QIMR Berghofer

Polygenic Score Workshop

Part 2: Evaluating Polygenic Scores

Jodi Thomas

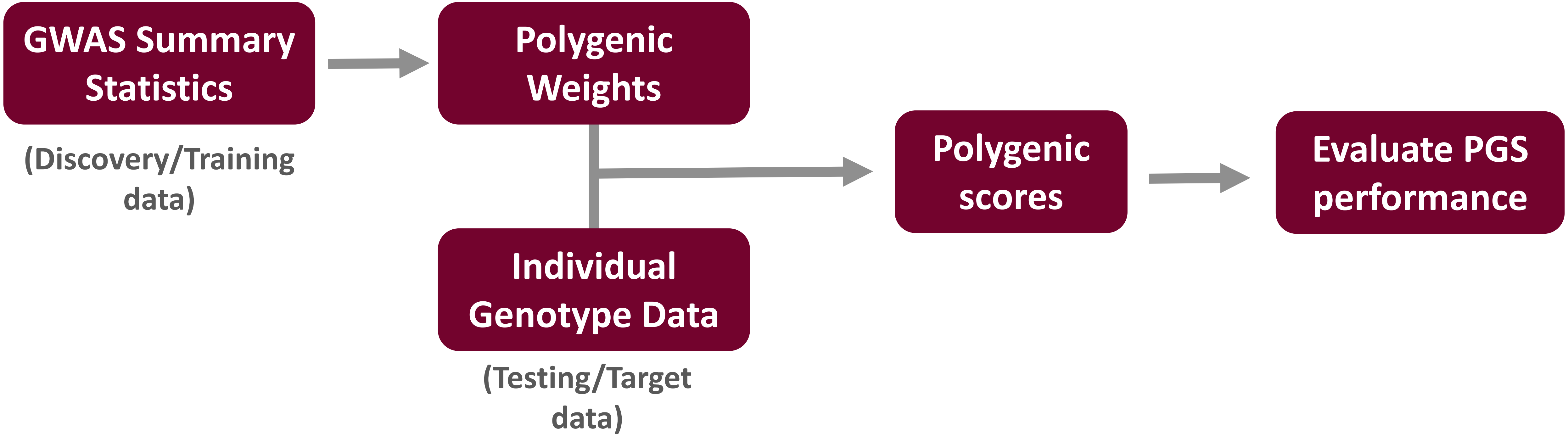
24th July 2025

Adelaide Medical School

Using Statistics to Evaluate PGS

- Data
- Statistics to evaluate PGS
  - Regression
  - R^2
  - Nagelkerke's R^2
  - R^2 on the liability scale
  - Odds Ratio by decile of PGS
  - AUC
- Applications
- Limitations
- Example

Data



GWAS Summary Statistics

(Discovery/Training data)

Polygenic Weights

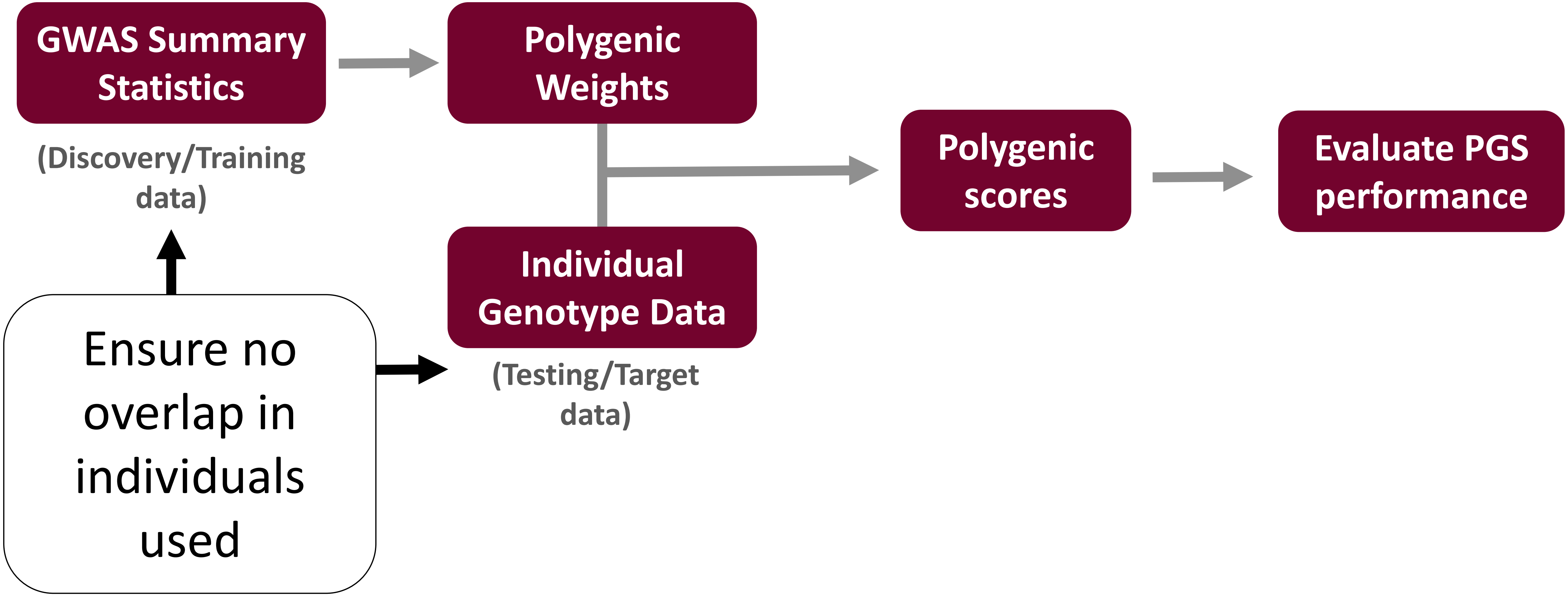
Individual Genotype Data

(Testing/Target data)

Polygenic scores

Evaluate PGS performance

Data



GWAS Summary Statistics

(Discovery/Training data)

Polygenic Weights

Individual Genotype Data

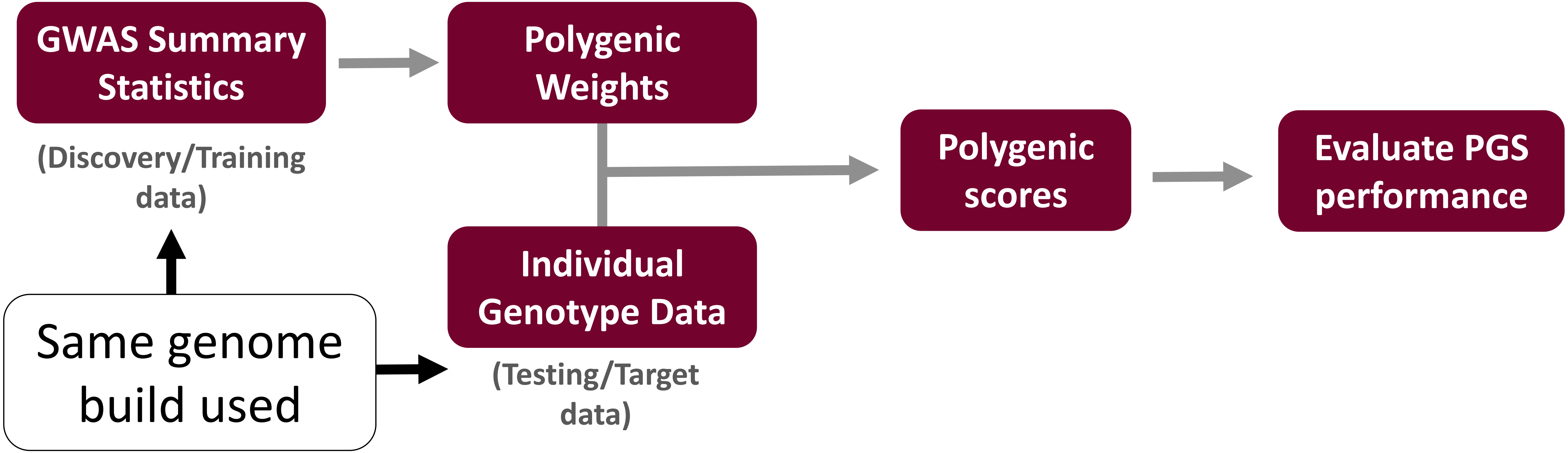
(Testing/Target data)

Polygenic scores

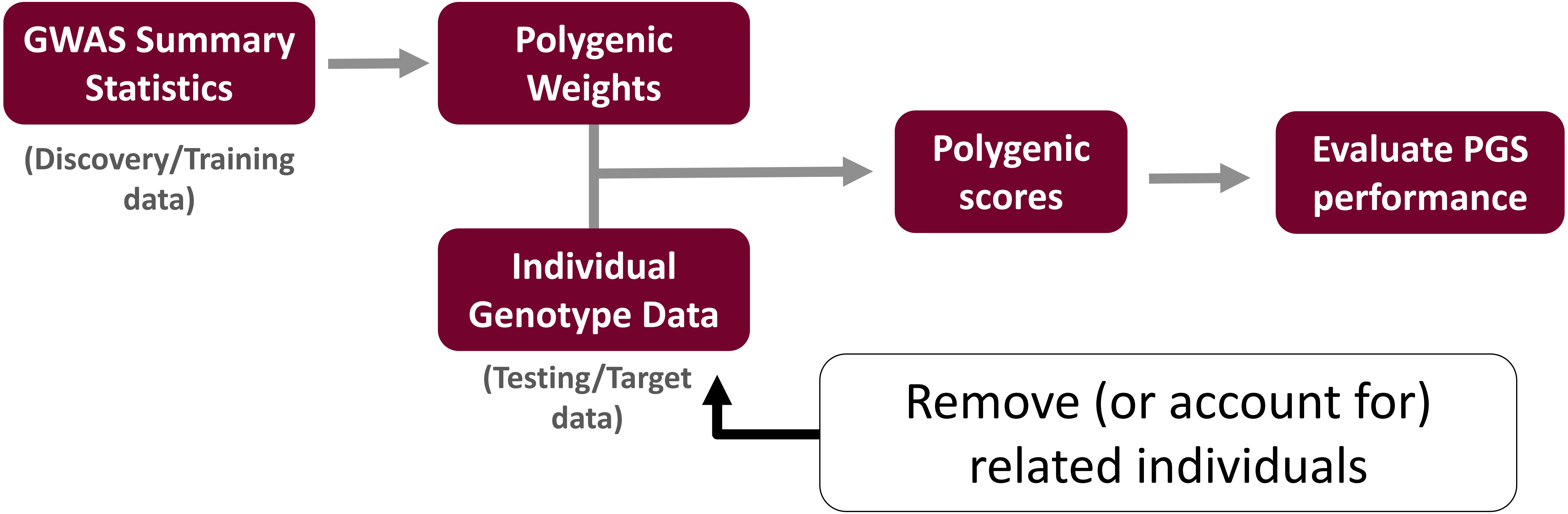
Evaluate PGS performance

Ensure no overlap in individuals used

Data


GWAS Summary Statistics
(Discovery/Training data)
Polygenic Weights
Individual Genotype Data
(Testing/Target data)
Polygenic scores
Evaluate PGS performance
Same genome build used

Data


GWAS Summary Statistics
(Discovery/Training data)
Polygenic Weights
Individual Genotype Data
(Testing/Target data)
Polygenic scores
Evaluate PGS performance
Remove (or account for) related individuals

Data

FID	IID	PGS	Sex	PC1	PC2	PC3	PC4	PC5	PC6	Epilepsy
HG00096	HG00096	-0.03122	1	0.000643	0.066432	-1.47E-02	-0.036	-0.01636	-0.02094	0
HG00097	HG00097	0.007768	2	0.001414	0.073602	8.82E-03	-0.02058	-0.01168	0.022409	0
HG00099	HG00099	-0.04946	2	0.002647	0.07177	-2.10E-02	-0.00609	-0.01414	-0.00713	0
HG00101	HG00101	-0.01496	1	0.001698	0.085445	-1.57E-02	-0.00289	-0.03352	-0.01412	1
HG00102	HG00102	-0.04752	2	0.004411	0.069636	1.76E-06	-0.02643	-0.04776	-0.03145	0
HG00103	HG00103	-0.02364	1	-0.00431	0.057179	-8.19E-03	-0.01349	0.015997	-0.01207	0

Standardised

QIMR Berghofer

Binary Traits

Continuous Traits

Regression

- Regression model: Trait ~ PGS + covariates

- Linear regression for continuous trait (e.g. height)
- Logistic regression for binary trait (e.g. case/control)

Regression

- Regression model: Trait ~ PGS + covariates

- Linear regression for continuous trait (e.g. height)
- Logistic regression for binary trait (e.g. case/control)

- Is there a significant association between trait and PGS?
- Is the association in the expected direction?

Continuous Traits

R^2 (Variance Explained)

- From the **linear regression**, estimate the R^2 that is attributable to the PGS

- R^2(full model) – R^2(covariates only model)

- Proportion of variance in the outcome variable explained by the PGS

Continuous Traits

R^2 (Variance Explained)

- Advantage
  - Comparable to SNP-based heritability (h^2)
- Limitation
  - R^2 can’t be compared across outcome traits on different scales (make sure to standardise outcome traits to allow comparison)

Binary Traits

Nagelkerke's R^2

- From the **logistic regression**, estimate Nagelkerke's R^2 that is attributable to the PGS

- Nagelkerke's R^2(full model) – Nagelkerke's R^2(covariates only model)

- A pseudo-R^2 value from 0 - 1
- A relative measure of model fit representing an approximation of explained variance

Binary Traits

Nagelkerke's R^2

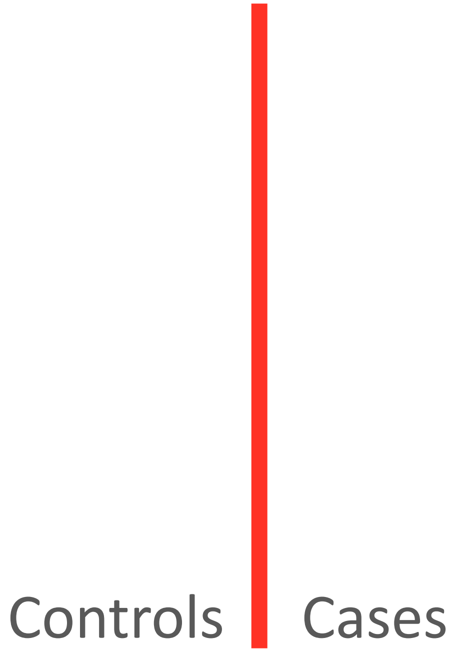
- Advantage
  - A familiar metric that is easy to compute
- Limitations
  - A relative measure of fit and can't be interpreted as 'proportion of variance explained' like linear R^2
  - On the binary observed scale
    - This depends on the case/control ratio in your sample
    - Can't be compared across studies with different case/control ascertainment
  - Is not comparable to SNP-based heritability (h^2)

Observed R^2

- Explains variation in case/control status
- Depends on case/control ratio in sample

Observed R^2

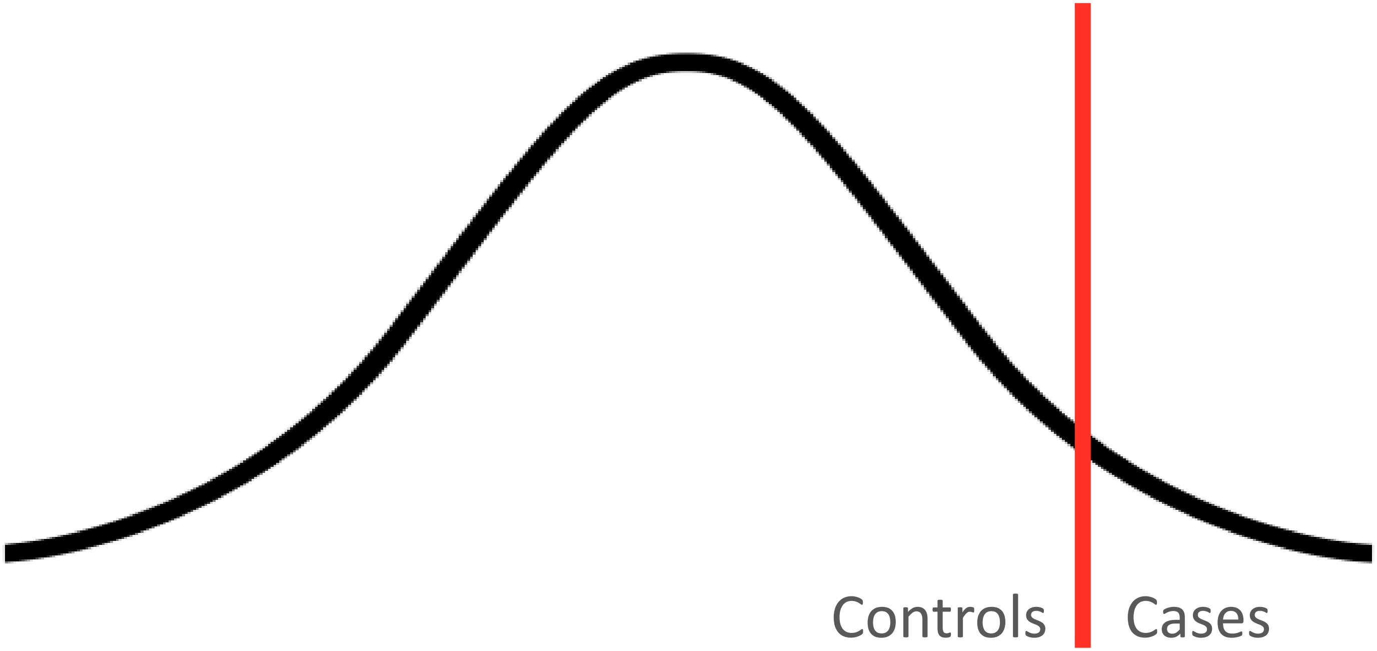
- Explains variation in case/control status
- Depends on case/control ratio in sample



Controls Cases

Binary Traits

Liability R^2

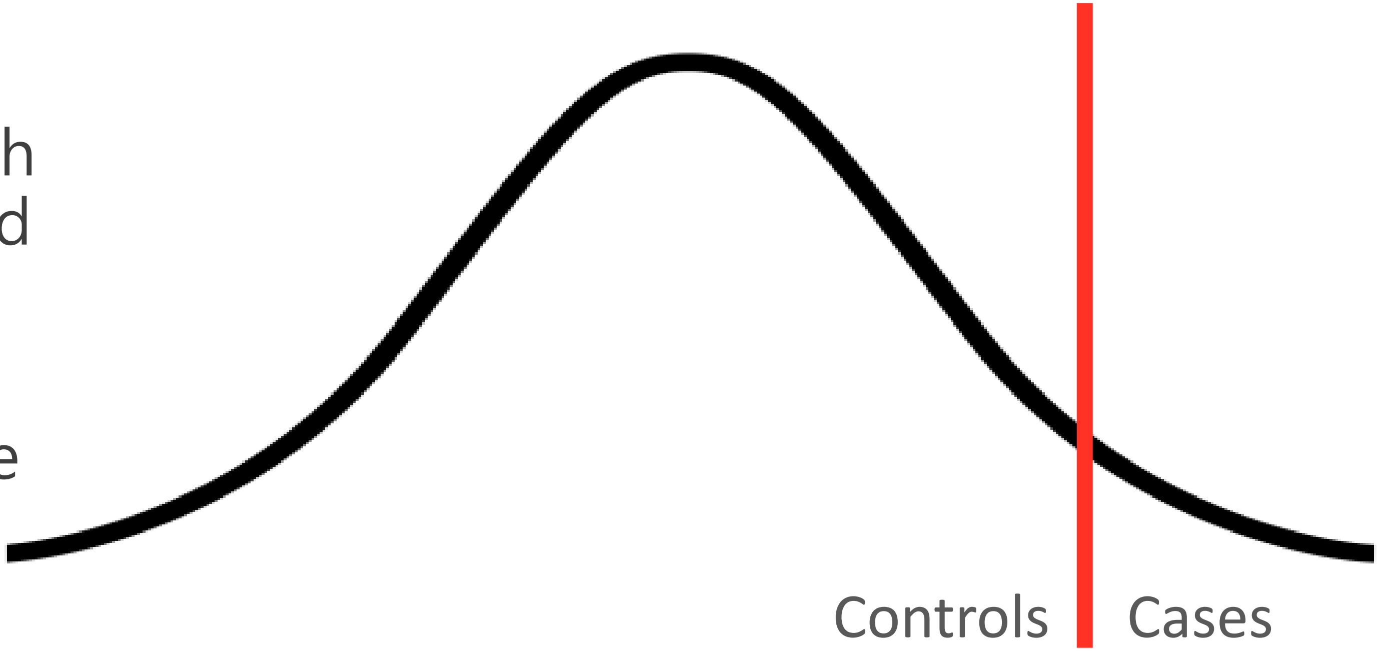


QIMR Berghofer

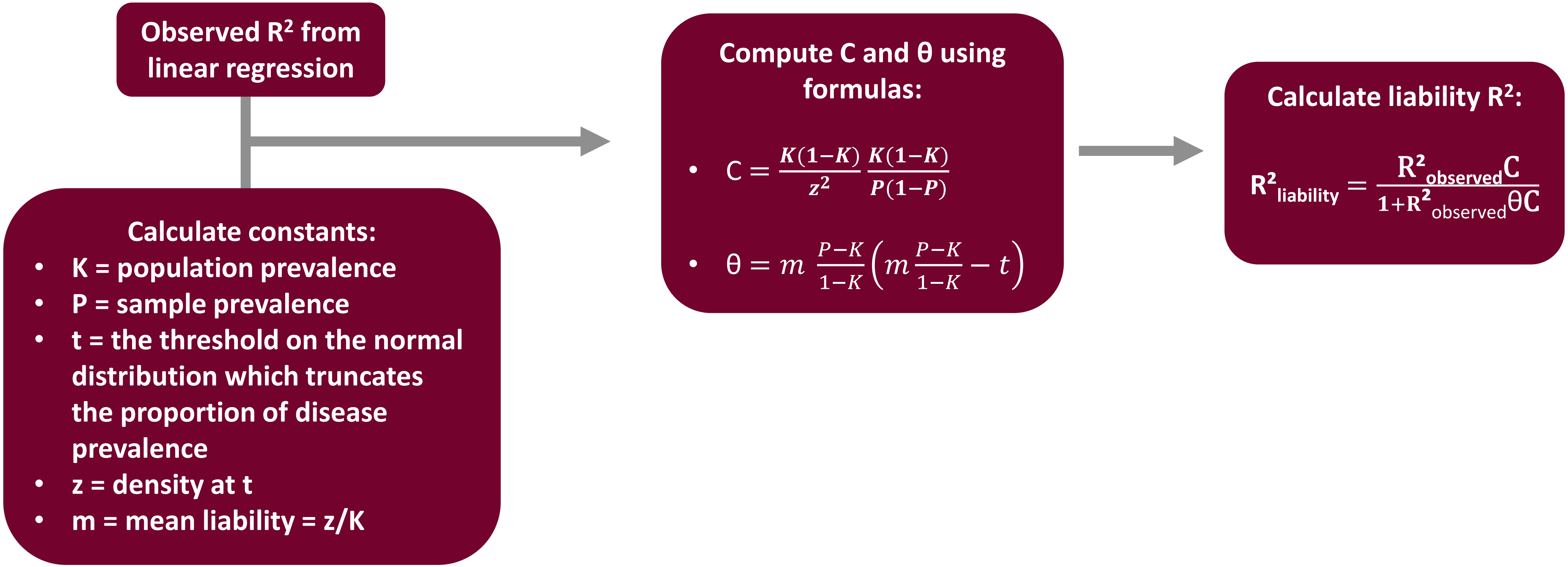
Binary Traits

Liability R^2

- Assumes there’s an underlying continuous risk for disease
- Liability R^2 estimates how much of that continuous, unobserved risk is explained by the PGS
- Adjusts for both population disease prevalence and sample ascertainment



Lee SH, et al., *A Better Coefficient of Determination for Genetic Profile Analysis.* Genetic Epidemiology, 2012. 36(3):214-224.



Observed R^2 from linear regression

Calculate constants:

- K = population prevalence
- P = sample prevalence
- t = the threshold on the normal distribution which truncates the proportion of disease prevalence
- z = density at t
- m = mean liability = z/K

Compute C and θ using formulas:

- $C = \frac{K(1-K)}{z^2}\frac{K(1-K)}{P(1-P)}$
- $\theta = m\,\frac{P-K}{1-K}\left(m\frac{P-K}{1-K} - t\right)$

Calculate liability R^2:

$$R^2_{\text{liability}} = \frac{R^2_{\text{observed}}C}{1+R^2_{\text{observed}}\theta C}$$

Liability R^2

- Steps
  - Run linear regression (even though outcome is binary)
  - Adjust R^2 to the liability scale
  - R^2 liability (full model) – R^2 liability (covariates only model)

- Proportion of variance in the unobserved liability (risk) for a binary trait that is explained by the PGS

Binary Traits

R^2 on the Liability Scale

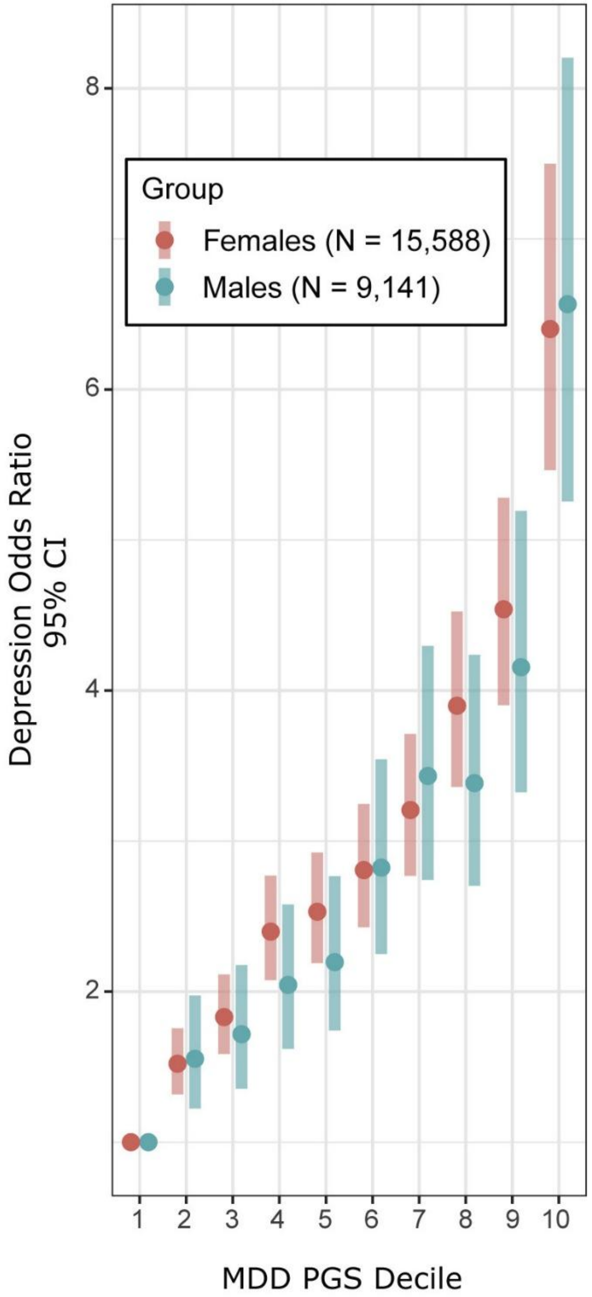
- Advantages
  - Can compare across studies
  - Comparable to SNP-based heritability (h^2)

- Limitations
  - Requires the population prevalence of the trait
  - Assumes an underlying liability threshold model

Odds Ratio by Decile of PGS

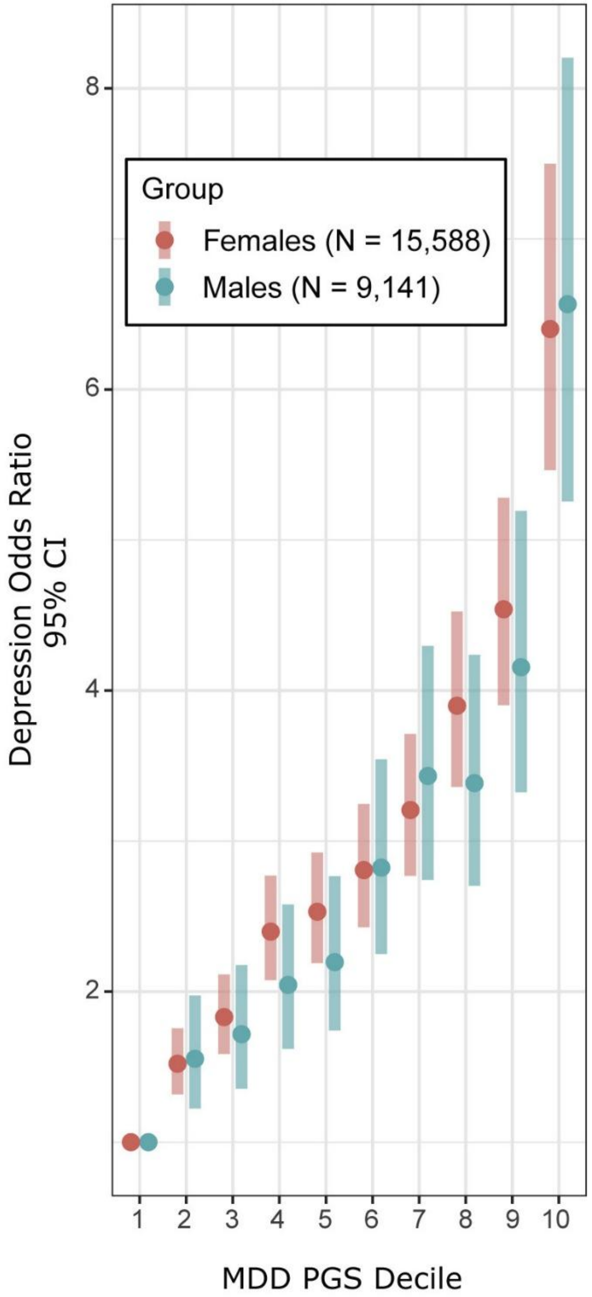
- Cut PGS distribution into deciles
- Run logistic regression: Trait(case/control) ~ PGS(deciles) + covariates
- Odds ratio of each PGS decile compared to the:
  - First (lowest) PGS decile
  - Middle PGS decile (more recently)

Binary Traits

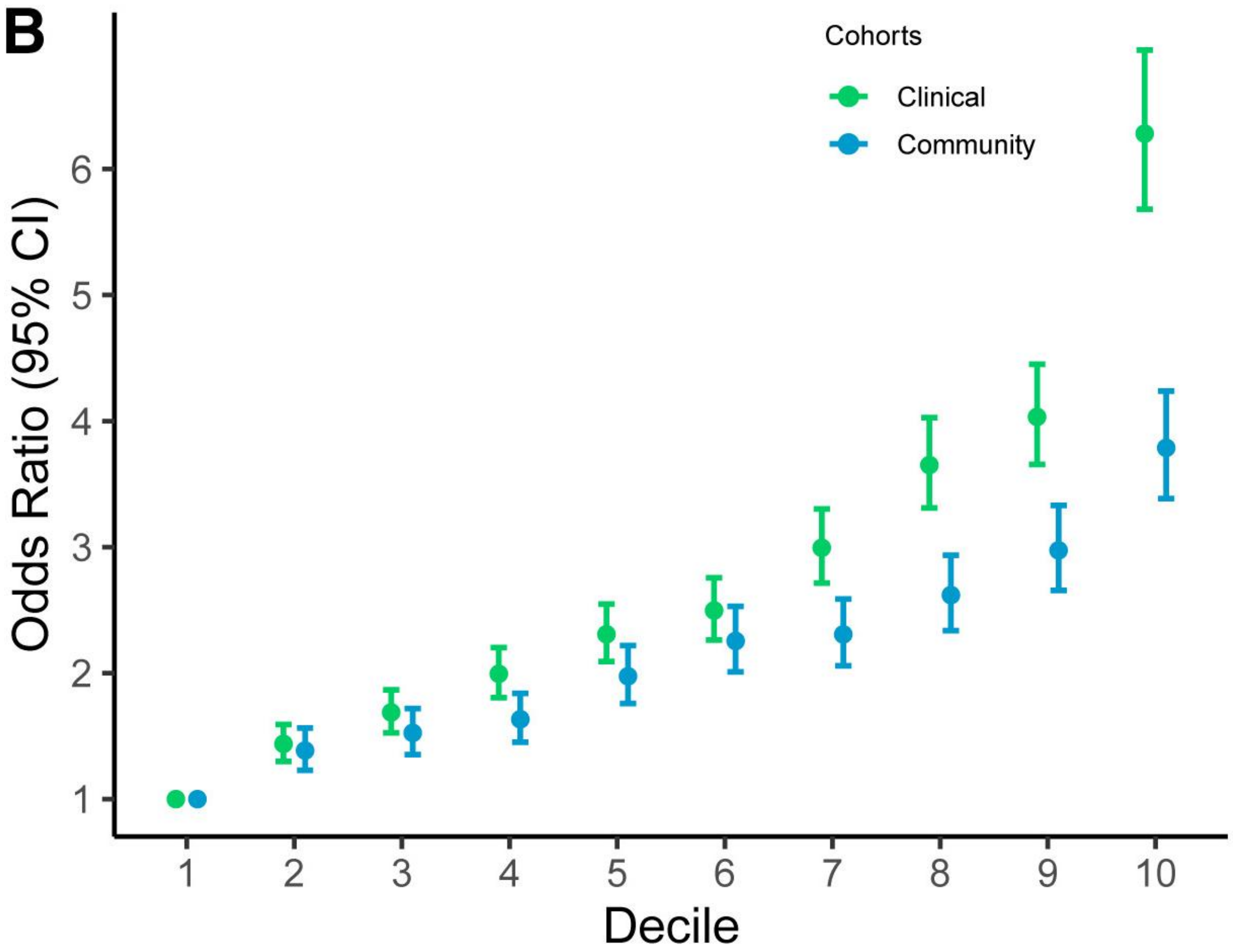


Mitchell BL, et al., *The Australian Genetics of Depression Study: New risk loci and dissecting heterogeneity between subtypes. Biol Psychiatry, 2022. 92(3):227-235.*

Binary Traits



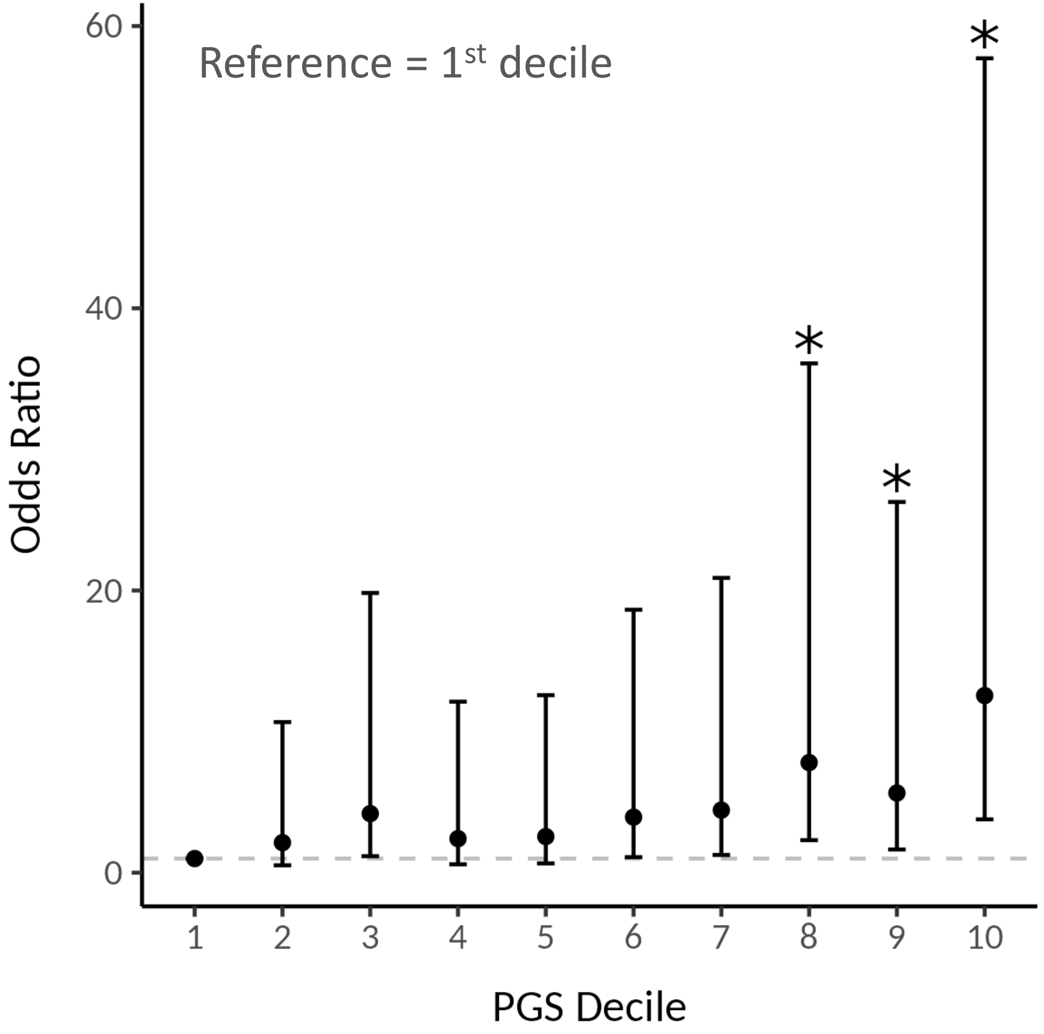
Mitchell BL, et al., *The Australian Genetics of Depression Study: New risk loci and dissecting heterogeneity between subtypes. Biol Psychiatry, 2022. 92(3):227-235.*

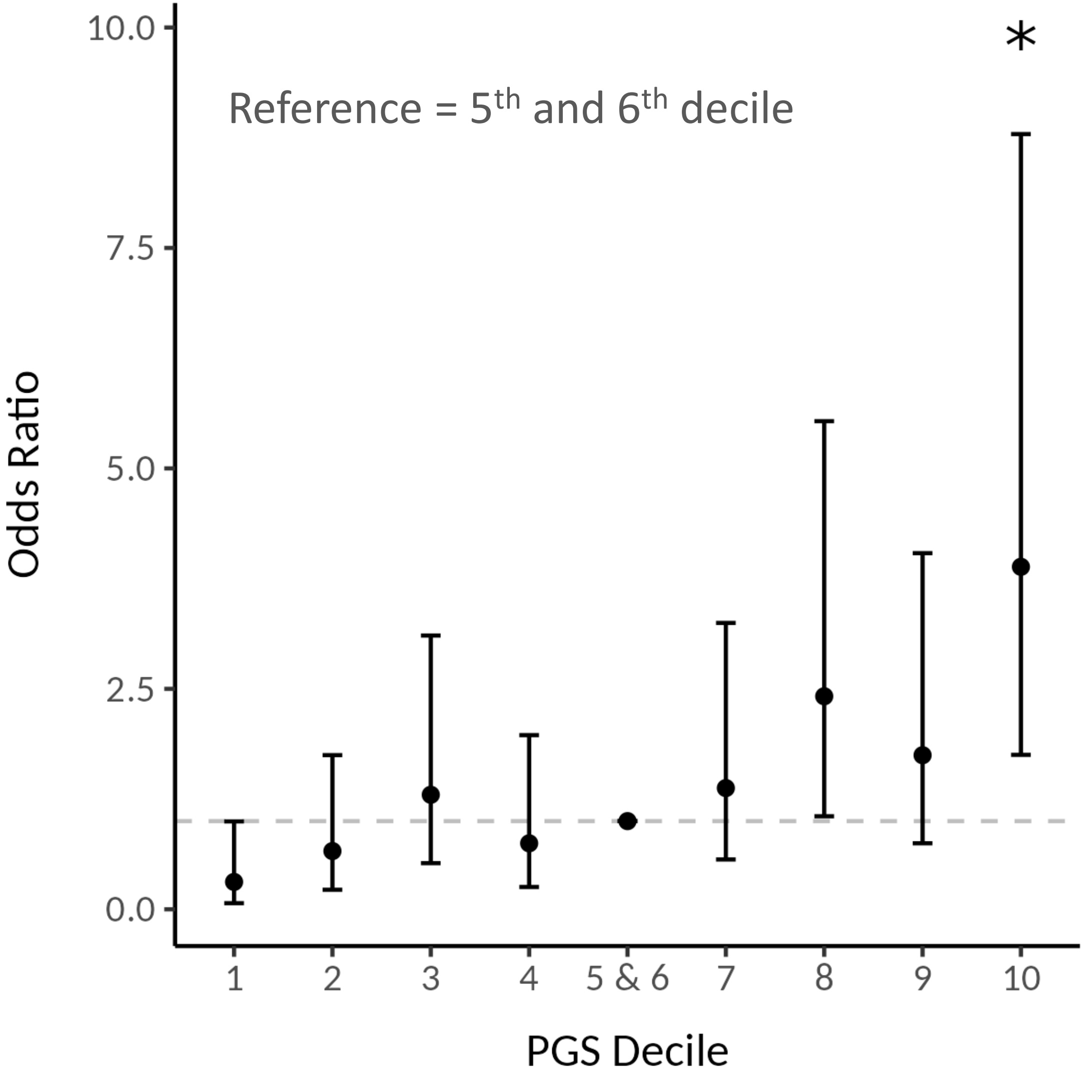


Adams MJ, et al., *Trans-ancestry genome-wide study of depression identifies 697 associations implicating cell types and pharmacotherapies. Cell, 2025. 188:1-13.*

Example from workshop (simulated data)

Binary Traits





Odds Ratio by Decile of PGS

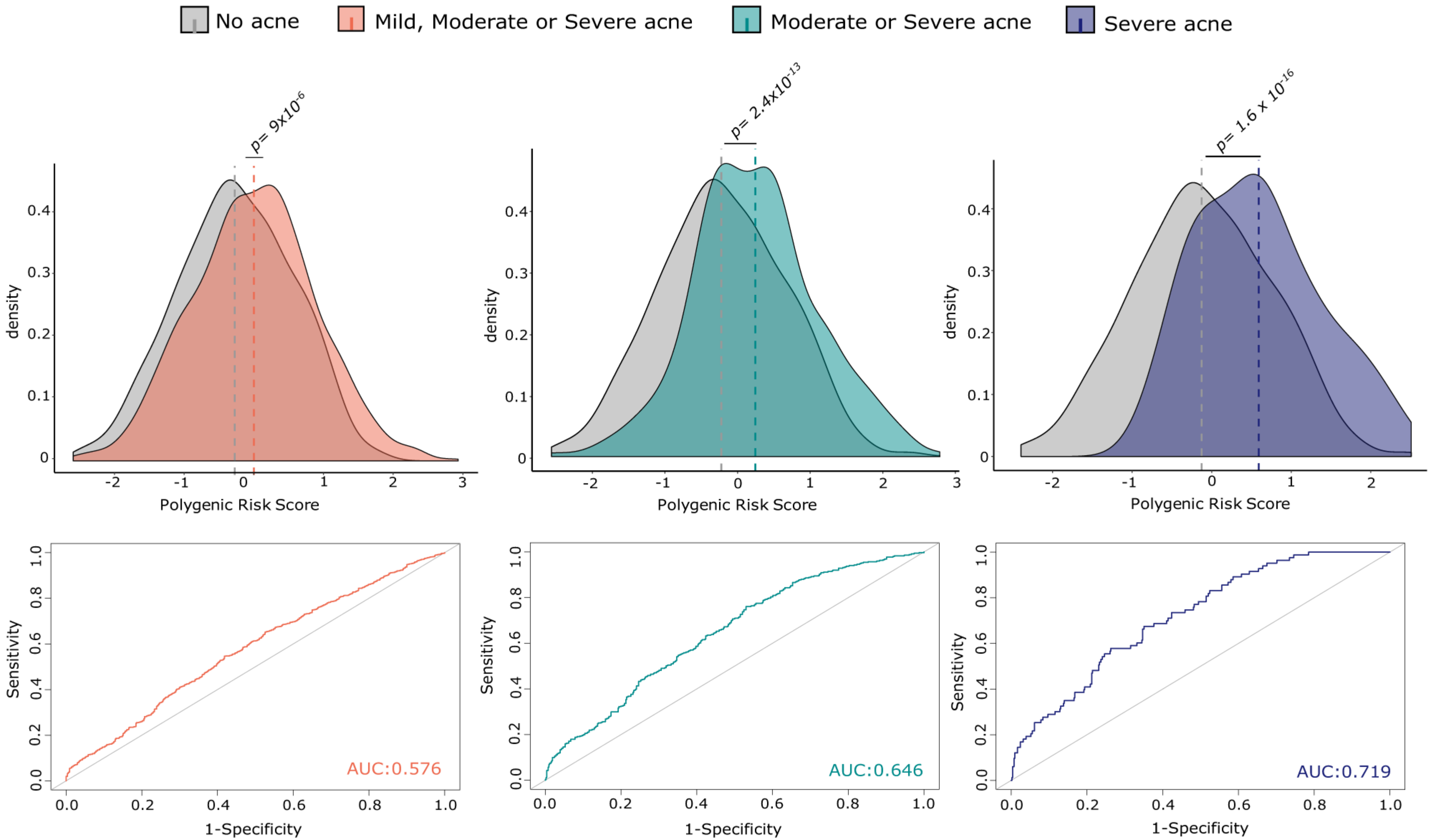
- Advantage
  - Practical and interpretable way to visualise risk stratification

- Limitations
  - Doesn’t take into account proportion of cases and controls in your data
    - Will look much more impressive if you have a data set with 50% cases and 50% controls, compared to a population sample.
  - Categorising PGS into deciles loses information

AUC

- Area Under the Receiver Operating Characteristic Curve
- Probability that a randomly selected case has a higher test score than a randomly selected control
  - 0.5 = no discrimination of cases and controls
  - 1 = perfect discrimination

Binary Traits



Mitchell BL, et al., *Genome-wide association meta-analysis identifies 29 new acne susceptibility loci. Nat Commun, 2022. 13(1):702.*

AUC

- Advantages
  - Well established measure
  - Independent to proportion of cases and controls in sample

- Limitations
  - Problem with genetic interpretation
    - PGS is a proxy (AUC is limited by how well the PGS captures genotype–phenotype associations)
    - The maximum AUC achievable depends on the heritability of the disease
    - A low AUC doesn’t necessarily mean genetics don’t matter

Best to report multiple measures

Make your results accessible to both geneticists and clinicians

Applications

Discovery Sample	Target Sample	Application	Biological Insight
Disorder A	Disorder A	Show polygenicity	PGS can predict outcomes even without genome-wide hits
Disorder A	Disorder B	Test pleiotropy	Reveals shared genetic architecture between conditions
Disorder A	Subtypes of Disorder A	Investigate heterogeneity	Subtypes may have distinct genetic contributions
Disorder A	Disorder A + environment data	Explore GxE	Genes may act differently depending on environmental context
Disorder A	Environmental exposure or presence of a trait	Explore gene-environment/trait correlations	Certain exposures/traits may be genetically linked
Disorder A	General population / clinical cohorts	Identify at-risk individuals	Risk stratification to identify individuals at high genetic risk

Limitations

- Individual-level prediction is not accurate enough for most phenotypes
    - Not reliable for individual diagnosis or decision-making
- Difficult to interpret what the PGS is truly capturing
    - Includes many variants with unknown function
    - Predictive power may reflect not only the causal effect of genetic variants but also gene-environment correlations, population stratification, indirect genetic effects, assortative mating
- Poor transferability across ancestries
    - PGS mainly developed in European populations and underperform when used in other ancestries

Example

Do polygenic scores (PGS) for Cerebral Palsy and comorbid traits associate with Cerebral Palsy?

These slides from the workshop have been removed due to being unpublished results

QIMR Berghofer

Dr. Jodi Thomas Research Officer

✉ jodi.thomas@qimrb.edu.au

Github Repository
Polygenic Score Workshop

https://github.com/jodithea/
Polygenic_score_worshop