PHG FOUNDATION

making science work for health

Polygenic score analysis: the test pipeline

Summary

- It can be useful to think of the process of obtaining or using a polygenic score as a pipeline
- The ways in which such pipelines are configured and function as part of clinical pathways can vary
- Achieving clarity as to which components contribute to a particular application is important in considering the necessary validation of individual components as well as that of the pipeline as a whole

In this briefing we provide an overview of the steps involved in generating a polygenic or integrated risk score and how they can come together and be thought of as a test pipeline.

The polygenic score test pipeline

Generating a polygenic score or an integrated score for an individual involves a series of steps. This includes testing to obtain genotype data or other data to inform integrated risk assessment, application of a prediction algorithm to the data and interpretation to obtain a score. There can be differences in the way each of these steps is conducted and how they are bought together.

- **Polygenic score:** output of a polygenic score model
- **Integrated risk score:** output of a multifactorial risk prediction model. This may include a polygenic score

A key factor impacting on the assessment and implementation of tests that provide or incorporate a polygenic score is failure to clearly describe and evaluate all their key components.

Key components of a pipeline

◆ Genotyping

Genotype data to feed into polygenic score analysis can be obtained through a variety of methods such as genotyping SNP panels, microarray, or next generation sequencing. This data could be generated 'in-house' in a clinical laboratory or may be provided from an external source. External sources of data could come from when individuals have had their DNA analysed as part of a commercial test or research project.

From a clinical or commercial laboratory perspective, it is usual practice to apply quality control steps to the data prior to the PGS analysis to ensure the information is appropriate and of sufficient quality, but these can vary.

Variations in the different elements of a test pipeline

	Genotyping	Polygenic score analysis	Integrated risk analysis	Digital interface
Variations	• Sequencing technology • Quality control measures • Imputation techniques	• Algorithm to calculate PGS • Methods to adjust for population stratification • Risk categories • Result that is reported	• Risk factors included • Algorithm for integrated risk score • Risk categories	• Measure reported: integrated, relative, or absolute risk? • Use and features of digital interface
Each determined by	Population	Purpose		Role

◆ Polygenic score analysis

Analysis of genotype data requires the application of a validated polygenic score model to genotype data to enable the calculation of a score. Different models are available for the same disease, and they can differ in the SNPs that are included and their weights.

The raw results of polygenic score analysis need to be converted to a risk score and there are several different outputs that can potentially be fed back. The reporting of absolute risk has been recommended as it is more interpretable and understandable than relative risk.

This requires additional steps, including consideration of the distribution of polygenic scores and disease incidence in the relevant population.

Polygenic scores can be generated and interpreted on their own or with other information as part of integrated risk assessment. The latter is more likely, especially for common complex diseases, where a variety of factors influence risk. Integrated risk prediction requires the further step of incorporating the polygenic score into an integrated risk prediction algorithm.

◆ Integrated risk assessment

New algorithms can be developed that calculate risk based on a polygenic score and other variables. Alternatively, as mentioned above, existing risk prediction algorithms that combine information on a variety of risk factors can be adapted to incorporate a polygenic score.

Risk factors included in integrated risk models and the way they are constructed, can vary. Other tests are usually needed to obtain information on these additional variables, such as a cholesterol test, a blood pressure test and/or a family history assessment.

◆ Additional software components

Additional software components may be included as part of pipelines to improve functionality and generalisability. These may include components that enable the use of prediction algorithms with different sources of genotype input data. This is to allow greater flexibility in the source genotype data that is used as part of the analysis pipeline.

A well-recognised shortcoming of many existing polygenic score models is their lower predictive performance in people not of European genetic ancestry. Mechanisms can be put in place to overcome this to some extent. Options being considered are either restricting use of models to specific groups or attempting to optimise a model to function across groups using statistical techniques. Software components may be needed to enable adjustments such as those that enable assignment of individuals to genetic groups.

◆ Presence of a digital interface

Prediction models can be packaged in the form of a digital tool to make them more user friendly. These tools can be considered the mechanism by which end-users interact and use information from a model. For example, they may enable easier inputting of model variables and visualisation of outputs, to allow interpretation by clinicians and patients. They can be designed to be incorporated into healthcare IT systems or they may be stand-alone.

◆ Differences in delivery of the pipeline

Currently, different approaches have been developed or are being investigated in bringing these steps together. This means that the test pipeline that is being proposed for implementation can vary. There is variability in the scope of testing (which populations, purpose, and role), mechanisms to access testing and the services provided. Companies such as Genomics PLC and Allelica (amongst others) require a health professional to order the test. Access to the CanRisk Tool, which has been developed in an academic setting is also restricted to health professionals.

The services provided in relation to the polygenic score analysis pathway can also differ. This may be a more comprehensive (end-to-end) service, including genotyping of a sample, use of algorithms developed in-house to calculate a polygenic score and report an integrated risk assessment to the health professional. Alternatively freely available polygenic score tools developed in a research setting can be used as part of a test delivery. In such cases, genotyping can be conducted separately, either by using a commercial service or an NHS laboratory and the results imported into the prediction algorithm.

Why is this a problem?

As described above products that provide or incorporate a polygenic score can be configured in different ways. This has created uncertainty in which components to evaluate, whether they should be looked at individually or together, and the way in which this should be done. This is creating a challenge for evidence generation and evaluation in support of their use, which has implications for both regulatory approval and implementation.

The pipeline as a solution

Considering the process of producing a polygenic score or an integrated score as a pipeline that can be configured in different ways can help addresses current complexities. For example:

- Allowing developers to better describe the product for implementation, which may be part of this pipeline or its entirety.
- Allowing a modular approach to evidence generation and assessment. Each component of the pipeline can be evaluated separately, which can be useful when they are shared across different applications. It also allows the examination of issues related to each component that are relevant for any application of polygenic score analysis.
- Linking evidence of the components to better understand the clinical validity and utility of particular applications. This also provides clarity on the relationship between the different components that inform polygenic score analysis (whether standalone or integrated).

Conclusion

Products that provide or incorporate a polygenic score bring together elements of molecular testing, prediction algorithms and digital tools. This is creating challenges in evidence generation, evaluation, and assessment. Considering these products by their component parts and proposed use together will aid evaluation and assessment.

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

Author: Dr Sowmiya Moorthie Published: October 2023

UNIVERSITY OF CAMBRIDGE