BMC Genomic Data

RESEARCH Open Access

Overestimated prediction using polygenic prediction derived from summary statistics

David Keetae Park[1†], Mingshen Chen[2†], Seungsoo Kim[3], Yoonjung Yoonie Joo[4], Rebekah K. Loving[6], Hyoung Seop Kim[7], Jiook Cha[5], Shinjae Yoo[8*] and Jong Hun Kim[9*]

Abstract

Background When polygenic risk score (PRS) is derived from summary statistics, independence between discovery and test sets cannot be monitored. We compared two types of PRS studies derived from raw genetic data (denoted as rPRS) and the summary statistics for IGAP (sPRS).

Results Two variables with the high heritability in UK Biobank, hypertension, and height, are used to derive an exemplary scale effect of PRS. sPRS without *APOE* is derived from International Genomics of Alzheimer's Project (IGAP), which records ΔAUC and ΔR^2 of 0.051 ± 0.013 and 0.063 ± 0.015 for Alzheimer's Disease Sequencing Project (ADSP) and 0.060 and 0.086 for Accelerating Medicine Partnership - Alzheimer's Disease (AMP-AD). On UK Biobank, rPRS performances for hypertension assuming a similar size of discovery and test sets are 0.0036 ± 0.0027 (ΔAUC) and 0.0032 ± 0.0028 (ΔR^2). For height, ΔR^2 is 0.029 ± 0.0037.

Conclusion Considering the high heritability of hypertension and height of UK Biobank and sample size of UK Biobank, sPRS results from AD databases are inflated. Independence between discovery and test sets is a well-known basic requirement for PRS studies. However, a lot of PRS studies cannot follow such requirements because of impossible direct comparisons when using summary statistics. Thus, for sPRS, potential duplications should be carefully considered within the same ethnic group.

Keywords Polygenic risk score, Complex genetic disease, Alzheimer's disease, Overestimation bias

*Correspondence:
Shinjae Yoo
sjyoo@bnl.gov
Jong Hun Kim
jh7521@naver.com
[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]Samsung Advanced Institute for Health Sciences & Technology (SAHIST), Sungkyunkwan University, Samsung Medical Center, 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 Lab. Computer Science and Math, Building 725, Room 2-189, Upton, NY 11973, USA
[9]Department of Neurology, Dementia Center, National Health Insurance Service Ilsan Hospital, 100 Ilsan-ro Ilsandong-gu, Goyang, Gyeonggi-Do 10444, South Korea

Background

Recently, genetic studies involving a polygenic risk score (PRS) have dramatically grown, and sophisticated tools and methodologies are being developed for its use [1–3]. Along with heritability, PRS has become an important metric for explaining complex genetic diseases (e.g. Alzheimer's disease, AD) and traits [4–8]. A typical PRS study involves both the discovery and test 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 are also 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 test set [10], which are well-known basic prerequisite for PRS studies. However, our preliminary analyses (Fig. 1A(i)) 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 test sets is clearly stated in the paper, the PRS performance of binary traits not statistically significant, while if not, results were highly variable or inflated [12–16]. 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 accuracy [17]. As a similar line of evidence, in a large-scale finnish study [12], 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 [18]. We suspect that the strictly preserved independence of subjects between different ethnicities (Fig. 1A(ii)) 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 is one of the reasons for the gap between within- and trans-ethnic generalization capacities of PRS models.

The clues mentioned above, 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 test 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

Figure 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 with 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: test) splits for ten-fold cross-validation for rPRS (Fig. 2B), while AMP-AD data are used as another test set for rPRS (Fig. 2C). An increasing degree of overlapping bias is observed with an expanding number of subjects in the test set being replaced by samples from the corresponding discovery set (Fig. 2D). We further demonstrate that sPRS also overestimates prediction performances (Figs. 1B and 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, we leverage a large number of samples in UK Biobank in terms of two phenotypes with higher heritabilities than AD [19–21], namely hypertension, and height, which 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 (Figs. 1C and 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

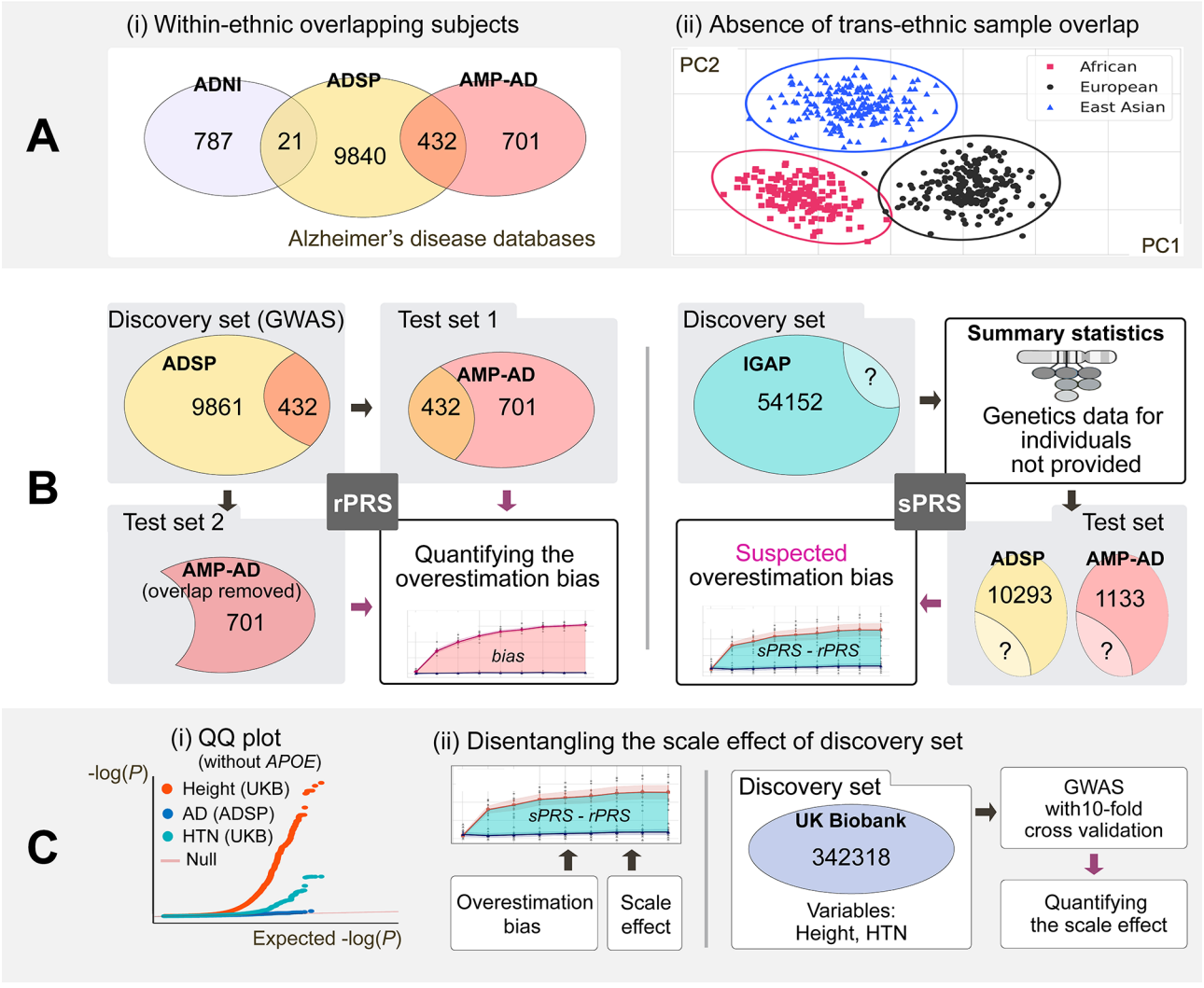

Fig. 1 Overview of the study. (**A**) (i) Overlapping subjects are observed between AD genetic initiatives. (ii) There is no overlapping subject across ethnicities. Until now, trans-ethnic applications of PRS have been limited. We suspect that subject overlap within an ethnicity is one of the key factors 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 by giving the summary statistics. However, the suspected inflation in the AD prediction performance (denoted by sPRS - rPRS) motivates further analysis of 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 = 342,318), the scale effect can be measured via computing the performance gains per sample unit. Cohort case counts and their percentages of the total were as follows: ADSP had 5687 (55.2%), AMP-AD had 696 (61.4%), IGAP had 17,008 (31.4%), and UK Biobank had 82,719 (24.2%)

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 (Supplementary Table 2), and ΔAUC still shows no statistical significance.

rPRS and sPRS Performances on AD prediction

Figure 2B and C, and Supplementary Table 3 show the comparison results of rPRS and sPRS. ADSP data are divided into the discovery and test datasets with 9:1 cross-validation for assessment of rPRS (Fig. 2B). Another test of rPRS is evaluated on non-overlapping

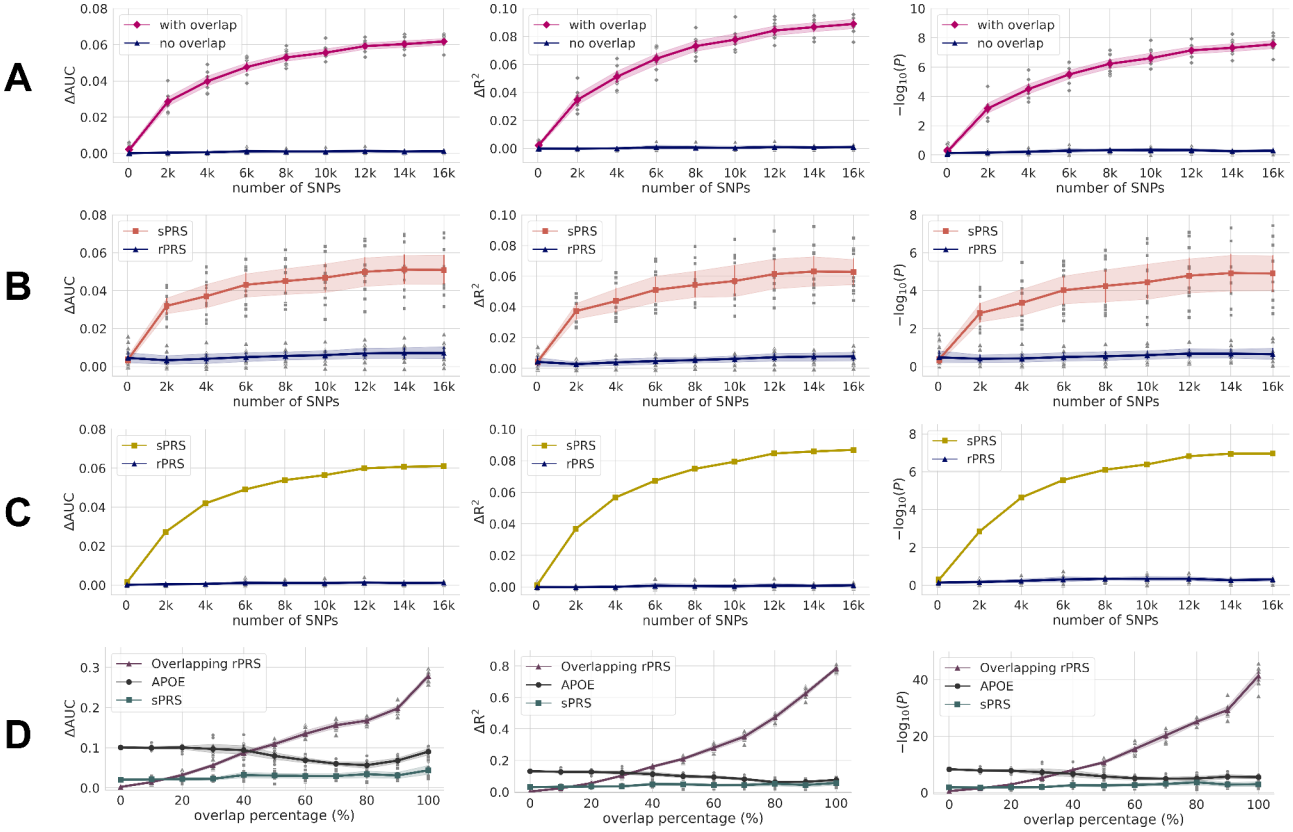

Fig. 2 PRS performance comparisons for Alzheimer's disease. ΔAUC and ΔR^2 denote the additive gain from introducing PRS term to Model II (refer to Materials and Methods for details). For convenience, we abbreviate the discovery and test sets as **D** and **T**, respectively. (**A**) AD prediction performances with and without subject overlap (**D**: ADSP, **T**: AMP-AD). All metrics of overlapping subjects are overestimated, growing in an increasing number of SNPs. (**B**) sPRS (**D**: IGAP, **T**: ADSP) is compared to rPRS (**D**: ADSP, **T**: ADSP). (**C**) AMP-AD data is another **T** for rPRS (**D**: ADSP) and sPRS (**D**: IGAP). **D** and **T** 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, **T**: AMP-AD), in which a subset of **D** replaces a growing number of subjects in **T** (see Results for details). The number of SNPs in the x-axis denotes number of the LD pruned SNPs selected in the order from the lowest P-value thresholds. That is, the lower number of SNP in the left side means the stricter P value threshold and the right-most side is the most generous P value threshold ($P < 0.5$)

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 C). sPRS on ADSP is evaluated on 10 test folds (i.e., sets), while, on AMP-AD, the whole data (n=1133) are regarded as a test 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±0.0052, 0.0077±0.0045) and (0.0013±0.00091, 0.0011±0.0018). *P*-values are not significant ($P>0.05$). In other words, when data independence 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±0.013, 0.063±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 test set. All ADSP data (n=10,293) are used to derive an rPRS model, which subsequently is evaluated on a mixture of AMP-AD (independent test set, n=692) and ADSP data randomly selected from test splits of ten-fold cross-validation (fully dependent test 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 increases 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 C).

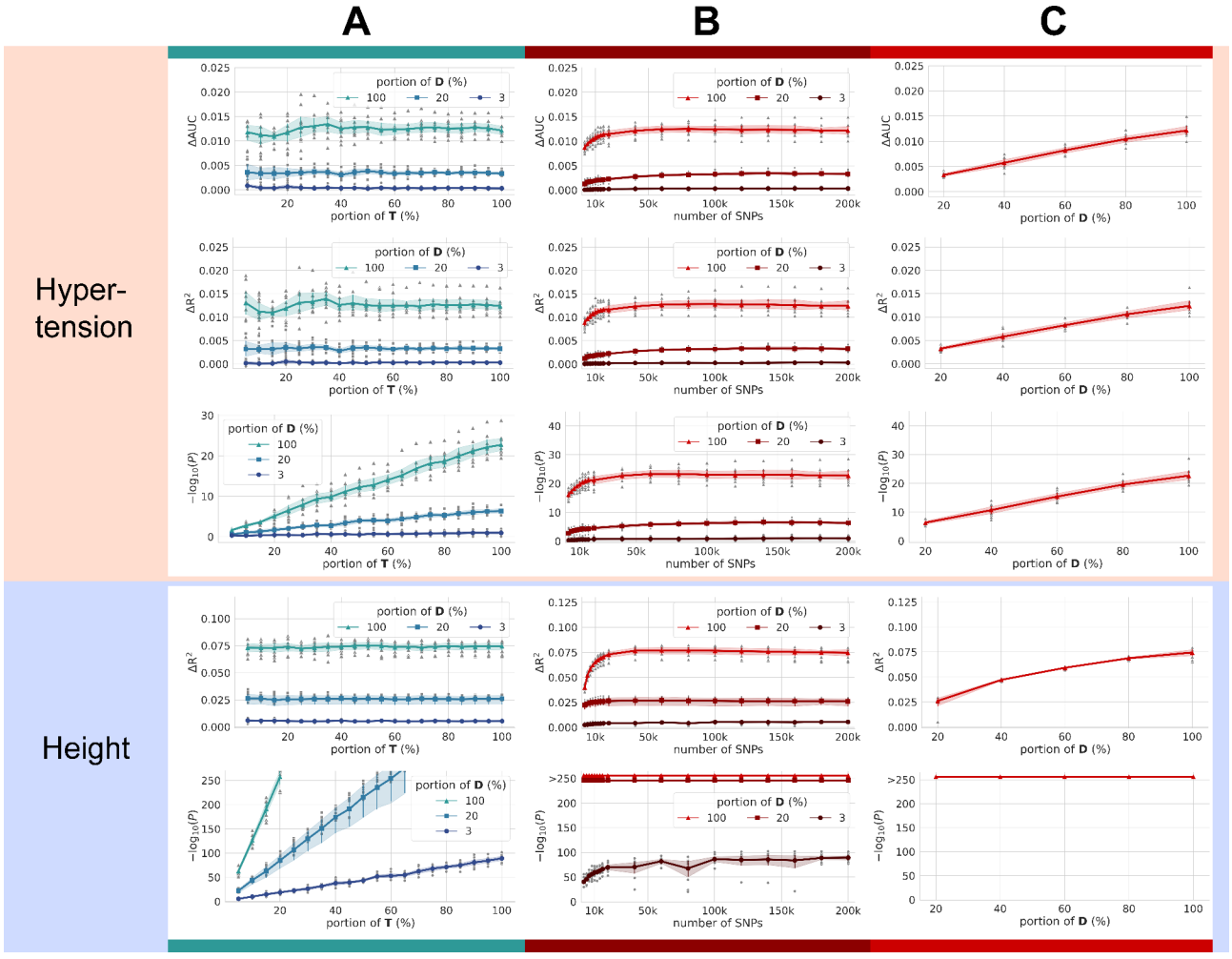

Fig. 3 PRS performance comparisons via UK Biobank. In this study, UK Biobank's primary purpose is to evaluate the scale effect, defined as the marginal gain of performance due to the size of the discovery set. To this end, two variables representative for high heritability, namely hypertension, and height, 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 test sets as **D** and **T**. 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. However, in the three sample sizes, a smaller subset of **T** 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(*P*). (**B**) When the number of SNPs varies with 100% of **T** used, most metrics show improvements until 50k SNPs are used, which plateaus. The number of SNPs in the x-axis denotes number of the LD pruned SNPs selected in the order from the lowest P-value thresholds. That is, the lower number of SNP in the left side means the stricter P value threshold and the right-most side is the most generous P value threshold ($P < 0.5$). (C) Although the size of **D** with 100% of **T** used 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**

To show that the AD characteristics of the test sets are maintained, performances of *APOE* ε4 are also displayed. When only *APOE* ε4 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. 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, we evaluate rPRS 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). Figure 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 C) 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, both the discovery and test sets require a substantially large number of subjects (Fig. S1). For instance, for a 60k-sized discovery set, more than 10k subjects are needed in the test set for sufficient power ($P<0.01$). Lassosum [2], 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.

The results can vary based on the case:control ratio of the discovery set. The IGAP study consisted of 17,008 cases and 37,154 controls, yielding a case:control ratio of 1:2.18, while the UK Biobank study had a case:control ratio of 1:3.14 (Table S1). Thus, to demonstrate that the lower PRS performances aren't a result of the case:control ratio, we kept the number of cases constant at 17,008, and varied the case:control ratio to 0.33, 1, and 3 (Fig. S2). The results from a case:control ratio of 1:3 are almost comparable to those observed with 20% of the discovery set, as depicted in Fig. 3. Furthermore, we examined the results according to both the MAF of the SNPs and the number of SNPs. The outcomes did not significantly differ across the range of the SNP's MAF interval (Fig. S2A). As for the results according to the number of SNPs, these were saturated from the outset and showed minimal variation (Fig. S2B and Fig. 3B). When the number of cases is held constant, the highest performance was observed at a case:control ratio of 1:3 (Fig. S2B). Therefore, if the number of cases in the discovery set remains fixed, the larger the total subject count, the higher the performance of the PRS.

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 needs to be suspected when the contribution of sPRS derived from discovery datasets with much smaller sample sizes than nation-wide biobank, by the referenced PRS methods [22], exceeds 1–2% of ΔR^2 for binary traits [17] 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, corresponding to 38.12% of AMP-AD data. Also, ADSP and ADNI have 21 overlapping subjects (Fig. 1A(i)). As genetic studies are conducted across multiple centers, the odds of having duplicated subjects are high among different initiatives. Therefore, the 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 performs significantly lower when data independence is explicitly controlled. For instance, when overlapping subjects are removed, PRS contributed less than 2% 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 clarified [12]. The PRS contributions of the 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 the finnish population and 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 performance

[19]. For example, hundreds of thousands of subjects in the discovery set would be required for PRS to be used in disease prediction [19]. 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 test set of 10k is required in the 60k discovery set for statistically significant AUC increment by PRS. In the studies registered in the PGS catalog, the median number of subjects in the test sets is 6995 and 24,573 in the discovery sets [23]. In addition, a systematic review of AD PRS studies reveals that the test set size ranges from 59 to 116,666 [24]. About one-third of them, the sample size is less than a thousand. Therefore, an abundance of prior sPRS studies [24–26] 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 test sets are from different countries or ethnicities [12–16]. However, a line of evidence suggests that trans-ethnic portability remains in many traits [27–29], even if the inherent differences in LD structures across races prevent causal variants from being correctly reflected in PRS [30]. For instance, Martin et al. reported equivalent performances within the confidence interval for intra- and trans-ethnic test sets in four out of five binary traits and 11 out of 17 non-binary traits of BioBank Japan [27]. Also, sPRS developed with Europeans showed a 50% discounted performance for East Asians and 25% for Africans [27]. The gap (or the variability) between intra-ethnic and trans-ethnic evaluations of PRS can be falsely increased by overestimation in a specific ethinic group study. The low trans-ethnic portability of PRS can be understood and overcomed only after excluding overestimation bias. That is, 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 ($P<5\times10^{-8}$) SNPs in core genes curated from multiple GWA studies [31–33], 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 regions [34], and PRS performances increase with a higher number of SNPs then plateau (Figs. 2 and 3) [35]. Therefore, PRS models from GWA SNPs may overfit to the discovery set. Hence, 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 study [27]. Heritability for height and hypertension was reported as 49.7% and 14.7%, respectively [21]. 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 [20, 21, 31]. Second, the number of SNPs used for UK Biobank is 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 of PRS studies not only for AD prediction but also for those of most binary complex genetic traits/diseases using subjects of less than national biobank scale.

Conclusion

As the risk of overestimation bias is evaluated in sPRS studies, care must be taken to prevent overlap, especially within the same ethnicity. Direct methods of calculating sample overlap are not always feasible, so indirect methods using summary statistics can be applied [36]. While applications of genetic studies continue to gain momentum and many countries create large-scale biobanks, 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 research [23]. When using summary statistics direct comparisons between a test set and large-scale data are difficult. 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 PRS [37–43].

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 AMP-AD (Supplementary Table 1). In the UK Biobank database [44], 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) [45]. 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

$$\mathrm{ModelI}: \mathrm{y} = \beta \mathrm{x}_{\mathrm{PRS}}$$

$$\mathrm{ModelII}: \mathrm{y} = \mathrm{x}_{\mathrm{co}}^{\mathrm{T}} \mathrm{a}_{\mathrm{co}} + \mathrm{x}_{\mathrm{PC}}^{\mathrm{T}} \mathrm{b}_{\mathrm{PC}}$$

$$\mathrm{ModelIII}: \mathrm{y} = \mathrm{x}_{\mathrm{co}}^{\mathrm{T}} \mathrm{a}_{\mathrm{co}} + \mathrm{x}_{\mathrm{PC}}^{\mathrm{T}} \mathrm{b}_{\mathrm{PC}} + \beta \mathrm{x}_{\mathrm{PRS}}$$

$$\mathrm{x}_{\mathrm{co}} = \begin{pmatrix} 1 \\ \mathrm{x}_{\mathrm{co},1} \\ ? \\ \mathrm{x}_{\mathrm{co},\mathrm{n}_1} \end{pmatrix}, \mathrm{a}_{\mathrm{co}} = \begin{pmatrix} \mathrm{a}_0 \\ \mathrm{a}_1 \\ ? \\ \mathrm{a}_{\mathrm{n}_1} \end{pmatrix}, \mathrm{x}_{\mathrm{PC}} = \begin{pmatrix} \mathrm{x}_{\mathrm{PC},1} \\ \mathrm{x}_{\mathrm{PC},2} \\ ? \\ \mathrm{x}_{\mathrm{PC},\mathrm{n}_2} \end{pmatrix}, \mathrm{b}_{\mathrm{PC}} = \begin{pmatrix} \mathrm{b}_1 \\ \mathrm{b}_2 \\ ? \\ \mathrm{b}_{\mathrm{n}_2} \end{pmatrix}$$

,Where $\mathrm{x}_{\mathrm{co}}, \mathrm{a}_{\mathrm{co}} \in \mathbb{R}^{\mathrm{n}_1+1}$ are vectors of general covariates (e.g., age and sex) and corresponding coefficients, respectively, while $\mathrm{x}_{\mathrm{PC}}, \mathrm{b}_{\mathrm{PC}} \in \mathbb{R}^{\mathrm{n}_2}$ are vectors of PCs and PC coefficients, respectively. Here, $\mathrm{x}_{\mathrm{PRS}}$ 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* ε4 status, and the sequencing centers. PCs are computed using the principal component analysis function of PLINK (v1.9) [45]. 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 fastPCA [46]. For hypertension of UK Biobank, body mass index is additionally 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 [47, 48]. Performance improvements from PRS are determined by subtracting AUC and R^2 of Model II from those of Model III, which we label as ΔAUC and ΔR^2. The statistical significance of ΔAUC is examined using DeLong's test [49]. 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 test splits based on each covariate in the statistical analyses using "StratifiedKFold" function from Python's (v3.8) "scikit-learn" (v0.24.1) package [50]. When testing a part of the cross-validated datasets, samples are balanced over covariates using the R (v4.0.3) "sampling" (v2.9) package [51].

Computation of PRS

For computing PRS, we select common (MAF ≥ 1%) SNPs and use summary statistics from discovery sets, followed by measuring PRS in test datasets. For rPRS, we calculated summary statistics by logistic regression. For sPRS, we downloaded summary statistics from IGAP web page (https://www.niagads.org/datasets/ng00036). After selecting SNPs with $P<0.5$ in the association tests using the discovery dataset, we perform clumping with the window of ±1Mbp and $r^2<0.1$. Clumping is performed using PLINK (v1.9) [45]. For AD genetic studies, we exclude any SNPs within 1Mbp 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, 22].

Abbreviations

AD	Alzheimer's disease
ADSP	AD Sequencing Project
AUC	Area under the curve
IGAP	International Genomics of Alzheimer's Project
LD	Linkage disequilibrium
PC	Principal component
PRS	Polygenic risk score

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12863-023-01151-4.

Additional file 1: Table S1. Demographic characteristics

Additional file 2: Table S2. PRS performance after excluding genetically close individuals from the test set

Additional file 3: Table S3. rPRS and sPRS results on AD

Additional file 4: Table S4. PRS performance comparisons for hypertension in UK Biobank

Additional file 5: Table S5. PRS performance comparisons for height in UK Biobank

Additional file 6: Table S6. Performance comparisons between PRS and Lassosum

Additional file 7: Fig. S1. The number of test set subjects required to gain statistical significance (P < 0.01) for hypertension using UK Biobank

Additional file 8: Fig. S2. Comparisons of PRS performance across different case:control ratios of discovery sets using hypertension phenotype of UK biobank

Additional file 9: Supporting material: material, methods, and additional references (docx)

Supplementary Material 10

Acknowledgements
Quality control filtering of the UK Biobank data was conducted by R. Mitchell, G. Hemani, T. Dudding, L. Corbin, S. Harrison, and L. Paternoster as described in the published protocol (doi: https://doi.org/10.5523/bris.1ovaau5sxunp2cv8rcy88688v). The authors also thank the International Genomics of Alzheimer's Project (IGAP) for providing summary results data for these analyses. The results published here are in part based on data obtained from the AD Knowledge Portal (https://adknowledgeportal.org).

Authors' 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.

Funding
This work was supported by the U.S. Department of Energy (DOE), Office of Science (SC), Advanced Scientific Computing Research program under award DE-SC-0012704 and used resources of the National Energy Research Scientific Computing Center, a DOE Office of Science User Facility supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231 using NERSC award ASCR-ERCAP0023081. Additionally, the computational resources for this study were supported by the internal funding of Ilsan Hospital.

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; https://doi.org/10.1038/sdata.2016.89), Mount Sinai Brain Bank (MSBB) study (accession syn3159438; https://doi.org/10.1038/sdata.2018.185), and Religious Orders Study and Memory and Aging Project (ROSMAP) Study (accession syn3159438; https://doi.org/10.1038/mp.2017.20). The UK Biobank Access Team should be contracted for UK Biobank data. Researchers can contact the UK Biobank Access Team by email at access@ukbiobank.ac.uk or through the UK Biobank website's Contact Us page.

Declarations

Ethics approval and consent to participate
All methods were carried out in accordance with relevant guidelines and regulations. The data used in this study were anonymized before its use. This study was approved by Ilsan hospital's institutional review board. The North West Multi-centre Research Ethics Committee has granted United Kingdom (UK) Biobank the Research Tissue Bank approval, which allows researchers to conduct their work under this approval without the need for separate ethical clearance [52]. J.H.K. granted administrative permission for data of Alzheimer's Disease Sequencing Project (ADSP) and Accelerating Medicine Partnership - Alzheimer's Disease (AMP-AD). J.H.K. and J.Y.C were approved for data of UK Biobank. This research has been conducted using the UK Biobank Resource under Application Number 32575. The written informed patient consent was received in ADSP, AMP-AD, UK) Biobank.

Consent for publication
Not applicable.

Competing interests
The authors declare that they have no competing interests.

Received: 12 February 2023 / Accepted: 16 August 2023
Published online: 14 September 2023

References

1. Euesden J, Lewis CM, O'Reilly PF. PRSice: polygenic risk score software. Bioinformatics. 2015;31(9):1466–8.
2. Mak TSH, Porsch RM, Choi SW, Zhou X, Sham PC. Polygenic scores via penalized regression on summary statistics. Genet Epidemiol. 2017;41(6):469–80.
3. Prive F, Arbel J, Vilhjalmsson BJ. LDpred2: better, faster, stronger. Bioinformatics. 2020;36(22–23):5424–31.
4. Yang J, Lee SH, Goddard ME, Visscher PM. GCTA: a tool for genome-wide complex trait analysis. Am J Hum Genet. 2011;88(1):76–82.
5. International Schizophrenia C, Purcell SM, Wray NR, Stone JL, Visscher PM, O'Donovan MC, Sullivan PF, Sklar P. Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. Nature. 2009;460(7256):748–52.
6. Khera AV, Chaffin M, Aragam KG, Haas ME, Roselli C, Choi SH, Natarajan P, Lander ES, Lubitz SA, Ellinor PT, et al. Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations. Nat Genet. 2018;50(9):1219–24.
7. Escott-Price V, Sims R, Bannister C, Harold D, Vronskaya M, Majounie E, Badarinarayan N, Perades G, IGAP consortia, Morgan K, Passmore P. Common polygenic variation enhances risk prediction for Alzheimer's disease. *Brain* 2015, 138(Pt 12):3673–3684.
8. Sims R, Hill M, Williams J. The multiplex model of the genetics of Alzheimer's disease. Nat Neurosci. 2020;23(3):311–22.
9. Choi SW, Mak TS, O'Reilly PF. Tutorial: a guide to performing polygenic risk score analyses. Nat Protoc. 2020;15(9):2759–72.
10. Wand H, Lambert SA, Tamburro C, Iacocca MA, O'Sullivan JW, Sillari C, Kullo IJ, Rowley R, Dron JS, Brockman D, et al. Improving reporting standards for polygenic scores in risk prediction studies. Nature. 2021;591(7849):211–9.
11. Tzoulaki I, Liberopoulos G, Ioannidis JP. Assessment of claims of improved prediction beyond the Framingham risk score. JAMA. 2009;302(21):2345–52.
12. Mars N, Koskela JT, Ripatti P, Kiiskinen TTJ, Havulinna AS, Lindbohm JV, Ahola-Olli A, Kurki M, Karjalainen J, Palta P, et al. Polygenic and clinical risk scores and their impact on age at onset and prediction of cardiometabolic diseases and common cancers. Nat Med. 2020;26(4):549–57.
13. Dikilitas O, Schaid DJ, Kosel ML, Carroll RJ, Chute CG, Denny JA, Fedotov A, Feng Q, Hakonarson H, Jarvik GP, et al. Predictive utility of polygenic risk scores for Coronary Heart Disease in three major racial and ethnic groups. Am J Hum Genet. 2020;106(5):707–16.
14. Duncan L, Shen H, Gelaye B, Meijsen J, Ressler K, Feldman M, Peterson R, Domingue B. Analysis of polygenic risk score usage and performance in diverse human populations. Nat Commun. 2019;10(1):3328.
15. Dube JB, Johansen CT, Robinson JF, Lindsay J, Hachinski V, Hegele RA. Genetic determinants of "cognitive impairment, no dementia. J Alzheimers Dis. 2013;33(3):831–40.
16. Marden JR, Walter S, Tchetgen Tchetgen EJ, Kawachi I, Glymour MM. Validation of a polygenic risk score for dementia in black and white individuals. Brain Behav. 2014;4(5):687–97.
17. Elliott J, Bodinier B, Bond TA, Chadeau-Hyam M, Evangelou E, Moons KGM, Dehghan A, Muller DC, Elliott P, Tzoulaki I. Predictive accuracy of a polygenic risk score-enhanced prediction model vs a clinical risk score for coronary artery disease. JAMA. 2020;323(7):636–45.
18. Bitarello BD, Mathieson I. Polygenic scores for height in Admixed populations. G3 (Bethesda). 2020;10(11):4027–36.
19. Dudbridge F. Power and predictive accuracy of polygenic risk scores. PLoS Genet. 2013;9(3):e1003348.
20. Holland D, Frei O, Desikan R, Fan CC, Shadrin AA, Smeland OB, Sundar VS, Thompson P, Andreassen OA, Dale AM. Beyond SNP heritability: polygenicity and discoverability of phenotypes estimated with a univariate gaussian mixture model. PLoS Genet. 2020;16(5):e1008612.

21. Canela-Xandri O, Rawlik K, Tenesa A. An atlas of genetic associations in UK Biobank. Nat Genet. 2018;50(11):1593–9.
22. Choi SW, O'Reilly PF. PRSice-2: polygenic risk score software for biobank-scale data. Gigascience 2019, 8(7).
23. Lambert SA, Gil L, Jupp S, Ritchie SC, Xu Y, Buniello A, McMahon A, Abraham G, Chapman M, Parkinson H, et al. The polygenic score catalog as an open database for reproducibility and systematic evaluation. Nat Genet. 2021;53(4):420–5.
24. Harrison JR, Mistry S, Muskett N, Escott-Price V. From polygenic scores to Precision Medicine in Alzheimer's Disease: a systematic review. J Alzheimers Dis. 2020;74(4):1271–83.
25. Oram RA, Patel K, Hill A, Shields B, McDonald TJ, Jones A, Hattersley AT, Weedon MN. A type 1 diabetes genetic risk score can Aid discrimination between type 1 and type 2 diabetes in young adults. Diabetes Care. 2016;39(3):337–44.
26. Harrison TM, Mahmood Z, Lau EP, Karacozoff AM, Burggren AC, Small GW, Bookheimer SY. An Alzheimer's Disease Genetic Risk Score Predicts Longitudinal Thinning of Hippocampal Complex Subregions in Healthy Older Adults. *eNeuro* 2016, 3(3).
27. Martin AR, Kanai M, Kamatani Y, Okada Y, Neale BM, Daly MJ. Clinical use of current polygenic risk scores may exacerbate health disparities. Nat Genet. 2019;51(4):584–91.
28. Domingue BW, Belsky DW, Harris KM, Smolen A, McQueen MB, Boardman JD. Polygenic risk predicts obesity in both white and black young adults. PLoS ONE. 2014;9(7):e101596.
29. Koyama S, Ito K, Terao C, Akiyama M, Horikoshi M, Momozawa Y, Matsunaga H, Ieki H, Ozaki K, Onouchi Y, et al. Population-specific and trans-ancestry genome-wide analyses identify distinct and shared genetic risk loci for coronary artery disease. Nat Genet. 2020;52(11):1169–77.
30. Amariuta T, Ishigaki K, Sugishita H, Ohta T, Koido M, Dey KK, Matsuda K, Murakami Y, Price AL, Kawakami E, et al. Improving the trans-ancestry portability of polygenic risk scores by prioritizing variants in predicted cell-type-specific regulatory elements. Nat Genet. 2020;52(12):1346–54.
31. Graff RE, Cavazos TB, Thai KK, Kachuri L, Rashkin SR, Hoffman JD, Alexeeff SE, Blatchins M, Meyers TJ, Leong L, et al. Cross-cancer evaluation of polygenic risk scores for 16 cancer types in two large cohorts. Nat Commun. 2021;12(1):970.
32. Belsky DW, Moffitt TE, Sugden K, Williams B, Houts R, McCarthy J, Caspi A. Development and evaluation of a genetic risk score for obesity. Biodemography Soc Biol. 2013;59(1):85–100.
33. Mavaddat N, Pharoah PD, Michailidou K, Tyrer J, Brook MN, Bolla MK, Wang Q, Dennis J, Dunning AM, Shah M et al. Prediction of breast cancer risk based on profiling with common genetic variants. J Natl Cancer Inst 2015, 107(5).
34. Freedman ML, Monteiro AN, Gayther SA, Coetzee GA, Risch A, Plass C, Casey G, De Biasi M, Carlson C, Duggan D, et al. Principles for the post-GWAS functional characterization of cancer risk loci. Nat Genet. 2011;43(6):513–8.
35. Ware EB, Schmitz LL, Faul J, Gard A, Mitchell C, Smith JA, Zhao W, Weir D, Kardia SL. Heterogeneity in polygenic scores for common human traits. bioRxiv 2017:106062.
36. Choi SW, Mak TSH, Hoggart CJ, O'Reilly PF. EraSOR: a software tool to eliminate inflation caused by sample overlap in polygenic score analyses. Gigascience 2022, 12.
37. Lambert JC, Ibrahim-Verbaas CA, Harold D, Naj AC, Sims R, Bellenguez C, DeStafano AL, Bis JC, Beecham GW, Grenier-Boley B, et al. Meta-analysis of 74,046 individuals identifies 11 new susceptibility loci for Alzheimer's disease. Nat Genet. 2013;45(12):1452–8.
38. Beecham GW, Bis JC, Martin ER, Choi SH, DeStefano AL, van Duijn CM, Fornage M, Gabriel SB, Koboldt DC, Larson DE, et al. The Alzheimer's disease sequencing project: study design and sample selection. Neurol Genet. 2017;3(5):e194.
39. Greenwood AK, Montgomery KS, Kauer N, Woo KH, Leanza ZJ, Poehlman WL, Gockley J, Sieberts SK, Bradic L, Logsdon BA, et al. The AD knowledge Portal: a repository for Multi-Omic Data on Alzheimer's Disease and Aging. Curr Protoc Hum Genet. 2020;108(1):e105.
40. Crane PK, Foroud T, Montine TJ, Larson EB. 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. 2017;13(12):1410–3.
41. Allen M, Carrasquillo MM, Funk C, Heavner BD, Zou F, Younkin CS, Burgess JD, Chai HS, Crook J, Eddy JA, et al. Human whole genome genotype and transcriptome data for Alzheimer's and other neurodegenerative diseases. Sci Data. 2016;3:160089.
42. Wang M, Beckmann ND, Roussos P, Wang E, Zhou X, Wang Q, Ming C, Neff R, Ma W, Fullard JF, et al. The Mount Sinai cohort of large-scale genomic, transcriptomic and proteomic data in Alzheimer's disease. Sci Data. 2018;5:180185.
43. De Jager PL, Ma Y, McCabe C, Xu J, Vardarajan BN, Felsky D, Klein HU, White CC, Peters MA, Lodgson B, et al. A multi-omic atlas of the human frontal cortex for aging and Alzheimer's disease research. Sci Data. 2018;5:180142.
44. Bycroft C, Freeman C, Petkova D, Band G, Elliott LT, Sharp K, Motyer A, Vukcevic D, Delaneau O, O'Connell J, et al. The UK Biobank resource with deep phenotyping and genomic data. Nature. 2018;562(7726):203–9.
45. Chang CC, Chow CC, Tellier LC, Vattikuti S, Purcell SM, Lee JJ. Second-generation PLINK: rising to the challenge of larger and richer datasets. Gigascience. 2015;4:7.
46. Galinsky KJ, Bhatia G, Loh PR, Georgiev S, Mukherjee S, Patterson NJ, Price AL. Fast principal-component analysis reveals convergent evolution of ADH1B in Europe and East Asia. Am J Hum Genet. 2016;98(3):456–72.
47. Nagelkerke NJD. A note on a general definition of the coefficient of determination. Biometrika. 1991;15:691–3.
48. Robin X, Turck N, Hainard A, Tiberti N, Lisacek F, Sanchez JC, Muller M. pROC: an open-source package for R and S + to analyze and compare ROC curves. BMC Bioinformatics. 2011;12:77.
49. DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. Biometrics. 1988;44(3):837–45.
50. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Louppe G, Prettenhofer P, Weiss R, et al. Scikit-learn: machine learning in Python. J Mach Learn Res. 2011;12:2825–30.
51. Tillé Y, Matei A. The R sampling package. In: *The Fifth International Conference on Establishment Surveys (ICES-V): 2016*.
52. UK Biobank research ethics approval. [https://www.ukbiobank.ac.uk/learn-more-about-uk-biobank/about-us/ethics].

Publisher's Note