

Doug Speed

Predictor Talk - PRS Challenge

Aarhus University, Denmark

CAGI7 SLIDE PACK GUIDELINES

I agree to my slides being used according to the following terms:

__Doug Speed__

- Authors: please include THIS slide at the beginning of your presentation to indicate that you agree with reuse of your slides in CAGI presentations, according to the terms below, and digitally sign by writing your name above.
- To disseminate the CAGI results, I encourage all CAGI participants to remix these CAGI slides for inclusion in their oral and poster presentations.
- CAGI participants may use or publish any content of these slides only in compliance with the CAGI Data Use Agreement.
- CAGI participants may not cite or publish any content of these slides except with the written permission of the originating author or in the primary CAGI publications.
- CAGI participants must acknowledge the original authors of these slides and CAGI7. Include the acknowledgement banner across the bottom.
- CAGI participants must include the CAGI credits slide in their presentations.
- Slides with a red slash (no remix) may not be included in a presentation unless all attendees are registered CAGI participants who signed the data use agreement.
- Slides with a red X over the entire slide may not be used in any circumstances.
- Slides with a blue slash (embargoed) may not be used unless the embargo is explicitly lifted.
- Slides without slash or X may be tweeted unless flagged with the "No tweet" icon.
- CAGI participants using these slides must explain the "No tweet" icon, which means that the slide content should not be disseminated outside the presenting conference hall. If the presentation venue permits pervasive tweeting, you may not include these slides in your talk.
- Note for slide authors: we expect all CAGI participants to abide by these restrictions and will make best effort to ensure they are followed. However, bear in mind that we have limited means of enforcing them, and therefore the restrictions cannot be guaranteed.



Please can I borrow five minutes from my next talk :)
(I checked with Ryan)

My Software is Called LDAK

DougSpeed.com

The Home of LDAK

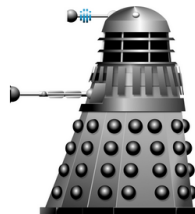
[Downloads](#) [Advice](#) [Association Testing](#) [Prediction](#) [REML / HE / PCGC](#) [SumHer](#) [More Features](#) [Other Details](#) [Heritability Model](#) [Help](#)

LDAK

Hi, my name is Doug Speed, and I am a researcher at the [Centre for Quantitative Genetics and Genomics](#) at Aarhus University. However, enough about me! This website exists to host my software LDAK, and I hope you find some of it useful.

LDAK is a software package for analysing association study data. In total, it has over 30 different functions (see [Main Arguments](#) for a full list). However, most likely, you will be interested in one of the following:

- > Using [LDAK-KVIK](#) to test SNPs and genes for association with a phenotype.
- > Using [MegaPRS](#) to construct polygenic scores from GWAS summary statistics.
- > Estimating SNP heritability by [analysing individual-level data](#).



LDAK: it's just better

My Software is Called LDAK

From 2012-2020, I focused on estimating SNP heritability
started with LDAK, a more accurate version of GCTA
then added SumHer, a more accurate version of LDSC

In 2021 added MegaPRS (currently working on MegaPRS2)

In 2024 added LDAK-KVIK

a tool for mixed-model association testing
can analyze quantitative and binary traits
copes with relatedness, structure and biobank-sized datasets

LDAK-KVIK has power of BOLT-LMM, but speed of REGENIE

My Software is Called LDAK

From 2012-2020, I focused on estimating SNP heritability
started with LDAK, a more accurate version of GCTA
then added SumHer, a more accurate version of LDSC

In 2021 added MegaPRS (currently working on MegaPRS2)

In 2024 added LDAK-KVIK

a tool for mixed-model association testing
can analyze quantitative and binary traits
copes with relatedness, structure and biobank-sized datasets

LDAK-KVIK has power of BOLT-LMM, but speed of REGENIE

[nature](#) > [nature genetics](#) > [articles](#) > article

Article | Published: 11 August 2025

LDAK-KVIK performs fast and powerful mixed-model association analysis of quantitative and binary phenotypes

[Jasper P. Hof](#) & [Doug Speed](#) 

[Nature Genetics](#) **57**, 2116–2123 (2025) | [Cite this article](#)

2883 Accesses | **3** Citations | **2** Altmetric | [Metrics](#)

Abstract

Mixed-model association analysis (MMAA) is the preferred tool for performing genome-wide association studies. However, existing MMAA tools often have long runtimes and high memory requirements. Here we present LDAK-KVIK, an MMAA tool for analysis of

MegaPRS (Original Version)

PRS take the form $P = X_1\beta_1 + X_2\beta_2 + \dots + X_m\beta_m$

where X_j contains genotypes for SNP j and β_j is its effect size

There are tens of different PRS tools

mainly differ based on prior distributions for β_j

The fundamental flaw of most PRS tools

is that their priors assume that all SNPs are equally important

I refer to this as assuming the “Uniform Heritability Model”

(each SNP expected to contribute equal heritability)

MegaPRS (Original Version)

PRS take the form $P = X_1\beta_1 + X_2\beta_2 + \dots + X_m\beta_m$

where X_j contains genotypes for SNP j and β_j is its effect size

There are tens of different PRS tools

mainly differ based on prior distributions for β_j

The fundamental flaw of most PRS tools

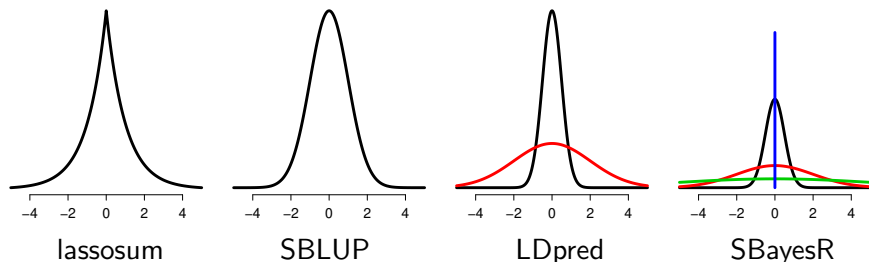
is that their priors assume that all SNPs are equally important

I refer to this as assuming the “Uniform Heritability Model”

(each SNP expected to contribute equal heritability)

MegaPRS (Original Version)

E.g., here are priors used by major PRS tools (in 2021)



Each tool assumes all SNPs are expected to contribute equal heritability
i.e., that $\mathbb{E}[h_j^2] = \mathbb{E}[\beta_j^2]$ is constant

This is sub-optimal, because we know that $\mathbb{E}[h_j^2]$ varies across genome
e.g., real data shows $\mathbb{E}[h_j^2]$ correlates with minor allele frequency

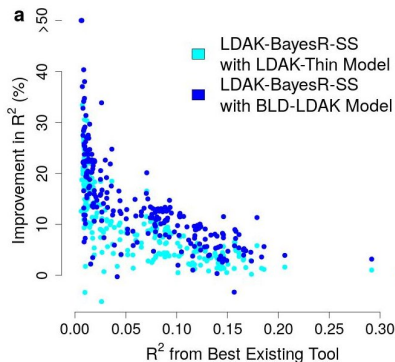
MegaPRS (Original Version)

MegaPRS contains new versions of lassosum, SBLUP, LDpred, SBayesR
that allow the user to specify the heritability model

We recommended using the MegaPRS version of SBayesR
and assuming the BLD-LDAK heritability model

Demonstrated MegaPRS using data from UK Biobank
considered 225 phenotypes (mixture of quantitative and binary)
average sample size was 280 000 individuals

MegaPRS (Original Version)



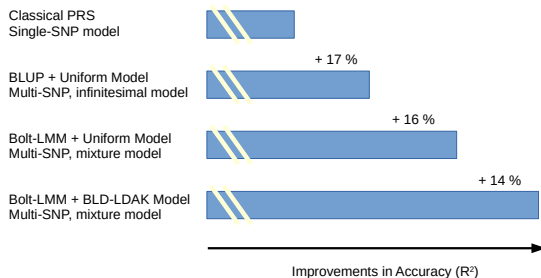
MegaPRS outperformed all existing tools for 223 out of 225 phenotypes

average increase in R^2 was 14%

equivalent to increasing sample size by one quarter

MegaPRS (Original Version)

I have described this as the “Third Major Advance” in PRS Accuracy



The improvement from using better a heritability model was similar to switching from Classical PRS to advanced tools or to switching from basic priors to mixture priors

New MegaPRS

New MegaPRS can analyze results from multiple GWAS

(whereas Original MegaPRS only uses results from one GWAS)

The GWAS results can be from different traits †

therefore, can include results for highly-correlated traits

The GWAS results can be from different ancestries †

therefore, can include results from multi-ancestry GWAS

The extra methodology is very similar to that used by MTAG

weights secondary traits based on correlation with target trait

† MTBLUP (PRS-CSx) also allow for multiple traits (ancestries)

Advantage 1 - Constructing PRS for Diseases

It is generally hard to construct accurate disease PRS
because diseases are binary and tend to have low heritability

Can improve accuracy by adding in GWAS results for correlated traits
ideally these additional traits should be easier to predict
(i.e., be quantitative and have high heritability)

E.g., for obesity, can use information from BMI
while for diabetes, can use information from glucose

Advantage 2 - PRS for Under-Represented Ancestries

PRS typically perform much worse for African and Asian datasets
reflecting that most GWAS use European individuals

However, we observe high genetic correlations across ancestries
so it should be possible to borrow information across ancestries
(although note that correlations tend to be lower for diseases)

Running New MegaPRS

Requires GWAS results - these are generally provided by the user
but LDK website provides results for 51 quantitative traits
that can be “directly plugged” into MegaPRS
(MegaPRS will decide which to use)

Summary statistics from LDK-KVIK

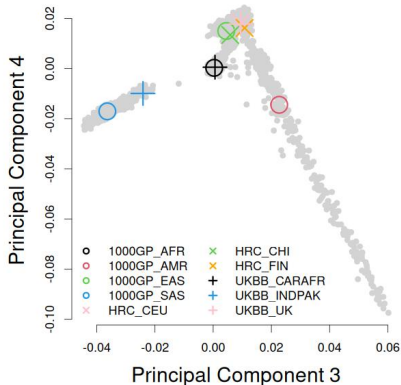
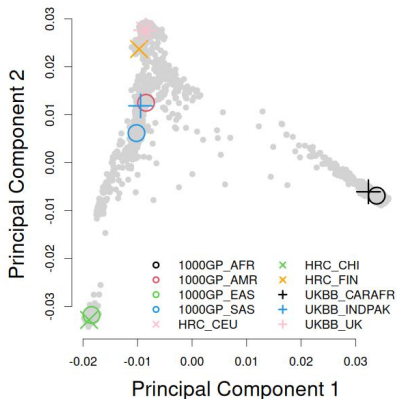
On this web page, we present the GWAS results of 62 quantitative traits from the UK Biobank using LDK-KVIK.

Index	Name	Field_ID	Sample_Size	h2_SNP	Num_Sig	Num_Clump	PRS_R2	Manhattan	Summary_Stats	PGS_Weights
1	Standing Height	50	458079	0.668	386876	5256	0.417	Manhattan	Summary Stats	PGS Weights
2	Mean Platelet Thrombocyte Volume	30100	445360	0.486	203778	3003	0.369	Manhattan	Summary Stats	PGS Weights
3	Platelet Count	30080	445365	0.405	156457	2226	0.247	Manhattan	Summary Stats	PGS Weights
4	Basal Metabolic Rate	23105	451092	0.365	139234	1835	0.196	Manhattan	Summary Stats	PGS Weights
5	HDL Cholesterol	30760	400615	0.356	84921	1262	0.222	Manhattan	Summary Stats	PGS Weights
6	Triglycerides	30870	437378	0.331	61474	790	0.133	Manhattan	Summary Stats	PGS Weights

Running New MegaPRS

Require reference panel or LD-matrix for each ancestry considered

LDAK website provides matrices for AFR, AMR, EUR, EAS and SAS



CAGI PRS Challenge - Helped by Jasper Hof

I started with legacy GWAS results (those provided by CAGI)
then searched for other GWAS

Main sources of GWAS results

Million Veterans Program (MVP) - EUR GWAS have $n \approx 400\text{ k}$

FinnGen (FG) - have $n \approx 450\text{ k}$

UK Biobank (UKBB) - have $n \approx 370\text{ k}$

Biobank Japan (BJ) - have $n \approx 80\text{ k}$

Note that MVP, FG and UKBB are population biobanks

(so diseases have low prevalences, and not carefully defined)

CAGI PRS Challenge - Helped by Jasper Hof

Breast Cancer - legacy GWAS has 120 k cases, newer version has 136 k added in FG (24 k), UKBB (17 k) and BJ (6 k)

Coronary Artery - legacy GWAS has 66 k cases, newer version has 190 k † added in MVP (124 k), FG (63 k), BJ (32 k)

IBD - legacy GWAS have 12 k or 25 k cases added in MVP (8 k), FG (11 k), UKBB (7 k)

Type 2 Diabetes - legacy GWAS has 27 k cases, newer version has 242 k † added in Glycated Haemoglobin (HbA1c) from UKBB ($n = 340\text{ k}$)

† GWAS were meta-analyses that included MGB (test dataset)

CAGI PRS Challenge - Helped by Jasper Hof

I'm giving a presentation because MegaPRS performed well

Would be interesting to know how much this is due to methods
how much due to increased sample size
and how much due to cheating

(I guess mainly the latter)

Final Slide

Thanks to Jasper Hof who helped test the CAGI models

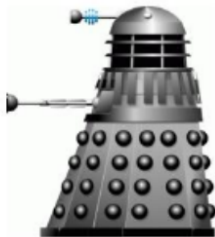
Funding from:

Independent Reseach Fund Denmark

EU Horizon 2020

Aarhus University Research Foundation

Lundbeck Foundation



www.ldak.org

If you are interested in visiting my group in Aarhus, please contact me