Doug Speed

Developing a Polygenic Risk Score Contest

Aarhus University, Denmark

CAGI 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 CAGI 6. 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, but bear in mind that we have limited means of enforcing them, and therefore the restrictions cannot be guaranteed.

Permission request regarding CAGI ** recordings and hybrid presentations

I understand that my presentation as associated with CAGI ** will be recorded and broadcasted via the internet through CAGI's designated technology provider.

As the copyright holder to this presentation entitled Developing a Polygenic Risk Score Contest, I hereby agree that the copyright for the online broadcast and recording of my presentation be released to CAGI and my recording be uploaded on to the CAGI website.

Overview of Talk

Topic 1 - Me

Topic 2 - You

About Me

Statistical geneticist

develop tools for analyzing GWAS data

Creator of LDAK, contains state-of-the-art tools for:

heritability analysis from individual-level data (LDAK)

heritability analysis from summary statistics (SumHer)

mixed-model association testing

gene-based association testing

making polygenic risk scores (PRS)

I "created" the heritability model

The Heritability Model

Informal definition

Most analyses of SNP data require assumptions regarding the expected importance of SNPs - the heritability model describes these assumptions

Mathematical definition

Suppose we assume the linear model $\mathbb{E}[Y] = X_1\beta_1 + X_2\beta_2 + \ldots + X_m\beta_m$

where Y denotes the phenotype and X_j denote the jth SNP

The expected heritability explained by SNP j is $\mathbb{E}[h_j^2] = \frac{\mathbb{E}[\beta_j^2]Var(X_j)}{Var(Y)}$

The heritability model specifies $\mathbb{E}[h_j^2]$ for each SNP

The Heritability Model

Informal definition

Most analyses of SNP data require assumptions regarding the expected importance of SNPs - the heritability model describes these assumptions

Mathematical definition

Suppose we assume the linear model $\mathbb{E}[Y] = X_1\beta_1 + X_2\beta_2 + \ldots + X_m\beta_m$

where Y denotes the phenotype and X_j denote the jth SNP

The expected heritability explained by SNP j is $\mathbb{E}[h_j^2] = \dfrac{\mathbb{E}[\beta_j^2]Var(X_j)}{Var(Y)}$

The heritability model specifies $\mathbb{E}[h_j^2]$ for each SNP

The Heritability Model

Most analyses in human statistical genetics assume $\mathbb{E}[h_j^2]$ is constant

all SNPs are expected to contribute equally

I call this the Uniform Model (aka the GCTA Model)

This model is used by the heritability softwares LDSC and MTG2

and by the association softwares Fast-LMM and Bolt-LMM

and by the prediction softwares BayesR, LDpred and lassosum

and by the genetic architecture softwares GENESIS and AI-Mixer

and by my first software, Sparse Partitioning

and is recommended by Tibshirani for lasso and ridge regression

etc, etc.

The Heritability Model

Most analyses in human statistical genetics assume $\mathbb{E}[h_j^2]$ is constant

all SNPs are expected to contribute equally

I call this the Uniform Model (aka the GCTA Model)

This model is used by the heritability softwares LDSC and MTG2

and by the association softwares Fast-LMM and Bolt-LMM

and by the prediction softwares BayesR, LDpred and lassosum

and by the genetic architecture softwares GENESIS and AI-Mixer

and by my first software, Sparse Partitioning

and is recommended by Tibshirani for lasso and ridge regression

etc, etc.

The Heritability Model

The Uniform Model arises whenever a software first standardizes SNPs

i.e., scales X_j so that $Var(X_j) = 1$

most software do not acknowledge (realise) they use this model

but this model is sub-optimal

The Uniform Model predicts that $\mathbb{E}[h_j^2]$ is independent of MAF

but we do not observe this on real data

instead we find that $\mathbb{E}[h_j^2]$ tends to increase with MAF

(it also depends on levels of LD, conservation and SNP function)

I recommend the Alpha -0.25 Model (aka the LDAK-Thin Model)

(or if super enthusiastic, the BLD-LDAK Model)

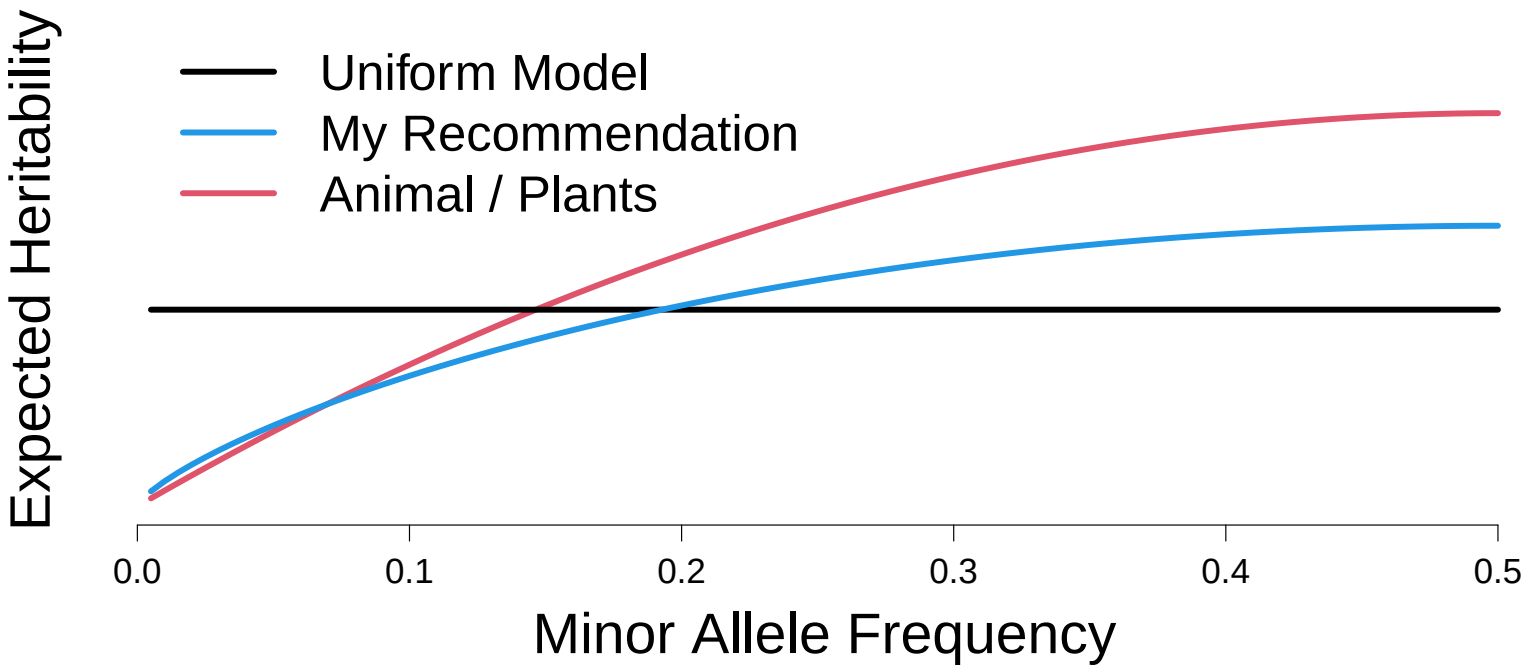
Relationship between $\mathbb{E}[h_j^2]$ and MAF

Consider models of the form $\mathbb{E}[h_j^2] \propto (p_j(1-p_j))^{1+\alpha}$, where p_j is MAF

Uniform Model obtained by setting $\alpha = -1$

In animal and plant genetics, $\alpha = 0$ is common (e.g., van Radan)

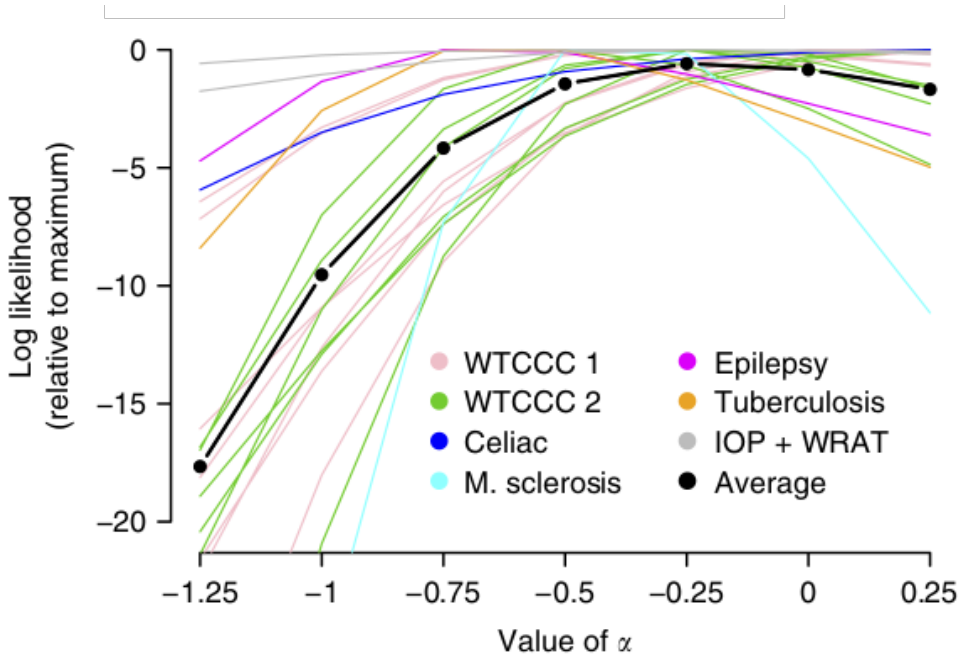
The Alpha -0.25 Model sets $\alpha = -0.25$

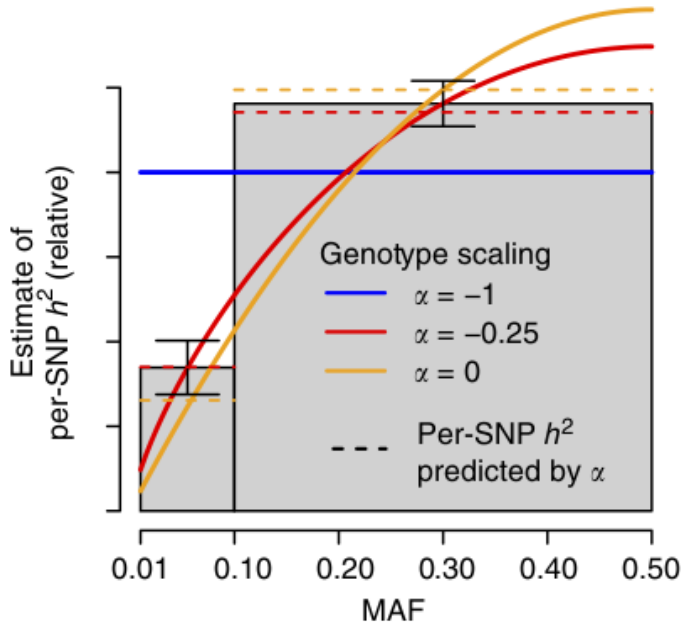


Relationship between $\mathbb{E}[h_j^2]$ and MAF

Determined best α by analyzing lots of human traits

and finding value that resulted in best model fit (likelihood)





Consistently find that $\alpha = -0.25$ fits data better than $\alpha = -1$

The Impact of the Heritability Model

The model has a big impact when estimating SNP heritability

estimates of h^2_{SNP} about 40% higher with Alpha -0.25 Model

The model has a big impact when estimating heritability enrichments

with Uniform Model, DHS estimated to explain 80% of h^2_{SNP}

with Alpha -0.25 Model, they are estimated to explain only 25%

The model has a limited impact when estimating genetic correlations

The model has a big impact when constructing PRS (next slides)

The Impact of the Heritability Model

The model has a big impact when estimating SNP heritability

estimates of h^2_{SNP} about 40% higher with Alpha -0.25 Model

The model has a big impact when estimating heritability enrichments

with Uniform Model, DHS estimated to explain 80% of h^2_{SNP}

with Alpha -0.25 Model, they are estimated to explain only 25%

The model has a limited impact when estimating genetic correlations

The model has a big impact when constructing PRS (next slides)

The Impact of the Heritability Model

The model has a big impact when estimating SNP heritability

- estimates of h^2_{SNP} about 40% higher with Alpha -0.25 Model

The model has a big impact when estimating heritability enrichments

- with Uniform Model, DHS estimated to explain 80% of h^2_{SNP}
- with Alpha -0.25 Model, they are estimated to explain only 25%

The model has a limited impact when estimating genetic correlations

The model has a big impact when constructing PRS (next slides)

The Impact of the Heritability Model

The model has a big impact when estimating SNP heritability

estimates of h^2_{SNP} about 40% higher with Alpha -0.25 Model

The model has a big impact when estimating heritability enrichments

with Uniform Model, DHS estimated to explain 80% of h^2_{SNP}

with Alpha -0.25 Model, they are estimated to explain only 25%

The model has a limited impact when estimating genetic correlations

The model has a big impact when constructing PRS (next slides)

Polygenic Risk Scores

PRS are genetic prediction models

estimate the genetic contribution toward's an individual's phenotype

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

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

i.e., a linear SNP-based prediction model

There are 100s of different tools for constructing PRS

some require individual-level data, some use summary statistics

vary based on their prior distribution for β_j

vary based on their algorithms (e.g., MCMC or Variational Bayes)

Polygenic Risk Scores

One way to measure the accuracy of a PRS is using R^2

the squared correlation between predicted and observed phenotypes

R^2 is between zero and heritability of trait

at present, most PRS have low R^2 (e.g., between 0.01 and 0.10)

Some people ask what is the point of this?

they say that only PRS with $R^2 > 0.2$ are useful

Some people think personalized medicine means we can accurately predict which individuals (from the population) will develop a particular diseases

this is not the case (and probably never will be)

Polygenic Risk Scores

One way to measure the accuracy of a PRS is using R^2

the squared correlation between predicted and observed phenotypes

R^2 is between zero and heritability of trait

at present, most PRS have low R^2 (e.g., between 0.01 and 0.10)

Some people ask what is the point of this?

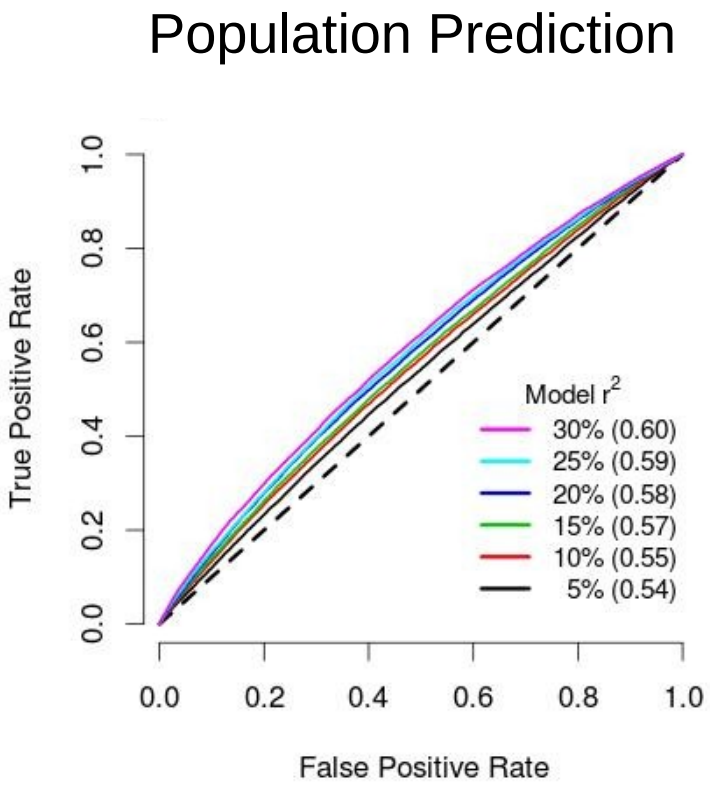
they say that only PRS with $R^2 > 0.2$ are useful

Some people think personalized medicine means we can accurately predict which individuals (from the population) will develop a particular diseases

this is not the case (and probably never will be)

Population vs Subset Prediction

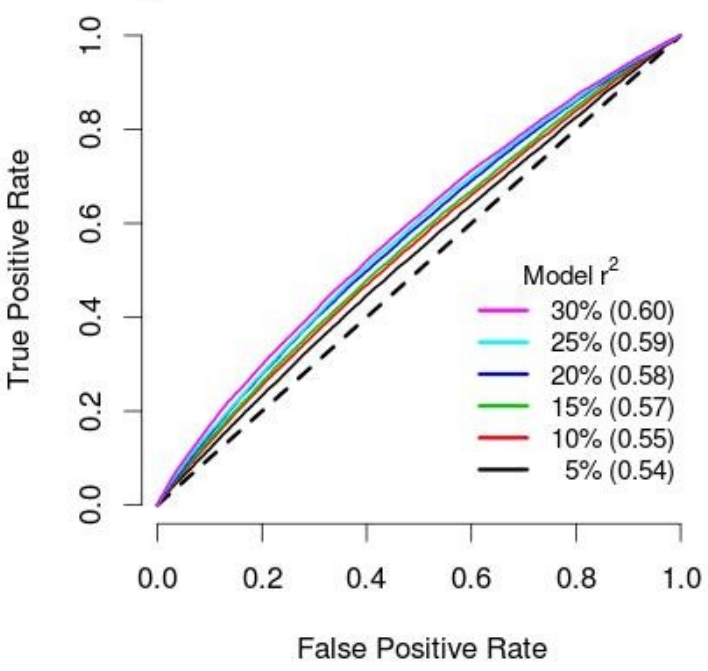
Prevalence of epilepsy is 0.005, so accurately predicting which 0.5% of individuals in the population will develop the disease is very hard



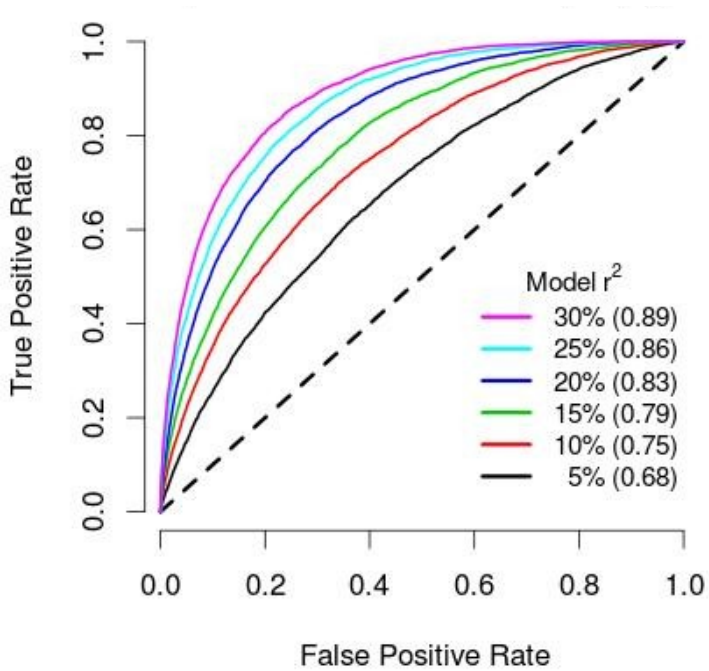
Population vs Subset Prediction

Prevalence of epilepsy is 0.005, so accurately predicting which 0.5% of individuals in the population will develop the disease is very hard

Population Prediction



Single-Seizure Prediction



But for individuals who experience a first seizure, prevalence is 0.5, so prediction becomes much easier / more effective.

Other Uses of PRS

Can improve our understanding of genetic architecture

below are three examples - sorry no time to describe

A PRS with $R^2 = 0.07$ showed that individuals with chronic schizophrenia had higher-than-average genetic liability to schizophrenia. High loading of polygenic risk in chronic schizophrenia. Molecular Psychiatry, 2016

A PRS with $R^2 = 0.02$ identified clinically-defined subtypes of autism that have significantly different genetic liabilities. Identification of common genetic risk variants for autism spectrum disorder. Nature Genetics, 2019.

PRS with $R^2 < 0.05$ demonstrated risk of developing emotional problems is moderated by an interaction between environmental sensitivity and type of parenting. Keers et al. Psychotherapy and psychosomatics, 2016.

My PRS Tools

I have developed about ten different PRS tools

2015 - Created MultiBLUP - was good at the time :)

2021 - Created three individual-level data and five summary statistic tools

ind-level - Ridge-Predict, Bolt-Predict, BayesR-Predict

sum-stats - Lasso-SS, Ridge-SS, Bolt-SS, SBayesR-SS, MegaPRS

2022 - Created QuickPRS (a lite version of MegaPRS)

2023 - Upgraded the recent tools

now simplified - require one or three commands (instead of five)

added new prior distribution (Elastic-Predict and Elastic-SS)

2021 PRS Tools

I started by examining eight of the most commonly-used PRS tools

Individual-level data:

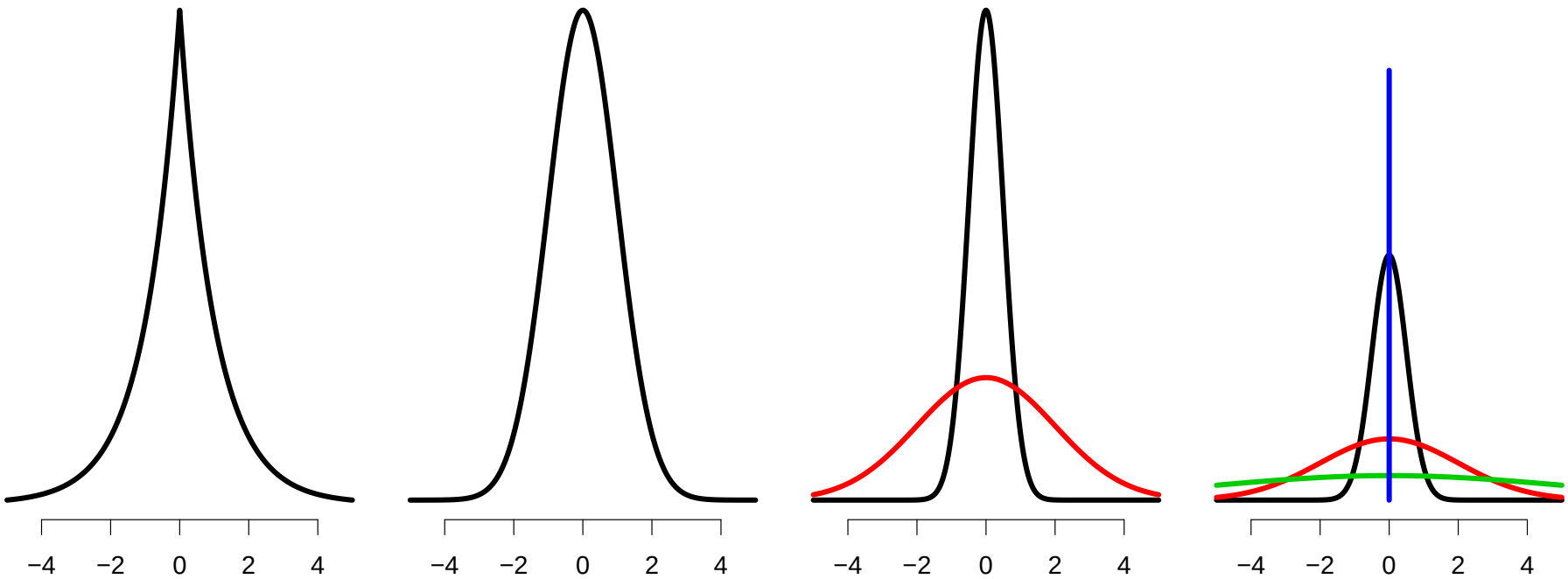
Lasso, Ridge Regression, Bolt-LMM and BayesR

Summary statistics:

lassosum, sBLUP, LDpred and SBayesR

2021 PRS Tools

These tools use a variety of prior distribution forms for effect sizes



Lasso / lassosum | Ridge / SBLUP | Bolt / LDpred | BayesR / SBayesR

However, they all assume the Uniform Model

Eight New Prediction Tools

I created generalized versions of the eight existing PRS tools

my versions allow the user to specify the heritability model

(Florian Privé created the generalized version of Lasso)

I also made the new versions more computationally efficient

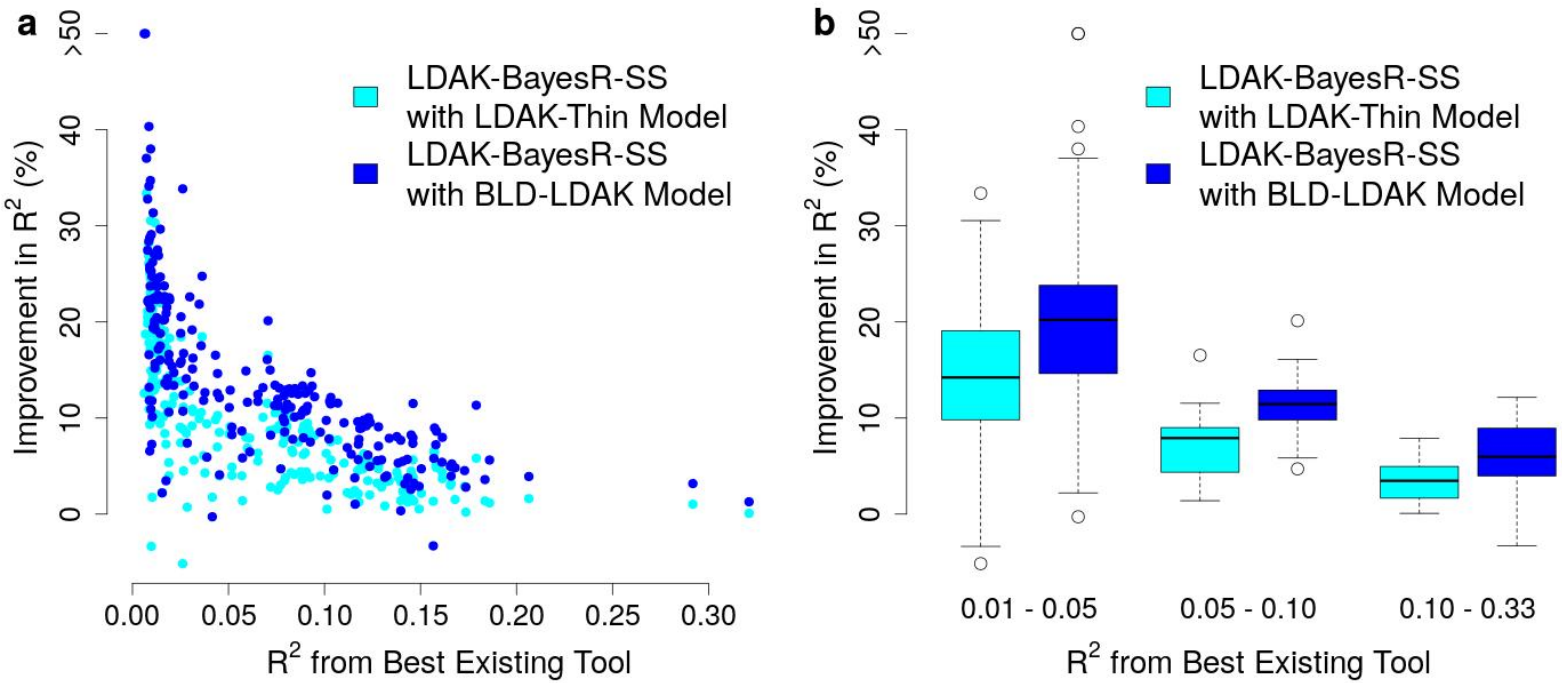
e.g., my version of BayesR was 50 times faster than original

summary statistic tools run within an hour (now within 15 minutes)

My versions are more robust

mainly due to switching from MCMC to Variational Bayes

Tested Tools on 225 Phenotypes from UK Biobank



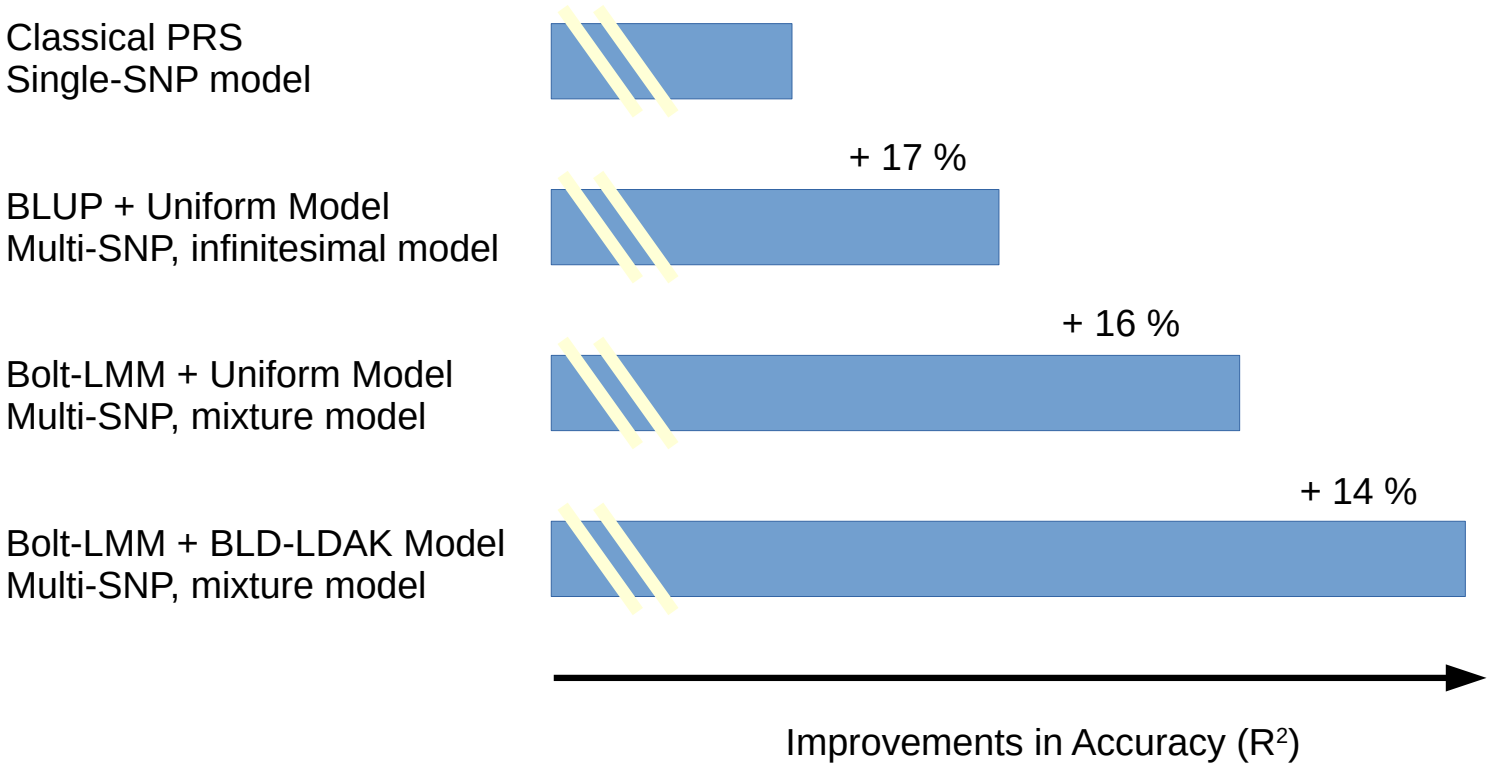
My new versions almost always out-performed the eight existing versions

higher R^2 for either 223 or 225 phenotypes (depending on tool)

average increase in R^2 was 14%

equivalent to increasing sample size by one quarter

This is the "Third Major Advance" in PRS Accuracy



Improvement from changing the heritability model is

similar to switching from Classical PRS to advanced tools

or to switching from basic priors to mixture priors

My Involvement in CAGI

I was invited to take part in the PRS challenge in CAGI 6 (thanks Sung)

- my software MegaPRS was one of the leading methods

I hope the PRS challenge appears in future editions of CAGI

- I have some ideas on how to improve the fairness
- how to improve the reliability
- and how to increase participation

Please note, that what Sung and Shamil organized last year was wonderful

1 - Improving Fairness

The PRS challenge might benefit from more rules (or categories)

Example A - exclusion of problematic SNPs

e.g., rare SNPs or those with ambiguous alleles (A/T, C/G)

otherwise entrants are required to guess what to do with these

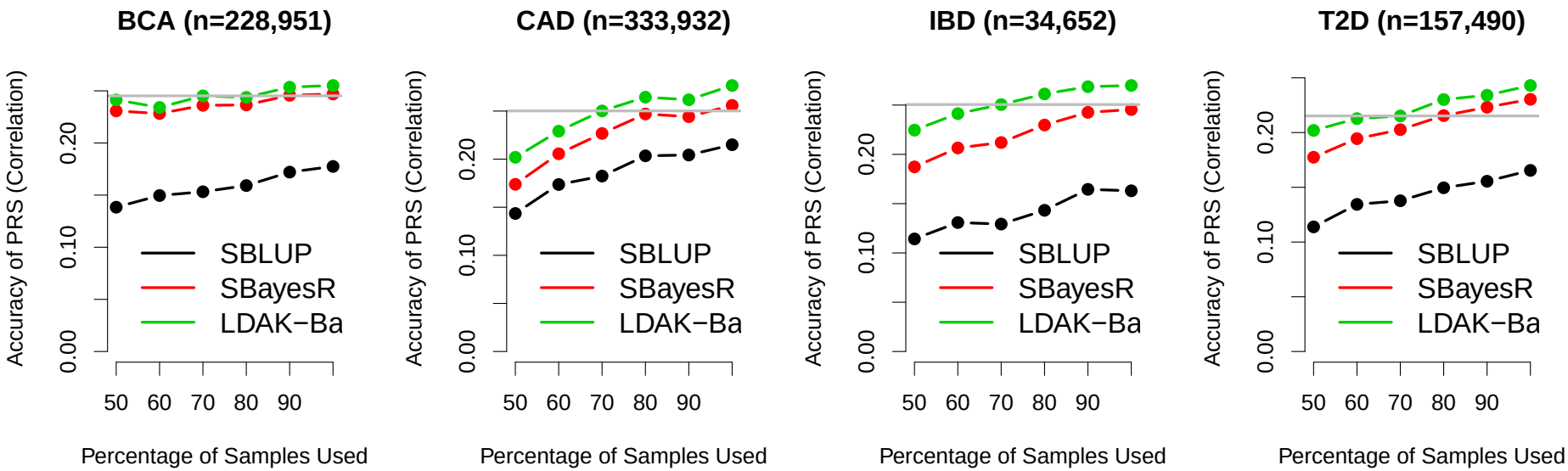
1 - Improving Fairness

The PRS challenge might benefit from more rules (or categories)

Example B - standardization of source GWAS

CAGI suggested four GWAS, but entrants were free to use others

otherwise, accuracy mainly reflects who has the largest sample size



2 - Improving Reliability

The PRS challenge would benefit from having more than four traits

The main limitation is availability of test data (used to measure PRS)

CAGI 6 used the Mass General Brigham (MGB) Biobank

that included four (suitable) phenotypes

Note that the test data should be relatively hard-to-access

e.g., it would be pointless using UK Biobank

3 - Improving Participation

I believe approximately 17 people registered for the CAGI 6 PRS challenge but only four or five submitted entries?

I think the main obstacles were lack of validation data and fear

Validation Data

A summary statistic PRS tool requires GWAS results (training data)

- however, many also require individual-level validation data
- this is used for deciding suitable parameter values

For example, most PRS tools require an estimate of h^2_{SNP}

- use training data to make PRS for different values of h^2_{SNP}
- then use validation data to see which PRS has highest accuracy

Note that recent PRS tools often avoid this by using "auto-tuning"

My software MegaPRS avoids this by using "psuedo cross-validation"

- found a way to "divide" summary statistics into training and test

Validation Data

A summary statistic PRS tool requires GWAS results (training data)

- however, many also require individual-level validation data
- this is used for deciding suitable parameter values

For example, most PRS tools require an estimate of h^2_{SNP}

- use training data to make PRS for different values of h^2_{SNP}
- then use validation data to see which PRS has highest accuracy

Note that recent PRS tools often avoid this by using "auto-tuning"

My software MegaPRS avoids this by using "psuedo cross-validation"

- found a way to "divide" summary statistics into training and test

Fear

Firstly, there is the risk of your tool being outperformed

or that your success will be ignored

Secondly, there is the risk that your PRS is hopeless

e.g., PRS tools are often sensitive to high-LD regions

tools that use MCMC can encounter convergence issues

(commom to see in method papers, never in applications)

Thirdly, the challenge requires trust in the organizers

e.g., that they can fairly assess PRS

which requires that they have performed adequate QG

and have accurately constructed (described) the phenotypes

How to Increase Participation

We DO NOT REQUIRE individual-level data to measure PRS accuracy

we can compute R^2 using summary stats (and a reference panel)

I propose that CAGI releases summary statistics from a subset of test data

(e.g., from 1000 individuals, or 10% of the test individuals)

Entrants can use these summary statistics as a validation dataset

and to reduce their fear of entering

(e.g., they can be confident their PRS performs OK)

Measuring PRS Accuracy from Summary Statistics

Want to compute $Cor(P, Y)$, where $P = \sum X_j \beta_j$

let $p_j = Cor(X_j, Y)$

$$Cor(P, Y) = \frac{Cov(P, Y)}{SD(P)SD(Y)} = \frac{\sum_j \beta_j Cov(X_j, Y)}{SD(P)SD(Y)} = \frac{\sum_j \rho_j SD(X_j)}{SD(P)}$$

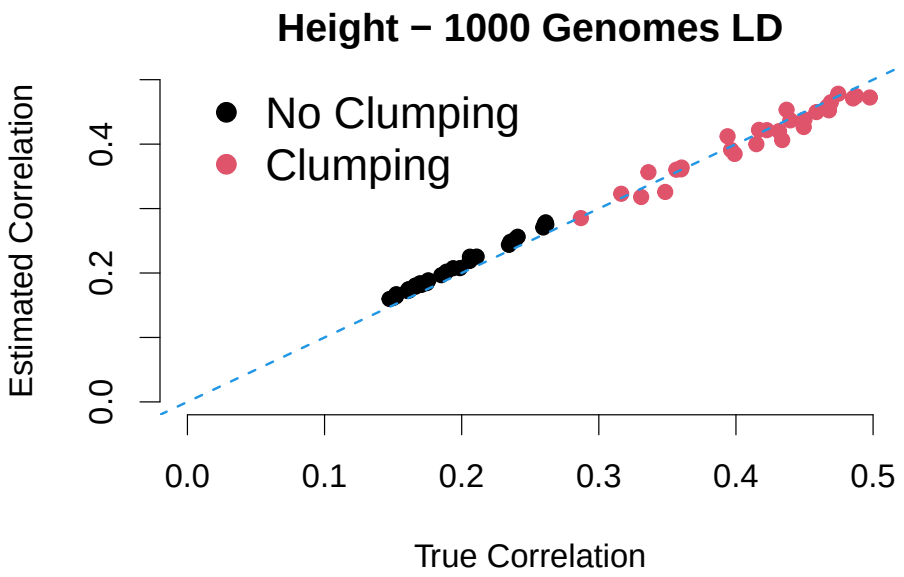
Can estimate $SD(P)$ and $SD(X_j)$ using the reference panel

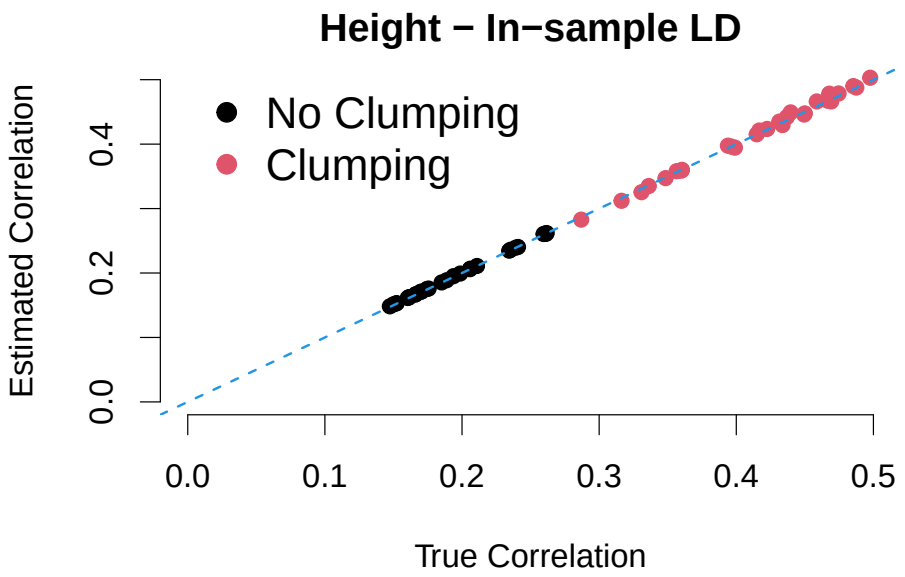
and recover p_j from the summary statistics ($\chi^2(1)$ statistic is np_j^2)

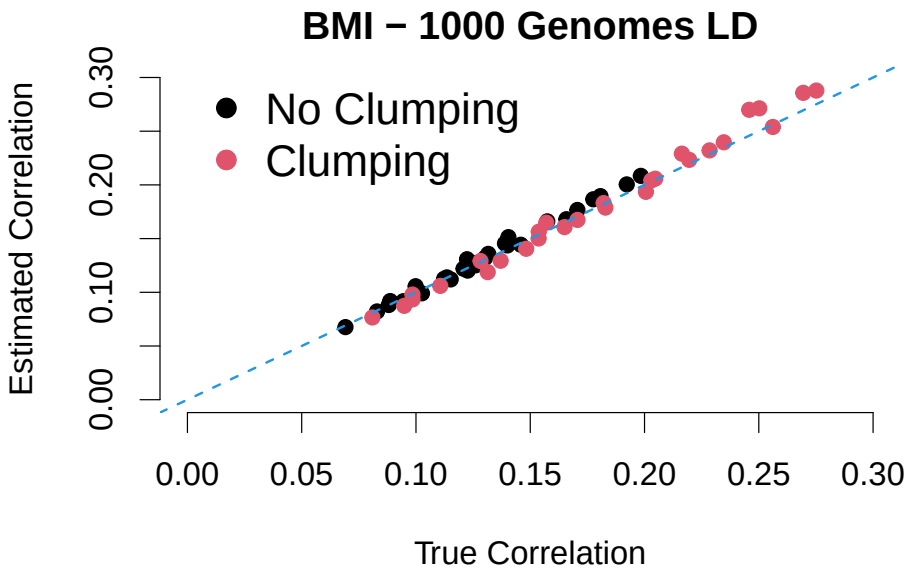
This approach requires access to a reference panel (probably OK)

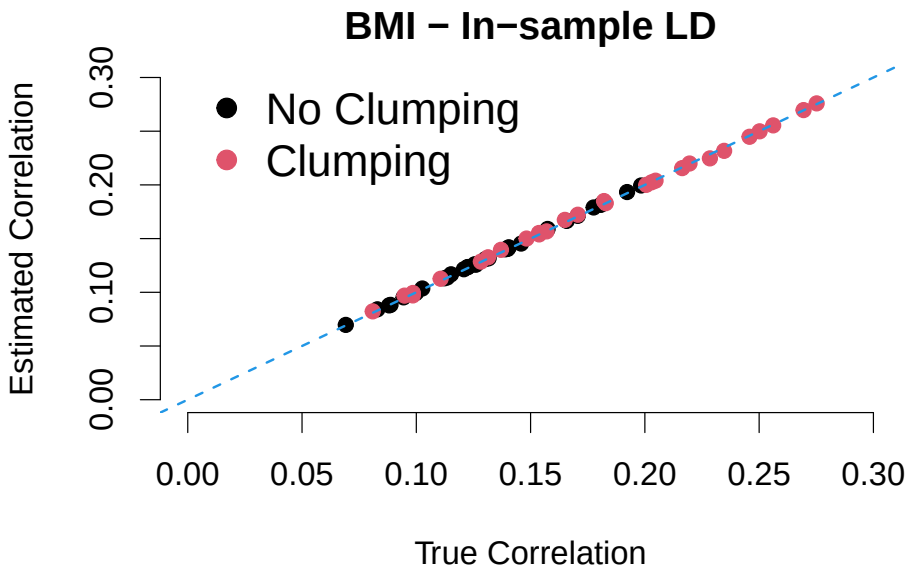
but could simplify further by releasing an LD-matrix

Idea works well for Height and BMI









Summary of Proposed Changes

1 - I suggest a few (minor) changes to rules

2 - I propose CAGI releases summary statistics from subset of test data

would also need to explain to particpants how to use these

(non-trivial, but I think OK)

If successful, I think this could increase participation

makes entry easier for those with methods that require validation

ensures PRS are tuned based on the exact phenotype

I think this would also enable testing of more traits

(although would likely need at least one more test dataset)

Final Slide

LDAK software is available at www.ldak.org

contains the new PRS tools plus much more

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, please contact me