Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations

Amir V. Khera et al., Nature Genetics (2019)

Lino Ferreira
ORC Journal Club | 14 February 2020

If we could identify people at high risk of developing a disease, we could screen them and encourage prevention.

Since common disease usually have an important genetic component, DNA sequencing can (hopefully!) help us.

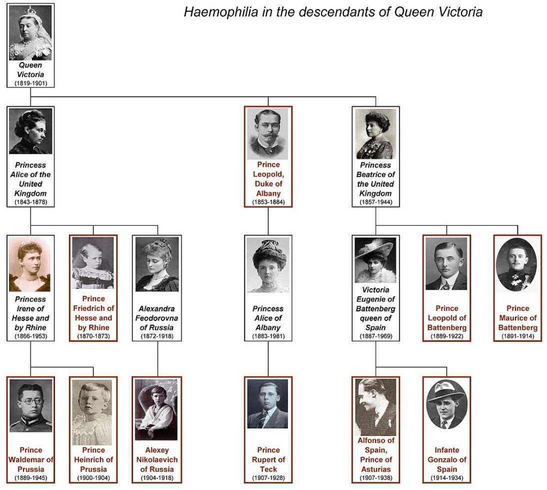
Disease genetics used to be easy...

Figure: Queen Victoria (1819–1901)

Haemophilia is a rare disease caused by a single mutation leading to low levels of clotting factors.

Photo: Alexander Bassano [Public domain]

Disease genetics used to be easy...



Haemophilia in the descendants of Queen Victoria

Queen Victoria (1819-1901)
Princess Alice of the United Kingdom (1843-1878)
Prince Leopold, Duke of Albany (1853-1884)
Princess Beatrice of the United Kingdom (1857-1944)
Princess Irene of Hesse and by Rhine (1866-1953)
Prince Friedrich of Hesse and by Rhine (1870-1873)
Alexandra Feodorovna of Russia (1872-1918)
Princess Alice of Albany (1883-1981)
Victoria Eugenie of Battenberg queen of Spain (1887-1969)
Prince Leopold of Battenberg (1889-1922)
Prince Maurice of Battenberg (1891-1914)
Prince Waldemar of Prussia (1889-1945)
Prince Heinrich of Prussia (1900-1904)
Alexey Nikolaevich of Russia (1904-1918)
Prince Rupert of Teck (1907-1928)
Alfonso of Spain, Prince of Asturias (1907-1938)
Infante Gonzalo of Spain (1914-1934)

Figure: Haemophilia in Queen Victoria's descendents

Image: Wikipedia / Shakko

Most **common diseases** are far more complex, depending on many genes and interactions with the environment.

Although some rare mutations with high impact have been found, the combined effect of many common genetic variants, each of small effect, is more important.

Even though complex diseases are hard...

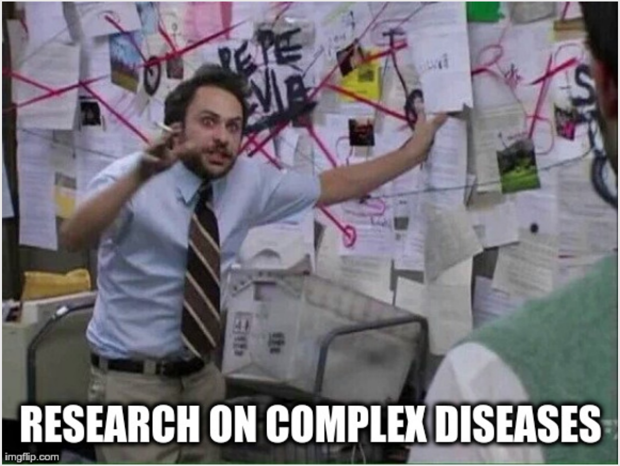


Figure: Meme

...polygenic scores might help

While

- we may not know which associated variants are causal
- the biological pathways may still be unclear

by looking at many SNPs simultaneously we might be able to accurately assess risk.

Advantages of polygenic scores:

- can be calculated for many diseases simultaneously
- can be assessed from time of birth

Genome-wide polygenic scores (GPS)

$$\begin{aligned}\text{GPS} = {} & \text{Weight}_1 \cdot \text{Mutation 1 indicator} + \\ & \text{Weight}_2 \cdot \text{Mutation 2 indicator} + \\ & \cdots + \\ & \text{Weight}_P \cdot \text{Mutation } P \text{ indicator} + \end{aligned}$$

Need to identify relevant variables and compute weights.

But GPSs don’t work very well...

So far, GPSs have had limited success (for example identifying 20% of the population with a 1.4-fold higher risk).

Why?

- small datasets (both for building and testing scores)
- methods not good enough

This paper

Try to improve the performance of GPSs by:

- using the results of **large recent GWAS studies**
- trying a **new method** (LDPred)

This paper

Focus on 5 diseases:

- coronary artery disease (CAD)
- atrial fibrillation
- type 2 diabetes (T2D)
- inflammatory bowel disease (IBD)
- breast cancer

Over **700 000 observations** in total.

Pruning and thresholding method

For a particular disease:

1. sort significant SNPs by p-value

Pruning and thresholding method

For a particular disease:

1. sort significant SNPs by p-value
2. start with the most significant and remove all other SNPs within 250 kb which are correlated with it above a given **threshold** (pruning)

Pruning and thresholding method

For a particular disease:

1. sort significant SNPs by p-value
2. start with the most significant and remove all other SNPs within 250 kb which are correlated with it above a given **threshold** (pruning)
3. keep only SNPs with p-value lower than a **second threshold** (thresholding)

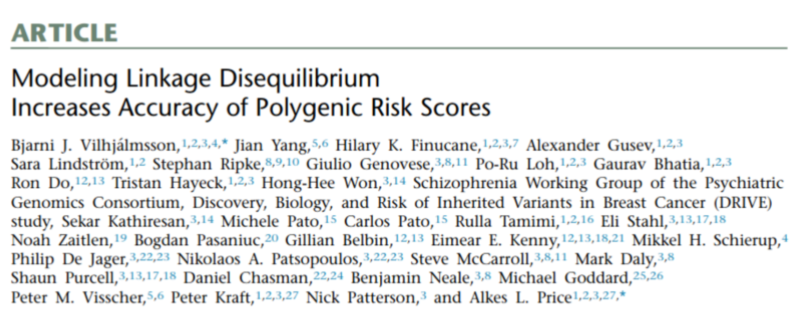
Pruning and thresholding method

For a particular disease:

1. sort significant SNPs by p-value
2. start with the most significant and remove all other SNPs within 250 kb which are correlated with it above a given **threshold** (pruning)
3. keep only SNPs with p-value lower than a **second threshold** (thresholding)
4. iterate across all significant SNPs

Obtain 24 candidate scores for each disease.

LDPred



ARTICLE

Modeling Linkage Disequilibrium Increases Accuracy of Polygenic Risk Scores

Bjarni J. Vilhjálmsson,[1,2,3,4,*] Jian Yang,[5,6] Hilary K. Finucane,[1,2,3,7] Alexander Gusev,[1,2,3] Sara Lindström,[1,2] Stephan Ripke,[8,9,10] Giulio Genovese,[3,8,11] Po-Ru Loh,[1,2,3] Gaurav Bhatia,[1,2,3] Ron Do,[12,13] Tristan Hayeck,[1,2,3] Hong-Hee Won,[3,14] Schizophrenia Working Group of the Psychiatric Genomics Consortium, Discovery, Biology, and Risk of Inherited Variants in Breast Cancer (DRIVE) study, Sekar Kathiresan,[3,14] Michele Pato,[15] Carlos Pato,[15] Rulla Tamimi,[1,2,16] Eli Stahl,[3,13,17,18] Noah Zaitlen,[19] Bogdan Pasaniuc,[20] Gillian Belbin,[12,13] Eimear E. Kenny,[12,13,18,21] Mikkel H. Schierup,[4] Philip De Jager,[3,22,23] Nikolaos A. Patsopoulos,[3,22,23] Steve McCarroll,[3,8,11] Mark Daly,[3,8] Shaun Purcell,[3,13,17,18] Daniel Chasman,[22,24] Benjamin Neale,[3,8] Michael Goddard,[25,26] Peter M. Visscher,[5,6] Peter Kraft,[1,2,3,27] Nick Patterson,[3] and Alkes L. Price[1,2,3,27,*]

Figure: Paper from Alkes Price’s group

LDPred

Motivation:

- P+T doesn’t account for linkage disequilibrium well enough and discards informative markers
- maybe we can do better by modelling the genetic architecture more explicitly

LDPred is Bayesian

Figure: Bayes ♡

Bayesian stats refresher

- All uncertainty is modelled through probability distributions
- Need prior distributions for model parameters
- Bayes' Theorem combines prior and data to give updated distribution of parameters:

$$\pi(\theta|\text{data}) \propto \Pr(\theta) \cdot \Pr(\text{data}|\theta)$$

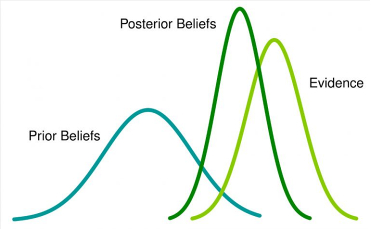


Figure: Updating prior with evidence

Image: Analytics Vidhya / NSS

Advantages of the Bayesian approach:

- can incorporate pre-existing knowledge in an explicit and coherent way
- can make probability statements about parameters

  e.g. *“there’s a 70% chance that this parameter is greater than 0”*

LDPred

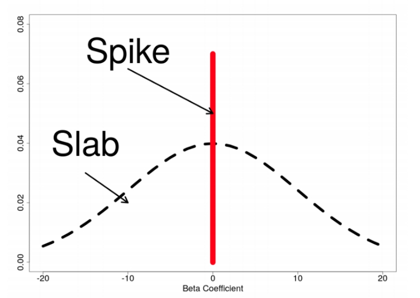


Figure: Spike-and-slab prior

Two parameters:

- fraction of causal markers
- heritability explained by genotypes

Image: Towards Data Science / Zarina Vakhitova

LDPred gives us the expected value (mean) for the effect of each SNP on the phenotype.

We use this to construct 7 GPSs for different values of the parameter for fraction of causal markers.

Our two methods (P+T and LDPred) give us 24 + 7 = 31 candidate scores for each of the 5 diseases.

Use a first dataset (UKBB phase 1) to choose the best-performing model (AUC) for each disease (validation).

Use a second dataset (UKBB phase 2) to get a final assessment of performance for the selected models (testing).

Results

- **CAD:** 8% of population with ≥ 3-fold higher risk
- **Atrial fibrillation:** 6.1% with ≥ 3-fold higher risk
- **T2D:** 3.5% with ≥ 3-fold higher risk
- **IBD:** 3.2% with ≥ 3-fold higher risk
- **Breast cancer:** 1.5% with ≥ 3-fold higher risk, 0.1% with ≥ 5 ×

Results

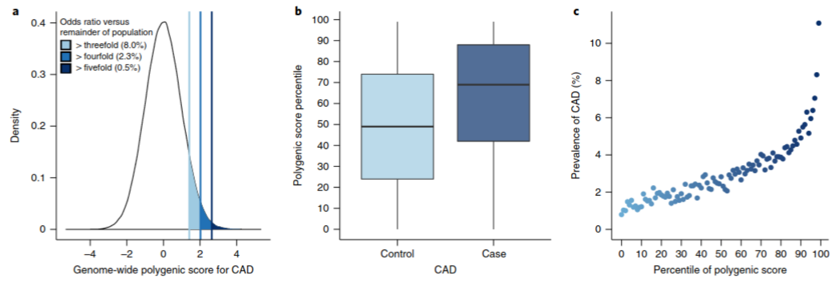


Figure: GPS for CAD

Take away

“Polygenic risk scores can now identify a substantially larger fraction of the population than is found by rare monogenic mutations, at comparable or greater disease risk.”

Big challenge

All this work was done with data from people of primarily European ancestry.

GPSs don’t work nearly as well for other ethnic groups:

- different allele frequencies, LD patterns and SNP effect sizes

Discussion

How can we communicate this type of results to patients?

Is it ever sensible and ethical to withhold information?

Should parents be able to choose whether to have their children tested?

How should we allocate resources between people with different levels of risk?

How should genetic risk stratification be integrated with other (e.g. environmental) risk factors?