Publicly Available Privacy-preserving Benchmarks for Polygenic Prediction

Menno J. Witteveen[1], Emil M. Pedersen[1], Joeri Meijsen[3], Michael Riis Andersen[4], Florian Privé[1], Doug Speed[2,6], and Bjarni J. Vilhjálmsson[1,5,6]

[1]National Centre for Register-based Research , Aarhus University, Denmark
[2]Center for Quantitative Genetics and Genomics , Aarhus University, Denmark
[3]Institute of Biological Psychiatry, Denmark
[4]Technical University of Denmark, Denmark
[5]Bioinformatics Research Centre, Aarhus University, Denmark
[6]These authors contributed equally
menno.j.witteveen@gmail.com

Abstract

Recently, several new approaches for creating polygenic scores (PGS) have been developed and this trend shows no sign of abating. However, it has thus far been challenging to determine which approaches are superior, as different studies report seemingly conflicting benchmark results. This heterogeneity in benchmark results is in part due to different outcomes being used, but also due to differences in the genetic variants being used, data preprocessing, and other quality control steps. As a solution, a publicly available benchmark for polygenic prediction is presented here, which allows researchers to both *train* and *test* polygenic prediction methods using only summary-level information, thus preserving privacy. Through simulations and real data, we show that model performance can be estimated with accuracy, using only linkage disequilibrium (LD) information and genome-wide association summary statistics for target outcomes. Finally, we make this PGS benchmark - consisting of 8 outcomes, including somatic and psychiatric disorders - publicly available for researchers to download on our PGS benchmark platform (http://www.pgsbenchmark.org). We believe this benchmark can help establish a clear and unbiased standard for future polygenic score methods to compare against.

1 Introduction

In recent years, interest in polygenic scores (PGS) has increased greatly, with researchers finding progressively more applications for polygenic scores in biomedical research, genetics, and epidemiology. Polygenic scores are now routinely used to examine the genetic relationship between outcomes, such as bipolar disorder and schizophrenia [35] and in phenome-wide association studies [5]. Polygenic scores can also be used to infer causal relationships [11] as well as improve power in genome-wide association studies (GWAS) [2, 29]. There is also a compounding body of evidence supporting the claim that polygenic scores can improve risk models in clinical applications [17, 14].

In this climate, there has been a proliferation of new approaches for creating polygenic scores, including linkage-disequilibrium (LD) clumping and p-value thresholding[9], Bayesian approaches [33, 13, 20, 46, 48], penalized regression [21, 32], and other machine learning methods [40]. Most of these new polygenic score methods claim to outperform previously proposed methods, making it confusing for users to choose an appropriate PGS approach for their application. However, this apparent performance paradox is due to challenges in properly determining what approaches are

superior, because different datasets, quality control, and preprocessing steps are used to determine performance. As shown in previous work, these processing steps can have a significant impact on the overall prediction accuracy for polygenic score methods [34], and thus lead to an incomplete comparison of PGS methods.

In the related field of Machine Learning these problems have been addressed by the introduction of publicly available benchmark datasets, which allow for fair comparison between approaches. This has been vital to the advancement of new and more powerful methods in the machine learning field. Such benchmark datasets, also called benchmarks, provide researchers with an identical dataset and accompanying evaluation metrics. This unified setup allows (and constrains) the researcher to both *train* and *test* models. Hence, benchmarks make published performance measures directly comparable, negating the need to redo analyses and also allowing for complete replication. Additionally, since many of these benchmarks are publicly available for anyone to download, not requiring special access of any kind, this poses minimal hurdles for researchers. Indeed, a clear and easy-to-use benchmark enables researchers to focus on improving their method instead of having to apply for data access and go through the sometimes onerous process of applying other PGS methods to the same data. However the creation of such publicly available benchmark datasets for polygenic prediction is challenging because of privacy concerns, as individual-level genotype data and health outcomes are only made available to approved researchers with restrictions.

As a solution to this challenge, we present a privacy-preserving and publicly available benchmark for polygenic prediction, which allows researchers to both *train* and *test* polygenic prediction methods using only summary-level information, which we define as linkage disequilibrium (LD) data and GWAS summary statistics, thus preserving privacy. GWAS summary statistics are usually made publicly available, and several public and easily accessible repositories exist, containing GWAS summary statistics for thousands of outcomes [44, 4]. Using both simulations and UK Biobank data (UKBB) for a diverse set of 8 external summary statistics, including both somatic and psychiatric disorders, we show that the squared correlation prediction accuracy can be estimated almost perfectly using summary-level test data only. We further used the benchmark data to compare a collection of commonly used polygenic scoring methods including PRS-CS, LDpred2, and SBayesR and observe a high concordance, with complete recovery of model rankings and almost perfect correlation of model performance measures. Finally, we make the benchmark data, necessary for training and testing, publicly available and encourage other researchers to consider using it to benchmark their methods.

2 Results

2.1 Constructing a Privacy-Preserving Benchmark [PPB]

Polygenic prediction methods typically use summary statistics and LD information as inputs for model training [33, 13, 20, 48]. However, this input data has not yet been used to construct benchmarks for PGS methods. Here we propose to use an alternative formulation of the squared Pearson correlation (R^2), which is regularly used as a performance measure for polygenic prediction. A similar formulation of the Pearson correlation and related measures has been proposed previously in the context of fitting hyper-parameters[21] and model selection [48, 41]. Staring from the original form of the performance measure, which is the squared Pearson correlation of the observed and predicted phenotypes, we can show that the prediction R^2 is equal to

$$R^2 = \frac{\left(\boldsymbol{\beta}^T \mathbf{z}\right)^2}{\boldsymbol{\beta}^T \mathbf{D} \boldsymbol{\beta}}.$$

Here, $\boldsymbol{\beta}$ is a vector of length M containing model weights of the polygenic prediction approach that is to be evaluated and M being the number of genetic variants, which typically is larger than 10^5 up to several million. $\mathbf{D} = \frac{1}{N}\mathbf{X}^T\mathbf{X}$ is the full M-by-M covariance matrix also called the LD matrix, and $\mathbf{z} = \frac{1}{N}\mathbf{X}^T\mathbf{y}$ denotes the summary statistic z-scores for the outcome of interest. Furthermore, $\mathbf{y}$ denotes the standardized phenotype and $\mathbf{X}$ the standardized genotype matrix, where each genetic variant has mean 0 and variance 1. A detailed derivation can be found in the supplementary information.

Although the equation above would allow us to perfectly estimate R^2 if we would have access to the full LD matrix, in practice, the full LD matrix $\mathbf{D}$ is very large, making it impractical to both compute and share. By default, we approximate $\mathbf{D}$ as a banded LD matrix by not including covariance beyond a genetic distance of 4 centimorgan (cM), in effect setting this covariance to zero. Additionally, we present results derived from other genetic distance cut-offs and LD blocks [3]. Hence, the missing parts (off-diagonal elements) of our approximation of $\mathbf{D}$ will be the cause of discrepancies between our privacy-preserving approach and the individual-level approach, meaning the original approach for determining R^2 by computing the squared correlation between the observed and predicted phenotypes. Interestingly, it is also possible to compute other performance measures using the same inputs, including the mean square error (MSE).

Experimental flow: methods & experiments. We compared computing performance measures using individual level data with our privacy-preserving benchmark [PPB] approach for a number of polygenic prediction methods, being PRS-CS, SBayesR, LDpred2 and lassosum, using both automatic and validation set based versions for hyper-parameter tuning where applicable. We looked both at concordance of absolute performance measure values and at relative performance characteristics (i.e. rankings of the methods).

We trained the collection of polygenic prediction methods with external summary statistics (real and simulated) and LD from the validation set, which was a randomly selected 10K subset of our preprocessed UK Biobank dataset (N=362,320 and 1,117,493 genetic HapMap3 variants)[33, 7]. For this benchmarking work, we followed and reproduced the experimental setup for PGS benchmarking as used in Privé *et al.* [33], unless otherwise specified. Real and simulated phenotypes from the UK Biobank and iPSYCH cohort [31, 6], were adjusted for sex, age and 10 principal components and the residuals were subsequently used. Then model selection (hyper-parameter tuning) was performed when required, using individual-level and privacy-preserving data modalities for their respective approaches (individual-level or PPB). This was followed by final model evaluation on the UK Biobank test dataset, yielding prediction R^2 for both individual-level and PPB approaches.

2.2 Simulations

The impact of LD reference To examine how the estimated R^2 depends on different LD references and genetic architectures, we performed simulations with simulated genotypes and phenotypes (see methods for details). We compared the prediction accuracy of three different simple polygenic scores, namely one where we used the true causal effects, one with marginal least squares effects (linear regression), and one where we applied a p-value threshold to the marginal least squares estimates. We then considered three different LD matrix references used to estimate the prediction R^2: one where we used the test data as LD reference (which should be optimal), one where training data was used as the LD reference, and finally an independent LD reference. The results for 1000 simulated phenotypes with heritability varying from 0 to 1 are shown in Supplementary Figure S1. We found that using the training data as LD reference results in biased prediction R^2 estimates, whereas using the target data as LD reference provides near optimal estimates. We also found that an independent LD reference (with samples not included in training nor testing data) yielded unbiased estimates for the prediction R^2. This does shed some additional light on our usage of the validation set for both LD and hyper-parameter selection, but as we show in this work this bias does not appear to lead to substantial performance deviations for the privacy-preserving approach if the validation data is solely used for selecting model hyper-parameters and not for estimating model performance.

Simulations using real genotypes In order to gain insight into comparative accuracy of the privacy-preserving approach we performed simulations, using real genotype data (from the UK Biobank and iPSYCH datasets) and simulated genetic effects with different levels of polygenicity and repeated each simulation 10 times. For the different levels of polygenicity we randomly picked a total of 10^3, 10^4, 10^5 or All to be causal variants. We then simulated phenotypes for the UK Biobank and iPSYCH cohorts, using a Liability Threshold Model (LTM), with a heritability of 0.4 and a prevalence of 20% [12] and performed GWAS in the iPSYCH cohort to get GWAS summary statistics for the prediction approaches in our comparison. We then treated these like summary statistics from real traits. We combined the summary statistics with the LD from the validation dataset and used this to train our models. After this we performed hyper-parameter selection using the validation dataset. Lastly, we evaluated final prediction performance for the simulated traits on the UKBB test dataset for both the individual-level and privacy-preserving based approaches.

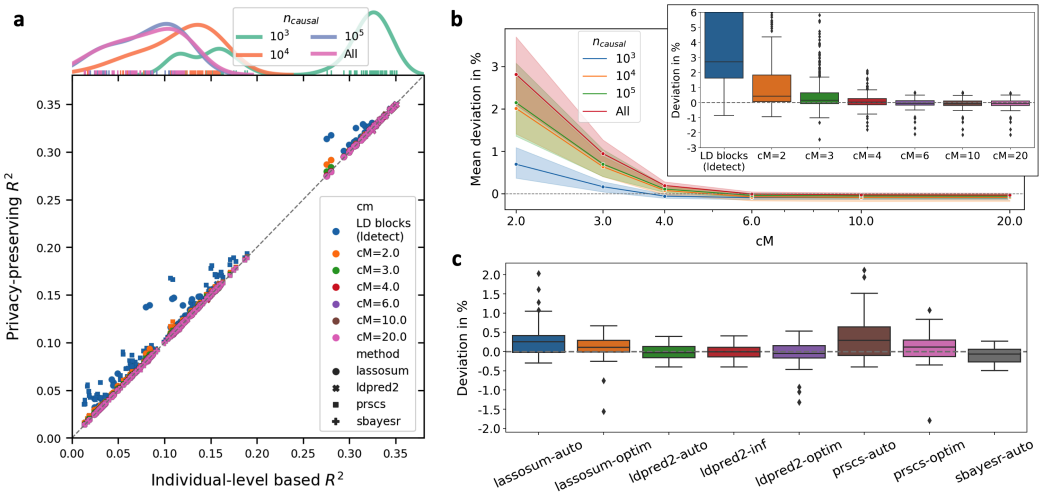

Figure 1: Accuracy of privacy-preserving prediction R^2 in simulations. a, Matched prediction R^2 computed using the privacy-preserving or individual-level approach for different LD approximation windows and blocks, prediction methods and effect-size distributions (shown by the density plot on the top). **b**, The difference in percentage of the privacy-preserving vs. the individual-level based approach for computing performance. Firstly shown as the mean of percentages grouped by effect-size distribution (number of causal variants) vs window size in centimorgan (cM), with 95% confidence intervals and secondly with boxplots for different window sizes (cM) and LD blocks (ldetect, [3]). **c**, Boxplots of differences in percentages for different polygenic prediction methods for a 4 cM window-size.

To investigate the power and potential limitations of a privacy-preserving approach for performance evaluation, the results of the simulations were analyzed (see Figure 1). Overall, concordance between our PPB and the individual-level approach was noteworthy, both in absolute and relative terms, if a sufficiently large LD window was used. In cases where the quantity of captured LD was small, the R^2 tended to be overestimated (Fig. 1a). We observed this overestimation for all effect-size distributions (+1.9% on average with 2 cM windows), but it was more pronounced for more polygenic ones (Fig. 1b) as measured by average percent deviation. For instance, with a window-size of 2 cM, the most polygenic effect-size distribution showed +2.82% average deviation versus +0.70% for the least polygenic one (All and 10^3 respectively). This overestimation rapidly became minimal with larger LD windows for all effect-size distributions. Furthermore, we observed that relying on the widely-used ldetect LD blocks [3] led to substantial overestimation of the prediction R^2 (+12.4% on average). These results suggest that a LD window-size of 4 cM provides a good balance between accuracy and LD data size, which one would like to keep of a manageable size for effective dissemination and use of the benchmark. Next, one can see the accuracy of the privacy-preserving approach for a number of prediction approaches in Figure 1c, with average deviations relatively close to zero across the prediction methods.

Another way to view the results is to consider the average correlation per effect-size distribution for the selected window-size of 4 cM (Table 1). We determined the average Spearman's and Pearson's correlation between the individual-level and privacy-preserving R^2 estimate over the different simulations to get an idea of both absolute and ranking accuracy within traits. As can be seen, Pearson's correlation is very close to one for all effect-size distributions. For the All effect-size experiments (all variants being causal), there was only one single experiment with a single rank flip ($6 \rightleftarrows 7$) of LDpred2-auto and LDpred2. However, in this particular case, there was only a 0.34% difference between the two approaches. Hence, based on the empirical evidence provided by these simulations, we can conclude that the overall concordance of the privacy-preserving approach aligns closely to the individual-level approach for an assortment of effect-size distributions.

n_{causal}	$\bar{\rho}_{Pearson}$	$\bar{\rho}_{Spearman}$
10^3	1.0000	1.0000
10^4	0.9999	1.0000
10^5	1.0000	1.0000
All	1.0000	0.9976

Table 1: The average Spearman's and Pearson's rho over the repeats of the experiment for different effect-size distributions using 4 cM windows. All rho's are very close to one with the Spearman's rho of the All effect-size, being due to one ranking flip for one simulation.

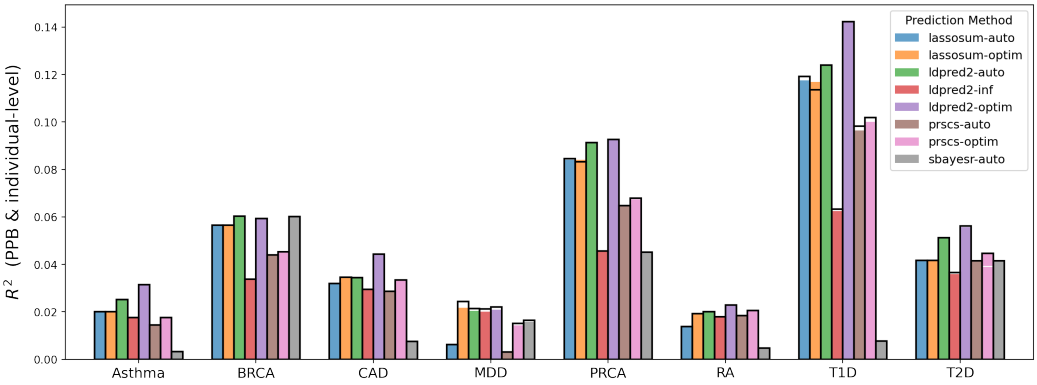

Figure 2: Barplot of the R^2 for different polygenic prediction approaches for 8 human traits. Both results for the individual-level and privacy-preserving approach are present in this barplot with the individual level being the colored bar, which represents the R^2 and the borders around the bars being the R^2 that was computed using the privacy-preserving approach. This mean that white spaces at the tops of the bars, like for Type 1 diabetes (T1D) or Major depression (MDD) represent overestimation of the R^2 by PPB. The meaning of the trait abbreviation can be found in Table 2.

2.3 Performance comparisons for real human traits

For real data analyses, 8 external GWAS summary statistics (details contained in Table S1) were used together with LD and other information from the validation set to train and optimize a collection of polygenic predictors. These predictors were then applied to the UK Biobank test set yielding prediction R^2 measures using the privacy-preserving and the individual-level approach, which are both contained in Figure 2 for their respective model-trait combination.

Trait	$\rho_{Pearson}$	$\rho_{Spearman}$
Asthma	1.0000	1.0000
Breast cancer (BRCA)	0.9999	1.0000
Coronary artery disease(CAD)	0.9999	1.0000
Major depression (MDD)	0.9979	1.0000
Prostate cancer (PRCA)	0.9998	1.0000
Rheumatoid arthritis (RA)	0.9999	1.0000
Type 1 diabetes (T1D)	0.9992	1.0000
Type 2 diabetes (T2D)	0.9997	1.0000

Table 2: Pearson's and Spearman's rho between the privacy-preserving and individual-level derived prediction R^2 within the respective traits. As can be seen the Spearman's rho has the highest possible value and hence the ranking is complete preserved.

In general, we observed good agreement between the two approaches. The only (small) discrepancy we observe is for Major depression (MDD), where we have a tendency to overestimate the R^2, which is an interesting point we will return to in later sections. We also observed small deviations including for Type-1 Diabetes (T1D), and initially a difference for Type 2 Diabetes (T2D) with PRS-CS ($+13.6\%$). Here, closer inspection reveals that this is due to PRS-CS[13] selecting a different hyper-parameter value based on the validation data when using the privacy-preserving approach, where the resulting R^2 values for the validation set differed by -0.27%. When using the same hyper-parameter value, the difference in the final test set estimates is only 0.71%.

Additionally, we looked at absolute and relative performance by rankings within traits, by computing Spearman's and Pearson's rho, which are shown in Table 2. We see very good agreement in terms of both absolute and relative terms, showing that the privacy-preserving approach is effective for comparing and ranking polygenic prediction methods across traits.

Finally, considering these real outcomes, we examined the average discrepancy per trait. For this we used the same approach for hyper-parameter selection in validation set (individual-level) for both the privacy-preserving and individual-level approach. We then plotted the mean deviation (measured in percentages) for different LD window-sizes in Figure 3. As can be seen clearly in this plot, all traits converge towards mean deviation of zero for increasing window-sizes, with the notable exception of Major depression, which converges to a prediction R^2 overestimation of about $+4.5\%$. Considering the fact that our simulations showed robust convergence towards a mean deviation of zero, this is a curious discrepancy from the trend, which we will revisit in the Discussion section. Also, the observed tail off for real traits further motivates our selection of the 4 cM window-size, being a good compromise between estimation performance and resulting dataset size.

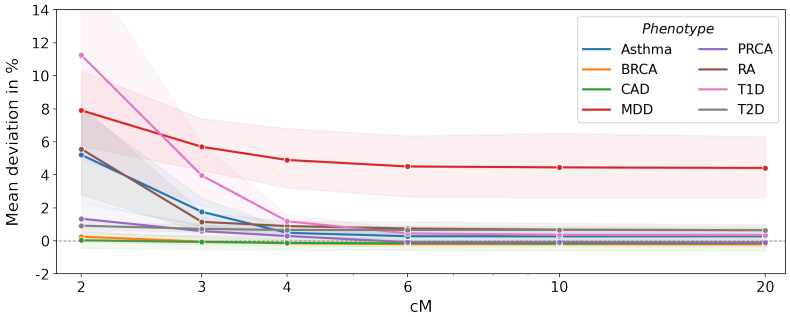

Figure 3: Mean deviation in % by trait for different window-sizes, with 95% confidence intervals. The sizes of the windows are indicated on the x-axis in centimorgan (cM) on a log-scale. The trait results were denoised for hyper-parameter selection deviations, by using the individual-level approach for both PPB and individual-level benchmarking.

3 Discussion

Here we propose a novel way to establish a linear polygenic prediction benchmark, that does not require any individual-level data to be shared. Using both simulations and real data we have demonstrated that, for our benchmark, performance can be determined with great accuracy, both in relative and absolute terms. We will shortly make this new benchmark dataset publicly available on our PGS benchmark platform (http://www.pgsbenchmark.org), where it can be downloaded by anyone. As shown, this benchmark allows for standardized unambiguous comparisons between the plethora of methods currently available. Additionally, we have created a publicly-accessible frequently-updated leaderboard on our platform in which one can see the approaches that are currently *state-of-the-art*, greatly simplifying researchers search for the best polygenic prediction approach. Additionally, the benchmark will allow researchers to much more easily reproduce each other's results. We believe this benchmark can be used as a clear and unbiased standard for future polygenic score methods to compare against.

Regarding possible limitations of our approach, obvious items to focus on would be to include better modeling of the potential assortative mating effects to account for the observed discrepancies (e.g. for the affected trait of Major depression). In fact, earlier experiments (data not shown) indicate that for certain other traits (Height and BMI) this effect is also present. This suggest the effect could be caused by assortative mating, since there is a significant body of evidence suggesting assortative mating effects for these effected traits [42, 23, 47, 36]. Also, we feel it is worth noting that this very discrepancy might be exploited as a measure of assortative mating. An expansion to the described benchmark could be to add a vetted collection of GWAS summary statistics in addition to the ones already supplied, which would allow for the evaluation of polygenic score combiner based approaches. Another enhancement would be to include ancestrally diverse summary-level data, which our benchmark is currently lacking, to enable analysis and improvement of cross-population prediction methods. We aim to address this limitation by publishing additional datasets for a diverse set of genetic ancestries for our benchmark to ameliorate the much-publicized risk that PGS could increase health-disparities [22]. Additionally, we could introduce recently proposed LD graphical models (LDGMs), which are sparse and efficient representations of LD, to more effectively model and share LD for diverse populations[27]. Another improvement would be to include privacy-preserving benchmarks for five common psychiatric disorders from the large Danish iPSYCH datasets.

Lastly, open benchmarks have played a crucial role in several fields, including those of computer vision (e.g. ImageNet [15, 18]), protein folding [1, 16, 39, 25], and more[43, 30]. Therefore, we believe that the impact of establishing open and privacy-preserving benchmarks could be profound, especially if one considers the downstream effects on medical science.

References

[1] Alquraishi, M. (2019). AlphaFold at CASP13. *Bioinformatics (Oxford, England)*, **35**(22), 4862–4865.

[2] Bennett, D. *et al.* (2021). Controlling for background genetic effects using polygenic scores improves the power of genome-wide association studies. *Scientific Reports*, **11**(1), 1–10.

[3] Berisa, T. and Pickrell, J. K. (2016). Approximately independent linkage disequilibrium blocks in human populations. *Bioinformatics (Oxford, England)*, **32**(2), 283–285.

[4] Buniello, A. *et al.* (2019). The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic Acids Research*, **47**(D1), D1005–D1012.

[5] Bush, W. S. *et al.* (2016). Unravelling the human genome-phenome relationship using phenome-wide association studies. *Nature Reviews Genetics*, **17**(3), 129–145.

[6] Bybjerg-Grauholm, J. *et al.* (2020). The iPSYCH2015 Case-Cohort sample: Updated directions for unravelling genetic and environmental architectures of severe mental disorders. *medRxiv : the preprint server for health sciences*.

[7] Bycroft, C. *et al.* (2018). The UK Biobank resource with deep phenotyping and genomic data. *Nature*, **562**(7726), 203–209.

[8] Censin, J. *et al.* (2017). Childhood adiposity and risk of type 1 diabetes: A mendelian randomization study. *PLoS Medicine*, **14**(8), e1002362.

[9] Choi, S. W. and O'Reilly, P. F. (2019). PRSice-2: Polygenic Risk Score software for biobank-scale data. *GigaScience*, **8**(7), 1–6.

[10] Demenais, F. *et al.* (2018). Multiancestry association study identifies new asthma risk loci that colocalize with immune-cell enhancer marks. *Nature Genetics*, **50**(1), 42.

[11] DiPrete, T. A. *et al.* (2018). Genetic instrumental variable regression: Explaining socioeconomic and health outcomes in nonexperimental data. *Proceedings of the National Academy of Sciences of the United States of America*, **115**(22), E4970–E4979.

[12] Falconer, D. S. (1965). The inheritance of liability to certain diseases, estimated from the incidence among relatives. *Annals of Human Genetics*, **29**(1), 51–76.

[13] Ge, T. *et al.* (2019). Polygenic prediction via Bayesian regression and continuous shrinkage priors. *Nature Communications*, **10**(1), 1–10.

[14] Inouye, M. *et al.* (2018). Genomic risk prediction of coronary artery disease in 480,000 adults: Implications for primary prevention. *Journal of the American College of Cardiology*, **72**(16), 1883–1893.

[15] Jia Deng *et al.* (2009). ImageNet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255. Ieee.

[16] Jumper, J. *et al.* (2021). Highly accurate protein structure prediction with AlphaFold. *Nature*, **596**(7873), 583–589.

[17] Khera, A. V. *et al.* (2018). Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations. *Nature Genetics*, **50**(9), 1219–1224.

[18] Krizhevsky, A. *et al.* (2012). ImageNet classification with deep convolutional neural networks. *Advances in Neural Information Processing Systems*, **25**.

[19] Lee, S. H. *et al.* (2012). A better coefficient of determination for genetic profile analysis. *Genetic epidemiology*, **36**(3), 214–224.

[20] Lloyd-Jones, L. R. *et al.* (2019). Improved polygenic prediction by Bayesian multiple regression on summary statistics. *Nature Communications*, **10**(1).

[21] Mak, T. S. H. *et al.* (2017). Polygenic scores via penalized regression on summary statistics. *Genetic Epidemiology*, **41**(6), 469–480.

[22] Martin, A. R. *et al.* (2019). Clinical use of current polygenic risk scores may exacerbate health disparities. *Nature Genetics*, **51**(4), 584–591.

[23] Mathews, C. A. and Reus, V. I. (2001). Assortative mating in the affective disorders: A systematic review and meta-analysis. *Comprehensive Psychiatry*, **42**(4), 257–262.

[24] Michailidou, K. *et al.* (2017). Association analysis identifies 65 new breast cancer risk loci. *Nature*, **551**(7678), 92.

[25] Moult, J. *et al.* (2020). Critical assessment of techniques for protein structure prediction, fourteenth round. *CASP 14 Abstract Book*.

[26] Nikpay, M. *et al.* (2015). A comprehensive 1000 genomes–based genome-wide association meta-analysis of coronary artery disease. *Nature Genetics*, **47**(10), 1121.

[27] Nowbandegani, P. S. *et al.* (2022). Extremely sparse models of linkage disequilibrium in ancestrally diverse association studies. *bioRxiv*.

[28] Okada, Y. *et al.* (2014). Genetics of rheumatoid arthritis contributes to biology and drug discovery. *Nature*, **506**(7488), 376.

[29] Orliac, E. J. *et al.* (2021). Improving GWAS discovery and genomic prediction accuracy in Biobank data. *bioRxiv : the preprint server for biology*, page 2021.08.12.456099.

[30] Panayotov, V. *et al.* (2015). Librispeech: An ASR corpus based on public domain audio books. In *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, volume 2015-Augus, pages 5206–5210.

[31] Pedersen, C. B. *et al.* (2018). The iPSYCH2012 case–cohort sample: New directions for unravelling genetic and environmental architectures of severe mental disorders. *Molecular psychiatry*, **23**(1), 6–14.

[32] Privé, F. *et al.* (2019). Efficient implementation of penalized regression for genetic risk prediction. *Genetics*, **212**(1), 65–74.

[33] Privé, F. *et al.* (2020). LDpred2: Better, faster, stronger. *Bioinformatics (Oxford, England)*, **36**(22-23), 5424–5431.

[34] Privé, F. *et al.* (2022). Identifying and correcting for misspecifications in GWAS summary statistics and polygenic scores. *bioRxiv : the preprint server for biology*, pages 1–18.

[35] Purcell, S. M. *et al.* (2009). Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. *Nature*, **460**(7256), 748–752.

[36] Robinson, M. R. *et al.* (2017). Genetic evidence of assortative mating in humans. *Nature Human Behaviour*, **1**(1), 1–13.

[37] Schumacher, F. R. *et al.* (2018). Association analyses of more than 140,000 men identify 63 new prostate cancer susceptibility loci. *Nature Genetics*, **50**(7), 928.

[38] Scott, R. A. *et al.* (2017). An expanded genome-wide association study of type 2 diabetes in Europeans. *Diabetes*, **66**(11), 2888–2902.

[39] Service, R. F. (2020). 'The game has changed.' AI triumphs at protein folding: In milestone, software predictions finally match structures calculated from experimental data. *Science (New York, N.Y.)*, **370**(6521), 1144–1145.

[40] Sigurdsson, A. *et al.* (2021). Deep integrative models for large-scale human genomics. *bioRxiv : the preprint server for biology*.

[41] Song, L. *et al.* (2019). Summaryauc: a tool for evaluating the performance of polygenic risk prediction models in validation datasets with only summary level statistics. *Bioinformatics*, **35**(20), 4038–4044.

[42] Torvik, F. A. *et al.* (2022). Modeling assortative mating and genetic similarities between partners, siblings, and in-laws. *Nature Communications*, **13**(1), 1–10.

[43] Wang, A. *et al.* (2018). GLUE: A multi-task benchmark and analysis platform for natural language understanding. *EMNLP 2018 - 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP, Proceedings of the 1st Workshop*, pages 353–355.

[44] Watanabe, K. *et al.* (2019). A global overview of pleiotropy and genetic architecture in complex traits. *Nature Genetics*, **51**(9), 1339–1348.

[45] Wray, N. R. *et al.* (2018). Genome-wide association analyses identify 44 risk variants and refine the genetic architecture of major depression. *Nature Genetics*, **50**(5), 668.

[46] Yang, S. and Zhou, X. (2020). Accurate and scalable construction of polygenic scores in large biobank data sets. *American Journal of Human Genetics*, **106**(5), 679–693.

[47] Yengo, L. *et al.* (2018). Imprint of assortative mating on the human genome. *Nature Human Behaviour*, **2**(12), 948–954.

[48] Zhang, Q. *et al.* (2021). Improved genetic prediction of complex traits from individual-level data or summary statistics. *Nature Communications*, **12**(1), 1–9.

Supplementary Information

Derivations for the PPB Method

This section goes into more detail regarding the derivation of procedure for arriving at our desired performance measures. We start by considering the following equation.

$$r = \frac{\sum_i (x_i - m_x)(y_i - m_y)}{\sqrt{\sum_i (x_i - m_x)^2}\sqrt{\sum_i (y_i - m_y)^2}}$$

which is the definition of the Pearson correlation coefficient. We can create a vectorized version, tailored to our use case, which is:

$$r = \frac{\frac{1}{N}(\hat{\mathbf{y}} - \mathbf{m}_{\hat{\mathbf{y}}})^T (\mathbf{y} - \mathbf{m}_{\mathbf{y}})}{\sqrt{\frac{1}{N}(\hat{\mathbf{y}} - \mathbf{m}_{\hat{\mathbf{y}}})^T (\hat{\mathbf{y}} - \mathbf{m}_{\hat{\mathbf{y}}})}\sqrt{\frac{1}{N}(\mathbf{y} - \mathbf{m}_{\mathbf{y}})^T (\mathbf{y} - \mathbf{m}_{\mathbf{y}})}} \quad , with \begin{cases} \hat{\mathbf{y}} = \mathbf{X}\boldsymbol{\beta} \\ c^2 = \frac{1}{N}\hat{\mathbf{y}}^T\hat{\mathbf{y}} \end{cases}$$

m is repeating vector, containing the mean. If we manage to tackle all the elements in this equation this will allow us to compute Pearson's r and therefore R^2 too. Also, if so desired, a prediction magnitude invariant mean squared error can be computed, using, $MSE = -2(R-1)$. A factor of $\frac{1}{N}$ on both sides of the division, because it will simplify upcoming derivations. Mind that we have full control over the standardization of $\mathbf{y}$ and chose to standardize it with a mean of 0 and standard deviation of 1. First, it is useful to notice that the following portion of the numerator can be expanded. For the computation of our final R^2 we also need to take into account the prevalence information for the respective trait. Therefore, we make use of a latent trait model correction as described in earlier work which takes the prevalence of a trait into account [19].

$$(\hat{\mathbf{y}} - \mathbf{m}_{\hat{\mathbf{y}}})^T (\mathbf{y} - \mathbf{m}_{\mathbf{y}}) = \hat{\mathbf{y}}^T\mathbf{y} - \underbrace{\hat{\mathbf{y}}^T\mathbf{m}_{\mathbf{y}} - \mathbf{y}^T\mathbf{m}_{\hat{\mathbf{y}}} + \mathbf{m}_{\hat{\mathbf{y}}}^T\mathbf{m}_{\mathbf{y}}}_{= 0, \text{ if } \mathbf{y} \text{ has a mean of zero.}}$$

Since $\mathbf{y}$ has a mean of zero, the numerator for r simplifies to $\frac{1}{N}\hat{\mathbf{y}}^T\mathbf{y}$. This is equivalent to $\boldsymbol{\beta}^T\left[\frac{1}{N}\mathbf{X}^T\mathbf{y}\right]$, being the product of the weights and the summary statistics, which we can easily compute. Next, we can split the $\frac{1}{N}$ factor into the roots of the denominator and realize that following holds since we chose to standardize $\mathbf{y}$ with a standard deviation of 1.

$$\sqrt{\frac{1}{N}(\mathbf{y} - \mathbf{m}_{\mathbf{y}})^T (\mathbf{y} - \mathbf{m}_{\mathbf{y}})} = 1$$

which further simplifies the equation for r. This leaves us with the task to determine a value for the empirical standard deviation of $\hat{\mathbf{y}}$, which we will denote with c, which we will now analyze.

$$\begin{aligned} c &= \sqrt{\frac{1}{N}(\hat{\mathbf{y}} - \mathbf{m}_{\hat{\mathbf{y}}})^T (\hat{\mathbf{y}} - \mathbf{m}_{\hat{\mathbf{y}}})} \\ c^2 &= \frac{1}{N}(\hat{\mathbf{y}} - \mathbf{m}_{\hat{\mathbf{y}}})^T (\hat{\mathbf{y}} - \mathbf{m}_{\hat{\mathbf{y}}}) \\ Nc^2 &= \hat{\mathbf{y}}^T\hat{\mathbf{y}} - 2\hat{\mathbf{y}}^T\mathbf{m}_{\hat{\mathbf{y}}} + \mathbf{m}_{\hat{\mathbf{y}}}^T\mathbf{m}_{\hat{\mathbf{y}}} \end{aligned}$$

Here, we re-encounter the mean of the prediction, $\mathbf{m}_{\hat{\mathbf{y}}}$, but in this case, its effect does not factorize out, like before. However, since we can standardize every variable in $\mathbf{X}$ to have a mean of 0, the quantity turns out to be 0 for that case.

$$\begin{aligned} \hat{y}_i &= x_1^i\beta_1 + x_2^i\beta_2 + x_3^i\beta_3 + \ldots + x_M^i\beta_M \\ \mathbf{m}_{\hat{\mathbf{y}}} \underset{\text{vectorize}}{\leftarrow} \frac{1}{N}\sum_i^N \hat{y}_i &= \frac{1}{N}[\left(x_1^1 + \ldots + x_1^i + \ldots + x_1^N\right)\beta_1 + \cdots + \left(x_j^1 + \ldots + x_j^i + \ldots + x_j^N\right)\beta_j \\ &\quad + \left(x_M^1 + \ldots + x_M^i + \ldots + x_M^N\right)\beta_M] \\ 0 &= \underbrace{\left(x_j^1 + \ldots + x_j^i + \ldots + x_j^N\right)}_{=0}\beta_j \end{aligned}$$

Computational verification confirmed this too. Therefore we can take $\mathbf{m}_{\hat{\mathbf{y}}}$ to be $\mathbf{0}$ and simplify our equation for c^2, which together with the previous simplifies to equation for r into the follow.

$$r = \frac{\frac{1}{N}\hat{\mathbf{y}}^T\mathbf{y}}{\sqrt{\frac{1}{N}\hat{\mathbf{y}}^T\hat{\mathbf{y}}}} \quad , with \quad \begin{cases} \hat{\mathbf{y}} = \mathbf{X}\boldsymbol{\beta} \\ c = \sqrt{\frac{1}{N}\hat{\mathbf{y}}^T\hat{\mathbf{y}}} \end{cases}$$

$$r = \frac{1}{c}\,\boldsymbol{\beta}^T \underbrace{\left[\tfrac{1}{N}\mathbf{X}^T\mathbf{y}\right]}_{\text{sumstats}}$$

With this equation, it becomes apparent that, the only non trivial part of computing r is finding a value of c. Thus our effort will now focus finding a good way of estimating c^2. For this, we will introduce some additional variables and a singular value decomposition (SVD) of a scaled $\mathbf{X}$, with the scaling being $\sqrt{N}$ because it unclutters the notation in later approaches.

$$c^2 = \tfrac{1}{N}\hat{\mathbf{y}}^T\hat{\mathbf{y}} = \boldsymbol{\beta}^T \underbrace{\left[\tfrac{1}{N}\mathbf{X}^T\mathbf{X}\right]}_{\mathbf{D}\text{, full LD}} \boldsymbol{\beta}$$

Hence:

$$R^2 = \frac{\left(\boldsymbol{\beta}^T\left[\frac{1}{N}\mathbf{X}^T\mathbf{y}\right]\right)^2}{\boldsymbol{\beta}^T\mathbf{D}\boldsymbol{\beta}} = \frac{\left(\boldsymbol{\beta}^T\widetilde{\boldsymbol{\beta}}\right)^2}{\boldsymbol{\beta}^T\mathbf{D}\boldsymbol{\beta}}$$

Extended results

Trait	GWAS reference	GWAS sample size	# GWAS variants	# matched variants
Breast cancer (BRCA)	Michailidou *et al.* [24]	137,045 / 119,078	11,792,542	1,114,424
Rheumatoid arthritis (RA)	Okada *et al.* [28]	29,880 / 73,758	9,739,303	656,087
Type 1 diabetes (T1D)	Censin *et al.* [8]	5913 / 8828	8,996,866	514,420
Type 2 diabetes (T2D)	Scott *et al.* [38]	26,676 / 132,532	12,056,346	1,108,760
Prostate cancer (PRCA)	Schumacher *et al.* [37]	79,148 / 61,106	20,370,946	1,115,688
Depression (MDD)	Wray *et al.* [45]	59,851 / 113,154	13,554,550	1,103,440
Coronary artery disease (CAD)	Nikpay *et al.* [26]	60,801 / 123,504	9,455,778	1,108,313
Asthma	Demenais *et al.* [10]	19,954 / 107,715	2,001,280	980,430

Table S1: External GWAS summary statistics. Summary of the 8 external GWAS summary statistics used for this work, with their respective abbreviations. The GWAS sample size is the number of cases / controls in the GWAS. Initial number of GWAS variants is given and the number of variants after matching with our PPB dataset from the UK Biobank.

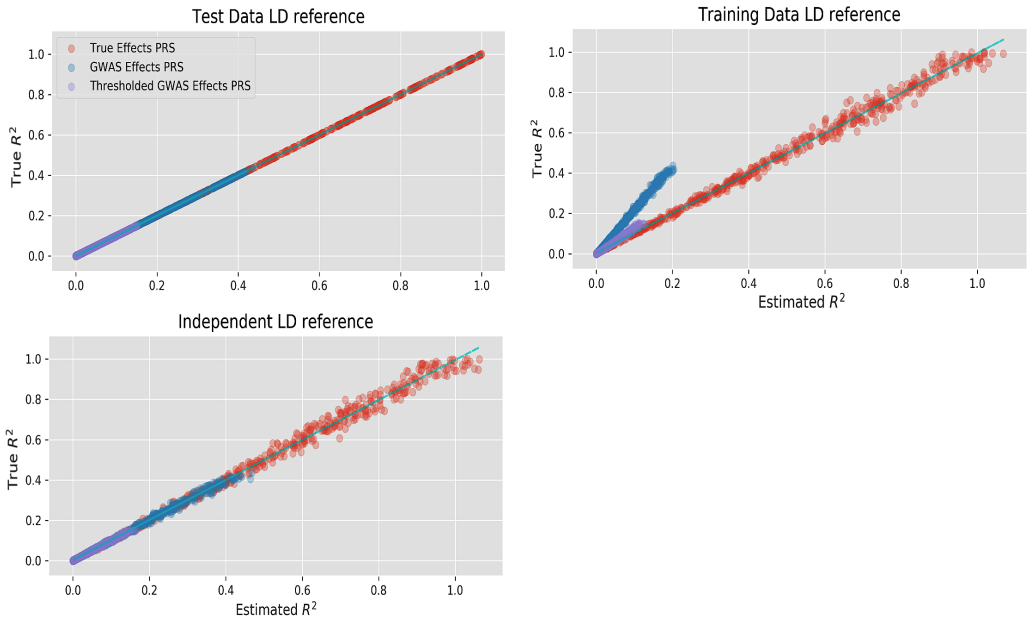

Figure S1: The impact of the LD reference on the PPB approach.