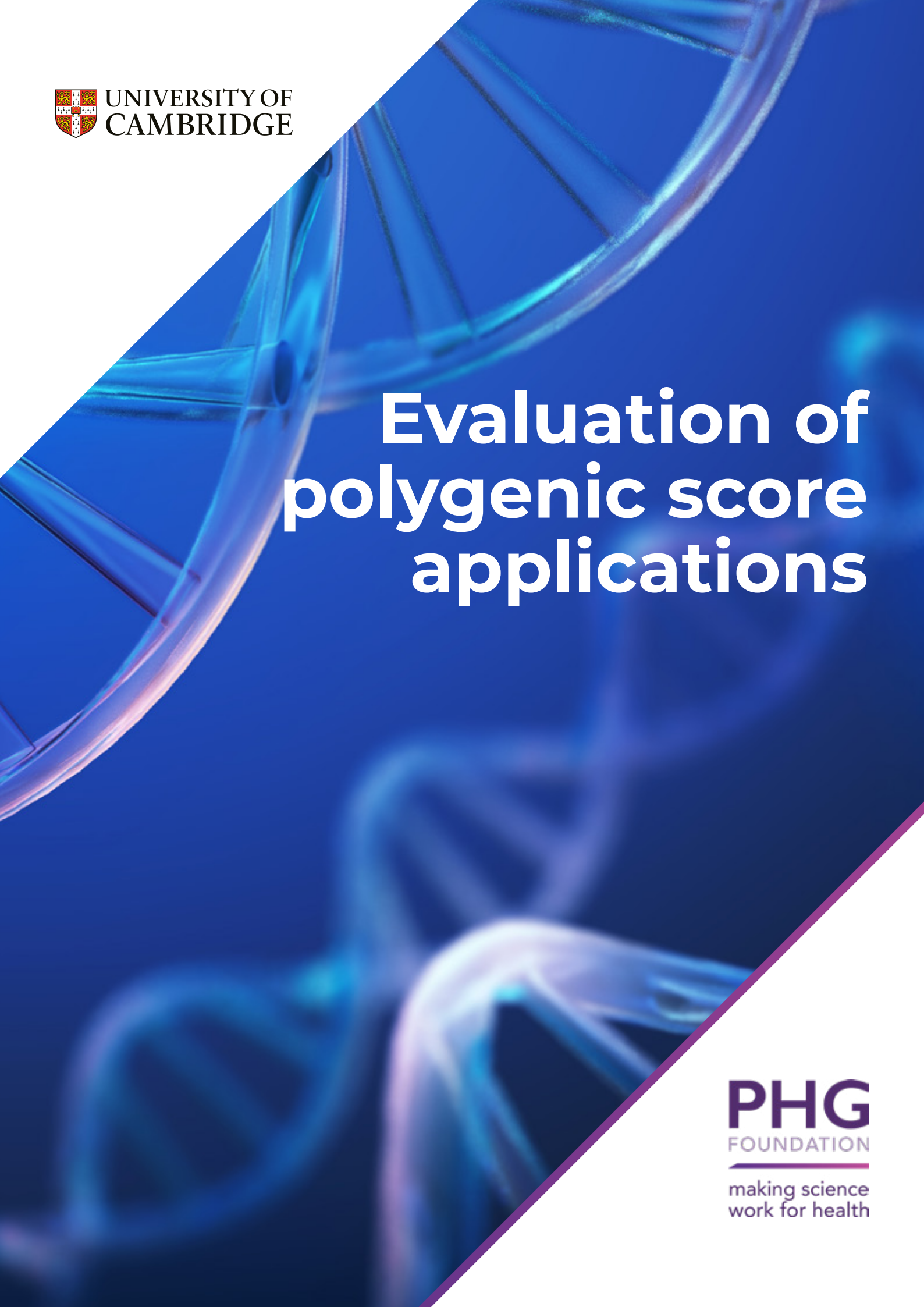


UNIVERSITY OF CAMBRIDGE

Evaluation of polygenic score applications

PHG FOUNDATION

making science work for health

Authors

Sowmiya Moorthie, Joanna Janus, Heather Turner, Chantal Babb de Villiers, Colin Mitchell and Mark Kroese

Acknowledgements

The PHG Foundation is grateful for the insight provided by individuals consulted during the course of this project.

We also thank Mary Ann Binuya, Prof Montserrat Garcia-Closas and Prof Paul Pharoah for reviewing and providing feedback on report drafts.

We are grateful for the PHG project team (Rebecca Bazeley, Ofori Canacoo, Dr Philippa Brice and Dr Laura Blackburn) for their contributions to this report

URLs in this report were correct as of June 2023

Written and produced by PHG Foundation

2 Worts Causeway, Cambridge, CB1 8RN, UK +44 (0)1223 761900

This report can be downloaded from

www.phgfoundation.org

Executive summary

Considerable progress has been made in uncovering common single nucleotide variants and developing mechanisms for genomic profiling. Using this knowledge routinely as part of clinical and public health practice is an ongoing aspiration. While products that enable conversion of genomic data into genome-based risk scores such as polygenic scores (PGS) are available, they are not widely used. Key barriers are uncertainty and a lack of evidence regarding the value of polygenic score information and how to approach evidence gathering and appraisal.

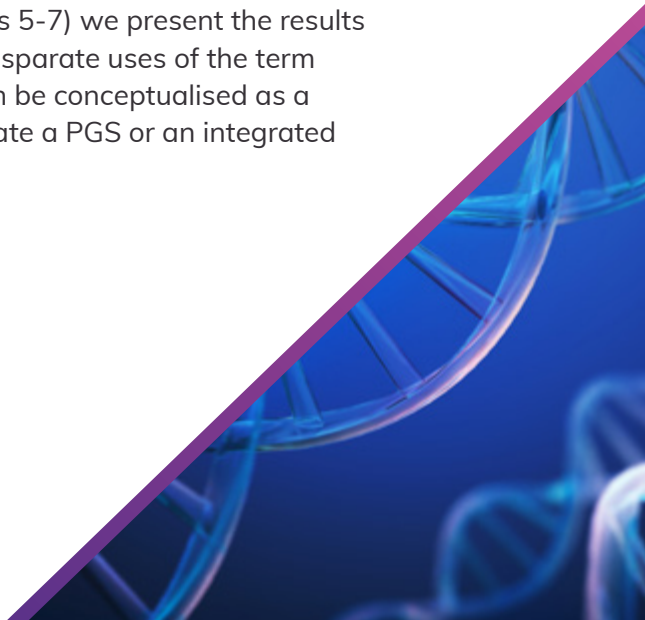
In this report, we discuss and present our analysis of the application of the principles of medical test evaluation to PGS based products. Medical test evaluation frameworks such as the ACCE framework can be used in evidence assessment and contribute to more informed and transparent decision making. Central to this process is consideration of context of use and application of an iterative process to examine evidence across domains of scientific, analytical, and clinical validity and utility. Evaluation of any PGS application will require evidence for and consideration of these different domains.

We demonstrate how specific factors drive these uncertainties about products that provide or incorporate a PGS. These include:

- Conflation of terminology relating to polygenic scores, models and algorithms.
- Inadequate description of specific applications, in relation to intended population, role and purpose as part of specific healthcare pathways.
- Failure to define and evaluate all the key elements of PGS applications.
- Lack of real-world evidence (RWE) for PGS applications.

Failure to adequately address these factors lead to a challenge for decision makers because, the existing evidence base (a) fails to show what information polygenic scores are providing (b) does not define with adequate precision how the product is to be used in health care or its intended purpose or objective or (c) how such use can be beneficial to the individual patient or to the health system as a whole. The consequence is that we are left with a body of evidence that is inadequate for the determination of the clinical validity or utility of a product in relation to its intended purpose. We have shown that it is possible to resolve these issues and address the needs of decision-makers through a more systematic approach to evidence generation.

Chapters 2-4 provide background information that may be useful and informative for those in different fields. This is to enable readers to develop a shared understanding of relevant topics. In the latter half of the report (Chapters 5-7) we present the results of our analysis which begins by providing clarity to the disparate uses of the term 'polygenic score 'and examines how polygenic scores can be conceptualised as a biomarker. We also describe processes that either calculate a PGS or an integrated



Executive summary

score including PGS and how they can be considered a test. We then demonstrate how existing evaluation frameworks can be applied to such products.

Polygenic scores are one way of assessing an individual's genetic risk to developing certain outcomes, including diseases. They are a proxy biomarker, calculated on the basis of an individual's genotype. Polygenic scores differ from traditional germline genetic markers in a variety of ways. Importantly, calculation of a score requires the use of algorithms which are developed from polygenic score models. Different models and algorithms must be created to predict different diseases and traits. In addition, different approaches may be taken to develop them, suited to each disease and or population of interest. This means that there is a variety of models and algorithms used to calculate a score. There are also differences in the way this information can be used and contribute to clinical practice. It is likely that in some contexts PGS will provide valuable clinical information and it is important to identify where this may be the case.

Currently there is ambiguity regarding how to apply regulations and carry out evaluation in support of products that provide a PGS. This is because there are different components that form the test pathway, namely, molecular testing to obtain genetic data; other clinical data; prediction algorithm(s) for analysis of this data; and digital tool(s) to enable data collation and feedback. This creates uncertainty in the nature, quality and quantity of evidence required for decision making across these components and in relation to a specific test strategy.

We have shown that existing frameworks can be applied to a specific product and application but require consideration of all the component parts. We propose considering these components as part of a test pipeline to allow the application of concepts and techniques from molecular test evaluation, prediction modelling and digital technology evaluation to each separate component of the pipeline. We demonstrate how analytical, scientific and clinical validity parameters can be assessed and some of the issues in determining these. Examination of these parameters across a PGS analysis pipeline can provide evidence of the performance of a PGS-based test. These can then inform the assessment on whether they meet the test's intended role and purpose.

As stated in our previous reports, clarity regarding the proposed PGS application for implementation helps to determine the evidence requirements, as well as the assessment of wider factors that may impact on its use and uptake. In this report we outline how better definition and descriptions of products can help achieve this and where issues in evidence generation currently lie.

Going forward, it will be important to achieve consensus amongst researchers, developers, health system decision makers and users as to which of the gaps which we

Executive summary

have identified are critical and how they can be addressed. Progress on establishing both the evidence required for the different components of the PGS test pipeline as well as the acceptable levels of evidence will be necessary for the successful clinical implementation and wider uptake and use of any PGS-based applications.

In conclusion, polygenic scores are likely to be useful under certain circumstances. Identifying these and creating optimal systems for their use requires a more focussed approach to evidence generation and appraisal which is currently lacking.

Contents

Introduction

1. Introduction

1.1 Project rationale

PHG Foundation previously completed a report on polygenic scores (PGS) and clinical utility [1], which provided a comprehensive background summary and analysis to support robust decision making around implementation of polygenic scores into health services. We worked with stakeholders to bring together key concepts around clinical utility and used our expertise in genetic test evaluation and regulation to consider how to assess the potential utility of polygenic score analysis most effectively in healthcare.

A key finding of the report was that a systematic approach to evidence generation and assessment is needed to better understand the value of PGS and products that calculate a PGS. However, several inter-related factors are hindering this process.

- There is a lack of detail as to the nature of such products and their intended purpose. This is needed to enable a clear assessment of the information they provide, taking into consideration the context of their use.
- There is a lack of clarity as to whether and how to apply existing evaluation frameworks to such products. This includes uncertainties regarding which elements of a process to produce a PGS or subsequent integrated risk score to evaluate.
- There is a lack of consensus on the type, quantity and quality of evidence required to support implementation.

These issues also serve as a challenge for the appropriate regulatory oversight of products that provide or incorporate a PGS and consideration of factors that will impact on their implementation within healthcare.

1.2 Methodology

Our approach in carrying out this project was to supplement our in-house expertise on the topics of genetic test evaluation and regulation, with literature review on the topics of prediction model evaluation and digital technology evaluation. We also engaged with external experts, in particular, researchers in the field of polygenic score model development and risk prediction, regulation and healthcare evaluation to inform this project. This allowed us to further our understanding of key topics and examine them from different perspectives. In addition, it allowed us to carry out analysis that brings together and utilises approaches from different fields.

1.3 Report structure

The first half of this report (Chapters 2-4) provides background information that may be useful and informative for those in different fields. This includes a broad overview of medical tests and their evaluation, followed by an introduction to genetic test evaluation and prediction models evaluation. This background information is provided to enable a shared understanding of concepts and frameworks that have informed our analysis. Those familiar with these topics may wish to move to the second half of the report.

The second half (Chapters 5-7) presents the results of our analysis which begins by providing clarity to the disparate uses of the term 'polygenic score' and examines how polygenic scores can be conceptualised as a biomarker. We also describe processes that either calculate a PGS or integrated score and consider how they can be conceptualised as a test. We then describe challenges in the regulation of such products. Finally, we consider the application of existing frameworks for evidence generation to such products, and outline the evidence base that may be considered in the assessment process.

Medical tests and their evaluation

2. Medical tests and their evaluation

In this chapter, we cover key background information relating to the broad term 'test', their role as part of clinical pathways and how tests are conceptualised for the purposes of evaluation. We also provide a brief overview of the development and evaluation of laboratory-based diagnostics and prediction models. These topics are covered to provide a summary of key concepts and terminology relating to medical tests and how they impact and influence the evaluation process.

Key points:

- The term 'test' is a broad term referring to different investigations that can help inform care of an individual.
- In the context of evaluation, the term 'test' has a more specific definition and can be further distinguished from technologies used to deliver testing (see Section 2.4).
- The terms 'prediction model', 'tool' and 'risk model' are often used interchangeably to refer to a broad range of products that examine multiple biomarkers and risk factors to provide an estimate of risk for particular outcomes.
- The extent to which prediction models are considered as a test for evaluation purposes can differ, leading to variability in the degree of scrutiny they undergo.
- How a test or testing strategy is defined will have an impact on how it is evaluated, the evidence requirements and specific considerations for their implementation.

2.1 The varied role of medical tests as part of healthcare pathways

The umbrella term medical tests encompasses a broad range of different investigations that inform care of an individual. In clinical settings, these investigations help to inform assessment regarding the probability of a particular outcome (the diagnostic process), enabling health professionals to rule in or rule out specific hypotheses about possible outcomes, including future risk of disease [2].

While this is referred to as the diagnostic process, it is not solely about determining the presence or absence of a condition. This process is aided by collating all available information relevant to a case. In addition to clinical features, this includes collection of demographic characteristics, symptoms and signs, physical examination and/or undertaking laboratory or imaging-based investigations to gain insights on particular biomarkers. Therefore, the term 'test' can be applied to a wide variety of processes, ranging from questionnaires to laboratory analyses that are used in information gathering to inform care and management of individuals.

Table 1 is an illustration of the commonly stated purpose for different categories of tests and their potential utility. As illustrated by the table, testing may be utilised at different time-points of a disease trajectory and inform healthcare decision making in a variety of ways. The majority of tests do not provide a definitive or conclusive answer but provide information that is often interpreted based on the context of use. Thus, context of use is important in deciding whether to utilise a particular test and in the interpretation of its results.

Medical tests usually undergo some form of evidence evaluation and assessment. However, the depth and format of the assessment will depend on the nature and purpose of particular tests. For example, questionnaires may be evaluated and assessed using different procedures in comparison to laboratory-based diagnostics.

In the UK, organisations that carry out evidence assessment include guidance developers (such as the National Institute for Health and Care Excellence, NICE), regulatory agencies (such as the Medicines and Healthcare products Regulatory Agency, MHRA), payers (e.g. individuals, laboratories, hospitals, departments of health), clinicians, patient interest groups, advisory committees (such as the National Screening Committee, NSC), amongst many others. The extent of the assessment will depend on the purpose of the assessment, the characteristics of the test, its intended use and the user group.

Table 1: Broad test categories and their potential utility.

Adapted from *The Essentials of Diagnostics series: Molecular Diagnostics, AdvaMedDx and DxInsights*, 2013 [3].

Broad test category	Potential uses	What they do	Potential utility
Screening	Disease risk Prediction	Evaluate likelihood of developing a particular condition	Could lead to lifestyle changes or treatment to minimise risk
	Early detection	Identify disease at an early stage	Reduce impact of disease or prevent progression if amenable to a treatment
Informing diagnosis and prognosis	Confirmatory diagnosis	Confirm or rule out specific diagnoses	Determine next steps in care
	Staging and prognosis	Determine severity of condition or predicted outcome	Determine treatment decisions
Management	Therapy selection	Predict effectiveness or potential side effects of treatments	Avoid unnecessary treatment
	Monitoring/ treatment assessment	Assess ongoing safety and effectiveness of treatments	Enables timely intervention to adjust or change treatment when necessary

There are a variety of mechanisms and resources available to aid in these assessments including basic checklists [4, 5], frameworks [6-10], reporting guidelines [11], standards [12, 13], and protocols [14]. In addition, specific groups, such as NICE [15] or the Clinical Laboratory Standards Institute (CLSI) [16], may develop their own methodologies. A key part of this process is clearly describing the test or test strategy that is to be evaluated and developing an understanding of the care pathway into which the test will be integrated.

2.2 Development and evaluation of diagnostics

Biomarker discovery and the development of technological platforms to analyse these biomarkers, are contributing to the availability of a wide variety of medical tests. The term 'diagnostics' is sometimes used as a broad term to refer to this field and to differentiate it from therapeutics development.

Just like the drug or therapeutic development process, diagnostics development can be a lengthy process [17]. It requires basic research for the development of technology platforms (e.g. sequencing, mass spectrometry, imaging) as well as biomarker discovery (e.g. genes, proteins), followed by translation of those biomarkers and discovery technologies into a tool or product. This additional development is needed to ensure that the biomarker and technology platforms meet user needs.

Diagnostics development can take a variety of forms. Existing technologies may be used to assess a novel biomarker, as is often the case with genomics or to evaluate a known biomarker using a new method (e.g. analysis of free foetal DNA). Novel techniques may be developed to measure a novel biomarker (e.g. polygenic score analysis) or an established biomarker (e.g. long read sequencing methods for genomic analysis).

Unlike therapy development, diagnostics development has no clearly established pathways and processes for evidence generation and integration into health systems [18]. This is in part due to the diversity and complexity of diagnostics, their users and providers, but also because development of diagnostics can be undertaken in a variety of sectors such as academia, commercial companies and health system laboratories. This means that a range of approaches may be used for the assessment of particular diagnostics and their integration into health systems depending on the intended use and user.

The potential introduction of novel diagnostics into healthcare pathways, as with uptake of therapeutics, requires due consideration of usefulness, benefits and risks. These are often assessed through technology evaluation processes and by ensuring compliance with regulatory frameworks.

There is variability in the degree to which diagnostics are scrutinised and the processes that are utilised and applied can vary. Nevertheless, the parameters that are assessed are broadly similar and include analytical validation of the technology, validation of the relationship between the biomarker and disease states (scientific validity) and further assessment of the use of the diagnostic in a specific context to determine test performance and outcomes (clinical validity and utility). These are discussed more fully in the next chapters of this report, along with how they may apply to polygenic score analysis.

2.3 Development and evaluation of clinical prediction models

In addition to individual biomarker based diagnostic tests, a cornerstone of medical practice is gathering information on an individual and estimating the probability of particular outcomes based on this knowledge [19]. In a similar way to history taking and individual biomarker-based clinical tests, prediction models can be used to assist this process by providing estimates of the likelihood of specific outcomes. In particular, they enable this by consideration of multiple biomarkers and risk factors to provide an estimate of risk of specific outcomes [19].

Variously referred to as clinical prediction rules, risk scores, decision rules, or prediction models, they are widely used within healthcare to inform decision-making [20, 21]. They can take variety of forms (e.g. tabular or web-based application), target differing disease endpoints (e.g. development, recurrence, death) and may be used in different contexts for individualised prediction.

Example uses include identifying populations to target for preventive measures (e.g. cardiovascular risk scores), informing clinical decisions such as referral to further testing or risk-stratifying individuals for different therapeutic strategies (e.g. APACHE III prognostic system [22]). They may also be used by individuals and clinicians to make informed choices about interventions (e.g. PREDICT for breast cancer [23]).

Development of prediction models can be a complex process involving a series of steps from identification and selection of predictors to model validation [24]. Recently, this field has developed considerably with the availability of larger and more comprehensive datasets that are enabling the creation of a wider variety of risk prediction models. It is not within the scope of this report to cover this field, and details of the processes involved in model development can be found elsewhere [25, 26] . Key points to note are that there are different methodologies from simple weightings to complex machine learning algorithms that can be used in model development [24, 27] and many different models may be developed targeted at the same clinical question.

As with the evaluation of individual biomarker-based assays and tests, it is important to determine whether particular models are valid and likely to have an impact on healthcare pathways. Key parameters of interest are a model's performance and the degree to which it has been validated in external datasets. Different measures can be used in assessing the predictive performance of models, including assessment of the agreement of predicted and actual outcomes (calibration) and assessment of the model's ability to separate different outcomes (discrimination) [25, 28]. Examination of model performance in external datasets enables assessment of the generalisability of a model, and thus the extent to which it may be applicable to particular populations in specific contexts.

The extent to which prediction models are considered as a diagnostic technology can vary, leading to corresponding variability in the degree of scrutiny that prediction models undergo. This variability is compounded by the fact that the term prediction model is applied to a relatively broad group of instruments from simple decision rules to complex clinical prediction algorithms based on machine learning. In addition, models, especially when they are complex, may be further developed into digital tools.

Tools can be considered as mechanisms through which end-users can input and collate individual level information, apply a prediction model and obtain a particular output (e.g. 10-year personalised risk of developing breast cancer). They typically consist of online, user-friendly interfaces such as web-based or mobile applications into which data can be entered and from which a tangible output will be returned. Examples of such tools include the Breast Cancer Risk Assessment Tool (BCRAT) [29], CanRisk [30] and IBIS [31].

While there is a large body of literature that discusses the development and validation of prediction models, there is still uncertainty as to how these principles apply to the evaluation of risk tools for clinical use.

2.4 Conceptualisation of tests for purposes of evaluation

The term medical test evaluation is often used to refer to the process that aims to inform decision making on the incorporation of novel diagnostics or testing strategies (including prediction models) into healthcare pathways. It may be a relatively straightforward process or a more complex endeavour.

Complexity may arise when a biomarker is novel, where gold standard equivalents do not exist, or there are multiple parameters being considered as part of the testing strategy, such as in the case of prediction models that bring together data from different tests.

Whether it is a complex or simple evaluation, a key part of medical test evaluation is understanding the purpose of a test or testing strategy (i.e. intended clinical applications e.g. screening) and its role (how it will alter the current pathway) [32]. These contextual factors enable a clearer assessment of the evidence base of particular technologies and biomarkers to determine how well they function or perform, as well as the consequences of their use.

The rise of technology platforms that can potentially be used in multiple clinical contexts or settings is leading to a greater need to clarify these different aspects of evaluation. This issue has been discussed in relation to genetic tests, where ambiguity around the intended purpose and role may arise when single platforms or technologies can be used for multiple purposes (Figure 1).

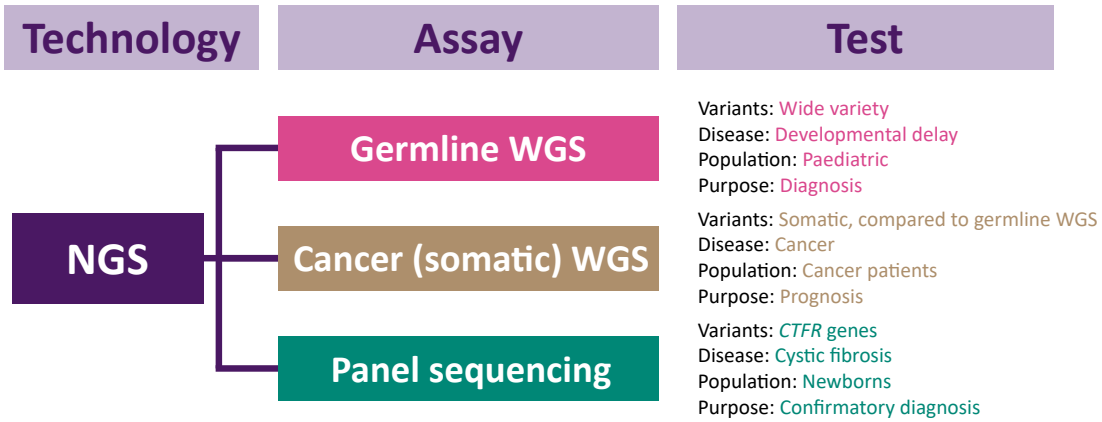
The use of different technologies to provide the same applications, and in certain instances the conflation of the technology with the application, can also create issues for determining intended purpose [33]. This led to the proposal to make a distinction between an assay and the test to enable greater clarity and more informative evaluation of genetic tests [34, 35]. This conceptualisation is also applied to other biomarker-based tests [32]. Technically, the method used to analyse a substance (or biomarker) in a sample is considered to be the assay. The test is described as the use of that assay:

- Within a specific disease context
- In a particular population
- For a particular purpose

This specific definition of a test in such a way enables linking it more clearly to its purpose and role as part of a specific healthcare pathway. This has important consequences in terms of medical test evaluation. Evaluating an assay can be restricted to validating a particular methodology, e.g. measuring a biomarker. This may demonstrate that a particular method used to analyse a substance or biomarker is reliable and robust within a range of conditions. In this regard, an assay is a scientific measurement of the biomarker, whilst a test is its interpretation.

Figure 1: One assay, multiple tests

This diagram illustrates how a single technology platform - Next Generation Sequencing (NGS) - can give rise to multiple assays. Each assay can be used to inform different questions, giving rise to different tests, each with differing purposes.



Evaluating a test is a more complex undertaking that encompasses wider considerations related to the use of the assay or model and its value, in addition to its technical performance. The use of this conceptual framework enables differentiation of the varying purposes of particular technologies (Figure 1), thereby enabling clearer assessment of the implications of an assay versus the test for healthcare pathways.

Consideration of the healthcare pathway is important for a number of reasons. The characteristics of the population and thus the prevalence of disease and case mix of patients (including disease stage) varies in different clinical contexts.

For example, a hospital-based population has very different characteristics to the general population. The prevalence of disease and case mix will impact on test performance characteristics, and the likelihood of obtaining a false result [36, 37]. This is referred to as spectrum bias, where in a high prevalence environment a positive result is more likely to be a true positive and a negative result to be a true negative. This is evidenced in secondary care, where there are more patients with severe disease who are more likely to be identified by a test. Therefore, any test evaluated in this context could overestimate predictive ability in a different clinical context (e.g. primary care).

An understanding of the clinical context also enables decision makers to determine if the predictive performance of the test is appropriate for the suggested use case. For example, tests used in the context of case finding usually optimise on test sensitivity in order not to miss potential disease. Whereas in a confirmatory diagnostic setting they are more likely to be optimised for specificity to avoid unnecessary treatment.

Considerations of the healthcare pathway also enables assessment of a novel test or testing strategy together with existing infrastructure (including tests that may be conducted in parallel or sequentially). Taking all these factors together also enables assessment of the value and contribution of information from a test to decision making processes.

As noted earlier in this report, due to variability in the extent that prediction models are considered as tests, the extent to which test evaluation concepts are applied to clinical prediction algorithms varies. Nevertheless, a key parameter in model development and their evaluation for clinical use requires a clear understanding of disease context, intended purpose and population for use. This ensures that appropriate data is used in model development and validation. Validation of prediction models is described further in Chapter 4.

2.5 Frameworks for evaluation of medical tests

Evaluation frameworks facilitate a systematic approach to evaluation of medical tests and can help mitigate against some of the challenges when assessing a novel diagnostic or testing strategy. Frameworks can enable multiple stakeholders to gain a shared understanding of an evaluation process and help to identify and agree upon appropriate objectives and methods.

Evidence appraisal is also supported by guidelines and standards developed to try and improve the quality of reporting of research activity. For example, guidelines that are applicable to test evaluation include the STAndards for the Reporting of Diagnostic accuracy studies (STARD) [5, 12]. In addition, the Prognosis Research Strategy (PROGRESS) group has proposed a number of methods to improve the quality and impact of model development [38], and a checklist for reporting on prediction or prognostic models - Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD) [39, 40].

Different frameworks and guidelines exist across fields and organisations. This has given rise to variation in terminology used to refer to similar concepts. For example as many modern technologies require software some frameworks exist that provide definitions and criteria to demonstrate software validation [41]. Together with the wide range of groups that have an interest in the outcomes of an evaluation process, this can contribute to misunderstanding. Nevertheless, a commonality amongst frameworks is ensuring that key parameters relating to the domains of analytical validity, clinical validity and clinical utility are addressed. Furthermore, this is often seen as an iterative or cyclical process as opposed to a simple linear one.

A wide range of groups have an interest in the outcomes of test evaluation processes, whether for laboratory diagnostics or prediction models. These include the individuals undergoing a test, healthcare professionals, healthcare laboratories, advisors to healthcare systems (for example organisations that set out guidelines or assessment units that evaluate the evidence), payers (e.g. individuals, laboratories, hospitals, departments of health) and regulatory bodies involved in approving healthcare products for use either by consumers or health systems.

Existing frameworks are a mechanism to bring together and critically appraise evidence to support decision making. Groups involved in decision making may have different thresholds with respect to the quality and quantity of evidence required to support the implementation, use or uptake of a novel diagnostic. They may also adopt different mechanisms to assess the evidence, based on their perspective and needs.

2.6 Summary

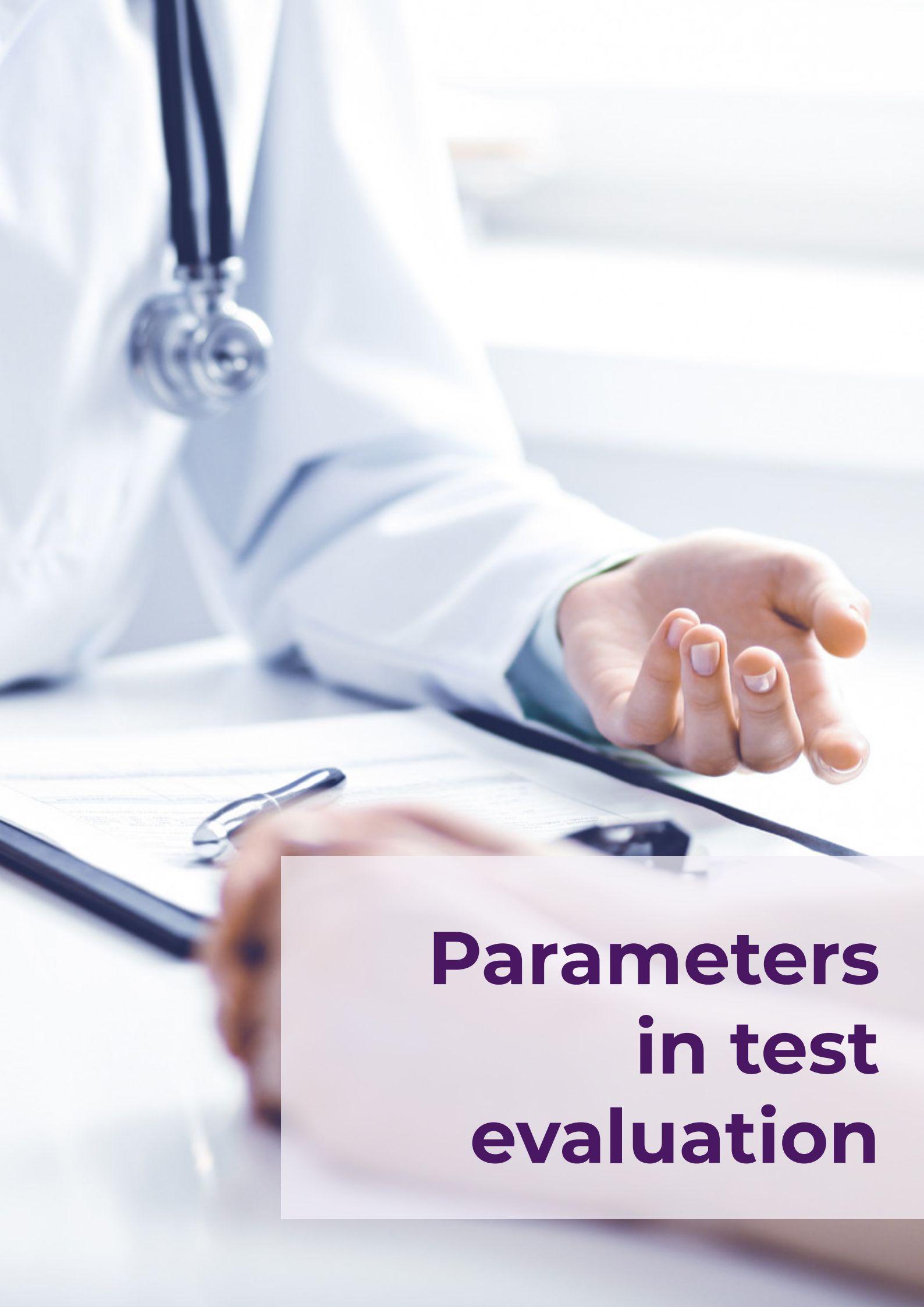
The umbrella term ‘medical tests’ encompasses a broad range of different investigations that inform care of an individual. This includes laboratory-based assays as well as prediction models. Appropriate evaluation can demonstrate their accuracy and impact on healthcare pathways, and aid decision making about their implementation.

Central to the evaluation process is understanding and defining medical tests in relation to their intended purpose and role as part of a healthcare pathway.

Evaluation can be a relatively simple process or a more complex endeavour, depending on the nature of the test and its intended use.

Laboratory assays such as next generation sequencing and prediction models can be complicated to evaluate due to the multiple components that make up or contribute to the testing strategy. This not only includes elements of sample or data collection, but also software used in analysis (such as bioinformatics pipelines or prediction algorithms and their associated tools). The validation and the function of these components may need to be considered as part of the evaluation process.

Methods and processes to address these issues continue to be developed, building upon existing frameworks for heath technology assessment and medical test evaluation. This is a continually evolving process as both technology and our understanding of disease biology progresses.



Parameters in test evaluation

3. Parameters in test evaluation

In this chapter, we provide an introduction to the parameters of test evaluation and expand upon mechanisms used to gather evidence, with a focus on genetic tests.

Key points:

- Test evaluation is an iterative process that examines evidence supporting analytical validity, clinical validity, and utility.
- Often evidence is not gathered in isolation for each of these domains and assessment requires consideration of the healthcare pathway.
- Diagnostic accuracy studies are the mechanisms by which clinical validity of tests are assessed, this is usually through comparison with a reference standard.
- Different metrics are reported by diagnostic accuracy studies to give an indication of the predictive and discriminative properties of a test.
- Principles of medical test evaluation can be applied to tests incorporating polygenic score analysis, however, specific aspects will need to be considered.

3.1 Key parameters in test evaluation

Evaluation frameworks exist for many types of tests with a large number developed specifically for molecular and genetic tests. This has been driven by technology development and the need to better understand the implications of novel diagnostics for healthcare.

Evaluation frameworks applied to genetic tests do not differ significantly from those applied to other tests, and aim to assess:

- Whether the test can accurately and reliably measure whether a variant is present
- If the test accurately measures or predicts the presence, absence or future risk of the clinical disorder
- The positive and negative impacts of carrying out the test
- The cost of testing
- The usefulness of the information obtained from using the test.

Genetic testing can be carried out for different purposes including for confirmatory diagnosis, prognostication or susceptibility testing.

Frameworks for evaluation of genetic tests differ in the 'type' of genetic tests they can be applied to. This diversity reflects the different considerations required in testing for different types of genetic variation (e.g. somatic, germline, etc.), clinical situations, and the levels of evidence that may be obtainable.

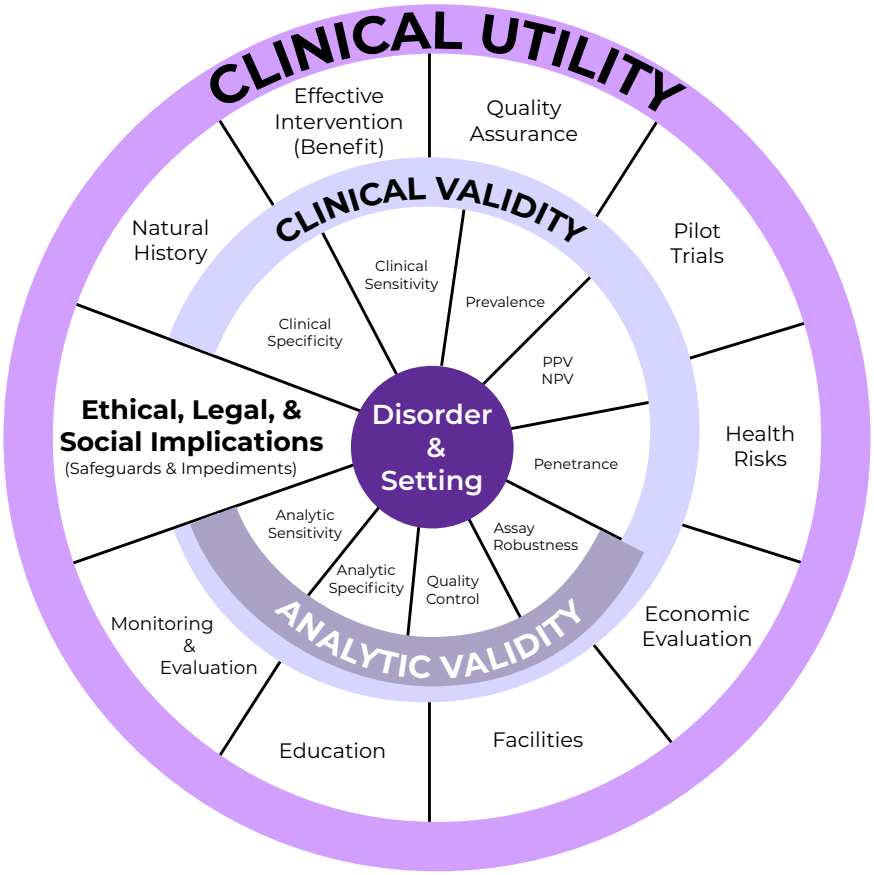
Many of these frameworks build on the ACCE (referring to Analytic validity, Clinical validity, Clinical utility and Ethical, legal and social implications) model (Figure 2), which is the most commonly used framework when evaluating a genetic test [9, 42]. This and other frameworks allow for a chain of evidence to be built to support decision making. As for other biomarker tests, genetic test evaluation frameworks outline critical parameters:

- **Analytic validity** of a test defines its ability to measure accurately and reliably the component of interest - the technical performance of the assay. For genetic tests, analytical validity defines the ability of the assay to measure accurately and reliably the genotype of interest. Analytical validity captures a wide range of key analytical performance characteristics of the test, such as analytical accuracy, precision, analytical sensitivity and specificity, reportable range of test results for the test platform, reference range and normal values [43].

- **Clinical validity** of a test defines its ability to detect or predict the presence or absence of the phenotype, clinical disease or predisposition to disease. Typically, clinical validity is described in terms of clinical sensitivity and specificity, and positive and negative predictive values, amongst other metrics. Exact metrics and thresholds for interpreting them will depend upon the disorder, setting and purpose of the test. This is because these parameters will have an impact on the mix of patients and prevalence of disease in these settings.
- **Clinical utility** of a test refers to the likelihood that the test will lead to an improved outcome. This is linked to the intended purpose of the test and may examine a range of factors such as effectiveness on clinical outcomes, feasibility of test delivery and/or cost-effectiveness. Final decisions with respect to clinical utility are usually based on holistic assessment of test performance characteristics, as well as consideration of a host of practical factors such as the impact and consequences of the test use on care pathways.

An additional parameter that is outlined as part of the ACCE framework is consideration of the **ethical, legal and social implications** of test use. This includes assessment of wider factors that may affect the likelihood that the test will lead to improved outcomes, such as access to testing and treatment, insurance, risk of discrimination and legal issues regarding consent and ownership of data.

Figure 2: ACCE model system for collecting, analysing and disseminating information on genetic tests.



Many frameworks present analytical and clinical validity, as well as clinical utility, as separate concepts, which can be demonstrated by using different types of studies and in a linear or sequential fashion. For practical purposes, these parameters of test evaluation may be established separately, with the basic components of analytical validity determined initially, before going on to demonstrate clinical validity and utility. However, these different parameters are interlinked, as illustrated by Figure 2.

Often the same study may be used to obtain information on multiple aspects of test evaluation. This especially applies to later phase studies, which may generate evidence relevant to both clinical validity and utility, as well as strengthening evidence on analytical validity. Interlinked with these parameters is scientific validity. This is sometimes described as part of clinical validity and denotes the relationship between a biomarker and an outcome. However, it can also be considered as the underpinning scientific evidence that enables interpretation of analytical and clinical validity. As such we describe this aspect of test evaluation separately in this report.

Below we describe in more detail how different characteristics of test function are determined in the context of currently provided genetic tests. It is not our intention to provide a detailed description of how all of these parameters are assessed across the genetic testing landscape, but to provide an illustration of the approaches that are currently taken.

3.2 Scientific validity

The National Institutes of Health Biomarkers Definitions Working Group define a biomarker as 'a characteristic that is objectively measured and evaluated as an indicator of normal biological processes, pathogenic processes, or pharmacologic responses to a therapeutic intervention' [44]. Different methods or instruments can be used to measure particular biomarkers (e.g. different sequencing technologies).

Effective use of a biomarker within healthcare requires biomarker validation (i.e. an understanding of the relationship between the condition of interest and the biomarker). This is sometimes referred to as the scientific validity of tests. The types of evidence and the strength of the disease association required for scientific validity can vary depending on the type of biomarker, type of test and test purpose.

For genetic tests, demonstrating an association between the genetic marker(s) and a trait is referred to as scientific validity. However, the quality and quantity of evidence supporting the association between specific genetic biomarkers and a trait can be variable. This has led to the development of frameworks and guidelines that can assist in determining the validity of specific gene-disease associations. These are discussed further below.

Determining gene-disease relationships – Mendelian disorders

Genetic tests used in the context of rare Mendelian disorders aim to either diagnose an existing condition or predict its future occurrence with a high degree of accuracy. Therefore, genetic analysis requires sufficient confidence of a gene-disease association, as well as the particular variants in that gene that are pathogenic.

Genetic tests may either analyse sequences that have already been classified as pathogenic or identify novel variants that need to be classified. For example, the use of next generation sequencing for whole genome or exome sequence analysis in those with rare diseases can lead to discovery of novel variants that require classification.

Over time, a variety of different resources and guidelines have been created to enable assessment of the scientific evidence supporting gene-disease and variant-disease associations. Numerous online databases of human genomic variation exist, including databases containing variant-disease associations (e.g. OMIM® [45], ClinVar [46]). In addition, initiatives such as the NIH funded ClinGen have developed a framework for reviewing data on gene-disease associations for Mendelian disorders [47]. This has enabled the creation of the Clinical Genome Resource, which provides clinicians and researchers with information on the clinical relevance of genes and variants [48]. These resources can be used to assess the scientific validity of particular gene/variant-disease associations.

The move from more targeted genetic analysis to the use of whole genome or exome analysis may uncover novel variants in diagnostic settings and has created challenges for determining gene-disease associations. Guidelines for the interpretation of findings from such analyses have been created to address this challenge. In 2015, the American College of Medical Genetics (ACMG) in collaboration with the Association for Molecular Pathology (AMP) published guidelines to enable a more systematic approach to variant interpretation for Mendelian disease diagnosis [49].

The ACMG framework details different levels of evidence for or against pathogenicity and outlines rules for combining evidence sets in order to classify variants into one of five categories. The ACMG-AMP guidelines have been further developed by the ClinGen Sequence Variant Interpretation (SVI) working group. These guidelines have been adopted internationally, including by the Association for Clinical Genomic Science (ACGS) in the UK in 2016 [50].

Disease-specific variant expert panels have also been established and are generating disease/gene specific guidelines (e.g. Huntington's disease, Lynch syndrome or cystic fibrosis) [51].

As these guidelines are established to standardise the weighting of evidence for different diseases to determine a causational relationship, they may not be relevant to all types of genetic testing, for example in common diseases, where genetic data can be used as a proxy for the true variant involved in disease. ClinVar has a specific sub-group to consider development of evidence-based assessment of actionability of PGS [52, 53].

Determining gene-disease relationships – other disease areas

The availability of curated databases to inform interpretation of gene/variant-phenotype associations also impacts on the use of genome analysis in other settings. This includes therapeutic decision making, such as in somatic changes in tumours, or the wider surveillance and management of risk, such as WGS for identification of antimicrobial resistance in bacteria. These differential uses impact on the guidelines that are developed for variant interpretation for these specific purposes. For example, standards and guidelines have been proposed by joint consensus of the Association for Molecular Pathology, American Society of Clinical Oncology and College of American Pathologists that categorise somatic variants into four tiers based on their level of evidence for clinical significance in cancer diagnosis, prognosis and/or therapeutics [54].

These above examples illustrate that different approaches are taken in establishing validity of different genomic biomarkers. Furthermore, there are issues and challenges when developing the scientific evidence base necessary to ensure the validity of the adoption of genomic tests for any application within healthcare.

Ongoing research and validation are essential for determining the accuracy with which available databases and tools can inform clinical decision-making. While existing frameworks may not be directly transferable to polygenic scores, they provide an illustration of the different approaches that are often needed for different biomarkers.

3.3 Analytical validity

An understanding of the relationship between a biomarker and disease is also influenced by the instrument(s) used in these measurements. Therefore, any instrument used in analysis of the biomarker should also be validated to show that it can accurately and reliably measure the intended analyte(s), e.g. proteins, genetic variants, lipids. This includes both the physical and computational elements of the instrument that the test comprises. Ideally, performance should be assessed in the context of test use, i.e. in the intended population of use and in real world settings e.g. using clinical sourced samples instead of or in addition to artificial samples when evaluating performance.

Analytical validity aims to determine whether the assay detects what it claims. Similar to other biomarker tests, the analytical validity of a genetic test is determined through interrogation of the accuracy and precision of both the wet and dry lab (bioinformatic) components of assays. Assays employed in genetic testing can take a variety of forms, and be used to interrogate a specific variant, a small-subset or a range of genes through to whole genome sequencing. Therefore, in determining analytical validity, it is often useful to define the assay type and region that will be interrogated (e.g. targeted variant, panel of genes, whole exome, whole genome, etc.).

It is important to note that different regions in the genome may be more or less difficult to sequence and, additionally, identification of certain types of genetic changes may be limited by the chosen technology. Examples of this include complex structural rearrangement, GC-bias, or regions which are highly repetitive or highly homologous to other regions in the genome, such as pseudogenes.

Regardless of assay type, the parameters for assessing the analytical validity of wet-lab procedures are well established. A more pressing issue is determining the analytical validity for the computational elements, particularly bioinformatic analysis of the sequencing data and the interpretation of variants identified. Furthermore, it is impossible to fully separate the wet and dry lab components of sequencing when considering analytical validity.

Clinical genome analysis is increasingly being delivered using NGS technologies, with different bioinformatics driven strategies used to address different clinical questions. Thus, while genetic tests can take a variety of forms, they are broadly based on using the same underlying technology and fundamental methodological principles. This can simplify some aspects of the validation process, especially as in early phases of the test development cycle, the focus of evaluation is on technical feasibility and analytical validity, which may be shared across tests. However, as large parts of current genome analysis pipelines are driven by bioinformatics (the mechanism by which raw sequence data is processed to enable interpretation), tools used for these processes also need to be validated.

Bioinformatic analysis is performed in pipelines with key stages including:

- Quality control
- Alignment
- Variant calling, variant filtering and variant interpretation
- Variant confirmation.

The purpose of effective bioinformatics analysis is to detect all actual/true/real genomic variation in the examined sequence and to accurately identify and discount false/unreal variation, allowing for subsequent accurate interpretation of true variants and, within a clinical context, reporting by a trained healthcare professional. Bioinformatic workflows are designed for a specific assay and may not be as accurate when used in combination with different wet lab protocols.

Important considerations include:

- Intended sample type
- Hybridisation versus amplicon gene panels
- Different sequencing platforms.

A recent study highlighted that a significant amount of the complexity of NGS workflows relates to decision making for bioinformatics pipelines, as well as explaining a significant amount of the variation in results from NGS tests [55].

The choice and customisation of these analysis pipeline(s) can have a profound effect on the interpretation of genomic information, for example resulting in miscalling of bases or misalignment of sequences. Low concordance identified when using different combinations of aligners and variant callers may be explained by a number of factors including choice of sequencing platform, variants of interest, and GC content [56]. Different bioinformatics tools are also available for analysis of DNA or RNA, germline or somatic, single nucleotide variants (SNVs), small insertions and deletions (indels), or larger complex genetic variants (such as copy number variants (CNVs) or structural variants) [57].

One approach to overcome the limitations of individual bioinformatic tools is to combine tools to increase sensitivity by integrating the results of multiple variant calling tools. However, any conflicting results between these tools can create additional complexity, whereby further work is needed to determine true variation prior to interpretation.

Criteria to define appropriate analytical validity and quality management systems (QMS) for the bioinformatic pipeline as a constituent of sequencing-based tests are essential to ensure the reliability and replicability of results. Improperly developed, validated, and/or monitored pipelines may generate inaccurate results that may have negative consequences for patient care. For this reason, clinical bioinformatics pipelines developed to analyse clinical NGS data require a robust quality assurance programme for both ongoing monitoring of metrics and pipeline updates.

Continual software updates and data sources for annotation makes the development, validation and deployment cycles of bioinformatics pipelines challenging. Bioinformatics pipelines require revalidation in an iterative process, performed using reference standards to ensure reproducibility, with systems to track versions and implementation dates [58].

Testing mechanisms within these revalidation processes, in an approach known as ‘deep testing’, enable identification of any errors or changes to test performance. Proficiency testing of laboratories ensure comparability and reproducibility of results across laboratories. However, there are limitations to the reliability of this process, including 1) only reporting actionable variants; 2) differences in the gene targets between NGS workflows (e.g. targeted panel versus WES); and 3) differences in assay analytical validity [57].

Establishing the analytical validity of assays such as NGS technologies where multiple elements influence data acquisition has been complex. The utilisation of several bioinformatics elements as part of the pipeline contributes to this complexity. This is due to the variety of bioinformatics tools available, as well as their constant evolution over time.

Currently, there is a high degree of variability in the approach taken in molecular genetics and pathology to establish and validate bioinformatics pipelines [59]. As we discuss later in the report, there are similarities between these elements and PGS analysis, which is also reliant on several computational elements. Experience from the assessment of NGS technologies can be useful in informing approaches for assessing the analytical validity of polygenic score analysis.

3.4 Clinical validity

As described above, the clinical validity of a test defines its ability to detect or predict the presence or absence of the phenotype, clinical disease or predisposition to disease in the context of its use. There are other terms that can be used to refer to this particular aspect of a test, including diagnostic accuracy, clinical performance and predictive ability.

In the sphere of genetic testing, clinical validity is the most commonly used term and refers to the predictive ability of an assay in a defined population for a particular purpose [9]. Clinical validity is often predicated on showing the association between biomarker and disease as well as the ability of an assay to detect the biomarker. This is because a test lacking sufficient analytical performance will not have reliable detection or predictive abilities, similarly a lack of scientific validity is likely to result in a test with poor diagnostic or predictive performance.

It should be noted that the definitions of some of these terms, especially those of prognosis and predictive, vary between different fields, in some cases being used interchangeably [60, 61]. For example, in the field of genomics, tests are typically categorised as diagnostic, when used to help diagnose an existing phenotype, or predictive, when used to predict the occurrence of a future disease phenotype [62]. As the convention in the test evaluation field is to use the term ‘prognostic tests’ to refer to those that predict future disease phenotype, we will use this term.

The term ‘predictive test’ is used to refer to those that predict patient response to a specific intervention or treatment. In this case a key outcome of interest is change in patient outcome rather than the ability of the test to detect or predict a condition of interest. Such tests are assessed using different studies and are not the focus of the discussion below.

Diagnostic and prognostic tests are typically evaluated following assessment of diagnostic accuracy studies. These studies produce metrics and evidence that demonstrate how well a test performs when used in the intended clinical context, and that the outcome of the test correlates with an outcome relevant to its intended purpose. Such studies are cross-sectional and compare the performance of the test being evaluated (the index test) to a reference standard.

The reference standard is usually that which is considered the 'gold standard'. For some conditions such as high blood pressure, well validated reference standards are available. However, in other cases there may be no clear reference standard available, for example if a condition is less well characterised and/or is hard to accurately characterise using existing methods. Sometimes the ideal reference standard may exist but not be practical for use in a such studies, for example if complex high-risk surgery is required. In these situations, an alternative reference standard may be used.

Sufficient evidence to demonstrate whether a particular test has clinical validity may be obtained from a single comprehensive study, or from a variety of studies demonstrating different test performance parameters. The choice of exact study design to generate the evidence needed for clinical validity will be influenced by the intended purpose of the test. Factors that need to be considered include:

- The position of the test in the clinical pathway
- Whether it will be used together with other tests
- Interventions that may follow
- Whether it is being used in a diagnostic or prognostic context.

Studies may also be designed to enable comparison of more than one test (comparative accuracy studies).

For prognostic tests, it is essential to define the time period over which patient follow-up occurs following the use of the index test. This could be determined based on the intended interventions following test use. For example, a test to predict the risk of developing prostate cancer used in men aged 40 may appear to be highly sensitive and specific if evaluated after a 50-year time period, by which time some men will have developed prostate cancer. But test sensitivity and specificity metrics calculated over this time frame are meaningless if the test is intended to be used to identify those men who are more likely to develop cancer over the next ten years.

Metrics or measures used to assess the performance of a test can vary [63, 64]. Measures that are provided will be influenced by whether the test results are dichotomous (e.g. test results are positive or negative or individuals are classified as diseased or non-diseased) or non-dichotomous. In either case, these measures are not fixed properties of a test and will be influenced by population characteristics, such as disease prevalence, setting and study design. The most commonly used metrics in relation to clinical validity of a test are clinical sensitivity and specificity, and positive and negative predictive values [65].

Sensitivity and specificity are useful baseline metrics to compare test performance, but may not be useful in helping clinicians understand whether a result for a particular individual is likely to be true or false, which is critical for clinical decision making [66]. Therefore, metrics such as PPV and NPV may be reported to enable clinicians and decision makers to understand the impact of carrying out the test.

These metrics will vary depending on the prevalence of the condition of interest in the relevant population. However, such analysis is only applicable where test results are dichotomous or there is a clear threshold for test positivity.

Where a test is non-dichotomous and/or has multiple different thresholds for positivity other measures may be more suitable. Receiver operator curves (ROC) can be used to visualise the relationship between sensitivity and specificity. The area under the receiver operating curve (AUROC or AUC) is often used to indicate the discriminatory performance of a test. The curve can be used to determine the specificity at a certain sensitivity cut-off point or vice versa, which can help decision makers in determining particular thresholds or cut-off values for test use. Test thresholds or cut-off values can be selected in advance of test performance evaluation, based on values that are useful for the intended clinical purpose. However, for novel tests or use cases, this may not always be apparent. In such cases, optimal thresholds may be reported based on data generated by the study.

In addition to measurements of specificity and sensitivity and predictive values, other information relating to test performance is also considered when assessing clinical validity. For example, the size of the confidence intervals and p-values related to different measurements can provide information of whether a test is significantly better at identifying a condition of interest than a comparator.

In the context of Mendelian disorders there are some specific additional considerations in relation to determination of clinical validity that are considered for all these uses. Factors such as penetrance (probability that a disease will appear when a disease related genotype is present), variable expressivity (range of severity of signs and symptoms) and pleiotropy (the same variant affecting multiple traits) will all have an impact on interpretation of studies, and on determining clinical validity [67].

It is also important to consider time between testing and outcome in appraisal of studies. This is especially important in the context of predictive genetic tests, that is, those that aim to assess the future risk of developing a specific condition.

As with other types of biomarkers, the types of studies that can inform evaluation of tests based on genetic biomarkers will depend on the intended use of the test as either a diagnostic, prognostic or predictive biomarker.

3.5 Clinical utility

Any new change of healthcare practice, including the introduction of tests, typically requires demonstration of value. Clinical utility has no singular or agreed definition and is a broad term that is used to denote usefulness or value. In the context of healthcare associated tests, there are two main considerations that influence decisions about usefulness, firstly on the information or data obtained from a test and how it can improve outcomes; and secondly on the value placed on implementing a test within the context of its use.

Some tests may be easier to use and implement in comparison with others, and this needs to be considered along with the information they provide. These aspects can be viewed from numerous perspectives: public health, clinical, personal or social; and are intricately linked with the purpose and context of testing.

Whilst demonstration of clinical utility is often considered the endpoint, the preceding steps do contribute to an overall assessment of usefulness. For example, analytical validity (the ability of a test to correctly detect an analyte) demonstrates one aspect of overall test performance, and it is unlikely that tests that perform poorly would have utility. In addition, assessment of scientific validity (biomarker-disease association) and clinical validity (test performance in a clinical setting) are important stepping stones towards demonstration of clinical utility. This is especially pertinent within the context of health services, where tests that are taken forward are usually expected to have sufficient evidence of analytical and clinical validity.

There is generally an absence of a set threshold for test performance parameters, as these are influenced by the context and purpose of the test. For example, in some scenarios a test with low sensitivity and specificity may be acceptable, but in others it may not. However, critical to the test evaluation process is understanding the disorder and healthcare setting or pathway into which the test or testing strategy will be integrated. This may mean that in some instances a genetic test may be used alongside other tests to inform decision making.

Final decisions with respect to clinical utility are usually based on a summative assessment of test performance characteristics, as well as consideration of a host of practical/pragmatic factors such as the impact and consequences of the test use on care pathways. Broad areas for further consideration therefore include the safety, effectiveness and efficacy of tests.

Evidence in support of clinical utility can relate to these different elements. It may come in different forms (i.e. quantitative or qualitative); and may potentially come from a variety of studies, as opposed to a single study, especially as test use can have both direct and indirect outcomes on healthcare pathways. Sources of evidence that are deemed acceptable may vary, ranging from diagnostic accuracy studies and RCTs to modelling studies and observational data, depending on the intended purpose and context of testing.

End-to-end studies, which capture evidence of both diagnostic accuracy and the different impacts of a test, can be arduous and expensive to undertake. Thus, decision makers may use a linked evidence approach to bring together test accuracy data with existing evidence on the condition and treatment pathways to determine the impact of test use. This means that while it is important to have evidence of clinical utility, the presence of clinical validity data can enable this assessment without the need for further studies.

3.6 Summary

There is widespread acceptance that genetic biomarkers are a useful part of healthcare practice. However, genetic biomarkers and genetic tests are not a single entity. While diseases have a genetic component, the exact nature of variants identified and linked to particular diseases and traits differs. Different techniques and technologies are used to measure specific types of genetic variation, from single nucleotide changes to the insertion or deletion of long stretches of DNA. In addition, the scientific evidence supporting association between genetic variants and observable traits also differs. This can create a challenge for assessing the scientific basis of particular tests and their usefulness as part of clinical practice.

As with other biomarker-based tests, the information from genetic tests are used under different circumstances to answer distinct questions. In the past, there has been a tendency for the focus in genetic test evaluation to be with the technology and what it can be used to identify. This is probably because in early phases of the test development cycle, the focus of evaluation is on technical feasibility and analytical validity, with later evaluation phases shifting to capture evidence relevant to clinical validity and clinical utility. Initial challenges in evaluation of sequencing technologies have been addressed by defining the clinical context for testing, the testing strategy, and standards for interpretation.

Frameworks and guidelines utilised by health professionals are undergoing continuous expansion and refinement. While genetic test evaluation frameworks have largely been used in the context of tests developed for Mendelian disorders, the principles they encompass can also be applied to the analysis of common low penetrance variants. We explore this more fully in Chapter 6 along with its intersection with risk prediction modelling, which is described in Chapter 4.

Evaluating risk prediction models

4. Evaluating risk prediction models

In this chapter, we provide an overview of the evaluation of risk prediction models and discuss differences in the conceptualisation of validity between genetic test evaluation and prediction model evaluation.

Key points:

- Validation of risk models, like evaluation of medical tests, requires consideration of context and purpose first and foremost.
- Model validation studies examine the predictive and discriminative properties of models using different datasets.
- There are overlaps in metrics reported by model validation and test evaluation studies in relation to test performance characteristics.
- Additional measures may be reported by model validation studies, including net-reclassification index (NRI) and integrated discrimination improvement (IDI).
- External validation of models can provide information on the clinical validity of a model, especially when there is alignment between the use case population and external validation dataset.
- Utility of models is assessed through examination of performance metrics or employing methods such as decision curve analysis (DCA).

4.1 Model validation studies

Evaluation is conducted at different stages in model development using statistical approaches to help inform its development as well as assessment of whether particular models are suitable for implementation. Model validation studies determine how well particular models predict risk and for which populations. This involves examining model performance in different datasets. Key parameters that are examined are model calibration and discrimination, these and other measures used to assess model performance are described in Section 4.2.

Validation studies can be differentiated by the datasets used. Internal validation refers to evaluation in datasets similar to that used in model development. For example, validation may be conducted using a subset of the data used to develop the model or using statistical approaches such as bootstrapping. Such studies can be used to test whether particular modelling approaches work and are optimal. They can also be used to determine if the model works for the target underlying population. As model fit and performance are usually better in the dataset used to develop the model, internal validation may provide optimistic results in relation to model performance.

External validation in an independent dataset is recommended to better determine performance and generalisability of the model [28]. External validation uses datasets outside of those utilised in model construction. This step can allow understanding of the generalisability, transportability and/or clinical validity of a model. Such studies also allow assessment of whether adjustment of the model can improve functionality in different settings or enable tailoring to particular circumstances.

External validation can provide information on the clinical validity of a model, especially when there is alignment between the use case population and external validation dataset [68]. It is often an implicit assumption that external validation datasets are equivalent to the population of use; however, this may not always be the case. Furthermore, while it is widely accepted that external validation of models is an important step prior to implementation, the extent of external validation needed prior to implementation is unclear [69].

4.2 Key parameters assessed in model validation studies

Key parameters that are assessed in model validation studies are calibration and discrimination. 'Model fit' or calibration examines discrepancies between predicted and observed values and is usually examined by plotting observed outcome against predicted risk. Well-calibrated prediction models have a slope equal to 1, with predicted risk falling along the reference line. If predictions are higher or lower than observations, there is deviation from the reference line. Calibration in an external dataset is important to assess if the model functions as well outside the datasets with which it was developed.

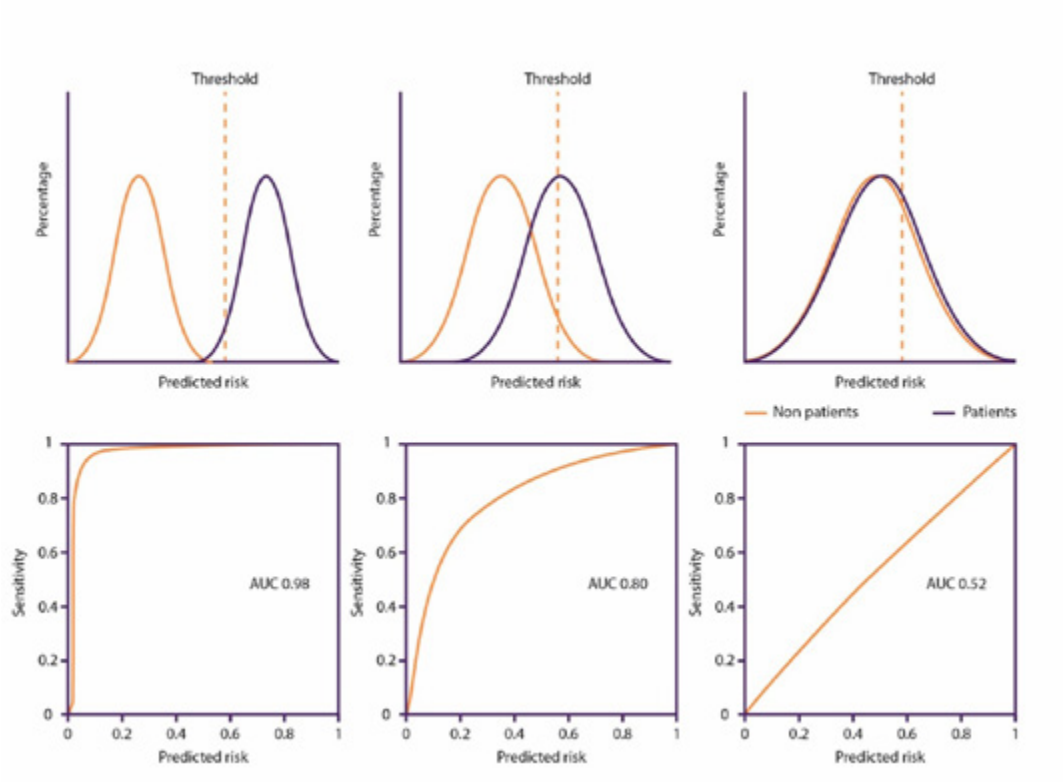
The discriminative ability of a model indicates the ability of the model to distinguish between those who will develop an outcome and those who will not. This can be measured using the C-statistic or C-index, which is equivalent to the area under the receiver operating curve (AUC) for binary outcomes. The measure ranges from 0.5 (no discriminative ability) to 1 (perfect discriminative ability). Models that produce risk distributions for cases and controls with less overlap result in a higher AUC (see Figure 3).

Similar to tests that are non-dichotomous, it may be useful to examine combinations of sensitivity and specificity for models at different thresholds. This is useful as clinical decisions are often made in categorical or dichotomous ways and have thresholds, so it can be useful to know the number of people above or below particular thresholds.

Importantly, AUC provides information on the discriminative capacity and does not summarise the clinical impact of the model, as there is no set threshold for AUC that can be considered optimal. The optimal AUC is dependent on the intended use of the model. Situations where accurate disease classification is needed (e.g. diagnosis or identification of sub-groups for an expensive intervention) will require a higher AUC. Conversely, population stratification into risk categories for differential management where interventions are inexpensive or have little harm may not require as high an AUC.

Figure 3: Risk distributions of cases (purple) and controls (orange) on the basis of a particular threshold (dotted line).

The area under the ROC curve (AUC) is shown below each risk distribution (adapted from Janssens & Martens. *Introduction to prediction research* 2018 [70]).



AUC is the most commonly used measure of predictive ability of a model. Increases in AUC are often examined to assess improvements in predictive ability of models when new predictors are added or to compare two models. However, this metric has been criticised as being an insensitive measure that is not able to fully capture all aspects of predictive ability [71]. This is because increases in AUC are usually small when predictors with small effect size are added, especially if the existing model is already able to discriminate between cases and controls effectively. This has led to the development of alternative metrics to evaluate model performance, such as increase in risk difference - integrated discrimination improvement (IDI) or net re-classification index (NRI).

IDI compares the difference in average predicted risk for cases and controls for two prediction models. For example, average risks in models with and without genetic factors can be compared. If the addition of genetic factors results in better separation of cases and controls, this would result in a positive IDI value. If there is an increase in risk differences between the two models, this serves to indicate that improved discriminative capacity has been achieved.

Model performance may also be evaluated by examining impact on reclassification of individuals across thresholds. Examination of reclassification using measures such as NRI can allow assessment of whether the addition of predictors results in differential classification of individuals across thresholds. For example, the addition of predictors (e.g. polygenic score) to an existing model (e.g. QRISK® [72]) can lead to changes in risk distribution and predicted risks, this in turn can lead to re-classification of events across thresholds, which in turn lead to different treatment decisions.

Measures such as the NRI assess the improvement in discrimination for specific risk thresholds, but are influenced by the value and number of thresholds [73].

4.3 Determining clinical utility of models

The clinical utility of a risk model depends first on its ability to accurately and correctly stratify a population. Correct stratification then allows for division into categories with sufficiently distinct risks resulting in an impact on provision of interventions. Division into various risk categories can be done for a number of reasons, such as to provide different interventions to those who fall into different categories. Categorisation depends on various factors, including the absolute risk of disease, the available strategies for disease prevention in the population and the risk-benefit implications of the interventions.

Similar to medical tests, utility of models may be assessed on the basis of metrics of diagnostic performance such as sensitivity and specificity [74]. This is possible when thresholds for performance and clinical impact are clear. However, in many cases this is not apparent.

Impact studies and decision analytic studies (DCA) have been proposed as mechanisms to examine the effect of model use on clinical pathways [75]. In DCA a clinical judgment of the relative value of benefits (treating a true positive case) and harms (treating a false positive case) associated with prediction models is made. As such, the preferences of patients or policy makers are accounted for by using a metric called threshold probability. A decision analytic measure called net benefit is then calculated for each possible threshold probability, which puts benefits and harms on the same scale [76].

4.4 Summary

Frameworks for the evaluation of risk prediction models exist and can be used in the appraisal of polygenic score models. Many of the parameters that are evaluated are broadly similar to those outlined in the previous chapter. However, terms such as scientific and analytical validity are rarely used in reference to prediction models. This is probably because most of the data parameters come from existing validated tests, such as a cholesterol test. Nevertheless, internal validation may cover some of these aspects through examining the predictive ability of different modelling approaches and their reproducibility. External validation then allows assessment of the predictive accuracy of models in alternate datasets, to examine their potential applicability and generalisability. These processes may lead to adjustments being made to the model to improve its calibration and discrimination, and thereby functionality in alternate datasets.

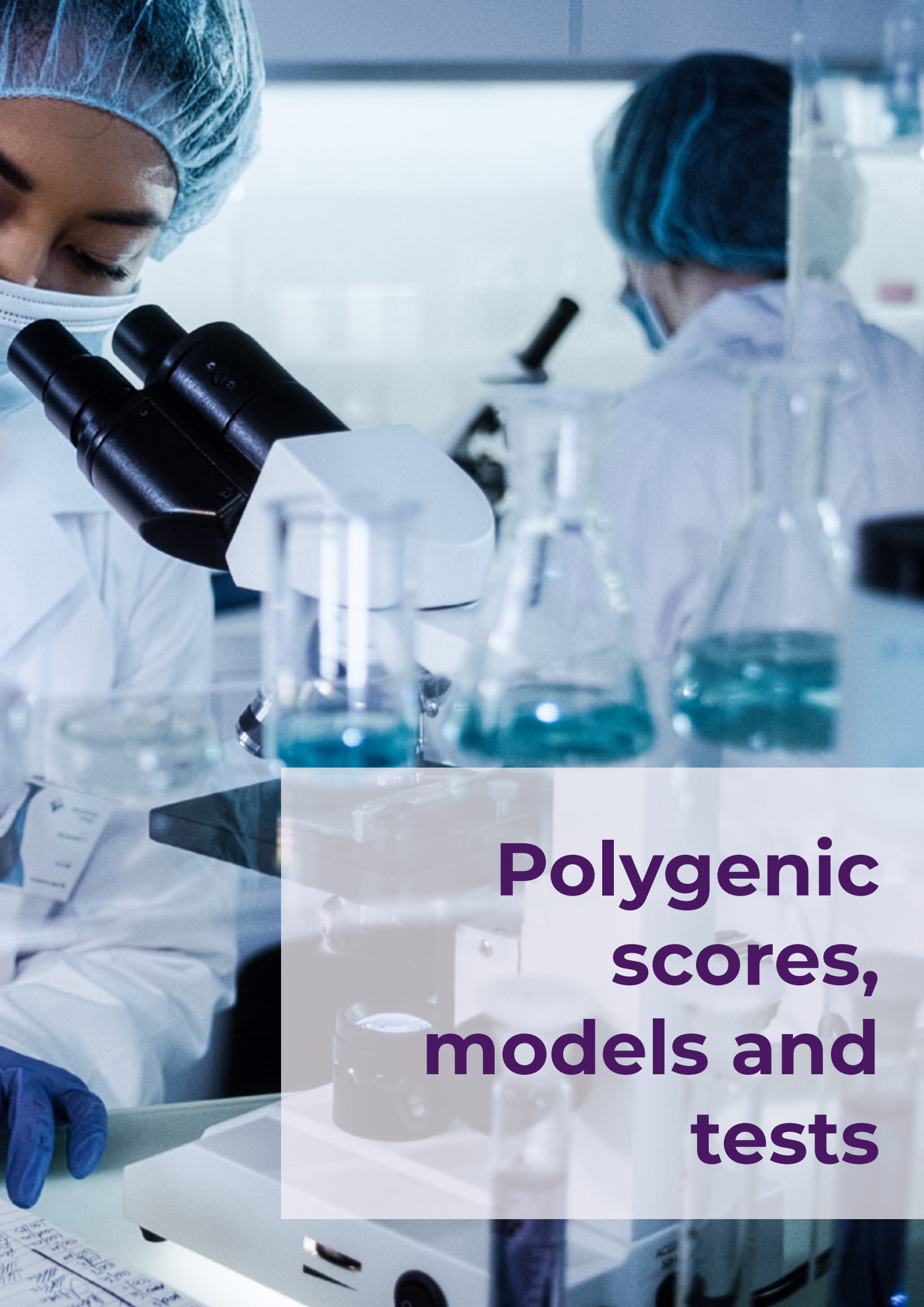
Model clinical validation studies provide information that is required for assessing clinical validity, by demonstrating how well a model can stratify a population into risk groups, on the basis of thresholds that are clinically relevant, such as those in existing guidelines. The use of these thresholds can enable measurement of how a particular model classifies individuals in comparison to the true disease status of the individual (i.e. true negative, true positive etc.). In practice such studies may be conducted using existing research cohorts prior to evaluation in real-world settings.

Evaluation of risk models, similar to the evaluation of medical tests, requires considering context and purpose first and foremost. Different parameters of model performance can then be appraised to assess if the model has been developed and validated appropriately.

The starting point of prognosis research should be the development of models to address a clinical need. Therefore, clinical validation parameters can in theory be more closely aligned with demonstrating utility. This means that in addition to parameters such as sensitivity and specificity, additional measures such as NRI and IDI may be reported. However, as mentioned previously, the variation in standards in reporting of model performance can create difficulties for those evaluating the evidence base [77].

There is large variability in the quality of models that are produced, and in the degree to which their predictive performance has been assessed and validated in appropriate datasets [21]. In addition, many models have been produced that do not provide a clinically relevant endpoint in relation to management. This has led to very few models being implemented or used as part of clinical practice [38, 78].

The introduction of models into clinical practice is not always a simple process and has often been compared to that for complex interventions due to the multiple interacting components that are needed, and which together impact on downstream outcomes [75]. These include model predictions, user understanding of model outputs, available interventions, and how use of the model may impact on administration or adherence to interventions.



Polygenic scores, models and tests

5. Polygenic scores, models and tests

In this chapter, we provide an introduction to polygenic scores, outline some of the terminology in this field, and include a description of the key components required for polygenic score analysis. The term polygenic score is often used interchangeably to describe the result produced by a polygenic score model, the underlying model that generates this score, as well as the overarching test strategy employed in such analysis.

The conflation around terminology for polygenic scores, and the components that make up each stage in the process, can make navigating test evaluation confusing. We attempt to clearly define each component to better identify the criteria that could enable evaluation.

Key points:

- Polygenic scores (PGS) can be considered a set of proxy genetic biomarkers. Each needs to be considered separately.
- The term polygenic score analysis can be used to describe the process of obtaining a PGS.
- PGS analysis can be used in different clinical contexts, giving rise to different tests.
- PGS analysis pipelines comprise different elements and can be composed in different ways.
- Achieving clarity as to what elements form or contribute to any particular pipeline or testing strategy is useful in considering how to approach validation of these components and the pathway as a whole.

5.1 What are polygenic scores?

Research has identified many common genetic variants in the form of single nucleotide polymorphisms (SNPs) associated with disease. However, each variant only has a small effect on disease risk. Given that individuals each have different sets of these variants, using these SNPs individually for risk prediction is untenable. However, examining their collective impact has been shown to be a potential mechanism to better utilise this information. This can be achieved by aggregating information across SNPs into a single score.

Different terms are used to refer to this score, such as polygenic risk score (PRS) or genetic risk score; we use the term polygenic score or PGS throughout this report. These scores can provide a single measurement of the cumulative effect of a large number of low-impact genetic changes for a specific trait or disease.

Polygenic scores for any given trait are normally distributed in a population, providing a spread of risks. For an individual, it may be informative to know their polygenic score and where they lie in the spread of risk for a disease.

Polygenic scores are a proxy marker of genetic liability to a trait, especially at the individual level [79]. This is because they are not simply a measurement of the presence or absence of pathogenic genetic variants, but rather a calculation based on analysis of different SNPs across the genome using a predictive algorithm.

At the individual level, polygenic scores are not deterministic or highly predictive [79, 80]. In addition, the pattern of inheritance of individual or combination of variants contributing to disease is not Mendelian, meaning that the implications for family members remain unclear.

Factors that contribute to the relatively lower predictive ability of PGS at the individual level include incomplete knowledge of causal SNPs and their impact on particular traits and diseases. Furthermore, as PGS are a calculated measure, any errors or uncertainties in the underlying datasets that were used in developing the predictive algorithm to calculate them will carry through to the final score. These factors impact on interpretation of information from analysis at the individual level.

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.

In addition, while polygenic scores can be considered stable due to their genetic underpinning, they are not simply a measurement of genetic variantion. Therefore, as scientific advances are made in uncovering true causal SNPs and methods in calculating polygenic scores evolve, interpretation of risk associated with the genetic variants and polygenic score analysis will shift. This means that similar to other genomic biomarkers, whilst the underlying variants contributing to a particular PGS can be considered 'stable', as our knowledge of the relationship between genomics and disease evolves, so will our ability to develop methods to accurately measure and interpret genomic risk profiles based on this information.

Taken together, this means that polygenic scores provide some information on genetic contribution to risk of disease. However, this only captures a proportion of overall risk, which may be further modulated, especially in the case of complex diseases by non-genetic factors. Nevertheless, PGS can be used as an additional piece of information along with other information about an individual to refine risk assessment.

Polygenic scores are likely to be more useful and informative for certain diseases and contexts than others. For example, the distinct genetic architectures of different cancers (i.e. different types and effects of SNPs) affect the current and future optimal performance of polygenic score models and their predictive ability [81]. Moreover, at the moment we still need to develop the evidence-base for use of PGS analysis within specified care pathways. This requires a better understanding of what the information from PGS analyses means or adds, and the implications of their use in different care pathways.

5.2 Polygenic score models

As described above, polygenic scores are a calculated measurement, usually expressed as a weighted sum of the SNPs associated with a disease. Historically, calculating a score relied on simple processes for weighting a few SNPs, but this has evolved over time as more SNPs associated with disease have been discovered. Current processes still calculate a PGS as a weighted sum of SNPs, but more complex statistical modelling is used to select SNPs and assign weights.

The algorithm developed for calculation of a score can also be referred to as a polygenic score model. However, as described below, this is a broad term used to refer to a variety of different types of models, many of which are developed for research as opposed to use in a clinical setting.

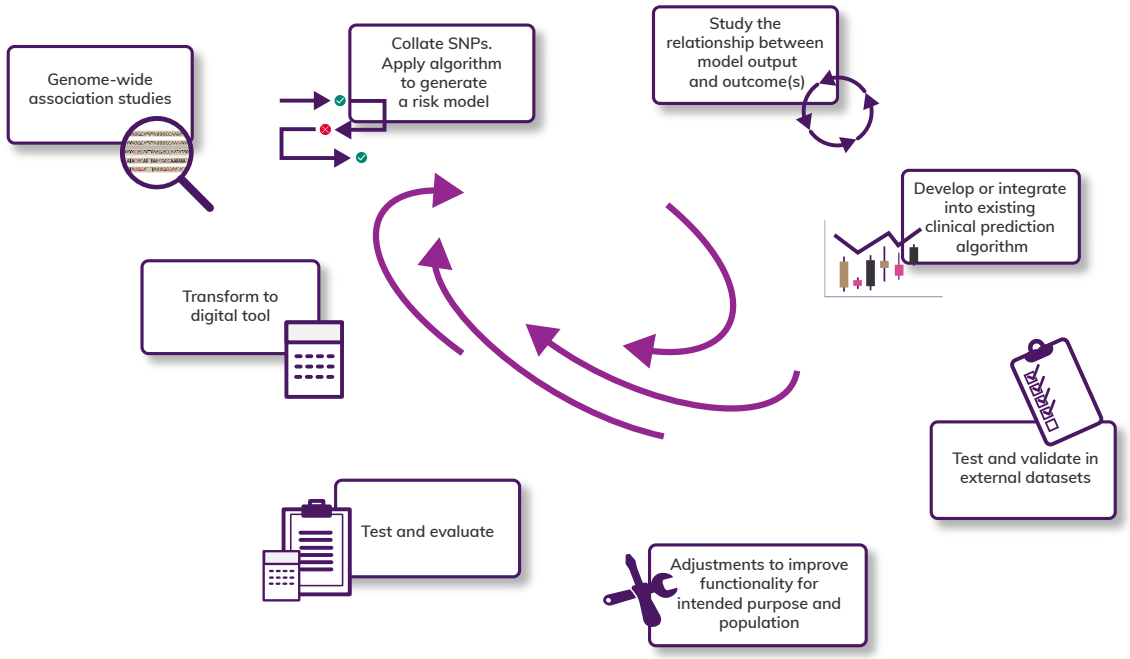
Polygenic score models developed for use in clinical settings function in a similar way to other clinical prediction algorithms or tools that collate risk factor information to estimate the likelihood of a particular outcome. In this case, they collate information across SNPs, and apply an algorithm or prediction model to estimate the likelihood of a particular outcome. In a similar way to the development of other clinical prediction algorithms, a number of steps contribute towards constructing PGS-based prediction algorithms (See Figure 4).

Basic scientific research is required to:

- Identify SNPs that can form the basis of a specific score
- Develop statistical methodologies that can be used to aggregate SNP information
- Examine the relationship between the PGS and specific traits.

On the basis of this knowledge, different risk models can be developed that may have relevance for clinical use. These may then be used as a basis to create a novel clinical risk prediction algorithm or integrate the new risk factor (i.e. PGS) into an existing clinical prediction algorithm.

Figure 4: PGS algorithm development cycle



5.3 The development of a PGS prediction algorithm

Figure 4 provides an overview of the steps in the development of a PGS prediction algorithm. We describe each of these steps in more detail below.

Identification of SNPs

Genome Wide Association Studies (GWAS) assess the association between common variants and particular traits. These studies undertake genotyping in cases and controls to uncover variants that are present more frequently in one group. They provide most of the source data for construction of PGS models; this information is used in variant selection as well as weighting.

While particular variants may be more frequent in cases or controls, it does not mean that they are causal. This is because non-causal variant(s) may be associated with a trait as a result of linkage disequilibrium. GWAS therefore may give an indication of an area where a SNP that is causal in disease is located, but often do not identify the exact causal variant. Further studies such as fine-mapping and functional studies need to be conducted in order to elucidate the exact causal SNP and its relationship to disease. For many diseases and traits, this has yet to be undertaken.

The primary output of these GWA studies is information on particular variants, the strength of the evidence of an association with the trait (p-value), an estimate of their effect size and direction of the effect. This information is used to inform the SNPs and weights used in PGS models. Significance thresholds are set in GWAS to ensure that identified associations are robust. However, it has been shown that the predictive ability of PGS models can be improved by including SNPs that fall below these thresholds.

Linkage disequilibrium

Linkage disequilibrium (LD) describes the non-random association between alleles at different loci on the same chromosome. Alleles in LD appear together more or less often than expected by chance.

Model building algorithms.

Polygenic score models are constructed using a variety of statistical methodologies to bring together information across variants and then used to calculate a polygenic score [82-84]. The main difference between various models is in the statistical methods used to choose variants and fine-tune the weights assigned to them. These statistical methods can be referred to as model building algorithms.

Model building algorithms are constantly evolving and will continue to do so. Different approaches may be more suitable for particular traits with particular genetic contributions and/or datasets. It is usual practice to develop a variety of models initially and investigate their predictive ability and the relationship between the polygenic score they produce and an outcome or trait of interest in internal and external datasets. These initial steps often test the modelling process as opposed to the final clinical prediction algorithm.

Polygenic score model: Generated using data usually from GWAS. The model is a set of SNPs and their weights.

Polygenic score: Application of a polygenic score model to calculate a number (a log relative risk) in an individual.

Development of a clinical prediction algorithm

The initial models calculate polygenic scores as relative risks and the comparator/reference used in relative risk estimation can vary. These relative risk models are used to conduct further research and develop understanding of the role of genetics in disease. If initial models are shown to be promising in research settings, they may be further validated and developed for use as clinical prediction algorithms. To enable this, initial models may need to be adapted to provide appropriate population specific risk prediction. For example, this will require consideration of the distribution of PGS and the outcome of interest in the target population [85]. The initial model may also be adjusted to enable functionality in different populations or contexts, such as different ethnicities (see next section).

Absolute risk models or integrated risk models (PGS and other risk factors) may also be constructed based on initial relative risk models and potentially using additional datasets. This means that when polygenic score models are discussed, it may be in reference to different forms of risk models along the continuum shown in Figure 5 [86].

The development, testing and evaluation processes all contribute to informing the final clinical prediction algorithm and tailoring its specific use. This may require the development of separate algorithmic elements to the 'core' PGS model. These elements can be further packaged in the form of a tool, which has additional computational elements to improve functionality and usability. For example, they may enable easy inputting of model variables and visualisation of outputs, to enable interpretation by clinicians and patients.

Adjustment for population stratification

It is well established that PGS models developed in those labelled as European ancestry populations (the majority of current models) are likely to have poorer predictive ability in populations of other genetic ancestries. It is postulated that differences between ancestries, in terms of the SNPs associated with outcomes and the strength of these associations, contribute to the differential performance of PGS models across ancestries. It is also likely that other factors, such as lifestyle, contribute to differences in predictive ability across ancestry groups [80, 87].

This presents a challenge for their wider use, as the populations in which such analyses is intended for use are often diverse and contain mixed genetic ancestry groups. The ideal scenario is implementation of models that can work across ancestries. However, this is unlikely, even when diverse datasets are used. This is because populations are not genetically homogenous entities, and social and environmental factors also shape and influence risk. There are also issues in the way individuals are currently assigned to particular ancestry groups that may cause further complexities [88].

These issues notwithstanding, statistical adjustment of models is common practice to improve functionality across different settings and populations. This approach is also being applied in the short-term to enable application of PGS models across ancestries. Options being considered are either restricting use of models to specific ancestry groups or attempting to optimise a model to function across ancestries using statistical techniques. Examples that have been employed include using principal components analysis (PCA) to adjust the results on the basis of population structure [89, 90]. Others have used methods to adjust SNP effect sizes to account for differences in population performance due to linkage disequilibrium [91].

While these measures are not perfect, they can aid in creating a model that is more applicable across populations. However, representative datasets are required to test that these adjustments work and are not giving biased predictions.

The polygenic score analysis pathway

Integration of polygenic scores into clinical practice requires robust, validated mechanisms to generate these scores. In practice, obtaining a polygenic score for an individual will involve a series of steps (Figure 5) [90]. This includes standardised processes for obtaining individual level genotype data, followed by the application of a validated prediction algorithm to that data to obtain a PGS score.

Decisions also need to be made on which, and how, the results of such analysis are reported back to clinicians and patients. PGS scores can be interpreted by themselves to determine risk of disease or be further integrated into existing or novel risk prediction algorithms. In using particular PGS algorithms, consideration will also need to be given to measures that can be taken to address differences between the population in which it was developed, and that in which it is being applied.

As discussed above, a well-recognised shortcoming of existing polygenic score models are their lower predictive performance in those not of European ancestry. However, mechanisms can be put in place to overcome this to some extent. These must be explicitly included and evaluated as part of any test pipeline.

These differing components or steps need to come together in an analysis pathway and function as a whole for test delivery [90] (Figure 5). A PGS-based test is thus much more than a risk model, comprising different elements that need to function together to provide consistent and robust results.

Key components that go into developing such a pathway are mechanisms to obtain genotype data, the best prediction algorithm to use, any adjustments to be made to improve functionality of the algorithm, the most appropriate output, and how this can be conveyed in an understandable manner. We describe each of these interlinking components below.

Genotyping

Genotype data to feed into PGS analysis can be obtained through a variety of methods such as genotyping SNP panels, microarray or next generation sequencing (NGS). 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 is of sufficient quality.

PGS prediction algorithm selection

The development of the genotyping assay for use as part of a pipeline or in considering criteria for QA of data from external sources, needs to consider the PGS prediction algorithm being employed. This is because the choice of SNPs may need to be balanced against any technical challenges in obtaining genotype data with a given method (e.g. microarray, NGS). The trend in construction and development of polygenic score models is towards inclusion of a greater number of SNPs. As noted above, the number of SNPs included for a particular disease is determined to some extent by the genetic architecture of the disease.

Construction of models includes optimising which SNP sets to include. In general, models based on a larger number of SNPs appear to have better predictive performance, although there is an attenuation of these gains with a subset of SNPs defining the greatest proportion of the association [81]. As a result, this improvement in predictive ability needs to be balanced against the benefit of the additional SNPs, as well as the marginal cost and practicalities of obtaining the extra genotype data.

The nature of PGS analysis means that there can be some flexibility in choice of SNPs. For example, if particular SNPs are difficult to genotype, an alternate SNP in linkage disequilibrium may be genotyped more reliably. Laboratories will need to make an informed decision around the genotyping assay and PGS model accounting for these challenges. The PGS algorithm that is taken forward may be selected on the basis of the feasibility of obtaining this genotype information or, conversely, the selected polygenic algorithm may determine what genotype information is collected.

Developers are creating additional software components that enable the use of PGS algorithms with different sources of genotype input data. This enables greater flexibility in the source genotype data that is used as part of the analysis pipeline [85].

Output and reporting

An additional part of the pipeline that needs careful consideration is in relation to the output of the analysis and how it is reported back to users, be they healthcare professionals or patients. The outputs and results of such analysis differ from traditional genetic analysis in several ways. The raw results of such analysis need to be converted to a risk score and there are several different outputs that can potentially be fed back (Figure 5). The reporting of absolute risk has been recommended as it is more interpretable and understandable [53]. This requires additional steps and taking into consideration the population distribution of PGS scores and disease incidence in that population.

Results from polygenic score analysis may be used by themselves or as part of an integrated risk prediction tool. In the former case, the score may be transformed and reported back as a dichotomous result (e.g. high vs. average risk) or a categorical variable [92, 93]. Where PGS information is used as part of an integrated risk tool, it is usually kept as a continuous variable.

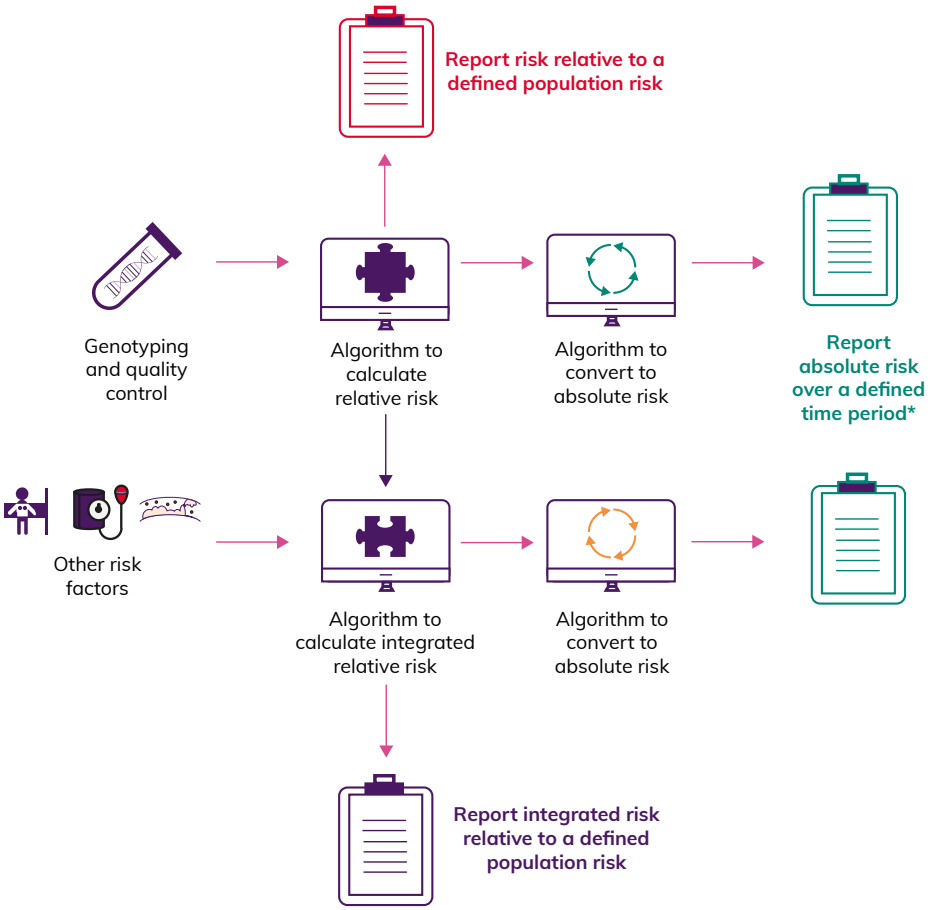
Several companies as well as research groups have developed automated computational modules or algorithms that enable conversion of genetic data into a risk score (either standalone PGS or integrated risk score). For example, IMPUTE.me is a web service developed by researchers that enables the public with access to their genetic data to upload it and carry out automated PGS analysis on a website [94].

Examples of integrated risk include the CanRisk tool and those developed by companies such as Genomics plc and Allelica. In order to be able to calculate an absolute risk, tools such as CanRisk require population specific incidence data, as well as data on the distribution of the risk factors that are part of the model. They may also incorporate population based competing mortality – i.e. the likelihood that an individual will die from another cause.

The additional data needed for risk calculation will be dependent on the nature of the absolute risk prediction model that is being utilised. Furthermore, developers have also created additional software steps that enable their models to be used with any form of input genotype data.

Figure 5: Components of the polygenic score analysis pipeline

The different components of a test pathway and options for reporting of results are shown below. *Absolute risk is the preferred option for reporting of results.



5.4 Summary

Polygenic scores differ from traditional germline genetic markers in a variety of ways. In the simplest sense, polygenic score analysis requires the use of algorithms to analyse inputted genotype data.

In a similar manner to NGS analysis pipelines, PGS analysis pipelines can be configured in different ways, and the elements included as part of this pipeline may differ. However, at its core is a clinical prediction algorithm, the nature of which differs depending on whether polygenic scores are intended to be used by themselves or as part of an integrated predictive algorithm.

Existing risk prediction algorithms that combine information on a variety of risk factors can be adapted to incorporate a polygenic score, or new algorithms can be developed that calculate risk based on a polygenic score and other variables. Separate assays may need to be conducted to obtain information on these additional variables, such as a cholesterol test, a blood pressure test and a family history assessment.

The focus of current reporting is on development of polygenic score models or integrated risk models and examination of their predictive ability for a particular trait. Whilst model development is a pre-requisite for test development, our analysis indicates that the progress being made in this area is often conflated with availability of a validated test. While these are not mutually exclusive processes, evaluation of particular modelling approaches should not be considered the same as the evaluation of a particular test or application.

Our previous reviews of the field identified that research progress for polygenic score-based applications across different disease areas is at various stages with respect to the development of a test, whether it is a discrete test or part of an integrated risk tool.

Currently, different approaches have been developed or are being investigated for delivery of this testing pathway. Furthermore, the computational elements of PGS analysis can be further packaged in the form of a tool, which has additional computational software elements for analysis of the genetic data and to improve functionality and usability. For example, they may enable use of a PGS algorithm in different populations or settings, or simplify inputting of model variables and visualisation of outputs, to enable interpretation by clinicians and patients.

Achieving clarity as to what elements form or contribute to any particular pipeline or testing strategy is useful in considering how to approach validation of these components and the pathway as a whole. In addition, as with NGS services, the infrastructure through which either the whole or elements of this pipeline are delivered is also likely to influence approaches to validation and evidence assessment thresholds.

Regulation of PGS analysis

6. Regulation of PGS analysis

Compliance with regulatory requirements is a prerequisite for a product to be made available on the market and for implementation within health services. It is therefore an important element of test evaluation and informs the assessment of clinical utility and test implementation from a health service perspective [1]. In this chapter we consider current regulatory frameworks, their evidence requirements in and some of the challenges this raises for products that provide a polygenic score.

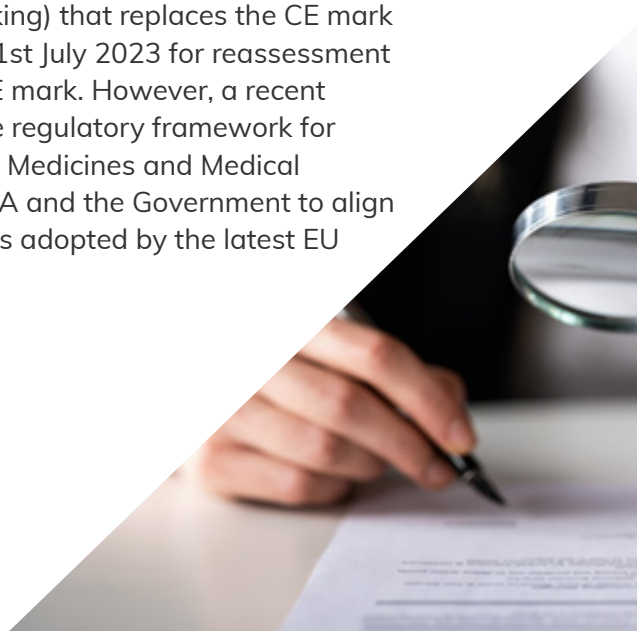
6.1 The types of regulation that may impact on tests

Many types of regulation impact on the development of assays and tests. These can apply to the test or assay itself, the setting in which the test is developed, or the expertise of the user offering or administering the test. The nature of regulatory scrutiny may also vary depending on the nature of an assay or tests, how they are provided or conducted, and their intended purpose.

The key regulations applying to the development of assays and tests are EU and UK medical device regulations. The application of these regulations is not straightforward. They can apply both at the level of the technology platform used to support the administration of an assay (for example a genomic sequencing platform), and/or the specific test to be utilised for a specified purpose within a population (for example, a specific genomic test used for diagnosis of a genetic disease in a given population). Meeting these regulatory criteria can be a complex process, and similar to the evaluation process for healthcare implementation requires test developers to consider the intended purpose and nature of the test.

6.2 Medical device regulation

When a product is placed on the market in the EU, it must have CE marking, namely accreditation that it has met appropriate levels of quality assurance, manufacturing practice and is fit for its intended purpose. In the EU for medical devices, this is governed by EU Medical Devices Regulation 2017/745 and EU IVD Regulation (IVDR) 2017/746. However, since 1st January 2021, Great Britain (GB) has relied on the 2002 Medical Devices Regulations that were based on previous EU Directives, through a bespoke Conformity Assessed mark system (UKCA marking) that replaces the CE mark [95]. In parallel, a grace period is currently in place until 1st July 2023 for reassessment of devices for the GB market that already have an EU CE mark. However, a recent MHRA consultation has proposed further changes to the regulatory framework for medical devices in the UK using powers contained in the Medicines and Medical Devices Act (2021). This signalled the desire of the MHRA and the Government to align with international best practice and many of the changes adopted by the latest EU Regulations.



Across these regulations, there is consensus that the primary way in which a device might qualify as a medical device is if it is intended for medical purposes. The interpretation of medical purpose is broad, and most tests used within health systems will qualify either as medical devices or *in vitro* diagnostic (IVD) devices, depending on the extent to which it is driven by data obtained *in vitro* by the examination of blood and samples.

6.3 Clinical evaluation as part of regulatory frameworks

The evidence appraisal for regulatory approval takes the form of a clinical evaluation report following critical evaluation of the relevant scientific literature and available clinical investigations, along with consideration of alternatives. There are synergies in the terminology, approaches and evidentiary standards of clinical evaluation reports with those of test evaluation. For example, clinical evaluation of medical devices requires evidence of scientific validity, analytical performance and clinical performance. These terms are analogous to those described earlier in this report in the context of test evaluation and clinical validity. Scientific validity refers to the association of an analyte to a clinical condition or physiological state and analytical performance refers to the ability of a medical device to correctly detect and measure a particular analyte. Clinical performance refers to the ability of the device to yield results that relate to a particular clinical condition or physiological state for the intended use and in accordance with the target population, and applicable to the intended user. In the context of the IVDR, the level of 'sufficient clinical evidence' required depends on three main factors*. First, the intended use of the device; second, the evaluation of interferences and cross-reactions; third, the acceptability of the risk benefit-ratio. This illustrates that evidence requirements are often proportionate to the nature and intended purpose of the device.

Evidence requirements for regulatory approval also faces many of the challenges as test evaluation, especially in the context of digital technologies, which have multiple components [6, 41]. In particular, the use of algorithms and software within devices or tests creates additional challenges and may also require developers to demonstrate that specific standards have been met.

One such challenge is deciding what constitutes 'a device' for the purposes of regulation, since software may be regulated as part of a device or as a separate device in its own right. Indeed, the definition of an in vitro medical diagnostic device in the *In Vitro* Diagnostic Medical Devices Directive (IVDD) uses the words "alone or in combination"† indicating that regulation can apply at the level of the individual components, or the whole with a key consideration being the interoperability of these components. While the Directive does not define 'software' as such, its Annexes have been amended and contain further provisions clarifying how algorithms and software should be 'validated in accordance with the state of the art taking into account the principles of the development life cycle, risk management, validation and verification'‡.

* Article 56(1), Regulation (EU) 2017/746.
† Council Directive 98/79/EC of the European Parliament and of the Council of 27 October 1998 on *in vitro* diagnostic medical devices [1998]. OJ L 331/1
‡ Council Directive 93/42/EEC of 14 June 1993 concerning medical devices [1993] OJ L 169/1

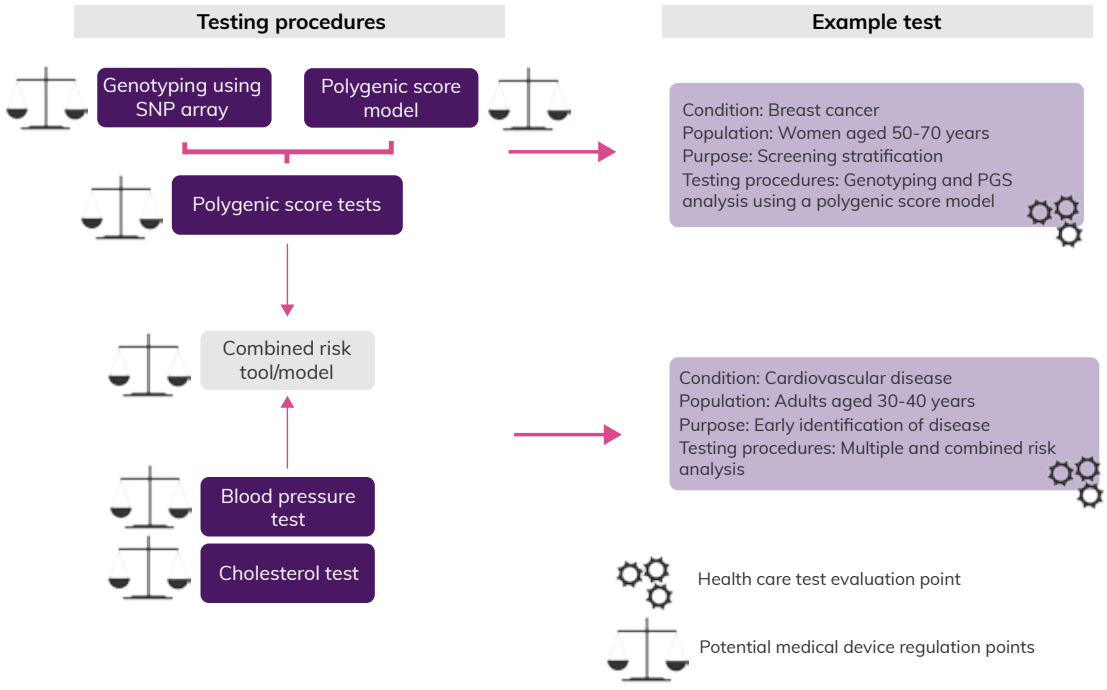
Another challenge is the potential for artificial intelligence or machine learning. Machine learning algorithms recognise and apply patterns in training data to new datasets to generate novel findings. However, the dynamic, highly adaptive nature of such algorithms means they are often opaque and increasingly intractable to traditional regulatory approaches. Machine learning algorithms are therefore prompting new regulatory approaches such as regulatory sandboxes and utilising synthetic data.

Where interventions include multiple different elements, commercial rights over each element in the form of intellectual property rights or trade secrets may inhibit transparency. The complexity of these systems, together with the cluster of rights over each element, creates additional regulatory challenges and uncertainties. In the next Section, we describe the specific challenges for PGS analysis.

6.4 Applying medical device regulation to PGS analysis

The application of medical device regulation to PGS analysis is ambiguous. Under current EU and UK medical device regulations, depending on how they are used, the target population, and the healthcare pathways involved, each of the steps in PGS analysis pipelines could be viewed as generating a discrete medical device (or IVD), itself subject to independent regulatory oversight. Alternatively, these components could be regarded as a single device with multiple components which must work interoperably (Figure 6).

Figure 6: Potential regulatory impact points in the development and use of polygenic scores, models, tests and tools.



One source of ambiguity is whether or not different elements of this pathway could qualify as an IVD medical device, as opposed to a medical device. This is because it is currently unclear how proximate to the analysis of a sample, or how heavily driven by that analysis, PGS analysis software and algorithms are required to be for qualification as an IVD medical device.

The latest guidance from the MHRA suggests that PGS analysis software may qualify as an IVD medical device if it is 'substantially driven' by IVD results, unless those are the results of 'historical' investigations that are 'unrelated to the software'.§ This may impact the nature of evidence required to demonstrate the safety and efficacy of the device(s).

Another ambiguity relates to the classification of the medical device or IVD medical device. There are four classes of medical devices assessed according to the risks arising from the intended purpose, but apps and software will generally fall into one or two classes:

- Class I - generally regarded as low risk
- Class IIa - generally regarded as medium risk.

If considered to be an IVD device, at present the UK medical devices regulations enable a high proportion of IVDs (~80%) to be placed onto the market on the basis of self-declaration.

The updated EU rules apply a risk-proportionate approach (more akin to medical device classifications) with four classifications from A to D (low to high risk). Class B is the default class for most IVD devices, but it is feasible that PGS-type devices would fall into Class C where there is a higher risk of harm, depending on the conditions involved and the influence of the output of the device on clinical decision making or screening programmes.

Under the new EU rules, all IVD devices that are intended for 'human genetic testing' will automatically fall into Class C¶. The MHRA and Government in the UK have signalled an intention to update the IVD classification rules in line with this risk-proportionate approach. However, devices intended for human genetic testing will not automatically be classified as medium-high risk. This will only be the case where there is a risk that an erroneous result could lead to a serious adverse event. The classification of a medical device or IVD device is crucial in setting the expected nature and level of evidence that will be sufficient to demonstrate safety and efficacy.

§ MHRA Guidance: Medical device stand-alone software including apps (including IVDMDs) p20: https://www.gov.uk/government/publications/medical-devices-software-applications-apps

¶ Regulation 2017/746 on *in vitro* diagnostic medical devices Annex VIII section 2.3(i)

6.5 Evidence requirements for PGS applications

Evidence requirements will be dependent on the nature of the PGS analysis and how they will be used for clinical or public health purposes. It is important to note that this evidence does not necessarily incorporate the full range of factors that may be relevant to assessing 'clinical utility'. For example, under the EU IVDR, clinical benefit (as opposed to utility) is defined as [lying in] 'providing accurate medical information on patients, where appropriate, assessed against medical information obtained through the use of other diagnostic options and technologies, whereas the final clinical outcome for the patients is dependent on further diagnostic and/or therapeutic options which could be available.'** Health economic factors, or those involved in health technology assessment, and clinical outcomes are excluded from the scope of clinical benefit.††

However, as outlined above the general requirement is for evidence of analytical or clinical performance (proportionate to the risk classification of the device) to demonstrate that the device performs in accordance with the generally acknowledged state of the art. Such requirements should enable regulators and notified bodies assessing devices to ensure that PGS devices meet appropriate performance standards, including demonstrating equity of impact and minimisation of bias according to the target population and the intended purpose.

Overall, the regulations require:

- A target patient group and intended purpose to be clearly specified and for scientific validity and analytical and clinical performance to have been clearly demonstrated for that population
- A balancing exercise that weighs risks (including those relating to bias) against potential benefits to the patient
- Vigilance through the lifecycle of the device, including capturing adverse events through post-market surveillance.

The expected reforms to medical device regulation in the UK, updating the framework and bringing it into line with international best practice, should enable a proportionate approach to different PGS devices, depending on their intended purpose. The challenge lies in further specifying when and in what form evidence is required for PGS devices, and ensuring consistent interpretation of the rules by manufacturers, notified bodies and the Regulator across the sector. This will require soft measures, including guidance and recommendations for PGS (or sub-categories of PGS), which both developers and those scrutinising new devices can follow, as opposed to changes in hard law and regulation.

** Recital 64 IVDR

†† MedTech Europe. Clinical Evidence Requirements for CE certification under the *In Vitro* Diagnostic Regulation in the European Union. 2ND Edition. November 2021.

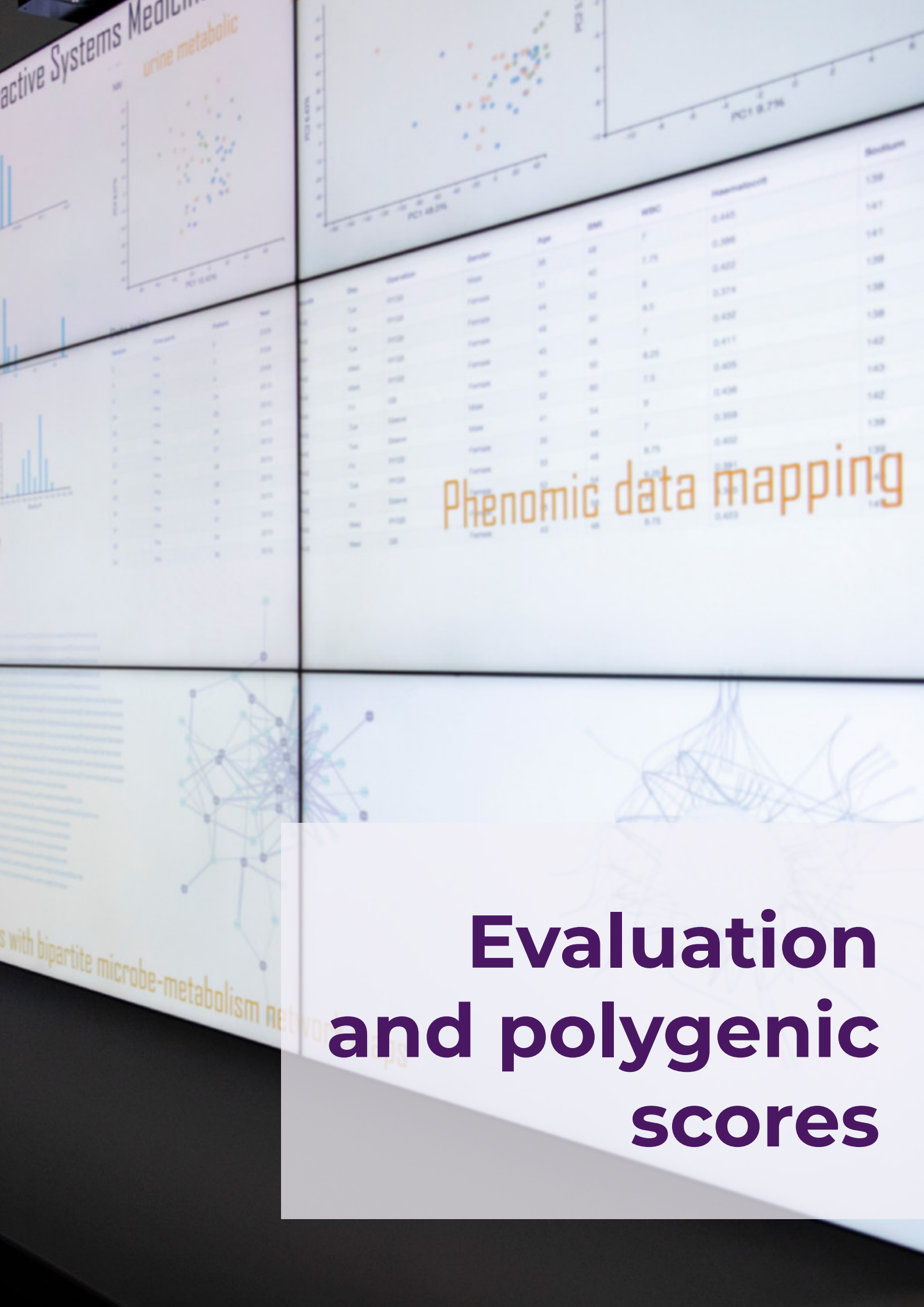
6.6 Summary

Regulations are important in ensuring safety, effectiveness and efficacy of tests and apply at multiple levels. The evidence that is gathered for regulatory approval is analogous to that required in medical test evaluation, especially when the intended purpose and population of a device are synonymous with that of a test.

Challenges exist for evidence generation in a regulatory context that are similar to those of test evaluation, with particular challenges in demonstrating clinical performance, due to the nature of PGS analysis as a device.

Drivers of this include the multiple components of PGS analysis, the status of algorithms or software within the device, the nature of potential users, and the overarching intended use and purpose.

Standards for evaluation of prediction algorithms and digital tools are less well developed, which has created issues for developers in determining both the studies that need to be undertaken to generate evidence, as well as the level of evidence required. An understanding of approaches to evaluation and evidence assessment of PGS-based applications or tests is needed to better inform regulatory evidence requirements.



Evaluation and polygenic scores

7. Evaluation and polygenic scores

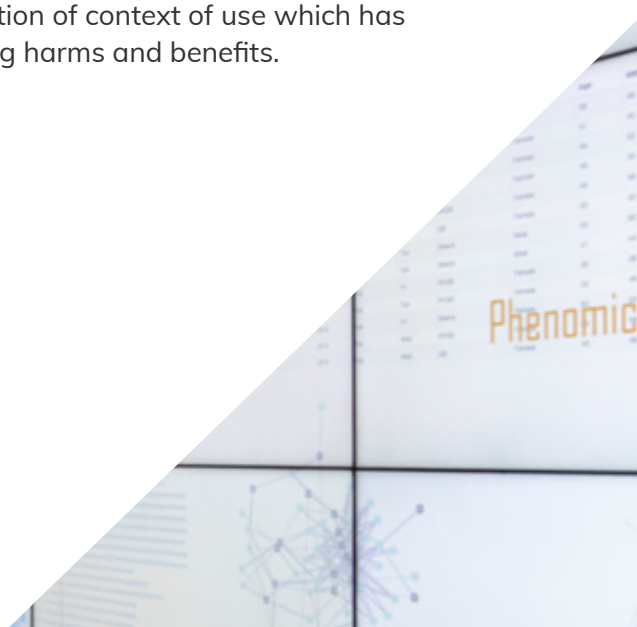
The previous chapters explored key concepts that contribute to evaluation. In this chapter we present our analysis, which brings together and builds on these elements and examines the types of evidence that can potentially contribute to the evaluation of products that incorporate a polygenic score. In doing so, we also consider the extent to which these principles can be applied to such products and describe the uncertainties. Our aim is to explain these uncertainties, in order to assist in the decision-making on which are the most important factors to consider prior to implementation of specific PGS applications for healthcare.

7.1 Terminology

Differences in definition of particular terms across fields and in common parlance can cause confusion. To achieve some clarity and enable application of test evaluation frameworks we have used the following definitions.

- **Assay:** A method for determining the presence or quantity of a component (e.g. genotyping and measurement of other risk factors)
- **Risk model:** Mathematical representation of a clinical situation able to predict different outcomes
- **Clinical prediction algorithm:** Derived from risk models and the "product" implemented in clinical practice to calculate a risk score
- **Tools:** Can take a variety of forms and are the mechanism through which prediction algorithms are utilised to enable individual level testing and risk scoring. They do so by enabling inputting of data for algorithms to function and presenting outputs. Algorithms and tools can have the potential to be used in different contexts and for different purposes
- **Test:** The use of a particular assay, clinical prediction algorithm or tool in a specific population for a specific purpose can be described as the test.

The above specific definition of a test enables consideration of context of use which has implications for evidence requirements and in considering harms and benefits.



7.2 Issues in evaluating polygenic score applications

The aim of medical test evaluation is to better determine if particular tests have value, and evidence supports their implementation within practice. Our previous report discussed the concept of clinical utility and has been summarised in earlier chapters of this report. Another important component for consideration prior to implementation and that feeds into considerations of clinical utility is clinical validity.

Clinical validity refers to the predictive ability of a test in a defined population for a particular purpose [9]. It is predicated on showing the association between biomarker and disease (scientific validity) as well as the ability of an assay to detect the biomarker (analytical validity). This is followed by demonstrating the predictive accuracy or test performance using prospective studies to establish test performance characteristics such as sensitivity, specificity, positive predictive value (PPV) and negative predictive value (NPV). Ideally, these prospective studies would be conducted mirroring the clinical use scenario, thus taking into consideration the setting, population, role and purpose.

Assessment of these parameters enables decision makers to consider the implications of test use and see if they are sufficient for a particular use case. There is generally an absence of a set threshold for test performance parameters, and the exact parameters that are examined, as these are influenced by the context and purpose of the test.

A key point is that clinical validity and utility are intricately linked to the setting, role and purpose of a test. Furthermore, it also requires clear articulation of the testing system or strategy. Therefore, in the context of PGS applications, the ideal scenario for assessing clinical validity would involve evaluating test performance of a defined PGS analysis pathway, in a prospective manner.

Such studies would utilise the index test at enrolment (i.e. PGS analysis) and individuals would then be monitored to see if they develop the outcome of interest using a reference standard test. However, this is challenging for several interrelated reasons.

Firstly, polygenic score analysis is a mechanism for risk assessment. Where this is already an established part of clinical practice, e.g. the use of QRisk® for cardiovascular disease prevention, the comparator or reference standard is clear. In this situation, the question may be more in relation to incremental improvements to an existing clinical prediction algorithm and associated tool through the addition of new data.

In contexts where risk assessment is not an established part of clinical practice, this can be more problematic due to uncertainties about whether risk assessment would be beneficial, and whether there is an appropriate comparator.

Secondly, many diseases for which polygenic score analyses are proposed take a long time to manifest. The follow-up of individuals for long periods of time to obtain information on the outcome is often not practical and is unlikely to yield results in a timely manner for decision makers. Finally, the extent to which prediction algorithms are considered as a test for evaluation purposes varies.

Evaluation is further complicated by the fact that polygenic score applications straddle the fields of genetic testing, prediction modelling and digital health technologies. This is because they encompass elements of molecular testing to obtain genetic data and a prediction model for analysis of this data. The latter may occur in two steps, depending on the application. For example, scores may be generated and interpreted by themselves or be generated for integration into an existing or novel prediction algorithm.

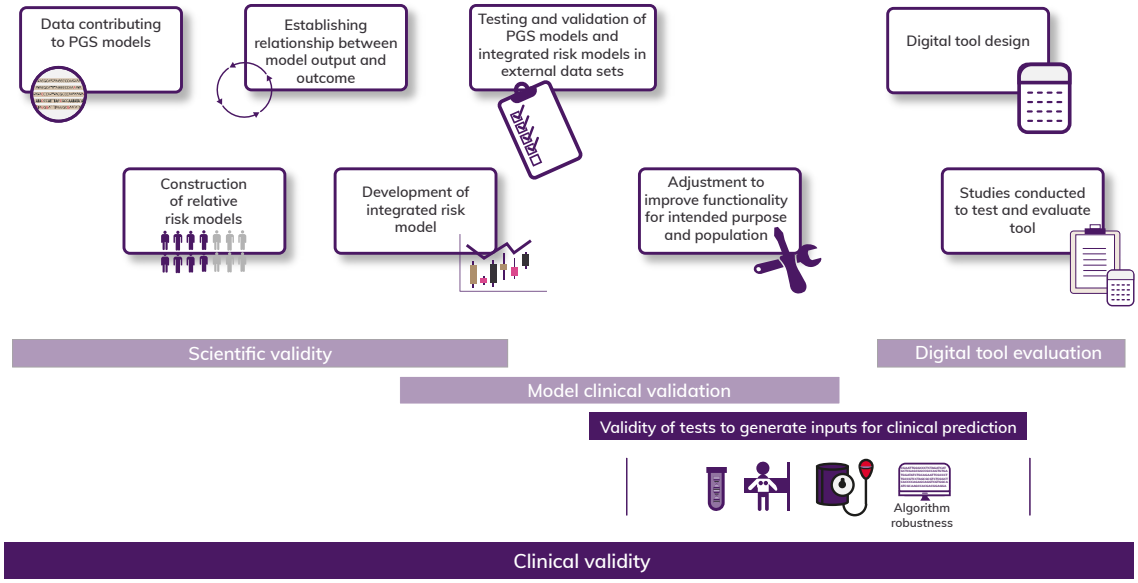
In addition, prediction algorithms may be part of digital tools for ease of use. This means that the composition of polygenic score analysis pathways can take a variety of forms, as described in Chapter 5. This can result in uncertainty regarding which elements to evaluate in terms of the process that produces the PGS or subsequent integrated risk score and the level of evidence required in support of each of these elements.

We have addressed this ambiguity by applying specific definitions, to enable the application of diagnostic test evaluation frameworks (see above) to this pipeline. Further, we have addressed the complexity of polygenic score analysis straddling different fields by:

- Approaching polygenic score analysis (whether standalone or integrated) as a pipeline composed of different elements
- Applying the relevant elements of test evaluation and/or prediction model evaluation to each of these elements
- Considering evidence generation in relation to the individual components as well as the whole system.

Taking this approach can provide clarity by cross-linking across evidence generated in relation to specific testing procedures (e.g. genotyping, cholesterol testing, blood pressure etc.), any prediction algorithms that are applied to this data and digital tools used to deliver the results of such analysis. In addition, it allows consideration across different parameters that contribute to decisions of clinical validity (Figure 7). We discuss each of these aspects in more detail below, along with the current issues in evidence generation and assessment.

Figure 7: Elements of evidence generation that need to be considered and linked.



7.3 Scientific validity and polygenic scores

Scientific validity is a key component of clinical validity. In relation to molecular genetic tests, this involves examination of the evidence supporting the relationship between the biomarker and a trait of interest. As described in Chapter 3, current frameworks used for demonstrating scientific validity of germline biomarkers cannot be applied to PGS analysis. This is because PGS analysis functions as a prediction model, and in this context, it is important to understand the evidence base supporting the relationship between the score generated by the model and a trait of interest. Therefore, scientific validity will require assessment of different aspects of the model construction and validation process, including:

- Data contributing to PGS models
- Assumptions made in building models
- The evidence supporting a relationship between a PGS and outcome
- Validity of a prediction model.

We expand on these points below and reflect on how they impact on interpretation of information from polygenic score research.

Data contributing to PGS models

A key consideration in prediction modelling is ensuring that the data used in model construction accurately reflects the population in which the risk prediction model is going to be used [25, 26, 38, 73]. Differences between the source data and target population can contribute to lower predictive performance.

A widely discussed issue is the fact that much of the data used in constructing PGS models are derived primarily from populations of European ancestry and may not be relevant or valid for other ancestries.

Other factors such as the trait measured and the age of the population in the GWAS may also be relevant and have an impact on the predictive accuracy of the PGS clinical prediction algorithm in other populations. This is because any errors or uncertainties in this data will be carried through in the model and impact on its results.

In assessing the scientific validity of particular PGS models, it is important to consider the GWAS data that has been used in its development. This includes assessment of the quality of the data source and whether the associations identified have been independently replicated. This can enable assessment of whether the source data used in construction of a model is relevant and valid. Standards exist in conducting and reporting results of GWAS, and resources such as the GWAS catalog enable assessment of the validity of these studies [96].

Assumptions made in model building algorithms

The validity of polygenic score models is tested at different stages throughout the development process and each step can contribute some information towards functionality in a clinical setting. The initial step tests different model building algorithms.

There is ongoing discussion on the relative merits of different model building approaches within the scientific community. These discussions indirectly affect considerations with regard to establishing scientific validity, in terms of biomarker disease relationship, as they make assumptions on how to bring together data from GWAS and the influence of SNPs on disease. As noted previously, it is now common practice in some model building approaches to include SNPs that do not meet significance thresholds for association in GWAS. This has been shown to improve predictive performance in some instances, however, some have questioned this approach [97].

Establishing the relationship between a PGS and outcome

GWA studies provide evidence for an association between individual SNPs and an outcome. The investigations described below help explore and establish the relationship between a PGS calculated by a particular model and an outcome. These are important to investigate, to establish which modelling approaches might work ‘best’ in bringing together SNP data, as well as in determining its potential usefulness for prediction.

An understanding of the relationship between PGS and an outcome is investigated using relative risk models. For example, scores are standardised and their distribution in cases and controls is examined to assess potential predictive capabilities. In most instances there is considerable overlap between the distributions of cases and controls. Therefore, it is usual practice to compare individuals who are considered high-risk versus those who are average or low-risk [98, 99].

The relationship between polygenic scores and particular outcomes may also be examined by performing regression analysis. This is usually conducted using different datasets to that used in model development. Key parameters that are assessed include measures of correlation, discrimination (e.g. AUC), calibration, proportion of variance in a trait explained and effect size estimates.

Examination across these metrics can give an indication of the relationship between the score and a trait, and its potential to have utility in clinical practice. For example, association testing gives some indication of the relationship between the polygenic score and trait of interest. Examination of the proportion of variance explained and AUC can give an indication of the predictive ability of the PGS. These associations need to be replicated in an independent dataset.

External validation is key in determining the functionality of any model in datasets outside those from which it was produced. While initial studies can provide support for a particular modelling approach and give indications of the predictive potential, external validation studies are key in assessing whether this can be achieved using alternate datasets.

Determining the validity of PGS prediction models

The investigations described above can also be used as a basis to select a particular relative risk model to take forward for clinical prediction. If integrated risk models are developed, they will need to be evaluated and validated following the processes described in Chapter 4.

As with the process above for SNPs, additional biomarkers are often selected on the basis of available data and their relationship with the outcome. It is also common practice to examine any correlation between the variables to be included in a model. Different approaches may be taken in building integrated prediction models, including assumptions made in the way the variables in them are brought together.

Assessment of validity of these models in internal datasets and then external datasets, enables understanding of their predictive potential. Where polygenic scores are included as part of existing integrated risk algorithms such as QRisk® [100] BOADICEA [101], these models undergo testing to examine performance with and without the novel biomarker.

Key gaps in understanding scientific validity

As may also be seen in other scientific fields, polygenic score model research and development includes both knowledge driven research efforts and those that are more application driven. Research that aims to advance scientific understanding may be focused on the modelling processes or using models to investigate the role of genetics in disease. More applied research will focus on the development of a product in the form of a prediction algorithm. The fact that all of this falls under the broad umbrella of polygenic score research, can lead to confusion amongst non-experts when examining the literature on PGS models, particularly in linking and assessing some of the elements described above.

Currently, making a judgement of the potential of particular models to function as clinical prediction algorithms is not a simple task. Assessing the parameters set out above is made difficult by inconsistencies in the way results of such investigations are reported and the lack of standardisation. A guideline has been published with the aim of improving reporting standards of prediction models [40]. A specific standards guideline for the reporting of polygenic scores is also available [52].

Adherence to these guidelines is not common practice, creating difficulties in determining whether models can be considered relevant and valid for use in clinical practice. In particular, it can be difficult to understand if 'model validation' is in relation to model building techniques, a relative risk model or an absolute risk model. It can also be difficult to establish whether validation is in external cohorts or as part of clinical validation.

7.4 Considering analytical validity aspects of a PGS test

Analytical validity defines the ability of the assay to correctly detect and measure a biomarker. In the context of PGS analysis pipelines, the analytical validity aspects can be considered in two steps. The first step is restricted to the laboratory components used to obtain genotype data. The second step assesses the ability to obtain a robust score using different combinations of methods to obtain genotype data and their use in a specific PGS clinical prediction algorithm, i.e. the ability to reliably generate the intended output based on the data provided.

Obtaining genotype data

Genotyping can be a relatively straightforward undertaking, especially when data on a few SNPs (e.g. 20) is required but can be more complex when it is across a larger number (e.g. hundreds of thousands). This is because if data is required on a small number of SNPs, it can be obtained using standard techniques such as genotype (SNP) panels or arrays. However, if data is required across a large number of SNPs, other methods may be more appropriate, such as next generation sequencing or the use

of genotyping arrays and imputation. In addition, challenges may arise in obtaining accurate genotype data if, for example, a proportion of SNPs are located in regions that are difficult to genotype or are not amenable to genotyping using particular types of assays.

Imputation describes the process of determining SNP variants that have not been directly genotyped in a sample of individuals, but are statistically inferred (imputed) based on knowledge of the haplotype (i.e. areas of high LD) from a reference sequence.

Evaluation of analytical validity of PGS analysis pipelines will be simpler where the genotyping assay captures all SNPs needed by a particular prediction algorithm. However, this is unlikely to be the case in practice, especially where genotyping needs to capture data across a large number of SNPs and imputation may be required. It is therefore important to assess performance of the genotyping assay across a number of samples to identify any issues in obtaining genotype data.

A further consideration is that arrays may perform differently depending upon the population, for example, there is evidence of shorter linkage disequilibrium (LD) in individuals of African ancestry [102]. As discussed, there are currently options and considerations for trying to improve the performance of polygenic scores across all populations; this may affect the SNPs that need to be genotyped to allow for these corrections to be done.

Decisions around genotyping assays may be dependent upon the choice of the specific PGS prediction algorithm to implement into the clinical service, and its purpose. The latter is likely to dictate how much precision is needed with respect to the scores and therefore the robustness of the score that is generated by the pipeline.

Data acquisition and robustness of polygenic scores

The combination of genotype method and model can have an impact on the robustness of the score. The use of genotyping arrays and imputation versus NGS has implications for the quality of genotype data. This in turn can have consequences for the score that is calculated, and how individuals are classified. Thus, as demonstrated by Hao et al, as part of analytical validation it is important to determine if assay choice will have an impact on the score and subsequent classification of individuals [90]. This may also inform key decisions in PGS assay development, especially if risk classification is not impacted by choice of genotyping method. For example, genotyping arrays are less costly than NGS, therefore the need for 'perfect' data may need to be balanced against the cost of obtaining it.

There is limited evidence available to assess how different PGS models perform when paired with different genotyping assays. Other factors may also affect the choice of genotyping method; for example, if the data will be used to generate more than one PGS or for other purposes (e.g. pharmacogenomics).

Key gaps in understanding analytical validity

In many research settings, predetermined genotype data is used for all constituent components of PGS model development, performance evaluation and validation. This includes basic scientific research to identify variants, development of the model, and testing of the model. This research is therefore not able to account for and assess the impact of issues in obtaining clinical grade genotype data [100].

There is still incomplete or inconsistent reporting on how missing data in biobank genotype datasets might impact on model development and performance and therefore outputs of PGS analysis at an individual level.

Research studies utilise imputation and other data science tools to minimise the impact of missing data, as a consequence of SNP target failures. Such processes need to be tested in clinical settings to ensure consistency in obtaining genotype data for PGS algorithm functioning. For example, it is not clear how much error imputation introduces at an individual level, and whether this is likely to impact on a PGS score [103]. The analytical validity also needs to account for inclusion of any steps as part of the pipeline to improve the functionality of models outside of the population(s) in which it was developed.

There is currently very little information in the public domain on the development of PGS analysis pipelines and assessment of their analytical validity. Those that have developed such pipelines acknowledge that it is not a trivial undertaking and requires balancing of various factors [90].

A key challenge in the development of a clinical assay is selection of an appropriately validated algorithm. The availability of numerous models combined with the lack of standardisation in reporting can hamper determining the scientific validity of models and their suitability for use as a clinical prediction algorithm. However, once the PGS algorithm has been selected, determining the analytical validity parameters for a PGS assay is comparatively straightforward, as there are established processes for genetic test validation [104].

A parameter that can be useful in developing the pipeline and determining analytical validity, is understanding the minimal acceptable SNP data for particular PGS prediction algorithms to function.

Currently, it is not clear what these parameters need to consider with different approaches likely to be taken by clinical laboratories. One point to consider would be that SNPs within the score are not weighted equally. This may mean that minimum parameters for accurate genotyping of specific SNPs known to define the greatest component of risk within the model are needed.

7.5 Establishing the clinical validity of models

The core of polygenic score analysis is a clinical prediction algorithm developed on the basis of risk modelling. Therefore, clinical validation of these algorithms is an important contributor to understanding clinical validity of the test pipeline.

As described in Chapter 4, clinical validation of a prediction model provides information on a variety of statistical parameters that can inform decision making on the performance of the model and its utility. Ideally, this is done through external validation using datasets that are similar or the same as the population in which the prediction model will be applied.

Several studies report on the external validation of polygenic score models or integrated risk models. However, as noted in Chapter 4, the relationship between external validation and demonstration of clinical validity may not always be apparent.

Uncertainties still exist in the interpretation of the findings of external validation studies and in determining whether a particular model is clinically valid. These issues are not specific to polygenic score models and there is ongoing development of guidance to aid decision makers in appraising prediction models [68, 105].

A particular issue in conducting external validation is in identifying appropriate datasets in which to assess the model. This is more of an issue for integrated risk models, because alternative datasets which contain all the model variables are needed. If models contain relatively new biomarkers such as PGS, or larger sets of variables, it can be difficult to identify datasets which contain information on all the variables in the model. For validation of PGS models, most often existing research cohorts are used in external validation, as this provides researchers with easier access to datasets in which to assess their model.

The use of existing datasets to validate models is pragmatic and has the advantage of enabling rapid assessment of a prediction model. However, the information from such studies needs to be interpreted with care in order to gain a clearer understanding of what this means in relation to the use case population. This is especially important if there are differences between research cohorts and the true use case population that could impact on interpretation of test performance measures. For example, the UK Biobank has been used to validate many polygenic score models, however, it is widely recognised that this particular data set is not representative of the general UK population.

The lack of ethnically diverse population cohorts is also an issue for clinical validation of models, and is contributing to uncertainties in how to interpret PGS information for those of different ancestries. The use of research cohorts also means that whilst many risk prediction models can provide outputs of 5-year, 10-year or lifetime risk, for pragmatic purposes external validation will only be for the maximum follow-up time of the cohorts, resulting in shorter term outputs such as 5-year risk being assessed.

Furthermore, research has shown that while algorithms used to calculate polygenic scores function equally well in terms of their predictive abilities at a population level, there can be variations in the calculation of polygenic scores at the individual level [80, 106]. This in turn can potentially affect how individuals are classified [107]. Differences in the datasets used to construct models or the way they are constructed may contribute to this. It has been proposed that guidelines are needed for constructing models to minimise these differences [107].

Key gaps in establishing clinical validation of risk models

As with other areas of prediction model research, there have been relatively few reports of comprehensive and high-quality clinical validation studies for PGS models, whether they are standalone or part of integrated risk models. Where they are available, appraisal of model validation studies requires consideration of whether the study populations are relevant and the degree to which they match the use case populations. This allows better assessment of the statistical parameters that are reported as part of this process.

Another key consideration is whether clinically relevant thresholds have been used in classifying people into different risk categories [74]. For certain applications, there may be specific clinical guidelines that set thresholds, which can be used in model validation. However, in many instances guidelines are not available, especially those that are based on PGS information alone.

While external validation is key in evaluating risk prediction models, it is unclear to what extent this needs to be demonstrated prior to implementation. External validation studies can be conducted across one or more datasets. If models are shown to have comparable performance across a range of datasets, it may provide greater evidence in relation to their robustness and generalisability, which may have implications for clinical implementation. There are currently no established standards for the level of external validation required to establish clinical validation. For instance, do such studies need to be conducted using datasets of a particular sample size, do they need to be replicated across different datasets and if so to what degree [69]?

7.6 Clinical validity of a test pathway

In comparison to model validation, pipeline validation or validation of a test pathway is often not explicitly addressed as part of the process to demonstrate clinical validity. In part, this is because different approaches have been taken in constructing this pipeline. In doing so, assumptions have been made in relation to analytical and scientific validity parameters.

Model validation studies often assume analytical validity and data acquisition processes can impact on the robustness of the pipeline. This may or may not be a significant issue and is dependent on the use case. For example, if information is used in decision making around certain behaviour changes, the degree of precision needed may not be as great as for a decision relating to invasive procedures, which have potentially greater immediate harms.

Often the rationale in selecting particular models is not made explicit. Validation of the whole analysis pipeline is a key step, as it will provide more information in relation to the 'real-world' functionality of the test system, however it is configured.

As prediction models are often used to develop new digital tools or incorporated into existing ones, they may require the assessment of additional parameters to ensure they are suitable for intended use. This includes usability and ensuring the format of the outputs enable appropriate interpretation by end-users. If healthcare provider and/or patient interactions with such tools are not appropriately considered, it may result in the development of devices that are incorrectly used.

Trials are underway that address some of these evidence requirements. For example, the HEART trial for cardiovascular disease recruited 1000 healthy volunteers aged 45-64 to assess the feasibility of incorporating PGS into existing risk assessment for cardiovascular disease [108]. The GenoVA study in the US is a RCT that aims to determine the clinical effectiveness of PGS to identify individuals at high risk for a variety of common diseases [109].

There are also trials such as MyPEBS, WISDOM and Perspective that are comparing standard breast cancer screening pathways to those that incorporate risk assessment (including PGS) [110-112]. As part of these trials, protocols and assays for obtaining and analysing genotype data in a clinical setting to provide a PGS for use in integrated risk assessment are being developed. Thus, they can provide additional evidence relating to the analytical, model clinical validation and test system validation parameters.

7.7 Summary

Polygenic score analysis pathways bring together elements of molecular testing, prediction algorithms and digital tools. This can create complexity for evaluation if there is a lack of clarity as to the nature of the test strategy and its intended objectives. While standards for evaluation of molecular tests are relatively well established, this is not the case for prediction algorithms or digital tools. The concept of validity for prediction models such as those used to calculate polygenic scores is also viewed differentially in research and healthcare evaluation. In addition, prediction algorithms and their associated digital tools have not always been evaluated as tests.

Our analysis shows that by breaking down the PGS analysis application to its component parts can improve the evaluation and assessment. This would require applying the concepts and techniques from molecular test evaluation, prediction modelling and digital technology evaluation to these components as well as the whole pathway to provide a summary assessment of clinical validity. Such an approach can also be more informative for regulatory purposes as well as understanding clinical utility because it enables:

- Distinguishing between assays, models, tools and tests
- A clear understanding of the purpose and population of use for each of these components
- Clarity about the relationship between the different elements that inform polygenic score analysis pathways (whether standalone or integrated).

Conclusion

8. Conclusion

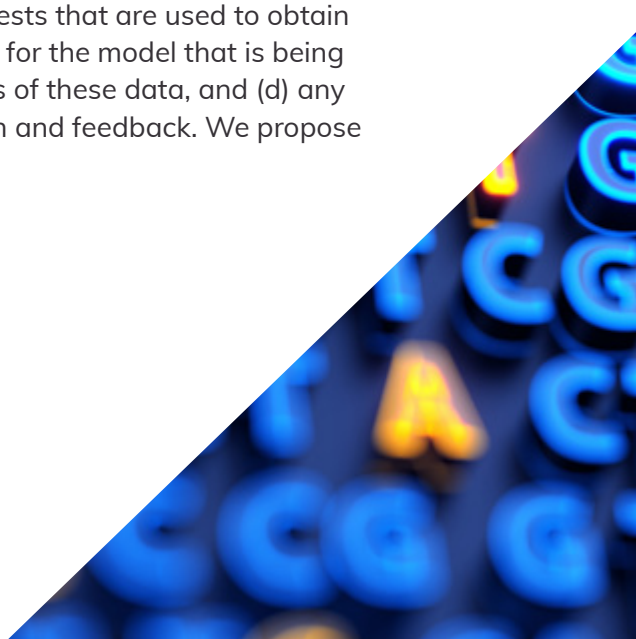
Research groups and companies have developed automated computational modules or algorithms that enable conversion of genetic data into a polygenic score for use on their own, or for incorporation with other risk factors to provide an integrated risk score. However, such products are still not widely used as part of healthcare. A key barrier to implementation has been uncertainty and lack of evidence regarding the value of polygenic score information. In this report we show how these uncertainties and evidence gaps come about by addressing four specific issues, resolution of which will serve to provide a more rational approach to evidence generation and appraisal for PGS applications. The issues are as follows:

- Conflation of terminology relating to polygenic scores, models and algorithms.
- Inadequate description of specific applications, in relation to intended population, role and purpose as part of specific healthcare pathways.
- Failure to define and evaluate all the key elements of PGS applications.
- Lack of real-world evidence (RWE) for PGS applications.

These are as yet unresolved and have a negative impact on the assessment and implementation of polygenic scores. They provide a challenge for decision makers because the existing evidence base (a) fails to show what information polygenic scores are providing (b) does not define with adequate precision how the product is to be used in health care or its intended purpose or objective or (c) how such use can be beneficial to the individual patient or to the health system as a whole. The consequence is that we are left with a body of evidence that is inadequate for the determination of the clinical validity or utility of a product in relation to its intended purpose.

In Chapter 5 and 7 of this report, we show how clarity with regards to terminology and better description of products and applications helps achieve more informative evidence, whether in relation to generation, evaluation or appraisal. Importantly, we have shown how existing evidence assessment frameworks can be applied and used for this purpose.

An essential element of our approach is to consider the different components that contribute to a specific product and application: (a) the tests that are used to obtain genetic data, (b) the source of other clinical data needed for the model that is being constructed (c) the nature of the algorithm(s) for analysis of these data, and (d) any other digital tool(s) that are used to enable data collation and feedback. We propose



considering these components as part of a test pipeline which will allow the bringing together and the application of concepts and techniques from molecular test evaluation, prediction modelling and digital technology evaluation to each separate component of the pipeline as well as the whole pathway.

Examination of analytical, scientific and clinical validity parameters across such components of a PGS analysis pipeline can provide evidence of the functional aspects of a PGS-based test. This can then inform the assessment of a test's performance with reference to the test's intended role and purpose. The advantage of this approach is that it enables cross-linking evidence across the different components of the PGS test pipeline, however they are configured.

We believe that our approach to such analysis allows us to identify issues that impact on the understanding of the analytic and scientific validity, as well as clinical validity of these pipelines. It will be important to achieve consensus amongst researchers, developers, health system decision makers and users as to which of the gaps which we have identified are critical and how they can be addressed. This is going to vary across different diseases and for particular applications. It is likely that for some areas the evidence gaps will persist, especially in relation to novel uses, but there is hope that at least for some diseases and for some applications the technique will provide extra utility. Progress on establishing both the evidence required for the different components of the PGS test pipeline as well as the acceptable levels of evidence will be necessary for the successful clinical implementation and wider uptake and use of any PGS-based applications.

In conclusion, polygenic scores are likely to be useful under certain circumstances. Identifying these and creating optimal systems for their use requires a more focussed approach to evidence generation and appraisal which is currently lacking.

References

9. References

1. Moorthie, S, Hall, A, Janus, J, et al. Polygenic scores and clinical utility. PHG Foundation. 2021.
2. Committee on Diagnostic Error in Health Care, Board on Health Care Services, Institute of Medicine, The National Academy of Sciences, Engineering, and Medicine,. 'In:' The Diagnostic Process. National Academies Press (US); 2015.
3. DxInsights/AdvaMedDx. Introduction to Molecular Diagnostics: The Essentials of Diagnostics Series. Available at: http://www.epemed.org/online/www/content2/108/469/3172/listdownloads/3175/507/ENG/dxinsights.pdf. 2013.
4. Critical Appraisal Skills Programme. CASP Checklists Available at: https://casp-uk.net/casp-tools-checklists/. 2022.
5. Bossuyt, P, Reitsma, J, Bruns, D, et al. STARD 2015: An Updated List of Essential Items for Reporting Diagnostic Accuracy Studies. British Medical Journal. 2015. 351(h5527).
6. Goldsack, J, Coravos, A, Bakker, J, et al. Verification, analytical validation, and clinical validation (V3): the foundation of determining fit-for-purpose for Biometric Monitoring Technologies (BioMeTs). npj Digital Medicine. 2020. 3(1): p. 55.
7. Mandelblatt, J, Ramsey, S, Lieu, T, et al. Evaluating Frameworks That Provide Value Measures for Health Care Interventions. Value Health. 2017. 20(2): pp. 185-192.
8. Augustovski, F, Alfie, V, Alcaraz, A, et al. A Value Framework for the Assessment of Diagnostic Technologies: A Proposal Based on a Targeted Systematic Review and a Multistakeholder Deliberative Process in Latin America. Value Health. 2021. 24(4): pp. 486-496.
9. Haddow, J, Palomaki, G. 'In:' ACCE: A Model Process for Evaluating Data on Emerging Genetic Tests. Oxford University Press; 2003.
10. Pitini, E, Baccolini, V, Migliara, G, et al. Time to Align: A Call for Consensus on the Assessment of Genetic Testing. Front Public Health. 2021. 9: p. 807695.
11. Doust, J A, Bell, K J L, Leeflang, M M G, et al. Guidance for the design and reporting of studies evaluating the clinical performance of tests for present or past SARS-CoV-2 infection. BMJ. 2021. 372: p. n568.
12. Noel-Storr, A H, McCleery, J M, Richard, E, et al. Reporting standards for studies of diagnostic test accuracy in dementia: The STARDdem Initiative. Neurology. 2014. 83(4): pp. 364-73.
13. Rennie, D. CONSORT revised--improving the reporting of randomized trials. JAMA. 2001. 285(15): pp. 2006-7.
14. Department of Health and Social Care. UK Government. Guidance. Protocol for evaluation of rapid diagnostic assays for specific SARS-CoV-2 antigens (lateral flow devices). 2021.
15. National Institute for Health and Care Excellence (NICE). NICE health technology evaluations: the manual Available at: www.nice.org.uk/process/pmg36. 2022.
16. Clinical Laboratory Standards Institute (CLSI). User Protocol for Evaluation of Qualitative Test Performance, 2nd Edition. 2021.
17. Walter, F M, Thompson, M J, Wellwood, I, et al. Evaluating diagnostic strategies for early detection of cancer: the CanTest framework. BMC Cancer. 2019. 19(1): p. 586.
18. Lijmer, J G, Leeflang, M, Bossuyt, P M. Proposals for a phased evaluation of medical tests. Med Decis Making. 2009. 29(5): pp. E13-21.

19. Moons, K G, Royston, P, Vergouwe, Y, et al. Prognosis and prognostic research: what, why, and how? BMJ. 2009. 338: p. b375.
20. Elovic, A, Pourmand, A. MDCalc Medical Calculator App Review. J Digit Imaging. 2019. 32(5): pp. 682-684.
21. Cowley, L E, Farewell, D M, Maguire, S, et al. Methodological standards for the development and evaluation of clinical prediction rules: a review of the literature. Diagn Progn Res. 2019. 3: p. 16.
22. Knaus, W A, Wagner, D P, Draper, E A, et al. The APACHE III prognostic system. Risk prediction of hospital mortality for critically ill hospitalized adults. Chest. 1991. 100(6): pp. 1619-36.
23. Wishart, G C, Azzato, E M, Greenberg, D C, et al. PREDICT: a new UK prognostic model that predicts survival following surgery for invasive breast cancer. Breast Cancer Res. 2010. 12(1): p. R1.
24. Royston, P, Moons, K G, Altman, D G, et al. Prognosis and prognostic research: Developing a prognostic model. BMJ. 2009. 338: p. b604.
25. Steyerberg, E W, Vergouwe, Y. Towards better clinical prediction models: seven steps for development and an ABCD for validation. Eur Heart J. 2014. 35(29): pp. 1925-31.
26. Steyerberg, E W. 'Ed.'Clinical prediction models : a practical approach to development, validation, and updating. Springer; 2009.
27. Van Calster, B, Wynants, L, Timmerman, D, et al. Predictive analytics in health care: how can we know it works? J Am Med Inform Assoc. 2019. 26(12): pp. 1651-1654.
28. Altman, D G, Vergouwe, Y, Royston, P, et al. Prognosis and prognostic research: validating a prognostic model. BMJ. 2009. 338: p. b605.
29. The Breast Cancer Risk Assessment Tool. 2023; Available from: https://bcrisktool.cancer.gov/.
30. CanRisk. 2023; Available from: https://www.canrisk.org/.
31. IBIS Breast Cancer Risk Evaluation Tool. 2023; Available from: https://ems-trials.org/riskevaluator/.
32. Horvath, A R, Lord, S J, StJohn, A, et al. From biomarkers to medical tests: the changing landscape of test evaluation. Clin Chim Acta. 2014. 427: pp. 49-57.
33. Burke, W, Zimmern, R L, Kroese, M. Defining purpose: a key step in genetic test evaluation. Genet Med. 2007. 9(10): pp. 675-81.
34. Zimmern, R L, Kroese, M. The evaluation of genetic tests. J Public Health (Oxf). 2007. 29(3): pp. 246-50.
35. Kroese, M, Zimmern, R L, Sanderson, S. Genetic tests and their evaluation: can we answer the key questions? Genet Med. 2004. 6(6): pp. 475-80.
36. Royal Statistical Society Working Group on Diagnostic Tests. Diagnostic tests working group report. 2021.
37. Montano, M. 'In:' 3 - Blood biomarkers: overview of existing serum test strategies for disease severity, risk for progression, therapeutic benchmark targets. Woodhead Publishing; 2014.
38. Steyerberg, E W, Moons, K G, van der Windt, D A, et al. Prognosis Research Strategy (PROGRESS) 3: prognostic model research. PLoS Med. 2013. 10(2): p. e1001381.
39. Tangri, N, Kent, D M. Toward a Modern Era in Clinical Prediction: The TRIPOD Statement for Reporting Prediction Models. American Journal of Kidney Diseases. 2015. 65(4): pp. 530-533.
40. Moons, K G M, Altman, D G, Reitsma, J B, et al. New Guideline for the Reporting of Studies Developing, Validating, or Updating a Multivariable Clinical Prediction Model: The TRIPOD Statement. Advances in Anatomic Pathology. 2015. 22(5): pp. 303-305.
41. Ordish, J, Murfet, H, Hall, A. Algorithms as medical devices. PHG Foundation. 2019.

42. Furness, P, Zimmern, R, Wright, C, et al. The evaluation of diagnostic laboratory tests and complex biomarkers. Summary of a Diagnostic Summit. 14-15 January, 2008. 2008.
43. Sun F, Bruening W, Erinoff E, et al. Addressing Challenges in Genetic Test Evaluation: Evaluation Frameworks and Assessment of Analytic Validity. Agency for Healthcare Research and Quality (US). 2011.
44. Atkinson, A J, Colburn, W A, DeGruttola, V G, et al. Biomarkers and surrogate endpoints: Preferred definitions and conceptual framework. Clinical Pharmacology & Therapeutics. 2001. 69(3): pp. 89-95.
45. Online Mendelian Inheritance of Man, O McKusick-Nathans Institute of Genetic Medicine, John Hopkins University (Baltimore, MD). 2023; Available from: https://omim.org/.
46. Landrum, M J, Lee, J M, Benson, M, et al. ClinVar: improving access to variant interpretations and supporting evidence. Nucleic Acids Res. 2018. 46(D1): pp. D1062-D1067.
47. Rehm, H L, Berg, J S, Brooks, L D, et al. ClinGen--the Clinical Genome Resource. N Engl J Med. 2015. 372(23): pp. 2235-42.
48. Strande, N T, Riggs, E R, Buchanan, A H, et al. Evaluating the Clinical Validity of Gene-Disease Associations: An Evidence-Based Framework Developed by the Clinical Genome Resource. Am J Hum Genet. 2017. 100(6): pp. 895-906.
49. Richards, S, Aziz, N, Bale, S, et al. Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. Genetics in Medicine. 2015. 17(5): pp. 405-423.
50. Ellard, S, Baple, E L, Callaway, A, et al. 'Ed.'ACGS Best Practice Guidelines for Variant Classification in Rare Disease 2020. Association for Clinical Genomics Science; 2020.
51. Association for Clinical Genomic Science. Best practice guidelines. Available at: https://www.acgs.uk.com/quality/best-practice-guidelines/. 2021. 2021(25/10/2021).
52. Wand, H, Lambert, S A, Tamburro, C, et al. Improving reporting standards for polygenic scores in risk prediction studies. Nature. 2021. 591(7849): pp. 211-219.
53. Abu-El-Haija, A, Reddi, H V, Wand, H, et al. The clinical application of polygenic risk scores: A points to consider statement of the American College of Medical Genetics and Genomics (ACMG). Genet Med. 2023. p. 100803.
54. Li, M M, Datto, M, Duncavage, E J, et al. Standards and Guidelines for the Interpretation and Reporting of Sequence Variants in Cancer: A Joint Consensus Recommendation of the Association for Molecular Pathology, American Society of Clinical Oncology, and College of American Pathologists. J Mol Diagn. 2017. 19(1): pp. 4-23.
55. Pan, B, Ren, L, Onuchic, V, et al. Assessing reproducibility of inherited variants detected with short-read whole genome sequencing. Genome Biol. 2022. 23(1): p. 2.
56. Liang, Y, He, L, Zhao, Y, et al. Comparative Analysis for the Performance of Variant Calling Pipelines on Detecting the de novo Mutations in Humans. Front Pharmacol. 2019. 10: p. 358.
57. SoRelle, J A, Wachsmann, M, Cantarel, B L. Assembling and Validating Bioinformatic Pipelines for Next-Generation Sequencing Clinical Assays. Arch Pathol Lab Med. 2020. 144(9): pp. 1118-1130.
58. Marshall, C R, Chowdhury, S, Taft, R J, et al. Best practices for the analytical validation of clinical whole-genome sequencing intended for the diagnosis of germline disease. NPJ Genom Med. 2020. 5: p. 47.

59. Roy, S, Coldren, C, Karunamurthy, A, et al. Standards and Guidelines for Validating Next-Generation Sequencing Bioinformatics Pipelines: A Joint Recommendation of the Association for Molecular Pathology and the College of American Pathologists. J Mol Diagn. 2018. 20(1): pp. 4-27.
60. Soares, M O, Walker, S, Palmer, S J, et al. Establishing the Value of Diagnostic and Prognostic Tests in Health Technology Assessment. Med Decis Making. 2018. 38(4): pp. 495-508.
61. Rector, T S, Taylor, B C, Wilt, T J. Chapter 12: systematic review of prognostic tests. J Gen Intern Med. 2012. 27 Suppl 1: pp. S94-101.
62. Burke, W. Genetic tests: clinical validity and clinical utility. Curr Protoc Hum Genet. 2014. 81: pp. 9 15 1-8.
63. Simundic, A M. Measures of Diagnostic Accuracy: Basic Definitions. EJIFCC. 2009. 19(4): pp. 203-11.
64. van Stralen, K J, Stel, V S, Reitsma, J B, et al. Diagnostic methods I: sensitivity, specificity, and other measures of accuracy. Kidney Int. 2009. 75(12): pp. 1257-1263.
65. Wright, C F, Kroese, M. Evaluation of genetic tests for susceptibility to common complex diseases: why, when and how? Hum Genet. 2010. 127(2): pp. 125-34.
66. Parikh, R, Mathai, A, Parikh, S, et al. Understanding and using sensitivity, specificity and predictive values. Indian J Ophthalmol. 2008. 56(1): pp. 45-50.
67. Kingdom, R, Wright, C F. Incomplete Penetrance and Variable Expressivity: From Clinical Studies to Population Cohorts. Front Genet. 2022. 13: p. 920390.
68. Sperrin, M, Riley, R D, Collins, G S, et al. Targeted validation: validating clinical prediction models in their intended population and setting. Diagn Progn Res. 2022. 6(1): p. 24.
69. Van Calster, B, Steyerberg, E W, Wynants, L, et al. There is no such thing as a validated prediction model. BMC Med. 2023. 21(1): p. 70.
70. Janssens, A C, Martens, F K. Prediction Research - An Introduction (Version 2.2, 2018). Available at: http://www.cecilejanssens.org/wp-content/uploads/2018/01/PredictionManual2.0.pdf. 2018.
71. Cook, N R. Use and misuse of the receiver operating characteristic curve in risk prediction. Circulation. 2007. 115(7): pp. 928-935.
72. QRISK®3 calculator. ClinRisk. Available at: https://www.qrisk.org/. Available from:
73. Steyerberg, E W, Vickers, A J, Cook, N R, et al. Assessing the performance of prediction models: a framework for traditional and novel measures. Epidemiology. 2010. 21(1): pp. 128-38.
74. Wynants, L, van Smeden, M, McLernon, D J, et al. Three myths about risk thresholds for prediction models. BMC Med. 2019. 17(1): p. 192.
75. Kappen, T H, van Klei, W A, van Wolfswinkel, L, et al. Evaluating the impact of prediction models: lessons learned, challenges, and recommendations. Diagn Progn Res. 2018. 2: p. 11.
76. Zhang, Z, Rousson, V, Lee, W C, et al. Decision curve analysis: a technical note. Ann Transl Med. 2018. 6(15): p. 308.
77. Grigore, B, Lewis, R, Peters, J, et al. Development, validation and effectiveness of diagnostic prediction tools for colorectal cancer in primary care: a systematic review. BMC Cancer. 2020. 20(1): p. 1084.
78. van Royen, F S, Moons, K G M, Geersing, G J, et al. Developing, validating, updating and judging the impact of prognostic models for respiratory diseases. Eur Respir J. 2022. 60(3).

79. 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.
80. 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.
81. Zhang, Y D, Hurson, A N, Zhang, H, et al. Assessment of polygenic architecture and risk prediction based on common variants across fourteen cancers. Nat Commun. 2020. 11(1): p. 3353.
82. Wang, Y, Tsuo, K, Kanai, M, et al. Challenges and Opportunities for Developing More Generalizable Polygenic Risk Scores. Annu Rev Biomed Data Sci. 2022.
83. 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.
84. Chatterjee, N, Shi, J, Garcia-Closas, M. Developing and evaluating polygenic risk prediction models for stratified disease prevention. Nat Rev Genet. 2016. 17(7): pp. 392-406.
85. Mavaddat, N, Ficorella, L, Carver, T, et al. Incorporating alternative Polygenic Risk Scores into the BOADICEA breast cancer risk prediction model. Cancer Epidemiol Biomarkers Prev. 2023.
86. Wand, H, Knowles, J W, Clarke, S L. The need for polygenic score reporting standards in evidence-based practice: lipid genetics use case. Curr Opin Lipidol. 2021. 32(2): pp. 89-95.
87. Mostafavi, H, Harpak, A, Agarwal, I, et al. Variable prediction accuracy of polygenic scores within an ancestry group. Elife. 2020. 9.
88. Cerdena, J P, Grubbs, V, Non, A L. Racialising genetic risk: assumptions, realities, and recommendations. Lancet. 2022. 400(10368): pp. 2147-2154.
89. Khera, A V, Chaffin, M, Zekavat, S M, et al. Whole-Genome Sequencing to Characterize Monogenic and Polygenic Contributions in Patients Hospitalized With Early-Onset Myocardial Infarction. Circulation. 2019. 139(13): pp. 1593-1602.
90. Hao, L, Kraft, P, Berriz, G F, et al. Development of a clinical polygenic risk score assay and reporting workflow. Nat Med. 2022.
91. Weissbrod, O, Kanai, M, Shi, H, et al. Leveraging fine-mapping and multipopulation training data to improve cross-population polygenic risk scores. Nat Genet. 2022. 54(4): pp. 450-458.
92. Lewis, A C F, Green, R C, Vassy, J L. Polygenic risk scores in the clinic: Translating risk into action. HGG Adv. 2021. 2(4): p. 100047.
93. Brockman, D G, Petronio, L, Dron, J S, et al. Design and user experience testing of a polygenic score report: a qualitative study of prospective users. BMC Med Genomics. 2021. 14(1): p. 238.
94. Folkersen, L, Pain, O, Ingason, A, et al. Impute.me: An Open-Source, Non-profit Tool for Using Data From Direct-to-Consumer Genetic Testing to Calculate and Interpret Polygenic Risk Scores. Front Genet. 2020. 11: p. 578.
95. UK Government. Using the UKCA mark from 1 Jan 2021. 2022; Available from: https://www.gov.uk/guidance/using-the-ukca-mark-from-1-january-2021
96. MacArthur, J, Bowler, E, Cerezo, M, et al. The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog). Nucleic Acids Res. 2017. 45(D1): pp. D896-D901.

97. Janssens, A C. Validity of polygenic risk scores: are we measuring what we think we are? Hum Mol Genet. 2019. 28(R2): pp. R143-R150.
98. Collister, J A, Liu, X, Clifton, L. Calculating Polygenic Risk Scores (PRS) in UK Biobank: A Practical Guide for Epidemiologists. Front Genet. 2022. 13: p. 818574.
99. Cupido, A J, Tromp, T R, Hovingh, G K. The clinical applicability of polygenic risk scores for LDL-cholesterol: considerations, current evidence and future perspectives. Curr Opin Lipidol. 2021. 32(2): pp. 112-116.
100. Weale, M E, Riveros-Mckay, F, Selzam, S, et al. Validation of an Integrated Risk Tool, Including Polygenic Risk Score, for Atherosclerotic Cardiovascular Disease in Multiple Ethnicities and Ancestries. Am J Cardiol. 2021. 148: pp. 157-164.
101. Lee, A, Mavaddat, N, Wilcox, A N, et al. BOADICEA: a comprehensive breast cancer risk prediction model incorporating genetic and nongenetic risk factors. Genet Med. 2019. 21(8): pp. 1708-1718.
102. Campbell, M C, Tishkoff, S A. African genetic diversity: implications for human demographic history, modern human origins, and complex disease mapping. Annu Rev Genomics Hum Genet. 2008. 9: pp. 403-33.
103. Shi, S, Yuan, N, Yang, M, et al. Comprehensive Assessment of Genotype Imputation Performance. Hum Hered. 2018. 83(3): pp. 107-116.
104. Reddi, H V, Wand, H, Funke, B, et al. Laboratory perspectives in the development of polygenic risk scores for disease: A points to consider statement of the American College of Medical Genetics and Genomics (ACMG). Genet Med. 2023. p. 100804.
105. Binuya, M A E, Engelhardt, E G, Schats, W, et al. Methodological guidance for the evaluation and updating of clinical prediction models: a systematic review. BMC Med Res Methodol. 2022. 22(1): p. 316.
106. Chen, S F, Dias, R, Evans, D, et al. Genotype imputation and variability in polygenic risk score estimation. Genome Med. 2020. 12(1): p. 100.
107. Clifton, L, Collister, J A, Liu, X, et al. Assessing agreement between different polygenic risk scores in the UK Biobank. Sci Rep. 2022. 12(1): p. 12812.
108. ClinicalTrials.gov. Bethesda (MD): National Library of Medicine (US) 2000. Identifier NCT05294419: The Healthcare Evaluation of Absolute Risk Testing Study [cited 19.01.2023]. 2023.
109. ClinicalTrials.gov Bethesda (MD) National Library of Medicine (US) 2000. Identifier NCT04331535: The Genomic Medicine at VA Study [cited 19.01.2023]. Available from: https://ClinicalTrials.gov/show/NCT04331535. 2023.
110. Esserman, L J, Study, W, Athena, I. The WISDOM Study: breaking the deadlock in the breast cancer screening debate. NPJ Breast Cancer. 2017. 3: p. 34.
111. Pons-Rodriguez, A, Forne Izquierdo, C, Vilaplana-Mayoral, J, et al. Feasibility and acceptability of personalised breast cancer screening (DECIDO study): protocol of a single-arm proof-of-concept trial. BMJ Open. 2020. 10(12): p. e044597.
112. ClinicalTrials.gov. Bethesda (MD): National Library of Medicine (US) 2000. Identifier NCT03672331: My Personalized Breast Screening [cited 19.01.2023]. Available at: https://clinicaltrials.gov/ct2/show/NCT03672331. 2023.

The PHG Foundation is a non-profit think tank with a special focus on how genomics and other emerging health technologies can provide more effective, personalised healthcare and deliver improvements in health for patients and citizens.

intelligence@phgfoundation.org

UNIVERSITY OF CAMBRIDGE

PHG FOUNDATION

making science work for health