POLYGENIC PREDICTION

Jian Zeng, j.zeng@uq.edu.au

1. **Fundamentals**: understanding, limitations, applications, challenges
2. **Evaluation**: statistics for quantitative and binary traits, visualization, pitfalls
3. **Conventional methods**: clumping & P-value thresholding (C+PT), best linear unbiased prediction (BLUP)
4. **Bayesian methods**: Bayes theorem, priors, posterior inference, sumstats-based methodology
5. **MCMC sampling (optional)**: spike-and-slab model, algorithm, technical details
6. **SBayesRC**: incorporating functional annotations, low-rank modelling, application

Fundamentals of Polygenic Prediction

Jian Zeng

j.zeng@uq.edu.au

Slide courtesy: Naomi Wray

What are we predicting?

Polygenic scores (PGS) predict individual genetic values of complex traits using **genome-wide** variations.

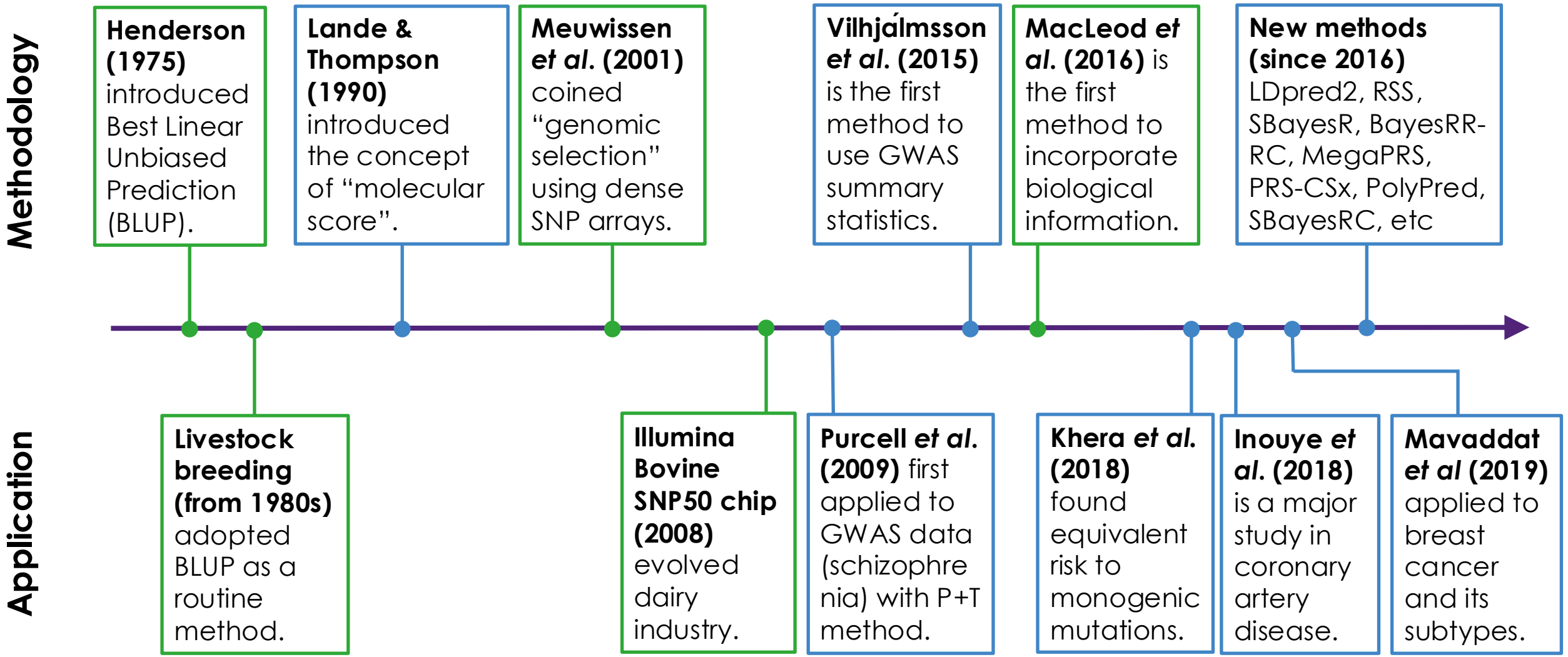
Polygenic risk scores (PRS) are predictors of the genetic susceptibilities of individuals to diseases.

Applications

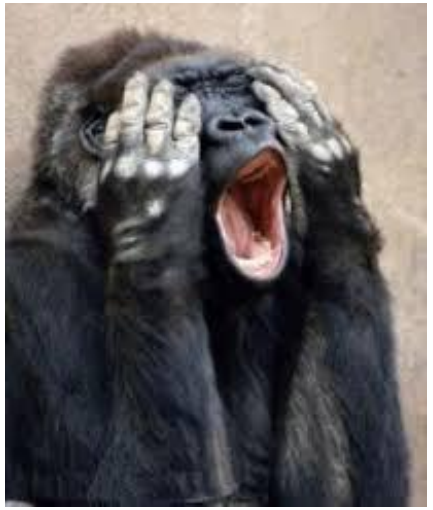
- Precision medicine (humans)
- Genomic selection (animals/plants)

Source: Strachan & Read Human Molecular Genetics 3.

A brief history of PGS in humans & agriculture



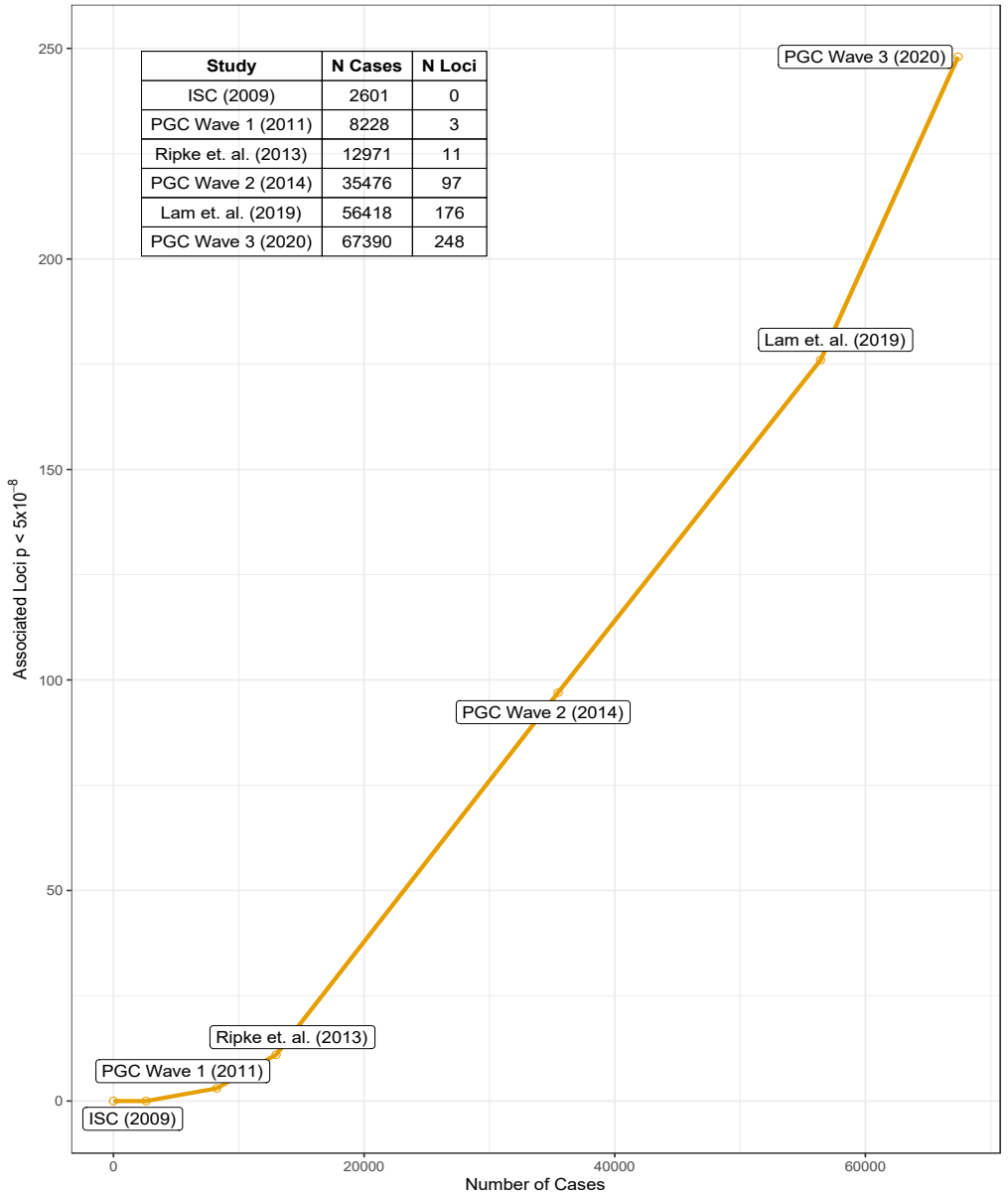
What's in a name?



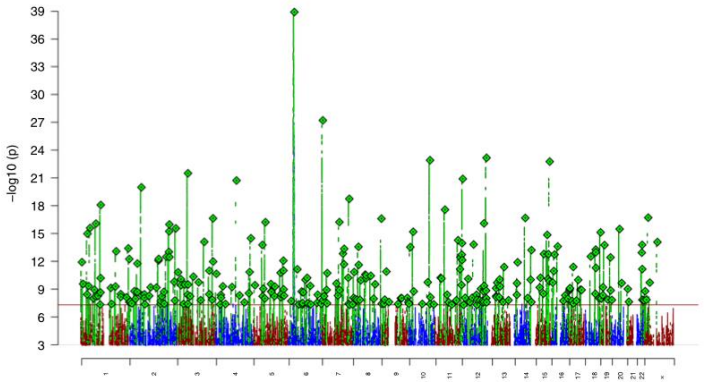
- **PRS**- Polygenic risk score
- **GPRS**- Genomic or genetic profile risk score
- **PGS** -Polygenic score
- **GRS** - Genetic risk score
- **rsPS** – restricted to significant polygenic score
- **gePS** – global extended polygenic score
- **Multi-SNP score** (usually this uses only single nucleotide polymorphisms (SNPs) that are genome-wide significant, hence the same as gePS)
- **MetaGRS** – a PRS constructed from genetic data for the disease/trait of interest plus from other correlated traits
- **MTAG-GRS/PRS** a PRS constructed from GWAS data from multiple correlated traits
- **Genetic score**
- **Genotypic score**
- **Allele score**
- **Profile score**
- **Linear predictor** (this of course is a generic term, but has been used to describe PRS when risk alleles are the only predictors)

Common diseases are polygenic

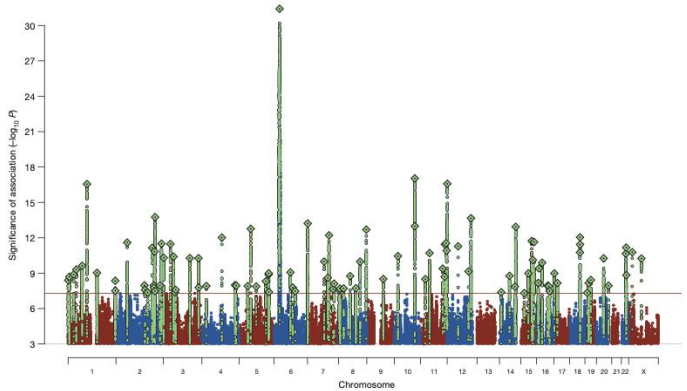
Study	N Cases	N Loci
ISC (2009)	2601	0
PGC Wave 1 (2011)	8228	3
Ripke et. al. (2013)	12971	11
PGC Wave 2 (2014)	35476	97
Lam et. al. (2019)	56418	176
PGC Wave 3 (2020)	67390	248



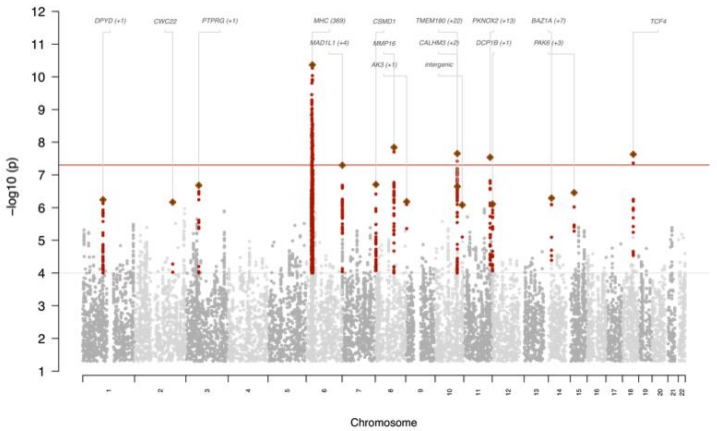
Schizophrenia



2022 PGC Wave 3

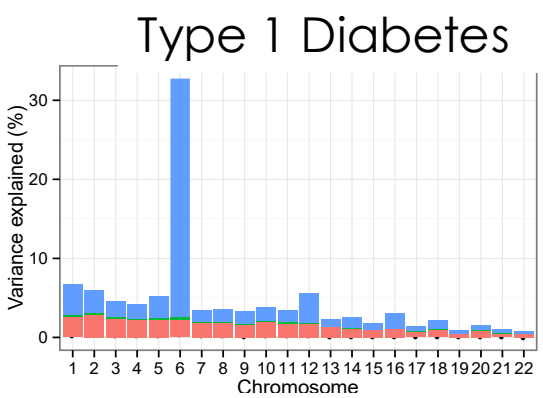


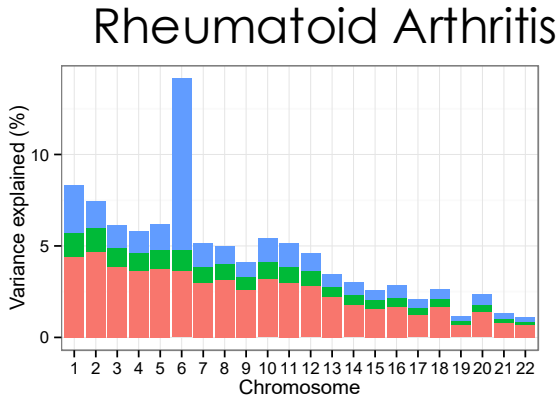
2014 PGC Wave 2

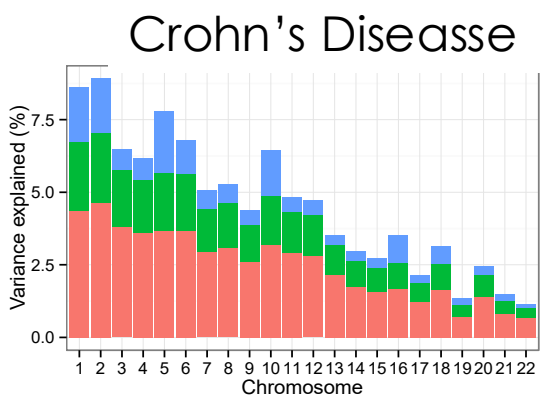


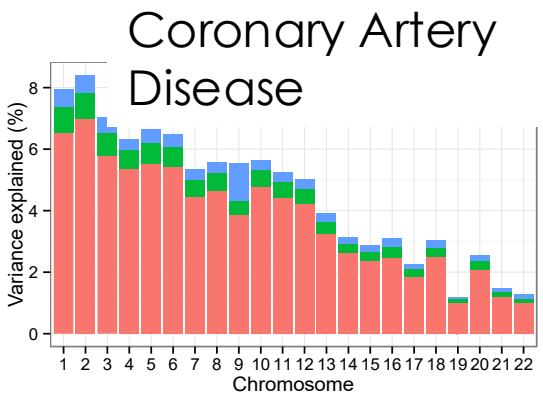
2011 PGC Wave 1

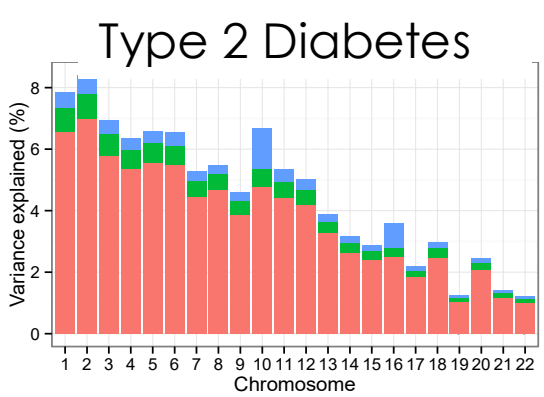
Many polygenic genetic architectures

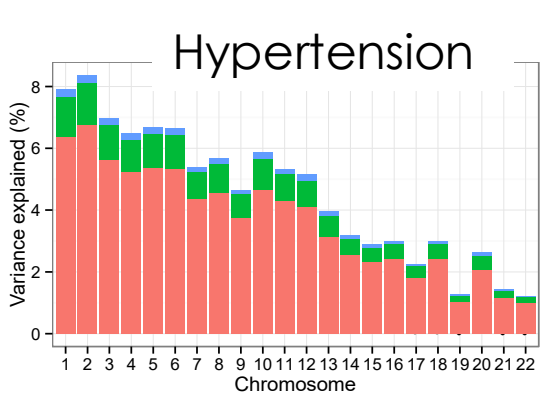


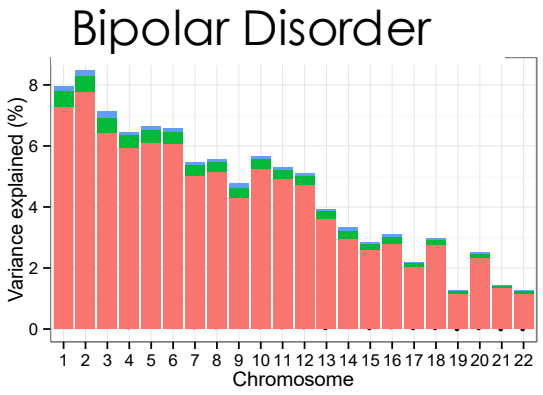


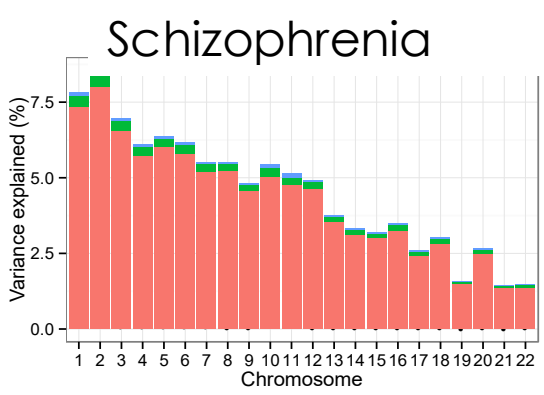












Mixture component

$10^{-4} \times \sigma^2_g$ small

$10^{-3} \times \sigma^2_g$ middle

$10^{-2} \times \sigma^2_g$ big

PLOS | GENETICS

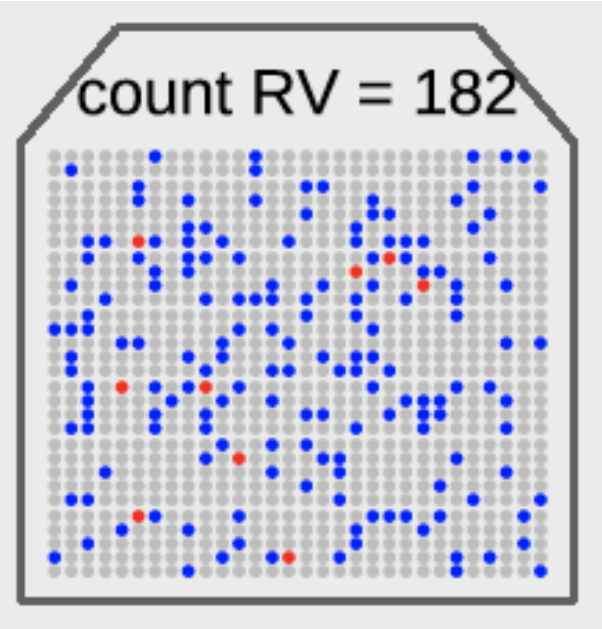
RESEARCH ARTICLE

Simultaneous Discovery, Estimation and Prediction Analysis of Complex Traits Using a Bayesian Mixture Model

Gerhard Moser[1]*, Sang Hong Lee[1], Ben J. Hayes[2,3], Michael E. Goddard[2,4], Naomi R. Wray[1], Peter M. Visscher[1,5]

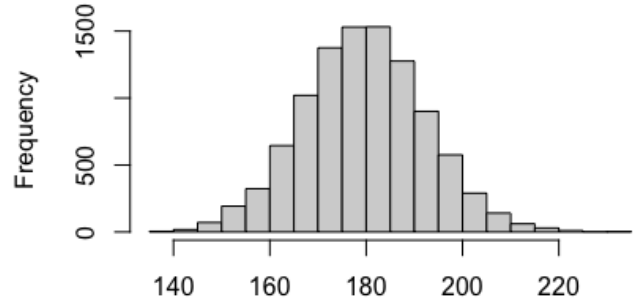
Many DNA variants contribute to genetic risk, and most have very small effects.

Polygenic disease for an individual



900 DNA polymorphic sites

RV =risk variant



Frequency of risk variant at each site: 0.1 (p)

Average person 900*2*0.1 = 180 risk variant

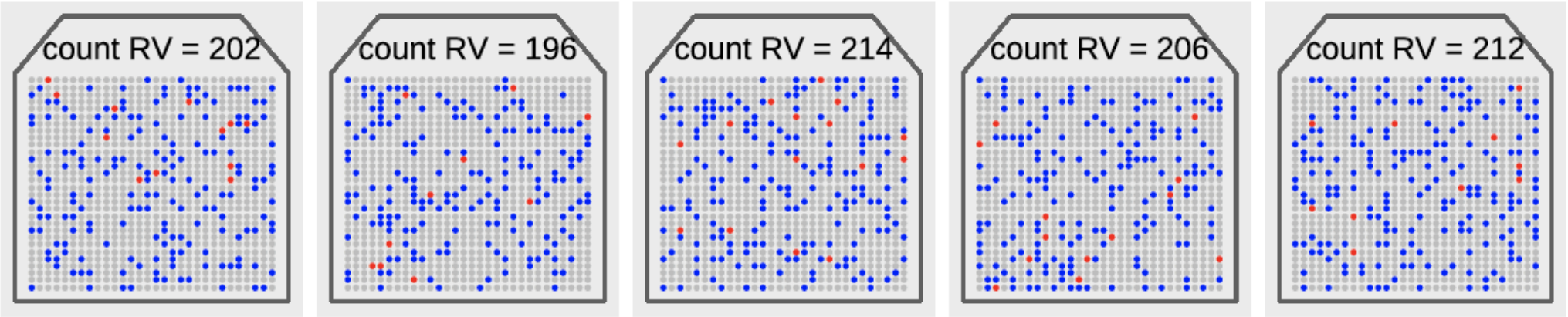
Mean +/- 3SD: 142 to 218

0 Grey: Homozygote no risk alleles (or equivalently 2 protective alleles)
1 Blue : Heterozygote one risk allele (and one non-risk/protective allele)
2 Red: Homozygote two risk alleles

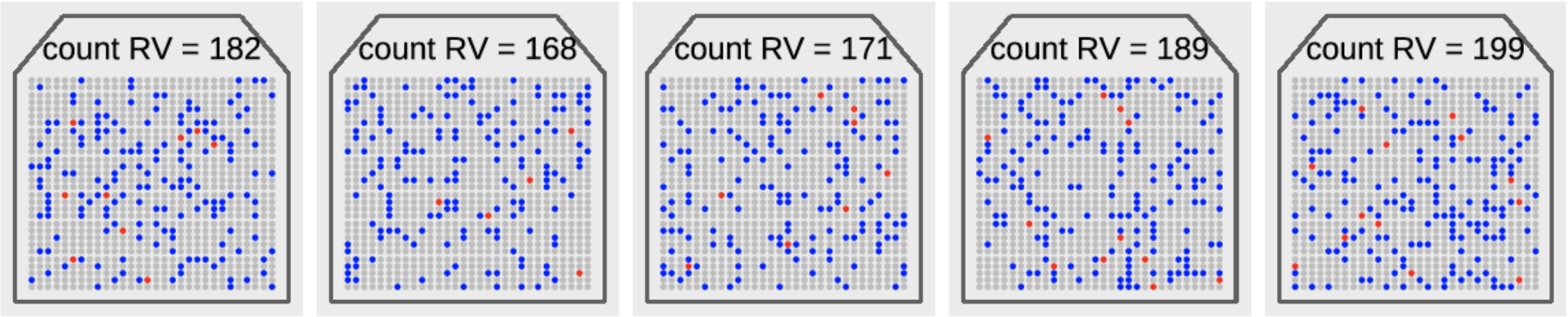
Toy example

Polygenic disease for an individual

Affected over lifetime

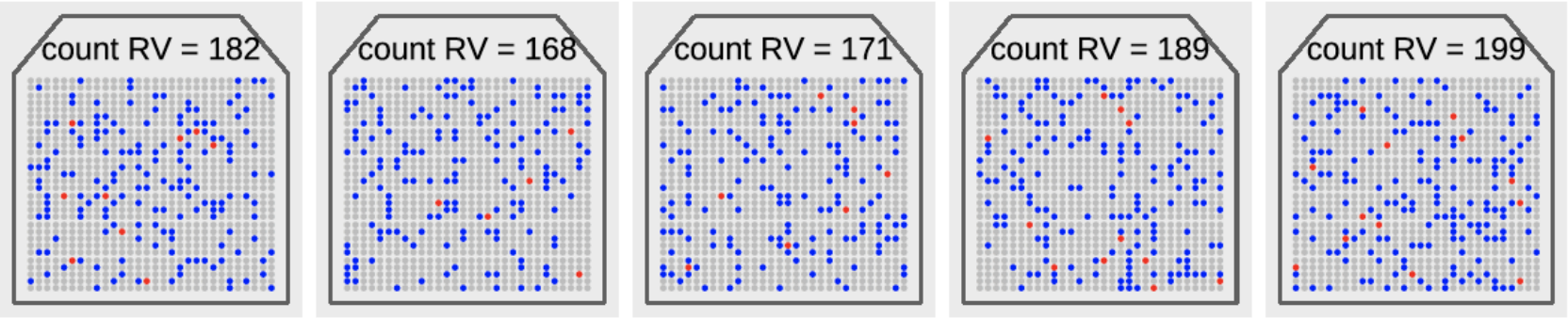


Not affected over lifetime



- We all carry risk variants for all diseases.
- Robustness
- Those affected carry a higher burden.
- Non-genetic factors contribute to risk too
- Each person carries a unique portfolio of risk alleles

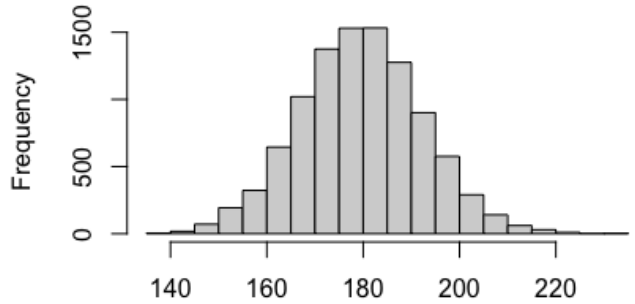
Polygenic score



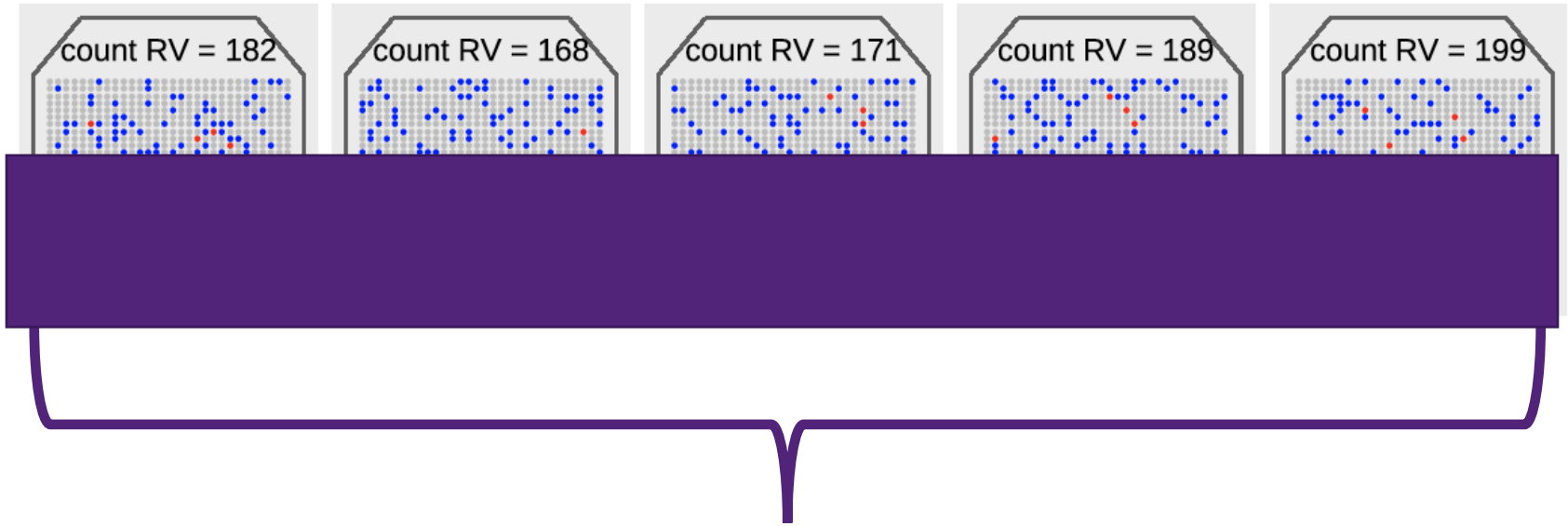
→ "True" polygenic score

Genetic variance between people attributed to all genetic factors V(A)

$h^2 = \frac{V(A)}{V(P)}$ heritability



Polygenic score



"True" polygenic score

Not all variants captured on genotyping arrays

Genetic variance between people attributed to all genetic factors V(A)

$h^2 = \frac{V(A)}{V(P)}$ heritability

Genetic variance between people attributed to all genetic factors associated with SNPs on genotyping arrays

$h^2_{SNP} = h^2_g = \frac{V(A:SNP)}{V(P)}$

$\mathrm{SNP-based\ heritability}$

Polygenic scores

In reality, risk variants have different effect sizes.

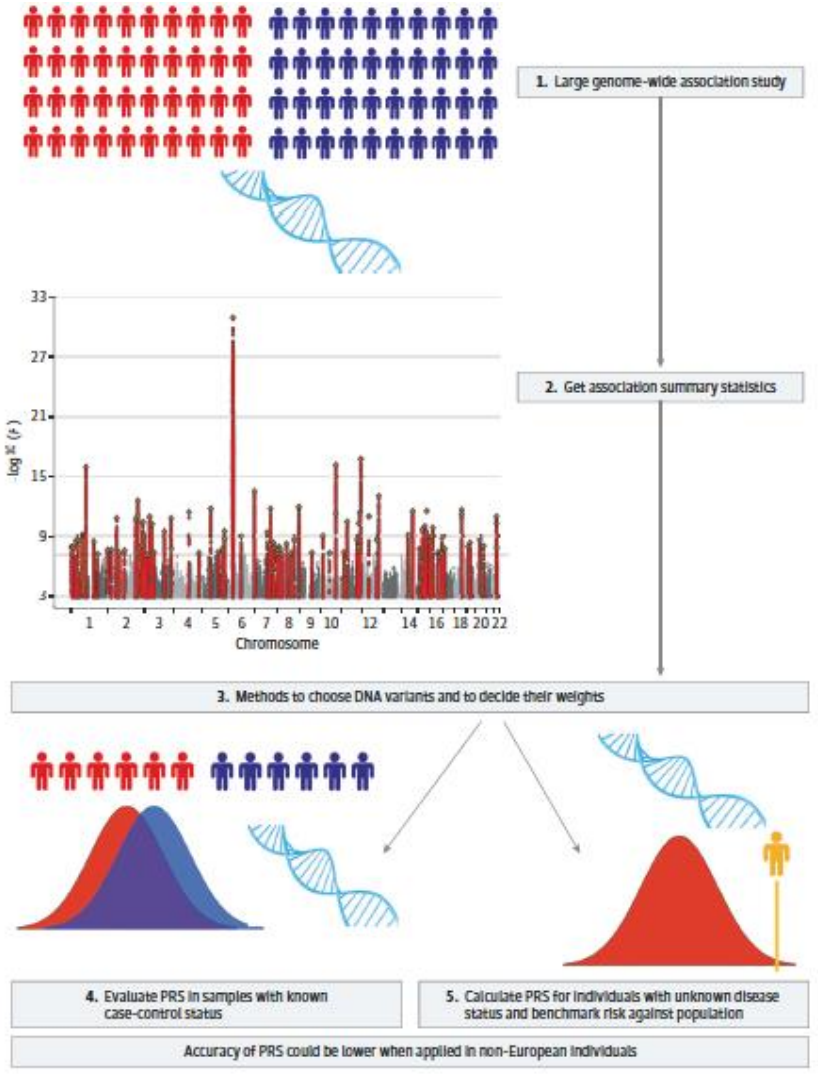
Therefore, PGS is a weighted count of risk alleles:

$$PGS = \widehat{\beta_1} x_{i1} + \widehat{\beta_2} x_{i2} + \widehat{\beta_3} x_{i3} + \cdots = \sum_{j=1}^{n_{SNP}} \widehat{\beta_j} x_{ij}$$

0, 1 or 2
Risk alleles

Which SNPs?

What weights?



- Don't need to know causal variants for prediction!
- Prediction can be based on correlated variants.

Evaluate
Y= b*PGS + e
R^2 = var(b*PGS)/Var(Y)

AUC statistic:
Probability that a case ranks higher than a control

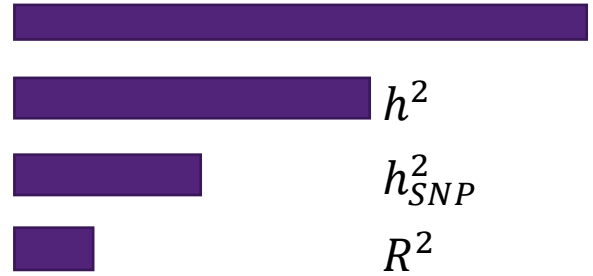
Limitations in prediction accuracy

- PGS have a **theoretical** upper limit dependent on the **heritability of the trait** (how much of the variance of trait values between people is attributed to genetic factors).

- PGS have a **technical** upper limit associated with the proportion of **variance tagged** by the DNA variants measured.

- PGS have a **practical** upper limit dependent on the **sample size of the discovery sample** used to estimate effect sizes of risk alleles, and the **quality** of the discovery sample.

- PGS can be pushed closer to the technical upper limit by the **statistical methodology** used to generate the optimal weighting given to the risk alleles, and new methods integrate new biological data.



Schizophrenia

Max:
25% Liability
AUC 0.84

Current:
11% Liability
AUC 0.74

Polygenic scores cannot be highly accurate predictors of phenotypes

Parameters determining the prediction accuracy

The expected value of prediction accuracy:

Variance explained by the predictor

$$R^2 = \frac{h_m^2}{1 + C}$$

h_m^2 : True variance explained by the predictor depends on the SNP set - subscript m.

C: captures the error in estimation

As C→ 0, R^2 → h_m^2

$$C \approx \frac{m}{N h_m^2}$$

- *N*: discovery sample size
- *m*: the number of SNPs (assume LD-independent)
- h_m^2: the SNP-heritability captured by *m* SNPs

Wray et al (2019) Complex trait prediction from genome data. Genetics

What is the maximum prediction accuracy we can get?

h_m^2 : True variance explained by the predictor depends on the SNP set - subscript m.

Variance explained by the predictor

$$R^2 = \frac{h_m^2}{1 + C}$$

C: captures the error in estimation

As C→ 0, R^2 → h_m^2

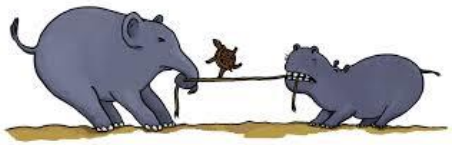
We want C to be as small as possible:

- C decreases as Discovery sample N increases
- C decreases as the number of SNPs in the SNP set m decreases

$$C \approx \frac{m}{N h_m^2}$$

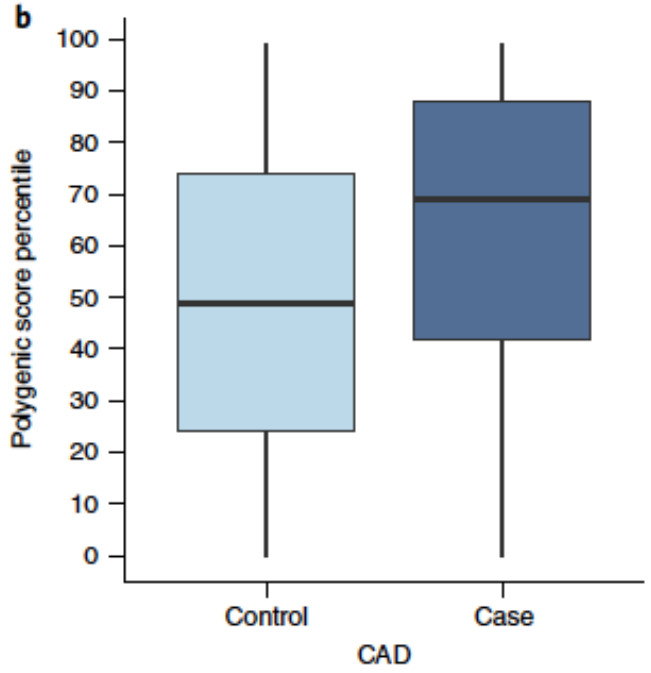
As m gets smaller, h_m^2 also gets smaller

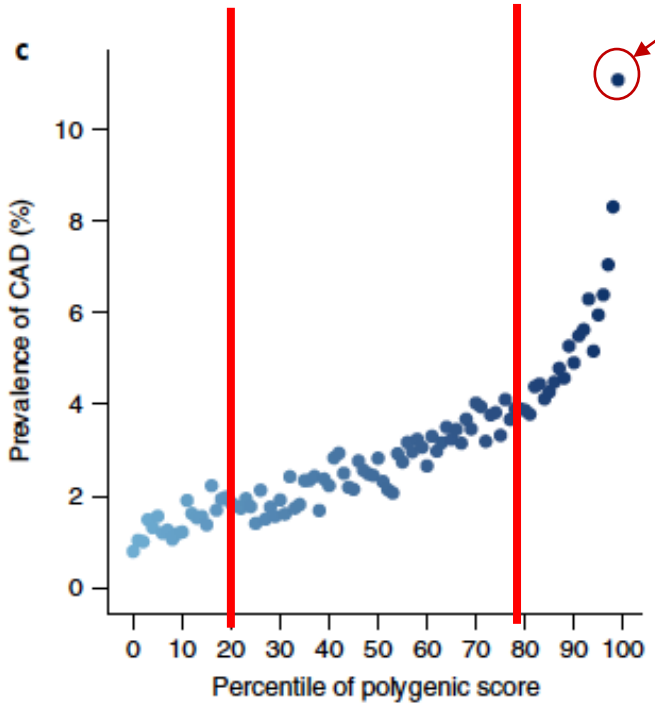
How to optimise m and h_m^2 to get max R^2 ?

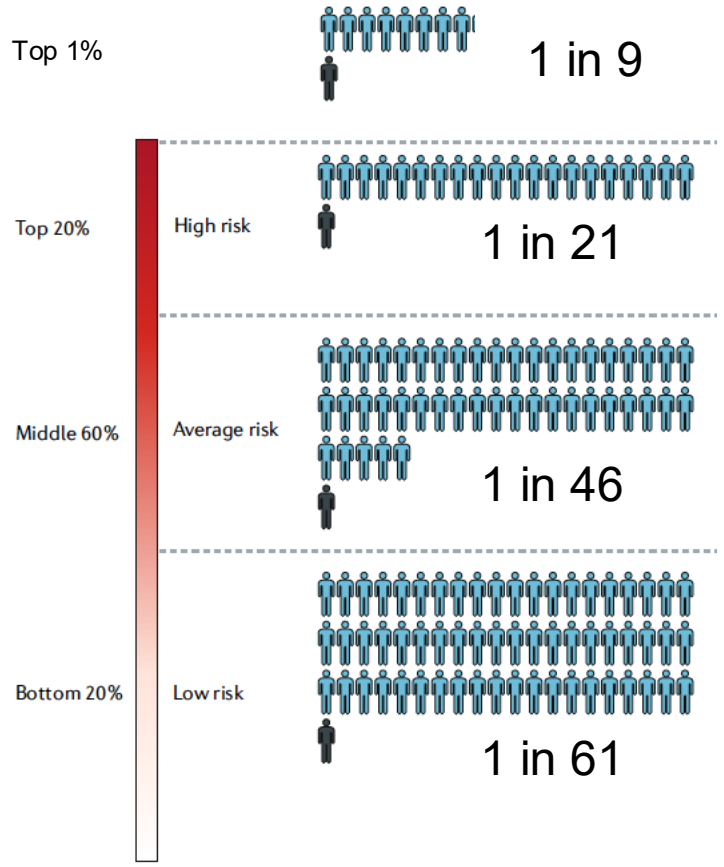


Wray et al (2019) Complex trait prediction from genome data. Genetics

How useful is PRS in practice?





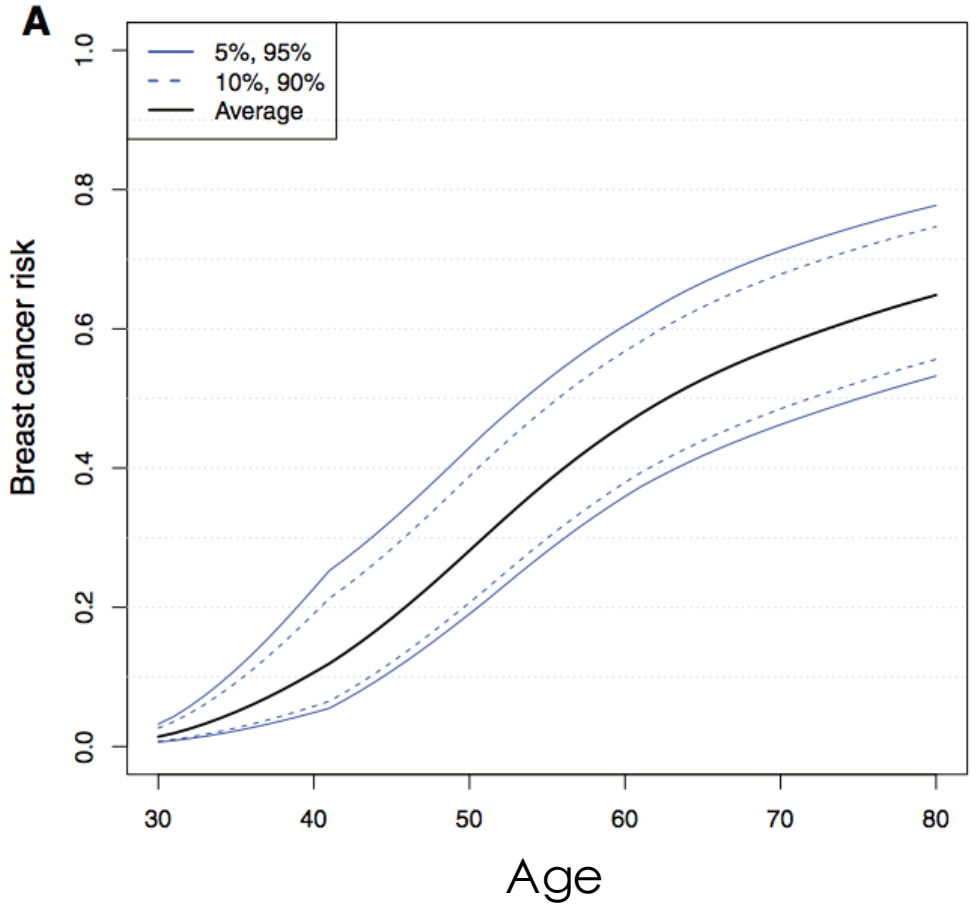


Khera et al (2018) Genome-wide polygenic scores for common diseases identify individuals with risk **equivalent to monogenic mutations**. Nature Genetics

Torkamani et al, Nat Rev Genetics, 2018

Combine PRS with known risk mutations

Breast cancer

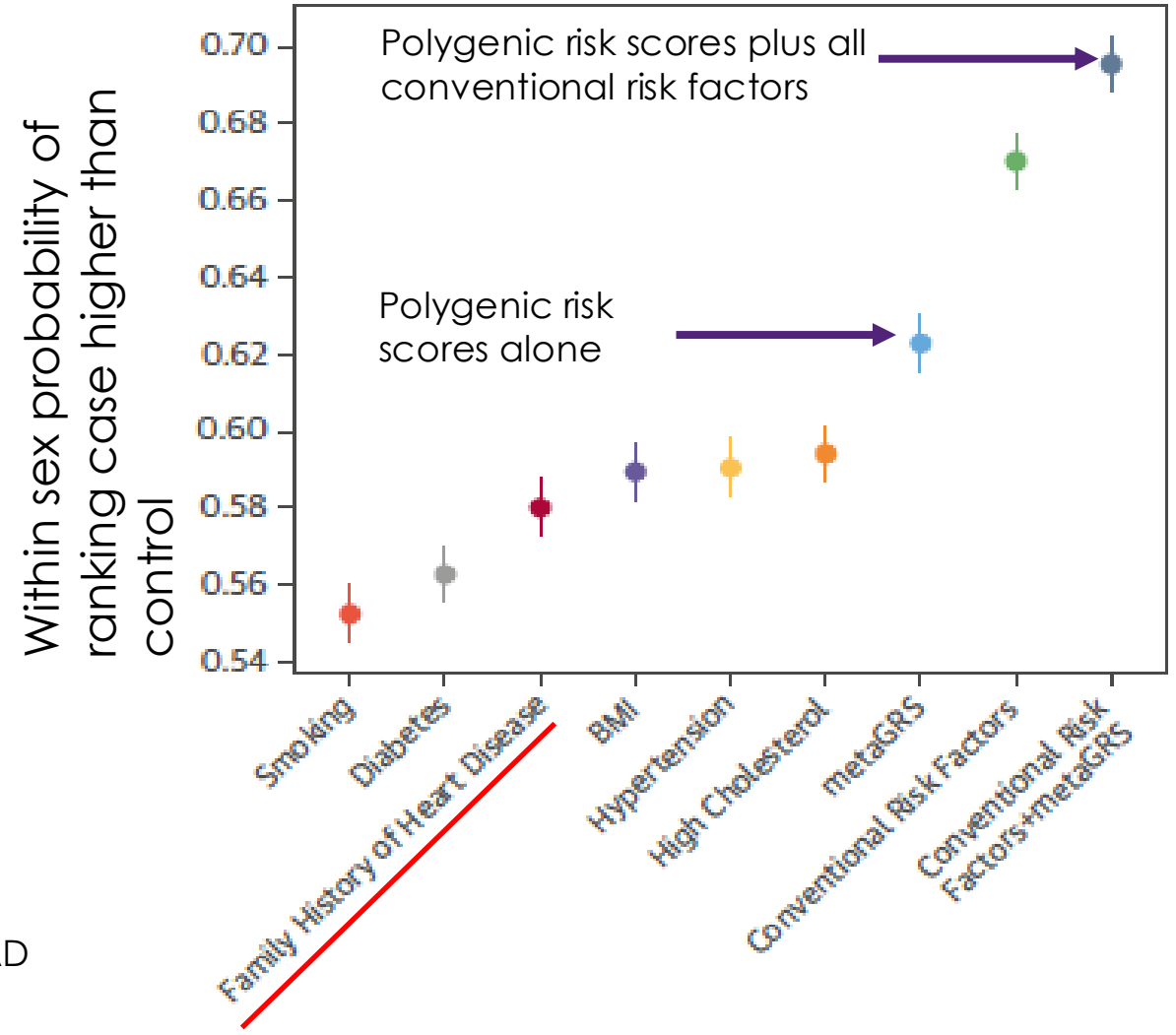


BRCA1 carriers

Kuchenbaecker et al: Evaluation of polygenic risk scores for breast and ovarian cancer risk prediction in BRCA1 and BRCA2 mutation carriers. J Natl Cancer Inst (2017)

Combine PRS with conventional risk predictors

Coronary Artery Disease



Inouye et al (2018) Genomic risk prediction of CAD in 480K adults. JACC

Family history

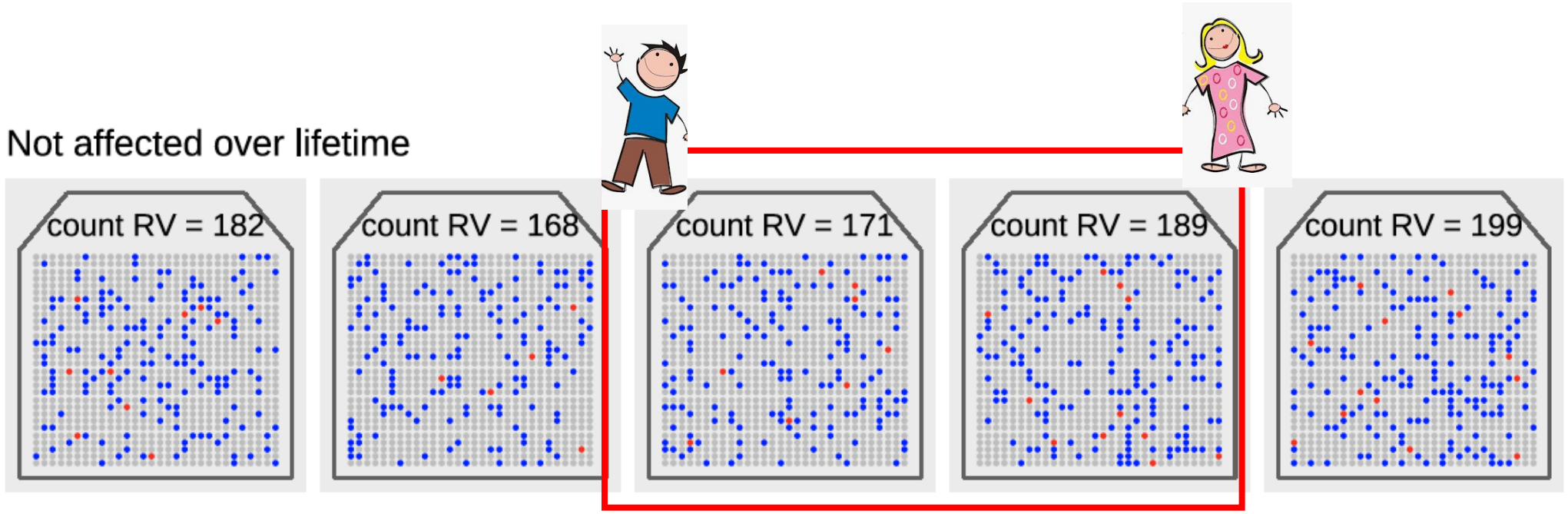
Will people withOUT known family history have high PRS?

Maybe, and that's important!

JAMA Psychiatry | Review

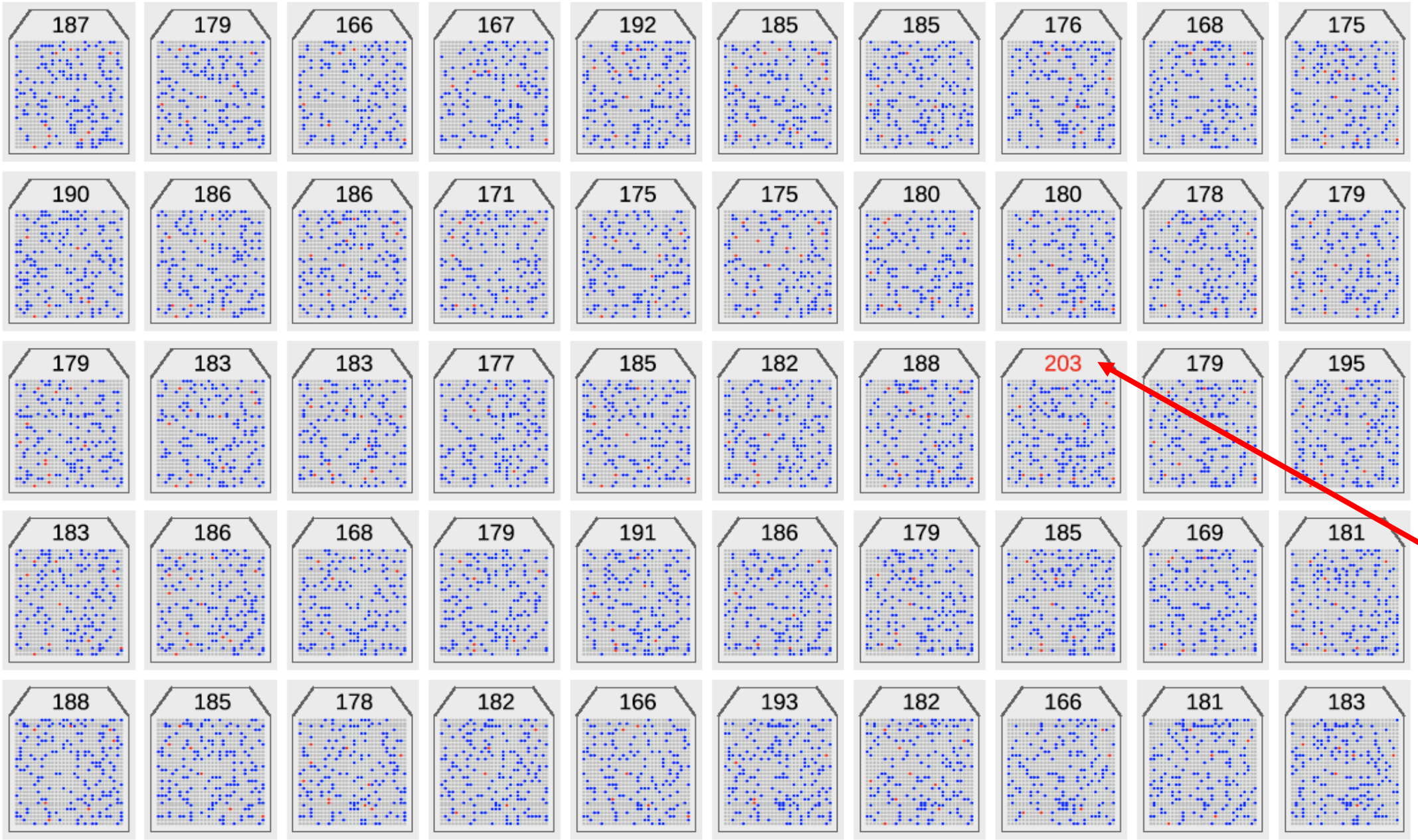
From Basic Science to Clinical Application of Polygenic Risk Scores
A Primer

Naomi R. Wray, PhD; Tian Lin, PhD; Jehannine Austin, PhD; John J. McGrath, MD, PhD; Ian B. Hickie, MD; Graham K. Murray, MD, PhD; Peter M. Visscher, PhD



Grey: Homozygote: Two non-risk/protective alleles – always passes a non-risk allele to child at the locus
Red: Homozygote: Two risk alleles – always passes a risk allele to child at the locus
Blue: Heterozygotes: One risk allele & one non-risk allele –
passes a risk allele 50% of the time & a non-risk allele 50% of the time

Children (Parents: 171 & 189)



Substantial genetic variance within the family

Children of these parents
Mean: 180
+/-3SD: 153-207

Population
Mean: 180
+/-3SD: 142-218

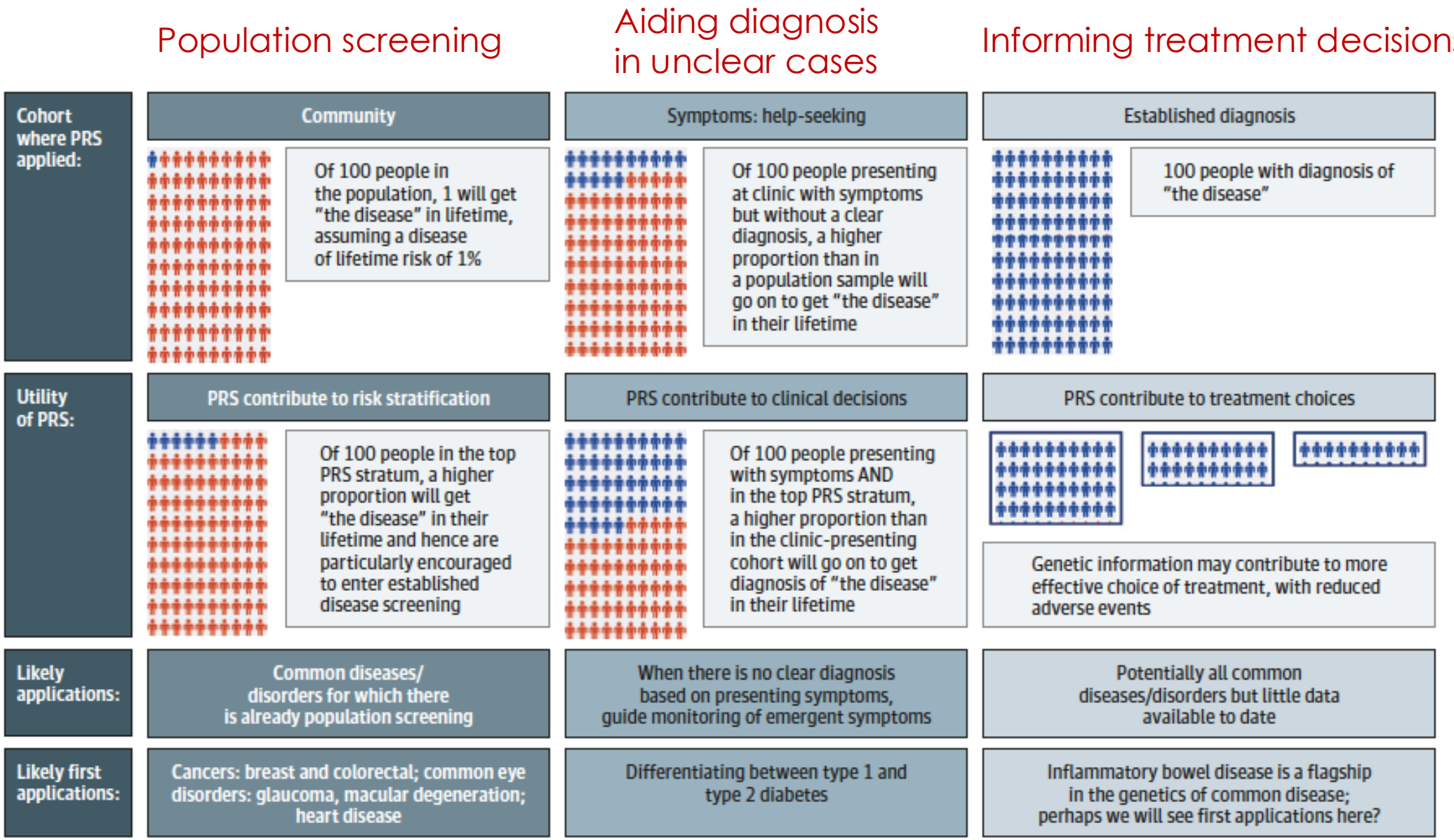
No family history, but by chance segregation of alleles has high genetic risk

Clinical applications

JAMA Psychiatry | Review

From Basic Science to Clinical Application of Polygenic Risk Scores
A Primer

Naomi R. Wray, PhD; Tian Lin, PhD; Jehannine Austin, PhD; John J. McGrath, MD, PhD; Ian B. Hickie, MD; Graham K. Murray, MD, PhD; Peter M. Visscher, PhD



Other applications of PGS

Identify correlates of genetic factors

e.g. Educational attainment PGS predicts early speech acquisition and is mediated by cognitive ability (Belsky et al., 2016).

Identify causal effects of genetic factors

Sibling data and family fixed effects → causal effect of PGS

Study GxE

e.g. Increase of compulsory schooling age in U.K. reduces BMI only among those with a high-BMI PGS (Barcellos, Carvalho, and Turley 2016)

Use as control variable

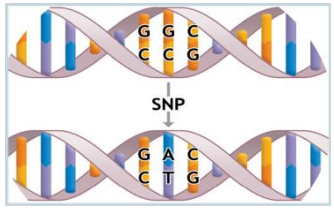
To control for confounding genetic factors or to increase statistical power for estimating the effect of a randomized treatment. If incremental R^2_{PGI} is 15%, then power increase is equivalent to 17% increase in sample size (Rietveld, 2013)

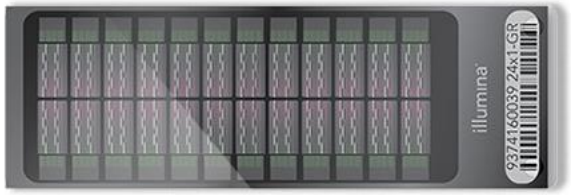
Genomic selection in livestock and crops

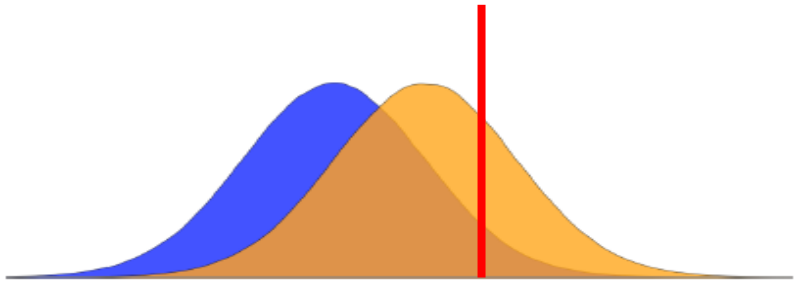
From Aysu Okbay

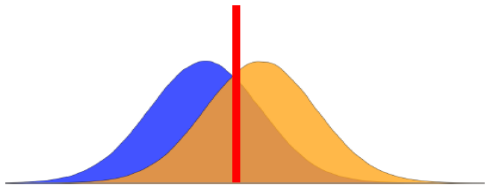
Justify for one disease and the rest come for free!

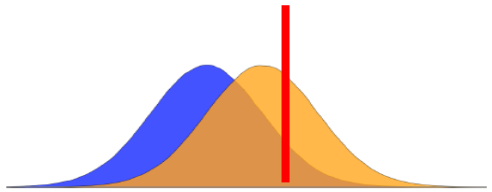
One disease

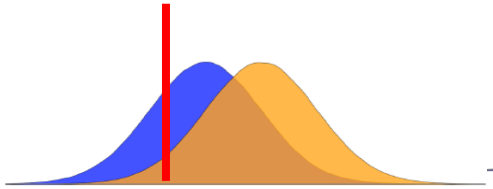


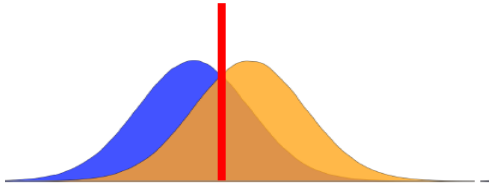


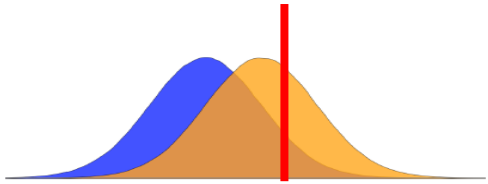


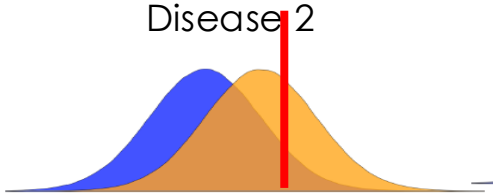

Disease 2

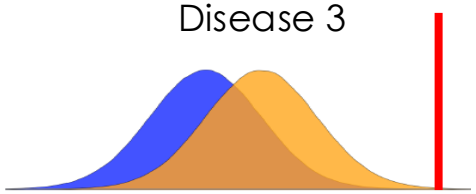

Disease 3

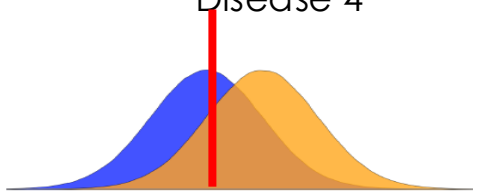

Disease 4

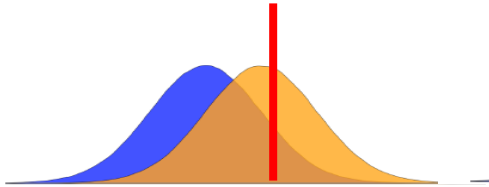

Disease 5

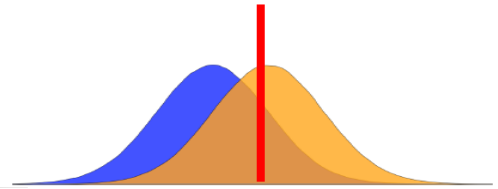

Disease 6

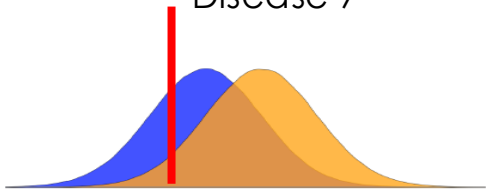

Disease 7

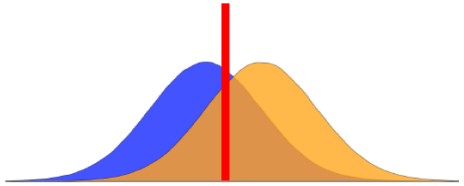

Disease 8

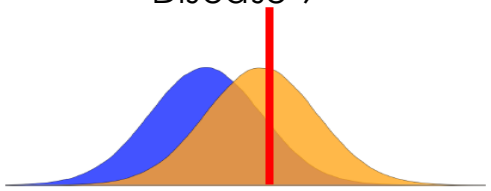

Disease 9

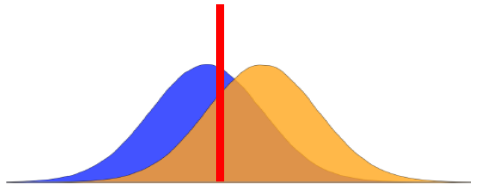

Disease 10

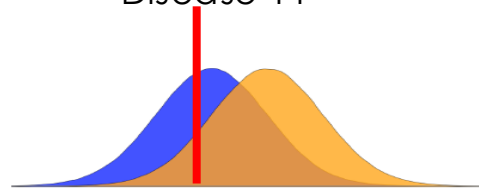

Disease 11


Disease 12


Disease 13


Disease 14


Disease 15

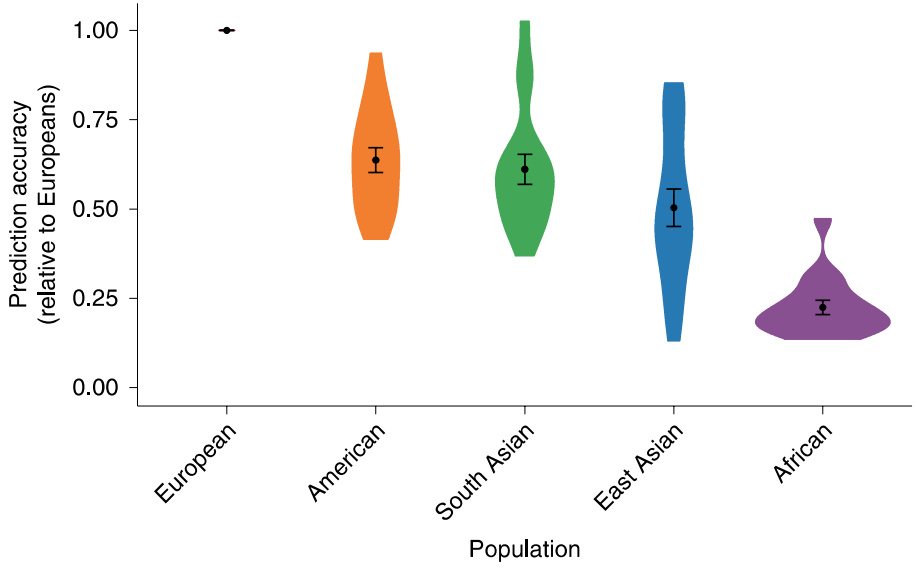

Disease 16

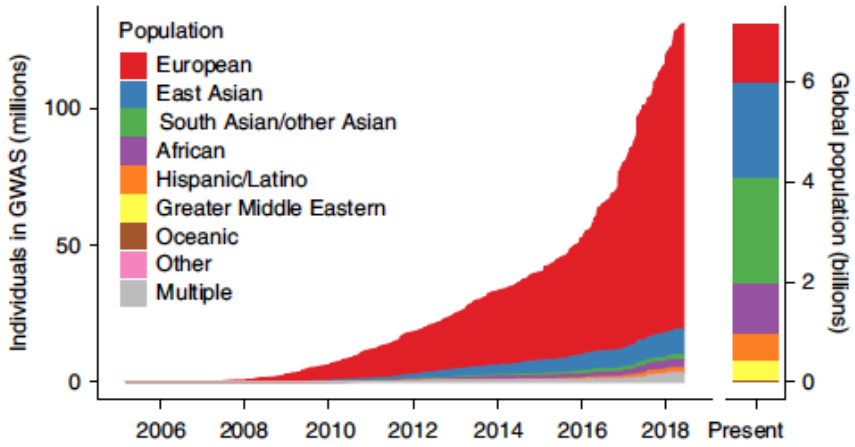
Poor trans-ancestry portability

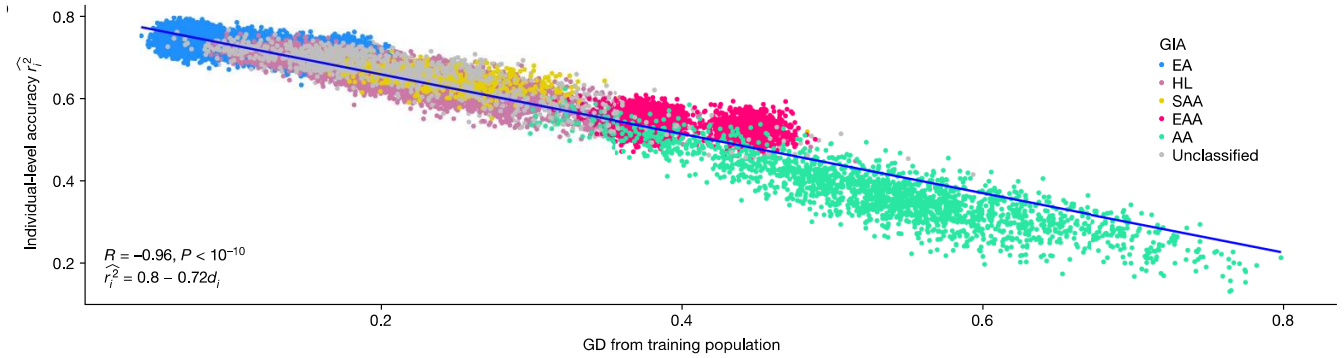
PERSPECTIVE
https://doi.org/10.1038/s41588-019-0379-x
nature genetics

Clinical use of current polygenic risk scores may exacerbate health disparities

Alicia R. Martin[1,2,3*], Masahiro Kanai[1,2,3,4,5], Yoichiro Kamatani[5,6], Yukinori Okada[5,7,8], Benjamin M. Neale[1,2,3] and Mark J. Daly[1,2,3,9]







nature
Explore content · About the journal · Publish with us
nature › articles › article
Article | Open Access | Published: 17 May 2023

Polygenic scoring accuracy varies across the genetic ancestry continuum

Yi Ding, Kangcheng Hou, Ziqi Xu, Aditya Pimplaskar, Ella Petter, Kristin Boulier, Florian Privé, Bjarni J. Vilhjálmsson, Loes M. Olde Loohuis & Bogdan Pasaniuc

Nature (2023) | Cite this article
11k Accesses | 200 Altmetric | Metrics

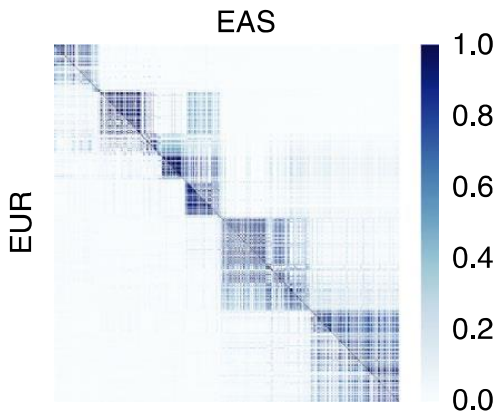
Poor trans-ancestry portability

Issues

- Same causal variants
  - Different allele frequencies
  - LD differences
  - Different effect sizes
- Different causal variants
  - GxE
  - Different phenotype

In *general:*
We expect common causal variants to be shared across ancestries

But correlation structure differs



Summary

- Polygenic scores are imperfect but useful genetic predictors.
- Their accuracy is fundamentally limited by heritability, SNP set, and sample size.
- A high PRS is mostly a consequence of genetic sampling.
- PRS have the potential to differentiate risk between family members who have the same family history information.
- Being evaluated in clinical settings and are often combined with other predictive measures to predict the total disease risk.
- One-off generation of genotypes provide PRS for multiple conditions.

PGS are not …

- Not diagnostic and never will be.
- Not absolute risk and do not provide a baseline or timeframe for the progression of a disease.
- Not and never will be stand-alone predictors of common diseases.
- Not equally applicable across populations – at least not yet.

Recommended reading

1. Wray NR, *et al*. Complex Trait Prediction from Genome Data: Contrasting EBV in Livestock to PRS in Humans: Genomic Prediction. *Genetics*. 2019 Apr;211(4):1131-1141. (**Review of polygenic prediction in livestock and humans**)

2. Wray NR, *et al*. From Basic Science to Clinical Application of Polygenic Risk Scores: A Primer. *JAMA Psychiatry*. 2021 Jan 1;78(1):101-109. (**Review of clinical application**)

3. Khera AV, *et al*. Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations. *Nat Genet*. 2018 Sep;50(9):1219-1224. (**Demonstration of utility**)

4. Martin AR, *et al*. Clinical use of current polygenic risk scores may exacerbate health disparities. *Nat Genet*. 2019 Apr;51(4):584-591. (**Poor cross-ancestry portability and consequences**)