

Doug Speed

The Future of CAGI

Just PRS right? Or do I get to decide everything?

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.



Overview

This talk is a word soup

a collection of random thoughts about PRS

presented without much conviction

hopefully my British accent helps ...

Reminder - PRS are linear SNP-based predictors of traits (diseases)

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

the aim of a PRS tool is to produce accurate estimates of β_j

This talk is a word soup

a collection of random thoughts about PRS

presented without much conviction

hopefully my British accent helps ...

Reminder - PRS are linear SNP-based predictors of traits (diseases)

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

the aim of a PRS tool is to produce accurate estimates of β_j

What is the Point of a PRS Challenge?

1 - compare the performance of PRS tools?

Comparing PRS tools is relatively easy (compared to other tools)
there is not much tool diversity

The **theoretical accuracy** of a PRS tool depends mainly on three factors

- the choice of prior distribution for β_j
- the choice of heritability model
- the algorithm used to estimate the posterior distribution of β_j

What is the Point of a PRS Challenge?

1 - compare the performance of PRS tools?

Comparing PRS tools is relatively easy (compared to other tools)
there is not much tool diversity

The theoretical accuracy of a PRS tool depends mainly on three factors

- the choice of prior distribution for β_j
- the choice of heritability model
- the algorithm used to estimate the posterior distribution of β_j

What is the Point of a PRS Challenge?

If the aim is simply to compare different PRS tools

then I think I could do a better job than CAGI on my own

I think Sung could also do better than CAGI

Florian, Tian, Jian (or other PRS tool creators) could each do better

Each of us would be better, because we “could” perform a fair comparison

i.e., test tools using exactly the same data / SNPs / quality control

Problem - each of us has a conflict of interest

Solution - give the job to Oliver Pain (Kings College London)

who created the (independent) GenoPred pipeline

What is the Point of a PRS Challenge?

If the aim is simply to compare different PRS tools

then I think I could do a better job than CAGI on my own

I think Sung could also do better than CAGI

Florian, Tian, Jian (or other PRS tool creators) could each do better

Each of us would be better, because we “could” perform a fair comparison

i.e., test tools using exactly the same data / SNPs / quality control

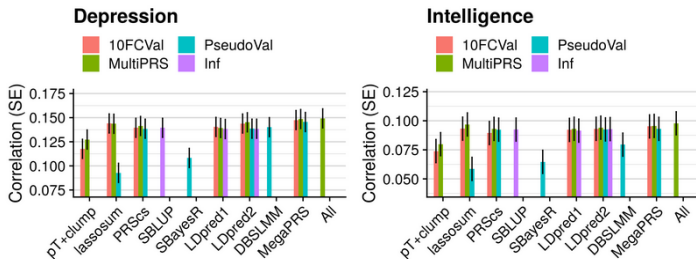
Problem - each of us has a conflict of interest

Solution - give the job to Oliver Pain (Kings College London)

who created the (independent) GenoPred pipeline

What is the Point of a PRS Challenge?

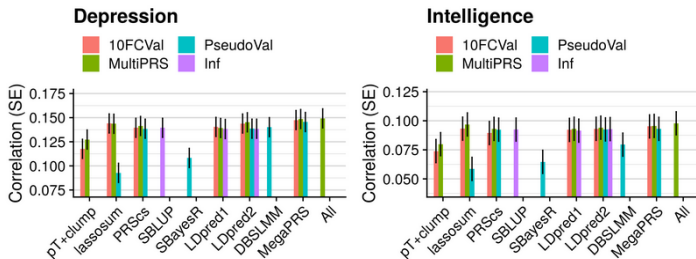
I expect the results would be very similar to those from GenoPred
i.e., the leading methods would have very similar performance



If you considered enough traits (and looked at ratios)
you could find significant differences
that would likely reflect the true ranking of PRS tools

What is the Point of a PRS Challenge?

I expect the results would be very similar to those from GenoPred
i.e., the leading methods would have very similar performance



If you considered enough traits (and looked at ratios)
you could find significant differences
that would likely reflect the true ranking of PRS tools

What is the Point of a PRS Challenge?

1 - compare the theoretical performance of PRS tools?

limited value - unlikely to get any surprises

2 - encourage methodological insights?

perhaps harder for PRS than other challenges

3 - increase awareness of the leading methods?

requires CAGI to become more well-known among PRS users

4 - many of us work better when we have a deadline / target

gave me an excuse to work this weekend

What is the Point of a PRS Challenge?

1 - compare the theoretical performance of PRS tools?

limited value - unlikely to get any surprises

2 - encourage methodological insights?

perhaps harder for PRS than other challenges

3 - increase awareness of the leading methods?

requires CAGI to become more well-known among PRS users

4 - many of us work better when we have a deadline / target

gave me an excuse to work this weekend

What is the Point of a PRS Challenge?

1 - compare the theoretical performance of PRS tools?

limited value - unlikely to get any surprises

2 - encourage methodological insights?

perhaps harder for PRS than other challenges

3 - increase awareness of the leading methods?

requires CAGI to become more well-known among PRS users

4 - many of us work better when we have a deadline / target

gave me an excuse to work this weekend

What is the Point of a PRS Challenge?

1 - compare the theoretical performance of PRS tools?

limited value - unlikely to get any surprises

2 - encourage methodological insights?

perhaps harder for PRS than other challenges

3 - increase awareness of the leading methods?

requires CAGI to become more well-known among PRS users

4 - many of us work better when we have a deadline / target

gave me an excuse to work this weekend

What is the Point of a PRS Challenge?

5 - encourage us to think more about the utility of PRS?

Most of us know the ultimate value of PRS (it's on the CAGI website)

“Polygenic risk scores (PRS) have potential clinical utility for risk surveillance, prevention and personalized medicine”

However, we are a long way off achieving this

main current use of PRS is investigating genetic architecture

Feels like the other challenges have more immediate value

(maybe I am biased, because I poorly understand other challenges)

What is the Point of a PRS Challenge?

6 - test the robustness / usability of PRS tools?

While PRS tools tend to have similar theoretical accuracies
their actual accuracies can be very different

E.g., a PRS tool fails if GWAS results and LD-matrix are not allele-aligned
or if the user performs inadequate quality control
while MCMC-based tools can sometimes “explode”

Should we forbid PRS tool developers from entering?
then the challenge results would reflect usability

What is **NOT** the Point of a PRS Challenge?

7 - test the googling skills of participants

MegaPRS did well, partly because I found the best data

If DECODE or 23&Me had entered
they probably would have won all contests

Results from the IIBDGC GWAS Mega2 GWAS are currently embargoed
anyone with access would have aced the IBD contest

What is **NOT** the Point of a PRS Challenge?

7 - test the googling skills of participants

MegaPRS did well, partly because I found the best data

If DECODE or 23&Me had entered
they probably would have won all contests

Results from the IIBDGC GWAS Mega2 GWAS are currently embargoed
anyone with access would have aced the IBD contest

New Collection of Thoughts



How Should We Measure the Performance of PRS?

1 - Accuracy (e.g., R^2 , AUC, proportion of cases in top percentile)

Note that subset-based measures more relevant for personalized medicine
population-measures more relevant for genetic architecture
(fortunately, there tends to be high concordance between two)

2 - Ideally, would also measure how much it costs to make them
actually making a PRS takes less than one day
but learning how to use a PRS tool can take months

How Should We Measure the Performance of PRS?

1 - Accuracy (e.g., R^2 , AUC, proportion of cases in top percentile)

Note that subset-based measures more relevant for personalized medicine
population-measures more relevant for genetic architecture
(fortunately, there tends to be high concordance between two)

2 - Ideally, would also measure how much it costs to make them
actually making a PRS takes less than one day
but learning how to use a PRS tool can take months

How Might We Improve PRS?

Will we develop better prior distributions?

yes, but improvement likely to be small (especially for large n)

Will we develop better heritability models?

yes, but improvement likely to be small (especially for large n)

Note that every SNP annotation database is a heritability model

e.g., CADD / FATHMM / EVE models > Uniform Model

(but best models use tens of annotations plus inter-genic SNPs)

Will we develop better algorithms for estimating the posterior?

yes, but improvement likely to be small

How Might We Improve PRS?

Will we develop better prior distributions?

yes, but improvement likely to be small (especially for large n)

Will we develop better heritability models?

yes, but improvement likely to be small (especially for large n)

Note that every SNP annotation database is a heritability model

e.g., CADD / FATHMM / EVE models $>$ Uniform Model

(but best models use tens of annotations plus inter-genic SNPs)

Will we develop better algorithms for estimating the posterior?

yes, but improvement likely to be small

How Might We Improve PRS?

Will we develop better prior distributions?

yes, but improvement likely to be small (especially for large n)

Will we develop better heritability models?

yes, but improvement likely to be small (especially for large n)

Note that every SNP annotation database is a heritability model

e.g., CADD / FATHMM / EVE models $>$ Uniform Model

(but best models use tens of annotations plus inter-genic SNPs)

Will we develop better algorithms for estimating the posterior?

yes, but improvement likely to be small

How Might We Improve PRS?

Will machine learning / AI help?

I think by definition, the answer is no

(caveat - I would have never predicted ChatGPT is possible)

The benefit of machine learning / AI comes from non-linear models

however, PRS are defined as linear SNP-based models

(might be some value in ensemble models)

Should we move from PRS to more general prediction models?

really hard to envisage how this would work

the assessment would probably kill Sung

How Might We Improve PRS?

Will machine learning / AI help?

I think by definition, the answer is no

(caveat - I would have never predicted ChatGPT is possible)

The benefit of machine learning / AI comes from non-linear models

however, PRS are defined as linear SNP-based models

(might be some value in ensemble models)

Should we move from PRS to more general prediction models?

really hard to envisage how this would work

the assessment would probably kill Sung

How Will We Actually Improve PRS?

1 - Make PRS tools easier to use (less likely to break)

Would help if we standardized the format of GWAS results

GWAS catalog tried, but has not been a success

2 - Make SNP weightings publicly available

PRS catalog exists (although overwhelmed with results)

plus sub-optimal, because weightings depend on SNP subset

3 - Make PRS publicly available

UK Biobank let users download PRS

much faster with small chance of failure (although PRS out-of-date)

How Will We Actually Improve PRS?

1 - Make PRS tools easier to use (less likely to break)

Would help if we standardized the format of GWAS results

GWAS catalog tried, but has not been a success

2 - Make SNP weightings publicly available

PRS catalog exists (although overwhelmed with results)

plus sub-optimal, because weightings depend on SNP subset

3 - Make PRS publicly available

UK Biobank let users download PRS

much faster with small chance of failure (although PRS out-of-date)

How Will We Actually Improve PRS?

1 - Make PRS tools easier to use (less likely to break)

Would help if we standardized the format of GWAS results

GWAS catalog tried, but has not been a success

2 - Make SNP weightings publicly available

PRS catalog exists (although overwhelmed with results)

plus sub-optimal, because weightings depend on SNP subset

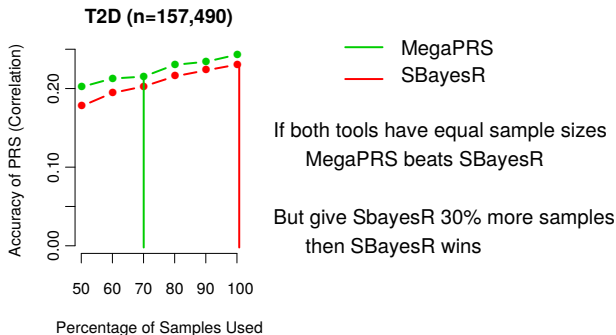
3 - Make PRS publicly available

UK Biobank let users download PRS

much faster with small chance of failure (although PRS out-of-date)

How Will We Actually Improve PRS?

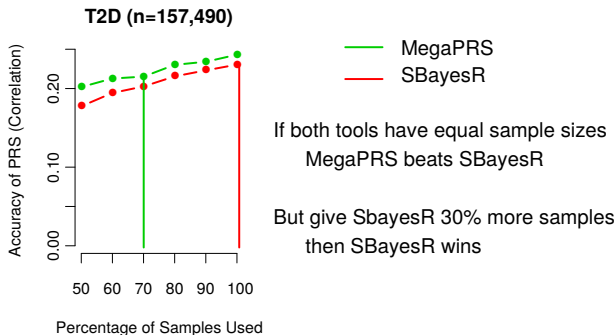
4 - Fastest way to improve PRS is to increase sample size (n)



Can increase n directly or indirectly (i.e., different traits / ancestries)
expect we will find better ways to integrate data sources
(e.g., MegaPRS > PRS-CSx, but unlikely to be optimal)

How Will We Actually Improve PRS?

4 - Fastest way to improve PRS is to increase sample size (n)



Can increase n directly or indirectly (i.e., different traits / ancestries)
expect we will find better ways to integrate data sources
(e.g., MegaPRS > PRS-CSx, but unlikely to be optimal)

How Will We Actually Improve PRS?

- 5 - Improve accuracy of PRS for Non-European ancestries
 - substantial room for methodological development
 - currently a “hot topic” (but therefore exploited)

Should CAGI have a non-European PRS challenge?

would require a multi-ancestry test dataset?

to reliably measure accuracy, requires ~ 5 k samples

What value would this CAGI challenge add to the PRS world?

CAGI's strength is motivating and bringing people together

and having people like Dean remind us why we get paid to research

how do we get participation from the (small) PRS community?

How Will We Actually Improve PRS?

- 5 - Improve accuracy of PRS for Non-European ancestries
 - substantial room for methodological development
 - currently a “hot topic” (but therefore exploited)

Should CAGI have a non-European PRS challenge?

would require a multi-ancestry test dataset?

to reliably measure accuracy, requires ~ 5 k samples

What value would this CAGI challenge add to the PRS world?

CAGI's strength is motivating and bringing people together

and having people like Dean remind us why we get paid to research

how do we get participation from the (small) PRS community?

How Will We Actually Improve PRS?

- 5 - Improve accuracy of PRS for Non-European ancestries
 - substantial room for methodological development
 - currently a “hot topic” (but therefore exploited)

Should CAGI have a non-European PRS challenge?

- would require a multi-ancestry test dataset?

- to reliably measure accuracy, requires ~ 5 k samples

What value would this CAGI challenge add to the PRS world?

- CAGI's strength is motivating and bringing people together

- and having people like Dean remind us why we get paid to research

- how do we get participation from the (small) PRS community?

New Collection of Thoughts



Doug's 2023 Suggestion for the PRS Challenge

I suggested the challenge should become very strict
e.g., tell participants which GWAS results to use

Would maximize chance that the best PRS tools came top

I no longer think this is a good idea

to achieve this aim, would need many more traits
as explained, we already know which tools are best

New Suggestions Part A (sorry, quite technical)

Here are some no-brainers

simple ways to make challenge fairer / easier for participants

- 1 - Remove ambiguous SNPs (alleles A & T or C & G) from bim file
- 2 - Provide minor allele frequencies (MAF)
- 3 - Provide both hg19 and hg38 genomic positions

New Suggestions Part B (also quite technical)

We should also make it easier for the organizers / assessors

1 - Require that participants submit finished models

would this disadvantage machine learning tools even more?

2 - Currently participants submit SNP weightings (estimates of β_j)

Sung uses PLINK (?) to compute PRS (calculate $\sum X_j \beta_j$)

he mentioned that this command sometimes fails

We can help by providing fake test genetic dataset

(e.g., fam / bed file for ten individuals)

participants can check PLINK command runs before submitting

New Suggestions Part B (also quite technical)

We should also make it easier for the organizers / assessors

1 - Require that participants submit finished models

would this disadvantage machine learning tools even more?

2 - Currently participants submit SNP weightings (estimates of β_j)

Sung uses PLINK (?) to compute PRS (calculate $\sum X_j \beta_j$)

he mentioned that this command sometimes fails

We can help by providing fake test genetic dataset

(e.g., fam / bed file for ten individuals)

participants can check PLINK command runs before submitting

New Suggestions Part B (also quite technical)

We should also make it easier for the organizers / assessors

1 - Require that participants submit finished models

would this disadvantage machine learning tools even more?

2 - Currently participants submit SNP weightings (estimates of β_j)

Sung uses PLINK (?) to compute PRS (calculate $\sum X_j \beta_j$)

he mentioned that this command sometimes fails

We can help by providing fake test genetic dataset

(e.g., fam / bed file for ten individuals)

participants can check PLINK command runs before submitting

New Suggestions Part C

1 - Get more traits

2 - Get new test dataset

CAGI 7 Challenge suffered due to data leakage
published GWAS for CAD and T2D included test dataset (MGB)

3 - Maybe use Danish dataset (e.g., Copenhagen Hospital Biobank?)
or AllofUs (cloud-based, so not trivial to use)

4 - Require that participants announce their GWAS in the forum?

New Suggestions Part D (quite crazy)

Could release GWAS results

- from single-SNP analysis of disease in test dataset

- non-sensitive - total size would be less than 100 Mb

Could also release a block-based LD-matrix

- could use Berisa and Pickrell LD blocks

- non-sensitive (?) - total size would be maybe 300 Mb

Then participants could perform the assessment themselves (almost)

New Suggestions Part E

Could make the CAGI PRS challenge continuous

E.g., participants can submit any time
entries would be assessed every 3 months (?)

CAGI becomes a resource for the community

CAGI (and the assessor) would get authorship + publicity

Acknowledgments

This talk is my fault, and my fault only

Last Slide

- 1 - What is the aim of the PRS challenge?
to test PRS tools or test data sources?
- 2 - Why does PRS challenge struggle compared to other challenges?
because it is new; due to size / reluctance of PRS community
- 3 - Simple changes that would help both participants and assessor
ambiguous SNPs, MAF, genomic positions, take test data
- 4 - Should we get new data?
new test data (Danish?), more traits, non-European data?
- 5 - Make the challenge continuous?