nature portfolio

Peer Review File

PGS Browser: a public platform for personalized polygenic score analysis and interpretation

Corresponding Author: Dr Mykyta Artomov

This file contains all reviewer reports in order by version, followed by all author rebuttals in order by version.

Version 0:

Reviewer comments:

Reviewer #1

(Remarks to the Author)
The manuscript presents a comprehensive evaluation of 3,168 polygenic score (PGS) models using data from PGS Catalogue using the FinnGen study, aiming to enhance personalized disease prediction by providing standardized benchmarking and accessible tools. The authors systematically analyze the performance of these PGS models across various complex diseases, demonstrating that integrating multiple scores can significantly improve predictive accuracy. They also provide publicly available resources, including ancestry-adjusted reference distributions and interactive predictive models through the PGS Browser, which facilitate equitable access to PGS analyses. Overall, the manuscript is well-structured and contributes valuable tools to individualized risk prediction. The PGS browser website is well-documented and user-friendly. The tutorial is comprehensive. I only have a few technical comments:

1. In Lines 160-163, the authors state that PGS models exhibit individual-level variability even when achieving similar AUROC scores. It would be beneficial to clarify whether this observation holds true for models with similar relative risk (90 vs. 50%), as the ability to stratify high-risk individuals is particularly meaningful in clinical applications.
2. In Lines 168-181, the authors provide a Spearman rank correlation matrix among all PGSs. Could the authors elaborate on how this matrix should be utilized to facilitate the application of PGSs in the context of individual-level variance?
3. The authors discuss two applications of the PGS atlas: investigating endpoints associated with a specific PGS and identifying PGSs associated with a given endpoint. However, both applications use depression as an example, which is somewhat expected. Can the authors provide more information on how these two utilities differ in terms of interpretation and potential application scenarios?
4. Lastly, does the integration of multiple PGSs improve prediction accuracy across different ancestries? This aspect would be important for assessing the generalizability of the findings.

Reviewer #2

(Remarks to the Author)
This study provides a valuable resource for the human polygenic score community by performing an orders-of-magnitude greater assessment of PGS performance in a very large population cohort that is independent of the ones used to derive the scores. Strengths of the study include the comprehensive scope (a truly impressive amount of work), the rigorous and careful state-of-the-art methodology, and the provision of an open-source web application for access to results. Key novel research findings include demonstration that predictive accuracy is often improved by incorporating multiple PGS, reciprocally demonstration that many PGS are also good predictors of other endpoints, and confirmation of relatively low (depending on your perspective) concordance among different scores for the same endpoint. We have five points of criticism, the last of which should require major revision.

First, the paragraph on lines 178-184 describing clinically relevant reporting standards (OR, TPR, S-PPV, relative and absolute risks) is too brief. These are the measures that would be used in clinics, but they are just stated without explanation. While the Supplement provides more details, even there it is difficult to discern what thresholds for inclusion in the predicted positive set are used – in one case RR is comparison of the 90th and 50th percentiles, but why those values, and is that standard for all assessments? This is particularly important for the PPV: most of the results in the manuscript are AUC values which span the full range of sensitivity and specificity, whereas these other standards need a cutoff and will be

highly dependent on it. We suggest a Box listing the 5 measures, stating the rationale, and the thresholds with some summary stats.

Second, more clarity on the full pipeline, criteria for exclusion as well as sample sizes of the cases and controls used for model evaluation should be provided as a supplement. The number of models retained for analyses and the filtering criteria is confusing to follow, with respect to overlap of variants, matched phenotype definitions, overlap between discovery samples and PGS test samples. For variance explained on the liability scale, it is essential to adjust for both the sample prevalence (in the analytic cohort) and the population prevalence of the disease. The approach described in the supplement fits a logistic model to compute R2 but does not appear to transform it to the liability scale using population prevalence parameters.

Third, it is unclear from the methods how the authors determined the optimal combination of PGSs in the multi-PGS models to achieve improved prediction, how rank-level inconsistencies across individuals were handled, and how heterogeneity in model construction was addressed. This is critical since, as noted, the concordance between multiple PGSs for the same trait or disease is low in the PGS Catalog (Supplementary Figures S4 - S6), owing to differences in GWAS, PGS construction methods, and phenotype definitions. For instance, only 30% overlap was observed in the top 2.5% of hypertension PGSs with similar AUCs. Even the top performing PGS models for coeliac disease with an AUC ~ 0.8 also show a high discordance.

Fourth, please give readers a sense of the clinical utility of the scores. An AUC value is fine for overall accuracy but is not used for patient reporting. How should readers interpret the values in Figure 2f – under what circumstances does a PPV of 0.1 have utility when prevalence is standardized to 0.05? How is this affected by the observed variance in PGS performance? Is a vitiligo OR of 4 meaningfully greater than a thyroid gland disorder OR of 1.5? I think this can be simply addressed in the Discussion, but it is essential for a paper claiming “equitable access” in the title – access to what exactly? Additionally, the authors should consider reporting credible intervals for the combined multi PGS and further discuss the implications of individual level uncertainty of PGS estimates.

Fifth, and this is the point of major revision, we think the ancestry comparisons (Figure 5, much of the supplement) should be removed from this paper. They are too preliminary and do not address a hypothesis sufficiently. Here are three issues. (i) Scaling the PGS essentially to have the same or similar mean and variance across populations makes some sense, and has been recommended by others, but it is far from clear that it is the right thing to do. If the PGS were instead eGFR, it would be akin to applying a “race correction factor” which is disallowed in the US but supported in Europe. Bottom line, even if the genetic risk is the same across populations, it is not obvious that a PGS derived from Europeans ought to have similar distributions in non-European populations. That is a research question not addressed here, and it seems dangerous to me to throw out the adjusted scores for others to use without addressing it. (ii) The adjustment is almost certainly population-specific since the PC will be different, as will the burden of disease in different settings. Thus, if someone in London or Los Angeles or Johannesburg or Rio were to use the adjusted scores, they would still need to readjust them for the local setting, and it is not clear that the FinnGen adjusted scores are the best starting point for that. (iii) There is also the issue that this adjustment only removes the systemic bias of ancestry proportions, it does not address the other aspect of transferability which is much reduced accuracy across ancestry groups with distance from Europe. The 19K non-Europeans in this study is too small, once partitioned and given small case numbers for most endpoints, to support any meaningful assessment of accuracy, so this important aspect is preliminary. Since the remainder of the paper is about providing a validated resource, and these adjusted scores are not validated, we strongly recommend that they be removed to another paper carefully dealing with these and other issues, leaving a clean message for this one.

Two minor corrections:
Starting on line 148, when you report the OR, state that this is the OR per standard deviation unit increase in the PGS (presumably).
Line 293: ref 10 should be ref 11.

Version 1:

Reviewer comments:

Reviewer #1

(Remarks to the Author)
The authors have addressed my major cocnerns. I only have a minor comment:
The website of PGS browser (https://www.pgs.nchigm.org) in line 628 seems not accessible. The correct website is https://pgs.nchigm.org.

Reviewer #2

(Remarks to the Author)
Thanks very much for your unusually comprehensive and thoughtful response to the reviews. I am not entirely in agreement with some of your arguments around the cross-ancestry comparisons, but appreciate that you addressed the issue, made some useful changes, and explained your reasoning. It is also clear that your resource is a major step forward for the community. I just hope that there is not a lot of uncritical transfer of scores that ignore other issues such as environmental

differences between cohort locations that also impinge on PGS utility. This is all material for debate and development though. Good luck with this.

REVIEWER COMMENTS

Reviewer #1 (Remarks to the Author):

The manuscript presents a comprehensive evaluation of 3,168 polygenic score (PGS) models using data from PGS Catalogue using the FinnGen study, aiming to enhance personalized disease prediction by providing standardized benchmarking and accessible tools. The authors systematically analyze the performance of these PGS models across various complex diseases, demonstrating that integrating multiple scores can significantly improve predictive accuracy. They also provide publicly available resources, including ancestry-adjusted reference distributions and interactive predictive models through the PGS Browser, which facilitate equitable access to PGS analyses. Overall, the manuscript is well-structured and contributes valuable tools to individualized risk prediction. The PGS browser website is well-documented and user-friendly. The tutorial is comprehensive. I only have a few technical comments:

We thank the reviewer for positive evaluation of our manuscript and helpful technical comments.

1. In Lines 160-163, the authors state that PGS models exhibit individual-level variability even when achieving similar AUROC scores. It would be beneficial to clarify whether this observation holds true for models with similar relative risk (90 vs. 50%), as the ability to stratify high-risk individuals is particularly meaningful in clinical applications.

We agree that relative risk at high PGS percentiles is more informative for clinical applications than AUROC. We therefore extended the analysis to explicitly compare PGS models with similar relative risks (RRs) among high-risk individuals (90 vs 50%). We find that substantial rank-level discordance persists even among models with comparable RRs, indicating that similar RR performance does not imply concordant identification of high-risk individuals. This clarification has been incorporated into the revised manuscript (**Lines 171-181**).

2. In Lines 168-181, the authors provide a Spearman rank correlation matrix among all PGSs. Could the authors elaborate on how this matrix should be utilized to facilitate the application of PGSs in the context of individual-level variance?

We thank the reviewer for this question. In brief, the Spearman correlation matrix can supplement the PGS model selection process by providing the information about the redundancy and complementarity among resulting scores. Highly correlated PGSs tend to rank individuals similarly and are therefore largely redundant, in which case users may select the model with the strongest predictive performance. In contrast, pairs of PGSs that both show strong performance for the same phenotype but low correlation are likely to capture partially distinct components of genetic liability, suggesting that their combination in a multi-PGS model may provide additional value. We have updated the main text accordingly to provide these details (**Lines 187-193**).

3. The authors discuss two applications of the PGS atlas: investigating endpoints associated with a specific PGS and identifying PGSs associated with a given endpoint. However, both applications use depression as an example, which is somewhat expected. Can the authors provide more information on how these two utilities differ in terms of interpretation and potential application scenarios?

We thank the reviewer for this comment and apologize for the incomplete illustration of the utility of PGS atlas. Although both examples use depression for illustration and share some overlapping use cases, the two applications address distinct analytical questions. We have updated the main text to clarify these differences (**Lines 238-246**). In brief:

The first, *PGS-centric application* ("given a specific PGS, which endpoints are associated with it?"), is intended to characterize a given PGS model rather than exact phenotype in terms of its target and secondary phenotypic associations. For example, when examining the depression PGS, this approach identifies other endpoints (e.g., cardiovascular disease) that are associated with that score. These observations can then be used to investigate which variants within the PGS model contribute to these associations, potentially providing additional insight into disease etiology.

The second, *endpoint-centric application* ("given an endpoint, which PGSs are associated with it?"), is designed to identify PGSs associated with a given FinnGen endpoint. This application supports risk-factor prioritization for the selected endpoint and can serve, for example, as a feature pre-selection step for multi-PGS disease prediction models. In the case of depression, this application highlights which PGSs from other traits are associated with depression risk and could therefore be incorporated into multi-PGS models of lifetime depression risk.

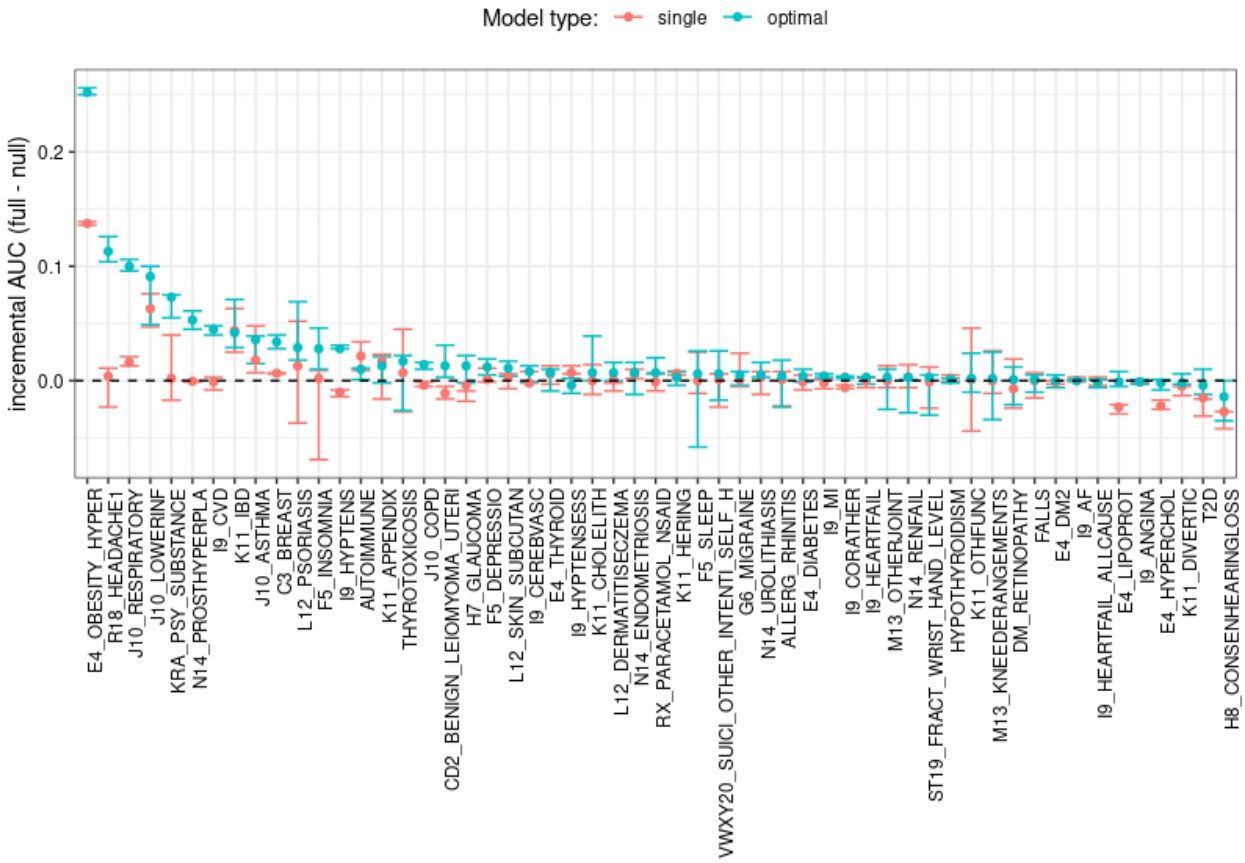
4. Lastly, does the integration of multiple PGSs improve prediction accuracy across different ancestries? This aspect would be important for assessing the generalizability of the findings.

We thank the reviewer for this suggestion and we agree that evaluating whether multi-PGS integration improves prediction accuracy across ancestries would be important for assessing generalizability of the results. To address this point, we conducted an additional analysis evaluating null models, best single-PGS models, and multi-PGS elastic-net classifiers in a non-European mixed cohort (n = 5,095) from a newer FinnGen release (R13). We combined different non-European ancestries into one set mainly due to the small sample sizes for individual ancestral groups. As a result, a non-European cohort included 456 individuals from African continental population, 2000 - Admixed American, 695 - East Asians, 352 - South Asians and 1592 - Other non-European admixed individuals. Additionally, to ensure sufficient statistical power, we restricted analysis to 55 phenotypes (FinnGen endpoints) with at least 100 incident cases (see **Review Figure 1** below).

Consistent with the results reported in the manuscript, the multi-PGS approach on average outperformed the best single-PGS models when evaluated in the non-European cohort.

However, while successfully replicating our primary finding, we note that the relatively small sample size and heterogeneous ancestry composition of this combined cohort limit the strength of conclusions regarding cross-ancestry generalizability. These limitations are now

explicitly discussed in the revised **Discussion** section (**Lines 410–413**).



Review Figure 1. Incremental ROC AUC for 55 FinnGen endpoints on a test set consisting of 5,095 non-European individuals.

Reviewer #2 (Remarks to the Author):

This study provides a valuable resource for the human polygenic score community by performing an orders-of-magnitude greater assessment of PGS performance in a very large population cohort that is independent of the ones used to derive the scores. Strengths of the study include the comprehensive scope (a truly impressive amount of work), the rigorous and careful state-of-the-art methodology, and the provision of an open-source web application for access to results. Key novel research findings include demonstration that predictive accuracy is often improved by incorporating multiple PGS, reciprocally demonstration that many PGS are also good predictors of other endpoints, and confirmation of relatively low (depending on your perspective) concordance among different scores for the same endpoint. We have five points of criticism, the last of which should require major revision.

We thank the reviewer for positive evaluation of our work and insightful comments.

First, the paragraph on lines 178-184 describing clinically relevant reporting standards (OR, TPR, S-PPV, relative and absolute risks) is too brief. These are the measures that would be

used in clinics, but they are just stated without explanation. While the Supplement provides more details, even there it is difficult to discern what thresholds for inclusion in the predicted positive set are used – in one case RR is comparison of the 90th and 50th percentiles, but why those values, and is that standard for all assessments? This is particularly important for the PPV: most of the results in the manuscript are AUC values which span the full range of sensitivity and specificity, whereas these other standards need a cutoff and will be highly dependent on it. We suggest a Box listing the 5 measures, stating the rationale, and the thresholds with some summary stats.

We thank the reviewer for this helpful comment and agree that the motivation for reporting these classification metrics and the rationale for the chosen thresholds were not sufficiently explained in the original manuscript. To address this, we have revised the section in the main text and added a summary table in the Supplement describing each measure, its interpretation and addressing threshold choice questions (**Lines 203-218**, **Supplementary Note, PGS evaluation measures, Table 1**). In addition, section "PGS Evaluation measures" in the supplement has a "practical interpretation" note with respect to every metric used in the manuscript.

Importantly, these metrics are not intended to define clinical decision thresholds in the current study. Rather, they are reported to characterize the statistical properties of PGS models beyond AUROC and to facilitate standardized comparison across models and phenotypes. Because measures such as PPV and relative risk indeed depend on the selected risk threshold, we standardized the threshold across models to enable consistent comparisons between PGS models. In the manuscript we therefore used the 90th percentile of the PGS distribution as an illustrative reference point, a commonly used threshold in the PGS literature (PMID: 38531883, 35414057, 38172535, 39218908).

At the same time, we emphasize that the appropriate cutoff is context-dependent and depends on the desired balance between precision and recall. For this reason, the PGS Browser provides an interactive interface in which these metrics can be explored across the full range of percentile thresholds (see https://pgs.nchigm.org/, "Classification measures" tab), allowing users to select thresholds appropriate for their specific research or clinical context.

Second, more clarity on the full pipeline, criteria for exclusion as well as sample sizes of the cases and controls used for model evaluation should be provided as a supplement. The number of models retained for analyses and the filtering criteria is confusing to follow, with respect to overlap of variants, matched phenotype definitions, overlap between discovery samples and PGS test samples.

We thank the reviewer for this helpful comment. To address this point, we revised the Results subsection "PGS Catalog and FinnGen data harmonization" to describe the full filtering pipeline more explicitly, including criteria related to variant overlap, phenotype matching, and exclusion of models with sample overlap between discovery and evaluation cohorts. We also redesigned **Fig. 1e** to make the sequence of inclusion and exclusion steps easier to follow, and expanded the Supplementary tables to report the numbers of cases and controls used in each evaluation experiment (**Lines 125-143**, see **Supplementary Table S2, S5, S7, S8**).

For variance explained on the liability scale, it is essential to adjust for both the sample prevalence (in the analytic cohort) and the population prevalence of the disease. The approach described in the supplement fits a logistic model to compute R2 but does not appear to transform it to the liability scale using population prevalence parameters.

We agree with the reviewer that, rather than reporting variance explained on the latent "logistic liability scale" as defined in Lee et al. (2012), it is more informative to report R2 on the normal liability-threshold scale, which incorporates analytic cohort and population prevalences. We therefore recalculated R2 on normal liability scale using FinnRegistry (n=7.2 mil.) as a primary source of population prevalence and updated the relevant figures, main text, and Supplement accordingly (see **Lines 195-201**, **Figure 2e**, **Supplementary Note, PGS evaluation measures**; **Supplementary Figure S7; Supplementary Table S4**).

Third, it is unclear from the methods how the authors determined the optimal combination of PGSs in the multi- PGS models to achieve improved prediction, how rank- level inconsistencies across individuals were handled, and how heterogeneity in model construction was addressed. This is critical since, as noted, the concordance between multiple PGSs for the same trait or disease is low in the PGS Catalog (Supplementary Figures S4 - S6), owing to differences in GWAS, PGS construction methods, and phenotype definitions. For instance, only 30% overlap was observed in the top 2.5% of hypertension PGSs with similar AUCs. Even the top performing PGS models for coeliac disease with an AUC ~ 0.8 also show a high discordance.

We thank the reviewer for this comment. We have revised the Methods section to explicitly describe how multi-PGS models were constructed and how optimal score combinations were determined (**Material and Methods, Risk prediction models**).

In brief, for each disorder we jointly modeled all candidate PGSs (n = 1,514) using an elastic-net-regularized regression. Model hyperparameters (the L1-L2 mixing parameter and overall regularization strength) were tuned using 10-fold cross-validation, and performance was evaluated on out-of-sample test sets. Elastic net is well suited for settings with multiple correlated predictors, as it balances sparsity and joint modeling: groups of correlated PGSs may be retained together with reduced coefficients when this improves predictive performance, or a single representative score may be selected when additional predictors are redundant and do not improve predictive performance. We therefore refer to the resulting set of retained scores as the "optimal" combination for a given disorder, as it maximizes predictive performance under the specified regularization and cross-validation scheme.

In this framework, rank-level discordance between individual PGSs is not treated as a problem to be resolved but rather as a potential source of complementary information. When multiple PGSs capture partially distinct components of genetic liability, their joint inclusion can improve prediction. Conversely, when correlated scores do not provide additional predictive signal, regularization shrinks redundant coefficients toward zero, yielding a sparse model.

This approach also addresses heterogeneity in PGS construction (differences in GWAS, modeling methods, and phenotype definitions), as all candidate scores are evaluated jointly within a single predictive framework and retained only when they contribute to model

performance.

Fourth, please give readers a sense of the clinical utility of the scores. An AUC value is fine for overall accuracy but is not used for patient reporting. How should readers interpret the values in Figure 2f? – under what circumstances does a PPV of 0.1 have utility when prevalence is standardized to 0.05? How is this affected by the observed variance in PGS performance? Is a vitiligo OR of 4 meaningfully greater than a thyroid gland disorder OR of 1.5?

We thank the reviewer for this constructive comment. We agree that additional context is needed to help readers interpret the potential clinical relevance of the reported performance metrics. To address this, we expanded the Supplementary section and added an explanatory table describing the interpretation and practical implications of the main evaluation measures used in this study. We also included explicit examples illustrating how these metrics can be interpreted in practice, including the scenarios raised by the reviewer ("Practical interpretation" subsections in the Supplement), and moved Figure 2f to the supplement to accompany the description (see **Supplementary Note, PGS evaluation measures, Table 1; Supplementary Figure S8**).

For example, a standardized PPV of 0.1 under a reference prevalence of 0.05 corresponds to approximately a two-fold enrichment of risk relative to baseline. Such enrichment may be useful for identifying subgroups for targeted screening, monitoring, or research recruitment. Standardization to 0.05 (~median prevalence across endpoints) allows cross-model comparison.

Differences in odds ratios across diseases (e.g., OR=4 vs OR=1.5) reflect differences in genetic risk stratification strength but should not be interpreted as directly comparable measures of clinical actionability. Their practical significance depends on additional factors, such as disease prevalence or disease severity.

The observed variance in PGS performance also affects the interpretation of threshold-based metrics. For traits with modest predictive performance, the same percentile thresholds typically produce smaller risk differences than for highly predictive traits, and therefore provide less value for individual-level decision making.

Importantly, the metrics reported in this study are intended primarily to contextualize the statistical properties of PGS models and enable standardized comparison between them, rather than to strictly define clinical decision thresholds.

I think this can be simply addressed in the Discussion, but it is essential for a paper claiming "equitable access" in the title – access to what exactly?

We thank the reviewer for this helpful comment and agree that the meaning of "equitable access" in the title required clearer explanation. In this work, "equitable access" refers to open access to PGS reference distributions and predictive models derived from national biobank imposing strict limitations on the data sharing. In practice, these resources are typically available only to researchers with direct access to FinnGen. By providing aggregated reference distributions and predictive models through the PGS Browser platform, we aim to make these resources accessible to researchers who do not have direct access to such large-scale cohorts.

To avoid potential ambiguity in the title, we have revised the Discussion section to clarify this point (**Lines 443-453**) and updated the manuscript's title to remove confusing term from the title and more clearly reflect the role of the platform: "PGS Browser: a public platform for personalized polygenic score analysis and interpretation".

Additionally, the authors should consider reporting credible intervals for the combined multi PGS and further discuss the implications of individual level uncertainty of PGS estimates.

We thank the reviewer for this helpful suggestion. As our modeling framework in this article is entirely frequentist rather than Bayesian, we added bootstrap-based confidence intervals for the delta AUCs for individual multi-PGS models (**Lines 288-315**), which appropriately quantify sampling uncertainty in model evaluation. For summary comparisons across endpoints, we additionally report the interquartile range for delta AUCs. In addition, we have expanded the Materials and Methods section to clarify that PGS-derived risk estimates represent probabilistic risk stratification rather than deterministic predictions for a given individual (**Material and Methods, Risk prediction models**). Uncertainty arises from multiple sources, including finite GWAS discovery sample sizes, probabilistic nature of PGS methods and multi-PGS model construction. As a result, individual-level predictions should be interpreted as risk estimates within a population context rather than precise estimates of an individual's absolute disease probability (**Lines 597-602**).

Fifth, and this is the point of major revision, we think the ancestry comparisons (Figure 5, much of the supplement) should be removed from this paper. They are too preliminary and do not address a hypothesis sufficiently.

We thank the reviewer for this comment. We partially agree that PGS performance in different ancestries in FinnGen are too preliminary and may not be sufficient to judge reliably about generalizability of results. We have updated this section of the manuscript by removing PGS evaluations involving non-European FinnGen samples from main and supplementary text. However, we respectfully disagree that analyses of performance in different ancestries in the UK biobank (**Supplementary Figure S16-20**) and ancestry adjustment procedures (**Figure 5**) should be removed from the results (please see detailed responses below for clarification).

Here are three issues. (i) Scaling the PGS essentially to have the same or similar mean and variance across populations makes some sense, and has been recommended by others, but it is far from clear that it is the right thing to do. If the PGS were instead eGFR, it would be akin to applying a "race correction factor" which is disallowed in the US but supported in Europe. Bottom line, even if the genetic risk is the same across populations, it is not obvious that a PGS derived from Europeans ought to have similar distributions in non-European populations. That is a research question not addressed here, and it seems dangerous to me to throw out the adjusted scores for others to use without addressing it.

We appreciate this thoughtful comment. We agree that forcing PGS to have identical mean and/or variance across ancestry groups is not biologically justified, and we did not intend to

imply that such normalization reflects true equivalence of genetic risk across populations. Our goal is narrower: to ensure that percentile-based interpretation of PGS is comparable across ancestries when scores are primarily derived from European training data.

Raw PGS distributions often differ substantially across populations due to differences in allele frequencies, linkage disequilibrium structure, and other technical factors. As a result, percentiles calculated from unadjusted PGS can reflect ancestry differences rather than relative genetic risk within a population. In downstream analyses, particularly when PGS percentiles are used as inputs to predictive models, this may lead models to partially capture ancestry-related variation rather than disease risk.

For this reason, we apply adjustment as a calibration step to improve percentile comparability, rather than as a biological correction implying that true genetic risk distributions must be identical across populations. The following ancestry adjustment aims to reduce ancestry-driven artifacts in percentile-based interpretation and in downstream modeling, *conceptually similar to standard covariate PC adjustment*. Furthermore, in this specific implementation we simply adopted the previously introduced adjustment techniques for PGS from the heavily cited literature (PMID: 35437332, 38374346).

We acknowledge that the broader question of whether PGS distributions should or should not differ across populations is an important research topic that remains unresolved and is beyond the scope of this work. To avoid misinterpretation, we have clarified in the manuscript that this adjustment is intended solely to facilitate cross-ancestry comparability of percentile-based analyses (see **Lines 413-419**; **Materials and Methods, PGS ancestry adjustment**).

(ii) The adjustment is almost certainly population-specific since the PC will be different, as will the burden of disease in different settings. Thus, if someone in London or Los Angeles or Johannesburg or Rio were to use the adjusted scores, they would still need to readjust them for the local setting, and it is not clear that the FinnGen adjusted scores are the best starting point for that.

We thank the reviewer for this comment and we believe there is a misunderstanding of how our standardized ancestry-adjustment procedure works. The PC space used for adjustment in our framework was not derived from FinnGen samples. Instead, we constructed the reference PC space using whole-genome sequencing data from the 1000 Genomes Project, which captures major continental populations. FinnGen individuals were subsequently projected into this predefined **fixed reference space**.

As a result, the ancestry adjustment is performed relative to a fixed and externally defined PC basis, rather than one derived from the FinnGen cohort itself. This ensures that the same adjustment procedure can be applied consistently to both the reference PGS distributions (derived from FinnGen) and to any external datasets. Consequently, samples from other cohorts (e.g., from London, Los Angeles, Johannesburg, or Rio de Janeiro) can be projected into the same reference PC space and processed using the adjustment procedure identical to the one used for the reference FinnGen data. Moreover, this procedure (using 1000G) would be almost the only choice to reliably capture any internal ancestral variation if the analyzed cohort is small (e.g. single patient) and researcher is lacking any other bigger cohorts to work with.

In addition, the coefficients estimated for PCs in regression for PGS adjustment were estimated using individuals representing multiple continental ancestry groups, which helps make the resulting calibration more broadly applicable across diverse populations.

To facilitate reproducibility and consistent application of this procedure, we provide an open-source command-line tool (*pgsb-cli;* see https://pgs.nchigm.org/, "Documentation" tab) that automates projection into the reference PC space and the subsequent adjustment steps. This ensures that external datasets can apply the same standardized procedure used in our analyses even for a single sample. We have updated Materials and Methods to clarify discussed points (**Materials and Methods, PGS ancestry adjustment**).

(iii) There is also the issue that this adjustment only removes the systemic bias of ancestry proportions, it does not address the other aspect of transferability which is much reduced accuracy across ancestry groups with distance from Europe.

The ancestry adjustment procedure is intended only to standardize score distributions across ancestries to make percentiles comparable, and not intended to resolve differences in predictive accuracy. Transferrability of the PGS performance will be an inherent limitation, however, this major research direction in the field is somewhat outside of the scope of current manuscript. We agree that this important note was not communicated explicitly in the text. To make this idea more explicit we have updated discussion and the materials and methods section (see **Lines 413-419**; **Materials and Methods, PGS ancestry adjustment**). We note that similar ancestry-adjustment strategies have been adopted in several recent observational clinical genomics studies and emerging clinical trials involving polygenic scores (PMID: 38374346, 35437332; ClinicalTrials.gov: NCT04331535), reflecting the current practical approach for enabling cross-population interpretation of PGS distributions.

The 19K non-Europeans in this study is too small, once partitioned and given small case numbers for most endpoints, to support any meaningful assessment of accuracy, so this important aspect is preliminary. Since the remainder of the paper is about providing a validated resource, and these adjusted scores are not validated, we strongly recommend that they be removed to another paper carefully dealing with these and other issues, leaving a clean message for this one.

We thank the reviewer for this comment and we agree with the reviewer that these analyses are preliminary and that the number of non-European individuals in FinnGen and number of cases for most endpoints are insufficient to support a reliable assessment of predictive accuracy across ancestries. In response to this concern, we have removed the corresponding PGS evaluations and ancestry comparisons involving non-European FinnGen cohorts from both the Results and the Supplementary Materials (former **Supplementary Figures S8-S9,** current **Supplementary Figure S14**, and the related sections in the **Main Text, Performance evaluation of PGS Catalog models in the FinnGen cohort**).

Two minor corrections:

Starting on line 148, when you report the OR, state that this is the OR per standard deviation unit increase in the PGS (presumably).
Line 293: ref 10 should be ref 11

We thank the reviewer for pointing out these issues. We have revised the main text (**Line 160**).