Overestimated Polygenic Prediction due to Overlapping Subjects in Genetic Datasets

Running title: Overestimation of Polygenic Risk Scores

David Keetae Park^{*1}, Mingshen Chen^{*2}, Seungsoo Kim³, Yoonjung Yoonie Joo,^{4,5} Rebekah K. Loving,⁶ Hyoung Seop Kim,⁷ Jiok Cha,⁵ Shinjae Yoo,⁸ Jong Hun Kim⁹

* Equally contributed to this study.

¹ Department of Biomedical Engineering, Columbia University, New York, USA

² Department of Applied Mathematics & Statistics, Stony Brook University, New York, USA

³ Department of Obstetrics and Gynecology, Columbia University Irving Medical Center, New York, NY, USA

⁴ Institute of Data Science, Korea University, Seoul, South Korea

⁵ Department of Psychology, Brain and Cognitive Sciences, AI Institute, Seoul National University, Seoul, South Korea

⁶ Department of Biology, California Institute of Technology, Pasadena, USA

⁷ Department of Physical Medicine and Rehabilitation, Dementia Center, National Health Insurance Service Ilsan Hospital, Goyang, South Korea

⁸ Computational Science Initiative, Brookhaven National Laboratory, New York, USA

⁹ Department of Neurology, Dementia Center, National Health Insurance Service Ilsan Hospital, Goyang, South Korea

Corresponding to:

Jong Hun Kim, MD, PhD

26 Department of Neurology, Dementia Center, National Health Insurance Service Ilsan Hospital
 27 100 Ilsan-ro Ilsandong-gu, Goyang, Gyeonggi-Do, 10444, South Korea
 28 Tel.: 82-31-900-0213, Fax.: 82-31-900-0343
 29 Email: jh7521@naver.com
 30
 31 Shinjae Yoo, PhD
 32 Computational Scientist, Brookhaven National Lab.
 33 Computer Science and Math, Building 725, Room 2-189, Upton, NY, 11973, USA
 34 Tel.: +1-631-344-4286, Fax.: 631-344-5751
 35 Email: sjyoo@bnl.gov
 36
 37

ABSTRACT

Recently, polygenic risk score (PRS) has gained significant attention in studies involving complex genetic diseases and traits. PRS is often derived from summary statistics, from which the independence between discovery and replication sets cannot be monitored. Prior studies, in which the independence is strictly observed, report a relatively low gain from PRS in predictive models of binary traits. We hypothesize that the independence assumption may be compromised when using the summary statistics, and suspect an overestimation bias in the predictive accuracy. To demonstrate the overestimation bias in the replication dataset, prediction performances of PRS models are compared when overlapping subjects are either present or removed. We consider the task of Alzheimer's disease (AD) prediction across genetics datasets, including the International Genomics of Alzheimer's Project (IGAP), AD Sequencing Project (ADSP), and Accelerating Medicine Partnership - Alzheimer's Disease (AMP-AD). PRS is computed from either sequencing studies for ADSP and AMP-AD (denoted as rPRS) or the summary statistics for IGAP (sPRS). Two variables with the high heritability in UK Biobank, hypertension, and height, are used to derive an exemplary scale effect of PRS. Based on the scale effect, the expected performance of sPRS is computed for AD prediction. Using ADSP as a discovery set for rPRS on AMP-AD, ΔAUC and ΔR^2 (performance gains in AUC and R^2 by PRS) record 0.069 and 0.11, respectively. Both drop to 0.0017 and 0.0041 once overlapping subjects are removed from AMP-AD. sPRS is derived from IGAP, which records ΔAUC and ΔR^2 of 0.051 ± 0.013 and 0.063 ± 0.015 for ADSP and 0.060 and 0.086 for AMP-AD, respectively. On UK Biobank, rPRS performances for hypertension assuming a similar size of discovery and replication sets are 0.0036 ± 0.0027 (ΔAUC) and 0.0032 ± 0.0028 (ΔR^2). For height, ΔR^2 is 0.029 ± 0.0037 . Considering the high heritability of hypertension and height of UK Biobank, we conclude that sPRS results from AD databases are inflated. The higher performances relative to the size of the discovery set were observed in PRS studies of several

diseases. PRS performances for binary traits, such as AD and hypertension, turned out unexpectedly low. This may, along with the difference in linkage disequilibrium, explain the high variability of PRS performances in cross-nation or cross-ethnicity applications, i.e., when there are no overlapping subjects. Hence, for sPRS, potential duplications should be carefully considered within the same ethnic group.

INTRODUCTION

Recently, genetic studies involving a polygenic risk score (PRS) have dramatically grown in number, and sophisticated tools and methodologies are being developed for its use¹⁻³. Along with heritability, PRS has become an important metric for explaining complex genetic diseases (e.g. Alzheimer's disease, AD) and traits⁴⁻⁸. A typical PRS study involves both the discovery and replication phases^{9,10}. In the discovery phase, two different methods are used to develop PRS. PRS is essentially derived from the raw genetic data, denoted as rPRS. Instead, the summary statistics from large-scale genetics studies or GWAS catalogs may be also be used, which we abbreviate as sPRS. Polygenic prediction performance is then evaluated by the marginal contribution of the PRS term in a regression model on target clinical application^{10,11}.

An underlying assumption of PRS models is that the subjects from the discovery set do not overlap with those of the replication set¹⁰. However, our preliminary analyses (Fig. 1a) demonstrate a significant number of identical subjects across multiple genetic datasets. The overlapping subjects may be identified and removed for rPRS using the raw genetic data, a challenge remains for sPRS in which raw data is inaccessible. Therefore, we posit that a strict level of independence across datasets is hard to achieve with sPRS. This may pose serious issues to related fields, since the subject-level dependence across datasets may not only inflate the polygenic prediction performance, but also prevent generalizable applications of the developed model.

Among prior studies, we identify multiple signs of potential inflation in PRS performance attributable to overlapping subjects. First, if the independence between discovery and replication sets is clearly stated in the paper, the PRS performance of binary traits not statistically significant, while if not, results were highly variable or inflated¹²⁻¹⁶. For instance, for models in which independence is explicitly controlled during development, even with a large-scale national biobank, PRS contributed less than 2% in the model accuracy¹⁷. As a similar line

of evidence, in a large-scale finnish study¹², polygenic predictions were not significant for datasets in which the independence is guaranteed, while in other groups without the guarantee performed significantly higher. Second, we argue that the low portability PRS in trans-ethnic applications may serve as another evidence of overlapping subjects developed for within-ethnic models. Low trans-ethnic portability, yet is not fully understood, has been attributed to different linkage disequilibrium (LD) structures, allele frequencies, and marginal effect size variations according to ancestries. We suspect that the strictly preserved independence of subjects between different ethnicities limits the prediction performance of trans-ethnic models (Fig. 1b). In other words, it is plausibly the level of dependence (i.e., overlapping subjects) between datasets that could explain the gap between within- and trans-ethnic generalization capacities of PRS models.

The aforementioned clues, albeit circumstantial, led us to a systematic investigation for detecting and quantifying the overestimation bias in sPRS due to overlapping subjects. On AD prediction, we first prove that PRS models overfit to overlapping subjects, resulting in overestimated prediction performances on the replication set. Then we extend our experiments using UK Biobank data to derive the scale effect of the inflation and brief guidelines for detecting the bias in sPRS.

RESULTS

Overview of the study

Fig. 1 illustrates an overview of our study design. We design an rPRS model that derives SNPs from ADSP, which is then replicated for all subjects from AMP-AD (see Fig. 1b for details and Fig. 2a for results). After removing subjects from AMP-AD that have close kinship with the ADSP study, predictions are made again on AMP-AD, and the two results with and without close subjects are compared. We also compare rPRS and sPRS. To this end, ADSP is divided

into 9:1 (discovery:replication) splits for ten-fold cross-validation for rPRS (Fig. 2b), while AMP-AD data are used as another replication set for rPRS (Fig. 2c). An increasing degree of overlapping bias is observed with an expanding number of subjects in the replication set being replaced by samples from the corresponding discovery set (Fig. 2d). We further demonstrate that sPRS also overestimates prediction performances (Fig. 1b, Fig. 2b, and Fig. 2c). We compare the sPRS prediction results against rPRS to indirectly infer the level of overfitting. However, the number of subjects in the IGAP study is larger than that of ADSP, and PRS predictions may not be directly comparable due to the scale effect, where a larger discovery set may result in better generalization capability.

To adjust for the scale effect, a large number of samples in UK Biobank is leveraged in terms of two phenotypes with higher heritabilities than AD¹⁸⁻²⁰, namely hypertension and height, which respectively are binary and non-binary variables. With a varying number of subjects in the discovery set, the rate of change per subject in rPRS accuracies is inferred. Finally, we estimate the level of overestimation bias in sPRS for AD prediction (Fig. 1c and Fig. 3).

PRS prediction performance after excluding genetically related individuals

432 identical subjects overlap between ADSP and AMP-AD (Fig. 1b). Using ADSP as the discovery set, rPRS on all subjects in AMP-AD results in ΔAUC of 0.069 ($p = 1.51 \times 10^{-10}$). After removing the overlapping individuals from AMP-AD, the ΔAUC decreases to 0.0017. Notably, ΔAUC loses its statistical significance ($p=0.57$). ΔR^2 shows a similar level of deflation, which drops from 0.11 to 0.0041 by removing the identical subjects (Fig. 2a). PRS performances are only slightly affected when close relatives are removed by applying a lower cutoff of PI_HAT, and ΔAUC still shows no statistical significance (Supplementary Table 2).

rPRS and sPRS Performances on AD prediction

Fig. 2b, Fig. 2c, and Supplementary Table 3 show the comparison results of rPRS and sPRS. ADSP data are divided into the discovery and replication datasets with 9:1 cross-validation for assessment of rPRS (Fig. 2b). Another test of rPRS is evaluated on non-overlapping data of AMP-AD (n=692) (Fig. 2c). We compute sPRS using summary statistics derived from the first stage of IGAP study (Fig. 2b and Fig. 2c). sPRS on ADSP is evaluated on 10 replication folds (i.e., sets), while, on AMP-AD, the whole data (n=1133) are regarded as a replication set. Unlike rPRS, sPRS is evaluated on all AMP-AD data as overlapping subjects are not identifiable against IGAP.

In rPRS, a respective set (ΔAUC , ΔR^2) of ADSP and AMP-AD data are (0.0071 $\pm$ 0.0052, 0.0077 $\pm$ 0.0045) and (0.0013 $\pm$ 0.00091, 0.0011 $\pm$ 0.0018). P-values are not significant ($p>0.05$). In other words, when the independence of data is guaranteed, PRS for AD displays unexpectedly low performance. In sPRS, a respective set (ΔAUC , ΔR^2) of ADSP and AMP-AD are as high as (0.051 $\pm$ 0.013, 0.063 $\pm$ 0.015) and (0.060, 0.086). The results of sPRS are significantly inflated in comparison to those of rPRS.

PRS performances are sensitively affected by dependency

Given the suspected inflation in sPRS, we wanted to show how the PRS results are sensitively affected according to the subject overlap, as well as estimate the number of overlapping subjects in sPRS. To this end, we simulate an increasing number of subjects from the discovery set to be added into the replication set. All ADSP data (n=10,923) are used to derive an rPRS model, which subsequently is evaluated on a mixture of AMP-AD (independent replication set, n=692) and ADSP data randomly selected from replication splits of ten-fold cross-validation (fully dependent replication set, n=692), for which the portion of the latter increases from 0 to 100% via an increment of 10%. sPRS is derived from the first stage data of the IGAP study and evaluated in the same way (Fig. 2d). As expected, all ΔAUC , ΔR^2 , and $-\log(P)$ monotonically increase in a growing portion of subject overlap. sPRS performances

remained relatively unchanged while maintaining the inflated values greater than those in independent rPRS (Fig. 2b and Fig. 2c). To show that the AD characteristics of the replication sets are maintained, performances of *APOE* $\epsilon 4$ are also displayed. Judging by the intersection of two lines representing rPRS and sPRS trends, we infer at least 10% of participants in AMP-AD or ADSP are included in IGAP. However, this holds true only if the number of subjects in discovery sets is equal. Meanwhile, the number of subjects in IGAP is five times larger than that of ADSP. Therefore, to confirm the suspected inflation in sPRS, we must investigate the scale effect of the discovery set.

Upper bounds of PRS performance derived from UK Biobank data

UK Biobank data are leveraged to infer the upper bounds for the scale effect of sPRS. Hypertension and height are selected as two target variables to investigate the scale effect due to their notably high heritability (Fig. 1c). We posit that AD prediction scores are bounded by both hypertension and height thanks to superior heritability and the scale effect. Subjects from UK Biobank are split by 9:1 following the ten-fold cross-validation scheme.

To investigate the scale effect of AD, rPRS is evaluated on three different sizes of the discovery set, corresponding to 9k, 60k, 300k (i.e., full data) and roughly equal to the size of ADSP, IGAP, and UK Biobank, respectively (Supplementary Table 4). Fig. 3 summarizes and visualizes the results. For hypertension, with a discovery set size of 60k subjects, a set of metrics (ΔAUC , ΔR^2) is (0.0033 ± 0.00047 , 0.0033 ± 0.00051). When the size is 300k, (ΔAUC , ΔR^2) is (0.012 ± 0.0014 , 0.012 ± 0.0017), which is still significantly smaller than the sPRS of ADSP and AMP-AD (Fig. 2b and Fig. 2c) corresponding to (0.051 ± 0.013 , 0.063 ± 0.015) and (0.060 , 0.086), respectively.

For height, ΔR^2 for 60k and 300k sizes are 0.028 ± 0.0015 and 0.075 ± 0.0056 (Fig. 3a and Supplementary Table 5). Relatively larger contributions of PRS in height compared to hypertension reflect the greater heritability. In both variables, PRS scores plateau at

approximately 50k of SNPs (Fig. 3b). The results of sPRS using IGAP (54k) as a discovery set are highly inflated compared to those of hypertension and height at a similar scale (60k). Scores display growth linear or sublinear to the size of the discovery set (Fig. 3c).

To obtain statistical significance for ΔAUC , a substantially large number of subjects are required for both the discovery and replication sets (Supplementary Fig.) For instance, for a 60k-sized discovery set, more than 10k subjects are needed in the replication set for sufficient power ($p < 0.01$). Lassosum², which uses the LD information, shows two-fold higher performance than PRS (Supplementary Table 6), but the linear pattern of the scale effect remains unchanged.

DISCUSSION

This study illuminates the overestimation bias in PRS studies. In AD prediction, we prove the presence of latently overlapping subjects for sPRS and demonstrate performance inflation. Developing sPRS without knowledge of the overlapping individuals raises concerns over overestimation and the model's generalizability. We argue that overestimation should be suspected when the PRS contribution by the referenced PRS methods²¹ exceeds 1–2% of ΔR^2 for binary traits¹⁷ and 7–8% of ΔR^2 for non-binary traits, which are the respective upper limits drawn by hypertension and height from UK Biobank.

As evidenced by the three separate AD studies sharing identical subjects, subject overlap may be prevalent in other cohort studies and large-scale meta-analyses. In our study, ADSP and AMP-AD data include 432 identical subjects, which corresponds to 38.12% of AMP-AD data. Also, ADSP and ADNI have 21 overlapping subjects (Fig. 1a). As genetic studies are conducted across multiple centers, the odds of having duplicated subjects are high among different initiatives. Therefore, planning of PRS studies requires considerable attention to exclude identical persons.

Several existing studies support the relationship between the independence of datasets and the possibility of the overestimation bias. Concretely, we observe the trend in which PRS is performing significantly lower when the independence of data is explicitly controlled. For instance, when overlapping subjects are removed, PRS contributed less than 2% in accuracy even with large discovery sets from a national biobank^{12,17}. In a large-scale finnish study for coronary artery disease (CAD), arterial fibrillation (AF), type 2 diabetes mellitus (T2DM), breast cancer (BrC) and prostate cancer (PrC), the independencies in CAD and AF were clarified¹². The PRS contributions of two diseases were statistically insignificant (0.3% and 0.9%, respectively). In contrast, the PRSs of T2DM, BrC, and PrC used the summary statistics derived from large meta-analyses which included finnish population, the contributions of PRS were significantly high (2%, 3.9%, and 2.9%, respectively).

We propose two potential signs when the overestimation bias should be suspected. The first is when a PRS model achieves surprisingly high performance without sufficient participants. The number of subjects and SNP heritability are important factors in PRS performance¹⁸. For example, hundreds of thousands of subjects in the discovery set would be required for PRS to be used in disease prediction¹⁸. Computing the PRS using UK Biobank for traits with prominent heritability sets the upper bounds for other traits with lower heritability. Thus, we reveal that a replication set of 10k is required in the 60k discovery set for statistically significant for AUC increment by PRS. In the studies registered in the PGS catalog, the median number of subjects in the replication sets is 6995 and 24,573 in the discovery sets²². In addition, a systematic review of AD PRS studies reveals that the replication set size ranges from 59 to 116,666²³. About one third of them, the sample size is less than a thousand. Therefore, an abundance of prior sPRS studies²³⁻²⁵ may not suffice as the number of samples required for statistically significant results.

Second, for particular phenotypes, high variability in performances across different nations or races using the same discovery set may be another sign of overestimation. PRS

performances plummet if the discovery and replication sets are from different countries or ethnicities¹²⁻¹⁶. However, a line of evidence suggests that trans-ethnic portability remains in many traits²⁶⁻²⁸, even if the inherent differences in LD structures across races prevent causal variants from being correctly reflected in PRS²⁹. For instance, Martin et al. reported equivalent performances within the confidence interval for intra- and trans-ethnic replication sets in four out of five binary traits and 11 out of 17 non-binary traits of BioBank Japan²⁶. Also, sPRS developed with Europeans showed a 50% discounted performance for East Asians and 25% for Africans²⁶. Therefore, the gap (or the variability) between intra-ethnic and trans-ethnic evaluations of PRS is attributable to overestimation bias, rather than a low trans-ethnic portability. For example, if the performance for a trans-ethnic application is preserved at less than 25% for intra-ethnic evaluation, overlapping bias might be suspected. In other words, sPRS suffers in trans-nation or trans-ethnic studies, since the subject-level independence strictly holds.

PRS also could be developed using GWA ($p < 5 \times 10^{-8}$) SNPs in core genes curated from multiple GWA studies³⁰⁻³², which we argue are not free from the overestimation bias. Using GWA SNPs is justified because it substitutes the P-value thresholding step required in conventional PRS studies for selecting SNPs. Also, GWA SNPs tend to show highly significant P-values, and therefore are regarded as reliable. Moreover, GWA SNPs information can be conveniently accessible via reviewing prior works, even if the authors do not release summary statistics. However, GWA SNPs are frequently found in the uninterpretable non-coding regions³³, and PRS performances increase with a higher number of SNPs then plateau (Fig. 2 and Fig. 3)³⁴. Therefore, PRS models from GWA SNPs may overfit to the discovery set. Hence, it follows that the odds of overestimation bias are high when the GWA SNPs are selected from multiple studies.

For the non-binary trait height, PRS has a greater contribution than binary trait hypertension. Similar trends have been observed in a prior study²⁶. Heritability for height and hypertension was reported as 49.7% and 14.7%, respectively²⁰. Although a superior heritability

of height partially explains the performance gap, characteristics of each phenotype may also play a role. For instance, while measuring height is straightforward, a diagnosis of hypertension can depend on age. Thus, a subset of the control group can later be diagnosed with hypertension.

One limitation in our study is the indirect derivation of the scale effect to prove the overestimation bias of sPRS in AD prediction. We justify the choice of our methods based on three reasons. First, hypertension and height have higher heritabilities than AD^{19,20,30}. Second, the number of SNPs used for UK Biobank is an order of magnitude larger than those used for AD data analysis. Finally, the number of available subjects in UK Biobank is five-fold larger than IGAP. Therefore, we argue that PRS performances of hypertension from UK Biobank are sufficient upper bounds not only for AD prediction but also for most binary complex genetic traits/diseases.

As the risk of overestimation bias is evaluated in sPRS studies, care must be taken to prevent overlap, especially within the same ethnicity. While applications of genetic studies continue to gain momentum and many countries create large-scale biobanks³⁵⁻³⁸, PRS developed from large meta-analyses that curate and merge data from several countries should be screened in advance to filter out overlapping subjects. Researchers often release summary statistics to help further research²². Direct comparisons between a replication set and large-scale data are difficult when using summary statistics. As such, we showcase both direct and indirect methods to probe overestimation bias within the same ethnicity—either of which, we argue, must be mandatory to improve PRS reliability.

METHODS

Participants

In this work, AD genetic studies—International Genomics of Alzheimer's Project (IGAP), Alzheimer's Disease Sequencing Project (ADSP), and AMP-AD—are used to demonstrate overestimation bias in PRS³⁹⁻⁴⁵. Non-Hispanic white individuals are used, and their cross-study genetic relatedness is revealed by principal component (PC) and identity-by-state analyses. After quality control, our final analyses include 10,293 participants from ADSP and 1,133 from - AD (Supplementary Table 1). In the UK Biobank database⁴⁶, a total of 342,318 white-British participants had hypertension and height records. Refer to the Supplementary Material for additional information about the study datasets, sequencing methods, and quality control processes.

Statistical analyses

We perform logistic regressions for binary traits and linear regressions for a continuous phenotype, using PLINK (v1.9)⁴⁷. Three different regression models are constructed. First, a simple regression model is used with PRS as the only covariate. Second, Model II denotes a multivariable regression without PRS, consisting of additional covariates highly related to the phenotypes. In Model III, we introduce PRS as an additional covariate to Model II. In both models, we control for 20 leading PCs

Model I: $y = \beta_{\text{PRS}}$

Model II: $y = x_{\text{co}}^T a_{\text{co}} + x_{\text{PC}}^T b_{\text{PC}}$

Model III: $y = x_{\text{co}}^T a_{\text{co}} + x_{\text{PC}}^T b_{\text{PC}} + \beta_{\text{PRS}}$

$$x_{\text{co}} = \begin{pmatrix} 1 \\ x_{\text{co},1} \\ \vdots \\ x_{\text{co},n_1} \end{pmatrix}, a_{\text{co}} = \begin{pmatrix} a_0 \\ a_1 \\ \vdots \\ a_{n_1} \end{pmatrix}, x_{\text{PC}} = \begin{pmatrix} x_{\text{PC},1} \\ x_{\text{PC},2} \\ \vdots \\ x_{\text{PC},n_2} \end{pmatrix}, b_{\text{PC}} = \begin{pmatrix} b_1 \\ b_2 \\ \vdots \\ b_{n_2} \end{pmatrix},$$

Where $x_{\text{co}}, a_{\text{co}} \in \mathbb{R}^{n_1+1}$ are vectors of general covariates (e.g., age and sex) and corresponding coefficients, respectively, while $x_{\text{PC}}, b_{\text{PC}} \in \mathbb{R}^{n_2}$ are vectors of PCs and PC coefficients,

respectively. Here, x_{PRS} denotes the PRS term. Throughout the manuscript, we focus on measuring the additive gain of PRS in Model III, on top of Model II.

For AD datasets, common covariates include sex, APOE $\epsilon 4$ status, and the sequencing centers. PCs are computed using the principal component analysis function of PLINK (v1.9)⁴⁷. For UK Biobank, age, sex, and array types (UK Biobank Axiom array or UK BiLEVE Axiom array) are considered as covariates. Here, we download 40 PCs pre-calculated with fastPCA⁴⁸. For hypertension of UK Biobank, body mass index additionally is included in the covariates. For binary traits, the areas under receiver operating characteristic (AUC) and Nagelkerke's pseudo- R^2 are used to assess model performances, which are calculated using "pROC" and "fsmB" packages of R (v4.0.3), respectively^{49,50}. Performance improvements from PRS are determined by subtracting AUC and R^2 of Model II from those of Model III, which we respectively label as ΔAUC and ΔR^2 . The statistical significance of ΔAUC is examined using DeLong's test⁵¹. For height, a non-binary trait, we compute the adjusted- R^2 via "lm" in the R program. PRS contributions are determined by comparing Model II and Model III with the extra sum of squares test.

Cross-validation

For ten-fold cross-validation tests, we balance the number of samples between the discovery and replication splits based on each covariate in the statistical analyses using "StratifiedKFold" function from Python's (v3.8) "scikit-learn" (v0.24.1) package⁵². When testing a part of the cross-validated datasets, samples are balanced over covariates using the R (v4.0.3) "sampling" (v2.9) package⁵³.

Computation of PRS

For computing PRS, we select common ($MAF \geq 1\%$) SNPs and use summary statistics from discovery sets, followed by measuring PRS in replication datasets. After selecting SNPs

with $p < 0.5$ in the association tests using the discovery dataset, we perform clumping with the window of ± 1 Mbp and $r^2 < 0.1$. Clumping is performed using PLINK (v1.9)⁴⁷. For AD genetic studies, we exclude any SNPs within 1 Mbp of the *APOE* (apolipoprotein E) region. When analyzing the effect of the number of SNPs on the results, the SNPs are selected in the order from the lowest P-value. We construct PRS with PRSice (v2.3) and Lassosum (v0.4.5)^{2,21}.

COMETING INTERESTS

The authors declare that they have no competing interests.

DATA AVAILABILITY

The dataset(s) supporting the conclusions of this article are available in webpages of UK Biobank (<https://www.ukbiobank.ac.uk/>), ADSP (accession phs000572.v1.p1; https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs000572.v1.p1), Mayo RNAseq study (accession syn5550404; <http://dx.doi.org/10.1038/sdata.2016.89>), Mount Sinai Brain Bank (MSBB) study (accession syn3159438; <http://dx.doi.org/10.1038/sdata.2018.185>), and Religious Orders Study and Memory and Aging Project (ROSMAP) Study (accession syn3159438; <http://dx.doi.org/10.1038/mp.2017.20>).

AUTHOR CONTRIBUTIONS

D.K.P., M.C., S.Y., and J.H.K. conceived and designed the study. D.K.P. and J.H.K. performed statistical analysis. D.K.P., M.C., and J.H.K. analyzed the genetic data. All authors discussed the results and implications and commented on the manuscript at all stages. M.C., S.K., Y.Y.J., R.K.L., H.S.K., J.C., and S.Y. gave technical support and conceptual advice. D.K.P., S.Y., and J.H.K. wrote the paper. All co-authors contributed to the final manuscript.

REFERENCES

1. Euesden, J., Lewis, C.M. & O'Reilly, P.F. PRSice: Polygenic Risk Score software. *Bioinformatics* **31**, 1466-8 (2015).
2. Mak, T.S.H., Porsch, R.M., Choi, S.W., Zhou, X. & Sham, P.C. Polygenic scores via penalized regression on summary statistics. *Genet Epidemiol* **41**, 469-480 (2017).
3. Prive, F., Arbel, J. & Vilhjalmsen, B.J. LDpred2: better, faster, stronger. *Bioinformatics* **36**, 5424-5431 (2020).
4. Yang, J., Lee, S.H., Goddard, M.E. & Visscher, P.M. GCTA: a tool for genome-wide complex trait analysis. *Am J Hum Genet* **88**, 76-82 (2011).
5. International Schizophrenia, C. *et al.* Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. *Nature* **460**, 748-52 (2009).
6. Khera, A.V. *et al.* Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations. *Nat Genet* **50**, 1219-1224 (2018).
7. Escott-Price, V. *et al.* Common polygenic variation enhances risk prediction for Alzheimer's disease. *Brain* **138**, 3673-84 (2015).
8. Sims, R., Hill, M. & Williams, J. The multiplex model of the genetics of Alzheimer's disease. *Nat Neurosci* **23**, 311-322 (2020).
9. Choi, S.W., Mak, T.S. & O'Reilly, P.F. Tutorial: a guide to performing polygenic risk score analyses. *Nat Protoc* **15**, 2759-2772 (2020).
10. Wand, H. *et al.* Improving reporting standards for polygenic scores in risk prediction studies. *Nature* **591**, 211-219 (2021).
11. Tzoulaki, I., Liberopoulos, G. & Ioannidis, J.P. Assessment of claims of improved prediction beyond the Framingham risk score. *JAMA* **302**, 2345-52 (2009).

12. Mars, N. *et al.* Polygenic and clinical risk scores and their impact on age at onset and prediction of cardiometabolic diseases and common cancers. *Nat Med* **26**, 549-557 (2020).
13. Dikilitas, O. *et al.* Predictive Utility of Polygenic Risk Scores for Coronary Heart Disease in Three Major Racial and Ethnic Groups. *Am J Hum Genet* **106**, 707-716 (2020).
14. Duncan, L. *et al.* Analysis of polygenic risk score usage and performance in diverse human populations. *Nat Commun* **10**, 3328 (2019).
15. Dube, J.B. *et al.* Genetic determinants of "cognitive impairment, no dementia". *J Alzheimers Dis* **33**, 831-40 (2013).
16. Marden, J.R., Walter, S., Tchetgen Tchetgen, E.J., Kawachi, I. & Glymour, M.M. Validation of a polygenic risk score for dementia in black and white individuals. *Brain Behav* **4**, 687-97 (2014).
17. Elliott, J. *et al.* Predictive Accuracy of a Polygenic Risk Score-Enhanced Prediction Model vs a Clinical Risk Score for Coronary Artery Disease. *JAMA* **323**, 636-645 (2020).
18. Dudbridge, F. Power and predictive accuracy of polygenic risk scores. *PLoS Genet* **9**, e1003348 (2013).
19. Holland, D. *et al.* Beyond SNP heritability: Polygenicity and discoverability of phenotypes estimated with a univariate Gaussian mixture model. *PLoS Genet* **16**, e1008612 (2020).
20. Canela-Xandri, O., Rawlik, K. & Tenesa, A. An atlas of genetic associations in UK Biobank. *Nat Genet* **50**, 1593-1599 (2018).
21. Choi, S.W. & O'Reilly, P.F. PRSice-2: Polygenic Risk Score software for biobank-scale data. *Gigascience* **8**(2019).
22. Lambert, S.A. *et al.* The Polygenic Score Catalog as an open database for reproducibility and systematic evaluation. *Nat Genet* **53**, 420-425 (2021).

23. Harrison, J.R., Mistry, S., Muskett, N. & Escott-Price, V. From Polygenic Scores to Precision Medicine in Alzheimer's Disease: A Systematic Review. *J Alzheimers Dis* **74**, 1271-1283 (2020).
24. Oram, R.A. *et al.* A Type 1 Diabetes Genetic Risk Score Can Aid Discrimination Between Type 1 and Type 2 Diabetes in Young Adults. *Diabetes Care* **39**, 337-44 (2016).
25. Harrison, T.M. *et al.* An Alzheimer's Disease Genetic Risk Score Predicts Longitudinal Thinning of Hippocampal Complex Subregions in Healthy Older Adults. *eNeuro* **3**(2016).
26. Martin, A.R. *et al.* Clinical use of current polygenic risk scores may exacerbate health disparities. *Nat Genet* **51**, 584-591 (2019).
27. Domingue, B.W. *et al.* Polygenic risk predicts obesity in both white and black young adults. *PLoS One* **9**, e101596 (2014).
28. Koyama, S. *et al.* Population-specific and trans-ancestry genome-wide analyses identify distinct and shared genetic risk loci for coronary artery disease. *Nat Genet* **52**, 1169-1177 (2020).
29. Amariuta, T. *et al.* Improving the trans-ancestry portability of polygenic risk scores by prioritizing variants in predicted cell-type-specific regulatory elements. *Nat Genet* **52**, 1346-1354 (2020).
30. Graff, R.E. *et al.* Cross-cancer evaluation of polygenic risk scores for 16 cancer types in two large cohorts. *Nat Commun* **12**, 970 (2021).
31. Belsky, D.W. *et al.* Development and evaluation of a genetic risk score for obesity. *Biodemography Soc Biol* **59**, 85-100 (2013).
32. Mavaddat, N. *et al.* Prediction of breast cancer risk based on profiling with common genetic variants. *J Natl Cancer Inst* **107**(2015).
33. Freedman, M.L. *et al.* Principles for the post-GWAS functional characterization of cancer risk loci. *Nat Genet* **43**, 513-8 (2011).

34. Ware, E.B. *et al.* Heterogeneity in polygenic scores for common human traits. *bioRxiv*, 106062 (2017).
35. Nagai, A. *et al.* Overview of the BioBank Japan Project: Study design and profile. *J Epidemiol* **27**, S2-S8 (2017).
36. Sakaue, S. *et al.* Trans-biobank analysis with 676,000 individuals elucidates the association of polygenic risk scores of complex traits with human lifespan. *Nat Med* **26**, 542-548 (2020).
37. Ramirez, A.H., Gebo, K.A. & Harris, P.A. Progress With the All of Us Research Program: Opening Access for Researchers. *JAMA* **325**, 2441-2442 (2021).
38. Chen, Z. *et al.* China Kadoorie Biobank of 0.5 million people: survey methods, baseline characteristics and long-term follow-up. *Int J Epidemiol* **40**, 1652-66 (2011).
39. Lambert, J.C. *et al.* Meta-analysis of 74,046 individuals identifies 11 new susceptibility loci for Alzheimer's disease. *Nat Genet* **45**, 1452-8 (2013).
40. Beecham, G.W. *et al.* The Alzheimer's Disease Sequencing Project: Study design and sample selection. *Neurol Genet* **3**, e194 (2017).
41. Greenwood, A.K. *et al.* The AD Knowledge Portal: A Repository for Multi-Omic Data on Alzheimer's Disease and Aging. *Curr Protoc Hum Genet* **108**, e105 (2020).
42. Crane, P.K., Foroud, T., Montine, T.J. & Larson, E.B. Alzheimer's Disease Sequencing Project discovery and replication criteria for cases and controls: Data from a community-based prospective cohort study with autopsy follow-up. *Alzheimers Dement* **13**, 1410-1413 (2017).
43. Allen, M. *et al.* Human whole genome genotype and transcriptome data for Alzheimer's and other neurodegenerative diseases. *Sci Data* **3**, 160089 (2016).
44. Wang, M. *et al.* The Mount Sinai cohort of large-scale genomic, transcriptomic and proteomic data in Alzheimer's disease. *Sci Data* **5**, 180185 (2018).

- 470 45. De Jager, P.L. *et al.* A multi-omic atlas of the human frontal cortex for aging and
471 Alzheimer's disease research. *Sci Data* **5**, 180142 (2018).
- 472 46. Bycroft, C. *et al.* The UK Biobank resource with deep phenotyping and genomic data.
473 *Nature* **562**, 203-209 (2018).
- 474 47. Chang, C.C. *et al.* Second-generation PLINK: rising to the challenge of larger and richer
475 datasets. *Gigascience* **4**, 7 (2015).
- 476 48. Galinsky, K.J. *et al.* Fast Principal-Component Analysis Reveals Convergent Evolution of
477 ADH1B in Europe and East Asia. *Am J Hum Genet* **98**, 456-472 (2016).
- 478 49. Nagelkerke, N.J.D. A note on a general definition of the coefficient of determination.
479 *Biometrika* **15**, 691-693 (1991).
- 480 50. Robin, X. *et al.* pROC: an open-source package for R and S+ to analyze and compare
481 ROC curves. *BMC Bioinformatics* **12**, 77 (2011).
- 482 51. DeLong, E.R., DeLong, D.M. & Clarke-Pearson, D.L. Comparing the areas under two or
483 more correlated receiver operating characteristic curves: a nonparametric approach.
484 *Biometrics* **44**, 837-45 (1988).
- 485 52. Pedregosa, F. *et al.* Scikit-learn: Machine Learning in Python. *The Journal of Machine*
486 *Learning Research* **12**, 2825–2830 (2011).
- 487 53. Tillé, Y. & Matei, A. The R sampling package. in *The Fifth International Conference on*
488 *Establishment Surveys (ICES-V)* (2016).

FIGURE LEGENDS

Fig. 1| Overview of the study. (A) (i) Overlapping subjects are observed between AD genetic initiatives. (ii) Without a direct comparison, there is no overlapping subject across ethnicities. Until now, trans-ethnic applications of PRS have been limited. We suspect that overlapping subjects within an ethnicity are one of the key factors to explain overestimated performances, which motivates this study. We divide PRS into two cases, where rPRS represents when the genetic information is provided and used as the discovery set and sPRS stands for the case when GWAS is pre-conducted and only summary statistics are provided. (B) For rPRS, overlapping subjects (n=432) between ADSP and AMP-AD are identified, which breaks the independence assumption and causes the overestimation bias. For sPRS, the overlapping ratio cannot be examined given the summary statistics. However, the suspected inflation in the AD prediction performance (denoted by sPRS - rPRS) motivates further analysis into the scale effect of the datasets because IGAP has a larger number of samples. (C) (i) Two new variables, hypertension and height, from the UK Biobank database are introduced to compute the upper bounds of the scale effect. Hypertension and height have a higher heritability than AD. Thus, they act as the upper bounds for AD over PRS performances (shown in the QQ plot). (ii) In AD, the gap between sPRS and rPRS (area shaded in green) is attributable to either the overestimation bias or the scale effect of the sample size of the discovery set. Because UK Biobank consists of a larger number of samples (n=342318), the scale effect can be measured via computing the performance gains per sample unit.

Fig. 2| PRS performance comparisons for Alzheimer's disease. ΔAUC and ΔR^2 denote the additive gain from introducing PRS term to Model II (refer to Materials and Methods for details). For convenience, we abbreviate the discovery and replication sets as D and R, respectively. (A) AD prediction performances with and without subject overlap (D: ADSP, R: AMP-AD). All metrics of overlapping subjects are overestimated, growing in an increasing number of SNPs. (B) sPRS (D: IGAP, R: ADSP) is compared to rPRS (D: ADSP, R: ADSP). (C) AMP-AD data are another R for rPRS (D: ADSP) and sPRS (D: IGAP). D and R of ADSP data are derived from tenfold cross-validation. In both (B) and (C), sPRS performances are significantly higher than rPRS, and we suspect that some participants of IGAP are identical to a subset of ADSP or AMP-AD. (D) A simulated study is conducted with rPRS (D: ADSP, R: AMP-AD), in which a growing number of subjects in R are replaced by a subset of D (see Results for details). As more subjects overlap between datasets, rPRS performance increases monotonically. Meanwhile, sPRS (D: IGAP, R: the mixture of ADSP and AMP-AD) results remain relatively consistent. Judging by the intersection point of rPRS and sPRS, at least 10% of AMP-AD participants are suspected to overlap with the IGAP study. When only APOE $\epsilon 4$ status is included for developing rPRS in the same manner, performances are relatively stable compared to rPRS, which indicates that the characters of AD are maintained irrespective of the AMP-AD and ADSP combination.

Fig. 3| PRS performance comparisons via UK Biobank. In this study, UK Biobank's primary purpose is to evaluate the scale effect, defined as the marginal gain of performance due to the size of the discovery set. To this end, two variables representative for high heritability, namely hypertension, and height, are analyzed. For experimental purposes, we intentionally design three discovery sets with different sizes, 300k, 60k, and 9k, which approximately correspond to the discovery set sizes of the full UK Biobank dataset, IGAP, and ADSP, respectively. For convenience, we abbreviate the discovery and replication sets as D and R. ΔAUC and ΔR^2 denote the additive gain from introducing the PRS term to Model II (refer to Materials and Methods for details). (A) A larger D size results in higher prediction performances (ΔAUC and ΔR^2), demonstrating the scale effect as hypothesized. Yet, in the three sample sizes, a smaller subset of R rarely degrades ΔAUC or ΔR^2 , but it had an impact on the significance level P, perhaps intuitively. As the highest heritability (Fig.1C) foretells, the height variable applied in PRS showed a greater impact on the prediction model than hypertension as indicated by higher ΔR^2 and $-\log_{10}(P)$. (B) When the number of SNPs varies with 100% of R used, most metrics show improvements until 50k SNPs are used, after which it plateaus. (C) Although the size of D shows a linear correlation with PRS performances, proving the hypothesized scale effect, the improvements are not dramatic. For instance, ΔR^2 increases by approximately 0.0000125 and 0.0000083 per 3k of D.