Polygenic risk scores

Are they ready for clinical implementation?

Rainer Malik

Molecular Diagnostics Symposium

Zürich

March 2023

LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

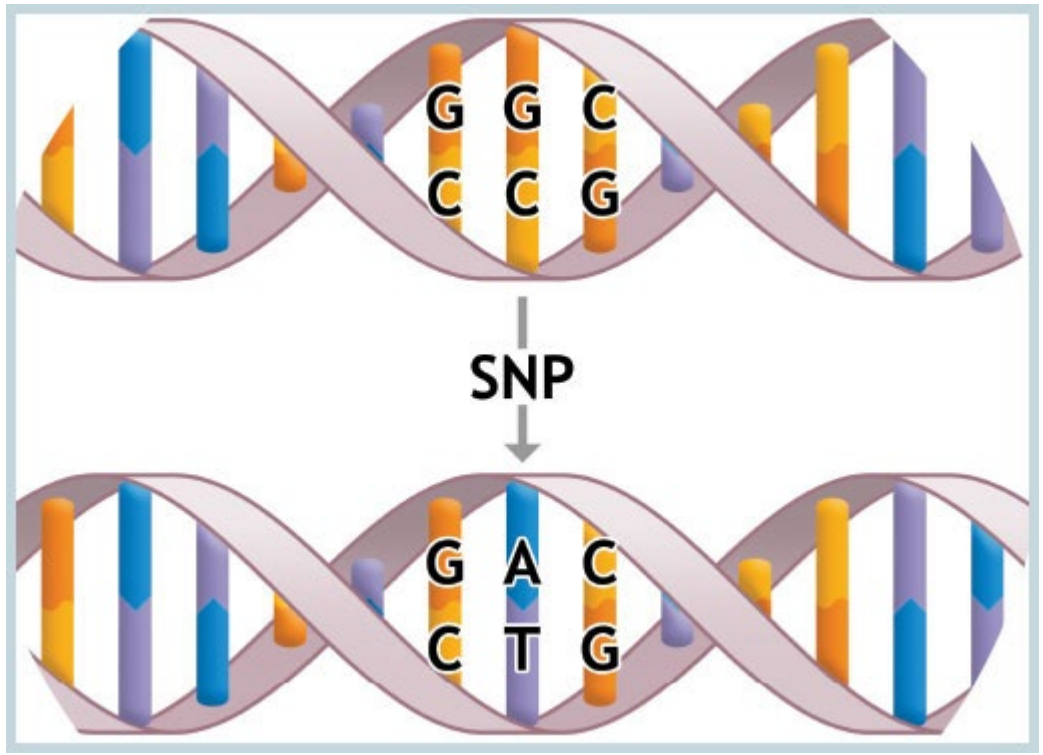
LMU KLINIKUM

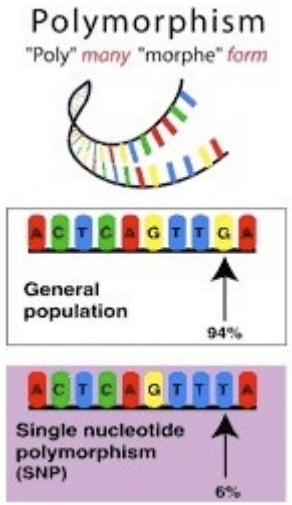
Institute for Stroke and Dementia Research (ISD)

Funded by: Solorz-Żak Research Foundation

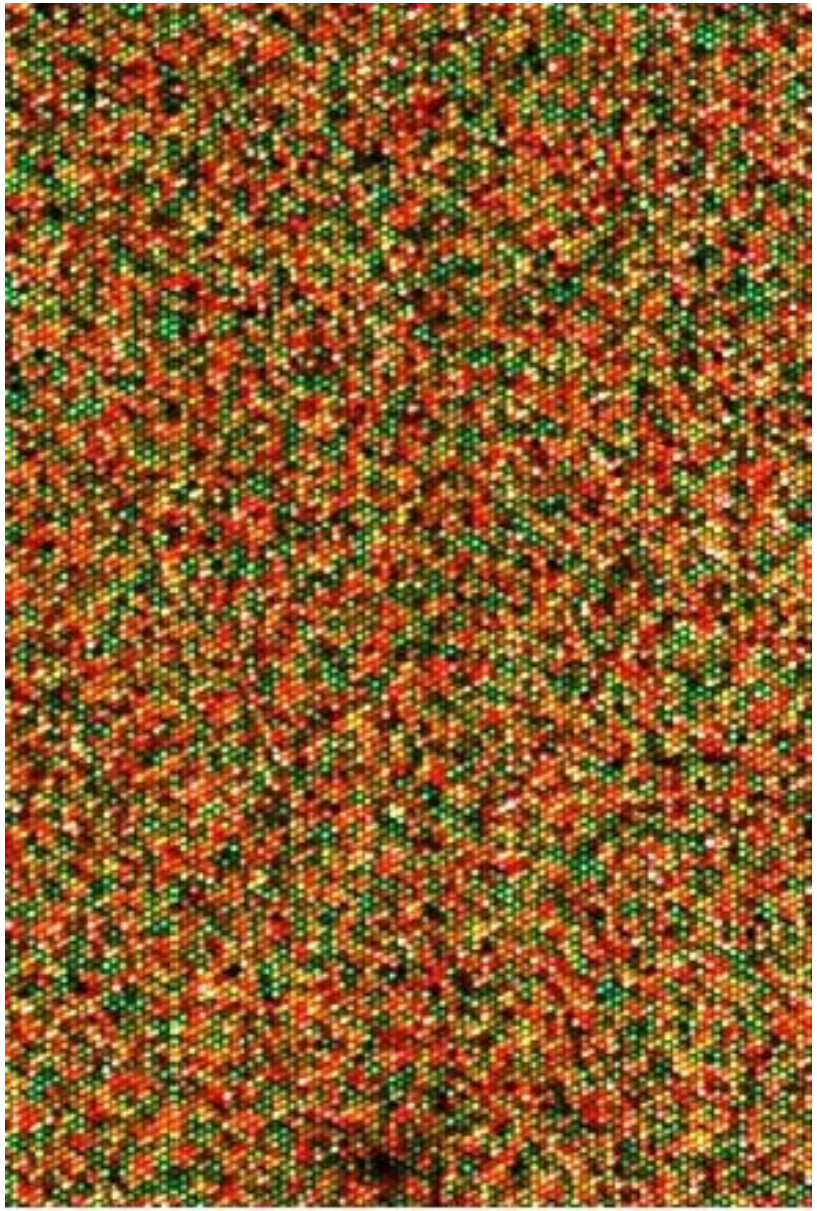
- Definitions
- Polygenic Risk Scores
  - Definition
  - History
  - Practical Examples
- Transferrability
- Translational perspective
- Summary
  - Pros
  - Cons

Definitions (Genotyping - SNP)



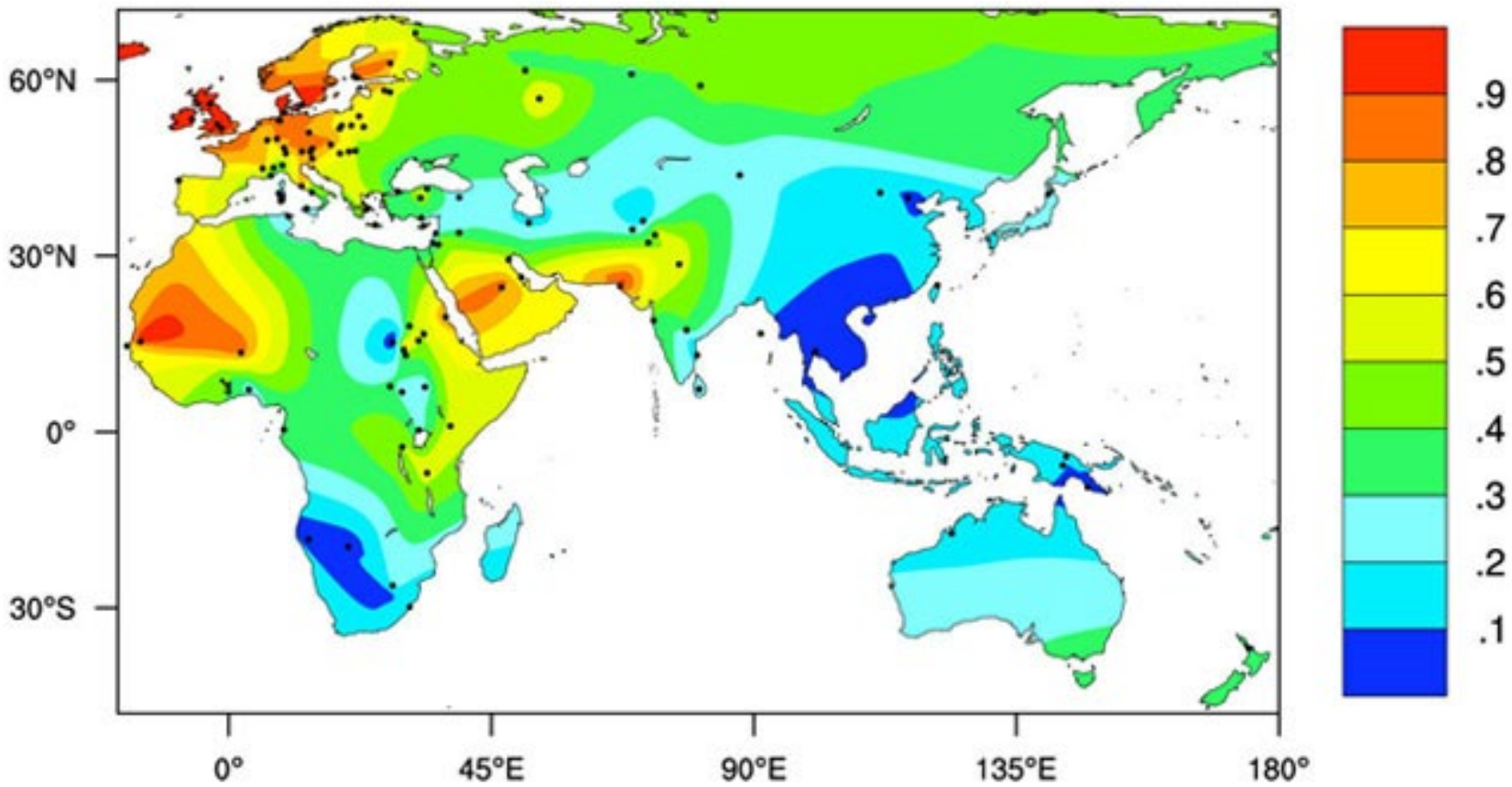


~40M – 50M SNPs in human genome



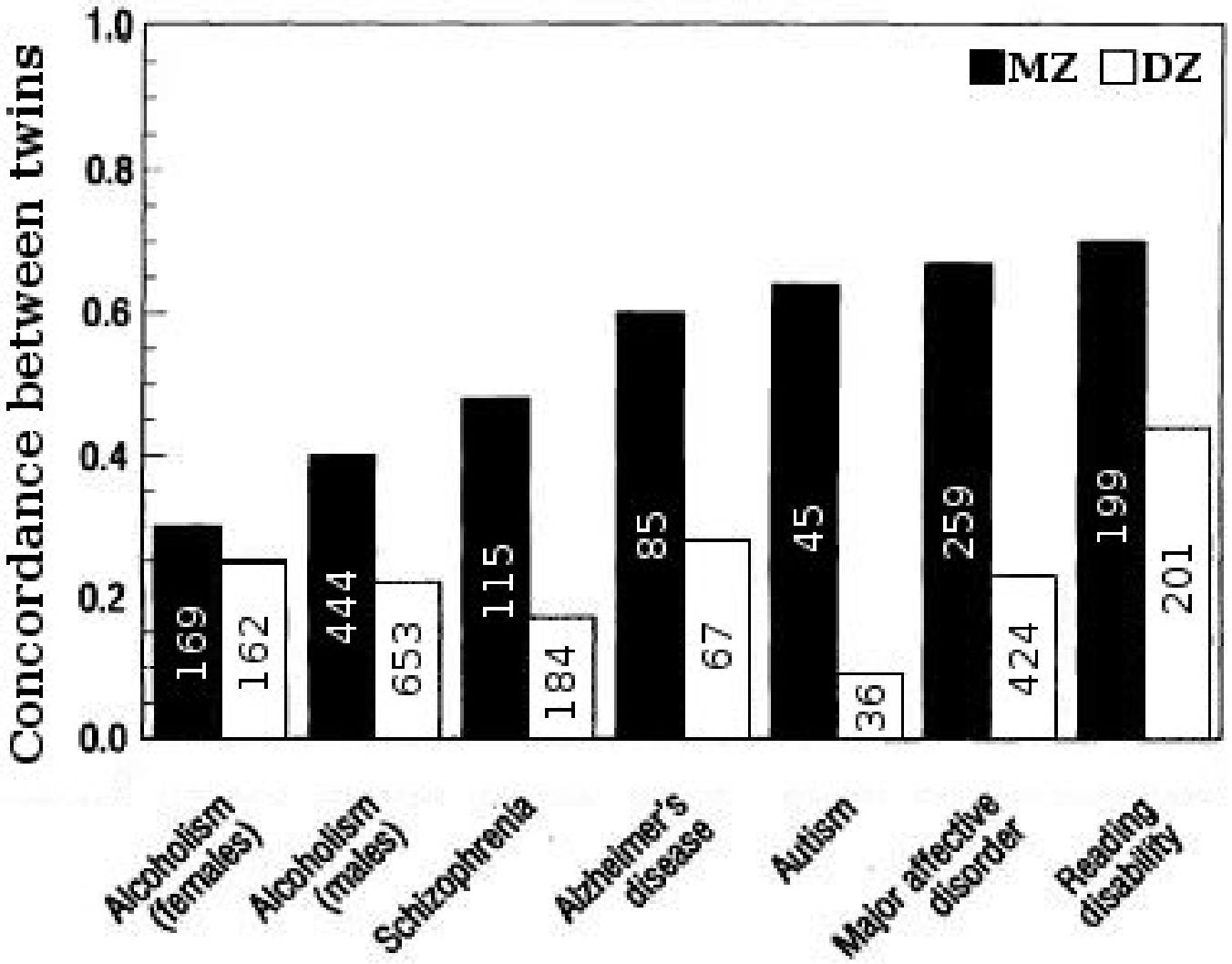
Genotyping chip with raw data -- red or green indicates identical copies of DNA inherited from both parents and yellow indicates copies from each parent are different.

Regional distribution - Lactose intolerance SNPs

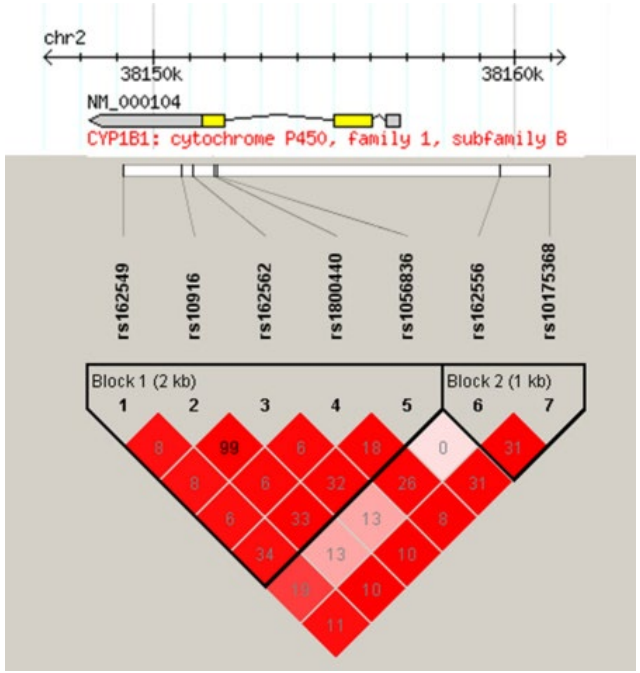


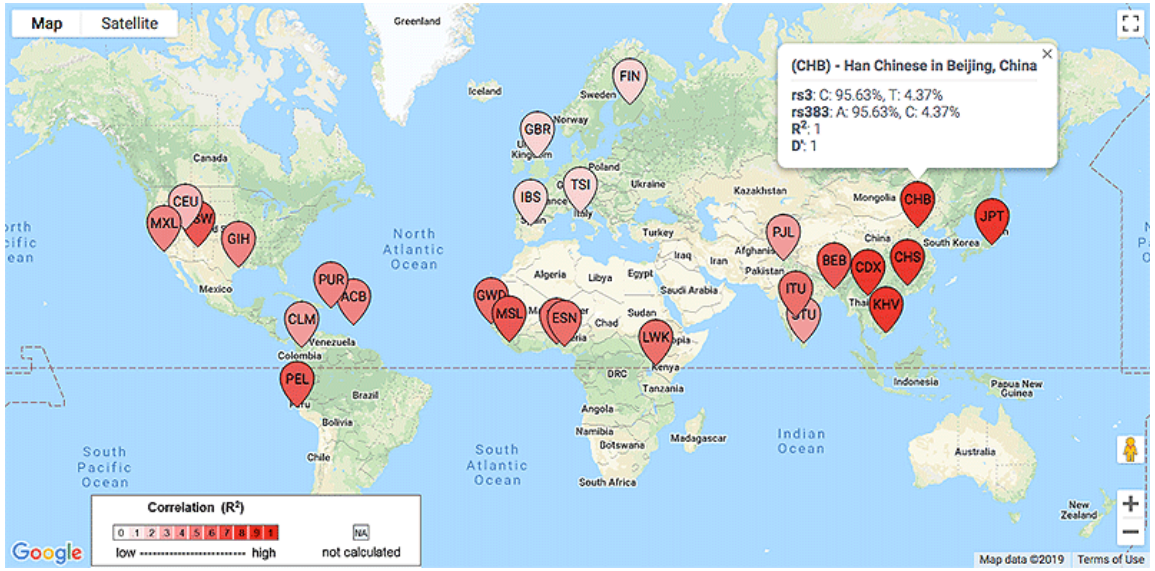
Itan et al. BMC Evolutionary Biology 2010, 10:36

Heritability



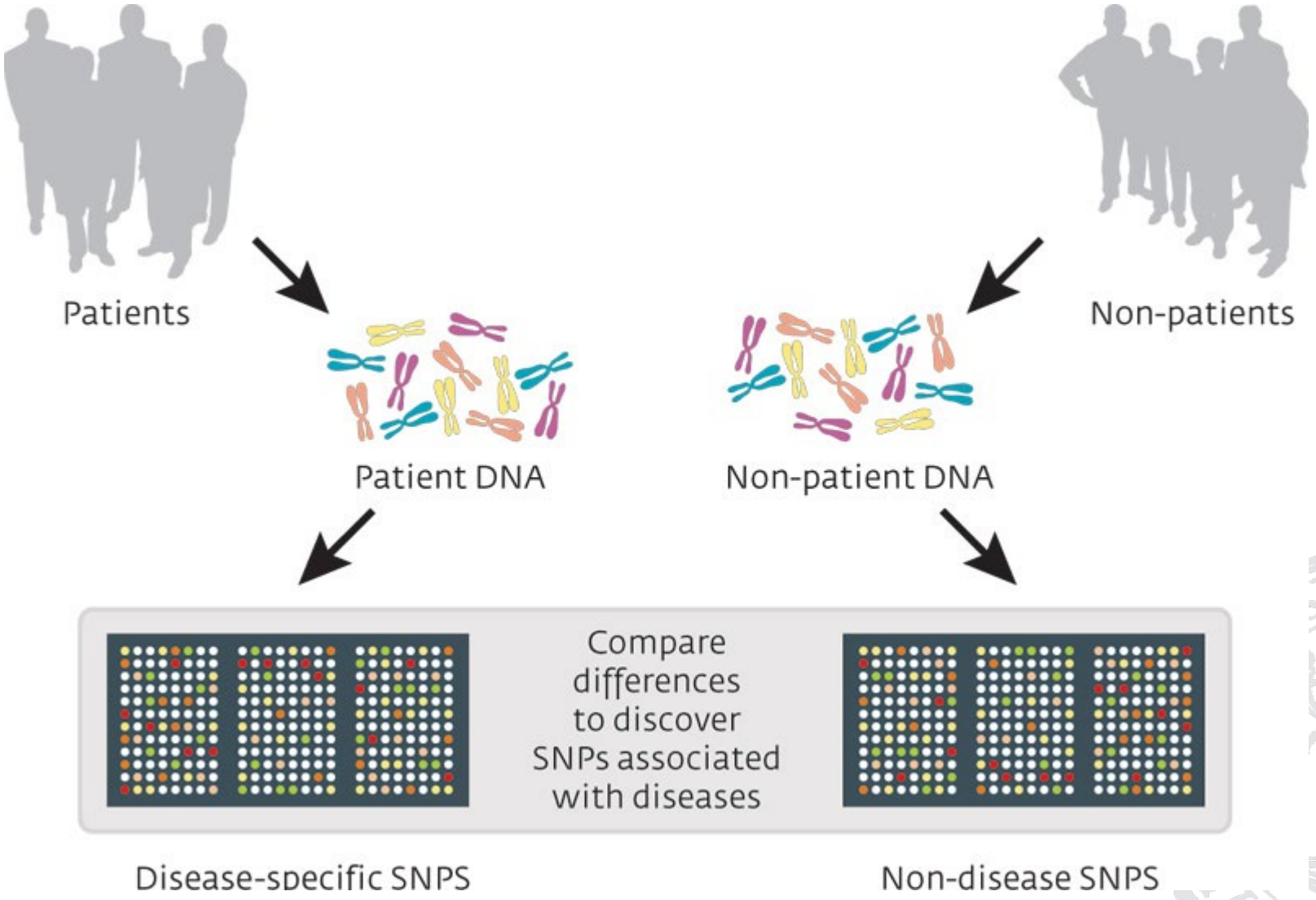
Linkage disequilibrium (LD)

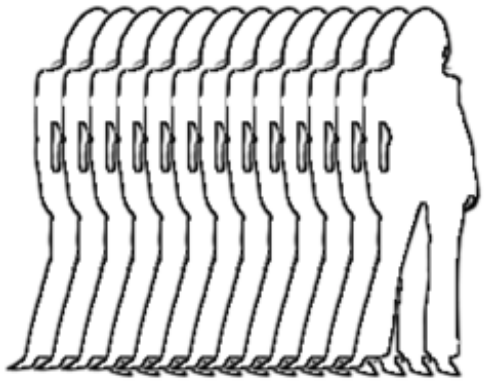




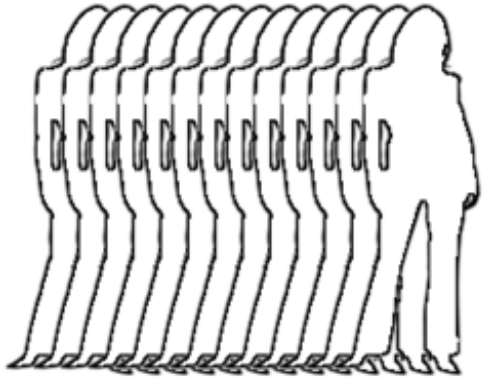
Institute for Stroke and Dementia Research

Genome-wide association studies





GC CC GG GC CC GC GC
GG CC GC GG GC GG

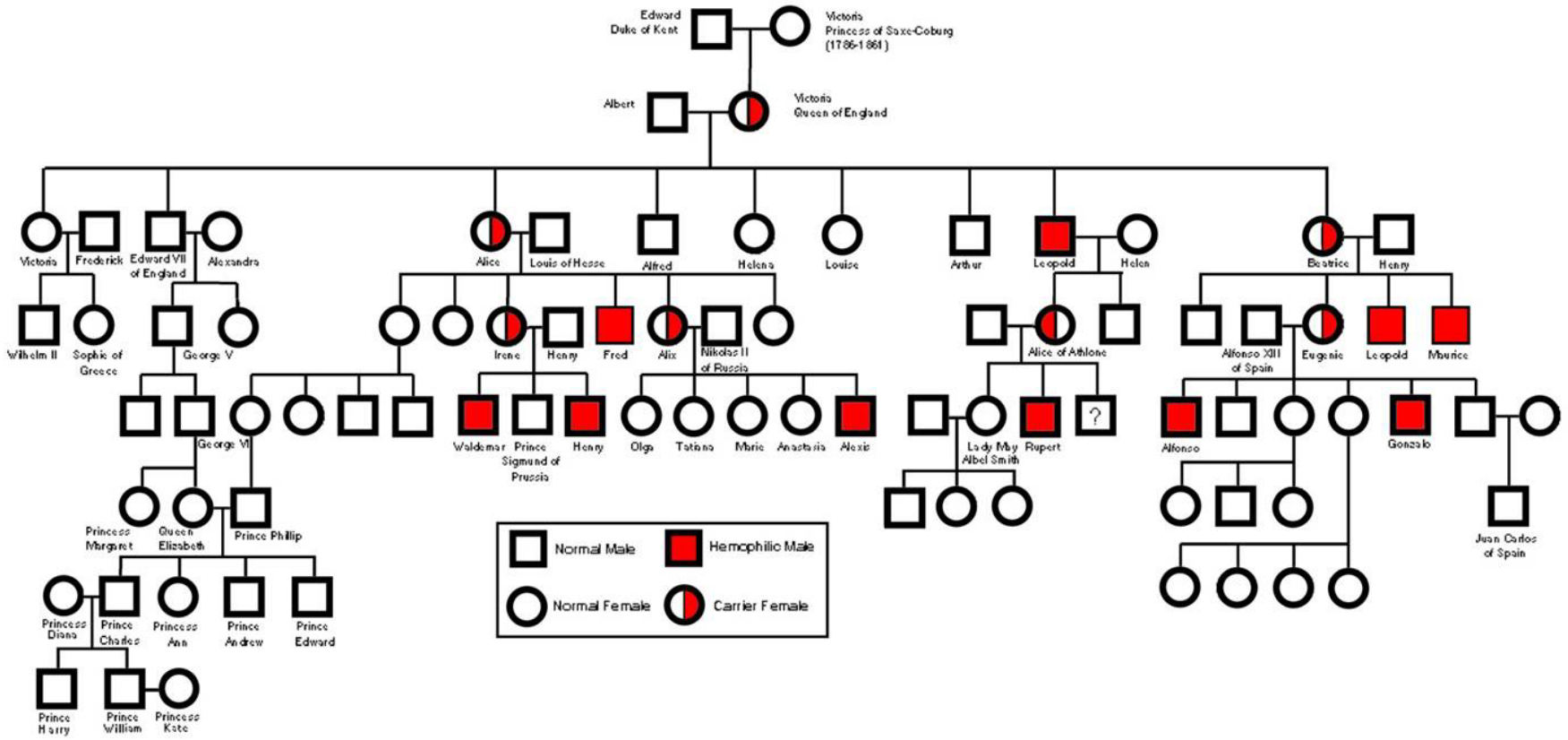


GC CC GC GC GG CC CC
CC GC GC GG GC GG

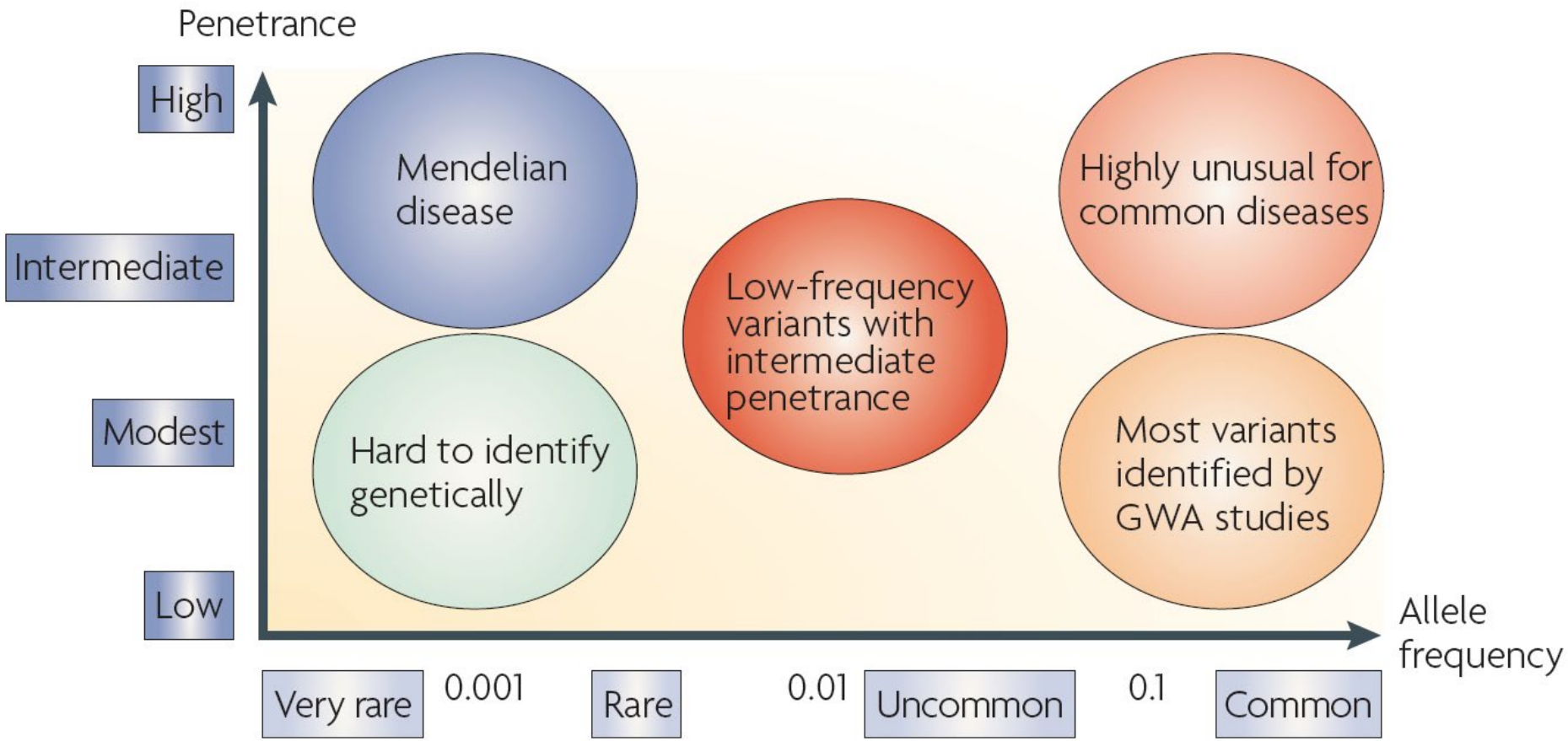
SNP1	*SNP2*	*SNP…*
Cases Count of G: 2104 of 4000 Frequency of G: 52.6%	**Cases** Count of G: 1648 of 4000 Frequency of G: 41.2%	*Repeat for all SNPs*
Controls Count of G: 2676 of 6000 Frequency of G: 44.6%	**Controls** Count of G: 2532 of 6000 Frequency of G: 42.2%	
P-value: $5.0 \cdot 10^{-15}$	**P-value:** 0.33	

OR= 1.15
Beta= ln(OR)=0.137

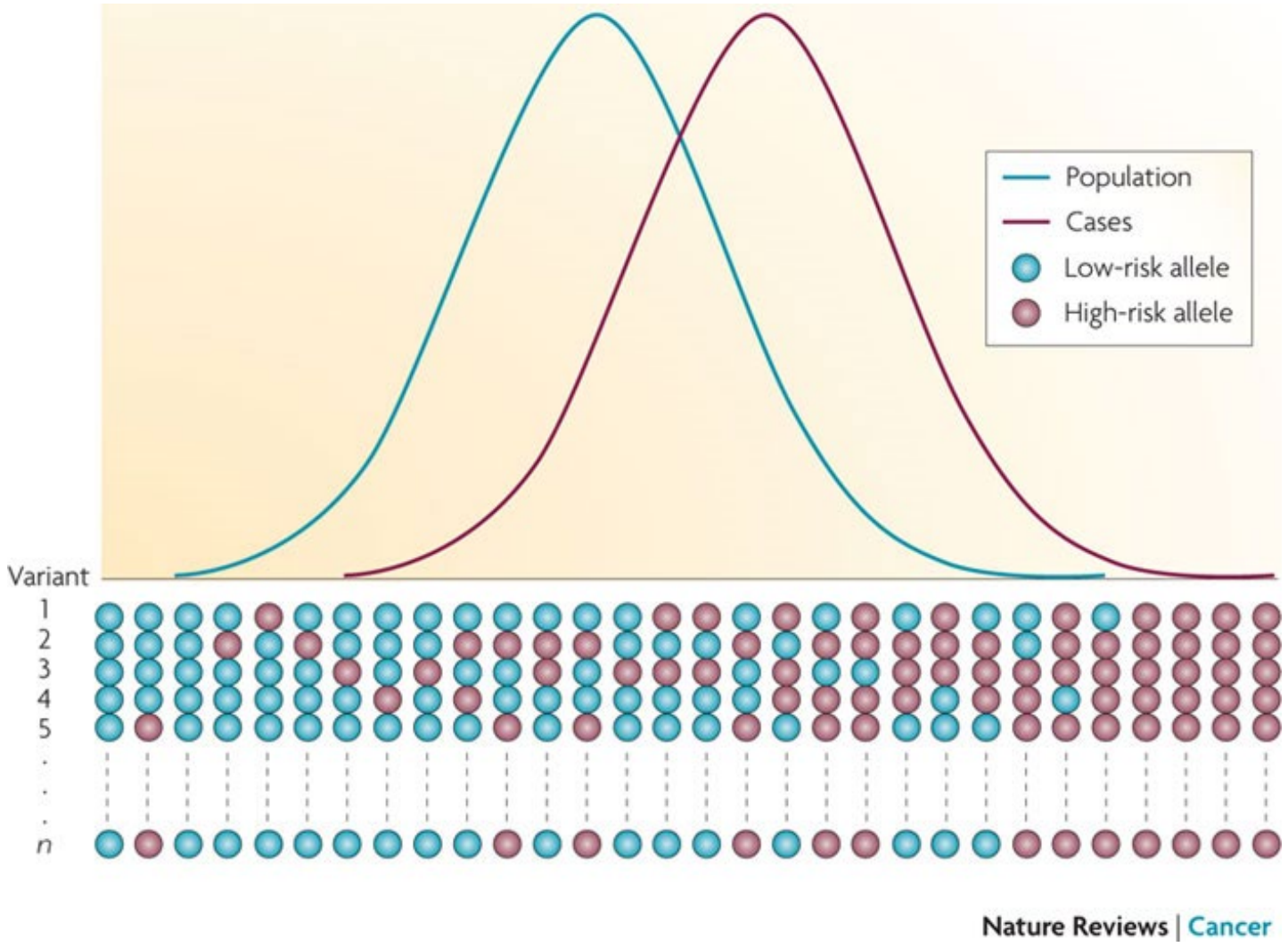
Mendelian disorder



rs398122990


Penetrance
High
Intermediate
Modest
Low
Mendelian disease
Hard to identify genetically
Low-frequency variants with intermediate penetrance
Highly unusual for common diseases
Most variants identified by GWA studies
Allele frequency
Very rare
0.001
Rare
0.01
Uncommon
0.1
Common

Polygenic


Population
Cases
Low-risk allele
High-risk allele
Variant
1
2
3
4
5
.
.
.
n
Nature Reviews | Cancer

Addendum: PRS, GRS etc

Box 1.

Definitions of relevant genetic risk prediction terms

Polygenic Score(s) (PGS):

a single value that quantifies an individual's genetic predisposition to a trait. Typically calculated by summing the number of trait-associated alleles in an individual weighted by per-allele effect sizes from a discovery GWAS, and normalized using a relevant population distribution. Sometimes referred to as a genetic score.

Polygenic Risk Score(s) (PRS):

a subset of PGS which is used to estimate risk of disease or other clinically relevant outcomes (binary or discrete). Sometimes referred to as a genetic or genomic risk score (GRS). See categories below.

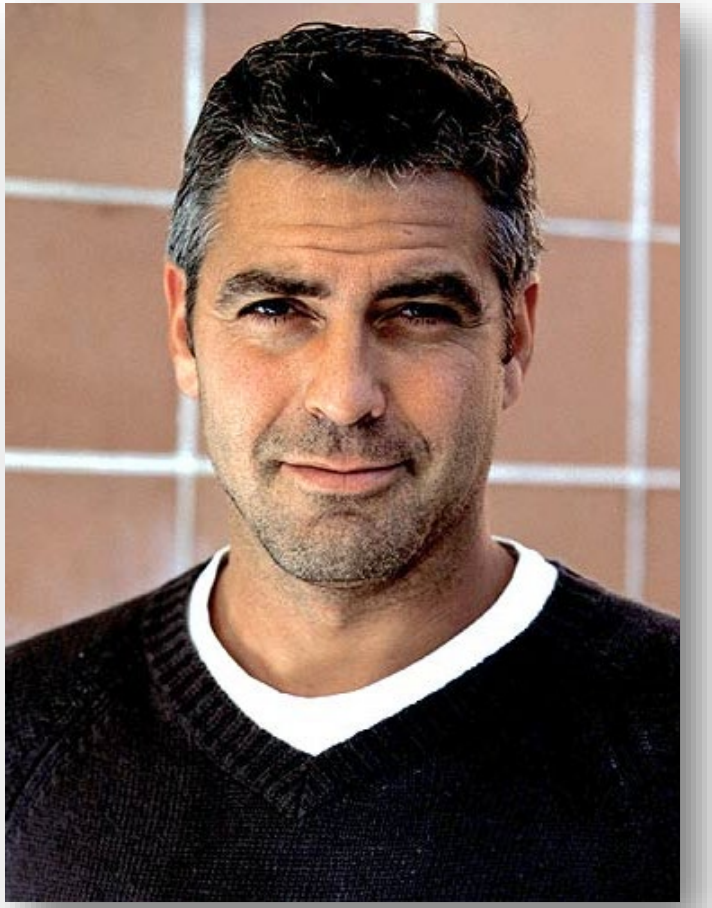
Wand et al., Nature, 2021

Who has a higher risk to suffer from stroke or MI within the next 10 years?

female, 27 yo

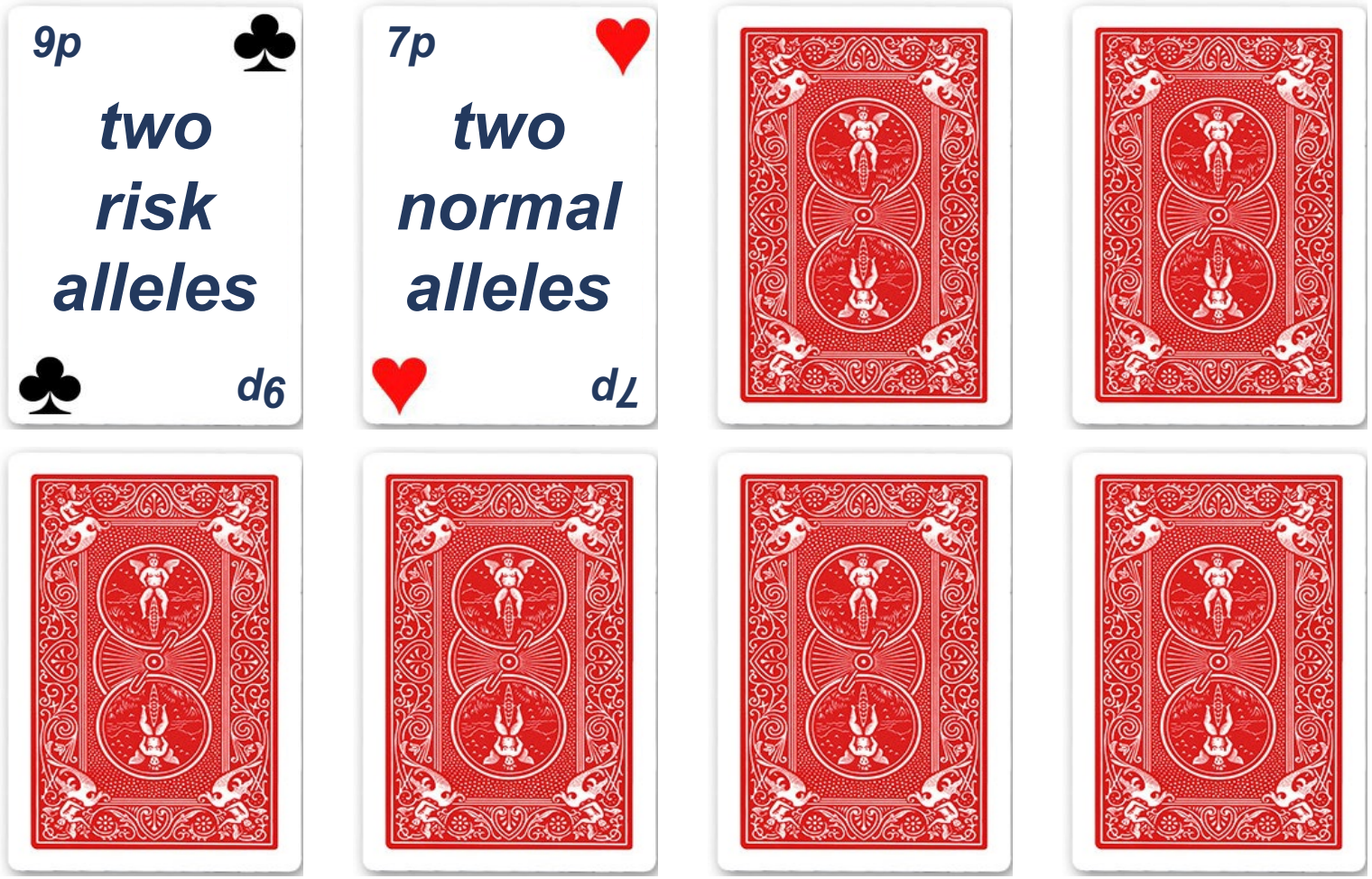
male, 78 yo, diabetic, hypertensive, smoker

Who has a higher risk to to suffer from stroke or MI within the next 10 years?

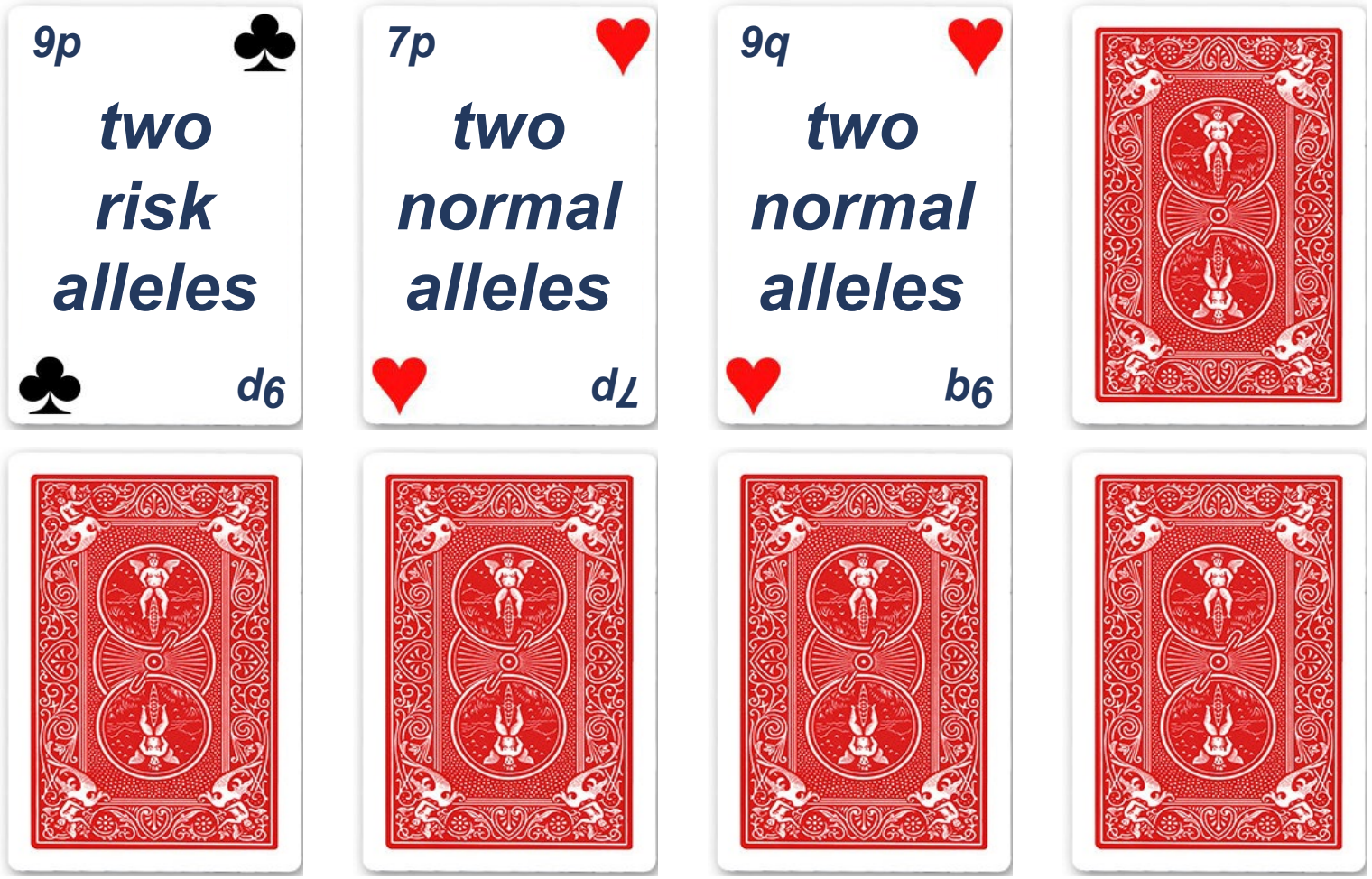


Computing individual probability

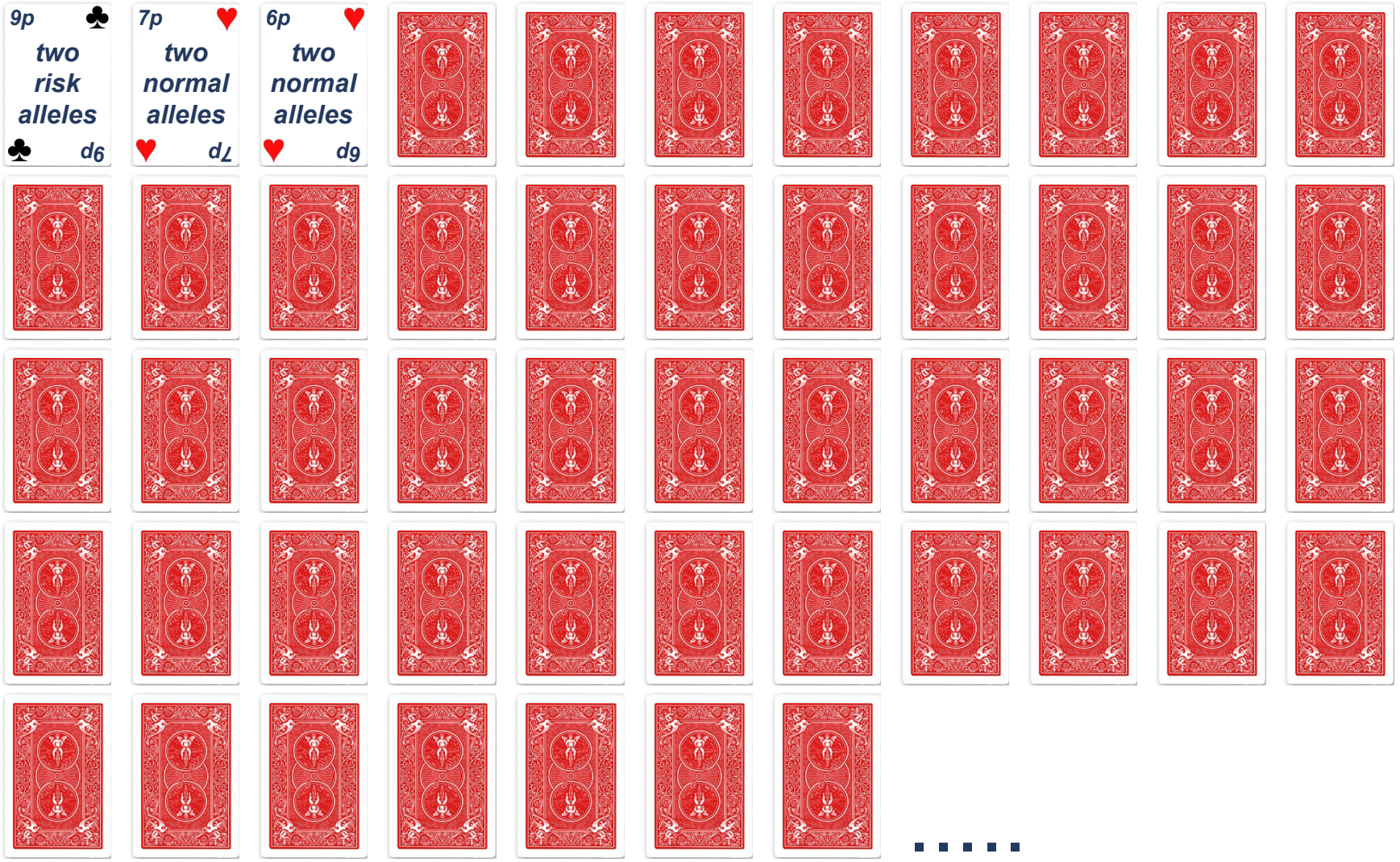
Computing individual probability



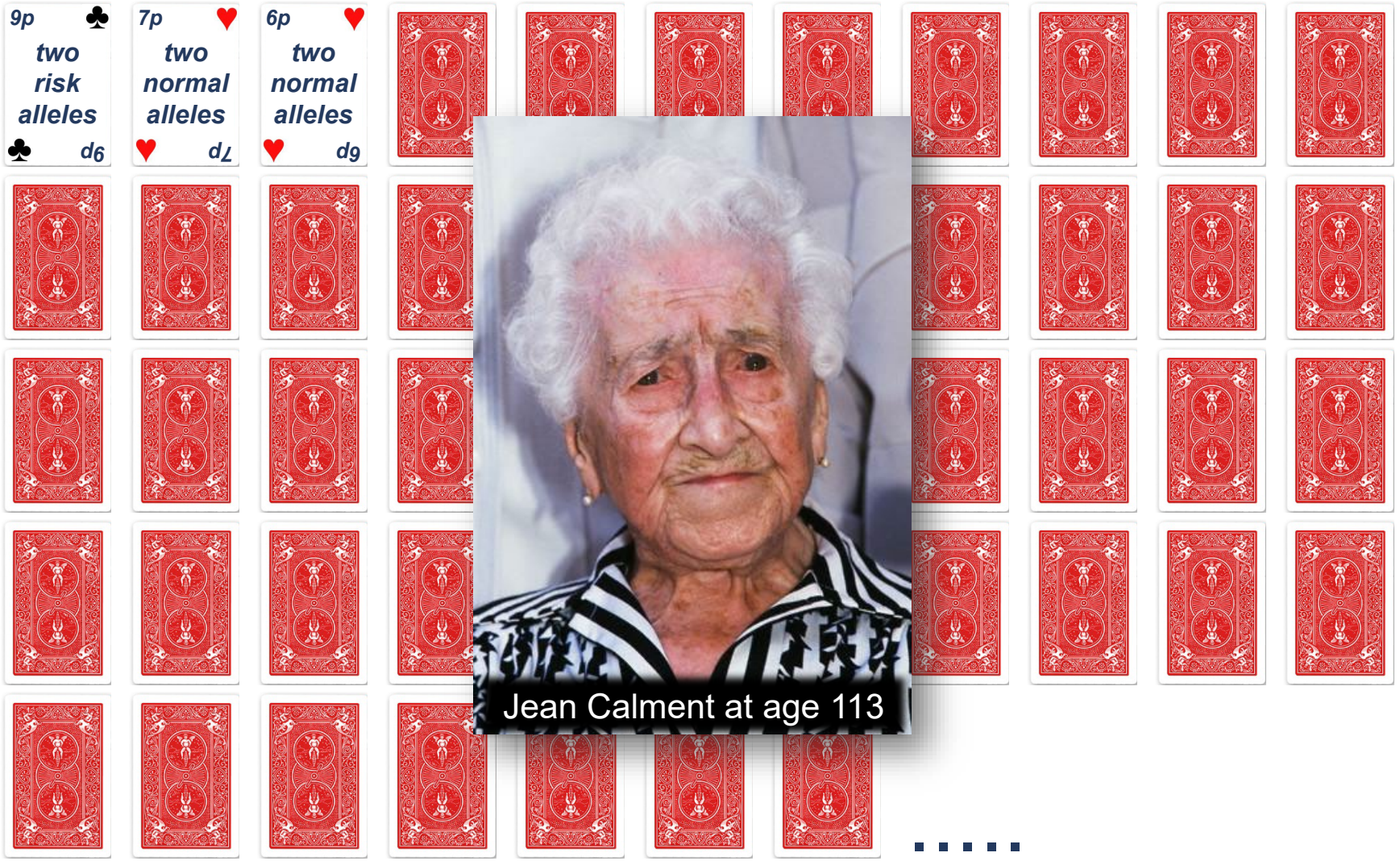
Computing individual probability



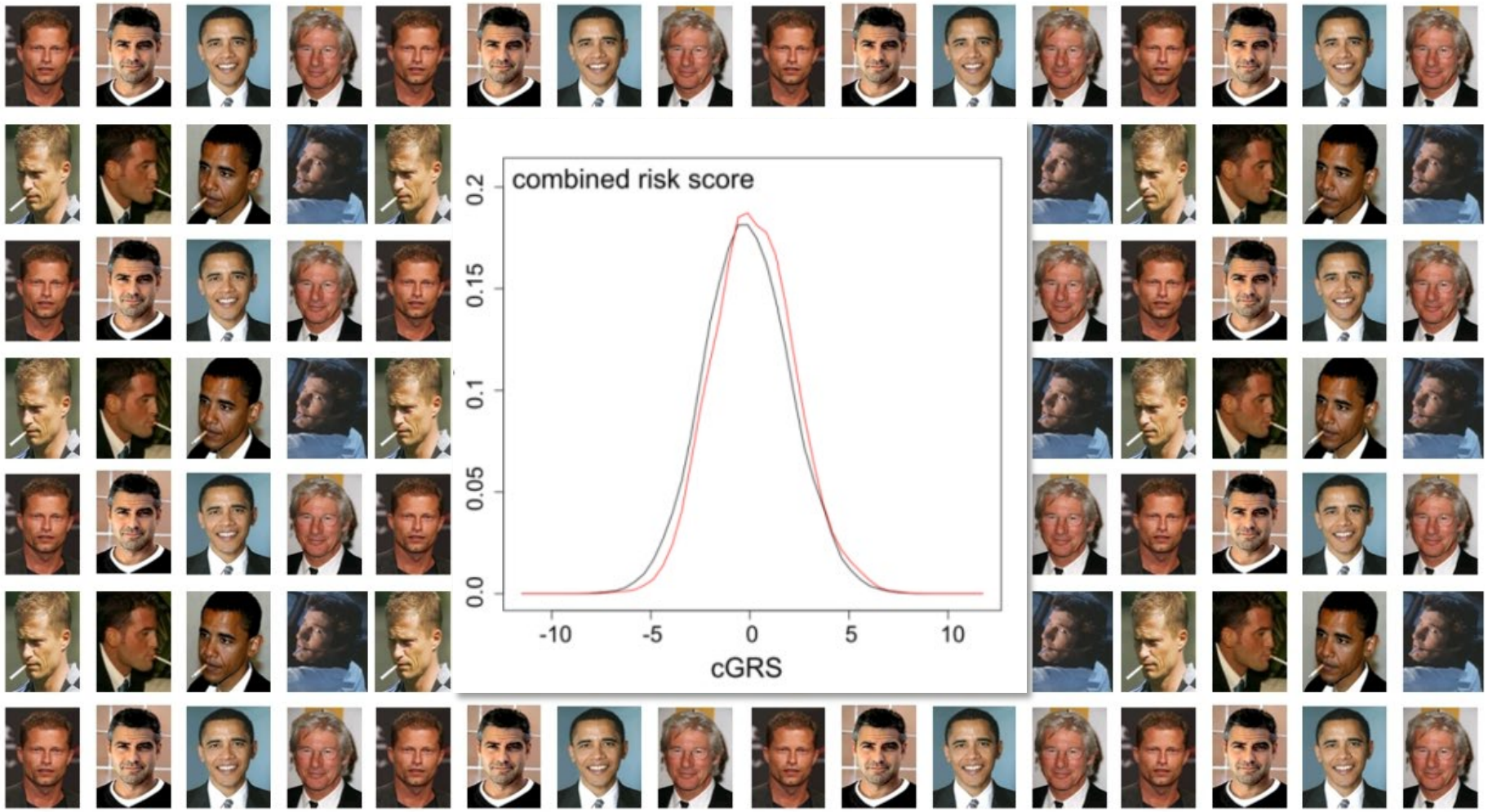
Computing individual probability



Computing individual probability



Most individuals have a moderate number of risk alleles



Malik et al. Stroke 2014

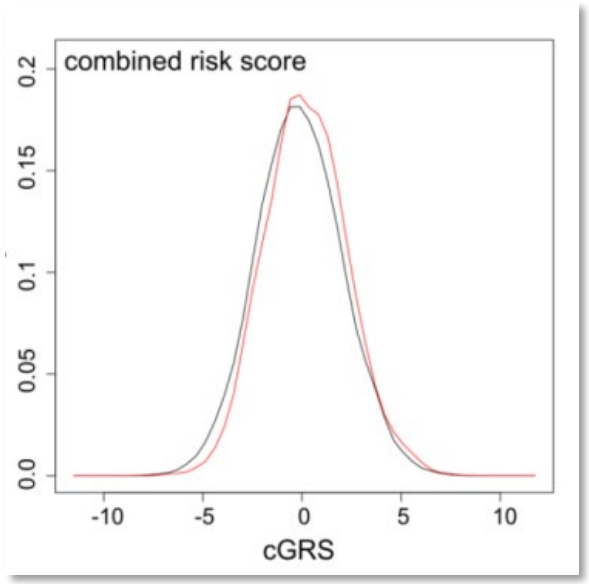
PRS calculation

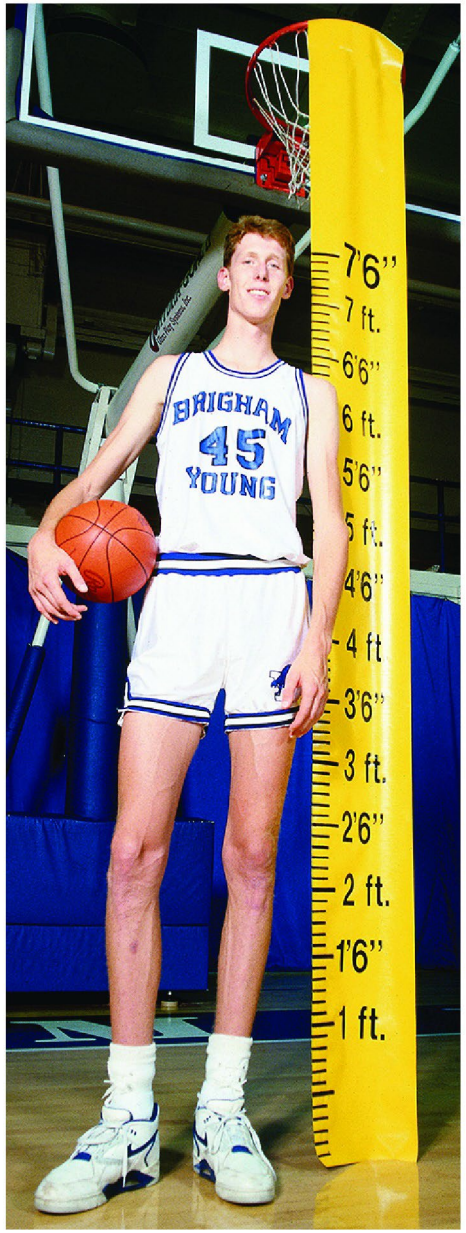
- **Step 1**:
  - Take SNP effects satisfying certain criteria (p-value cutoff, LD independence) from GWAS from discovery cohort for a certain phenotype with associated effect size

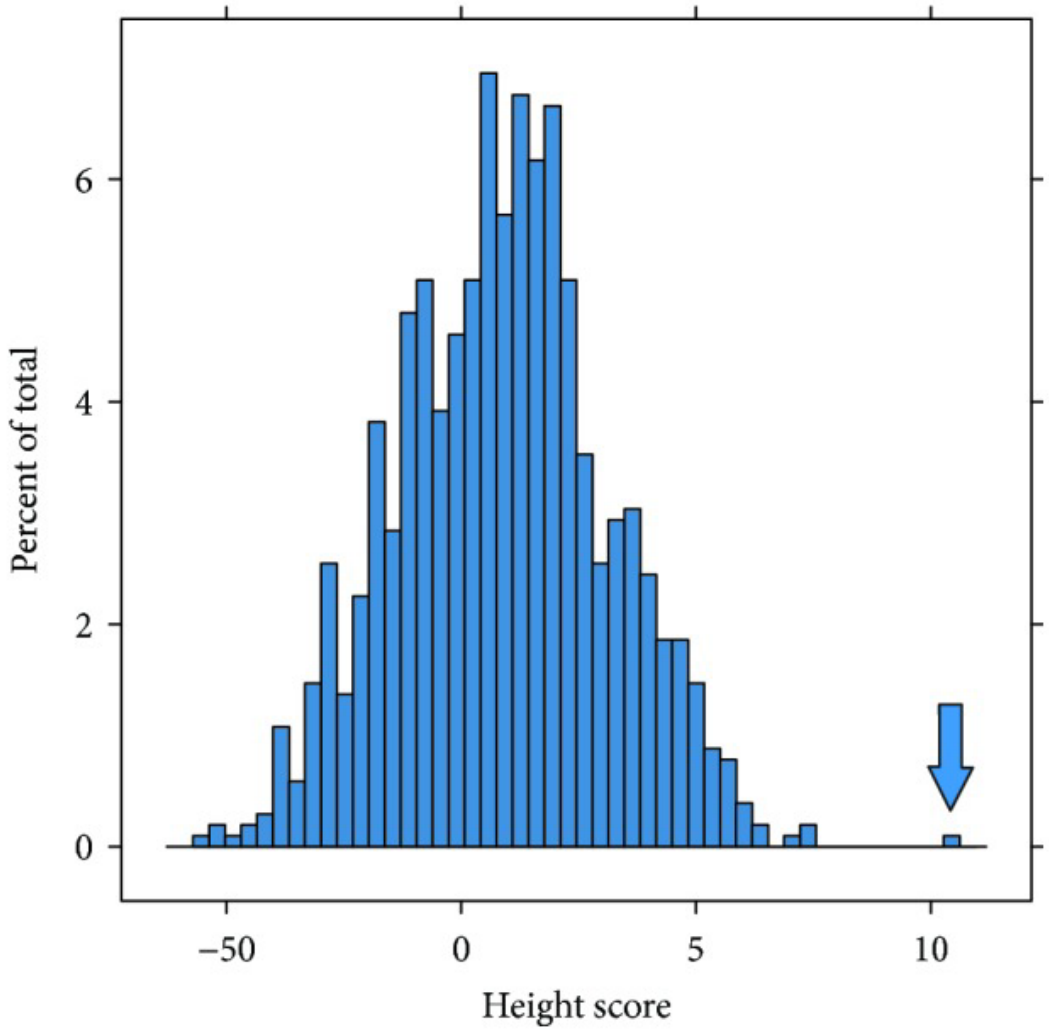
  - SNP list:
    - SNP1 (A)  0.84
    - SNP2 (T)  0.21
    - …  0.31
    - …
    - …
    - …
    - …
    - …
    - …
    - …
    - …

- **Step 2**:
  - For each individual in **INDEPENDENT** target population, calculate risk score by summing number of alleles, weighted by effect size and dividing through number of alleles:
    - Individual 1: SNP1 (AA), SNP2 (TC): 0.156
    - Individual 2: SNP1 (TT), SNP2 (TT): 0.175
    - Individual 3: SNP1 (AT), SNP2 (CC): 0.03425

  - Result: Risk score of each individual – centered at 0, normal distribution







Institute for Stroke and Dementia Research

Sexton et al., Int J Genomics, 2018

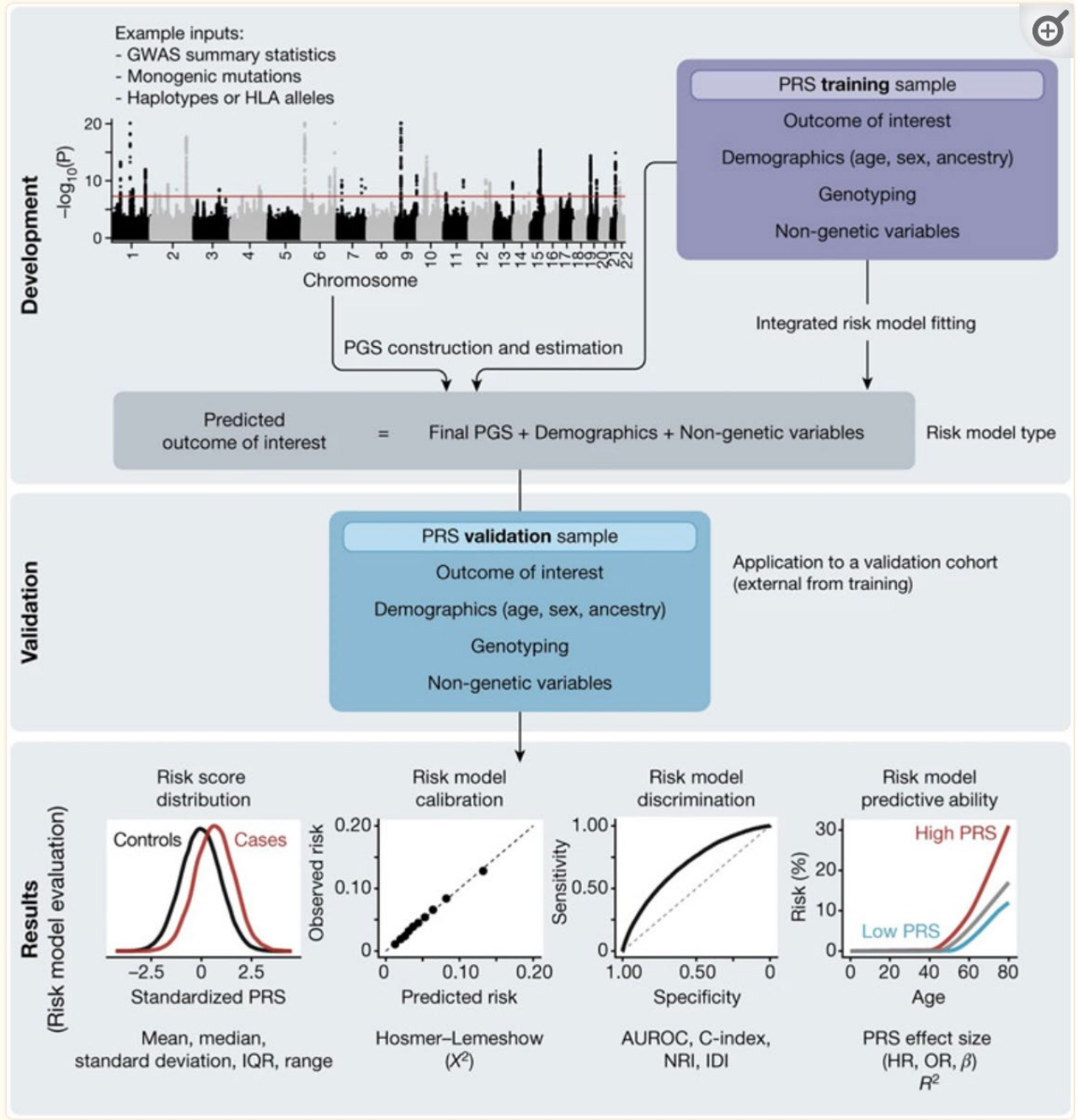
- Step 3:
    - Does the overall „score“ associate with biomarker, incidence, case/control status etc. in the idenpendent cohort?

linear regression (for continuous traits), logistic regression (for case/control analysis), Cox regression (for time to event analysis)

Model:

$$\text{Outcome} \sim \text{PRS} + \text{covariates}$$

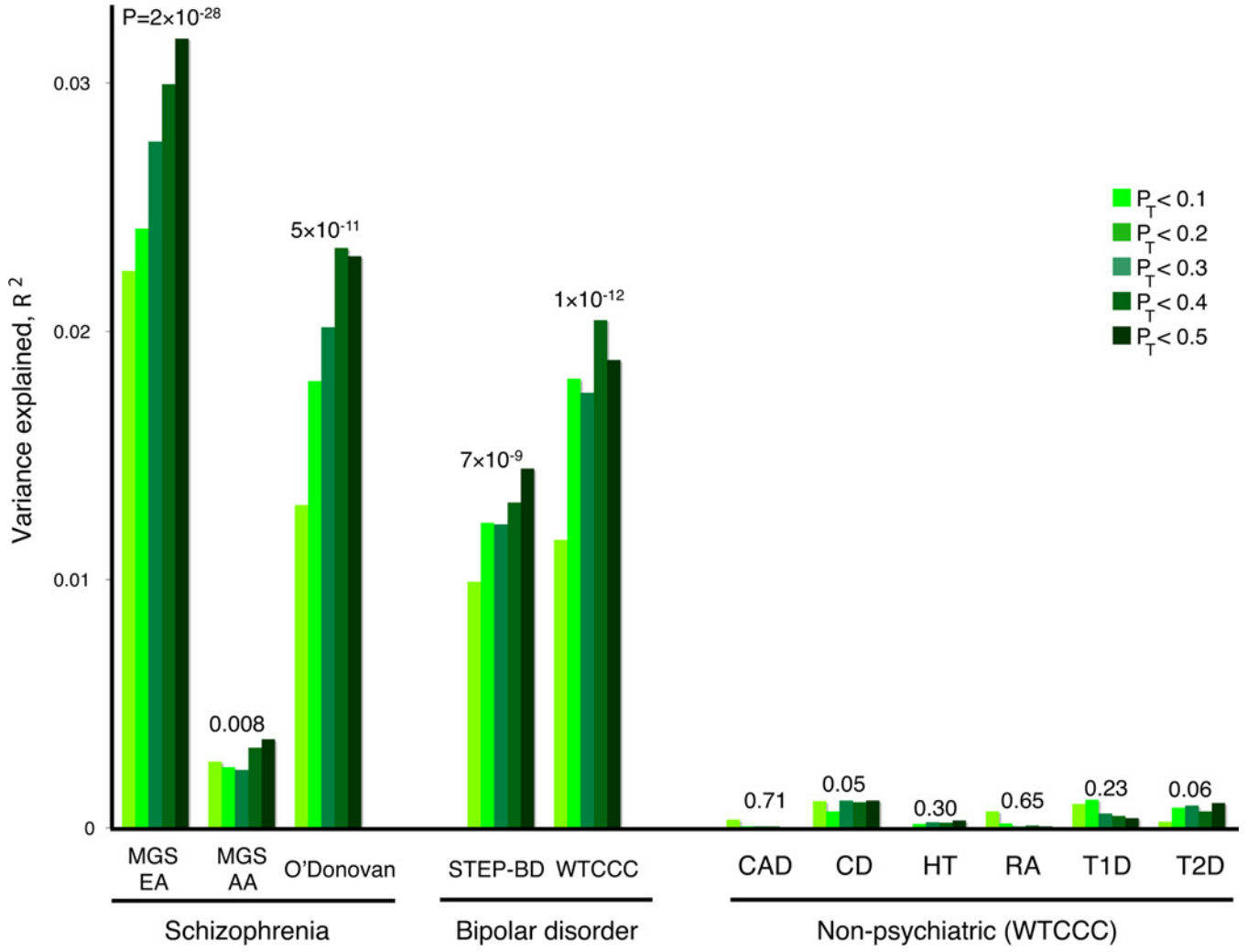
Outcome measures: AUC, c-index, variance explained (R2), p-value of association



Institute for Stroke and Dementia Research

Wand et al., Nature, 2021

One of the first high impact PRS on schizophrenia



Institute for Stroke and Dementia Research

Purcell et al., Nature, 2009

OPEN ACCESS Freely available online

PLOS | GENETICS

Power and Predictive Accuracy of Polygenic Risk Scores

Frank Dudbridge*

Faculty of Epidemiology and Population Health, London School of Hygiene and Tropical Medicine, London, United Kingdom

Abstract

Polygenic scores have recently been used to summarise genetic effects among an ensemble of markers that do not individually achieve significance in a large-scale association study. Markers are selected using an initial training sample and used to construct a score in an independent replication sample by forming the weighted sum of associated alleles within each subject. Association between a trait and this composite score implies that a genetic signal is present among the selected markers, and the score can then be used for prediction of individual trait values. This approach has been used to obtain evidence of a genetic effect when no single markers are significant, to establish a common genetic basis for related disorders, and to construct risk prediction models. In some cases, however, the desired association or prediction has not been achieved. Here, the power and predictive accuracy of a polygenic score are derived from a quantitative genetics model as a function of the sizes of the two samples, explained genetic variance, selection thresholds for including a marker in the score, and methods for weighting effect sizes in the score. Expressions are derived for quantitative and discrete traits, the latter allowing for case/control sampling. A novel approach to estimating the variance explained by a marker panel is also proposed. It is shown that published studies with significant association of polygenic scores have been well powered, whereas those with negative results can be explained by low sample size. It is also shown that useful levels of prediction may only be approached when predictors are estimated from very large samples, up to an order of magnitude greater than currently available. Therefore, polygenic scores currently have more utility for association testing than predicting complex traits, but prediction will become more feasible as sample sizes continue to grow.

Citation: Dudbridge F (2013) Power and Predictive Accuracy of Polygenic Risk Scores. PLoS Genet 9(3): e1003348. doi:10.1371/journal.pgen.1003348

Editor: Naomi R. Wray, Queensland Institute of Medical Research, Australia

Received May 24, 2012; **Accepted** January 16, 2013; **Published** March 21, 2013

Funding: This work was funded by the Medical Research Council, grant number G1000718 (www.mrc.ac.uk). The funder had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The author has declared that no competing interests exist.

* E-mail: frank.dudbridge@lshtm.ac.uk

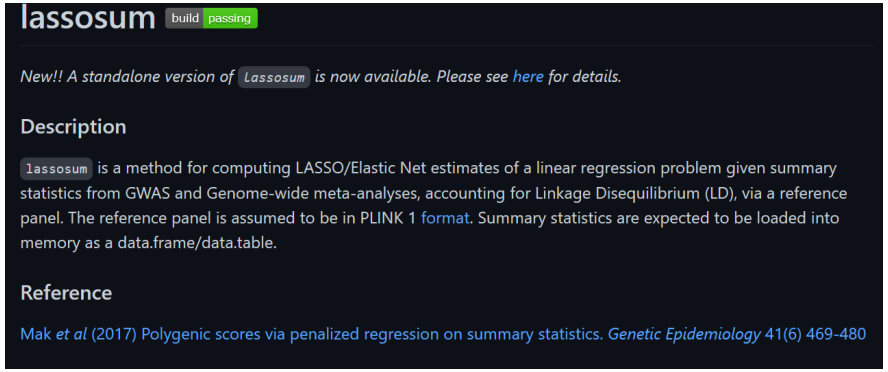
Practical considerations

- **Many methods are available that compute PRS directly from GWAS summary statistics**

PRSice-2: Polygenic Risk Score software

PRSice (pronounced 'precise') is a Polygenic Risk Score software for calculating, applying, evaluating and plotting the results of polygenic risk scores (PRS) analyses. Some of the features include:

1. High-resolution scoring (PRS calculated across a large number of P-value thresholds)
2. Identify Most predictive PRS
3. Empirical P-values output (not subject to over-fitting)
4. Genotyped (PLINK binary) and imputed (Oxford bgen v1.2) data input
5. Biobank-scale genotyped data can be analysed within hours
6. Incorporation of covariates
7. Application across multiple target traits simultaneously
8. Results plotted in several formats (bar plots, high-res plots, quantile plots)
9. PRSet: function for calculating PRS across user-defined pathways / gene sets



Background

LDpred-2 is one of the dedicated PRS programs which is an `R` package that uses a Bayesian approach to polygenic risk scoring.

Installing LDpred-2

Note

The script used here is based on LDpred 2 implemented under bigsnpr version 1.4.7

Note

For more details, please refer to LDpred 2's homepage

You can install `LDpred` and its dependencies in `R` with the following command:

```
install.packages("remotes")
library(remotes)
remotes::install_github("https://github.com/privefl/bigsnpr.git")
```

PRS catalogue

The Polygenic Score (PGS) Catalog

An open database of polygenic scores and the relevant metadata required for accurate application and evaluation.

Search the PGS Catalog

Examples: breast cancer, glaucoma, BMI, EFO_0001645

New tool!

We just released **pgsc_calc**: a reproducible workflow to calculate both PGS Catalog and custom polygenic scores. > See more information

Explore the Data

In the current PGS Catalog you can **browse** the scores and metadata through the following categories:

Polygenic Scores	Traits	Publications
3,349	584	426

Trait: **Ischemic stroke**

Trait Information	
Identifier	HP_0002140
Description	Acute ischemic stroke (AIS) is defined by the sudden loss of blood flow to an area of the brain with... Show more >
Trait category	Other trait
Synonym	Ischaemic stroke
Mapped terms	2 mapped terms

View in NHGRI-EBI GWAS Catalog

Associated Polygenic Score(s)

Filter PGS by Participant Ancestry

Individuals included in: All Stages combined [G, D, E]

G - Source of Variant Associations (GWAS)
D - Score Development/Training
E - PGS Evaluation

List of ancestries includes: --

Display options:
Show European ancestry data
Show only Multi-ancestry data

Ancestry legend

Multi-ancestry (including European)
Multi-ancestry (excluding European)
African
East Asian
South Asian
Additional Asian Ancestries
European
Greater Middle Eastern
Hispanic or Latin American
Additional Diverse Ancestries
Not Reported

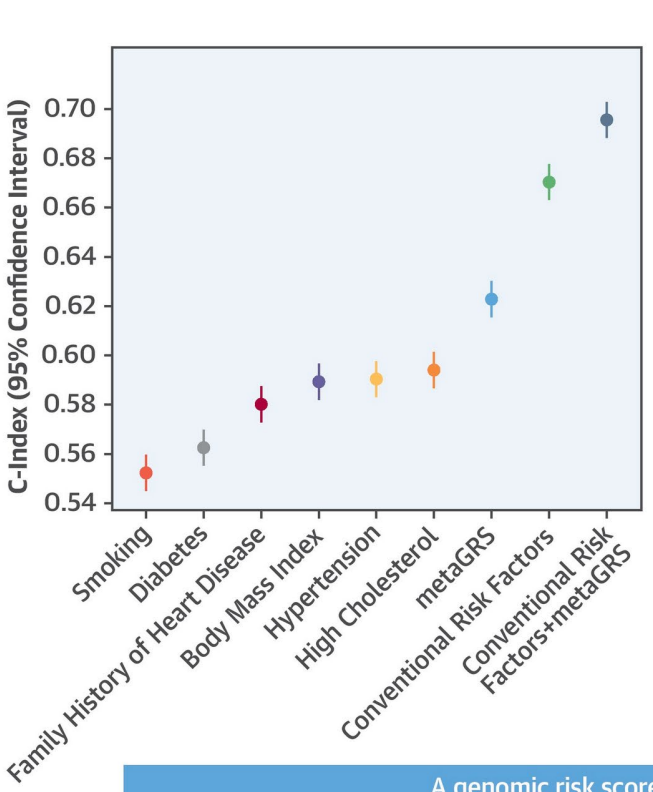
Search

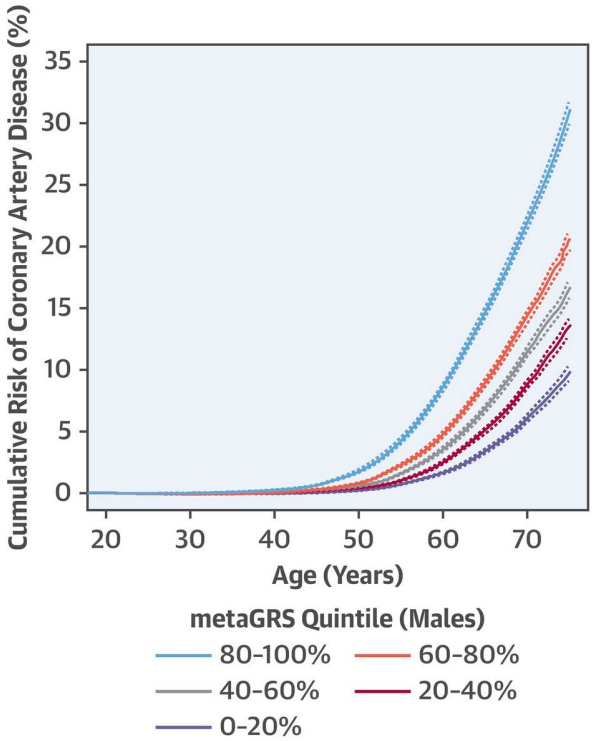
Polygenic Score ID & Name	PGS Publication ID (PGP)	Reported Trait	Mapped Trait(s) (Ontology)	Number of Variants	Ancestry distribution	Scoring File (FTP Link)
PGS000039 (metaGRS_ischaemicstroke)	PGP000027 » Abraham G et al. Nat Commun (2019)	Ischaemic stroke	stroke, Ischemic stroke	3,225,583	G D E	
PGS000665 (GRS_32)	PGP000125 » Marston NA et al. Circulation (2020)	Ischemic stroke	stroke, Ischemic stroke	32	G - E	
PGS000911 (PRS_IS)	PGP000239 » O'Sullivan JW et al. Circ Genom Precis Med (2021)	Ischemic stroke	stroke, Ischemic stroke	530,933	G D E	
PGS002724 (GIGASTROKE_PGS_EUR)	PGP000333 » Mishra A et al. Nature (2022)	Ischemic stroke	stroke, Ischemic stroke	1,213,574	G D E	
PGS002725 (GIGASTROKE_PGS_EAS)	PGP000333 » Mishra A et al. Nature (2022)	Ischemic stroke	stroke, Ischemic stroke	6,010,730	G D E	

Institute for Stroke and Dementia Research

Cardiovascular disease (Kathiresan, Inouye)

CENTRAL ILLUSTRATION: Genomic Risk Score for Coronary Artery Disease





A genomic risk score for coronary artery disease

Greater association with future coronary artery disease than any single conventional risk factor
Independent of yet complements conventional risk factors
Provides meaningful lifetime risk estimates of coronary artery disease
Quantifiable at or before birth and shows potential for risk screening in early life

Inouye, M. et al. J Am Coll Cardiol. 2018;72(16):1883-93.

Integrated GRS - metaGRS

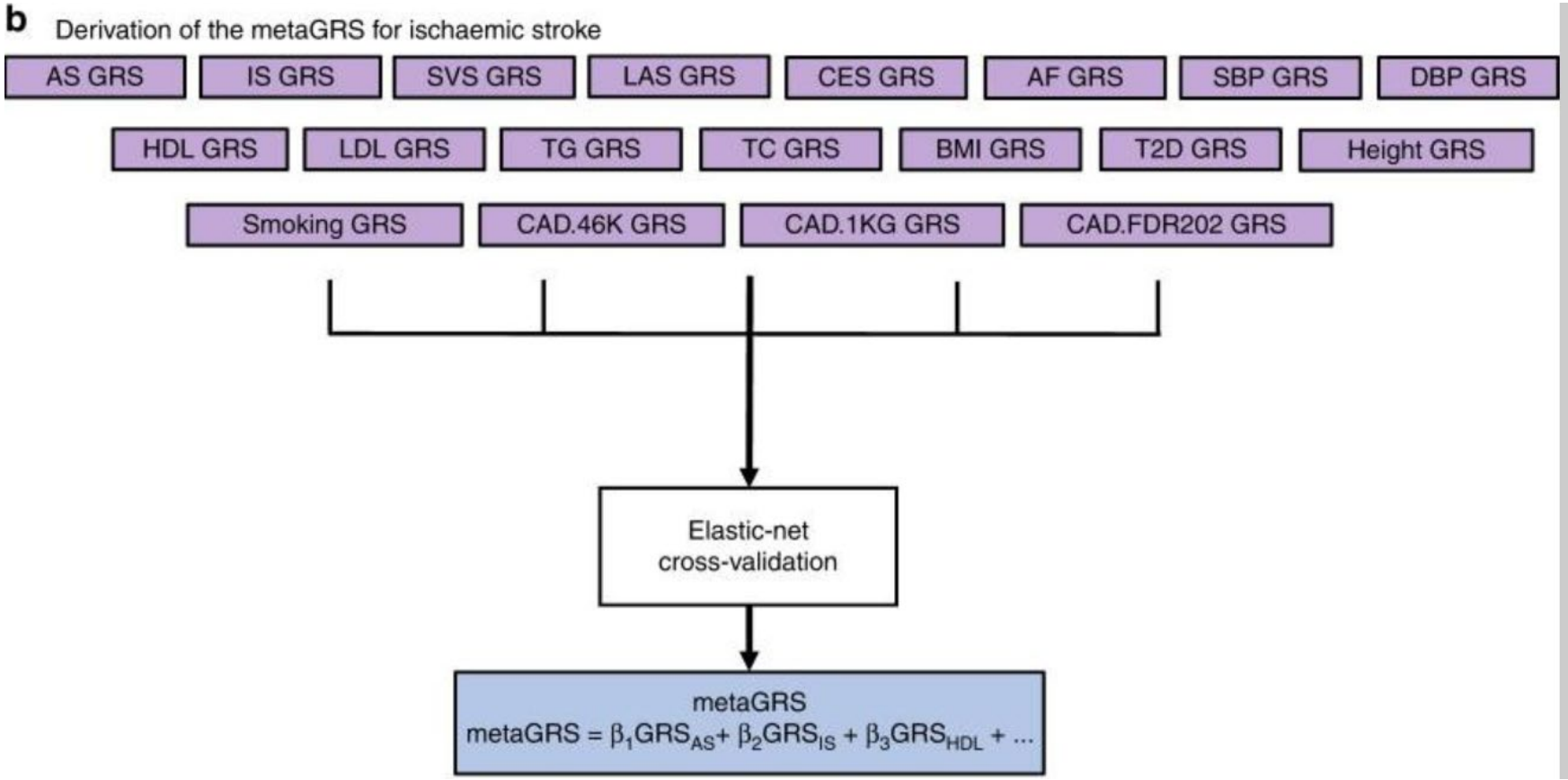
Integrated Risk Model:

a risk model for the outcome of interest which combines PRS with other risk factors, such as demographics (often age and sex), anthropometrics, biomarkers, and clinical measurements.

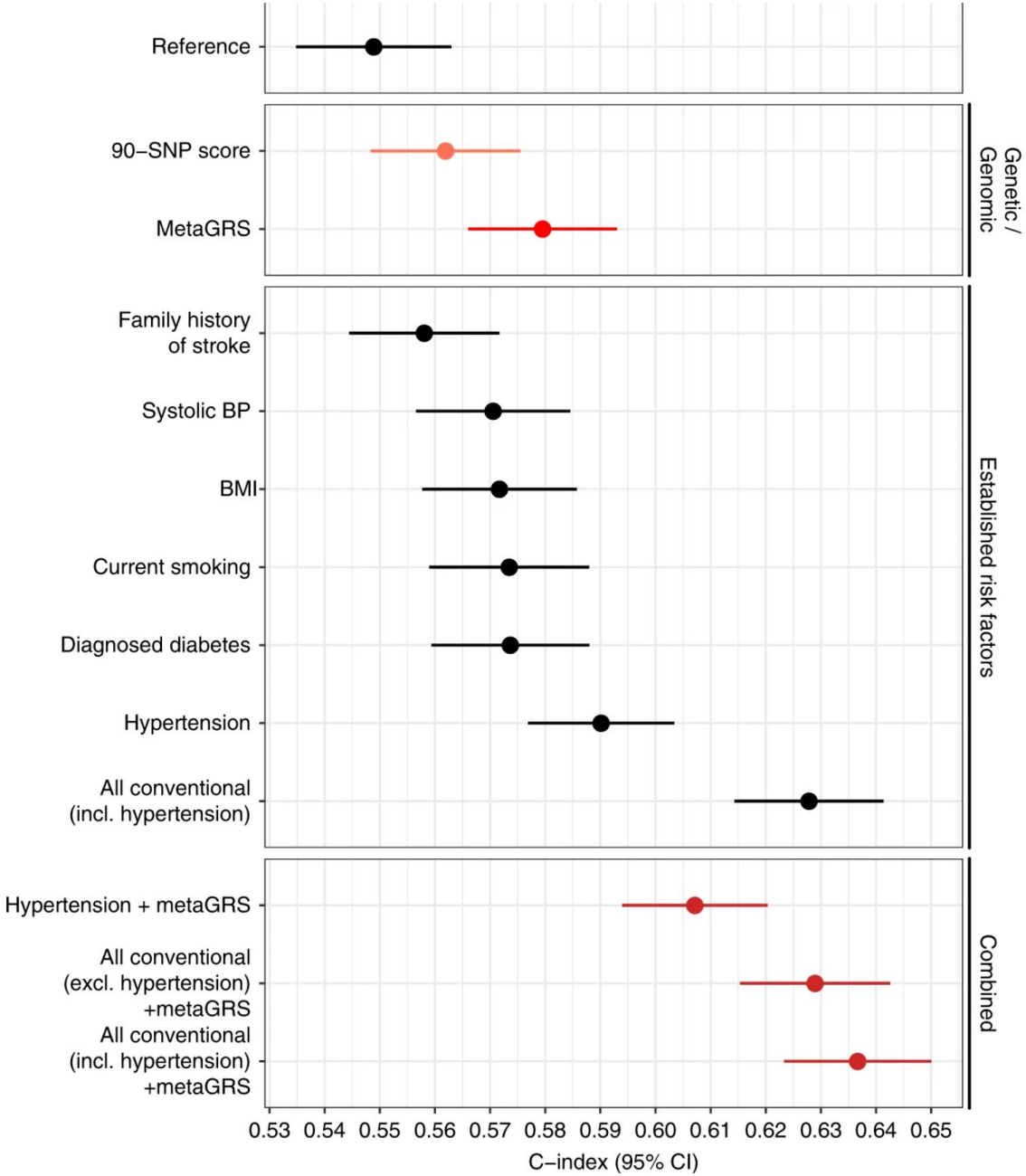
Categories of use for PRS and/or integrated risk models The addition of PRS to existing risk models has several potential applications, summarized below. Each aims to improve individual or subgroup classification such that there is clinical benefit.

Wand et al., Nature, 2021

metaGRS (Abraham)



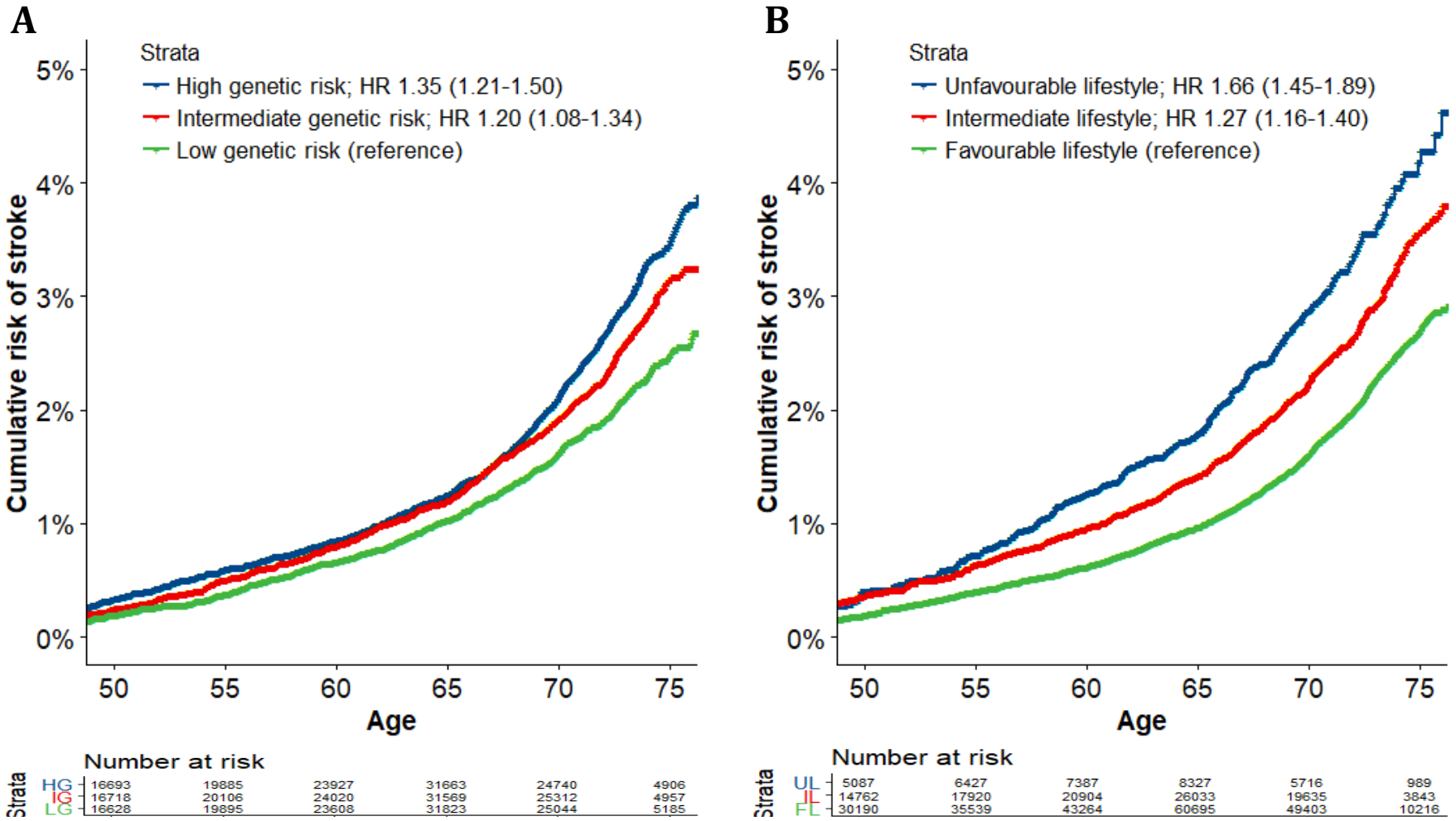
Abraham*, Malik* et al., Nature Communications, 2020



Institute for Stroke and Dementia Research

Abraham*, Malik* et al., Nature Communications, 2020

Comparison to lifestyle

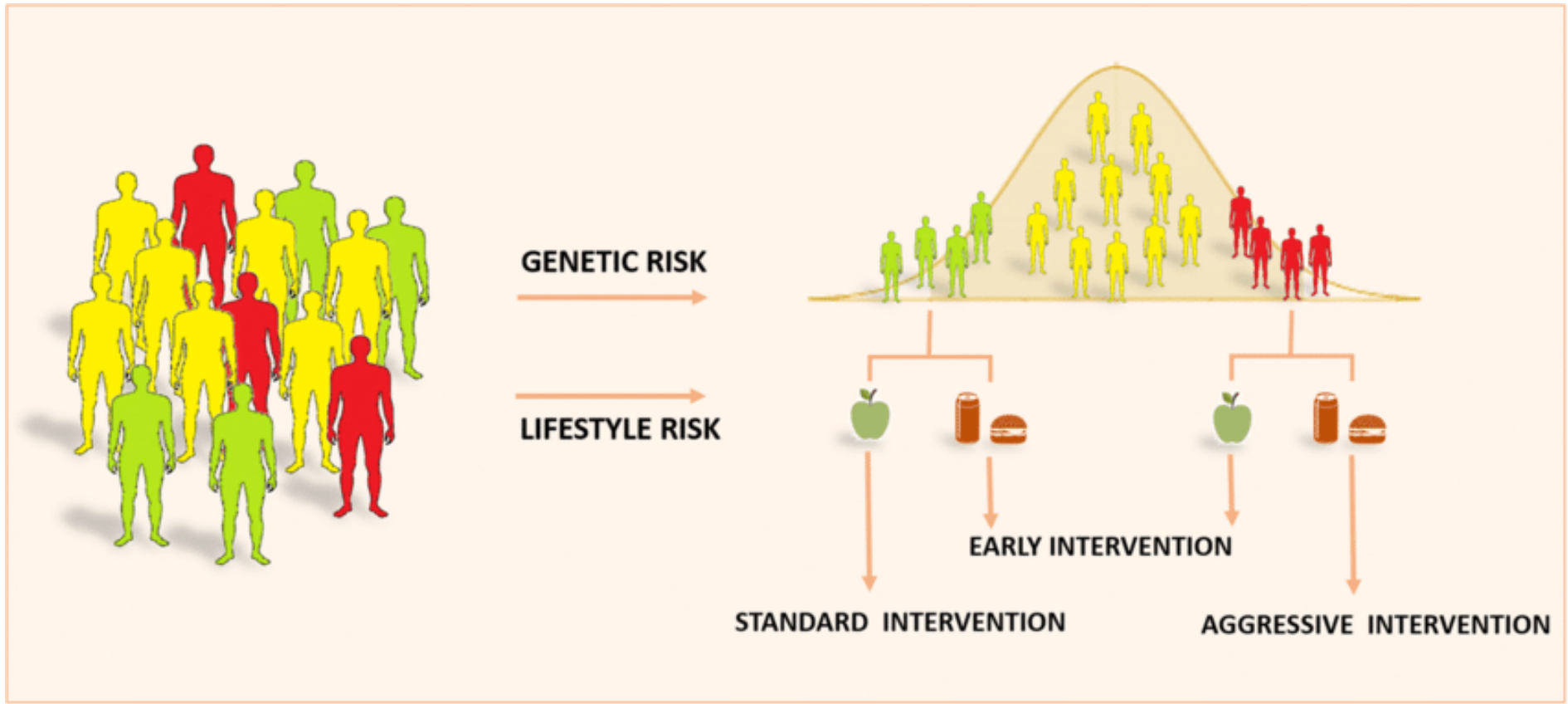


Institute for Stroke and Dementia Research

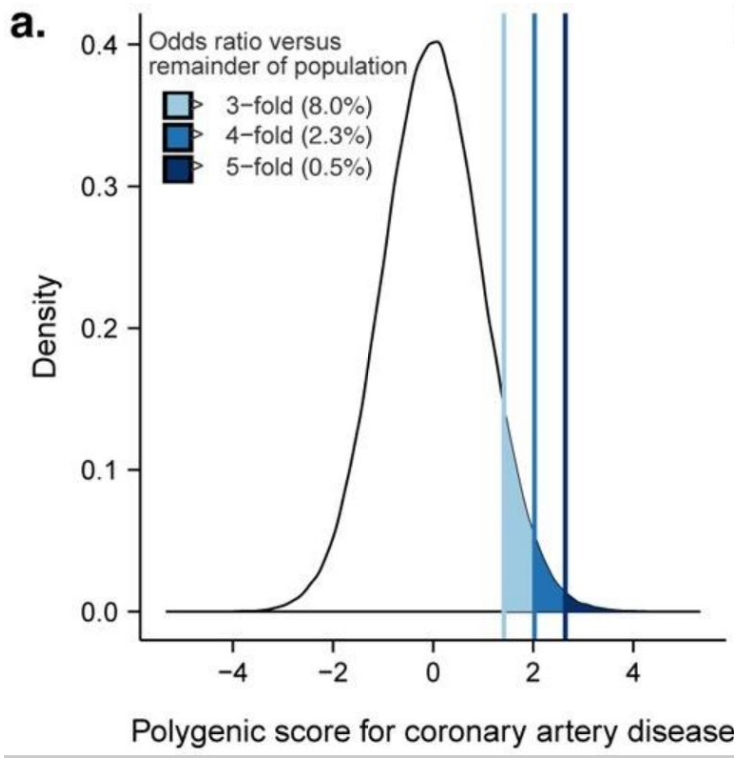
Rutten-Jacobs, Larsson, Malik et al., BMJ, 2018

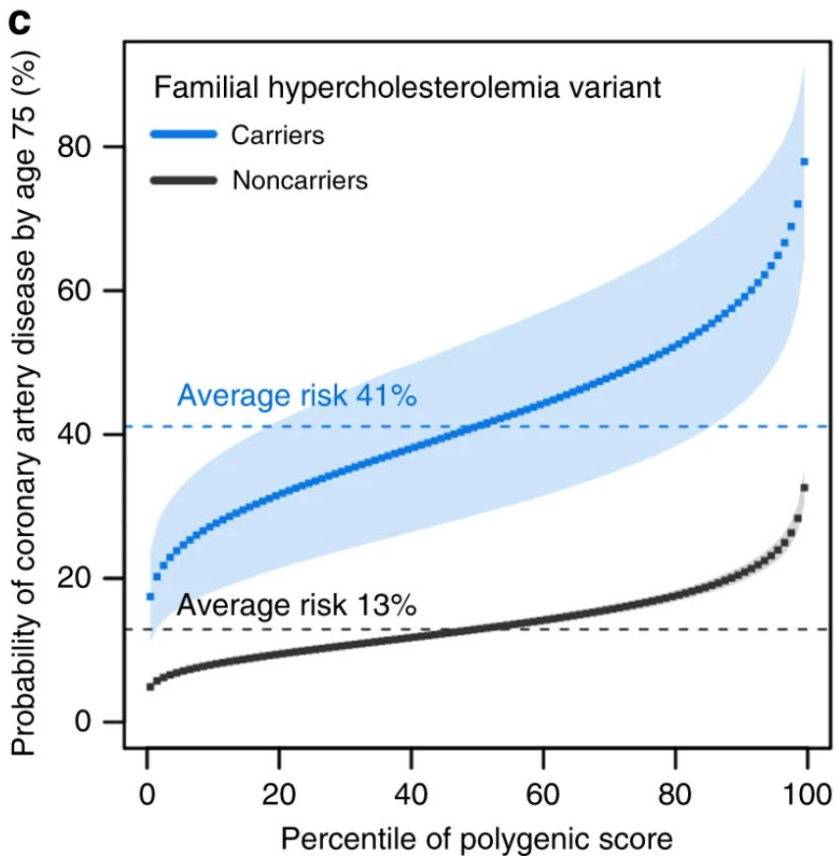
Genetic risk	Lifestyle		
	Favourable	Intermediate	Unfavourable
Low			
Hazard ratio* (95% CI)	1 (reference)	1.36 (1.14 to 1.63), $P=7.3\times10^{-04}$	1.84 (1.44 to 2.35), $P=8.0\times10^{-07}$
8 year cumulative incidence† (%) (95% CI)	0.54 (0.47 to 0.60)	0.74 (0.63 to 0.85)	0.95 (0.74 to 1.17)
Intermediate			
Hazard ratio* (95% CI)	1.26 (1.09 to 1.46), P=0.002	1.62 (1.37 to 1.92), $P=3.2\times10^{-08}$	1.85 (1.46 to 2.37), $P=5.4\times10^{-07}$
8 year cumulative incidence† (%) (95% CI)	0.67 (0.60 to 0.74)	0.82 (0.71 to 0.93)	0.92 (0.72 to 1.12)
High			
Hazard ratio* (95% CI)	1.44 (1.25 to 1.66), $P=7.0\times10^{-07}$	1.70 (1.44 to 2.01), $P=8.1\times10^{-10}$	2.30 (1.84 to 2.87), $P=3.3\times10^{-13}$
8 year cumulative incidence† (%) (95% CI)	0.78 (0.70 to 0.86)	0.91 (0.78 to 1.04)	1.11 (0.89 to 1.33)

Rutten-Jacobs, Larsson, Malik et al., BMJ, 2018


GENETIC RISK
LIFESTYLE RISK
EARLY INTERVENTION
STANDARD INTERVENTION
AGGRESSIVE INTERVENTION

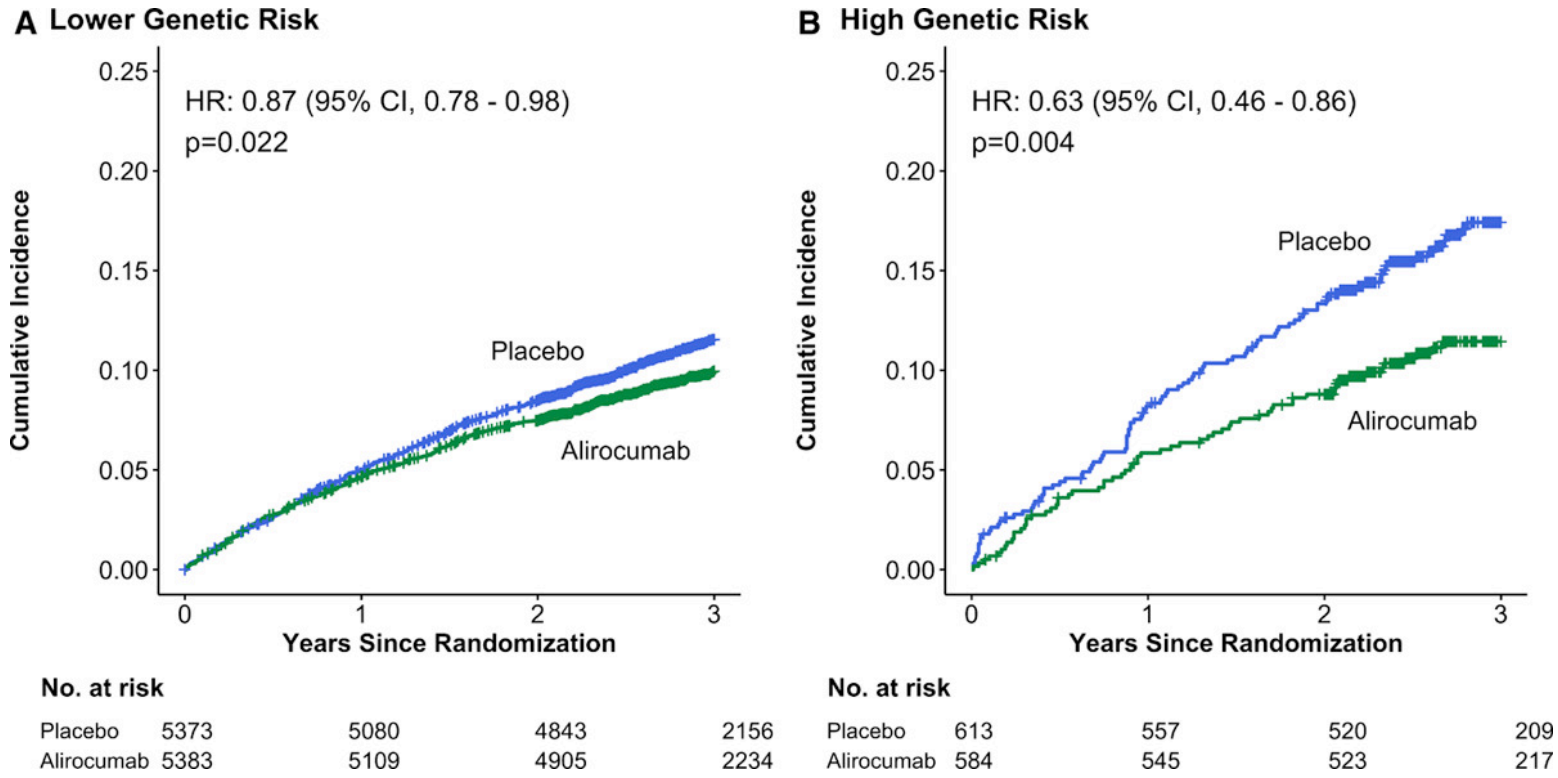
Which risk is higher? Monogenic vs. Polygenic





Khera et al., Nature Genetics, 2018
Fahed et al. Nature Communications, 2020

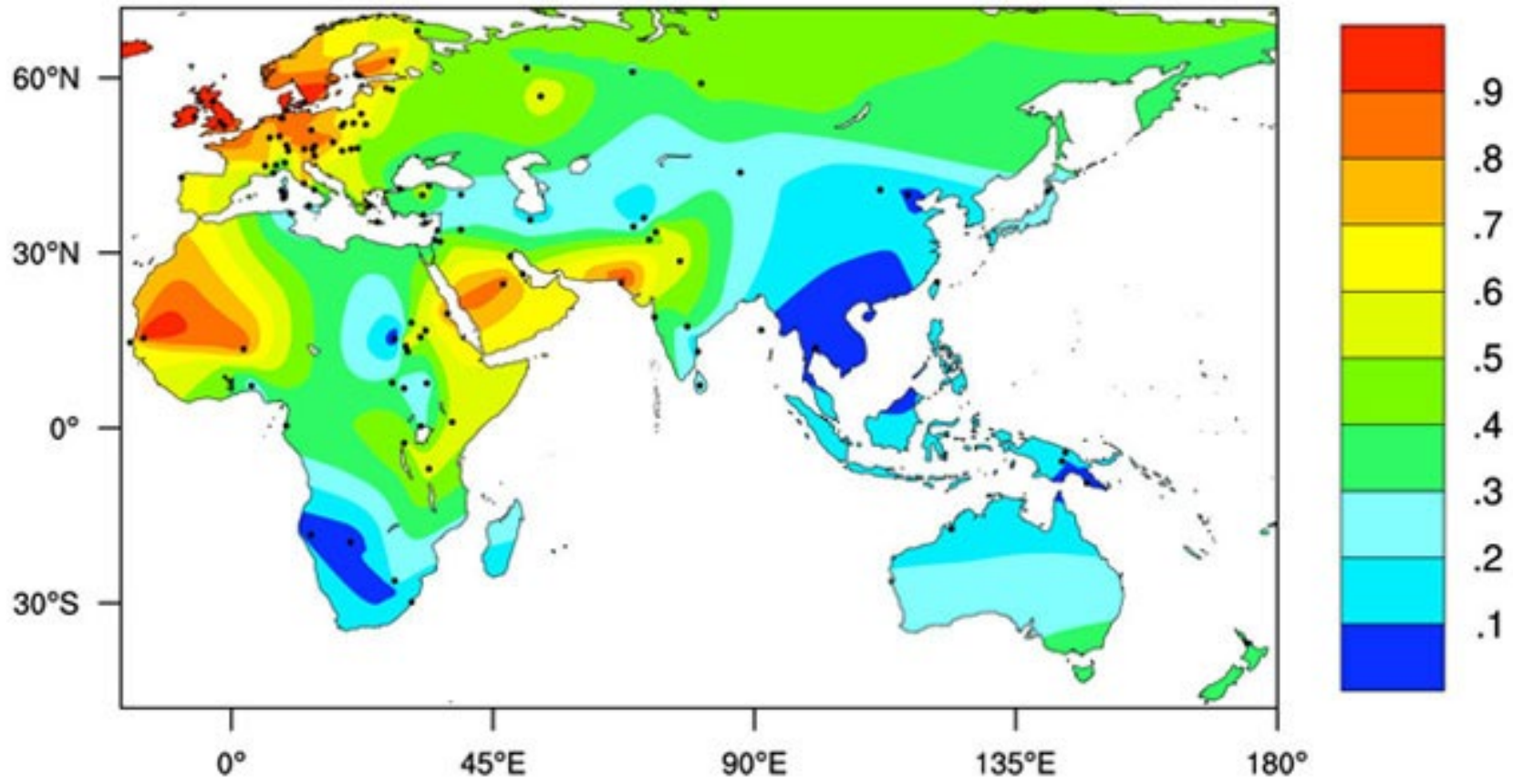
ODYSSEY OUTCOMES – secondary prevention after MI with PCSK9 inhibitor on top of statins



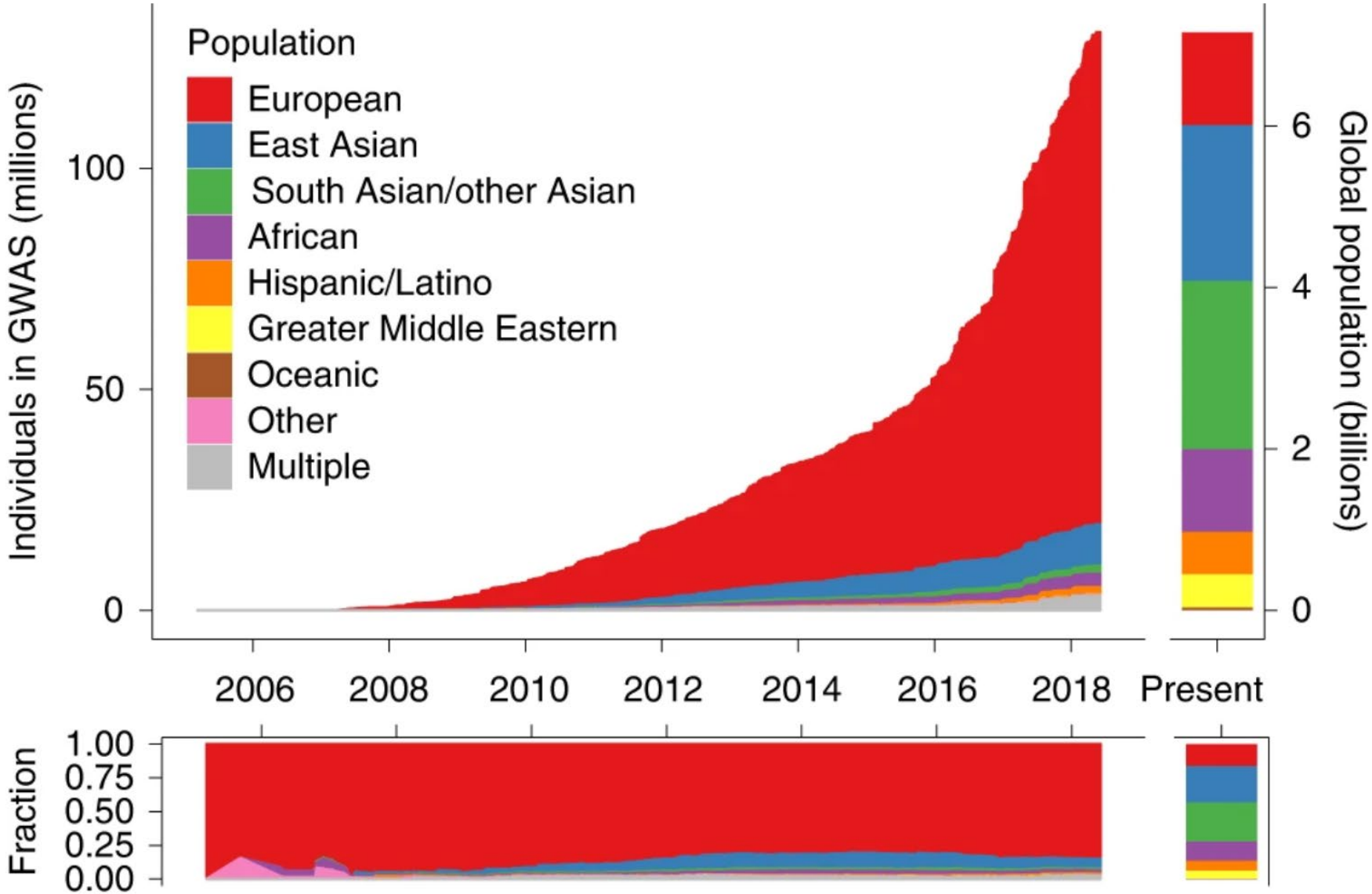
Institute for Stroke and Dementia Research

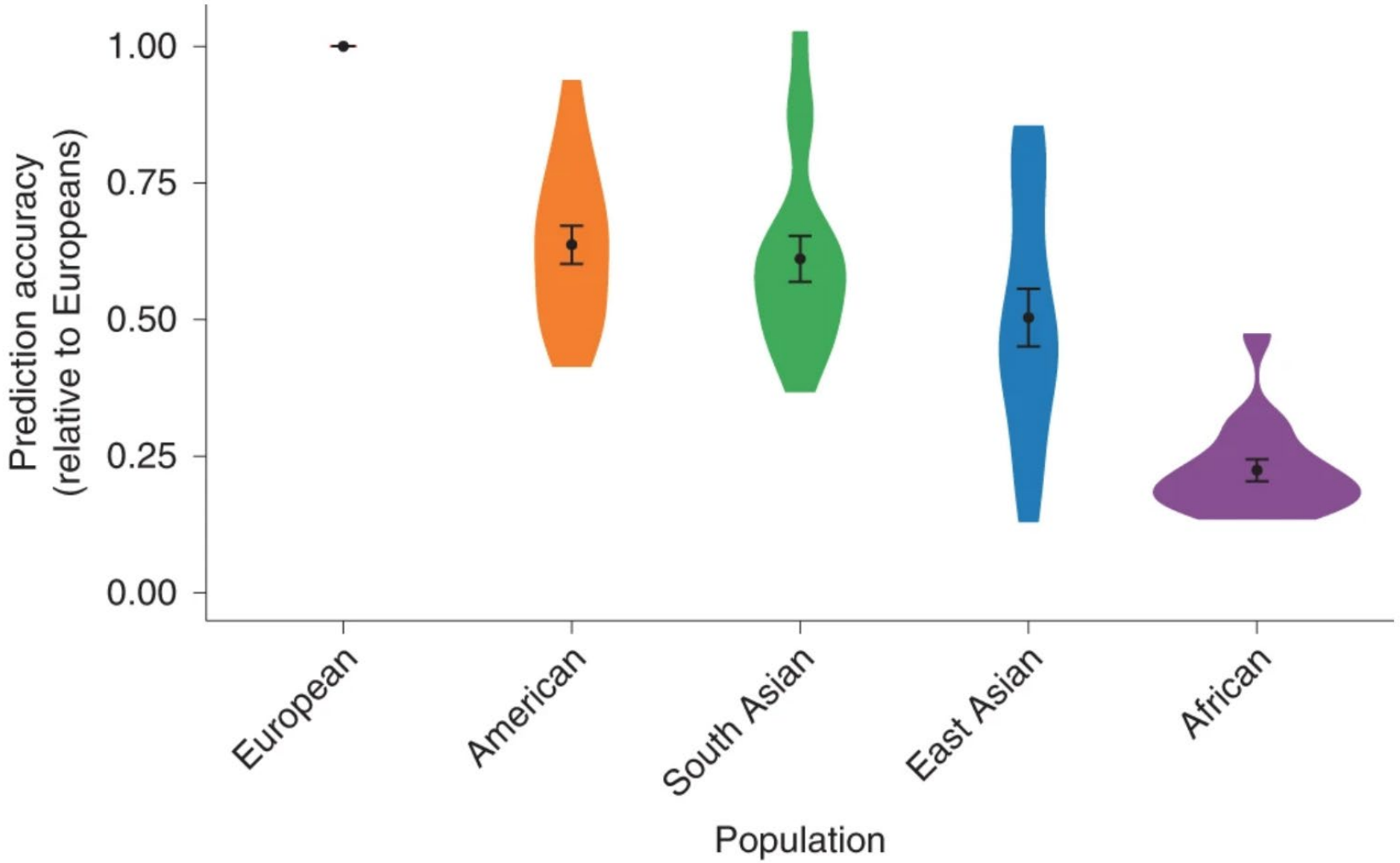
Damask et al., Circulation, 2019

Transethnic considerations



Itan et al. BMC Evolutionary Biology 2010, 10:36


Population
European
East Asian
South Asian/other Asian
African
Hispanic/Latino
Greater Middle Eastern
Oceanic
Other
Multiple
Individuals in GWAS (millions)
100
50
0
2006
2008
2010
2012
2014
2016
2018
Present
Global population (billions)
6
4
2
0
Fraction
1.00
0.75
0.50
0.25
0.00

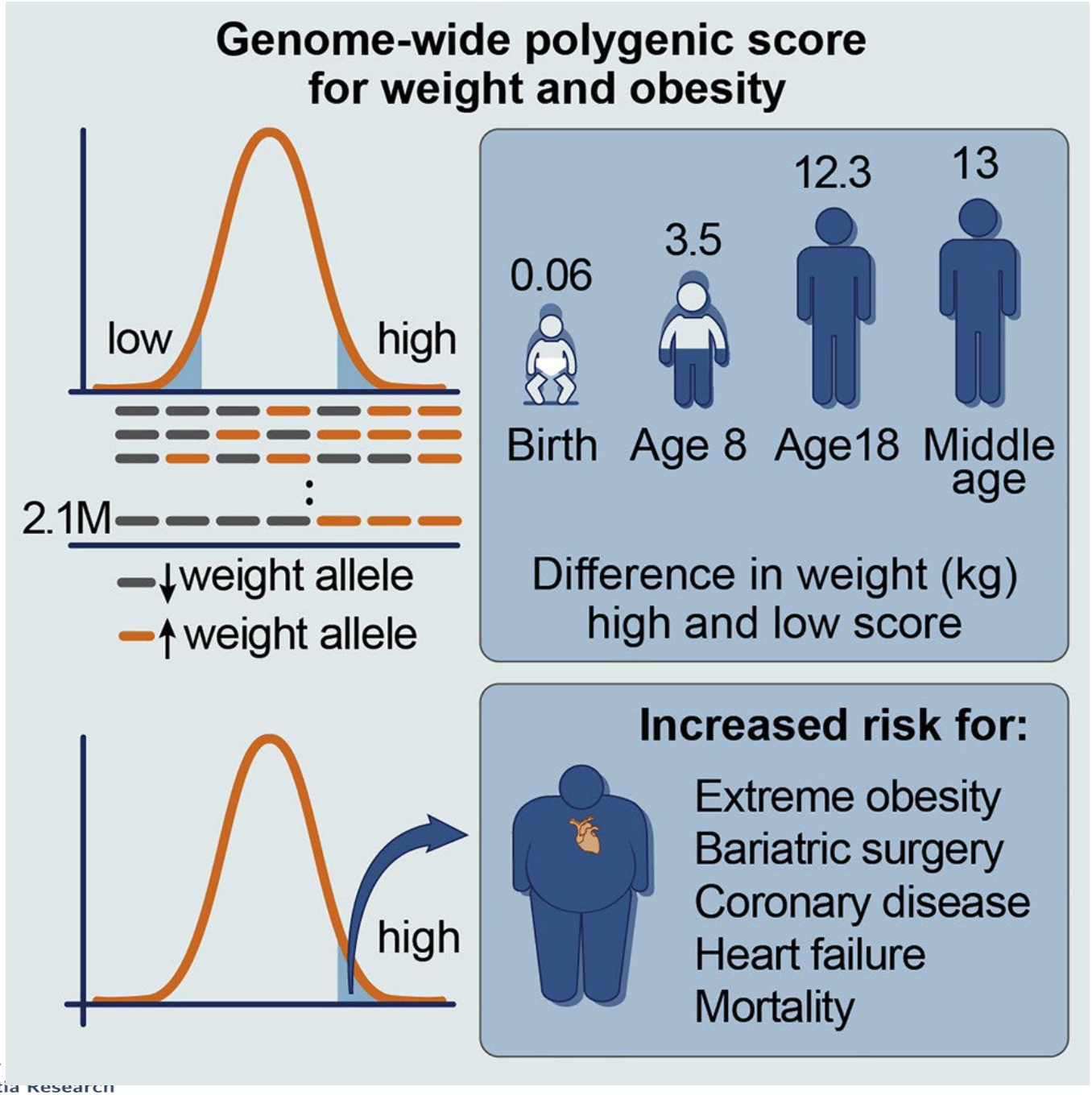

Prediction accuracy
(relative to Europeans)
1.00
0.75
0.50
0.25
0.00
European
American
South Asian
East Asian
African
Population

Clinical implementations

CAD		
Risk factors	Mendelian risk factors	FH mutations: *LDLR, APOB, PCSK9*
	Other factors	Age, sex and family history Systolic blood pressure, LDL or non-HDL cholesterol and BMI *Lifestyle*: smoking, diet and physical activity
Potential clinical utility for PRS		• Adds accuracy to clinical risk predictors (e.g. Framingham risk score, ACC/AHA13 (16)) • Useful for defining most benefit from statin prescription (17,18) • Useful for estimating lifetime risk trajectories (27,56)

Obesity & BMI

Risk factors	Mendelian risk factors	*MC4R* mutations
	Other factors	Age, sex and family history *Lifestyle*: diet and physical activity
Potential clinical utility for PRS		• Targeting lifestyle interventions and potential treatments (e.g. bariatric surgery) to those at most risk of developing obesity • BMI PRS is enriched in those who have undergone bariatric surgery in UK Biobank (32) • Predicting weight gain trajectories (32,90,91) • Useful as a risk predictor of other diseases where obesity is a causal risk factor (79)



Institute for
and Dementia Research

Khera et al., Cell, 2019

What does the future look like?

CANCER POLYGENIC RISK SCORE

Publication number: 20190345566
Abstract: The present disclosure relates to a method of determining a risk of developing breast cancer in a subject, the method comprising identifying whether at least 95 single nucleotide polymorphisms (SNPs) from Table A is present in a biological sample from the subject, wherein the presence of a risk allele of a SNP from Table A indicates that the subject has an increased risk of breast cancer, and wherein the presence of an alternative allele indicates that the subject has a decreased risk of breast cancer.
Type: Application
Filed: July 12, 2019
Publication date: November 14, 2019
Inventors: Amit V. KHERA, Derek KLARIN, Sekar KATHIRESAN

ATRIAL FIBRILLATION POLYGENIC RISK SCORE

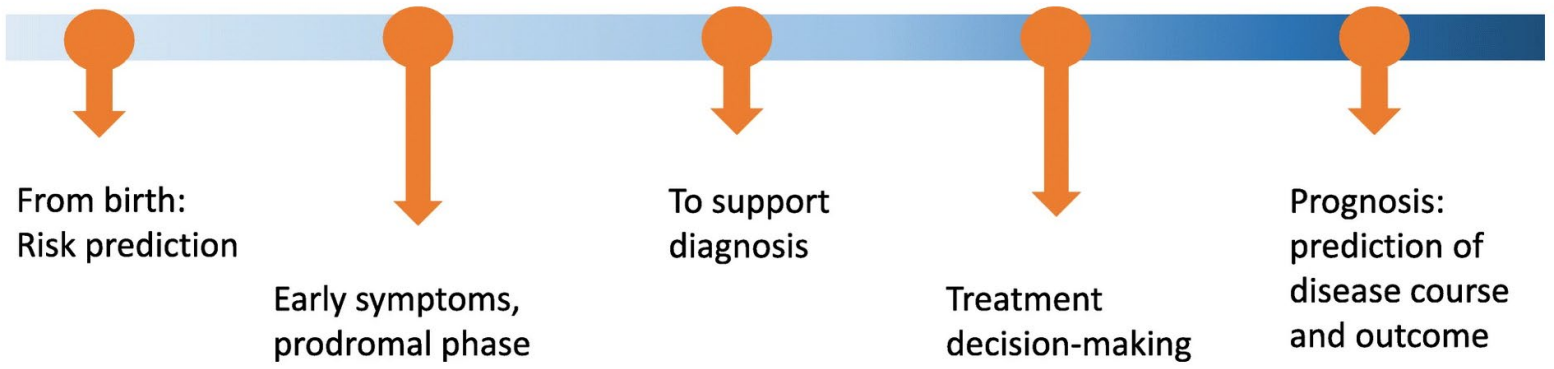
Publication number: 20190345557
Abstract: The present disclosure relates to a method of determining a risk of developing atrial fibrillation in a subject, the method comprising identifying whether at least 95 single nucleotide polymorphisms (SNPs) from Table A is present in a biological sample from the subject, wherein the presence of a risk allele of a SNP from Table A indicates that the subject has an increased risk of atrial fibrillation, and wherein the presence of an alternative allele indicates that the subject has a decreased risk of atrial fibrillation.
Type: Application
Filed: July 12, 2019
Publication date: November 14, 2019
Inventors: AMIT V. KHERA, DEREK KLARIN, SEKAR KATHIRESAN

INFLAMMATORY BOWEL DISEASE POLYGENIC RISK SCORE

Publication number: 20190341125
Abstract: The present disclosure relates to a method of determining a risk of developing inflammatory bowel disease in a subject, the method comprising identifying whether at least 50 single nucleotide polymorphisms (SNPs) from Table A is present in a biological sample from the subject, wherein the presence of a risk allele of a SNP from Table A indicates that the subject has an increased risk of inflammatory bowel disease, and wherein the presence of an alternative allele indicates that the subject has a decreased risk of inflammatory bowel disease.
Type: Application
Filed: July 12, 2019
Publication date: November 7, 2019
Inventors: Amit V. KHERA, Derek KLARIN, Sekar KATHIRESAN

Pros

- „First risk factor“
- Identify population at high risk from birth – targeted intervention (statins etc.)
- High polygenic risk better predictor than most „conventional risk factors“
- PRS additional risk factor in constructing clinical scores etc.
- Effect sizes similar to monogenic mutations
- Genotyping of SNPs is cheap(ish)



Lewis et al., Genome Biol, 2020

Cons

- Small improvement in general prediction of events (AUC, c-index)
  - Not so applicable for general population
- Not (completely) transferrable to other ethnicites
- „just another risk factor“
- Based on „common SNPs“ (MAF>1%)
- Large GWAS studies needed to derive PRS (nowadays not a huge problem)
- Methodological: Overestimation of effect sizes (winner‘s curse) when using only a handful of SNPs
- Not independent of other „classical“ risk factors

Summary

- Genotyping is useful to identify very high risk population – this population will be very small (2%-5%)
- For these individuals, PRS is as detrimental as monogenic mutation
- For others, not very useful, hardly can discriminate
- Also: Not useful for non-Europeans or even mixed ancestry
- Primary prevention: Identify high risk individuals – e.g. start treatment at borderline LDL levels
- Secondary prevention: Individuals with high PRS might benefit guiding the intensity of preventive therapies against recurrent events.

Acknowledgements

ISD
Martin Dichgans
Marios Georgakis

Institute for Stroke and Dementia Research (ISD)