	**Selected learning rate**	1000 (100,1000)	100 (100,100)	100 (10,100)	1000 (100,1000)	100 (100,100)	100 (100,100)
	Selected iteration	3 (2,14)	10 (8,11)	3 (2,7)	3 (2,8)	12 (9,13)	4 (3,6)
TL-PRS-CS	**Selected learning rate**	1000 (1000,1000)	1000 (1000,1000)	100 (100,100)	1000 (1000,1000)	1000 (1000,1000)	100 (100,1000)
	Selected iteration	7 (6,8)	3 (2,4)	10 (7,13)	8 (7,9)	4 (3,4)	14 (3,15)

Table S4. The selected learning rates and iterations of TL-PRS-Lsum and TL-PRS-CS in simulations. Two different percentages of causal variants (0.1% and 1% causal variants) and three different cross-population genetic correlations (0.4, 0.7, and 1.0) were considered. The value in each cell is the median and the values in the parentheses represents the 1st and 3rd quartile of the distribution.

Target population	trait	Best approach (rank 1)	Rank 2	Rank 3
South Asian	HDL	*MTL-PRS-CS*	*MTL-PRS-Lsum*	*TL-PRS-CS(UKBB)*
	LDL	*MTL-PRS-Lsum*	Lsum-multi	*MTL-PRS-CS*
	BMI	Lsum-multi	*MTL-PRS-Lsum*	*MTL-PRS-CS*
	TG	Lsum-multi	PRS-CSx	*MTL-PRS-CS*
	SBP	*MTL-PRS-Lsum*	*MTL-PRS-CS*	PRS-CSx
	DBP	*MTL-PRS-CS*	*TL-PRS-CS(UKBB)*	*MTL-PRS-Lsum*
	HGT	*MTL-PRS-Lsum*	*MTL-PRS-CS*	*TL-PRS-Lsum(UKBB)*
	CAD	*MTL-PRS-CS*	PRS-CSx	Lsum-multi
	T2D	*MTL-PRS-CS*	*MTL-PRS-Lsum*	PRS-CSx
African	HDL	*MTL-PRS-CS*	PRS-CSx	*MTL-PRS-Lsum*
	LDL	*TL-PRS-Lsum(BBJ)*	*MTL-PRS-Lsum*	*TL-PRS-Lsum(UKBB)*
	BMI	*MTL-PRS-Lsum*	Lsum-multi	*MTL-PRS-CS*
	TG	*MTL-PRS-CS*	*TL-PRS-CS(UKBB)*	PRS-CSx
	SBP	*TL-PRS-CS(UKBB)*	PRS-CSx	*MTL-PRS-CS*
	DBP	*MTL-PRS-CS*	*TL-PRS-CS(UKBB)*	PRS-CSx
	HGT	*MTL-PRS-Lsum*	lsum-multi	*MTL-PRS-CS*
	CAD	PT(UKBB)	*MTL-PRS-CS*	PT-multi
	T2D	*MTL-PRS-CS*	PRS-CS(UKBB)	*TL-PRS-CS(UKBB)*

Table S7. The top three methods for all nine traits in the South Asian and African ancestries in terms of predicted R^2. Single-source prediction methods (PT, Lsum, TL-PRS-Lsum, PRS-CS, TL-PRS-CS) based on UKBB and BBJ GWAS results and multi-source PRS methods (PT-multi, Lsum-multi, MTL-PRS-Lsum, PRS-CSx, MTL-PRS-CS) were included in the comparison and our approaches were highlighted using italics.

(a)

	0.1% Causal			1% Causal		
Genetic Correlation	**0.4**	**0.7**	**1**	**0.4**	**0.7**	**1**
PT	0.051 (0.011)	0.159 (0.026)	0.319 (0.019)	0.042 (0.007)	0.126 (0.012)	0.251 (0.016)
Lsum	0.060 (0.012)	0.188 (0.023)	0.380 (0.020)	0.053 (0.008)	0.157 (0.016)	0.317 (0.014)
TL-PRS-Lsum	0.205 (0.017)	0.267 (0.023)	0.389 (0.018)	0.083 (0.011)	0.175 (0.014)	0.321 (0.014)
PRS-CS	0.050 (0.012)	0.165 (0.024)	0.331 (0.017)	0.045 (0.006)	0.133 (0.016)	0.268 (0.015)
TL-PRS-CS	0.073 (0.013)	0.177 (0.024)	0.333 (0.016)	0.064 (0.009)	0.146 (0.015)	0.270 (0.016)

(b)

	0.1% Causal			1% Causal		
Genetic Correlation	**0.4**	**0.7**	**1**	**0.4**	**0.7**	**1**
PT	0.049 (0.011)	0.156 (0.023)	0.309 (0.017)	0.034 (0.007)	0.104 (0.015)	0.210 (0.014)
Lsum	0.060 (0.011)	0.186 (0.023)	0.373 (0.019)	0.045 (0.007)	0.134 (0.016)	0.268 (0.014)
TL-PRS-Lsum	0.190 (0.021)	0.262 (0.019)	0.382 (0.017)	0.069 (0.009)	0.152 (0.017)	0.273 (0.014)
PRS-CS	0.047 (0.012)	0.148 (0.023)	0.304 (0.020)	0.038 (0.006)	0.112 (0.017)	0.222 (0.014)
TL-PRS-CS	0.070 (0.013)	0.161 (0.023)	0.307 (0.020)	0.059 (0.008)	0.127 (0.018)	0.226 (0.014)

Table S8. Prediction accuracy of single-source polygenic prediction methods in simulations. Two different percentages of causal variants (0.1% and 1% causal variants) and three different cross-population genetic correlations (0.4, 0.7 and 1.0) were considered. Heritability was fixed at 50%. Prediction accuracy was measured by the squared correlation (R^2) between the simulated and predicted phenotypes in the testing dataset, averaged across 20 simulation replicates. The number in the parentheses indicates the standard deviation of R^2 across simulation replicates. (a) Simulation with 100,000 European GWAS sample size; (b) Simulation with 50,000 European GWAS sample size.