# Translating Polygenic Risk Scores for Medicine & Public Health

Frederick J. Boehm

March 19, 2024

# My background

# Education and Training

- M.D., University of Wisconsin-Madison, 2007
- M.S. (Population Health Sciences), University of Wisconsin-Madison, 2007
- Postdoctoral training (Statistical genetics), University of Washington, 2009
- Ph.D. (Statistics), University of Wisconsin-Madison, 2019
- Postdoctoral training (Systems Genetics), University of Massachusetts Medical School, 2021
- Postdoctoral training (Biostatistics & Cardiovascular Medicine), University of Michigan, 2024

# Research Interests

- Statistical genetics & genomics
- Statistics & data science education
- Causal inference in observational studies (including genetics studies)
- Efficient & scalable statistical methods for big data
- Clinical & public health applications of genetics and genomics
- Reproducible research & open science
- Collaboration & team science

# Introduction

# Overview

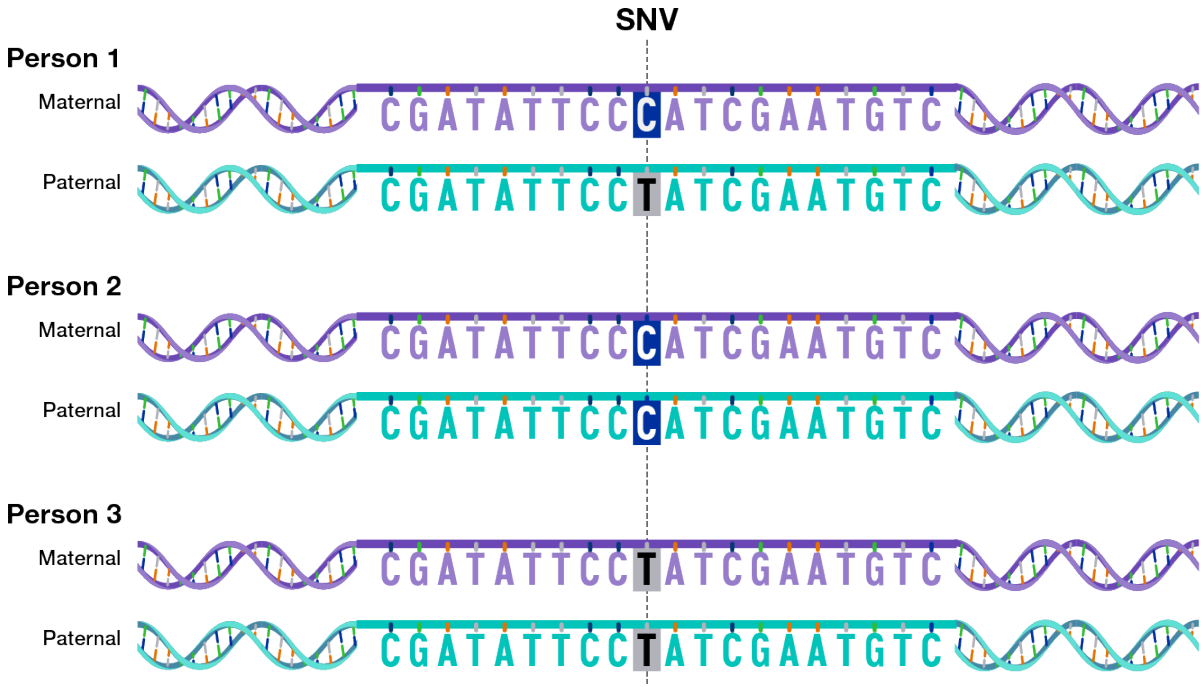
- Polygenic scores predict a single person's genetic risk for a specific trait or disease
  - Used for risk stratification
- Genotypes at genetic markers (SNPs) from across the genome are used to calculate PGS
- Statistical properties of PGS are relatively underexplored, despite their impact on risk stratifications
- Examining variability in PGS point estimates is an initial step in understanding statistical properties & performance of PGS

# Human Genetics Advances

- Working Draft of Human Genome Sequence, 2000
- First Genome-wide Association Study (GWAS), 2005
  - Age-related macular degeneration: 96 cases & 50 controls, 105,980 SNPs (Klein et al. 2005)
- Rare variant association tests, ~ 2009 to 2012

# Single Nucleotide Polymorphisms (SNPs)

- Genetic sequence differences between two individuals - at a single DNA position



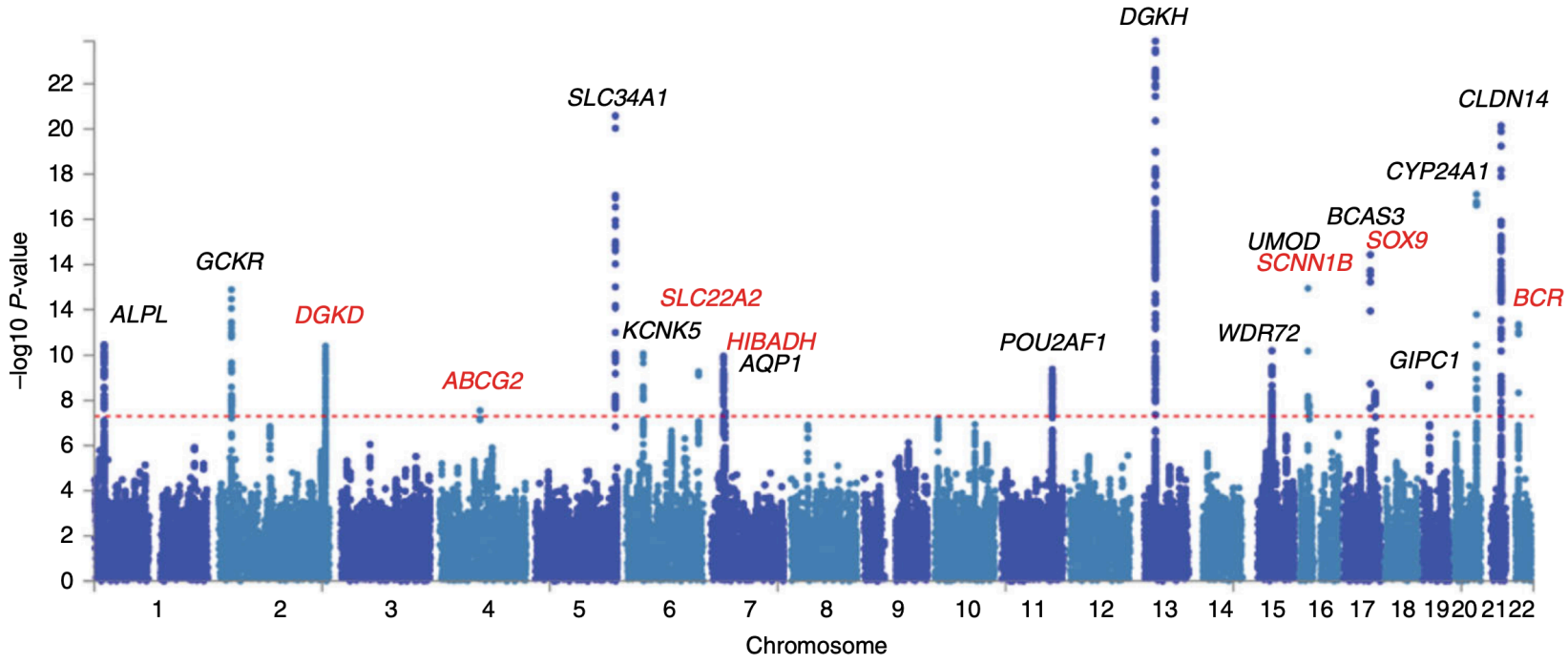

https://www.genome.gov/sites/default/files/inline-images/Genomic%20variation_SNV.png

# Genome-wide Association Studies (GWAS)

- GWAS are used to identify genetic variants associated with a trait (W. T. C. C. Consortium 2007)
- Millions of genetic variants, one at a time, are tested for association with a trait (W. T. C. C. Consortium 2007; Uffelmann et al. 2021)

$$\underset{n\times 1}{\text{Trait values}} = \underset{n\times 1}{G}\, b_G + \underset{n\times p_c}{C}\ \underset{p_C\times 1}{B_C} + \underset{n\times 1}{\epsilon}$$

# Genome-wide Association Studies (GWAS)




Kidney stone disease GWAS from Howles et al. (2019)

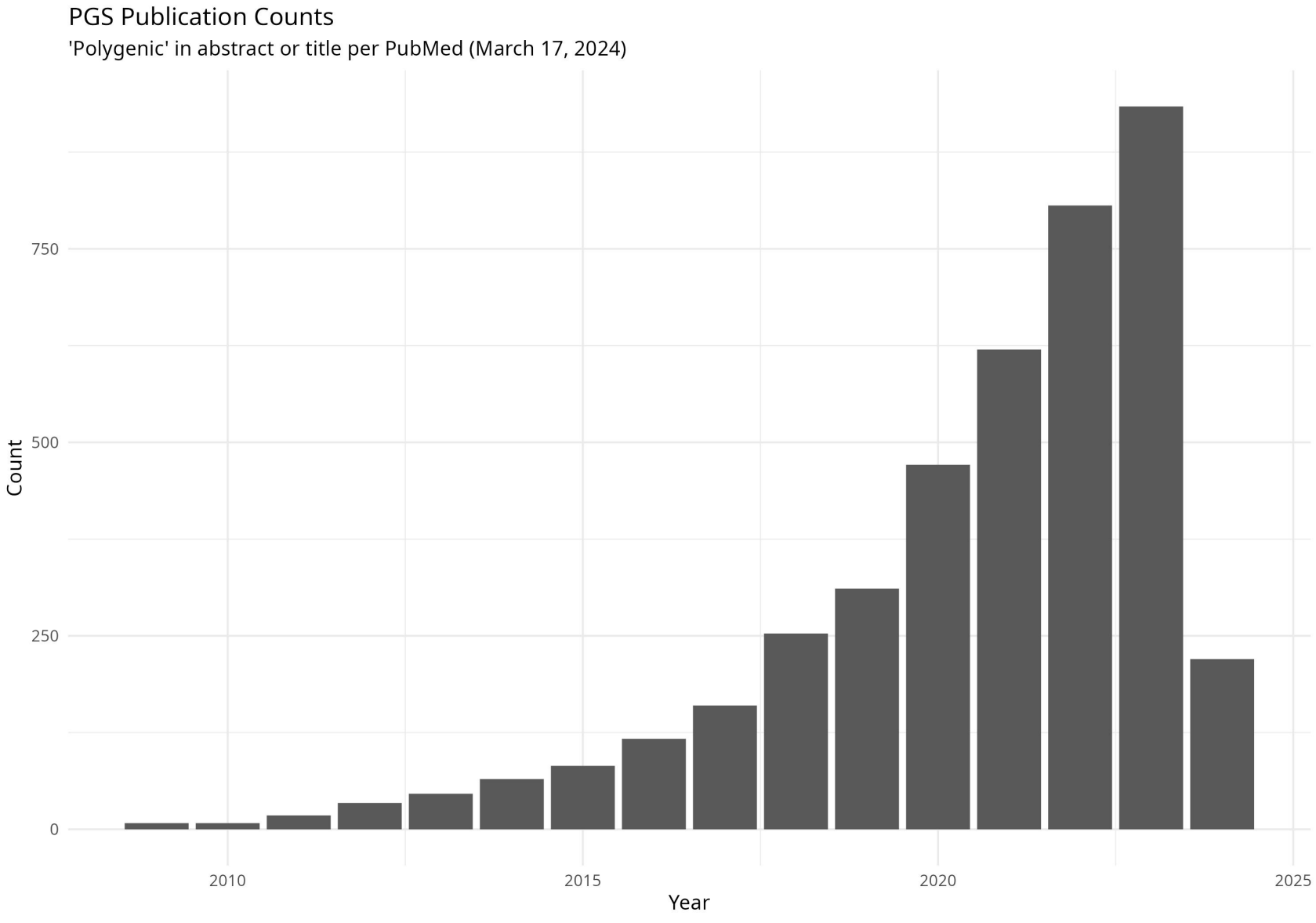
# Biobanks: Genotypes and Phenotypes

- Genotypes and phenotypes for hundreds of thousands of people
- Biobank examples include: UK Biobank, All of Us (USA), Biobank Japan, and FinnGen - with many others starting around the world
- Enable GWAS in diverse populations

# Polygenic Scores (PGS)

- Polygenic scores (PGS) use genetic variants' trait effects to predict trait values (Wray et al. 2021)
- PGS for some diseases predict disease as effectively as Mendelian gene variants (Khera et al. 2018)

# PGS Publication Counts



# PGS Methods

# Calculating PGS

- Weighted sums of risk allele count (0, 1, 2) for each individual
- Weights use estimated effect sizes from a genome-wide association study (GWAS)
- Summarizes a person's genetic risk for disease in a single number

$$P\hat{G}S_i = \sum_j g_{ij}\hat{\beta}_j$$

- Different PGS methods use different weights and different sets of SNPs

# Calculating PGS

- Early methods for calculating PGS used a p-value threshold to select variants (Wray, Goddard, and Visscher 2007; Dudbridge 2013; Wray et al. 2014; Euesden, Lewis, and O'reilly 2015; Chatterjee, Shi, and García-Closas 2016; T. I. S. Consortium 2009)
  - eg., only include variants with trait association p-value $<$ 5e-8
- Many recent (DBSLMM, PRS-CS, ldpred2) methods use all genome-wide variants (Yang and Zhou 2020; Ge et al. 2019; Privé, Arbel, and Vilhjálmsson 2020)

# DBSLMM (Yang and Zhou 2020)

- Deterministic Bayesian sparse linear mixed model
- Computationally efficient and scalable to large biobank data sets

$$Y = XB + \epsilon = X_l B_l + X_s B_s + \epsilon$$

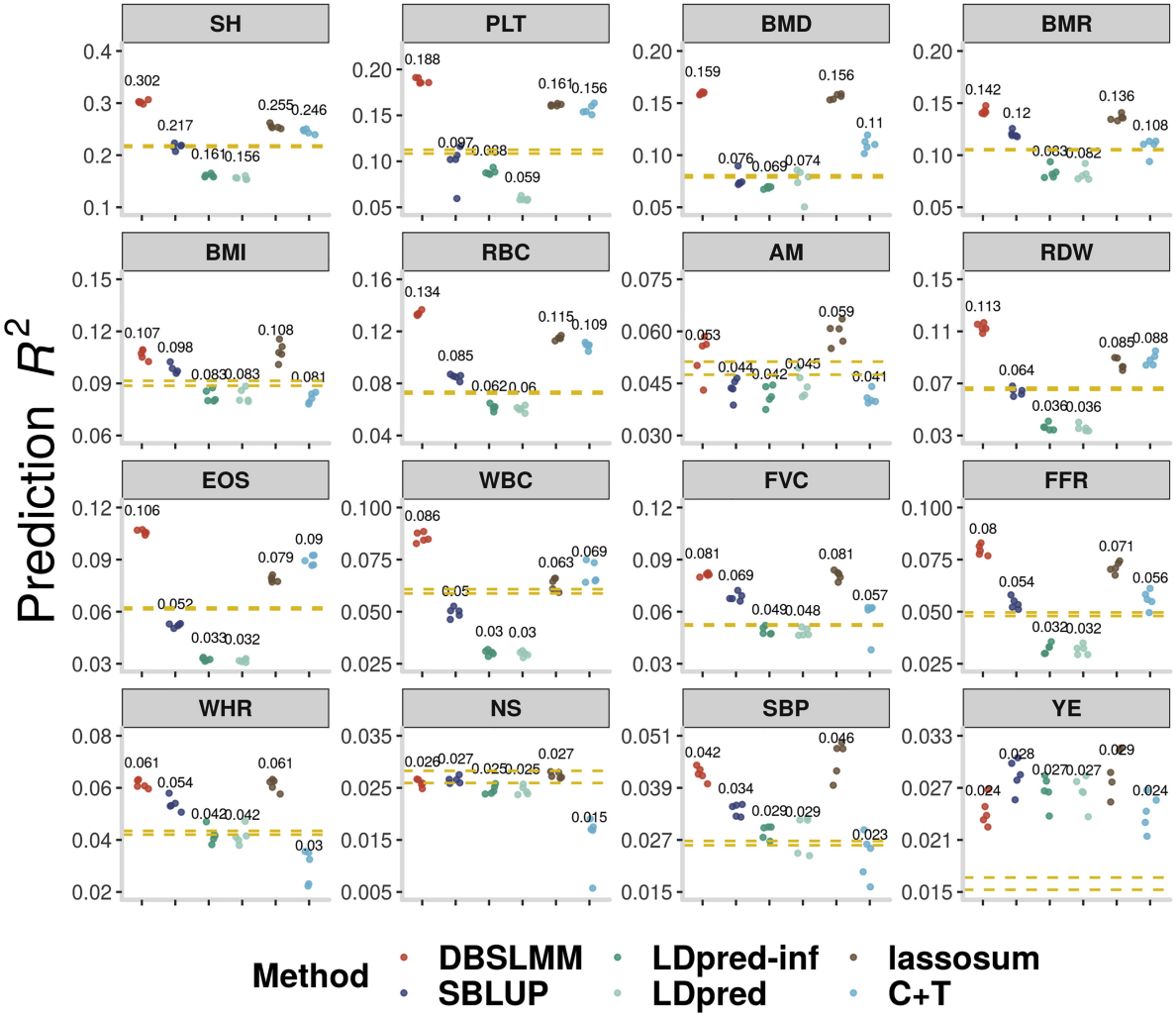
# DBSLMM (Yang and Zhou 2020)

- Identify large-effect SNPs via genome-wide association study
  - Treat large effects as fixed effects in the statistical model
  - Estimate large effects with high precision
- Remaining SNPs are assigned to the small-effect group
  - Treat small effects as random effects
  - While individual small effects are hard to estimate, collective estimation of their "polygenic" effects is possible

# DBSLMM (Yang and Zhou 2020)

- Performs well across a range of simulation settings
  - Heritability, polygenicity, effect size distribution
- Performs well in real data applications
  - Outperforms (in terms of prediction accuracy) other PGS methods in UK Biobank data
  - C+T, LDpred, SBLUP, lassosum

# DBSLMM (Yang and Zhou 2020)



# Jackknife+ & Crossvalidation+ for Prediction Intervals (Barber et al. 2021)

# Prediction Intervals for PGS

- (Quantitative trait) PGS are point estimates of trait values
- Uncertainty in the point estimates is often not reported
  - May impact PGS clinical utility
- Current PGS uses include risk stratification
- Uncertainty in PGS point estimates may risk stratification and ultimately lead to misclassification of subjects

# Jackknife method for prediction intervals in regression

- Split the data into a training set and a test set
- For each subject in the training set, fit the $n_{training}$ leave-one-out linear models $\hat{\mu}_{(-i)}$
- Calculate the $n_{training}$ absolute residuals: $|y_i - \hat{\mu}_{(-i)}(x_i)| = R_i^{LOO}$

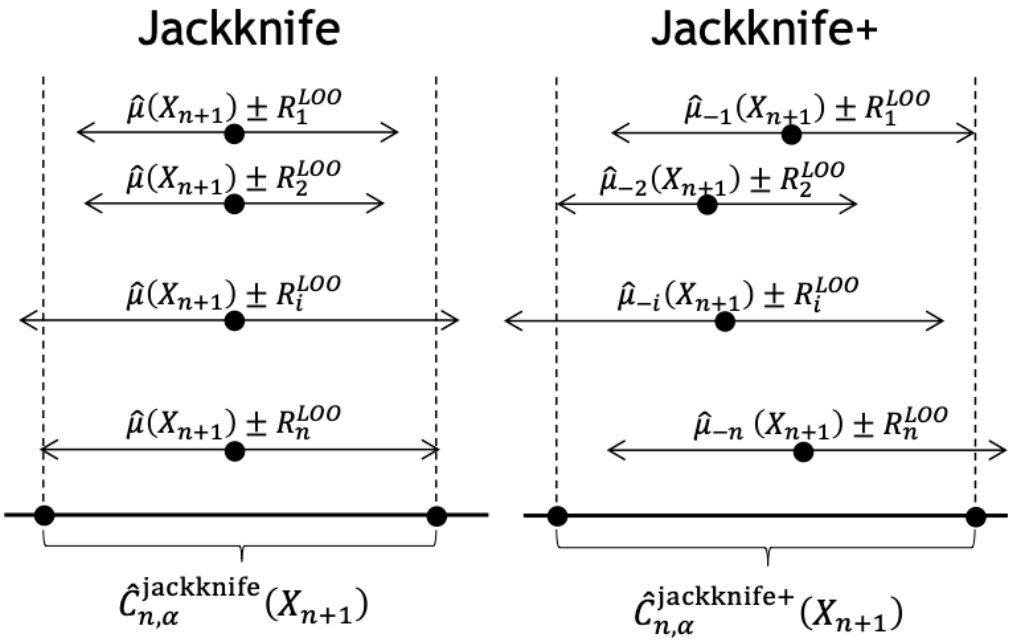
# Jackknife method for prediction intervals in regression

- Fit a linear model $\hat{\mu}$ using all training set subjects
- Calculate the $n_{training}$ values: $\hat{\mu}(x_{test}) + R_i^{LOO}$ and $\hat{\mu}(x_{test}) - R_i^{LOO}$
- Take the quantiles of the collection of $n_{training}$ values from $\hat{\mu}(x_{test}) + R_i^{LOO}$ and $\hat{\mu}(x_{test}) - R_i^{LOO}$

# Jackknife+ method

- A modification of the Jackknife method
- Uses the same leave-one-out residuals as Jackknife
- Uses the leave-one-out predictions at the test point to account for the variability in the fitted regression function

# Jackknife+ method



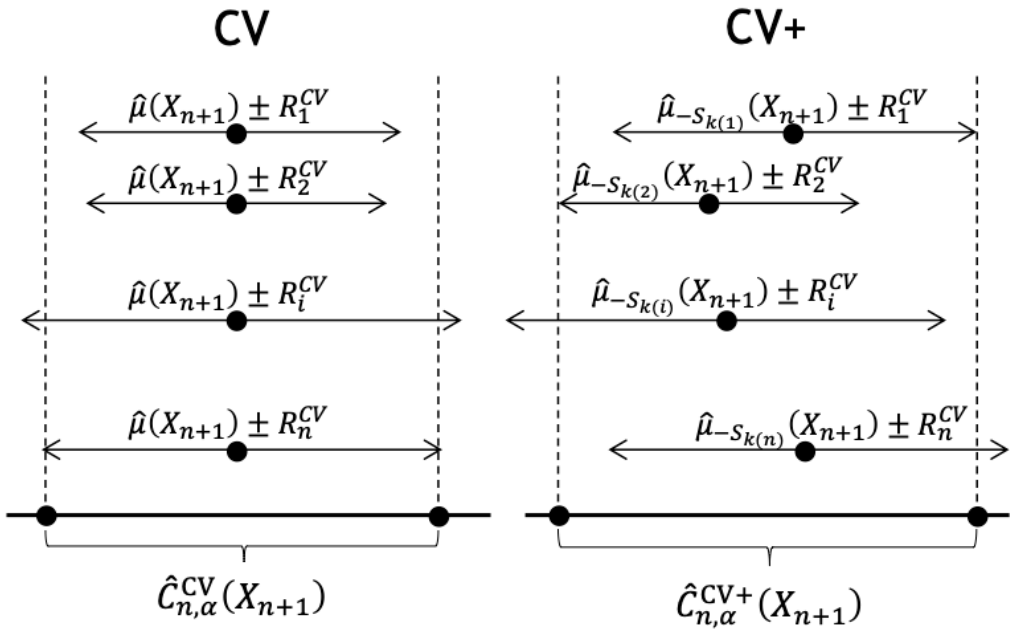

https://mapie.readthedocs.io/en/latest/theoretical_description_regression.html

# Jackknife+ Properties

- Jackknife+ interval for a new "test" point contains the true response with probability at least $1 - 2\alpha$
  - Proof leverages exchangeability of $n + 1$ points (Barber et al. 2021)

# K-fold CV+

- K-fold CV+ is a modification of K-fold cross-validation
- Compared to Jackknife+, K-fold CV+ requires fewer ($K$) model fits, and tends to produce wider intervals
  - Due to smaller effective sample size (compared to $n-1$ of Jackknife+)




https://mapie.readthedocs.io/en/latest/theoretical_description_regression.html

# K-fold CV+ Procedure

- Split training set into K folds
- For each fold, fit model with subjects from remaining K-1 folds
- Predict trait values for subjects in the held-out fold
- Calculate absolute residuals $R_i^{LOO}$ for every training set subject

# K-fold CV+ Procedure

- For every test set subject, use the K fitted values and the $n_{training}$ absolute residuals to get $n_{training}$ values from fitted values plus residuals
  - Similarly, get $n_{training}$ values from fitted values minus residuals
- Get quantiles of the collections of:
  - $n_{training}$ fitted values plus residuals
  - $n_{training}$ fitted values minus residuals
- Use these quantiles as the prediction interval for the test set subject

# K-fold CV+ Prediction Intervals for PGS

- Requires only $K$ model fits (one per fold)
- Tend to get wider intervals (compared to Jackknife+) due to decreased sample size
- Can be used with any method that produces PGS point estimates
  - Ding et al. (2022) also constructs PGS intervals, but they require `ldpred2` construction method
  - Their approach doesn't work with, for example, `DBSLMM` or widely used Clumping & Thresholding method

# UK Biobank PGS Intervals

# Design

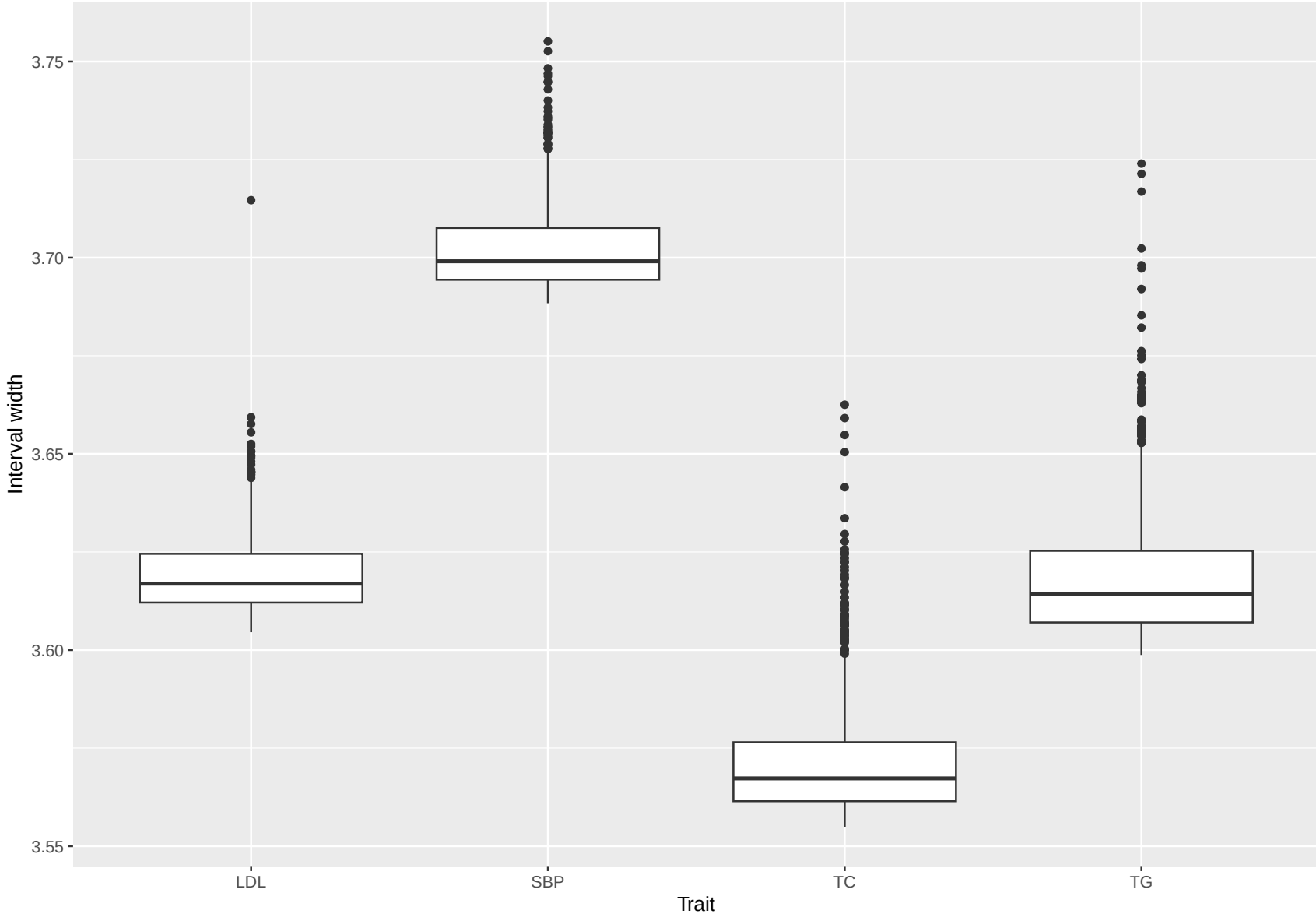
- ~300,000 UK Biobank subjects
- 5-fold CV+ for four cardiovascular traits:
  - LDL cholesterol (LDL)
  - Systolic blood pressure (SBP)
  - Total cholesterol (TC)
  - Triglycerides (TG)

# UKB PGS Intervals with DBSLMM

UKB PGS Interval Coverages

| Trait | Coverage |
|---|---|
| systolic blood pressure | 0.93 |
| total cholesterol | 0.93 |
| low-density lipoprotein cholesterol | 0.93 |
| triglycerides | 0.94 |

# UKB PGS Intervals with DBSLMM



# Next Steps

- Expand to other traits
- Compare coverages and interval widths for other PGS methods
- Characterize effect of fold number on CV+ intervals
- Examine training set sample size effect on CV+ intervals
- Study performance across diverse heritability and polygenicity settings

# Future Directions

# Future Directions

- How to integrate PGS with other risk factors for clinical and public health applications?

# Framingham Coronary Artery Disease Risk Factors

- Age
- Total cholesterol
- High-density lipoprotein cholesterol
- Blood pressure
- Treatment for high blood pressure
- Smoking status

# Coronary Artery Disease & PGS

- PGS risks for coronary artery disease differ from those of clinical risk factors, as quantified by Framingham risk scores (Abraham et al. 2016)
- PGS & Framingham risk scores complement each other & together improve risk prediction (Abraham et al. 2016)
- PGS & Framingham risk scores are nearly uncorrelated (Abraham et al. 2016)

# Improving PGS Methods

1. Model SNP-SNP interactions
2. Use SNP functional annotations
3. Accommodate SNP-environment interactions
4. Jointly model common & rare variants

# Modeling SNP-SNP Interactions

- Current PGS methods ignore SNP-SNP interactions
- SNP-SNP interactions may be important for some traits
- Large number of SNP-SNP interactions ($\sim \binom{10^6}{2}$)

# Modeling SNP-SNP Interactions

- Use sparsity-inducing priors (eg., spike-and-slab) for SNP-SNP interaction effects
- Use variational methods vs. sampling-based strategies (Markov chain Monte Carlo) to decrease computational burdens

# Modeling SNP-SNP Interactions

- Epistasis (SNP-SNP interaction) Factor Analysis (Tang, Freudenberg, and Dahl 2023)
- Factors polygenic epistasis into interactions between a few epistasis factors (latent polygenic components)

# Conclusions

# Conclusions

- Polygenic scores (PGS) use genetic variants' trait effects to predict trait values
- PGS methods use different weights and different sets of SNPs
- CV+ Prediction Intervals for PGS aid in characterizing uncertainty in PGS point estimates

# Conclusions

- Integrating PGS with other risk factors for clinical and public health applications improves risk prediction
- Interplay among biostatistics & clinical research will bring PGS to clinical and public health applications

Thank you!

# References


Abraham, Gad, Aki S Havulinna, Oneil G Bhalala, Sean G Byars, Alysha M De Livera, Laxman Yetukuri, Emmi Tikkanen, et al. 2016. "Genomic Prediction of Coronary Heart Disease." European Heart Journal 37 (43): 3267–78.

Barber, Rina Foygel, Emmanuel J. Candes, Aaditya Ramdas, and Ryan J. Tibshirani. 2021. "Predictive Inference with the Jackknife+." Annals of Statistics 49 (1): 486–507. https://doi.org/10.1214/20-AOS1965.

Chatterjee, Nilanjan, Jianxin Shi, and Montserrat García-Closas. 2016. "Developing and Evaluating Polygenic Risk Prediction Models for Stratified Disease Prevention." Nature Reviews Genetics 17 (7): 392–406.

Consortium, The International Schizophrenia. 2009. "Common Polygenic Variation Contributes to Risk of Schizophrenia and Bipolar Disorder." Nature 460: 748–52.

Consortium, Wellcome Trust Case Control. 2007. "Genome-Wide Association Study of 14,000 Cases of Seven Common Diseases and 3,000 Shared Controls." Nature 447 (7145): 661–78.

Ding, Yi, Kangcheng Hou, Kathryn S Burch, Sandra Lapinska, Florian Privé, Bjarni Vilhjálmsson, Sriram Sankararaman, and Bogdan Pasaniuc. 2022. "Large Uncertainty in Individual Polygenic Risk Score Estimation Impacts PRS-Based Risk Stratification." Nature Genetics 54 (1): 30–39.

Dudbridge, Frank. 2013. "Power and Predictive Accuracy of Polygenic Risk Scores." PLoS Genetics 9 (3): e1003348.

Euesden, Jack, Cathryn M Lewis, and Paul F O'reilly. 2015. "PRSice: Polygenic Risk Score Software." Bioinformatics 31 (9): 1466–68.

Ge, Tian, Chia-Yen Chen, Yang Ni, Yen-Chen Anne Feng, and Jordan W Smoller. 2019. "Polygenic Prediction via Bayesian Regression and Continuous Shrinkage Priors." Nature Communications 10 (1): 1776.

Howles, Sarah A, Akira Wiberg, Michelle Goldsworthy, Asha L Bayliss, Anna K Gluck, Michael Ng, Emily Grout, et al. 2019. "Genetic Variants of Calcium and Vitamin d Metabolism in Kidney Stone Disease." Nature Communications 10 (1): 5175.

Khera, Amit V., Mark Chaffin, Krishna G. Aragam, Mary E. Haas, Carolina Roselli, Seung Hoan Choi, Pradeep Natarajan, et al. 2018. "Genome-Wide Polygenic Scores for Common

# Extra slides

# Clumping + Thresholding

- Clumping:
  - Identify independent genetic variants
  - Remove variants that are highly correlated with other variants
- Thresholding:
  - Select variants that pass a p-value threshold in GWAS
  - Weight variants by their effect size

# Jackknife method

- Jackknife is a resampling method for estimating bias or variance
- Idea: use the distribution of the leave-one-out statistics to estimate the distribution of the statistic of interest
- For each observation, calculate the statistic with that observation removed

# Jackknife method

- Example: Calculate the sample mean from a collection of observations
- 4 adults' heights: $64, 70, 72, 70$
- $\bar{x}_{(-1)} = \frac{70+72+70}{3} = 70.67$
- $\bar{x}_{(-2)} = \frac{64+72+70}{3} = 68.67$
- $\bar{x}_{(-3)} = \frac{64+70+70}{3} = 68$
- $\bar{x}_{(-4)} = \frac{64+70+72}{3} = 68.67$
- $\bar{x}_{Jackknife} = \frac{\bar{x}_{(-1)}+\bar{x}_{(-2)}+\bar{x}_{(-3)}+\bar{x}_{(-4)}}{4} = \frac{70.67+68.67+68+68.67}{4} = 69$