Unpacking polygenic scores

PHG
FOUNDATION

making science
work for health

Summary

◆ The term 'polygenic scores' encompasses different genetic indicators

◆ A polygenic score provides a single measure of the cumulative effect of multiple or individually low-impact genetic changes

◆ A polygenic score for an individual is calculated by applying an algorithm to their genotype data

◆ Information from polygenic scoring is not deterministic or highly predictive by itself

◆ Polygenic scores provide additional data that can potentially contribute to decision making in a variety of healthcare contexts

A polygenic score can be used to help determine risk of developing a specific disease. This briefing provides an overview of what polygenic scores are and what they can and cannot tell us about the risk of disease.

Common genetic variants and disease

The likelihood and severity of all diseases are affected by genetic and non-genetic factors and we rarely have complete knowledge of all these factors. Genetic variants that contribute to disease are often grouped as either rare or common variants. They may also be grouped as strongly and weakly deterministic and penetrant, with common variants generally coming within the latter category.

Known rare variants can be considered a small sub-set of the genetic underpinning of a disease. Information from rare variants is already used in the treatment and management of people with cancer or rare diseases.

Common genetic variants are defined as those that have a frequency of greater than 1% in the population. The most well studied of these are single nucleotide polymorphisms or SNPs. These are changes to a single base in the DNA molecule, for example, cytocytosine (C) instead of thymine (T) in the genetic code.

How is a polygenic score calculated?

Through genome-wide association studies research has identified many SNPs linked with disease. However, each individually only has a small effect. To understand their impact on disease risk, they need to be examined collectively.

Using data from genome-wide association studies, polygenic score models are created that specify a set of SNPs and their weights. These statistical models are used to create algorithms that can be used in clinical practice to calculate a risk score.[1] Many models have been created aimed at different diseases and outcomes.

A polygenic score for an individual is calculated by applying a validated algorithm to their genotype data. Algorithms have been developed that calculate a score on the basis of as few as 10 SNPs to over a million SNPs.

Characteristics of polygenic scores

Polygenic scores can be considered as an indicator of genetic liability to a disease, as they are a calculation based on analysis of different SNPs across the genome using a predictive algorithm. Research has shown that polygenic scores provide an estimate of disease risk in a given population, but have more uncertainty when applied to an individual.[2]

While polygenic scores provide an estimate of genetic contribution to the risk of disease, this only captures a proportion of overall genetic risk, i.e. that attributed to identified common genetic variants. For many diseases, risk will also be modulated by other genetic and non-genetic factors, especially in the case of common complex diseases (e.g. diabetes).[3]

In contrast to the identification of rare pathogenic genetic variants, information from polygenic scores is neither deterministic nor highly predictive, especially by itself. In addition, the inheritance of these variants does not follow predictable patterns. Taken together this means that implications for family members are not straightforward and need to be carefully interpreted.

Whilst all polygenic scores share common features they can also differ from each other. The number of genetic variants, the magnitude of their effects, as well as their frequency in the population and interaction between them and the environment is called the genetic architecture.

Genetic architecture varies across diseases and population groups. Interpreting information from a polygenic score is therefore both condition specific and population specific. Different models and algorithms are created to predict specific conditions, and may be applicable to different sub-groups of the population.

Therefore, polygenic scores can be considered as an umbrella term for different genetic indicators. Some will be more informative for certain diseases and contexts than others.

Considerations when using polygenic score information

Polygenic scores provide an estimate of risk or probability of developing a condition. For any given disease they are "normally" distributed in a population.

For an individual, it may be informative to know their polygenic score and where they lie in the spread of risk for a disease. However the score must be interpreted with caution. The spread and mean of this distribution is likely to vary between different population sub-groups - for example a population of people over 50 or women with a family history of breast cancer - just as it does for biomarkers such as cholesterol, height, weight and blood pressure.

Figure 1: Example distribution of polygenic scores across a population
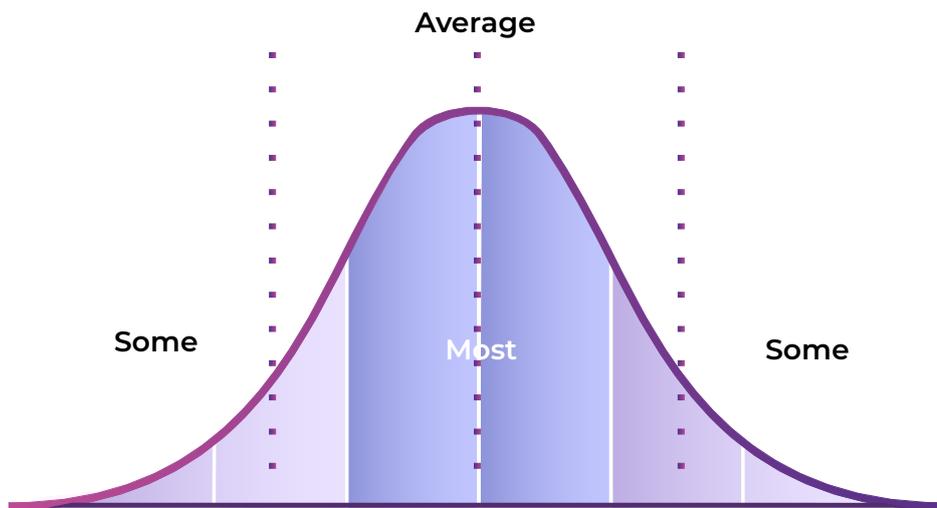
The scores provide an estimate of risk, which means:

◆ There will be cases of the condition among those classified low risk
◆ There will be people in high-risk populations who do not develop the condition
◆ The highest number of cases will arise in the average risk population (because that is where the largest number of people are)

The risk of developing a disease will also vary between individuals and changes over time as other factors such as age and exposures throughout life will also affect risk. This means that polygenic scores can be used to identify sub-populations that might be at increased risk of an outcome (such as developing a specific disease), but there will be uncertainty as to the exact individuals within that population that may develop that particular outcome.

As polygenic scores are a calculated measure, any errors or uncertainties in the datasets used in developing the algorithm that calculated them will carry through to the final score. These factors impact on interpretation of information from analysis at the individual level and across different population groups (such as different ethnicities, ages, geographies). Similar issues arise for the use of most algorithms in healthcare.

Making the most of polygenic scores

Even though polygenic scores are not a perfect measure of genetic risk, they can potentially provide additional information, along with other details about an individual, to refine risk assessment and guide clinical decisions.[4]

Their utility as part of healthcare pathways will be variable, as they are likely to be more useful and informative for certain diseases and contexts than others.

The evidence base for the use of polygenic scores within specified care pathways requires development, including better understanding of how to define and evaluate products that generate and use a polygenic score.

References

1. Babb de Villiers, C, Kroese, M, Moorthie, S. Understanding polygenic models, their development and the potential application of polygenic scores in healthcare. J Med Genet. 2020. 57(11): pp. 725-732.

2. Ding, Y, Hou, K, Burch, K S, et al. Large uncertainty in individual polygenic risk score estimation impacts PRS-based risk stratification. Nat Genet. 2022. 54(1): pp. 30-39.

3. Wray, N R, Lin, T, Austin, J, et al. From Basic Science to Clinical Application of Polygenic Risk Scores: A Primer. JAMA Psychiatry. 2021. 78(1): pp. 101-109

4. Lewis, C.M., Vassos, E. Polygenic risk scores: from research tools to clinical instruments. Genome Med 12, 44 (2020). https://doi.org/10.1186/s13073-020-00742-5

For more quick guides to polygenic scores and their implementation, go to phgfoundation.org

Author: Dr Sowmiya Moorthie Published: October 2023

UNIVERSITY OF CAMBRIDGE