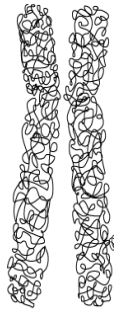


Introduction to Polygenic Risk Scores (PRS) and Mendelian Randomization (MR)

Linda Kachuri, MPH, PhD

Postdoctoral Scholar, Department of Epidemiology & Biostatistics

Genetic and Molecular Epidemiology: EPID 217

February 16, 2021

UCSF

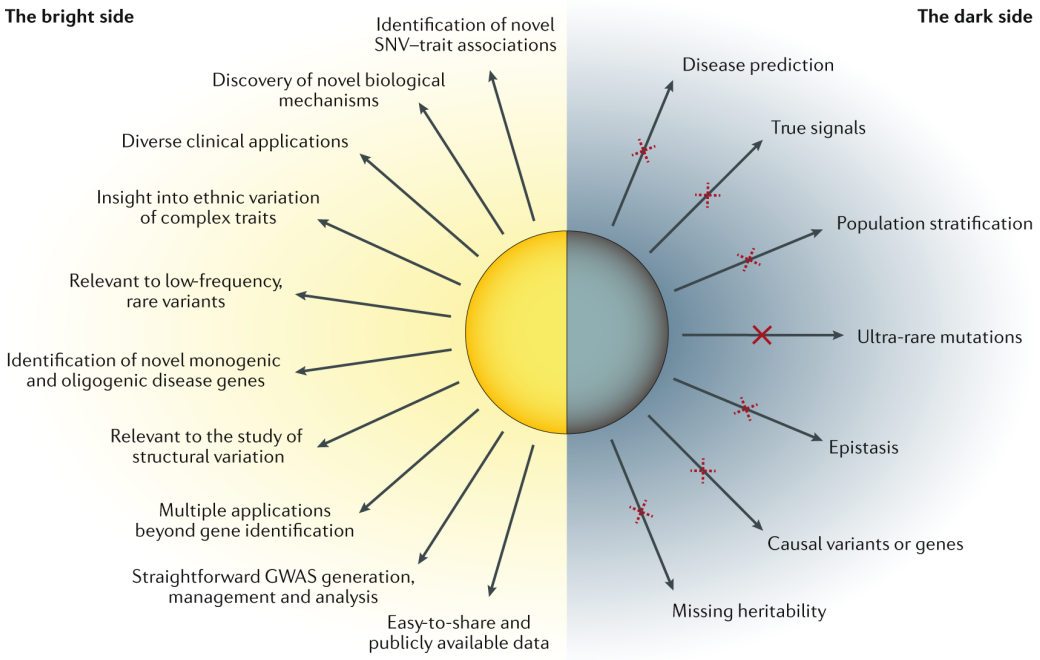
Introduction

Motivation

GWAS have yielded a wealth of germline variants associated with different **traits** and **diseases**

However, the translational impact of GWAS results has been limited

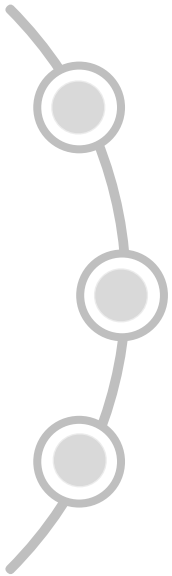
- Predict disease occurrence?
  - Polygenic risk scores (PRS)
- New insights into etiology?
  - PRS, Mendelian randomization



Tam et al. *Nature Reviews Genetics* (2019) 20: 467 – 484

Outline

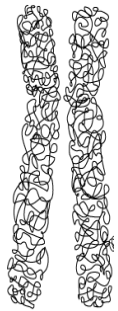
Lecture Overview



I. Polygenic risk score (PRS) basics:
   - Understand the goals of a PRS analysis
   - Learn how to construct and apply a PRS

II. Critically evaluating PRS studies:
   - Discuss Lambert et al. (2019) article

III. From association to causation: Mendelian Randomization (MR)
   - Understand the goals of an MR analysis
   - Understand similarities and differences between MR and PRS



Part I:

Polygenic risk score (PRS) basics

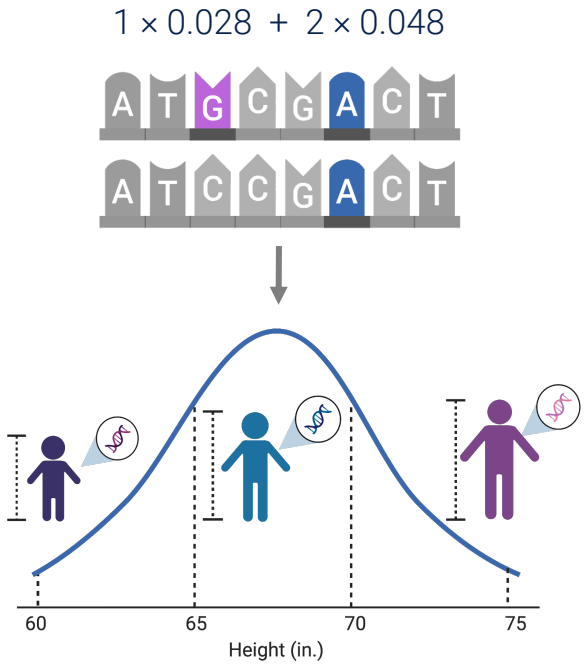
1.

PRS

Introduction

What is a polygenic risk score (PRS) / genetic risk score (GRS)?

- Quantitative measure of an individual's cumulative genetic risk (disease) or predisposition (traits)
- Quantitative summary of one's genetic vulnerability (disease) or propensity (trait)
  - Sum of risk alleles weighted by their effect sizes (beta coefficients from linear or logistic regression models or other effect measures)
  - Typically comprised of independent variants reaching a given p-value threshold



UCSF

PRS

Introduction

What can a PRS be used for?

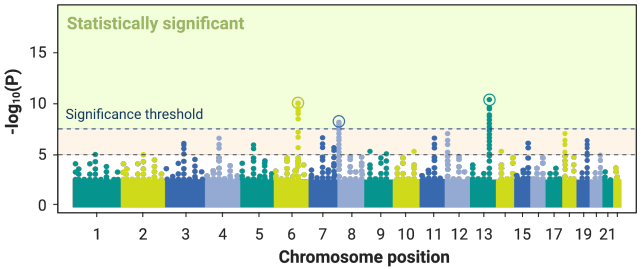
- Genetic prediction: prioritize individuals with a genetic predisposition to adverse health outcomes for appropriate interventions
  - High PRS for breast cancer → earlier or more frequent screening
  - Low PRS for vitamin D → increased dietary supplementation
- Insight into biological pathways: shared genetic basis between traits
  - High PRS for BMI → Increased risk of cardiovascular disease
  - High PRS for cigarettes/day → Decreased forced vital capacity (lung function)

2.

PRS

Overview

Input data: GWAS summary stats



1. Select associated variants
2. Obtain risk alleles and effect sizes

Target data: independent population

1. Calculate PRS: weighted allele sum
2. Evaluate associations with outcome(s)

UCSF

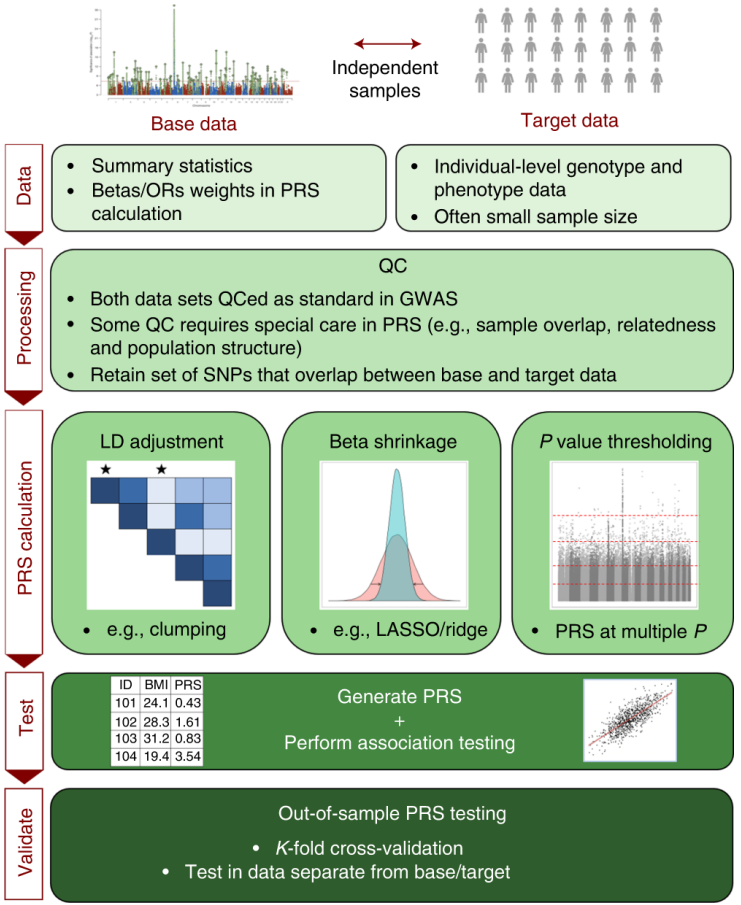
2.

PRS

Workflow

High-level analytic decisions:

- How to select variants for inclusion in the PRS?
  - Consider association thresholds (p-values), LD structure, variant frequency, functional features
- How to assign weights to variants in the PRS?
  - Are the variants independent or do the weights (betas) need to be shrunk to account for LD?
- Before you PRS conduct GWAS QC in the target sample
- Data processing is crucial for calculating the PRS
  - Determine the effect allele and align betas appropriately
  - Obtain LD proxies for missing variants



Choi et al. *Nature Protocols* (2020) 15: 2759 – 2772

PRS

Data Processing

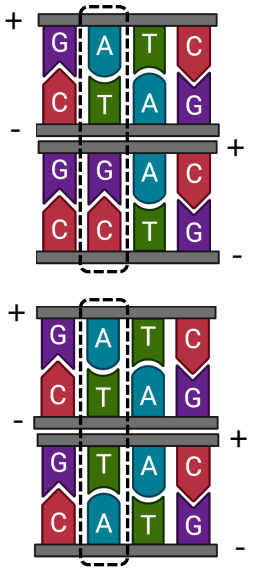
Step 1: obtain or generate GWAS summary statistics for your trait

- Conduct literature search to identify all relevant studies and collate results
- Identify a single data source: largest/most recent GWAS meta-analysis

Step 2: perform basic data checks on the input summary statistics

- ✓ SNP identifiers (rs ID, chr:pos), chromosome, and position are present
  - Watch out for differences in genome build between datasets
- ✓ Both alleles are reported (A1/A2) and **effect allele (EA)** is not ambiguous
  - Exclude A/T, T/A, G/C, C/G variants or resolve strand ambiguity
- ✓ For continuous traits weights are **β** from linear regression
  - Watch out for units: scales across studies (cm/in), transformations (log)

  For dichotomous traits weights are derived from odds ratios: **β=log(OR)**
- ✓ Only use associations from well-imputed SNPs (in both datasets)



rs123: A/G [T/C]
effect = A
MAF (A)= 0.25

rs456: A/T [T/A]
effect = A ?
MAF (A)= 0.25
MAF (A)= 0.49

PRS

PRS Calculation

Clumping and Thresholding (C+T)

- Select the variant with the lowest p-value per LD region
  - **Clumping ≠ pruning:** selects one SNP from correlated pair, typically with higher MAF, agnostically (without considering p-values)
- P-value thresholds: traditional PRS use $P<5\times10^{-8}$, but adding more variants results in a more powerful PRS and improves predictive performance
  - **P-value tuning:** calculate PRS over a range of p-values and select one that maximizes prediction
  - However, this incurs a high multiple testing burden, requires separate train, test, and validation datasets, and access to full GWAS summary statistics
- LD thresholds: traditional PRS use independent variants (e.g.: LD<0.1)
  - Inclusion of variants in LD with original effect sizes leads to overfitting

Privé et al. *AJHG* (2019) 6(5): 1213 – 1221

https://privefl.github.io/bigsnpr/articles/pruning-vs-clumping.html

PRS

PRS Calculation

Standard PRS calculation for C+T:

$$PRS = \sum_{j=1}^{n} x_{ij}\,\hat{\beta}_j$$

- $\hat{\beta}_j$ ← Effect size of SNP j (marginal effect rom additive model)
- x_{ij} ← Genotype or dosage for SNP j in individual i

Alternative PRS calculation for C+T that accounts for uncertainty in SNP effect size

- Crude shrinkage that is helpful when PRS inputs are sourced from multiple studies

$$PRS_{IV} = \sum_{j=1}^{n} x_{ij}\,\frac{\hat{\beta}_j}{SE(\hat{\beta}_j)}$$

- $SE(\hat{\beta}_j)$ ← Standard error of SNP j effect size

Kachuri et al. *Nature Communications* (2020) 11: 6084

2.

PRS

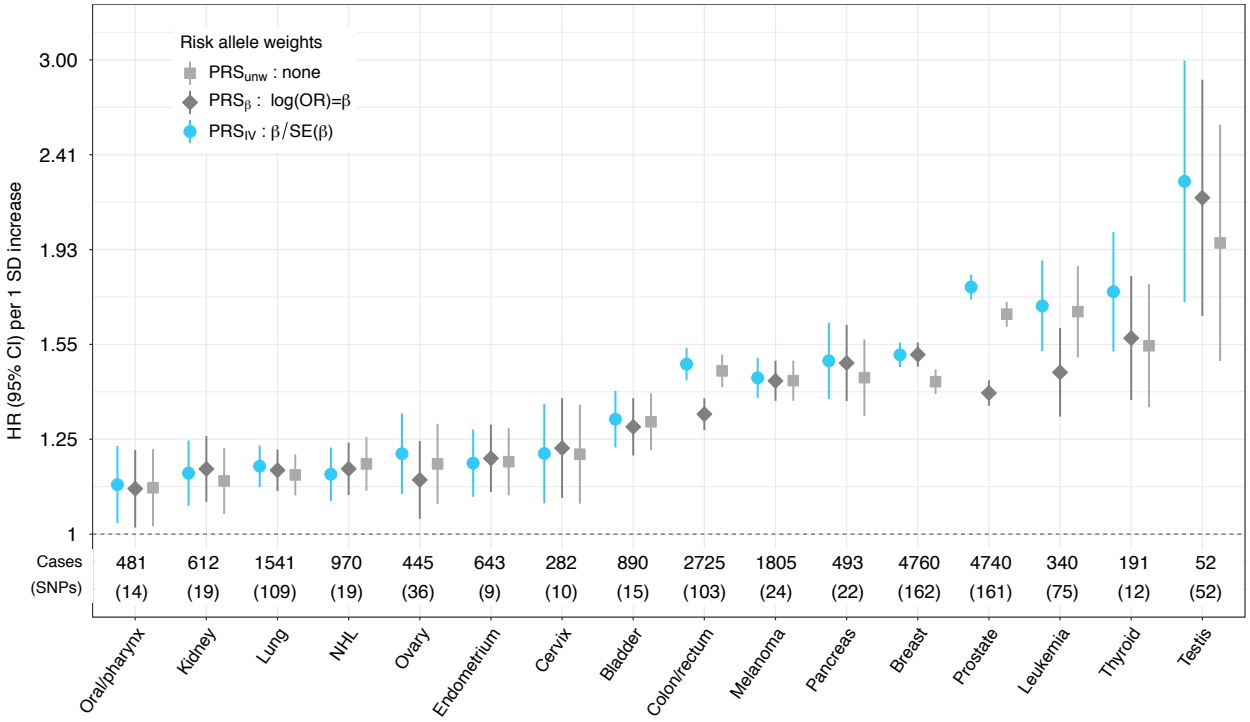
PRS Calculation

PRS for 16 cancers using C+T

- Derived from systematic search of all published GWAS
- $P<5\times10^{-8}$ and $r^2<0.30$

Applied to UK Biobank (external population)

PRS_{IV} shows stronger or equivalent associations to standard PRS for most cancers



Kachuri et al. *Nature Communications* (2020) 11: 6084

UCSF

PRS

PRS Calculation

More advanced methods using penalization to account for LD structure

- Best Linear Unbiased Predictor (BLUP)
  - Jointly estimated effect sizes for variants using linear mixed models
  - Assumes an infinitesimal (Gaussian) genetic architecture where all variants are causal, but some effect sizes are shrunk close to 0
  - Requires individual-level data to calculate genetic variance-covariance matrix, which is computationally intensive for large datasets
- More recent extensions: sBLUP
  - Removes the requirement for individual-level data
  - Uses external LD reference to obtain BLUP (joint-fit) effect sizes
  - Conceptually similar to LDPred

Yang et al. *Am J Hum Genet* (2011) 88(1): 76 – 82
Robinson et al. *Nat Hum Behavior* (2017) 1: 0016

PRS

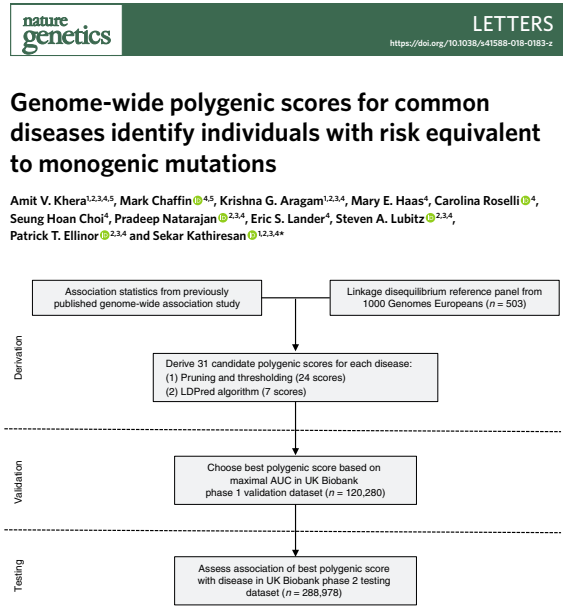
PRS Calculation

More advanced methods using penalization to account for LD structure

- LDPred: Bayesian version of BLUP using external LD reference
  - Posterior mean effect size of each SNP is estimated by using a prior:

$$\beta_i \sim iid \begin{cases} N\left(0, \frac{h_g^2}{Mp}\right) \text{ with probability } p \\ 0 \text{ with probability } (1-p) \end{cases}$$

  - Where h_g^2 is the heritability, M is the number of variants, p is the proportion of causal variants, which is unknown
  - Requires training and testing to tune p, computationally intensive
- LDPred-2: new and improved LDPred
  - Claims to better handle long-range LD (HLA region)
  - Includes a sparse model where some weights are shrunk to 0
  - Parallelized computation by chromosome



Vilhjálmsson et al. *Am J Hum Genet* (2015) 97(4): 576 – 592
Privé et al. *Bioinformatics* (2020) btaa1029

PRS

PRS Calculation

More advanced methods using penalization to account for LD structure

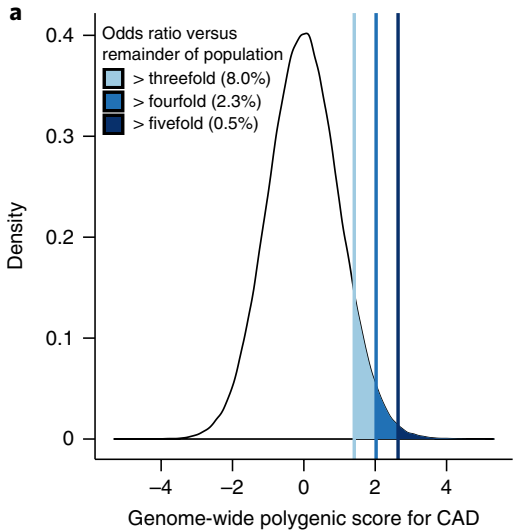
- Using individual-level data
  - Re-estimate joint SNP effects using LASSO, Elastic Net, Ridge, or any other shrinkage method
  - Consider pre-selecting variants ($p<0.01$ and LD $r^2<0.8$) to reduce computational burden (but not going as far as C+T)
  - Requires adequate training, testing, and validation sample size
- Using summary statistics: lassosum
  - Weights obtained by applying elastic net penalization
  - Tuning parameters derived using an external LD reference panel

Mak et al. Genetic Epidemiology (2017) 41(6): 469 – 480

PRS

Association Testing

- Raw PRS is usually mean centered and scaled to have SD=1
- Models should include appropriate covariates (age, PC's)
- For continuous traits fit linear regression models:
  - Estimate β per unit increase in PRS
  - Examine the variance explained (R^2) by the PRS
  - Partial R^2: correlation between the trait and PRS after having residualized both of them using a set of covariates
- For dichotomous (disease) traits use logistic regression:
  - Estimate OR per unit increase in PRS
  - When comparing OR's across PRS deciles or other cut-points use the average-risk group as the reference
  - Assess predictive performance: AUC, pseudo R^2, Brier score, NRI



Khera et al. *Nature Genetics* (2018) 50: 1219-1224

3.

PRS

Association Testing

Applied example: breast cancer

Are all these PRS methods really that different?

Table 2. Association between PRS and Breast Cancer Risk in the Validation Set and Prospective Test Datasets

	Validation Set			Prospective Test Set		
	OR[a]	95% CI	AUC	OR[a]	95% CI	AUC
77 SNP PRS (PRS_{77})						
Overall BC	1.49	1.44–1.56	0.612	1.46	1.42–1.49	0.603
ER-positive	1.56	1.49–1.63	0.623	1.52	1.48–1.56	0.615
ER-negative	1.40	1.30–1.50	0.596	1.35	1.27–1.43	0.584
313 SNP PRS (PRS_{313})						
Overall BC	1.65	1.59–1.72	0.639	1.61	1.57–1.65	0.630
ER-positive	1.74	1.66–1.82	0.651	1.68	1.63–1.73	0.641
ER-negative	1.47	1.37–1.58	0.611	1.45	1.37–1.53	0.601
3,820 SNP PRS (PRS_{3820})						
Overall BC	1.71	1.64–1.79	0.646	1.66	1.61–1.70	0.636
ER-positive	1.81	1.73–1.89	0.659	1.73	1.68–1.78	0.647
ER-negative	1.48	1.37–1.59	0.611	1.44	1.36–1.53	0.600

Mavaddat et al. *Am J Hum Genet* (2019) 104(1): 21 – 340

PRS

Association Testing

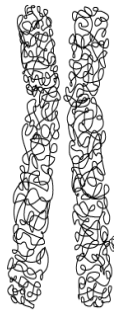
Applied example: breast cancer

Are all these PRS methods really that different?

Table 3. Comparison of PRS Methods on Breast Cancer PRS Performance in MGI and UKB

Method Tuning Parameter	# SNPs	Pseudo-R^2	Brier Score	AAUC (95% CI)	Odds Ratio Continuous PRS (95% CI)	Odds Ratio Top 1% (95% CI)
MGI Cohort						
Lassosum: s = 0.5, λ = 0.0043	*118,388*	*0.0592*	*0.134*	*0.641 (0.625,0.656)*	*1.70 (1.59,1.81)*	2.48 (1.63,3.77)
P&T: $p \leq 4 \times 10^{-4}$	3,038	0.0532	0.135	0.635 (0.618,0.651)	1.64 (1.54,1.75)	2.84 (1.88,4.28)
Fixed threshold: $p \leq 5 \times 10^{-5}$	1,307	0.0521	0.135	0.634 (0.618,0.65)	1.63 (1.53,1.74)	2.75 (1.81,4.17)
Fixed threshold: $p \leq 5 \times 10^{-6}$	712	0.0484	0.135	0.629 (0.612,0.645)	1.60 (1.50,1.70)	3.06 (2.04,4.59)
Fixed threshold: $p \leq 5 \times 10^{-7}$	464	0.0476	0.135	0.627 (0.611,0.644)	1.60 (1.50,1.70)	*3.38 (2.28,5.02)*
Fixed threshold: $p \leq 5 \times 10^{-8}$	334	0.0462	0.136	0.625 (0.609,0.641)	1.58 (1.49,1.69)	3.32 (2.24,4.93)
Fixed threshold: $p \leq 5 \times 10^{-9}$	264	0.0455	0.136	0.624 (0.608,0.64)	1.58 (1.48,1.68)	2.56 (1.68,3.90)
UKB Cohort						
Lassosum: s = 0.9, λ = 0.0043	*286,144*	*0.0487*	*0.0807*	*0.643 (0.637,0.65)*	*1.70 (1.65,1.75)*	*4.02 (3.46,4.67)*
P&T: $p \leq 1 \times 10^{-4}$	1,682	0.0401	0.0811	0.63 (0.623,0.637)	1.61 (1.57,1.66)	3.69 (3.16,4.31)
Fixed threshold: $p \leq 5 \times 10^{-6}$	712	0.0402	0.0811	0.628 (0.62,0.635)	1.61 (1.57,1.65)	3.32 (2.83,3.90)
Fixed threshold: $p \leq 5 \times 10^{-5}$	1,307	0.0392	0.0812	0.627 (0.62,0.634)	1.60 (1.56,1.64)	3.49 (2.98,4.08)
Fixed threshold: $p \leq 5 \times 10^{-7}$	464	0.0384	0.0812	0.626 (0.618,0.633)	1.59 (1.55,1.63)	3.81 (3.27,4.44)
Fixed threshold: $p \leq 5 \times 10^{-8}$	334	0.0361	0.0813	0.622 (0.615,0.63)	1.57 (1.53,1.61)	3.69 (3.16,4.31)
Fixed threshold: $p \leq 5 \times 10^{-9}$	264	0.0347	0.0813	0.62 (0.612,0.627)	1.55 (1.51,1.59)	3.28 (2.79,3.86)

Fritsche et al. *Am J Hum Genet* (2020) 41(6): 469 – 480



Part II:

Critically evaluating PRS studies

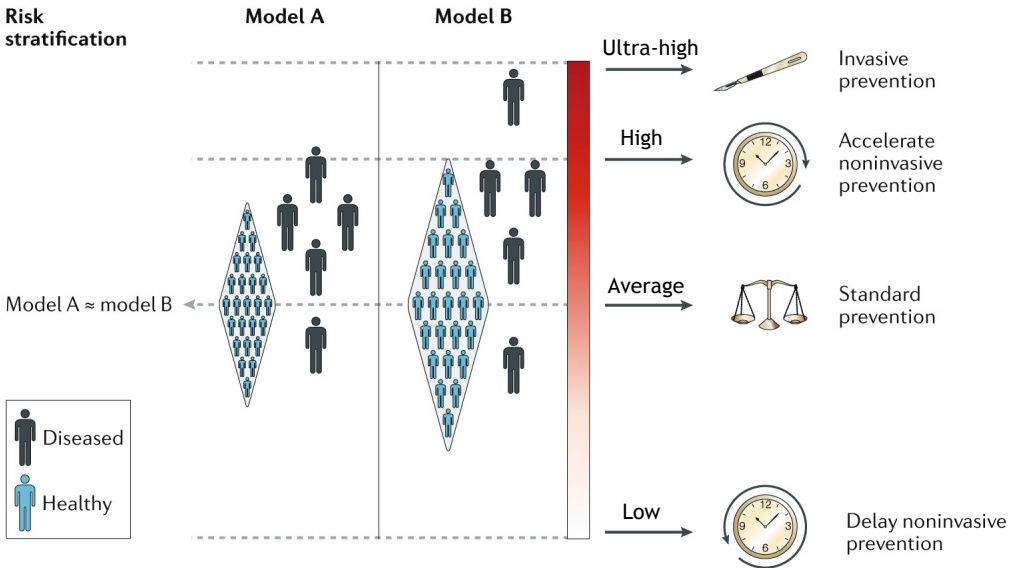
Presentation and discussion of Lambert et al. (2019)

PRS

Clinical Utility

Critically thinking about PRS utility

- Risk prediction: how well does the model distinguish between disease-free and affected individuals?
  - Classification problem
  - Equivalent AUC for Model A vs. B
- Risk stratification: how well does the model distinguish between gradients in disease risk?
  - Clinical decisions rely on specific absolute risk thresholds
  - Improved risk stratification in Model B



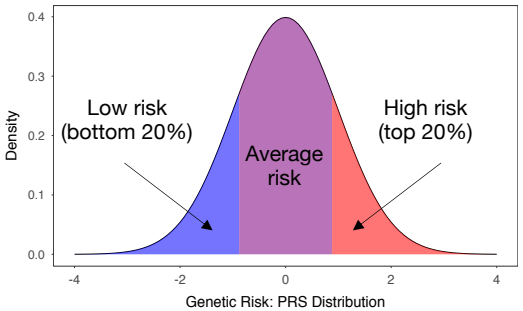
Torkamani et al. *Nature Reviews Genetics* (2018) 19: 581 – 590

4.

PRS

Clinical Utility

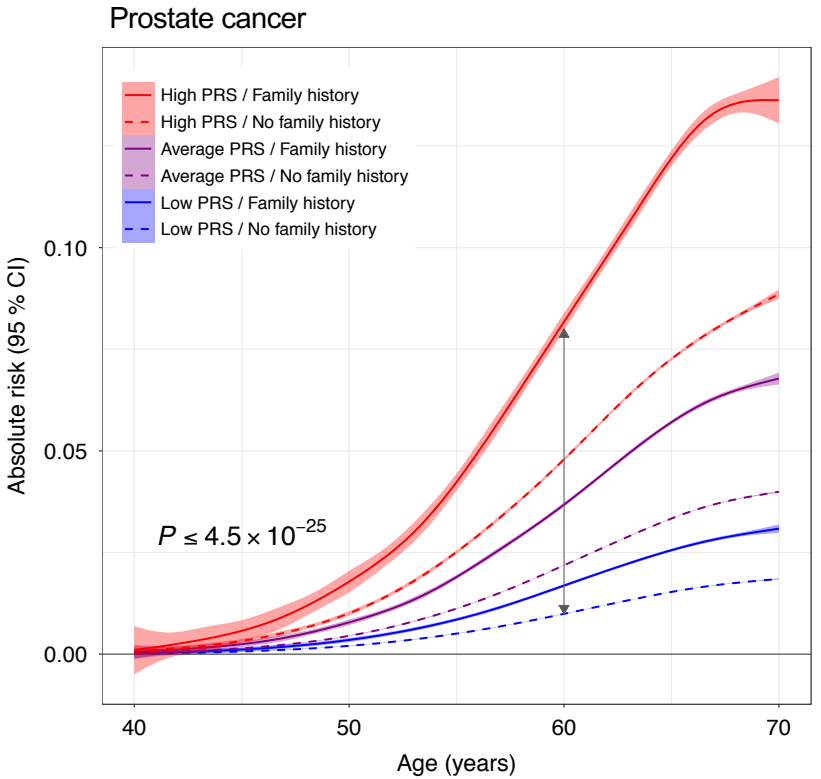
Critically thinking about PRS utility



+

Family history of prostate cancer (yes/no)

→



PRS significantly improves **risk stratification** compared to family history alone

Kachuri et al. *Nature Communications* (2020) 6084

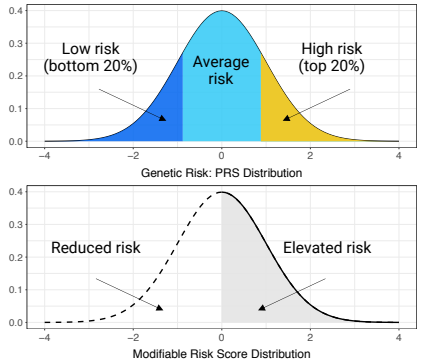
UCSF

4.

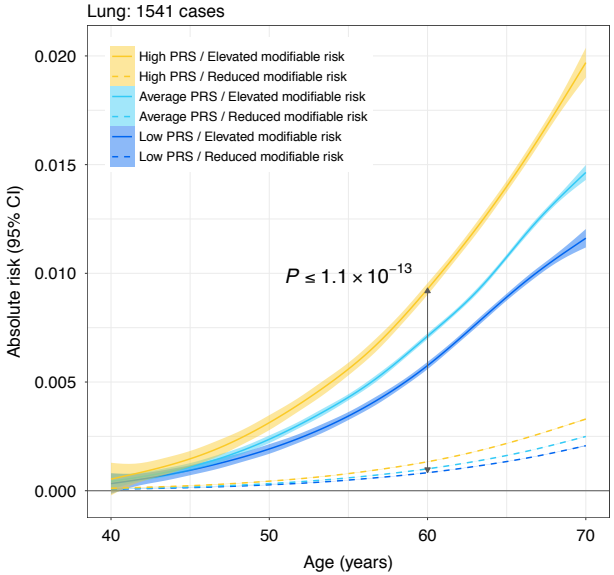
PRS

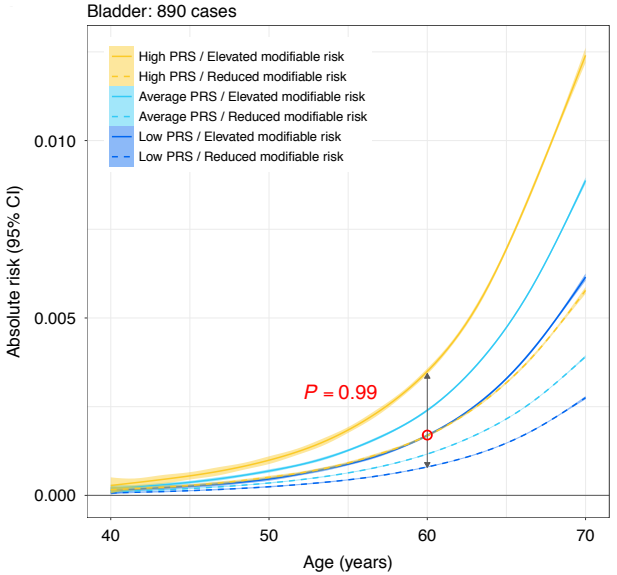
Clinical Utility

Critically thinking about PRS utility



Modifiable risk factors determine levels of risk, but PRS refines risk categories





Kachuri et al. *Nature Communications* (2020) 6084

UCSF

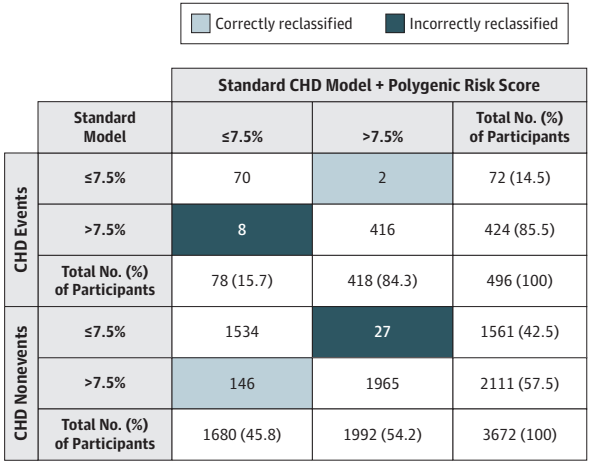
4.

PRS

Clinical Utility

Critically thinking about PRS utility

- **What does meaningful improvement look like?**
  - Net reclassification improvement (NRI) at clinically-actionable thresholds
  - What if no intervention exists?
- **Cost effectiveness: does the improvement in performance justify additional costs?**
  - Individual and system-level costs
  - Is the PRS intended to complement or replace existing tools?



NRI in CHD events:
(-8+2)/496 = -0.012

NRI in CHD non-events:
(-27+146)/3672 = 0.032

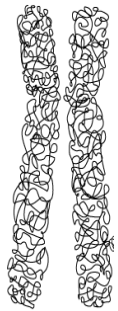
Overall NRI:
-0.012 + 0.032 = 0.020

Table 5. Reclassification Based on the Net Reclassification Improvement

		Net Reclassification Improvement (95% CI)[b]
Model 1[a]	Model 2[a]	ARIC
Age + sex	+ Pooled cohort risk group[3]	0.020 (−0.015 to 0.091)
Age + sex	+ Polygenic risk score	− 0.022 (− 0.036 to 0.021)
Age + sex + pooled cohort risk group[c]	+ Polygenic risk score	0.018 (− 0.012 to 0.036)

Moseley et al. *JAMA* (2020) 323, 7: 627 – 635

UCSF



Part III

From association to causation: Mendelian Randomization

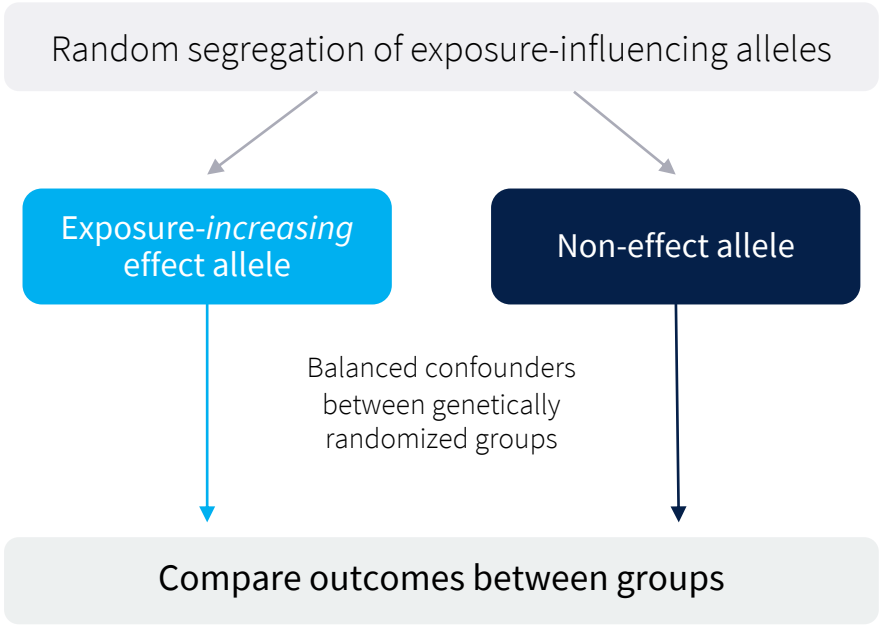
1.

Introduction

Motivation

Mendelian Randomization (MR)

- Approach for inferring causal relationships between phenotypes using genetic variants as instruments
- Overcomes certain limitations of observational studies:
  - **Reverse causation, confounding**
- But also presents new challenges:
  - **Genetic confounding, interpretation**

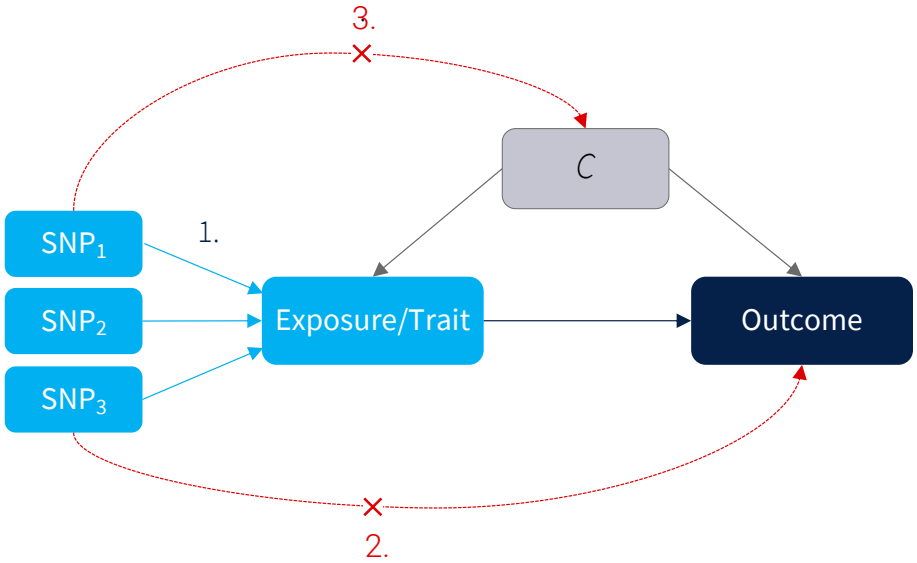


Introduction

Assumptions

Fundamental MR assumptions:

1. Instruments must be strongly associated with the exposure of interest
   - Is the exposure heritable?
2. Instruments should not affect the outcome via other pathways:
   - Violated by horizontal pleiotropy
3. Instruments must not be association with confounders
4. Consistency: instruments must have the same effect in all individuals



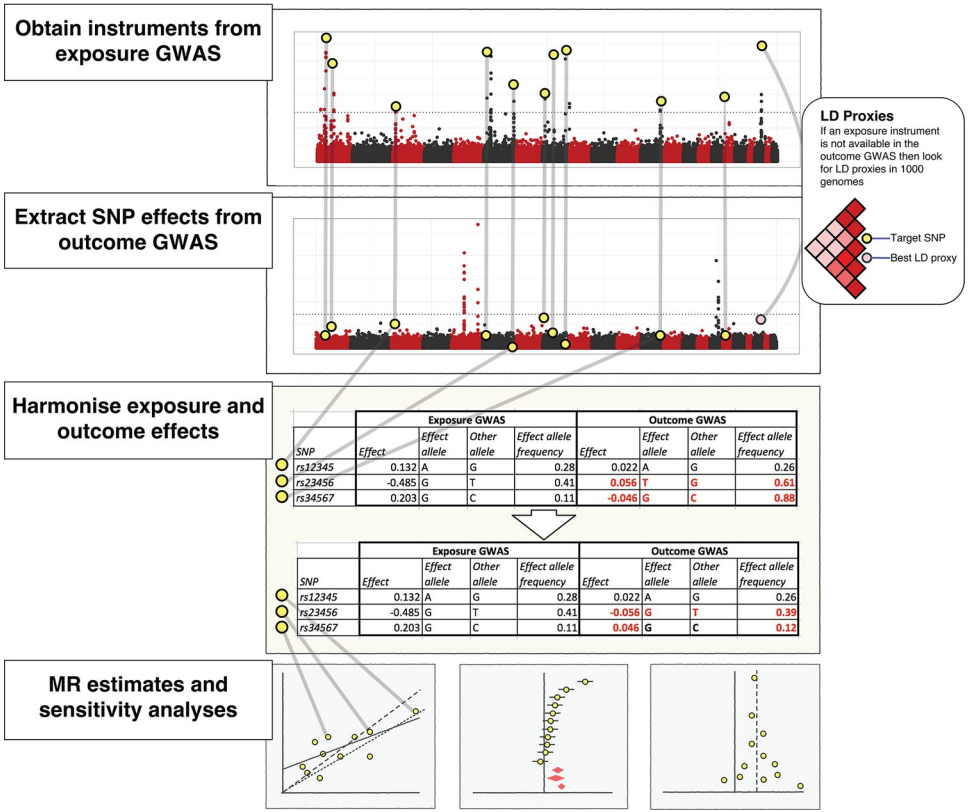
2.

Approach

Two-Sample MR

Two-sample analysis using summary statistics:

- Genetic associations with exposure (instruments) are obtained from one GWAS and applied to a separate outcome GWAS
- Individual-level data are not necessary
- Allows you to investigate associations with risk factors that are not available in your target study
- Analytic considerations similar to PRS



Hemani et al (2018) *eLife*7:e34408

UCSF

2.

Approach

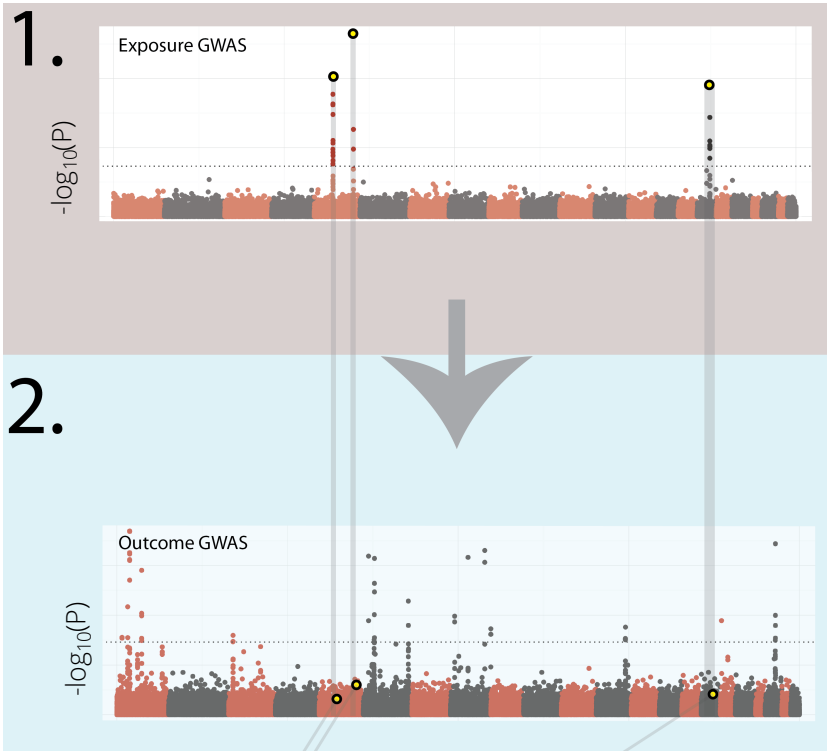
Two-Sample MR

Step 1: identify genetic instruments

- Select genetic variants with strong and consistent associations from the largest available GWAS **(most commonly: $P < 5\times10^{-8}$**, instrument strength: F>10)
- Most methods use independent variants, but some incorporate the SNP LD matrix

Step 2: apply genetic instruments to target dataset

- **No overlap between exposure and outcome GWAS**
- Partial overlap, especially for controls, is less problematic and can be corrected analytically



From: https://mrcieu.github.io/TwoSampleMR/articles/introduction.html

UCSF

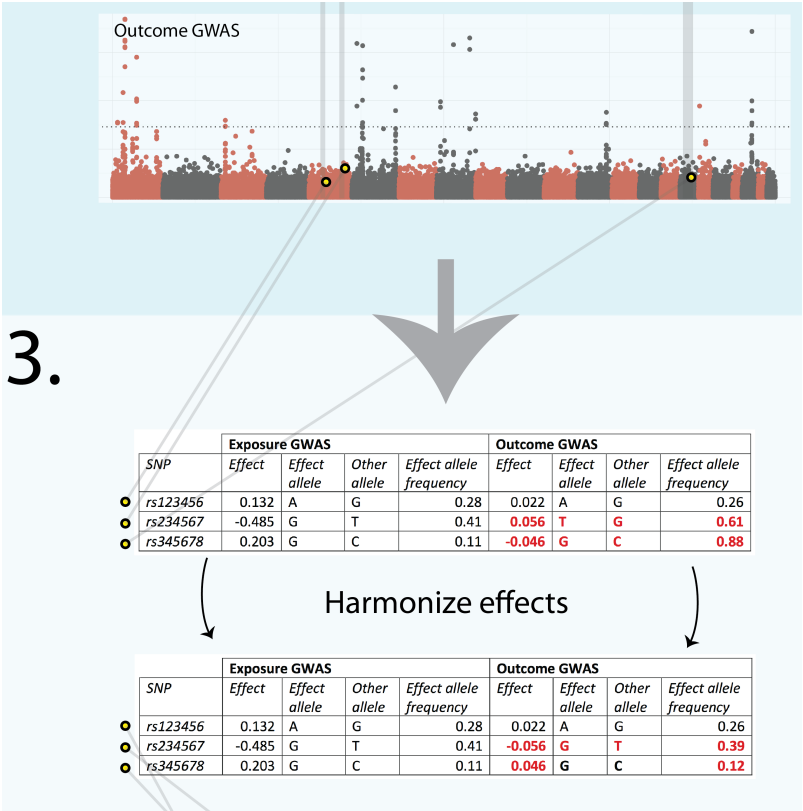
2.

Approach

Two-Sample MR

Step 3: harmonize the data appropriately

- Ensure that regression coefficients for each genetic variant in the exposure and outcome datasets correspond to the same allele
- Successful data harmonization requires well-annotated summary statistics
- Check for mismatched alleles, non-inferable strand ambiguous SNPs, imputation quality information (INFO scores)



	Exposure GWAS				Outcome GWAS			
SNP	*Effect*	*Effect allele*	*Other allele*	*Effect allele frequency*	*Effect*	*Effect allele*	*Other allele*	*Effect allele frequency*
rs123456	0.132	A	G	0.28	0.022	A	G	0.26
rs234567	-0.485	G	T	0.41	**0.056**	**T**	**G**	**0.61**
rs345678	0.203	G	C	0.11	**-0.046**	**G**	**C**	**0.88**

	Exposure GWAS				Outcome GWAS			
SNP	*Effect*	*Effect allele*	*Other allele*	*Effect allele frequency*	*Effect*	*Effect allele*	*Other allele*	*Effect allele frequency*
rs123456	0.132	A	G	0.28	0.022	A	G	0.26
rs234567	-0.485	G	T	0.41	**-0.056**	**G**	**T**	**0.39**
rs345678	0.203	G	C	0.11	**0.046**	**G**	**C**	**0.12**

From: https://mrcieu.github.io/TwoSampleMR/articles/introduction.html

UCSF

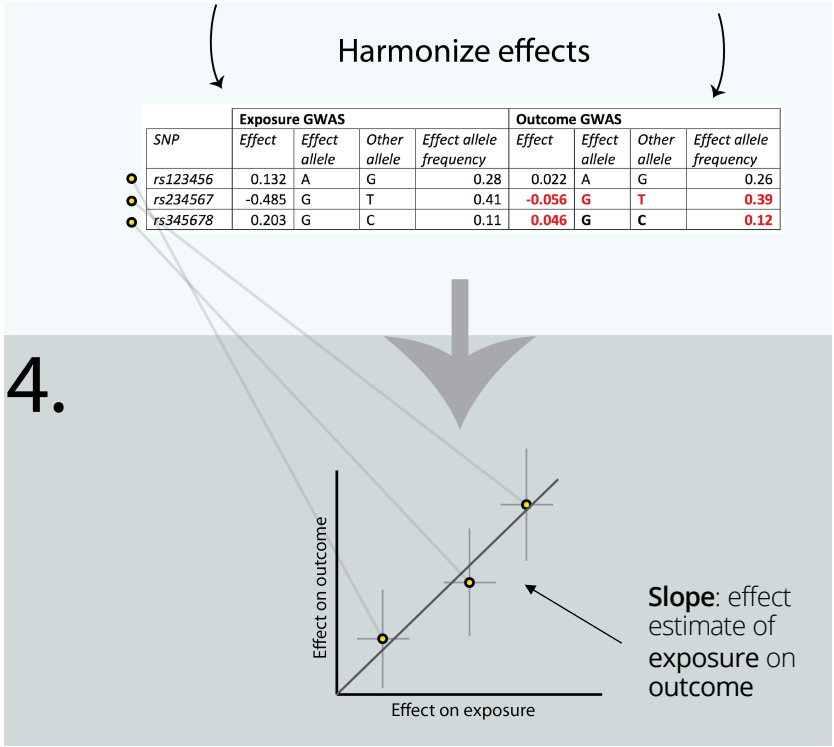
2.

Approach

Two-Sample MR

Step 4: Perform MR analysis

- Estimate genetically predicted exposure effects on the outcome
- Conduct extensive sensitivity analyses and examine multiple MR diagnostics to ensure that the results are robust and unbiased



Adapted from: https://mrcieu.github.io/TwoSampleMR/articles/introduction.html

UCSF

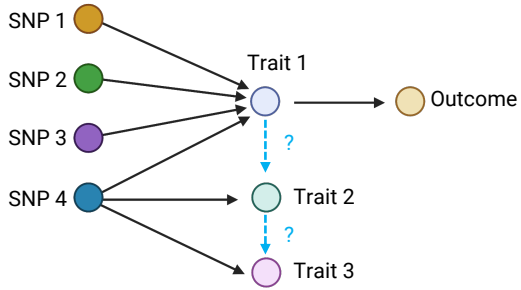
2.

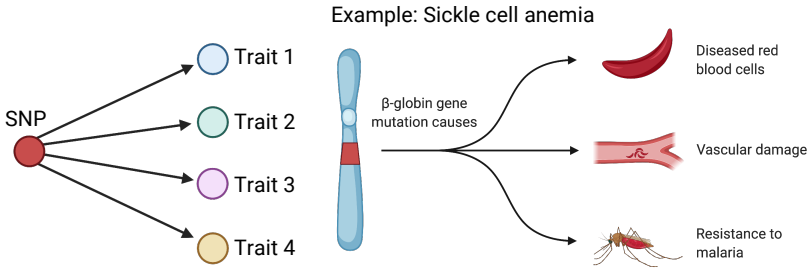
Approach

Threats to Validity

A closer look at MR assumptions:

- Distinction between vertical and horizontal pleiotropy
  - MR relies on vertical pleiotropy : SNPs associated with multiple phenotypes in the same pathway
  - MR results are confounded by horizontal pleiotropy: SNPs associated with multiple phenotypes in distinct biological pathways
- Underlying biological pathways are difficult to establish, so need to rely on statistical methods
- MR assumptions cannot be conclusively verified, so causal inference is not “guaranteed”





2.

Approach

Threats to Validity

MR Egger: test and correct for directional horizontal pleiotropy

- Parameters estimated:
  - Intercept (β_0): average pleiotropic effect across all variants
  - If $\beta_0 \neq 0$ at p<0.05, report the corrected MR Egger effect estimate (β_{1E})
- Magnitude of pleiotropic effects is independent of instrument strength
  - InSIDE (Instrument Strength Independent of Direct Effect) assumption
- Negligible uncertainty in SNP-exposure estimates
  - NOME (No Measurement Error) assumption
  - Violations will inflate intercept test and bias causal effect towards 0

Bowden et al. *Stat Medicine.* (2017) 36: 1783 – 1802

2.

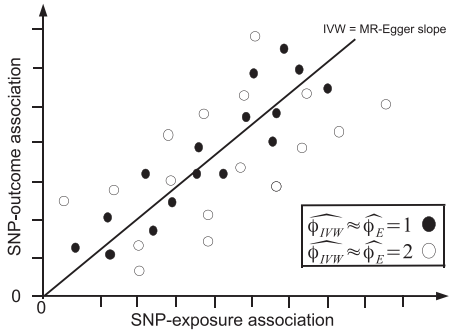
Approach

Threats to Validity

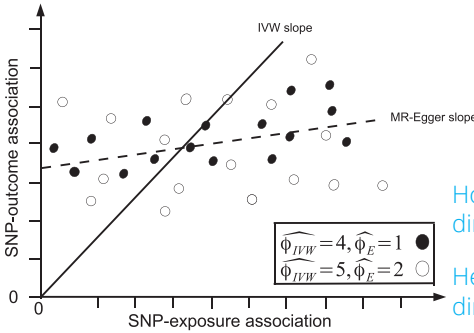
Heterogeneity (Q statistic): indicative of balanced horizontal pleiotropy

- Modified version of standard meta-analysis Cochran's Q
- **Quantifies heterogeneity across SNP-specific IVW effect estimates**
- Consider filtering out instruments with significant Q p-values (at an appropriate threshold)

$$Q = \sum_{j=1}^{L} \left(\frac{\hat{\Gamma}_j}{\sigma_{Yj}} - \frac{\hat{\beta}_{IVW}\gamma_j}{\sigma_{Yj}} \right)^2 = \sum_{j=1}^{L} w_j(\hat{\beta}_j - \hat{\beta}_{IVW})^2$$



No pleiotropy
Balanced pleiotropy



Homogeneous directional pleiotropy

Heterogenous directional pleiotropy

Bowden et al. *Stat Medicine.* (2017) 36: 1783 – 1802

UCSF

Approach

Threats to Validity

Other threats to valid inference:

- Weak instrument bias /regression dilution bias
  - Low proportion of variation explained, may be due to noise, measurement error
  - Assumption that confounders are balanced between genetically “randomized” groups can break down
- Selection bias in the exposure or outcome GWAS
- Non-random or assortative mating
- Winner’s curse in the exposure GWAS will bias MR estimates towards the null
- Inappropriate covariate adjustment
  - MR of BMI where SNP effects on BMI are adjusted for waist to hip ratio

2.

Analysis

Example

Workflow Overview

Initial set of instruments: $P_{DISC}<5\times10^{-8}$ and $P_{REP}<0.05$

- Exclude palindromic SNPs (A/T or G/C) with intermediate allele frequencies
- Exclude SNPs with MAF<0.01
- Exclude SNPs with EAF discrepancy>0.10

Estimate causal effects using the initial set of instruments

- Check main MR diagnostics
  - Directional pleiotropy: Egger $\beta_0 \neq 0$
  - Balanced pleiotropy (heterogeneity): Cochran's Q p<0.05
  - Weak instrument bias: $I_{GX}^2<1$

- Identify potentially invalid instruments (outliers) contributing to heterogeneity: Cochran's Q (p<0.05)

Sensitivity analyses:

- Consistency across multiple MR estimators
- Consistent effects in leave-one-out (LOO) analyses

UCSF

3.

Analysis

MR Estimators

Using multiple estimation approaches helps assess the robustness of the observed results

1. Mean-based methods: **average causal effect across all instruments**
   - Assume that all instruments are valid: 0% breakdown level
   - **Inverse variance weighted (IVW) and maximum-likelihood (ML)**
   - **Multiplicative random effects IVW:** accounts for heterogeneity across causal effects
     - Assumes that pleiotropy is balanced, without affecting the slope
2. Consensus methods: **causal effect is the value taken by the largest number of instruments**
   - "Plurality valid" assumption
3. Outlier robust methods: **identify potentially invalid instruments and correct for their influence on the causal effect estimate**

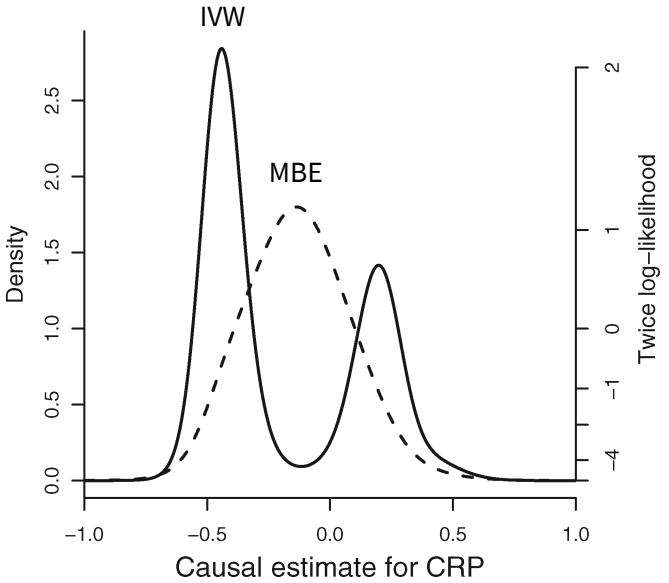
Pierce & Burgess *Am J Epidemiol*. 2013 178(7): 1177–1184
Burgess et al. *Genet Epidemiol*. 2013 37(7): 658–665

Analysis

MR Estimators

2. Consensus methods:

- Weighted median estimator (WME)
  - Median of the weighted empirical distribution function of SNP-specific causal effect estimates
  - Unbiased when up to 50% of the weights come from invalid instruments
- Mode-based estimator (MBE)
  - Causal effect is based on the mode of the smoothed empirical density of individual SNP-specific causal effect estimates



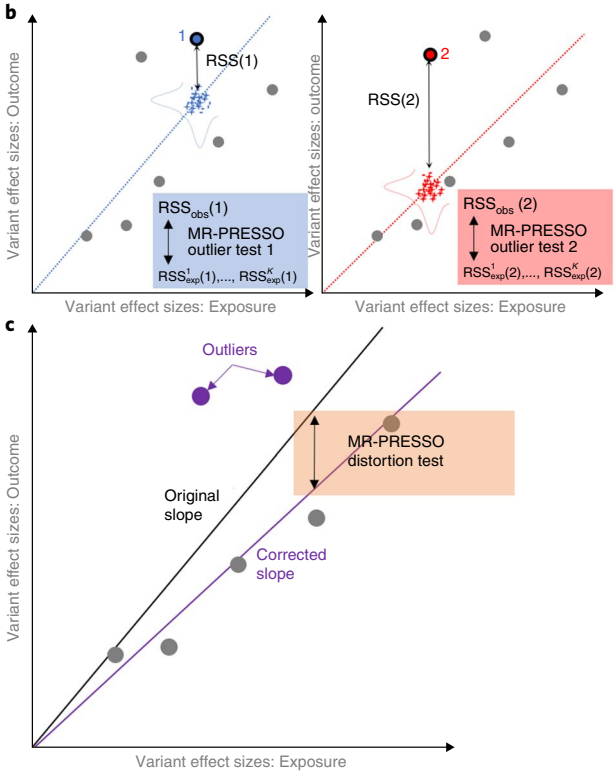
Bowden et al. *Genet Epidemiol*. 2016 May; 40(4): 304–314
Burgess et al. *Int J Epidemiol*, 2018; 47(4): 1242–1254

Analysis

MR Estimators

3. Outlier robust methods

- Outlier detection based on heterogeneity
- MR pleiotropy residual sum and outlier (MR PRESSO)
  - Estimates the distortion caused by outliers and corrects for it analytically
- MR Robust Adjusted Profile Score (MR RAPS)
  - Influence of outliers is controlled by loss functions
  - Empirical Bayes shrinkage adaptively down-weights weak instruments



Verbanck et al. *Nature Genetics* 2018 50(5): 693-698
Zhao et al. arXiv: (2018) 1801.09652

4.

MR Example

Lung Cancer

Relationship between impaired pulmonary function and lung cancer risk is well established, but is it causal?

- Both disease phenotypes are caused by cigarette smoking
  - **Other etiologic pathways also plausible (immune- and inflammation-mediated)**
- Shared genetic susceptibility loci between lung cancer and COPD
- COPD is a risk factor in never smokers

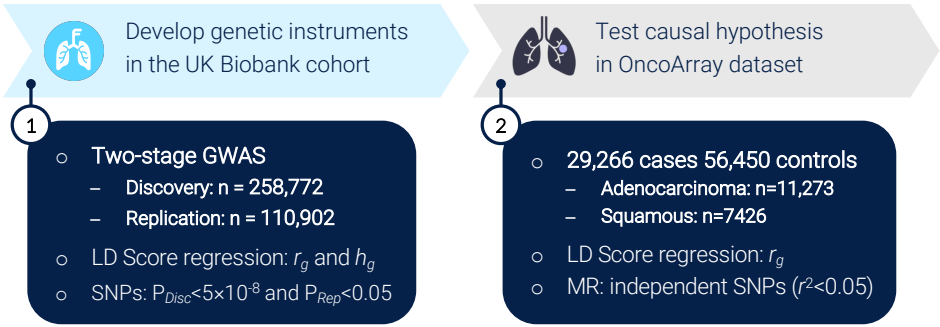
nature COMMUNICATIONS

ARTICLE

https://doi.org/10.1038/s41467-019-13855-2 OPEN

Immune-mediated genetic pathways resulting in pulmonary function impairment increase lung cancer susceptibility

Linda Kachuri et al.#

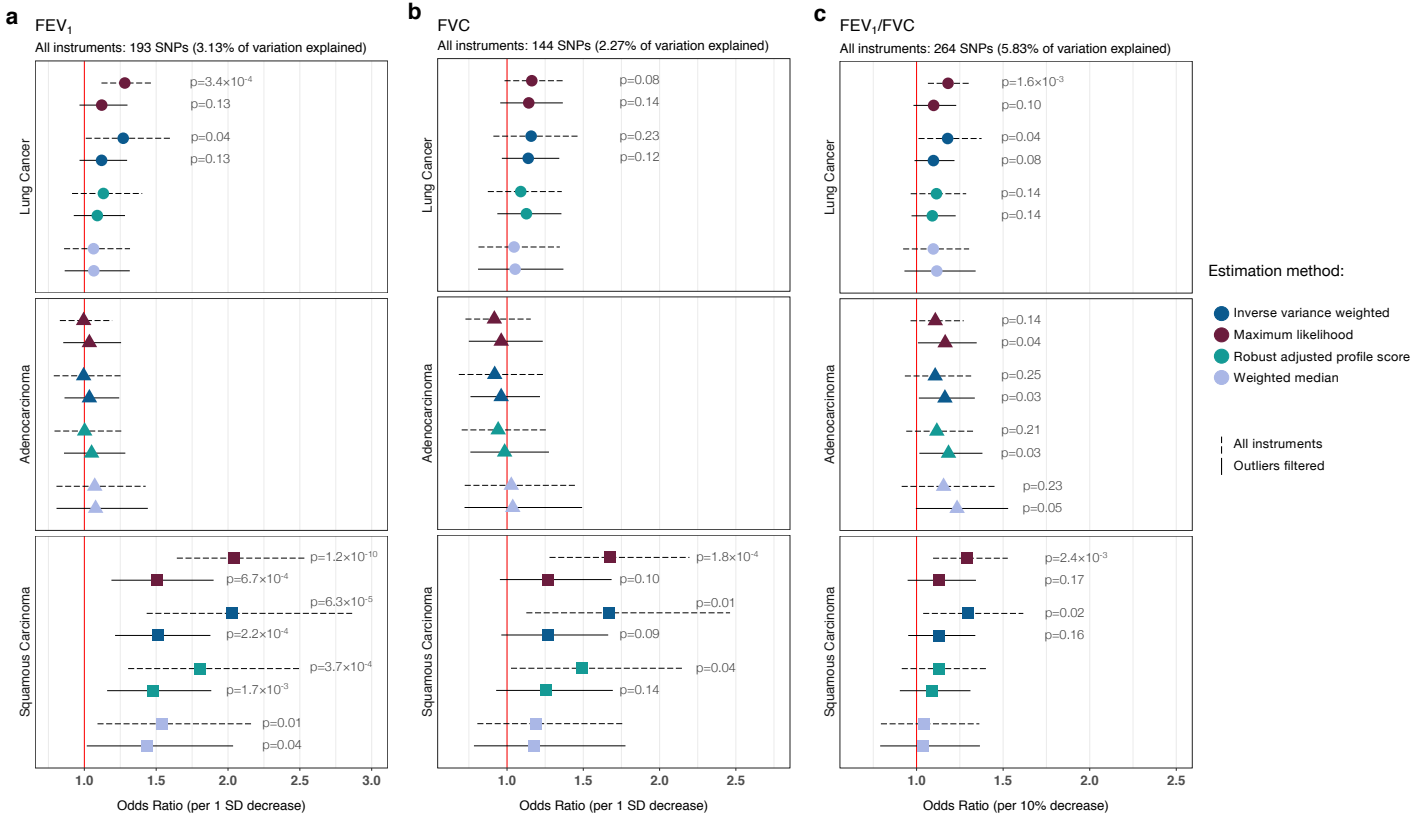


UCSF

MR Example

Lung Cancer

- Reduced FEV_1 increases risk of squamous lung carcinoma
- Evidence for FVC is inconsistent
- Reduced FEV_1/FVC increases risk of lung adenocarcinoma



Summary

MR and PRS

Similarities	Differences
○ **Inference: identify cross-trait associations** ○ Two-sample approach to avoid overfitting and ensure valid inference ○ Thresholds for selecting variants/instruments ○ Processing of summary statistics ○ General data QC considerations ○ Challenges with analyses in ancestrally diverse populations (more on that next week!)	○ **Inference: causation (MR) vs. association (PRS)** ○ MR is less powerful than PRS and have stricter assumptions to support causal inference ○ **MR can inform intervention targets** ○ **PRS can be used for risk prediction and risk stratification** – Provides individual-level risk profiles – MR does not use individual-level data

Summary

Resources

Resources for PRS

- R package for C+T and LD-Pred2: https://github.com/privefl/bigsnpr
- GCTA for BLUP: https://cnsgenomics.com/software/gcta/#BLUP
- PRS tutorial with examples for multiple methods: https://choishingwan.github.io/PRS-Tutorial/

Resources for MR

- R packages with examples: https://mrcieu.github.io/TwoSampleMR and https://cran.r-project.org/web/packages/MendelianRandomization
- Web-based app: http://www.mrbase.org

Other useful links:

- Processed GWAS data: https://gwas.mrcieu.ac.uk
- Look at LD and find proxies: https://ldlink.nci.nih.gov also available as R package (LDlinkR)
- PRS power and sample size estimation: https://github.com/DudbridgeLab/avengeme/

Thank you

Questions?

Linda.Kachuri@ucsf.edu @Linda_Kachuri

wittelab.ucsf.edu

Extra Slides

MR Diagnostics

I^2_{GX} : Quantifies collective strength across L instruments and estimates the degree of regression dilution bias due to exposure measurement error

$$I^2_{GX}{}' = (Q_{GX} - (L-1))/Q_{GX}$$

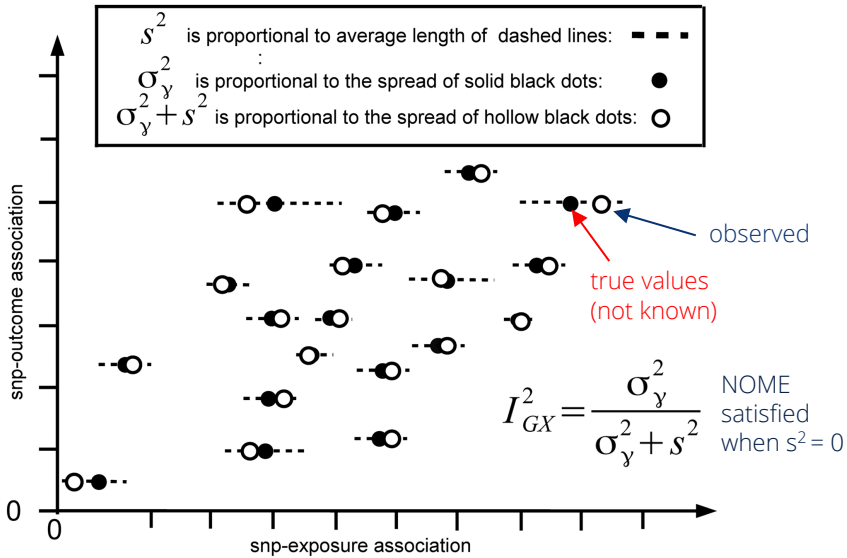
- When I^2_{GX} ~ 1 attenuation due to NOME violation is negligible
- When I^2_{GX} ~ 0.90 the bias in $\hat{\beta}_{1E}$ due to measurement error is ~10%
- Can be used to correct MR Egger estimates by the degree of NOME violation

SNP-exposure association

Mean SNP-exposure association

$$Q_{GX} = \frac{\sum_{j=1}^{L} (\hat{\gamma}_j - \bar{\hat{\gamma}})^2}{\sigma^2_{Xj}}$$

Variance of SNP-exposure estimates



Bowden et al. *Int J Epidemiol.* 2016 45(6):1961-1974

UCSF