Cell Genomics Transparent Peer Review Record

Integrative polygenic risk score improves the prediction accuracy of complex traits and diseases

Buu Truong[1,2], Leland E. Hull[3,4], Yunfeng Ruan[1,2], Qin Qin Huang[5], Whitney Hornsby[1,2], Hilary Martin[5], David A. van Heel[6], Ying Wang[1,7,8], Alicia R. Martin[7,8], S. Hong Lee[9],

Pradeep Natarajan[1,2,4,*]

Summary

Initial submission:	Received : 8/14/23
	Scientific editor: Judith Nicholson
First round of review:	Number of reviewers: 2 Revision invited : 9/15/23 Revision received : 1/12/23
Second round of review:	Number of reviewers: 2 Accepted : 2/20/24
Data freely available:	Yes
Code freely available:	Yes

This transparent peer review record is not systematically proofread, type-set, or edited. Special characters, formatting, and equations may fail to render properly. Standard procedural text within the editor's letters has been deleted for the sake of brevity, but all official correspondence specific to the manuscript has been preserved.

Referees' reports, first round of review

Reviewer #1: Truong et al proposed a comprehensive approaches for integrating polygenic risk score (PRS) either trait-specific (PRS-mix) or cross-trait (PRS-mix+) for complex traits and diseases and evaluated the PRS performance in European and South Asian ancestries. Compared to the previously best-performing PRS for each trait, PRS derived from the new method significantly improved the mean prediction accuracy by up to 1.7-fold (PRS-mix+). Of note for coronary artery disease (CAD), their method remarkably improved the prediction accuracy up to 3.27-fold over the established methods for combining PRSs from correlated traits. From a clinical point of view, the article elucidated the potential of leveraging different PRSs to substantially model robustness and performance beyond pre-existing PRS for a target population.
Major Comments:
1.As we know, the majority of PRSs were derived based on GWASs in European populations and the transferability of PRSs across ancestry groups was limited. The author indicated that leveraging the PRSs either trait-specific or cross-trait significantly improved the mean prediction accuracy. Would the improvement in prediction accuracy for a target population be greater by integrating ancestry-specific PRSs?
2.Recently, PRSs have shown promise in improving risk prediction and stratification over guideline-recommended clinical risk scores for CAD. In this article, the author mainly evaluated the clinical utility of the integrating PRS with established risk factors versus the traditional model with clinical risk factors. I guess whether the PRS also improve the predictive accuracy over traditional guideline framework, such as the PCE, QRISK, could the author add relevant analyses?
3.As the author mentioned, there have been numerous emerging studies to combine PRS. The author only compared LDpred2 with the framework in the present study. Could the author compare more benchmarking combination methods to improve the robustness of their study?
4.In lines 699-702, the author described about the details of simulations. I wonder what the PRS1-PRS6 represented? Trait-specific PRS or other? Could the author add more details for them?
5.PRS-mix+ incorporates genetically correlated traits to better capture the human genetic architecture. I would guess what's the exact inclusion criteria for correlated traits? Could the author add more details?

Reviewer #2: The manuscript describes a very comprehensive study that uses publicly available resources (PGS Catalog) to train two methods: PRSmix, a single-trait multi-PRS method and PRSmix+, a multi-trait multi-PRS method. The proposed methods show significant increase in prediction accuracy in a simulations study and across multiple outcomes and cohorts.

The need for the two new methods is justified by filling-up multiple missing pieces in the polygenic risk score literature: genetic prediction across diverse populations, optimization of existing data usage and combining multiple PRS methods. This approach leverages the PGS Catalog and clearly benefits from the fact that the two discovery cohorts used All of Us and Genes &amp; Health have not been included in many published studies yet, and thus are not part of any study in the PGS Catalog. I think the strength of the manuscript is not the novelty of the methods as very similar strategies have been previously published. The authors include these other methods in the results and discussion, but do not clearly state this in the introduction. In my opinion, the novelty of the paper comes from 1) the diversity of traits analysed 2) the comparisons across cohorts and 3) the R package that makes the methods easily accessible to other researchers (but it might need a more explicit warning to not use this method on datasets that have been included in the PGS Catalog).

However, I have a few criticism that refer generally to the manuscript.

1. In all method comparisons, the authors continuously refer to the "best PGS Catalog" but do not specify which study it was each time. I think it would be very informative to researchers interested in each of the specific traits to know which PRS the authors are benchmarking their method against. For PRSmix, it is interesting to know if this best PRS Catalog is the one with largest sample size, newest study, meta-analysis or not, etc. For PRSmix+, it is even more relevant whether the "Best PGS Catalog" is from the same predicted trait or a genetically correlated one.

2. In the same line as the comment above, there is no information on how many PRS from the 2600 end up in the final model. The tables with the individual weights are available for some models (Supplementary tables 13-16) but a comment in the results sections/including it in the figures would be more transparent towards e.g. over-fitting (or including the PRS trait name as a column in the tables). As all models are based on a single split of the data into training/testing, there is also no information on the robustness of the small-weighted scores. Cross-validation of the data could help provide a sense of the robustness of their results. Additionally, highlight the PRS that were more generally useful across traits would be highly informative. I feel like there is a missed opportunity to highlight GWAS/PRS studies in the PGS Catalog that are generally useful for cross-trait prediction. It is great that their models increase prediction accuracy, but the results presented in the paper are somehow lacking digging deeper on where is that prediction accuracy coming from.

3. The authors define partial R2 as "difference of R2 between the model with PRS and covariates including age, sex, and 10 PCs versus the base model with only covariates"That is not the definition of partial R2, so its use is misleading for comparison across papers. What the authors are defying is the difference in r2 between full and covariate model. The correct definition of partial R2 divides this difference by the full model r2 (Apologies for citing Wikipedia: https://en.wikipedia.org/wiki/Coefficient_of_determination#Coefficient_of_partial_determination). Please consider rephrasing or changing to correct partial R2. The authors use the correct terminology in the context of AUC in lines 412-413.

4. I think there is some slightly misleading discussion related to the purpose of using cross-trait models or specifically their PRSmix+ model. The authors claim that the additional value of including correlated traits in the model comes from pleiotropy. I agree pleiotropy plays a role, but there are many other reasons mostly related to cohort-specific effects that could increase prediction. In the same line, the authors attribute the results of low cross-cohort transferability of R2 between the two cohorts analysed to differences in genetic ancestry (227-282). However, there are multiple other sources of variability between the two cohorts that would decrease prediction accuracy e.g. work from the Global Biobank (Wang et al. 2023).

Comments by section

Simulations

Comment 1. The simulations work is very interesting and shows the potential to improve prediction in small samples sizes 200-5000 with PRSmix. In line 200, the text says "We observed that PRSmix demonstrated a saturation of improvement from Ntraining= 10,000." but Figure 2 shows saturated improvement in all simulated scenarios from Ntraining &gt;= 5,000. Same with PRSmix+, the improvement is shown from 5,000 in Figure 2 but the text says 30,000.

Comment 2. Both models PRSmix/PRSmix+ seem to be particularly useful for traits with low heritability, for which more PRS translate to better prediction.

Comment 3. Line 695 says there are 7 traits heritability h2 equal to 0.05, 0.1, 0.2, and 0.5. Is this a typo or is there some information missing? Are there 7*4 traits? The number of simulated causal SNPs (1000, p = 0.001) is very small compared to real polygenic outcomes, expect the number of samples needed to predict more polygenic traits with the same accuracy would be much larger. Please state why you "removed individuals with PC1 and PC2 &gt; 3 standard deviation from the mean." in the simulations methods. In line 711, please state again the proportion of sample to each of the 3 data sets.

Combining trait-specific PRS improves prediction accuracy (PRSmix)

Comment 4. In Figures 3 and 4, it would be useful to include the average number of non-zero weighted PRS overall and for each category of trait. In general, it would be very informative to know how PRS were included in each PRSmix and how many of those ended up with a weight of 0 after training the elastic net. Additionally, the PRS weight distribution per trait would be useful to understand where the improvement is coming from. It is a very different scenario that 10 PRSs in a model have a large weight and then there are 200 extra PRS with neglectable weight that do not contribute to increasing prediction but are "cohort specific".

Comment 5. The caption in Figure 3 has conflicting information on how the whiskers for the bar estimates were derived.

Comment 6. In the results for PRSmix+ in real data, the authors pre-select the traits from the 2,600 list to train the elastic net models filtering on their theoretic power. There is no comment or results on how many actual PRS are retained after this filter. Comment 4 from the PRSmix part also applies to PRSmix+.

Comment 7. Line 281 states "However, using a linear combination with the matched ancestry still demonstrated a better performance than using transferred weights." These results are highly relevant to the topic of prs transferability. They should be highlighted in the discussion, as it indicates that methods that attempt to re-adjust prs weights cross-ancestry are less effective for prediction than directly getting the weights form the matched ancestry.

Prediction accuracy and predictive improvement across various types of traits

Comment 8. In the section "Prediction accuracy and predictive improvement across various types of traits" I don't find the category "other conditions" particularly interesting as it seems very heterogeneous in its content. Therefore, the finding that the prediction in that category is the highest does not seem relevant.

Comparison with previous combination methods

Comment 9. In Figure 5. It is very surprising that Elastic net has increased prediction over PRSmix (LDpred2 + PGS Catalog) for T2D. Can the authors comment further on why this could be the case?

Comment 10. In Figure 5. Elastic net vs. PRSmix does not show a significant improvement in prediction for BMI, CAD and depression. This indicates that currently there is no valuable information in the GWAS Catalog that improves just doing an elastic net model with 26 PRS. I am curious if this is the case for most of the results in this paper. Again, showing highlighting which PRS have large weights in PRSmix/PRSmix+ could help answer this question.

Comment 11. Although the authors highlight the PRSmix+ results for T2D in Figure 5, a closer look at supplementary table 6 shows that the most pleiotropic effects captured by the multi-trait model are coming from the depression PRSmix+, with a x2 prediction accuracy boost over all other methods. In terms of pleiotropy/heterogeneity, psychiatric disorders are at a different league than T2D.

Comment 12. The authors say in line 344 that "we benchmarked PRSmix and PRSmix+ against the previous methods"However, they only benchmark their methods explicitly to wMT-SBLUP. I don't expect more comparisons, but I think it is misleading to cite methods that are not being compared and that would yield potentially similar results to PRSmix+.

Clinical utility for coronary artery disease

Comment 13. I think the clinical evaluation of the models in the context of CAD is very complete and gives a good overview of the different performance metrics that other researchers could be interested in.

Comment 14. I don't agree with the authors'; conclusion from the "plateau" analysis described in

Supplementary figure 7. The simulations included only 6 PRS with genetic correlation with the outcome of 0.4-0.8. The CAD PRSmix included ~27 PRSs, based on Figure 6, but again it is unclear which ones were actually non-zero in the final model. I think it is misleading to conclude that 5,000 samples are required to train a model with max prediction accuracy, as this will enormously depend on many factors: snp-heritability, polygenicity, heterogeneity in the definition of the disease etc. And finally, the composition of the PGS Catalog. CAD is a highly studied trait as the authors argue, but I think this 5,000 samples rule will not be the reality many other traits.

Comment 15. In line 551, the authors mention “Third, we did not validate the mixing weights in an independent cohort.”. I am confused as the authors evaluated the cross-cohort performance of PRSmix/PRSmix+ with results in Supplementary Table 4. I this cross-ancestry not All of Us vs. Genes &amp; Health cohort? If not, I have misunderstood all cross-ancestry results. If yes, then the authors have validated the results in an independent cohort.

Small comments

- Line 84. “To better capture the genetic architecture of the outcome traits, we proposed PRSmix, a framework to combine PRS from the same trait with the outcome trait.” . I disagree the purpose of training on the same trait is to better capture its genetic architecture. Either to be truthful to the phenotype definition or
- Line 95. “other conditions as the prediction accuracies varied in each group”.
- Line 132-133 “We selected the most common traits from the PGS Catalog which have the highest number of PRS.”
- Line 151 “best single PRS from the training”….. more like the top weighted?
- Line 231 “who had undergone whole genome sequencing” is this relevant?
Supplementary info
- The description for ST4 does not include the performance measure and seems cut.
- Again, the description for ST5 does not include info enough to understand the table. What do the ratios represent?
- ST6 contains info from the figure that does not belong in the table “The whiskers demonstrate 95% confidence intervals of mean prediction accuracy. BMI, Body mass index; CAD, coronary artery disease; T2D, type 2 diabetes. GWAS, genome-wide association study”.

Authors’ response to the first round of review

Reviewer #1: Truong et al proposed a comprehensive approaches for integrating polygenic risk score (PRS) either trait-specific (PRS-mix) or cross-trait (PRS-mix+) for complex traits and diseases and evaluated the PRS performance in European and South Asian ancestries. Compared to the previously best-performing PRS for each trait, PRS derived from the new method significantly improved the mean prediction accuracy by up to 1.7-fold (PRSmix+). Of note for coronary artery disease (CAD), their method remarkably improved the prediction accuracy up to 3.27-fold over the established methods for combining PRSs from correlated traits. From a clinical point of view, the article elucidated the potential of leveraging different PRSs to substantially model robustness and performance beyond pre-existing PRS for a target population.

Response: We appreciate the overall enthusiasm and constructive feedback.

Major Comments:

1. As we know, the majority of PRSs were derived based on GWASs in European populations and the transferability of PRSs across ancestry groups was limited. The author indicated that leveraging the PRSs either trait-specific or cross-trait significantly improved the mean prediction accuracy. Would the improvement in prediction accuracy for a target population be greater by integrating ancestry-specific P

Response: We have now added the analyses integrating ancestry-matched PRSs to the target population. Due to the limited number of 100% South Asian ancestry PRS in PGS Catalog, we only applied this approach to European ancestry individuals within the All of Us Research Program. Overall, the results did not show that ancestry-matched combination was better than combining all scores from all ancestries available in the PGS Catalog. This is generally consistent with other PRS studies indicating that multi-ancestry data is generally superior to single ancestry data for PRS performance, at least due to larger sample sizes (Graham S et al Nature 2021).

Changes in manuscript: on line 299-305, we added: "We also compared PRSmix and PRSmix+ using all scores or only ancestry-matched scores from PGS Catalog (Supplementary Figure 5). We observed that combining using all scores improved the prediction accuracy better than combining only ancestry-matched scores to the target population. With PRSmix, using all scores improved the prediction accuracy 1.2-fold (95%CI = [1.09; 1.31]; P-value=0.0002) compared to using ancestry-matched scores alone. With PRSmix+, using all scores improved the prediction accuracy 1.12-fold (95%CI = [1.07; 1.18]; P-value=2.14x10-5) compared to using ancestrymatched scores alone (Supplementary Figure 5)."

2. Recently, PRSs have shown promise in improving risk prediction and stratification over guideline-recommended clinical risk scores for CAD. In this article, the author mainly evaluated the clinical utility of the integrating PRS with established risk factors versus the traditional model with clinical risk factors. I guess whether the PRS also improve the predictive accuracy over traditional guideline framework, such as the PCE, QRISK, could the author add relevant analyses?

Response: We added the analysis by benchmarking with QRISK3, the guideline-supported clinical risk calculator in the UK. The result shows that the combined PRS still demonstrated a better Net Reclassification Improvement.

Changes in manuscript: On line 425-433, we revised the result with addition with QRISK3 and clinical risk factors as the baseline model "In European ancestry, the CAD PRSmix+ integrative score improved the continuous net reclassification of 33% (95% CI: [22%; 44%]; P-value = 4.15 x 10-9) compared to PRSmix (30%; 95% CI: [20%; 44%]; P-value = 1.4 x 10-10) and the best PRS from the PGS Catalog (24%; 95% CI: [13%; 36%]; P-value = 2.05 x 10-5). In South Asian ancestry, the integrated score with PRSmix+ showed significant continuous net reclassification of 27% (95% CI: [16%; 39%]; P-value = 3.69 x 10-6) compared to PRSmix (23%; 95% CI: [11%; 34%]; P-value = 6.56 x 10-5) and the best PGS Catalog (11%; 95% CI: [-0.3%; 23%]; P-value = 0.05). Our results also demonstrated an improvement in net reclassification for models without clinical risk factors."

3. As the author mentioned, there have been numerous emerging studies to combine PRS. The author only compared LDpred2 with the framework in the present study. Could the author compare more benchmarking combination methods to improve the robustness of their study?

Response: To our knowledge, wMT-SBLUP would be the relevant method to combine multiple PRS, and Abraham et al Nature Communications 2019 named as metaPRS also used ElasticNet to combine PRS with a chosen set of traits. Additionally, Albiñana et al, Nature Communications 2023 would be the recent combination framework which used L1-regression and Xgboost to combine the scores. However, Albiñana et al. demonstrated that linear and non-linear combinations of PGS gave comparable prediction results. ElasticNet has been used in common metaPRS frameworks (Abraham et al Nature Communications 2019, Krapohl et al, Mol. Psychiatry 2018). In our work, we added an extended utility without pre-selecting set of correlated trait but employed ElasticNet to agnostically select secondary traits.

Changes in manuscript:

On line 537-541, we emphasized: "Krapohl et al.32–34 and Abraham et al.8 proposed to use Elastic Net to combine the scores developed from summary statistics, and correlated traits were selected with prior knowledge. However, these strategies consider scores developed from particular methods using predefined summary statistics. Our framework utilizes all PRSs available in the PGS Catalog which were optimized for their target traits."

On line 606-607, we highlighted from Albiñana et al: "there is no statistical significance difference between linear and non-linear combinations for neuropsychiatric diseases."

We also updated the reference of Albiñana et al from Biorxiv to Nature Communications 2023 (ref. 13).

4. In lines 699-702, the author described about the details of simulations. I wonder what the PRS1-PRS6 represented? Trait-specific PRS or other? Could the author add more details for them?

Response: We simulated PRS1-3 as trait-specific scores and PRS4-6 as cross-trait scores by introducing the genetic correlation between the scores. We have now better clarified this distinction.

Changes in manuscript: On line 767-770, we clarify by revising the sentence: "The genetic components were simulated as a linear combination of 6 PRSs where PRS1, PRS2, and PRS3 were considered trait-specific scores with genetic correlations equal to 0.8. PRS4, PRS5, and PRS6 were simulated as cross-trait scores with genetic correlation equal to 0.4."

5. PRS-mix+ incorporates genetically correlated traits to better capture the human genetic architecture. I would guess what's the exact inclusion criteria for correlated traits? Could the author add more details?

Response: We agnostically evaluated the variance explained by the PRS of secondary traits for the primary traits using R2. We also considered the theoretical significance and power derived using Momin's et al AJHG, 2023. Explicit genetic correlation would require a full GWAS summary statistics with effect sizes, standard error. However, the PGS Catalog only provides the adjusted SNP effect sizes from various PRS methods. Therefore, we used variance explained (R2) as the proxy for accuracy to predict the outcome trait.

Changes in manuscript: On line 693-700, we added "Summary-based methods to estimate genetic correlation between traits require full GWAS summary statistics including marginal effect sizes and standard error whereas the PGS Catalog only provides the adjusted SNP effect sizes from various PRS methods. We employed the predictive R2 to estimate variance explained of PRS for the outcome trait to select the secondary traits. PRS demonstrating theoretically significant predictive R2 and high power were selected for combination. We selected high-power scores defined as power > 0.95 with P-value <= 0.05 for the combination with Elastic Net."

Reviewer #2: The manuscript describes a very comprehensive study that uses publicly available resources (PGS Catalog) to train two methods: PRSmix, a single-trait multi-PRS method and PRSmix+, a multi-trait multi-PRS method. The proposed methods show significant increase in prediction accuracy in a simulations study and across multiple outcomes and cohorts. The need for the two new methods is justified by filling-up multiple missing pieces in the polygenic risk score literature: genetic prediction across diverse populations, optimization of existing data usage and combining multiple PRS methods. This approach leverages the PGS Catalog and clearly benefits from the fact that the two discovery cohorts used All of Us and Genes & Health have not been included in many published studies yet, and thus are not part of any study in the PGS Catalog. I think the strength of the manuscript is not the novelty of the methods as very similar strategies have been previously published. The authors include these other methods in the results and discussion, but do not clearly state this in the introduction. In my opinion, the novelty of the paper comes from 1) the diversity of traits analysed 2) the comparisons across cohorts and 3) the R package that makes the methods easily accessible to other researchers (but it might need a more explicit warning to not use this method on datasets that have been included in the PGS Catalog).

Response: We appreciate the overall enthusiasm and constructive feedback.

Changes in manuscript: On line 78-81, we added: "Additionally, recent studies have selected scores based on prior knowledge of clinical risk factors to the main traits8,14,15. However, this strategy may neglect important information from other traits. Our study leverages the diversity of traits analyzed across cohorts and PRS methodologies."

However, I have a few criticism that refer generally to the manuscript.

1. In all method comparisons, the authors continuously refer to the "best PGS Catalog" but do not specify which study it was each time. I think it would be very informative to researchers interested in each of the specific traits to know which PRS the authors are benchmarking their method against. For PRSmix, it is interesting to know if this best PRS Catalog is the one with largest sample size, newest study, meta-analysis or not, etc. For PRSmix+, it is even more relevant whether the "Best PGS Catalog" is from the same predicted trait or a genetically correlated one.

Response: We have added Supplementary Table 5 to describe best PGSs from PGS Catalog with the available information from PGS Catalog including sample sizes, release date and ancestries. The data on whether meta-analysis or not for original GWAS is not clearly provided on the PGS Catalog (some scores provided this information while most of the rest did not). To mimic common usage, our study used the best trait-specific score in the training sample as the "best PGS Catalog" and evaluated its performance on the testing sample.

Changes in manuscript:

On line 154-155, we added: "we selected the best-performing PRS from the set of traits matched with the outcome trait from the training set and evaluated by incremental R2 in the testing set."

On line 275-284, we added: "PRSmix yielded an equivalent number of non-zero mixing weights between European ancestry (median=8; interquartile range = [5; 12]) and South Asian ancestry (median=8; IQR=[3;14]). However, PRSmix+ demonstrated a higher number of non-zero weights in European ancestry (median = 55; IQR = [30;76]) compared to South Asian ancestry (median=32; IQR=[11;49]). The median absolute mixing weights were similar between European ancestry and South Asian ancestry (Supplementary Fig. 4). We note that most of the best PRS across traits were developed from 2021 onward. The details of the most recent trait-specific PRS, PRS with largest sample sizes, best PRS being compared and score with highest weights for European ancestry and South Asian ancestry are provided in Supplementary Table 5 and Supplementary Table 6, respectively."

We added Supplementary Table 5 to describe the best PRS and the scores with the highest weight in PRSmix, PRSmix+ for European and South Asian ancestry.

2. In the same line as the comment above, there is no information on how many PRS from the 2600 end up in the final model. The tables with the individual weights are available for some models (Supplementary tables 13-16) but a comment in the results sections/including it in the figures would be more transparent towards e.g. over-fitting (or including the PRS trait name as a column in the tables). As all models are based on a single split of the data into training/testing, there is also no information on the robustness of the small-weighted scores. Cross-validation of the data could help provide a sense of the robustness of their results. Additionally, highlight the PRS that were more generally useful across traits would be highly informative. I feel like there is a missed opportunity to highlight GWAS/PRS studies in the PGS Catalog that are generally useful for cross-trait prediction. It is great that their models increase prediction accuracy, but the results presented in the paper are somehow lacking digging deeper on where is that prediction accuracy coming from.

Response: In our analysis, in the training cohort, we indeed performed 5-fold cross-validation to select the best combination. We then used the cross-validated model to perform prediction in the testing cohort. On line 641, we mentioned about 5-fold cross-validation in the training set to estimate mixing weight. In our package, we also integrated cross-validation to estimate the mixing weights of the PRSs. We demonstrate the source of increased prediction accuracy for CAD and stroke in Supplementary Figure 7 and Supplementary Figure 10. The best scores in the linear combination are provided in Supplementary Table 5 and Supplementary Table 6 for European ancestry and South Asian ancestry, respectively. These tables give information about the best score for each model which helps to explain the improvement in prediction accuracy.

Changes in manuscript: We added Supplementary Table 4 to provide the number of PRSs ended up in the final model.

On line 275-284, we added: "PRSmix yielded an equivalent number of non-zero mixing weights between European ancestry (median=8; interquartile range = [5; 12]) and South Asian ancestry (median=8; IQR=[3;14]). However, PRSmix+ demonstrated a higher number of non-zero weights in European ancestry (median = 55; IQR = [30;76]) compared to South Asian ancestry (median=32; IQR=[11;49]). The median absolute mixing weights were similar between European ancestry and South Asian ancestry (Supplementary Fig. 4). We note that most of the best PRS across traits were developed from 2021 onward. The details of the most recent trait-specific PRS, PRS with largest sample sizes, best PRS being compared and score with highest weights for European ancestry and South Asian ancestry are provided in Supplementary Table 5 and Supplementary Table 6, respectively."

We added Supplementary Table 5 and Supplementary 6 to highlight the PRS that contribute to the models for European and South Asian, respectively.

To comment on the scores that contribute to stroke compared to previous study from Abraham et al., we added "compared to previous work conducted on stroke such as usual walking pace, arthropathies, lipoprotein(a) (Supplementary Fig. 10 and Supplementary Table 17)."

3. The authors define partial R2 as "difference of R2 between the model with PRS and covariates including age, sex, and 10 PCs versus the base model with only covariates". That is not the definition of partial R2, so its use is misleading for comparison across papers. What the authors are defying is the difference in r2 between full and covariate model. The correct definition of partial R2 divides this difference by the full model r2 (Apologies for citing Wikipedia: https://en.wikipedia.org/wiki/Coefficient_of_determination#Coefficient_of_partial_determination). Please consider rephrasing or changing to correct partial R2. The authors use the correct terminology in the context of AUC in lines 412-413.

Response: We rephrased "partial R2" to "incremental R2".

Changes in manuscript: On line 138, 155, 222, 266, 315, 357, 742, 743 we revised “partial R2” into “incremental R2”.
On line 724-729, we added the description of incremental-R2 and liability-R2 for continuous traits and binary traits, respectively: “The prediction accuracy (R2) was calculated as incremental R2 which is a difference of R2 between the model with PRS and covariates including age, sex, and 10 PCs versus the base model with only covariates. Incremental-R2 indicates the difference between the full model and the covariate-only model which isolated the explanatory power of PRS2. Prediction accuracy for binary traits was assessed with liability-R2 where disease prevalence was approximately estimated as a proportion of cases in the testing set.‟

4. I think there is some slightly misleading discussion related to the purpose of using crosstrait models or specifically their PRSmix+ model. The authors claim that the additional value of including correlated traits in the model comes from pleiotropy. I agree pleiotropy plays a role, but there are many other reasons mostly related to cohortspecific effects that could increase prediction. In the same line, the authors attribute the results of low cross-cohort transferability of R2 between the two cohorts analysed to differences in genetic ancestry (227-282). However, there are multiple other sources of variability between the two cohorts that would decrease prediction accuracy e.g. work from the Global Biobank (Wang et al. 2023).

Response: We agree that there are many possible reasons. In our study, we have used crossvalidation to estimate the mixing weights of PRSs. We added the discussion of possible sources for increased prediction accuracy.

Changes in manuscript: On line 526-531, we added: “Additionally, the variability between the two cohorts would decrease prediction accuracy. There might be more contributing factors that influence the prediction accuracy such as sample sizes of the PGS panel and polygenicity of the traits6,32, ancestral consistency between discovery GWAS and LD reference panels in PRS methods33, ancestry proportions in the discovery GWAS33, and cohort-specific contexts34”

Comments by section

Simulations

1. The simulations work is very interesting and shows the potential to improve prediction in small samples sizes 200-5000 with PRSmix. In line 200, the text says "We observed that PRSmix demonstrated a saturation of improvement from Ntraining= 10,000." but Figure 2 shows saturated improvement in all simulated scenarios from Ntraining >= 5,000. Same with PRSmix+, the improvement is shown from 5,000 in Figure 2 but the text says 30,000.

Response: We apologize for this typo. We agree that the saturated improvement in all simulated scenarios from Ntraining >= 5,000. We revised the results from 10000 to 5000 individuals required for training.

Changes in manuscript: On line 204 and 205, we revised “10000” and “30000” into “5000”.

2. Comment 2. Both models PRSmix/PRSmix+ seem to be particularly useful for traits with low heritability, for which more PRS translate to better prediction.

Response: We thank the reviewer for this comment and agree with this notion. We discussed when there was only a marginal improvement of PRSmix+ over PRSmix and discussed that PRSmix+ would demonstrate a higher benefit compared to PRSmix for traits with low heritability on line 565-575: “We observed that in cases of highly heritable traits or high performance with a single PRS, there was only marginal improvement of PRSmix+ over PRSmix. In this scenario, PRSmix could provide similar predictive performance while being less timeconsuming because trait-specific PRS inputs only are required. However, for traits with lower heritability PRSmix+ shows a marked improvement over PRSmix and would be preferred.”

3. Comment 3. Line 695 says there are 7 traits heritability h2 equal to 0.05, 0.1, 0.2, and 0.5. Is this a typo or is there some information missing? Are there 7*4 traits? The number of simulated causal SNPs (1000, p = 0.001) is very small compared to real polygenic outcomes, expect the number of samples needed to predict more polygenic traits with the same accuracy would be much larger. Please state why you "removed individuals with PC1 and PC2 > 3 standard deviation from the mean." in the simulations methods. In line 711, please state again the proportion of sample to each of the 3 data sets.

Response: We thank the reviewer for pointing out this typo. We simulated 4 traits with h2 equal to 0.05, 0.1, 0.2 and 0.5. Via simulation, we aimed to evaluate the power of the linear combination across various values of heritability with a similar/fixed number of causal SNPs. As expected, from our simulation, we consistently showed that traits with higher heritability required a lower number of samples for the linear combination. We agree with the reviewer that number of samples needed to predict more polygenic traits with the same accuracy would be much larger as shown in the added formula. This has been well-established in the theoretical prediction accuracy mentioned in Wang et al. Cell Genomics 2023, Vilhjalmsson et al. AJHG 2015.

We removed individuals with PC1 and PC2 > 3 standard deviation from the mean to remove outliers of the inferred genetic ancestry. We also stated again the proportion of sample for each of 3 data sets.

Changes in manuscript:

On line 762-763, we revised to: "Overall, we simulated 4 traits with heritability $h2$ equal to 0.05, 0.1, 0.2, and 0.5."

We added reference 6 and 32 on line 528-529 to highlight the effect of polygenicity of the traits and training sample sizes on the prediction accuracy: "There might be more contributing factors that influence the prediction accuracy such as sample sizes of the PGS panel and polygenicity of the traits6,32"

On line 764-766, we added: "We removed individuals with PC1 and PC2 > 3 standard deviation from the mean to remove outliers of the inferred genetic ancestry."

On line 778-781, we added: "1) GWAS (200,000 individuals – 60%) 2) training set (130,000 individuals – 38%): training the mixing weights with a linear combination and 3) testing set (7000 individuals – 2%): testing the combined PRS. We incorporated PRS1, PRS2 and PRS3 to assess the trait-specific PRSmix framework"

Combining trait-specific PRS improves prediction accuracy (PRSmix)

4. Comment 4. In Figures 3 and 4, it would be useful to include the average number of non-zero weighted PRS overall and for each category of trait. In general, it would be very informative to know how PRS were included in each PRSmix and how many of those ended up with a weight of 0 after training the elastic net. Additionally, the PRS weight distribution per trait would be useful to understand where the improvement is coming from. It is a very different scenario that 10 PRSs in a model have a large weight and then there are 200 extra PRS with neglectable weight that do not contribute to increasing prediction but are "cohort specific".

Response: We thank the reviewer for the comment. We added a separate Supplementary Figure 4, Supplementary Table 4 to demonstrate the counts of non-zero mixing weights and the median of the absolute mixing weight distribution in European and South Asian ancestries. Indeed, the combination in European ancestry contains a higher number of non-zero-weight scores and a higher median of the absolute weights distribution than the combination in South Asian ancestry. This may explain a more efficient power of linear combination in Europeans relative to South Asians. We revised Supplementary Table 17 and Supplementary Table 18 to a long format to list the weights for each score in the combination.

Changes in manuscript: We added Supplementary Figure 4 and on line 275-280, we added: "PRSmix yielded an equivalent number of non-zero mixing weights between European ancestry (median=8; interquartile range = [5; 12]) and South Asian ancestry (median=8; IQR=[3;14]). However, PRSmix+ demonstrated a higher number of non-zero weights in European ancestry (median = 55; IQR = [30;76]) compared to South Asian ancestry (median=32; IQR=[11;49]). The median absolute mixing weights were similar between European ancestry and South Asian ancestry (Supplementary Fig. 4)."

5. Comment 5. The caption in Figure 3 has conflicting information on how the whiskers for the bar estimates were derived.

Response: We removed the conflicting information.

Changes in manuscript: In the caption of Figure 3, we removed "The whiskers reflect the maximum and minimum values within the 1.5 × interquartile range."

6. Comment 6. In the results for PRSmix+ in real data, the authors pre-select the traits from the 2,600 list to train the elastic net models filtering on their theoretic power. There is no comment or results on how many actual PRS are retained after this filter. Comment 4 from the PRSmix part also applies to PRSmix+.

Response: Similar to comment 4, we added the boxplot of the counts of non-zero mixing weights and the max absolute mixing weight in European and South Asian ancestries (Supplementary Figure 4).

Changes in manuscript: on line 275-280, we added: "PRSmix yielded an equivalent number of non-zero mixing weights between European ancestry (median=8; interquartile range = [5; 12]) and South Asian ancestry (median=8; IQR=[3;14]). However, PRSmix+ demonstrated a higher number of non-zero weights in European ancestry (median = 55; IQR = [30;76]) compared to South Asian ancestry (median=32; IQR=[11;49]). The median absolute mixing weights were similar between European ancestry and South Asian ancestry (Supplementary Fig. 4)."

7. Comment 7. Line 281 states "However, using a linear combination with the matched ancestry still demonstrated a better performance than using transferred weights." These results are highly relevant to the topic of prs transferability. They should be highlighted in the discussion, as it indicates that methods that attempt to re-adjust prs weights cross-ancestry are less effective for prediction than directly getting the weights form the matched ancestry.

Response: We agree that this is a strength of this framework and we have now added a fourth discussion point better emphasizing the cross-ancestry utility with PRSmix

Changes in manuscript: On line 521-526, we added: "Fourth, we showed that using a linear combination in a matched ancestry still demonstrated a better performance than using transferred pre-trained weights from another ancestry. This indicates that methods that attempt to re-adjust PRS weights cross-ancestry are less effective for prediction than directly obtaining the weights from the matched ancestry. Additionally, we showed that providing the linear combination model with all scores from all ancestries demonstrated a better predictive accuracy than using only ancestry-matched PRSs to the outcome trait."

Prediction accuracy and predictive improvement across various types of traits

8. Comment 8. In the section "Prediction accuracy and predictive improvement across various types of traits" I don't find the category "other conditions" particularly interesting as it seems very heterogeneous in its content. Therefore, the finding that the prediction in that category is the highest does not seem relevant.

Response: Although the "other condition" group is heterogeneous, we demonstrated that this group has a lowest prediction accuracy and best point-estimated improvement in the combination framework. This pattern is clearly seen in European ancestry. In South Asian, the "other conditions" group did not show high significant improvement due to a small number of traits in this category in Genes and Health data, we mentioned on line 335-338.

Comparison with previous combination methods

9. Comment 9. In Figure 5. It is very surprising that Elastic net has increased prediction over PRSmix (LDpred2 + PGS Catalog) for T2D. Can the authors comment further on why this could be the case?

Response: T2D is a trait that demonstrated a high pleiotropic effect with other complex traits and diseases. PRSmix (LDpred2+PGS Catalog) has a lower prediction accuracy because it incorporated trait-specific scores which might not be sufficient to improve the performance of T2D PRS. T2D is a common highly polygenic condition shared genetic information across other cardiometabolic risk factors as well as social/lifestyle factors. Threfore, multi-trait PRS methods which perform cross-trait combination for T2D would greatly benefit from employing other traits' information. Our revised Supplementary Table 17 and 19 gives information about the features in the model. This is mentioned and discussed on line 369-371 and 548-552.

Changes in manuscript: We added additional discussion on line 552-559: "We demonstrated that T2D demonstrated a greater prediction accuracy when incorporating information from multiple traits. T2D is a common highly polygenic condition correlated with other cardiometabolic risk factors as well as social/lifestyle factors. Furthermore, with a limited predefined list of correlated traits, we showed that cross-trait combination may give a similar performance as combining trait-specific scores in PGS Catalog."

10. Comment 10. In Figure 5. Elastic net vs. PRSmix does not show a significant improvement in prediction for BMI, CAD and depression. This indicates that currently there is no valuable information in the GWAS Catalog that improves just doing an elastic net model with 26 PRS. I am curious if this is the case for most of the results in this paper. Again, showing highlighting which PRS have large weights in PRSmix/PRSmix+ could help answer this question.

Response: We thank the reviewer for this comment. In the main text, ElasticNet demonstrated a similar power to PRSmix which indicates that additional secondary based on selected traits does not have a better performance compared to only including trait-specific scores.

Changes in manuscript: On line 557-565, we added: "Furthermore, with a limited pre-defined list of correlated traits, we showed that cross-trait combination may give a similar performance as combining trait-specific scores in PGS Catalog (BMI, CAD, depression in Fig. 5). Across different traits, we demonstrated that PRSmix+ and PRSmix has an overall better power compared to other methods. Cross-trait combination for height does not significantly improve prediction accuracy whereas T2D demonstrated a higher accuracy with any cross-trait combination methods. Intuitively, height has been known as a well-established highly polygenic trait thanks to its enormous sample sizes in a recent study21 with well-powered scores from the PGS Catalog."

11. Comment 11. Although the authors highlight the PRSmix+ results for T2D in Figure 5, a closer look at supplementary table 6 shows that the most pleiotropic effects captured by the multi-trait model are coming from the depression PRSmix+, with a x2 prediction accuracy boost over all other methods. In terms of pleiotropy/heterogeneity, psychiatric disorders are at a different league than T2D.

Response: We thank the reviewer for pointing out this interesting note. Indeed, our method agnostically assigns and penalizes the weights contributed by the PRSs. Instead of curating contributing traits as shown in several previous studies, we holistically demonstrated that permitting the model to decide the scores to combine would give us a broader picture of the contributing factors to improve prediction accuracy.

Changes in manuscript: On line 552-557, we added: "We demonstrated that T2D demonstrated a greater prediction accuracy when incorporating information from multiple traits. T2D is a common highly polygenic condition correlated with other cardiometabolic risk factors as well as social/lifestyle factors. Previous PRSs for T2D may not closely consider pleiotropic effect from correlated traits to improve PRS for T2D. PRS for T2D could still have room for improvements when further incorporate genetically correlated traits in the future."

12. Comment 12. The authors say in line 344 that "we benchmarked PRSmix and PRSmix+ against the previous methods". However, they only benchmark their methods explicitly to wMT-SBLUP. I don't expect more comparisons, but I think it is misleading to cite methods that are not being compared and that would yield potentially similar results to PRSmix+.

Response: We revised "previous methods" to wMT-SBLUP and a combination of PRSs developed by LDpred2

Changes in manuscript: On line 367-368, we added: … against wMT-SBLUP using summary statistics and a combination of PRSs developed by LDpred2…

Clinical utility for coronary artery disease

13. Comment 13. I think the clinical evaluation of the models in the context of CAD is very complete and gives a good overview of the different performance metrics that other researchers could be interested in.

Response: We thank the reviewer for this comment.

14. Comment 14. I don't agree with the authors' conclusion from the "plateau" analysis described in Supplementary figure 7. The simulations included only 6 PRS with genetic correlation with the outcome of 0.4-0.8. The CAD PRSmix included ~27 PRSs, based on Figure 6, but again it is unclear which ones were actually non-zero in the final model. I think it is misleading to conclude that 5,000 samples are required to train a model with max prediction accuracy, as this will enormously depend on many factors: snpheritability, polygenicity, heterogeneity in the definition of the disease etc. And finally, the composition of the PGS Catalog. CAD is a highly studied trait as the authors argue, but I think this 5,000 samples rule will not be the reality many other traits.

Response: In simulation, we aimed to evaluate the performance with different genetic architecture with heritability ranged from 5% to 40%, and with different sample size for the training of the linear combination. The overall trend demonstrated that PRSmix and PRSmix+ with sample size larger than 5000 or 10000 demonstrated a saturated improvement. CAD and most other traits in our study are highly polygenic trait and the real-data result agrees with our simulation. We agree with the reviewer that snp-heritability, polygenicity, heterogeneity in the definition of the disease etc may influence the sample sizes to evaluate. However, our simulation highlighted the most important information including heritability and the genetic correlation between traits to combine. Elastic Net would also be helpful in leveraging the heterogeneity of various scores that provided in the PGS Catalog that it can impose a penalty on the features that does not contribute to the combination.

Changes in manuscript:

On line 460, we added "Moreover, we observed that there was a modest improvement for PRSmix from the training size of 5000 in both European and South Asian ancestries (Supplementary Fig.

9)”.
On line 463-469, we added “To obtain maximized prediction accuracy in real-data, SNPheritability, polygenicity, heterogeneity in the definition of the disease may influence the sample sizes need for combination. However, our simulation highlighted the most important information including heritability and the genetic correlation between traits to combine. Improvements from the combination benefit from the imposing penalty by the ElasticNet on unimportant features across multiple scores in the combination.”
On line 469, we added “with CAD” to emphasize the application in CAD.
15. Comment 15. In line 551, the authors mention "Third, we did not validate the mixing weights in an independent cohort.". I am confused as the authors evaluated the crosscohort performance of PRSmix/PRSmix+ with results in Supplementary Table 4. I this cross-ancestry not All of Us vs. Genes & Health cohort? If not, I have misunderstood all cross-ancestry results. If yes, then the authors have validated the results in an independent cohort.
Response: We thank the reviewer for this comment. We indeed did perform the cross-ancestry evaluation between All of Us and Genes & Health cohort.
Changes in manuscript: On line 597, we removed this limitation.
Small comments
- Line 84. "To better capture the genetic architecture of the outcome traits, we proposed PRSmix, a framework to combine PRS from the same trait with the outcome trait." . I disagree the purpose of training on the same trait is to better capture its genetic architecture. Either to be truthful to the phenotype definition or
Changes in manuscripts: On line 86, we revised the sentence to: To aggregate the genetic information across different sources…
- Line 95. "other conditions as the prediction accuracies varied in each group".
Changes in manuscript: we removed “as the prediction accuracies varied in each group.”
- Line 132-133 " We selected the most common traits from the PGS Catalog which have the highest number of PRS."
Changes in manuscript: On line 135-136, we revised to “We selected traits from the PGS Catalog which have the highest number of PRS.”
- Line 151 "best single PRS from the training set" … more like the top weighted?
Changes in manuscript: On line 153-155, we revised the sentence to “For a fair comparison with the proposed framework, we selected the best-performing PRS from the set of traits matched with the outcome trait from the training set and evaluated by incremental R2 in the testing set.”
- Line 231 "who had undergone whole genome sequencing" is this relevant?
Changes in manuscript: On line 233-234, we revised the sentence to “we used whole genome sequencing data from European ancestry participants in the All of Us research program, and imputed Genes & Health participants of South Asian ancestry.”
Supplementary info
- The description for ST4 does not include the performance measure and seems cut.
Changes in manuscript: We added to the description of Supplementary Table 7 (this was ST4): “The prediction accuracy was assessed as incremental R2 and liability R2 for continuous and binary traits, respectively. The incremental R2 is a difference of R2 between the model with PRS and covariates including age, sex, and 10 PCs versus the base model with only covariates. Prediction accuracy for binary traits was assessed with liability R2 where disease prevalence was approximately estimated as a proportion of cases in the testing set.”
- Again, the description for ST5 does not include info enough to understand the table. What do the ratios represent?
Changes in manuscript: On the description of ST8 (this was ST5), we added: “The ratio was calculated by the ratio between prediction accuracy of PRSmix or PRSmix+ and the best PGS Catalog.”
- ST6 contains info from the figure that does not belong in the table "The whiskers demonstrate 95% confidence intervals of mean prediction accuracy. BMI, Body mass index; CAD, coronary artery disease; T2D, type 2 diabetes. GWAS, genome-wide association study."
Changes in manuscript: We removed “The whiskers demonstrate 95% confidence intervals of mean prediction accuracy.”

Referees' reports, second round of review

Reviewer#1 I would like to thank the authors for their efforts on improving this work. The major comments have been answered satisfactorily. I don't have further comments.

Reviewer#2 I would like to congratulate the authors for their great work and for providing thorough replies to all my comments. I have no further comments.

Authors' response to the second round of review

Reviewer #1: I would like to thank the authors for their efforts on improving this work. The major comments have been answered satisfactorily. I don't have further comments.

We would like to thank the reviewers for the constructive suggestions and comments.

Reviewer #2: I would like to congratulate the authors for their great work and for providing thorough replies to all my comments. I have no further comments.

We would like to thank the reviewers for the constructive suggestions and comments.