RápidoPGS: A rapid polygenic score calculator for summary GWAS data without validation dataset

Guillermo Reales[1,2], Martin Kelemen[1,3], Chris Wallace[1,2,4]

[1] Cambridge Institute of Therapeutic Immunology & Infectious Disease (CITIID), Jeffrey Cheah Biomedical Centre, Cambridge Biomedical Campus, University of Cambridge, Puddicombe Way, Cambridge CB2 0AW, UK.
[2] Department of Medicine, University of Cambridge School of Clinical Medicine, Cambridge Biomedical Campus, Cambridge, CB2 2QQ, UK.
[3] Wellcome Sanger Institute, Hinxton, Cambridgeshire, UK.
[4] MRC Biostatistics Unit, University of Cambridge, Cambridge Institute of Public Health, Forvie Site, Robinson Way, Cambridge Biomedical Campus, Cambridge, CB2 0SR, UK.

Correspondence: Dr. Chris Wallace (cew54@cam.ac.uk)

Abstract

Polygenic scores (PGS) aim to predict complex traits at an individual level based on genetic data. Computation of PGS is based on simply ascertained genome-wide association summary statistics but typically requires an independent test dataset to tune PGS parameters, and for the more sophisticated methods, the computation of LD matrices. Internally tuned methods have recently been proposed that obviate the need for a test dataset, but they remain computationally intensive to run.
Here we present RápidoPGS, a flexible and fast method to compute PGS without the need to compute LD-matrices, requiring only summary-level GWAS datasets.
Based on fine-mapping principles, RápidoPGS computes the posterior probability that each variant is causal, which in turn is used to shrink effect sizes adaptively as a function of LD and strength of association. We show by summary and individual-level validation that RápidoPGS performs well in comparison with another well-established internally-trained method (median AUC difference [RápidoPGS - LDpred2-auto] = -0.02205, range = [-0.1184, 0.0217]), with at least 50-fold improved speed. RápidoPGS is implemented in R, and can work with user-supplied summary statistics or download them directly from the GWAS catalog. We propose that RápidoPGS can be used to rapidly screen a set of candidate traits for utility, before more computationally intensive methods are applied to selected traits.

Main text

Genome-wide association studies (GWAS) have been widely successful at identifying a large number of genetic variants (usually single nucleotide polymorphisms, or SNPs), associated with a wide range of diseases and complex traits[1]. Most genetic variants have a small individual effect on the tested traits and thus have low predictive power[2–4]. However, simultaneously evaluating the effects of all common SNPs has the potential to explain much of the heritability for complex diseases and phenotypes[2,5].

Polygenic scores (PGS) estimate individual propensity to a phenotype (eg. disease) by a summing individuals genotypes (coded 0, 1, 2), weighted by their effect sizes, calculated from GWAS data

$$\sum_i \beta_i G_i$$

where β_i is the estimated effect for variant *i*, and G_i the genotype. Challenges relate to the nature of GWAS data, to determine the set of SNPs (i) to use. Linkage disequilibrium (LD) between variants, if ignored, would mean double-counting the effects of causal variants with multiple variants in high LD. Error in the estimated effect sizes is most pronounced for small effects, for which true association may not be distinguishable from noise about a null result. Different approaches have developed to deal with these. Initial approaches selected only the most strongly associated variant in each genome wide-significant peak, but it was soon realised that predictive accuracy could be improved by reducing the significance threshold, to capture more truly associated variants even at the cost of including some false associations (which should add noise, but not bias, to any prediction), and methods developed to select variants through using an external validation set to tune the significance parameter, as well as using automated LD-pruning algorithms to select independent variants rather than visual inspection of Manhattan plots[6].

More statistically sophisticated approaches have since been developed to average over multiple variants in LD without double counting (which is more accurate than selecting just one), and to shrink effect estimates by a continuous weight, w_i [7–9]

$$\sum_i w_i \beta_i G_i$$

As with the “prune and threshold” approach, such methods initially required external independent validation datasets to tune their parameters, which can present a further barrier. Recently, automated methods have been developed that tune their parameters on the same dataset used for training via hierarchical Bayesian models[7,9] which perform nearly as well as their externally tuned counterparts. However, the mathematics of dealing with LD while allowing the capturing of signals from nearby causal variants requires storage inversion of large matrices, which is computationally intensive, and tuning, internally or externally, adds a further layer of iterative computation. Some approaches mitigate this by adding a thinning step, discarding a random subset of SNPs to reduce the burden, but constructing PGS still takes hours.

PGS construction is related to fine mapping. If we knew exactly the causal variants for a trait, an optimal PGS would comprise the estimated effects of just those variants. Fine mapping methods estimate probabilities that a variant is causal given observed data, and so in the absence of knowing the exact set of true causal variants, a natural PGS might be constructed as above, with w_i set to probability variant *i* is causal. The basic fine mapping approach is very fast, because it makes a simplifying assumption that only a single causal may exist in any LD-defined region. While unrealistic, it has been shown to perform well, and

of relevance to PGS, allows the definition of fine mapping probabilities using predefined LD regions and without the burden of processing large LD matrices[10]. Distinguishing noise around null associations from true associations is dealt with by two parameters: a prior probability that a random SNP is causal, and the prior distribution of effect sizes at true causal variants. While in a PGS these would be open to tuning, in fine mapping sensible default values have been chosen that reflect existing knowledge from the breadth of GWAS studies already conducted[11].

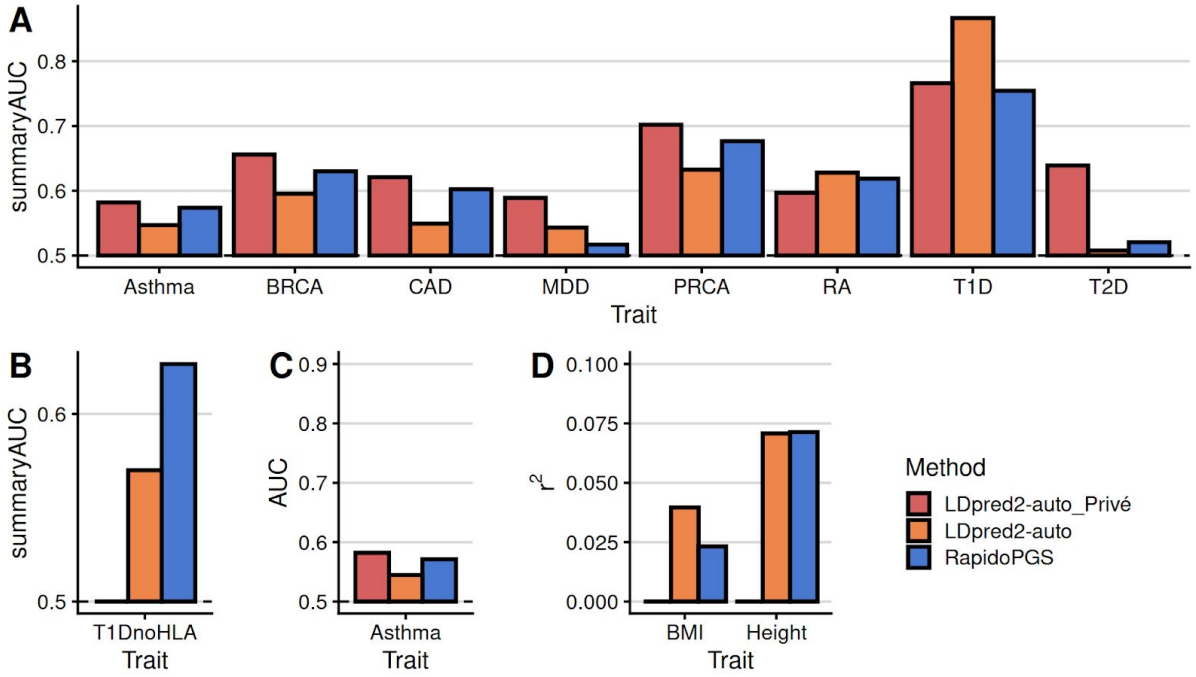
Here we present RápidoPGS, a lightweight and fast (*rápido*, in Spanish) method to compute PGS based on this basic fine mapping approach, that only requires a summary statistics dataset, and that can generate weights for millions of SNPs in only a few seconds. RápidoPGS works fully in R, requiring few dependencies. A PGS can be quickly constructed from any GWAS summary statistic dataset, or any GWAS PubMed ID, if harmonised datasets are available at GWAS catalog.

We used RápidoPGS to compute PGS models for eight case-control and two quantitative traits from publicly available GWAS summary statistics (Suppl. Table 1) and compared its performance to that of PGS generated using an automated tuning method, LDpred2-auto,[6] in GWAS data from large biobanks (UK Biobank and Finngen, Suppl. Table 1). Note that LDpred2-auto uses SNP thinning to limit computational burden, and we found that we could not raise the thinning parameter above 10% (i.e. discard 90% of SNPs at random) without segfault errors, thus all are comparisons are against this setting. For binary traits, we used summaryAUC[12], a method for PGS evaluation using summary statistics as validation datasets. SummaryAUC measures were similar for RápidoPGS and LDpred2-auto models (median difference [RápidoPGS - LDpred2-auto] = 0.0267, range = [-0.1123, 0.0566], Figure 1A), with the exception of type 1 diabetes (T1D). T1D has multiple very strong associations in the HLA region[12], which is well known for its extensive long range LD architecture[14]. We suspected this was the cause of the underperformance, and indeed, removing the HLA region (Chr6:20-40Mb) produced more similar (but lower) AUCs for the two methods (Figure 1B). This warns that RápidoPGS may underperform in the presence of multiple very strong signals in LD.

Both LDpred2-auto and RápidoPGS tended to underperform compared to the AUC values reported by Privé et al.[7]. We checked whether this was due to using summaryAUC, which suggests trimming to the strongest 20,000 SNPs for accuracy, but found individual-level estimates of AUC in UK Biobank gave similar results (Figure 1C), suggesting that the issue related to our inability to raise the LDpred2-auto thinning parameter above 0.1. The RápidoPGS AUC, even trimmed to 20k SNPs, was typically only a little lower than the reported AUC values (median difference [RápidoPGS - LDpred2-auto] = -0.02205, range = [-0.1184, 0.0217]).

For the quantitative traits, we used individual level data to calculate r^2 in UK Biobank, and again found similar performance for RápidoPGS compared to LDpred2-auto (Figure 1D).

Figure 1. RápidoPGS has comparable performance to LDpred2-auto. A. AUC for case-control traits in large biobank datasets (UKBB for all except FinnGen for CAD). AUC was estimated by summaryAUC on LDpred2-auto and RápidoPGS. summaryAUC suggests trimming to the top 20k SNPs for accuracy. For comparison purposes, we added AUC reported by Privé et al.[6]. **B.** Repeated on the T1D dataset with HLA region removed (Chr6:20-40Mb). No Privé results available for this subset. **C.** For asthma, we also computed AUC by individual genotypes in UKBB. **D.** r^2 results from BMI and height computed using individual-level data from UK Biobank. No Privé results available for these traits.

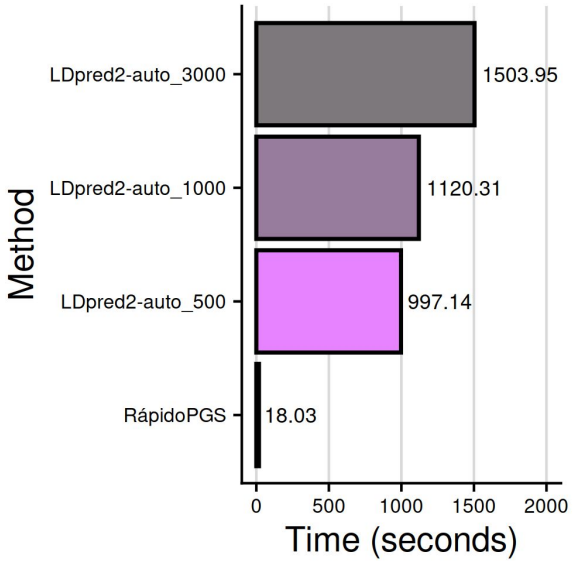


RápidoPGS includes an optional step to generate a sparse PGS, by identifying a subset of SNPs with largest weights and recalculating the score. We found that, in general, the larger the number of SNPs, the higher AUC, and post-threshold recalculation generally, but not always, improves performance (Suppl. Fig. 1). In general, 10^{-4}-thresholded (median 231,493 SNPs) with recalculation performed better than any other alternative, although in most cases 20k threshold (ie. 20,000 SNPs with largest absolute weights) performed almost as well, suggesting that smaller PGS, which may be easier to use in subsequent calculations, may be sufficient for the rapid screening we envisage RápidoPGS may be useful for.

We benchmarked the speed of RápidoPGS and LDpred2-auto, using repeated analysis of the asthma GWAS dataset[13]. Since the number of iterations may affect LDpred2 speed, we run LDpred2-auto using 500 (LDpred2-auto default), 1000, and 3000 (the number we used for our computations) iterations, and 15 cores. RápidoPGS does not require parallelisation nor high computational power and can be run on a regular laptop, although we ran it here for comparison on the same high performance computing system as LDpred2. RápidoPGS took 18 seconds to finish, while the fastest LDpred2-auto (500 iterations) took ~997 seconds. LDpred2-auto (3000 iterations) took ~1504 seconds to finish (Figure 2). RápidoPGS is a

standard R package, and does not require pre-installation or download of any external dataset (eg. a reference panel).

Figure 2. Timings of RápidoPGS and LDpred2-auto (with variable numbers of iterations), calculating PGS for Asthma.



Having PGS scores computed easily and quickly can be advantageous in a context in which there is a need for rapid assessment of many traits. Our work suggests that fine mapping approaches offer an alternative method to generate the weights w_i in the weighted PGS calculation. Fine mapping methods do require setting parameters (typically, as prior parameters describing the frequency and distribution of true causal effect sizes), but sensible values for these have been learnt from analysis of the many GWAS datasets to date, and thus avoids the need to tune for individual datasets, which may be an insurmountable barrier for rare traits where a pair of training and test GWAS do not exist. While we chose a simple fine mapping approach here, other approaches have been developed which work do not make the simplifying single causal variant assumption yet still work with reasonable speed on GWAS summary data[15], and these more sophisticated approaches may potentially increase the performance of fine-mapping PGS here to be competitive with PGS-inspired methods.

We conclude that RápidoPGS is 50 times or more faster than the most competitive automated PGS method with only a small loss in accuracy, suited for rapid screening of many potentially related traits, before maximising the accuracy of scores for specific traits of interest.

Declaration of interests

The authors declare no competing interests.

Acknowledgements

This work was supported by the Wellcome Trust (WT107881, 203950/Z/16/A) and the MRC (MC_UU_00002/4).

This research has been conducted using the UK Biobank Resource under Application Number 45669.

Web resources

R - https://cran.r-project.org/
RápidoPGS - https://github.com/GRealesM/RapidoPGS
LDpred2 - https://privefl.github.io/bigsnpr/articles/LDpred2.html
summaryAUC - https://github.com/lsncibb/SummaryAUC
Neale's lab (UKBB and FinnGen summary GWAS datasets) - http://www.nealelab.is/uk-biobank/
LD detect datasets - https://bitbucket.org/nygcresearch/ldetect-data/src
GWAS catalog - https://www.ebi.ac.uk/gwas/
1000 Genomes Project - https://www.internationalgenome.org/

Data and code availability

Summary statistics datasets used are publicly available in their respective publications (see Suppl. Table 1). Code used in the analyses is available at GitHub (https://github.com/GRealesM/RapidoPGS_paper).

Description of Supplemental Data

Supplemental data contains Supplementary Materials Methods, 1 figure, and 3 tables.

References

1. Buniello, A., MacArthur, J. a. L., Cerezo, M., Harris, L.W., Hayhurst, J., Malangone, C., McMahon, A., Morales, J., Mountjoy, E., Sollis, E., et al. (2019). The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. Nucleic Acids Res. *47*, D1005–D1012.
2. Yang, J., Benyamin, B., McEvoy, B.P., Gordon, S., Henders, A.K., Nyholt, D.R., Madden, P.A., Heath, A.C., Martin, N.G., Montgomery, G.W., et al. (2010). Common SNPs explain a large proportion of the heritability for human height. Nat. Genet. *42*, 565–569.
3. Dudbridge, F. (2013). Power and predictive accuracy of polygenic risk scores. PLoS Genet. *9*, e1003348.
4. International Schizophrenia Consortium, Purcell, S.M., Wray, N.R., Stone, J.L., Visscher, P.M., O'Donovan, M.C., Sullivan, P.F., and Sklar, P. (2009). Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. Nature *460*, 748–752.
5. Bulik-Sullivan, B.K., Loh, P.-R., Finucane, H.K., Ripke, S., Yang, J., Schizophrenia Working Group of the Psychiatric Genomics Consortium, Patterson, N., Daly, M.J., Price, A.L., and Neale, B.M. (2015). LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. Nat. Genet. *47*, 291–295.
6. Privé, F., Vilhjálmsson, B.J., Aschard, H., and Blum, M.G.B. (2019). Making the Most of Clumping and Thresholding for Polygenic Scores. Am. J. Hum. Genet. *105*, 1213–1221.
7. Privé, F., Arbel, J., and Vilhjálmsson, B.J. (2020). LDpred2: better, faster, stronger. BioRxiv 2020.04.28.066720.
8. Vilhjálmsson, B.J., Yang, J., Finucane, H.K., Gusev, A., Lindström, S., Ripke, S.,

Genovese, G., Loh, P.-R., Bhatia, G., Do, R., et al. (2015). Modeling Linkage Disequilibrium Increases Accuracy of Polygenic Risk Scores. Am. J. Hum. Genet. *97*, 576–592.
9. Ge, T., Chen, C.-Y., Ni, Y., Feng, Y.-C.A., and Smoller, J.W. (2019). Polygenic prediction via Bayesian regression and continuous shrinkage priors. Nat. Commun. *10*, 1776.
10. Maller, J.B., McVean, G., Byrnes, J., Vukcevic, D., Palin, K., Su, Z., Howson, J.M.M., Auton, A., Myers, S., Morris, A., et al. (2012). Bayesian refinement of association signals for 14 loci in 3 common diseases. Nat. Genet. *44*, 1294–1301.
11. Wallace, C. (2020). Eliciting priors and relaxing the single causal variant assumption in colocalisation analyses. PLOS Genet. *16*, e1008720.
12. Song, L., Liu, A., Shi, J., Gejman, P.V., Sanders, A.R., Duan, J., Cloninger, C.R., Svrakic, D.M., Buccola, N.G., Levinson, D.F., 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*, 4038–4044.
13. Demenais, F., Margaritte-Jeannin, P., Barnes, K.C., Cookson, W.O.C., Altmüller, J., Ang, W., Barr, R.G., Beaty, T.H., Becker, A.B., Beilby, J., et al. (2018). Multiancestry association study identifies new asthma risk loci that colocalize with immune-cell enhancer marks. Nat. Genet. *50*, 42–53.
14. Sanchez-Mazas, A., Djoulah, S., Busson, M., Le Monnier de Gouville, I., Poirier, J.-C., Dehay, C., Charron, D., Excoffier, L., Schneider, S., Langaney, A., et al. (2000). A linkage disequilibrium map of the MHC region based on the analysis of 14 loci haplotypes in 50 French families. Eur. J. Hum. Genet. *8*, 33–41.
15. Newcombe, P.J., Conti, D.V., and Richardson, S. (2016). JAM: A Scalable Bayesian Framework for Joint Analysis of Marginal SNP Effects. Genet. Epidemiol. *40*, 188–201.