NHS
Genomic AI Network

Digital Health Technology Evaluation of Genomic AI Tools

Genomic AI Network

Contents

Executive Summary

Introduction

Work Package 3 (WP3) of the Genomic AI Network (GAIN) focused on piloting and evaluating artificial intelligence tools across different stages of the NHS genomic medicine pathway. The aim was to explore how AI technologies can support clinical decision-making, data interpretation, and patient identification while operating within appropriate governance and safety frameworks.

The WP3 portfolio included four complementary projects addressing key challenges in genomic medicine. **Genollama (WP3A)** developed an automated AI pipeline to extract and standardise genomic biomarker data from electronic patient records. Deployed at Guy's and St Thomas' NHS Foundation Trust, the tool enables structured use of genomic test report data and supports downstream applications such as clinical trial recruitment and research analytics.

MendelScan (WP3B) focused on proactive case-finding for rare diseases using electronic health record analysis. The tool, an MHRA-registered Class I medical device, scans primary care records to identify patients who may benefit from genomic testing or specialist referral. During the GAIN programme, additional algorithms were developed to identify patients with monogenic inflammatory bowel disease and rare epilepsy, supporting earlier diagnosis and intervention.

Language AI for Inherited Cardiac Conditions (WP3C) applied natural language processing to identify patients with inherited cardiac diseases from unstructured NHS clinical data. Using the CogStack platform, the system extracts phenotypic information from clinical records and aligns patients to National Genomic Test Directory criteria, improving identification and referral for genomic testing. Retrospective validation and shadow-mode deployment demonstrated strong model performance and feasibility for integration into clinical pathways.

GenePy (WP3D) explored a computational method for compressing pathogenic signals from large genomic sequencing datasets into gene-level scores. This approach supports variant prioritisation by helping clinical scientists more quickly identify candidate disease genes in whole genome and exome sequencing data. The method has been implemented at scale within the Genomics England Research Environment and is being evaluated for its potential to reduce manual curation time and identify previously missed diagnoses.

Together, these projects demonstrate how AI can support multiple aspects of genomic medicine, from identifying patients who may benefit from testing to improving the interpretation and use of genomic data. The WP3 pilots also highlight the diversity of AI maturity within the NHS ecosystem, ranging from research-stage analytical tools to locally deployed digital health technologies and regulated clinical decision-support systems.

Purpose

This document provides a structured evaluation of Work Package 3 within the Genomic AI Network. Its purpose is to consider how the projects align with recognised national standards for digital health technologies, and to capture practical lessons for the safe development and deployment of AI within NHS Genomic Medicine.

WP3 includes tools at different levels of technical maturity, regulatory readiness and clinical integration. By evaluating them within a common framework, this document offers a comparative snapshot of how AI

technologies behave across the genomic pathway, from early-stage infrastructure tools to regulated medical devices.

Beyond programme reporting, this evaluation is intended to serve as a practical reference. It highlights the types of governance, safety, evidence and implementation considerations that any AI tool developer working in the NHS should anticipate from the outset.

NICE Evidence Standards

The evaluation is structured around the Evidence Standards Framework for Digital Health Technologies developed by the National Institute for Health and Care Excellence (NICE).

NICE provides national guidance to improve health and care quality in England. Its Evidence Standards Framework sets out expectations for digital health technologies across 21 standards, covering areas such as:

- Design factors
- Describing value
- Demonstrating performance
- Delivering value
- Deployment considerations

The framework categorises digital health technologies into tiers based on their intended purpose and potential clinical risk, recognising that higher-risk tools require stronger evidence and more robust governance arrangements. As the WP3 projects are at different stages of maturity, not all NICE standards are directly applicable to every tool. However, the framework provides a broad and consistent set of considerations that align with national expectations for digital health technologies within the NHS.

NHS Genomic AI Network

Summary table

NICE Standard	WP3A	WP3B	WP3C	WP3D
Standard 1: the digital health technology (DHT) should comply with relevant safety and quality standards	☑	☑	☑	☑
Standard 2: incorporate intended user group acceptability in the design of the DHT	☑	☑	☑	☑
Standard 3: consider environmental sustainability	☑	☑	☑	☑
Standard 4: consider health and care inequalities and bias mitigation	☑	☑	☑	☑
Standard 5: embed good data practices in the design of the DHT	☑	☑	☑	☑
Standard 6: define the level of professional oversight	☑	☑	☑	☑
Standard 7: show processes for creating reliable health information	N/A Not relevant – tier B/C focus. Not applicable to tier A scope as written. However, evidence of standard still seen.	☑	☑	☑
Standard 8: show that the DHT is credible with UK professionals	N/A Not relevant – downstream of the project scope	☑	☑	N/A
Standard 9: provide safeguarding assurances for DHTs where users are considered to be in vulnerable groups, or where peer-to-peer interaction is enabled	N/A Not relevant – tier b/c. not applicable to tier A however evidence of standard is still seen.	N/A Not relevant – clinician facing CDS; no patient P2P.	☑	N/A Not relevant – clinician facing CDS; no patient P2P.
Standard 10: describe the intended purpose and target population	☑	☑	☑	☑
Standard 11: describe the current pathway or system process	☑	☑	☑	☑
Standard 12: describe the proposed pathway or system process using the DHT	☑	☑	☑	☑
Standard 13: describe the expected health, cost and resource impacts compared with current care or system processes	☑	☑	☑	☑
Standard 14: provide evidence of the DHT's effectiveness to support its claimed benefits	N/A Not relevant – tier C only.	☑	☑	☑
Standard 15: show real-world evidence that the claimed benefits can be realised in practice	☑	☑	☑	☑
Standard 16: the company and evaluator should agree a plan for measuring usage and changes in the DHT's performance over time	☑	☑	☑	☑
Standard 17: provide a budget impact analysis	☑	☑	☑	☑
Standard 18: for DHTs with higher financial risk, provide a cost-effectiveness analysis	N/A	N/A	N/A	N/A
Standard 19: ensure transparency about requirements for deployment	☑	☑	☑	☑
Standard 20: describe strategies for communication, consent and training processes to allow the DHT to be understood	☑	☑	☑	☑
Standard 21: ensure appropriate scalability	☑	☑	☑	☑

Work Package 3A Evaluation

Executive summary

This project sought to develop an automated AI-driven pipeline to extract existing genomic biomarker data in the electronic patient record, with a view to unleashing the potential that this data holds for clinical care and research. To date, such a pipeline has been developed, and the associated model, Genollama, has been deployed to standardise (i.e. surface biomarkers in) genomic test reports at GSTT. The outputs of the process are primed to support downstream activities such as clinical trial recruitment. Evaluation has shown positive pipeline performance and positive performance on standardisation tasks. Next steps are to deploy the model for a wider range of use cases within the trust and at other trusts, positioning it as a standard part of genomics workflows. There is also significant potential for Genollama to impact complex processes such as automated decision support.

Note that for the purpose of this evaluation we consider Genollama to be a Tier A Digital Health Technology (DHT).

Standard 1: the digital health technology (DHT) should comply with relevant safety and quality standards

The DHT (Genollama) complies with relevant safety and quality standards: it is a tool that, at present, is used by teams within a single hospital trust (GSTT) – with no new data collected, and no data leaving the trust – and thus operates within existing ethical, data and governance frameworks. Should the tool be deployed in additional settings going forward, template governance documents (e.g. a template DPIA) have been developed for the use of tools of this nature, and are available for use. Despite operating within existing frameworks, bespoke DCB0129 and DCB0160 documents have been completed to complement the existing governance setup (see appendix).

Standard 2: incorporate intended user group acceptability in the design of the DHT

The requirement for a schema standardisation tool has been motivated by the views and interests of a broad range of stakeholders. This includes clinicians, data engineers and data scientists across a variety of healthcare settings. These individuals have also directly informed the design and development of the tool, including the structure of the synthetic data used for training, the structure of the target schema and the mechanisms by which the model can be applied to data at scale. For use cases like clinical trial recruitment, the thoughts and experiences of individuals working in the domain have been sought to, for example, understand how the tool might complement existing workflows.

When deploying at other sites, recommendations and considerations for incorporating intended user group accessibility into Genollama naturally extend to conversations with patients about the use of a tool that, ultimately, may impact care pathways. At these sites, recommendations also extend to the types of input data expected, and the types of existing data that should inform the structure of synthetic data if further fine-tuning is to be performed.

Standard 3: consider environmental sustainability

Artificial Intelligence (AI) technology can have high energy demands, so one should always be mindful of the environmental impact associated with its use. Although Genollama is an example of an AI-based technology – specifically an example of a fine-tuned Large Language Model (LLM) – reducing environmental impact is actually at the heart of its design: as the model runs locally with relatively modest compute (GPU) requirements, its energy demands and impact are far smaller, if not negligible, compared to the demands and impact of the commercial cloud-based models that might otherwise be used in its

place. These gains do not come at the expense of output quality. At the same time, it is worth acknowledging that cloud-based models are used to generate training data, however this is a one-shot process that is compensated for by the lower energy demands described. Still, as the energy demands and impact of the DHT are not zero, further optimisations in this regard continue to be a priority.

Standard 4: consider health and care inequalities and bias mitigation

The primary downstream use case facilitated by Genollama is automated clinical trial recruitment, a process that currently, as with any manual activity, is subject to the constraints of the individual performing the task, such as cognitive load. These constraints indirectly introduce bias into the recruitment process, something that inadvertently leads to inequalities in care. Therefore, much of the purpose of the tool is to mitigate bias and reduce inequalities.

Despite this focus, the tool itself is potentially impacted by bias, particularly given the reliance on open-source foundation models (Llama) that may not have representative training data. This is offset in this project via fine-tuning against synthetic data, the form of which is controlled – in consultation with stakeholders (described) – to ensure the absence of bias. Consultation with patients in future deployment settings (described) can also be used as an opportunity to identify and mitigate bias in any source data.

Standard 5: embed good data practices in the design of the DHT

Genollama is fine-tuned with synthetic free-text data to structured data pairs, to guide domain-specific schema standardisation. The structure of the synthetic free-text data was determined with a range of stakeholders (described), to ensure both accurate and contemporary document structures are included. Moreover, the model development pipeline developed through this project and others facilitates the ongoing refinement of these structures based on additional deployment use cases.

Structured data is generated from this free-text data by a separate language model (Claude Sonnet 4.5), which is also provided with the target schema. The comprehensive nature of this model for this task is evident in a variety of third-party benchmarks, but project-specific validation is also performed against all the samples automatically using schema conformance checks, and manually by genomics experts. Once again, the quality of the outputs of the model hinges largely upon the quality of the source genomic test data, and specifying minimum source data quality requirements would be a key activity upon deployment at further sites.

Standard 6: define the level of professional oversight

The stakeholder discussions informing the design of Genollama have included conversations with organisations like NHS England and those developing the Unified Genomic Record (UGR), and conversations with the Genomic Medicine Service (GMS) and the National Disease Registration Service (NDRS). We would look to these same individuals/organisations to provide professional oversight going forward. As any direct impacts on care would be downstream on the tool, and to ensure acceptability to the professionals mentioned, this oversight would likely come in the form of periodic trend reviews, to ensure Genollama is still appropriately configured. This most naturally maps to periodic discussions on the appropriateness of the synthetic free-text data and the schema used. In the short-term, the views and inputs of these individuals could also be leveraged to understand the potential downstream clinical utility of Genollama's outputs.

Standard 7: show processes for creating reliable health information

Although the Genollama is classified as Tier A, and although interactions with the DHT's outputs happen via downstream platforms, ensuring that reliable health information is created starts within the tool itself. Genollama is the output of a model training pipeline that has been shown to produce tools that standardise health records with a high degree of accuracy, in the order of >99%. More specifically, the pipeline generates models that score well in both precision (>98%) and recall (>96%), and the trade-off between the two (f1; >97%).

To complement inherited statistical performance, the tool's outputs are reviewed by domain experts both during development and once in use (discussed). Such a process also confirms the comprehensiveness and validity of both the outputs and the model development process. Evaluation is an ongoing process, and despite the strong current performance of the model(s), going forward, new dimensions of evaluation might be identified, and even more domain expertise could be leveraged to further validate model performance under these new metrics.

Standard 8: show that the DHT is credible with UK professionals

The input of professionals on the DHT is essential. To complement existing stakeholder input, future work would likely convene a focus group of such professions to help demonstrate the impact of the model. In addition to (further consultation with) the stakeholders already noted, this group would likely include: a cross-section of individuals connected to genomics in the UK health and social care system, including clinical geneticists and scientists, specialist nurses and those already assessing new genomic technologies (e.g. NICE); professionals connected to downstream activities, such as, in the case of clinical trial recruitment, current/previous trial sponsors and investigators, as well as representatives from regulatory bodies (e.g. the MHRA); and industry representatives who are building similar models, either for general purpose use or healthcare specific.

Standard 9: provide safeguarding assutrials for DHTs where users are considered to be in vulnerable groups, or where peer-to-peer interaction is enabled

Genollama is designed to sit upstream of any functions that would benefit from standardised data output (e.g. clinical trial recruitment). As such, whether Genollama's downstream users are considered to be in vulnerable groups, or whether peer-to-peer interactions are enabled downstream (necessitating safeguarding assurances), depends very much upon the clinical pathway in which it is deployed (i.e. changing Genollama's classification from A to B or C). Despite this fact, every effort has been made, via stakeholder involvement and by following established best practice in data use (discussed), to ensure that Genollama has been developed with the needs and requirements of a wide range of users and use cases in mind.

Standard 10: describe the intended purpose and target population

The purpose of Genollama is to standardise genomic test report output to support (the automation of) a variety of downstream activities. The primary target activity is clinical trial matching/recruitment, where the standardised fields enable automated recruitment, or, at the very least, assist with recruitment. The target population is therefore those recruiting for (molecularly targeted) clinical trials, be this at the point of care or as a part of wider teams. Given the size of the clinical trials field, the size of this target population is significant, and, while uptake is to some degree dependent on the quality of downstream tooling, the low technical barrier to entry in directly using the tool's text-based outputs suggests high uptake. Naturally some barriers to entry remain for those who may be deploying the tool and/or any associated trial matching technology.

Genollama can also support other downstream activities, such as matching patients to guidelines for the purpose of (automated) decision support. Additional activities like this bring their own user populations and thus broaden the impact of the tool.

Standards 11 and 12: describe the current pathway or system process. Describe the proposed pathway or system process using the DHT

The Genollama logic model (attached) provides a structured way to describe the potential system pathway and processes, and future considerations to evaluate the innovation. It outlines the existing processes in genomic data handling - largely manual extraction of biomarkers, unstructured clinical text, and labour-intensive curation by clinical scientists and data teams. These highlight the current pathway, where genomic insights rely on time-consuming review of reports, inconsistent documentation, and variable data quality across primary and secondary care.

The model includes enhanced descriptions of outcomes and ways to measure them, illustrating how automated NLP-based biomarker extraction, structured phenotype generation, and integration into clinical workflows could streamline data processing. Activities such as automated annotation, retrospective record population, and standardisation of genomic information demonstrate how Genollama replaces or enhances manual steps. The outputs and outcomes sections show system-level changes: improved data accuracy, reduced staff burden, faster MDT decision-making, and enhanced clinical trial matching. By mapping inputs, activities, outputs, outcomes, and impacts, the model enables a clear articulation of how GenoLlama transforms the pathway from manual, fragmented processes into a more efficient, accurate, and scalable digital workflow.

Standard 13: describe the expected health, cost and resource impacts compared with current care or system processes

The GenoLlama logic model clearly sets out the anticipated health, cost, and resource impacts by contrasting the manual, resource-intensive current system with the automated, standardised pathway enabled by the DHT. It identifies how current care relies on clinicians and data teams manually reviewing genomic reports, extracting biomarkers, and coding unstructured text - activities that consume significant clinical scientist time and delay downstream processes such as MDT discussions, trial matching, and diagnostic pathways.

The model then describes how GenoLlama automates these tasks, enabling faster and more reliable extraction of biomarkers, cancer history, and molecular features. This supports expectations of improved accuracy, reduced variability, and more complete national datasets. Subject to a real-world evaluation, these changes could translate directly into health impacts, such as faster diagnostic assessments, improved treatment planning, enhanced eligibility identification for clinical trials, and more efficient patient flow through genomic pathways.

The model also outlines potential cost and resource impacts, including reduced manual curation time, lower trial recruitment costs, reduced unnecessary testing, and better use of specialist staff. By mapping activities to outputs, outcomes, and impacts, the logic model provides an evidence-based structure for describing how GenoLlama delivers health benefits and system-level efficiencies compared with existing processes.

Standard 14: provide evidence of the DHT's effectiveness to support its claimed benefits

Evidence of the DHT's effectiveness to support its claimed benefits is covered in the responses to other standards in this document, including Standards 7 and 15.

Standard 15: show real-world evidence that the claimed benefits can be realised in practice

Genollama has been used to standardise (convert from free text to text under a structured schema, surfacing biomarkers) historic GeneWorks data (genomic test reports) held at GSTT. Outputs have been tested for schema adherence, with 92% conformance (n=1000). Of the 8% that do not conform, the majority can be addressed with basic post processing. Note that for these outputs, the original data remains intact for downstream use cases. This demonstrates a successful pilot of the tool, and that it is able to perform to its expected level.

The structured data produced by the pilot is an impactful asset in its own right, as it enables a variety of offline analytics use cases. These use cases will likely be supported by the trust-wide Snowflake analytics platform, for which standardised GeneWorks data is an ingestion target. While the standardisation process and its outputs are the best evidence for the realisation of claimed benefits, and while the silent evaluation of Genollama in this way aligns with the expected evidence gathering mechanisms, it was also deemed important to understand the likely impact on downstream functions, particularly clinical trial recruitment. Therefore, the eligibility criteria for a historic clinical trial – 'EPIK-O', with molecular targets that include BRCA mutation – were identified with the GSTT clinical trials team, recreated and applied in

an automated fashion against the standardised genomics data, with evidence emerging that this approach could emulate manual recruitment, suggesting improvements in the speed of recruitment and lowering clinical team load, and even potentially wider patient recruitment.

Standard 16: the company and evaluator should agree a plan for measuring usage and changes in the DHT's performance over time

The Genollama schema and the expected structure of synthetic free-text data are assets that are likely to change over time, e.g. with the arrival of new management systems, tests, nomenclature, etc. For any Genollama deployment, an annual revalidation with domain experts would be recommended to ascertain the current suitability of each of these assets. Unforeseen environmental changes (e.g. the unanticipated deprecation of a key management system) should also be identified by domain experts and signal early revalidation, as should the arrival of significant new use cases for the structured output. Many of the processes that were used to derive the current schema and synthetic data, such as consultation with stakeholders, should drive the update of these assets and revalidation. All downstream functions and connected stakeholders dependent on the tool's output should be made aware of any updates via, for example, package updates.

Standard 17: provide a budget impact analysis

The attached logic model provides a structured foundation for developing a budget impact analysis by identifying the costs, resource use, and system changes associated with both current care and the proposed GenoLlama-enabled pathway. It outlines the inputs required for implementation - such as IT infrastructure, staff training, governance arrangements, and integration work - which can be translated directly into upfront and ongoing costs within a budget impact analysis.

The model also identifies activities currently performed manually, such as biomarker extraction, reviewing genomic reports, structuring free text, and preparing datasets. These represent baseline resource use and can be costed using staff time, clinical scientist capacity, and administrative overhead. By contrast, the GenoLlama pathway replaces or reduces many of these manual steps through automation. The outputs and outcomes - such as reduced manual curation time, faster diagnostics, fewer unnecessary tests, and improved trial recruitment efficiency - provide quantifiable areas where cost savings and resource efficiencies can be modelled.

Additionally, the logic model highlights wider impacts such as improved data quality, enhanced national submissions, and reduced risk of failed trial recruitment. These help to estimate downstream financial benefits. Overall, the model provides the necessary structure for comparing costs and savings, enabling a robust budget impact analysis.

Standard 19: ensure transparency about requirements for deployment

Assuming the full deployment of Genollama, rather than the development of a downstream platform (e.g. for clinical trial recruitment), the two primary requirements relate to infrastructure and data, respectively. Using the deployment at GSTT as a guide, infrastructure requirements are a GPU with at least 24 GB of VRAM. Non-GPU setups are viable, however sites can expect significantly slower inference times. The second natural requirement is access to the source data to standardise, which should have a suitable level of completeness with respect to the Genollama schema. Although the tool is robust to incomplete data, the absence of any of the expected information would render the standardisation process redundant. The same is true of source document (free text) structure; although, once again, Genollama is, by design, robust to a variety of structures, styles that fall wildly outside the content of the training data may necessitate central retraining (discussed). Both completeness and structure will necessitate clinical expertise, making this also a resource requirement for deployment.

Standard 20: describe strategies for communication, consent and training processes to allow the DHT to be understood

In terms of generally understanding the tool and the role it may play in patient pathways, Genollama benefits from sitting alongside a suite of similar models that serve as further examples of the schema standardisation approach (e.g. Oncollama for oncology letters).

In addition to this, technical individuals tasked with deploying the model benefit from open source (licensed), versioned assets (e.g. the schema) via GitHub that document design considerations and choices, as well as fully documented deployment packages. Model cards also provide, among other things, a means of expressing more detailed information about Genollama and the foundation model upon which is it based (Llama), as well as information about any future iterations of the model. Healthcare professionals working either directly with the tool, a downstream function (e.g. trial recruitment), or interacting with the tool's output in other context benefit from existing educational materials (e.g. slide decks) that describe the rationale behind the development of Genollama and an accessible description of the steps taken to do so.

Standard 21: ensure appropriate scalability

Within the current use case at GSTT, Genollama has suitable throughput against a large document corpus with relatively modest computational requirements, suggesting good transferability to a wide range of use cases. Should throughput not be acceptable – e.g. to a different input data setup, or different output presentation requirements for new downstream use cases -- these modest requirements mean that the tool should not require significant additional computational expense to scale. Steps have/are being taken to support scaling from a software perspective also, such as containerising inference packages such that the model could be scaled within a cluster. Scaling of human resources is also a key consideration, and is something that can be supported here via accessible and easily distributed deployment guidelines, which form part of the general communication and documentation of the tool (discussed).

NHS
Genomic AI Network

Work Package 3B Evaluation

Executive summary

Work Package 3B of the NHS Genomic AI Network (GAIN) focuses on the deployment and evaluation of MendelScan to address the "diagnostic odyssey" faced by rare disease patients. GAIN is a national community designed to connect experts across genomic and AI landscapes, fostering commercial partnerships to deploy clinical interventions that accelerate patient diagnosis. This project specifically leverages the expertise of the Genomic Medicine Service Alliances (GMSAs) to bridge the gap between AI-driven data insights and clinical action.

MendelScan

MendelScan, manufactured by Mendelian, is an MHRA-registered Class I MDD software medical device. It operates as a clinical decision support system designed for proactive, asynchronous case-finding. By scanning electronic health records (EHRs), the technology identifies patients with specific clinical patterns that suggest they would benefit from further clinical review, genomic testing, or specialist referral. Initially funded by an Innovate UK grant, MendelScan was a winner of the 2023 NHS AI in Health and Care Award. This facilitated a 20-month AI Award Study, which generated extensive evidence regarding the tool's clinical utility, safety, and technical effectiveness across millions of UK patient records.

GAIN Project Developments: mIBD and Epilepsy

During the GAIN project timeframe, Mendelian developed two new validated algorithms to expand MendelScan's diagnostic reach:

- **Monogenic IBD (mIBD):** Developed in collaboration with the University of Oxford academics, this algorithm identifies high-risk patients suitable for the National Genomic Test Directory (GTD) R15 panel.
- **Rare Epilepsy:** Refined with experts from Oxford and Queen Square, this tool prioritises patients meeting GTD R59 panel criteria, specifically targeting those with drug-resistant phenotypes who are most likely to benefit from genomic intervention.

These algorithms have the potential to assist the NHS in transitioning from a reactive model to a proactive one, ensuring that patients recorded in primary care with markers for these complex conditions are identified for appropriate genomic testing.

Standard 1: the digital health technology (DHT) should comply with relevant safety and quality standards

MendelScan is an MHRA-registered Class I medical device, compliant with the 93/42/EEC Medical Device Directive (MDD). This certification confirms that the device meets essential requirements for safety and performance for its intended use as a patient health record information system. All MendelScan's disease- or phenotype-specific algorithms are validated, versioned and "locked" before deployment.
The technology adheres to several key UK and international standards to ensure clinical and technical safety:

- **Clinical Risk Management:** Complies with DCB0129, ensuring a formal clinical risk management process is embedded in the product's development and maintenance.
- **Quality Management:** The device is manufactured and operates under a Quality Management System aligned with BS EN ISO 13485:2016.
- **Technical Safety:** Follows BS EN 62304:2006+A1:2015 for medical device software lifecycle processes and BS EN ISO 14971:2019 for the application of risk management to medical devices.
- **Data Security:** Annually completes NHS Digital's Data Security and Protection Toolkit (DSPT), meeting the National Data Guardian's 10 data security standards for handling NHS patient data. See Mendelian's listing here.

This multi-layered compliance framework ensures that MendelScan, and all its algorithms, provides a safe and secure environment for identifying undiagnosed rare disease patients within the NHS.

Standard 2: incorporate intended user group acceptability in the design of the DHT

Robust, user-centred design and iterative stakeholder engagement was used in the development of MendelScan.

Patient and Public Involvement (PPI)
A dedicated PPI group, including rare disease advocates and carers, quarterly shapes ethical frameworks. Their input ensured "patient recontact" documents addressed emotional distress with supportive, clear terminology. This aligns with the 2024 Public Perception Survey conducted as part of the AI Award Study, where 92.8% of the 254 respondents endorsed technology-led case-finding. While "cold-calling" was an initial concern, another component in the AI Award Study found patients felt gratitude rather than anxiety, describing the proactive identification as "a light at the end of the tunnel" after years of diagnostic uncertainty. Furthermore, the vast majority of the public support their own GP practices using MendelScan, reinforcing the high level of social license and user-group acceptability.

Clinical Acceptability and Safety
Clinical acceptability and safety have been ensured through vast engagement and co-development of the algorithms with world-leading disease experts within the NHS as well as the incorporation of published guidelines and peer-reviewed literature. For example, for the new mIBD and Epilepsy algorithms, Mendelian integrated expert feedback from the following Key Opinion Leaders (KOLs):

- **Monogenic IBD (mIBD):** The mIBD algorithm logic was developed in collaboration with the University of Oxford's Department of Paediatrics to ensure the tool identifies high-risk patients suitable for the National Genomic Test Directory (GTD) R15 panel.
    - **Professor Holm Uhlig (Professor of Paediatric Gastroenterology):** As a world-leading expert and Director of the Centre for Human Genetics, Professor Uhlig provided the foundational clinical oversight. His leadership of the NIHR Paediatric IBD BioResource ensured that MendelScan's logic reflects the latest research into inborn errors of immunity and mucosal barrier failure, ensuring the tool targets the most clinically relevant phenotypes.
    - **Dr James Charlesworth (Paediatric Registrar & Immunologist):** Dr Charlesworth (PhD in immunology) provided technical guidance on the "treatment atlas" for mIBD. His expertise ensured the algorithm focuses on identifying patients where a genetic diagnosis would justify molecularly targeted treatments, therapies that are often missed or even contraindicated in classical IBD (e.g., anti-TNFs), thereby directly improving clinical management.
- **Rare Epilepsy:** Consultations with Epilepsy KOLs were pivotal in shifting the algorithm from a broad phenotype search to a high-yield clinical decision support tool.
    - **Dr Usha Kini (Consultant Clinical Geneticist, Oxford):** Following meetings with Dr Kini, the algorithm was refined to focus on objective EHR proxies rather than subjective clinical features like dysmorphism, which are poorly captured in primary care data. This ensured the tool remained robust in a real-world GP setting. Furthermore, Dr Kini's audit data, showing that 25% of patients had a direct change in management following an R59 genetic test, provided the evidence base for the algorithm's high clinical utility.
    - **Dr Oliver Ziff (Neurology Registrar, Queen Square):** Dr Ziff provided a specialist prescribing hierarchy (referencing Queen Square standards) which allowed the algorithm to identify "drug-resistant" epilepsy. By identifying patients on 2+ anti-seizure medications whose underlying aetiology remains unknown, the tool targets the cohort most likely to benefit from genomic intervention.

- **Sophie Muir (Timothy Syndrome Alliance):** Engagement with patient advocates ensured that the "actionable insights" generated by the tool addressed the diagnostic odyssey, framing the output as a supportive pathway toward specialist care.

  Real world clinical acceptability was established via multiple pilots, examples of these include an Innovate UK-funded pilot conducted in Lower Lea Valley in 2019 and 2020, two NHS GMSA Transformation Projects across 14 GP practices in 2024 and across a network of GP practices covering over 700k patients in the AI Award Study.

  In total various algorithms have been deployed on over 11-million UK primary care patient records, leading to the identification of around 1,500 potential undiagnosed rare disease patients. Healthcare Professionals (HCPs) confirmed the tool is "not onerous," integrating seamlessly into workflows with a review time of just 5–10 minutes per case. HCPs praised report quality as "more comprehensive than hospital letters," with 50% of flags receiving positive clinical feedback, directly informing management decisions and planned further action was indicated in 30% of these cases.

Standard 3: Consider Environmental Sustainability

MendelScan is designed to drive system-wide efficiency by identifying rare disease patients significantly earlier in their diagnostic journey. Findings from the AI Award Study demonstrate that the technology can avoid thousands of years of the "diagnostic odyssey" across multiple conditions. For example, the tool is projected to save over 3,800 years of diagnostic delay for Alpha-1-antitrypsin deficiency and 2,400 years for Paroxysmal nocturnal haemoglobinuria at a UK population scale.

By accelerating diagnosis, MendelScan removes avoidable healthcare resource utilisation, including repetitive primary care appointments and secondary care investigations. This optimisation led to the technology being awarded the 2023 Horizon Prize (MIT Solve), which specifically recognises innovations that reduce the environmental impact of rare disease healthcare. By minimising the "diagnostic odyssey," MendelScan reduces the total carbon footprint associated with long-term, inconclusive patient journeys within the NHS.

Technically, the platform maintains environmental sustainability by continuously optimising its AI architectures to reduce server load and energy consumption. These digital efficiencies, combined with modelled annual cost savings for conditions like DiGeorge syndrome and BWS, ensure MendelScan supports the NHS "Net Zero" commitment while maximising clinical output per unit of resource.

Standard 4: consider health and care inequalities and bias mitigation

MendelScan is designed to reduce health inequalities by providing an objective, data-driven "safety net" that identifies rare disease patterns regardless of a patient's ability to self-advocate or their proximity to specialist centres. Recognising that the "diagnostic odyssey" is often amplified by socio-economic factors, Mendelian has undertaken extensive health equity research using the Optimum Patient Care Research Database (OPCRD), covering 23 million de-identified records.

To mitigate algorithmic bias, the technology underwent rigorous analytical validation in a case-control study involving over 2 million controls. This process ensured that the algorithms perform consistently across diverse datasets and geographical regions. Furthermore, Mendelian evaluates diagnostic precision at a group level, utilising the Index of Multiple Deprivation (IMD) to monitor whether diagnosis rates vary across different levels of deprivation. By automating the identification of rare symptoms in

primary care records, MendelScan helps to standardise care, ensuring that patients in under-served or highly deprived areas receive the same proactive screening as those in more affluent regions, directly aligning with the NHS Long Term Plan's commitment to tackling health inequalities.

Standard 5: embed good data practices in the design of the DHT

MendelScan adheres to stringent data protection protocols, ensuring full compliance with the UK General Data Protection Regulation (GDPR) and the Data Protection Act 2018. As a Class I Medical Device, the technology operates under a robust Data Protection Impact Assessment (DPIA) and a formal Data Protection Procedure. Mendelian annually achieves "Standards Met" on the NHS Digital Data Security and Protection Toolkit (DSPT), satisfying the National Data Guardian's ten data security standards.

The technology is designed for secure, asynchronous processing of electronic health records (EHRs). Patient data is handled via secure SFTP or encrypted APIs, with storage restricted to UK-based servers. Access is strictly governed by a "least privilege" model and audited through the Technical Data Access Audit. Furthermore, the company undergoes regular independent verification of its technical defences against cyber threats. By ensuring that all data processing is lawful, transparent, and secure, MendelScan maintains the high level of trust required to operate across the NHS.

MendelScan's validated algorithms are validated using the Optimum Patient Care Research Database (OPCRD), one of the UK's largest repositories of de-identified electronic health records (EHRs). This dataset comprises longitudinal, SNOMED CT and CTV3-coded data from over 23 million patients across 1,100 primary care practices, ensuring the training environment reflects the diversity of the UK population in terms of age, geography, and socio-economic status.

For each disease-specific algorithm, 'ground truth' was established through a rigorous process of clinical expert review and historical case identification. Mendelian utilise a case-control methodology, where confirmed cases of rare diseases were matched against over 2 million controls to determine diagnostic thresholds. No synthetic data was used in the validation of these algorithms; performance metrics are entirely based on real-world clinical evidence.

MendelScan adheres to MHRA Good Machine Learning Practice (GMLP) by utilising a robust methodology that prioritises clinical causality over simple pattern recognition. By maintaining a strict separation between training, validation, and holdout datasets, effectively eliminating data leakage and ensuring that performance estimates remain unbiased when applied to external populations. This approach, validated against a high-fidelity dataset representing 24% of the UK population, minimises overfitting to specific samples and maximises the generalisability of "locked" algorithms. Furthermore, by codifying the "true patient journey" through expert-backed logic, these models offer full explainability; they signpost clinicians to the specific clinical proxies driving a prediction, thereby ensuring the output is directly actionable and holds genuine clinical utility.

Standard 6: define the level of professional oversight

MendelScan is designed as a clinical decision support tool that functions as an adjunct to, rather than a replacement for, professional medical judgement. The level of professional oversight is structured through a mandatory "human-in-the-loop" architecture, ensuring every output is reviewed by a qualified healthcare professional (HCP) before any clinical action or patient contact occurs.

The anticipated level of oversight involves a case-by-case expert review. When a patient is flagged, a clinical report is generated for the GP or a designated clinical reviewer. As specified in MendelScan's

Intended Use Statement, clinicians are instructed not to rely solely on these reports but to use them as a "safety net" to trigger further investigation. The AI Award Study demonstrated that this oversight is proportionate and acceptable, with GPs spending an average of 5–10 minutes reviewing each flagged case to validate the findings against the full patient record.

To ensure long-term calibration, Mendelian conducts periodic overarching reviews via Post-Market Surveillance (PMS) and the technical data access audit. These processes monitor for "overridden" outputs where clinicians choose not to refer a flagged patient, allowing for continuous refinement of the algorithms to ensure they remain aligned with best clinical practice and the risk profile of a Tier C technology.

Standard 7: show processes for creating reliable health information

MendelScan ensures all health information, particularly the disease-specific clinical summaries provided to GPs, is valid, accurate, and aligned with the best available clinical evidence. The information is derived from high-quality sources, including NICE guidelines, peer-reviewed literature, and consensus statements from recognised UK patient and professional organisations (e.g., Genetic Alliance UK).

The process for creating and maintaining this information is governed by a formal Design Review. Each validated algorithm is accompanied by a clinical report template that undergoes a "human-in-the-loop" review. To ensure the information remains accurate and comprehensive, reports are designed to be "standard" clinical documents that do not require extra qualifications for a doctor to comprehend.

Maintenance is managed through a defined lifecycle: reports are reviewed and updated by clinical experts at defined intervals. Specifically, the system includes a check to ensure that the criteria and clinical context used in any generated report have been reviewed or updated within the last 12 months. This ensures that as medical knowledge of rare diseases evolves, the information provided to the NHS remains aligned with current best practice and relevant to the target population.

Standard 8: show that the DHT is credible with UK professionals

Expert-Led Clinical Foundations
The credibility of MendelScan’s disease-specific algorithms, such as those for monogenic IBD (mIBD) and Rare Epilepsy, is anchored in direct collaboration with leading NHS specialists. By integrating the expertise of Professor Holm Uhlig and Dr James Charlesworth (University of Oxford), the mIBD algorithm aligns with the National GTD R15 criteria and targets phenotypes where genetic diagnosis enables precision, molecularly targeted treatments. Similarly, the Rare Epilepsy algorithm was refined through consultation with Dr Usha Kini (Oxford) and Dr Oliver Ziff (Queen Square), shifting the tool towards objective EHR proxies and drug-resistance hierarchies. This ensures the technology is not merely a data-pattern recognition tool, but a clinically backed decision support system that understands the "true patient journey."

Validation through NHS Integration
The tool’s credibility in practice is evidenced by its deployment across over 11 million UK primary care records, identifying approximately 1,500 potentially undiagnosed patients. MendelScan has been a focal point of two GMSA Transformation Projects and an Innovate UK pilot, covering 14 practices and a pilot as part of the AI Award Study covering 700,000 patients.

Standard 9: provide safeguarding assurances for DHTs where users are considered to be in vulnerable groups, or where peer-to-peer interaction is enabled

This standard is not applicable as the users of this technology are health care professionals.

Standard 10: describe the intended purpose and target population

MendelScan acts as a clinical decision support system designed to assist in the proactive identification of individuals with potential signs and symptoms of undiagnosed or misdiagnosed rare and hard-to-diagnose diseases recorded in their EHRs. Its intended purpose is to process real world electronic health records (EHRs) asynchronously to identify patterns and anomalies that may warrant further clinical investigation or specialist referral.

The target population comprises all patients registered within health settings at population scales. Inclusion criteria involve any patient with a medical history within primary or secondary care EHRs. Subgroups of particular interest include those in the "diagnostic odyssey" phase, patients with high healthcare utilisation but no definitive diagnosis.

The expected uptake profile is projected at scale through Integrated Care Systems (ICSs) and / or regional or national Secure Data Environments (SDEs).

Standards 11 and 12: describe the current pathway or system process. Describe the proposed pathway or system process using the DHT

The Current Pathway: The Diagnostic Odyssey
The current clinical pathway for rare diseases in the UK is predominantly reactive and fragmented. Patients typically present to primary care with non-specific, multi-system symptoms that do not clearly align with common conditions. This initiates the "diagnostic odyssey," a cycle of repeated GP appointments, inappropriate secondary care referrals, and inconclusive investigations. National data suggests that, in patients who do ultimately receive a diagnosis, it takes an average of over five years, involving multiple specialist boundaries, during which time a patient's condition may irreversibly deteriorate. This pathway is characterised by high resource wastage and significant patient distress due to the "information challenge" faced by GPs in identifying over 10,000 rare diseases.

The Proposed Pathway: Proactive Case-Finding
MendelScan allows for the introduction of a proactive, data-driven pathway. The technology is implemented as an adjunct "safety net" that complements existing care without replacing clinical staff with the following steps:

- **Identification:** MendelScan asynchronously scans primary care EHRs to identify patients whose historical clinical data match rare disease patterns for the deployed algorithms.
- **Clinical Review:** Instead of waiting for a patient to present with further disease progression MendelScan generates a comprehensive report for the reviewing HCP who then conducts a deeper review of the patient's full EHR. The AI Award Study confirms this step requires only 5 to 10 minutes of professional oversight.
- **Informed Testing or Referral:** The HCP uses their clinical review to make an informed decision on possible patient recontact and next steps, which may entail offering the patient a specific test or a referral to an appropriate specialist.

Impact on Infrastructure and Workforce
Implementation requires minimal infrastructure changes as MendelScan integrates with existing NHS digital architecture and EHR platforms. While it introduces an additional step (the report review), it

reduces long-term workforce pressure by avoiding thousands of unnecessary GP and specialist appointments. Training is straightforward, and is discussed further in Standard 20.
The logic model attached shows how MendelScan is anticipated to change activity across the whole rare disease pathway. It sets out the inputs, activities, outputs, outcomes, and impacts required to compare current care with the enhanced pathway enabled by AI-driven case finding. For health impacts, the model identifies measurable improvements such as increased detection of rare diseases, earlier diagnosis, faster initiation of appropriate treatment, reductions in diagnostic odyssey, and more equitable case identification across regions and deprivation groups. These provide a comparison with the slower, less consistent identification seen in usual care. For resource impacts, the model helps to quantify the shift in workload at each stage - for example GP time for second reviews, changes in referral volumes, genomic or biochemical testing activity, and reductions in inappropriate referrals or unnecessary investigations - allowing assessment of where resource use increases or decreases relative to current processes. Finally, the model outlines likely cost impacts such as reduced costs from unnecessary activity, avoided complications from late diagnosis, and laboratory savings from fewer non-diagnostic tests. Together, these elements offer a structured way to estimate system-wide benefits attributable to MendelScan's introduction.

Standard 13: describe the expected health, cost and resource impacts compared with current care or system processes

The Current Cost of the Diagnostic Odyssey
The current reactive pathway for rare diseases is characterised by a significant "diagnostic odyssey," with patients waiting an average of over five years for a definitive diagnosis. This delay is not only detrimental to patient health but also incurs substantial avoidable costs for the NHS through repetitive primary care consultations and non-targeted investigations. Research conducted with Imperial College Health Partners (ICHP) across a subset of 426 rare disease-specific ICD-10 codes confirms the scale of this burden:

- **Higher Hospital Costs:** Rare disease patients cost the NHS an average of £13,064 in the decade prior to diagnosis, more than double the £5,910 average for the general hospital population.
- **Paediatric Impact:** In patients diagnosed at 10 years of age or younger, the cost disparity is even more pronounced, with rare disease patients costing £9,327 compared to just £2,240 for age-matched controls.
- **Inefficient Resource Use:** Prior to diagnosis, rare disease patients experience significantly higher hospital activity, including an average of 34.6 outpatient appointments compared to 22.3 in the rest of the population. This cohort also undergoes a higher volume of costly, invasive investigative procedures, such as laparoscopies, spinal punctures, and CT scans.

Anticipated Impact of MendelScan
Proactive Case-Finding MendelScan shifts the system towards proactive identification, delivering health benefits through earlier intervention and improved care coordination.

- **Truncating the Odyssey:** Modelling demonstrates that MendelScan can avoid thousands of years of diagnostic delay across the UK population, as examples 3,842 years for alpha-1-antitrypsin deficiency and 2,483 years for paroxysmal nocturnal haemoglobinuria if these algorithms were to be deployed at a UK population scale.
- **Clinical Utility in Epilepsy and mIBD:** For Rare Epilepsy, where a genetic diagnosis carries profound management implications, in a retrospective analysis MendelScan identifies approximately 40% of severe epilepsy cases with a monogenic aetiology. In mIBD, identifying the correct genetic driver allows for molecularly targeted treatments and avoids standard IBD therapies or management (such as anti-TNFs or surgery) that may be ineffective or detrimental.
- **Direct Cost Savings:** By reducing inappropriate healthcare activity, modelling for DiGeorge Syndrome alone indicates potential UK population scale savings of over 6,200 GP appointments

and 2,000 secondary care appointments, equating to at least £641,332 in direct savings. While MendelScan introduces a minimal initial resource requirement (the 5–10 minute clinician review time), this is significantly offset by the prevention of long-term hospitalisations and unnecessary procedures. These benefits align with the NHS Long Term Plan by optimising specialist resource allocation and reducing the structural inefficiencies of rare disease management.

The logic model attached provides a structured, end-to-end view of how MendelScan changes the rare-disease pathway, making it possible to describe expected health, cost, and resource impacts compared with current care. It maps the chain from inputs (EHR data, GP time, genomic testing capacity) through activities (algorithmic case-finding, GP review, referrals) to outputs and longer-term impacts. This allows clear comparison with existing processes where rare-disease identification is slower, more variable, and often dependent on chance presentations. The model identifies health impacts such as increased appropriate referrals, more genomic/biochemical testing, earlier diagnosis, reduced diagnostic odysseys, and optimised treatment initiation, all of which can be contrasted with current care that often involves delayed recognition and fragmented pathways. On resource use, the model specifies where activity will rise or fall, for example, additional GP review time, changes in specialist clinic workload, and reductions in inappropriate referrals and unnecessary tests - enabling assessment of net resource shifts against the baseline system. Finally, it outlines cost impacts such as reduced unnecessary activity, avoided costs of complications from late diagnosis, and laboratory savings, helping future quantification of the economic difference between MendelScan enabled care and standard pathways.

Standard 14: provide evidence of the DHT's effectiveness to support its claimed benefits

MendelScan has demonstrated its effectiveness through extensive analytical validation and real-world clinical utility evaluations. Under the AI Award Study, 34 algorithms underwent rigorous analytical validation using the Optimum Patient Care Research Database (OPCRD), which contained 23 million de-identified UK primary care records at that time. This study matching confirmed cases against over 2 million controls provided high-quality evidence of the tool's accuracy and reliability in a setting directly relevant to the NHS.

The technology's impact on clinical management is evidenced by its ability to significantly reduce the time to diagnosis. AI Award Study modelling projected that MendelScan could avoid thousands of years of diagnostic delay across the UK, including 3,842 years for alpha-1-antitrypsin deficiency and 2,483 years for paroxysmal nocturnal haemoglobinuria. Real-world evaluation in the GMSA Transformation Projects across 14 practices showed that 1 in2 flags received positive feedback from clinical reviewers, with GPs confirming the reports led to appropriate specialist referrals.

Furthermore, the AI Award Study provided qualitative evidence of clinical utility. UK HCPs described the MendelScan reports as "more comprehensive than hospital letters," noting they provided the necessary evidence to confidently act on rare disease suspicions that would otherwise remain undetected in standard time-pressured appointments. By acting as a proactive "safety net," MendelScan delivers reliable test results that directly influence the diagnostic pathway and improve patient-relevant outcomes. Specifically, with regards to the new mIBD and Epilepsy algorithms, these have demonstrated effectiveness as population-level case-finding tools through rigorous analytical validation using OPCRD.

Monogenic IBD (mIBD) Algorithmic Validation

The effectiveness of the mIBD algorithm (Iteration 2, v7.0) was validated by codifying the GTD R15 panel and clinical consensus guidelines into automated logic.

- **Performance in Real-World Data:** In a scan of 28 million records, the algorithm identified 189,880 patients with a diagnostic code for IBD.
- **High-Yield Case Finding:** The tool successfully isolated 107 cases exhibiting both an IBD diagnosis and a rare phenotype associated with mIBD (e.g., Chronic Granulomatous Disease), equating to a flagging rate of 3.8 cases per million EHRs.

- **Scoring Accuracy:** Utilising a clinical point threshold, the algorithm identified fewer than 500 records with a score of less than 5, matching the high index of suspicion required for further investigation. Expert quality control (QC) confirmed an 87.7% accuracy rate for patients meeting the clinical threshold for a full EHR screen.

Rare Epilepsy Algorithmic Validation
Validation for the Rare Epilepsy algorithm focused on the GTD R59 panel criteria to identify the approximately 40% of severe epilepsy cases with a monogenic aetiology.

- **Objective Stratification:** Within a 200,000-patient epilepsy cohort, the algorithm identified 12,048 patients diagnosed before the age of two who meet the R59 panel criteria for genetic testing..
- **Predictive Markers:** Effectiveness was further validated by mapping disease severity through a specialist prescribing hierarchy. In a cohort of 1,974 patients diagnosed since 2010, the algorithm tracked those progressing to a third anti-seizure medication, a journey taking an average of 4.08 years, to highlight drug-resistant cases for genomic review.
- **Clinical Yield:** This objective methodology is supported by audit data showing a 25% diagnostic yield and subsequent management change for patients identified via these criteria.

Standard 15: show real-world evidence that the claimed benefits can be realised in practice

MendelScan's performance as a population-level case-finding tool is supported by extensive RWE from NHS primary care deployments. The evidence demonstrates that the tool successfully identifies discriminatory clinical patterns in EHR data to assist HCPs in identifying patients for clinical review.

Validation of Case-Finding and Decision Support
Deployment across 14 practices in two GMSA Transformation Projects (covering 200,000 patients) provided the primary evidence base. An independent evaluation by St George's University of London confirmed that MendelScan successfully identifies patients who warrant further investigation, fulfilling its intended purpose as a decision-support tool. Crucially, the RWE supports Mendelian's claim that MendelScan does not act as a "screening" tool but rather as a "case-finding" tool; it identifies those with established clinical markers in their records who have nonetheless been overlooked in routine care.

Clinical Utility and HCP Acceptability
The AI Award Study, conducted by Health Innovation East, provided qualitative RWE regarding the tool's integration into time-pressured workflows:

- **Evidence-Based Review:** HCPs reported that MendelScan reports were "more comprehensive than hospital letters," providing the necessary context (the "why" behind the flag) to allow for an independent clinical decision.
- **Workflow Efficiency:** RWE confirmed that the tool is not used as the sole basis for diagnosis; instead, clinicians spend an average of 5 to 10 minutes reviewing the flagged patterns to determine if a referral is appropriate.
- **Positive Clinical Action:** In practice, 50% of flags received positive clinical feedback, with 30% of reviewed cases resulting in a direct change in management or the initiation of a new diagnostic pathway.

Safety and Patient Perspective
RWE from the AI Award Study found no evidence of unintended negative impacts, such as "over-medicalisation" or patient anxiety. Instead, patients expressed gratitude for the proactive identification of

their discriminatory patterns, describing the insight as a "light at the end of the tunnel." These findings confirm that MendelScan achieves its intended health benefits, identifying potentially overlooked patients for further review, while remaining firmly within the "Inform" (Tier C) clinical management framework. Neither the mIBD nor epilepsy algorithms have been deployed yet, therefore no real-world evidence exists for these at present.

Standard 16: the company and evaluator should agree a plan for measuring usage and changes in the DHT's performance over time

Mendelian maintains a comprehensive plan for measuring the ongoing usage and performance of MendelScan, governed by its Post-Market Surveillance (PMS) Plan and Quality Management System. To track real-world impact, the system collects anonymised, aggregate data on service user outcomes, such as the number of patients identified, the conversion rate of "flags" to clinical referrals, and the accuracy of those referrals as confirmed by GPs.
Algorithm performance is monitored through a series of technical and clinical controls:

- **Performance Tracking:** Automated technical audits log data access and processing accuracy to ensure system stability.
- **Algorithm Re-versioning:** Algorithms are updated in subsequent versions (e.g., PNH v3.5) following structured design reviews. Regular system checks ensure that no report is generated using clinical criteria that have not been reviewed or updated within the last 12 months.
- **Feedback Loops**: A formal process is in place to capture feedback directly from participating HCPs, allowing for the detection of any decreasing performance or unintended outcomes in specific patient subgroups.
- **Reporting:** Changes in performance and significant updates are reported to evaluators and commissioning bodies through quarterly clinical reviews and annual PMS reports.

This independent overview process ensures that MendelScan remains calibrated against the latest clinical best practice while maintaining high performance across diverse UK populations.

Standard 17: provide a budget impact analysis

The attached logic model provides a foundation for developing a budget impact analysis by identifying the costs, resource use, and system changes associated with both current care and the MendelScan supported pathway. It makes it possible to translate each step into budgetary consequences. For instance, the model specifies the time required for MendelScan first reviews, the approx. 20 minutes of GP time for second reviews, and the increase in genomic or biochemical testing, all of which represent direct costs that can be compared with current care pathways. At the same time, the logic model highlights offsetting savings such as reductions in inappropriate referrals, reductions in unnecessary genomic/biochemical testing, and reduced healthcare resource costs from unnecessary activity. These elements can be costed and used to estimate net financial impact. The logic model also outlines longer-term cost impacts, including avoided complications from late diagnosis, earlier treatment initiation, and reduced diagnostic odysseys, enabling scenario modelling across multiple budget years. Overall, the logic model provides important detail needed to build a robust budget impact analysis comparing MendelScan with existing processes.
Mendelian has modelled MendelScan’s budget impact within various diseases yet only looking at the diagnostic journey. This characterised by a modest initial investment in proactive screening that yields significant downstream savings by truncating the "diagnostic odyssey". This analysis was based on a

target population of all patients registered in UK primary care, with validation datasets covering 23 million records (~24% of the UK population), which was then extrapolated up to the full UK population.

Direct Costs and Resource Use

Direct costs for implementing MendelScan include the software licensing, secure IT integration (e.g., via EMIS IM1), and the clinician review time. Real-world evidence from the AI Award Study confirms that the primary workforce impact is limited to 5–10 minutes per flagged patient for clinical review.

Cost Savings and Efficiency

Compared to current reactive care, MendelScan reduces the volume of inappropriate healthcare activity. Population-scale modelling for various disease algorithms indicates substantial direct cost savings. For example:

- **DiGeorge Syndrome:** Estimated savings of at least £641,332 by avoiding 6,202 GP appointments, 2,067 secondary care appointments, and 10 hospitalisations.
- **Paroxysmal Nocturnal Haemoglobinuria (PNH):** Projected savings of £770,366 through the avoidance of 7,450 GP visits and 2,483 specialist consultations.
- **Beckwith-Wiedemann Syndrome (BWS):** Anticipated savings of £160,806.

These figures represent conservative estimates based on direct NHS tariff costs for appointments and procedures. The cumulative budget impact expected across all algorithms demonstrates that MendelScan is a cost-effective intervention that improves system efficiency by directing specialist resources to the patients most likely to benefit, thereby reducing the multi-year financial burden of undiagnosed rare diseases on the NHS.

Standard 19: ensure transparency about requirements for deployment

MendelScan is designed for seamless integration into existing NHS primary care digital infrastructure. The technology functions as a cloud-based Software as a Medical Device (SaMD) that asynchronously processes EHR data.

Data Requirements and Format

The system requires input data in structured formats from primary care EHR systems such as EMIS and TPP or structured or pre-processed unstructured secondary care data. To ensure accuracy and interoperability, data must be standardised using clinical ontologies or terminologies (SNOMED CT, CTV3, LOINC, HPO or ICD). The input data includes longitudinal patient records, encompassing clinical codes, medications, and laboratory results, which are processed via secure SFTP or encrypted APIs.

Data Flow and Integrity

A formal data flow map governs the deployment, ensuring that patient data always remains within UK-based servers. To maintain data quality, the system includes automated checks to verify that the analysis is performed on the latest available instance of a patient's medical record. MendelScan is designed with a high level of tolerance for the complex, "noisy" nature of real-world EHR data; however, it relies on coded entries rather than free text to maintain its validated performance levels.

Infrastructure Requirements

The minimum infrastructure requirement for deployment is a standard NHS clinical workstation with a modern web browser (e.g., Chrome) and reliable internet connectivity to access the generated clinical reports. MendelScan can be deployed into secure data environments. Or, with the right data sharing

agreements in place, the computational processing can be handled within Mendelian's secure, UK-hosted cloud environment.

Standard 20: describe strategies for communication, consent and training processes to allow the DHT to be understood

Communication with Health and Care Professionals
The primary communication tool for HCPs is the MendelScan Clinical Report, which is designed as a "standard" clinical document intended to be intelligible without extra qualifications. Each report clearly identifies the specific clinical patterns or "flags" detected, the source of the disease information used (e.g., NICE guidelines), and the context of the identified anomalies. This allows clinicians to interpret the output within the broader context of the patient's record.

Training and Education
Mendelian provides structured training for end users to ensure they can interpret the technology's outputs correctly. Training emphasises that the tool is an adjunct, not a replacement for clinical judgment, and that no diagnostic or referral decision should be made solely based on the report. This is reinforced in practice contracts, which mandate that all flags be reviewed by a qualified HCP.

Patient Communication and Consent
Patient-facing materials, such as information sheets and consent forms, have been refined through Patient and Public Involvement (PPI) to ensure they use clear, non-medicalised language. For example, terms like "positive" were modified to avoid confusion regarding whether the result meant "good news" or a potential diagnosis. While MendelScan processes data asynchronously as part of clinical care, any proactive follow-up initiated by the GP follows standard NHS consent protocols. The AI Award Study confirmed that patients found these communications acceptable, particularly when signposted to additional support for the emotional impact of a potential rare disease diagnosis.

Standard 21: ensure appropriate scalability

MendelScan is built upon a cloud-native architecture designed to scale dynamically in response to the demands of large-scale NHS populations. To ensure performance reliability, Mendelian conducts regular load testing that simulates high-concurrency scenarios, verifying that the system can process millions of patient records without degradation in speed or clinical accuracy. This testing is directly calibrated to the expected uptake across the entire NHS, nationally or via regional Integrated Care Boards (ICBs), where entire regional populations may be screened simultaneously.

Primary Care EHR Providers
In primary care, the two main EHR providers are Optum (EMIS) and TPP. TPP was not engaged directly, yet Mendelian is able to use a bulk extract mechanism to gain access to EHRs on the TPP system. This is a clunky mechanism, yet it is sufficient. Mendelian have previously engaged extensively with EMIS (now Optum) exploring the ability to deploy MendelScan inside their platform, this is technically feasible, yet costs are prohibitive. Mendelian is accredited to extract EMIS data via the NHS IM1 mechanism.
A significant challenge to scaling via these routes is the requirement for individual GP practice engagement, buy in and then contracting. In many instances Mendelian has opted to work with larger GP supergroups (Modality Partnership and Operose Health) and primary care networks (PCNs). Yet ultimately this technology should be deployed via larger regional or national data aggregators.

Third-Party Technology Partners

Mendelian has established a partnership with Optimum Patient Care (OPC), a not-for-profit social enterprise that aggregates NHS primary care EHR data for both research and clinical utility projects. Together with OPC, NHS clinical teams can run real-world primary care rare disease case-finding projects over almost 10-million primary care records from GP practices across England who have opted in for this service. This allows impressive scale, yet the OPC costs are high and prohibitive in most scenarios. Mendelian is in the process of exploring partnership opportunities with a number of third-party NHS suppliers who may facilitate large scale data access for NHS-led regional or national projects. These include Magentus and Eclipse and a number of shared care record providers such as Graphnet and Orion.

NHS Secure Data Environments (SDEs)

Furthermore, Mendelian is actively exploring the use of NHS Secure Data Environments (SDEs) for regional and national deployments. By operating within SDEs, MendelScan can leverage high-performance compute resources while maintaining the highest standards of data security.

Within the GAIN project, Mendelian stakeholders engaged deeply with the Wessex (acting as the "builder" platform) and Kent, Medway and Sussex (KMS) SDEs. A dedicated task team met weekly for 10 months to coordinate the deployment, successfully completing the technical mapping and documentation required for integration.

The engagement identified a clear distinction between data insight and clinical action, which defined the project's trajectory:

- **Phase 1: Insight Generation:** This involved using de-identified data within the SDE to generate aggregate and individual reports on undiagnosed rare diseases. From a technical and Data Access Committee (DAC) perspective, this phase was feasible and met the standard requirements for SDE operation.
- **Phase 2: Clinical Intervention:** This phase required the GMSA to use Phase 1 insights to contact GP practices for patient interventions.

Despite the technical success of Phase 1, Phase 2 encountered significant "translational" challenges that ultimately stalled the deployment:

Governance & Remit Limitations

A fundamental friction point emerged regarding the remit of an SDE. While SDEs are designed to authorise access to data for research, they do not currently have the mandate to authorise interventional clinical studies. Because Phase 2 moved into "Direct Care," it fell outside the traditional SDE research governance framework. Furthermore, moving from de-identified research data back to identifiable patient data for GP intervention required a level of IG clearance (such as CAG251) that the current SDE structure was not positioned to facilitate for this specific use case.

Clinical Stakeholder Alignment

The SDE’s engagement with Local Medical Committees (LMCs) revealed that technical readiness does not equate to clinical buy-in. The proposal faced resistance from GP leaders due to:

- **Unclear GP Practice Value Proposition**: A perceived lack of articulated benefits for individual busy practices.
- **Primary Care Resources**: A need for more granular detail on the proposed interventions to ensure they wouldn't overwhelm primary care.

- **Trust Risk**: Concerns that re-identifying patients via the SDE might undermine the "de-identified for research" promise made to practices when they originally contributed data to the SDE.

Public and Patient Perspectives
Engagement with the Patient and Public Advisory Group (PPAG) was broadly positive regarding the potential benefits of finding undiagnosed patients. However, they highlighted the need for a highly sensitive communication plan when contacting identified patients, marking another layer of complexity for the delivery model.

Route Forward
Mendelian will prioritise direct engagement with GP forums and system leaders to build a clinical "pull" for the technology. By securing clinical buy-in and defining a clear direct-care pathway first, the SDE can then act as the secure engine to power a pre-approved clinical workflow, rather than being the body responsible for authorising the intervention itself.

Work Package 3C Evaluation

Executive summary

Inherited cardiac conditions (ICC) are a major cause of heart failure and sudden cardiac death in young and otherwise healthy individuals. Although diagnostic criteria are well established, many patients present only after catastrophic events such as cardiac arrest or aortic dissection. In routine NHS practice, subtle warning signals are often embedded within ECG archives, imaging reports, lipid profiles, and clinic correspondence but remain unrecognised because they are dispersed across multiple systems and largely unstructured. Approximately 80 percent of NHS clinical data exists as free text, making systematic, scalable identification of eligible patients extremely challenging.

Work Package 3C (WP3C) was designed to address this gap by developing and evaluating language AI approaches to identify patients with ICC using routinely collected NHS data. Leveraging the NHS-developed CogStack platform and advanced natural language processing, the programme focused on extracting structured phenotypes from unstructured records, integrating Human Phenotype Ontology coding, aligning patients to National Genomic Test Directory criteria, and streamlining referral to specialist cardiology and clinical genetics services.
WP3C established secure technical and clinical governance infrastructure and developed phenotyping models for hypertrophic cardiomyopathy, aortopathy, and familial hypercholesterolaemia, subsequently extending to dilated cardiomyopathy. Retrospective validation was undertaken across partner Trust datasets with external cross-site testing. Model performance was high across conditions, with strong balanced accuracy and stable transportability between Trusts. This was followed by shadow-mode prospective deployment to evaluate real-world performance, safety and pathway integration.

The pilot demonstrates that AI can systematically identify patients previously unrecognised as having an ICC and flag individuals who meet National Genomic Test Directory criteria for genetic testing but had not been referred. Evaluation has combined algorithm performance metrics, pathway analysis and early implementation of integrated ICC and genetics models, enabling real-world validation of AI-identified cohorts and assessment of downstream clinical activity.

An embedded Patient and Public Involvement and Engagement programme invited all patients attending pilot pathways to participate in interviews and focus groups, ensuring evaluation of patient experience, trust, transparency and equity considerations and informing iterative refinement of implementation.
In parallel with rapid advances in large language models and generative AI, WP3C has evolved to include a further workstream focused on foundation models, multimodal data integration across imaging, ECG and genomics, automated referral support, and dynamic risk stratification to enable proportionate follow-up and prioritisation.

Next steps focus on translating validated prototypes into prospective real-world clinical deployment. Multi-site external validation and formal real-world effectiveness and budget impact evaluation are ongoing in partnership with the AI Centre for Value Based Healthcare at Guy's and St Thomas'. Implementation is progressing through structured shadow-mode deployment to rigorously assess model performance, safety, bias and workflow integration using live data. Subject to evaluation, this will transition to phased embedding within the live clinical environment, enabling controlled assessment of impact on referral patterns, genetic testing uptake, pathway efficiency, health outcomes and system sustainability prior to broader NHS scale-up.

Standard 1: the digital health technology (DHT) should comply with relevant safety and quality standards

The WP3C digital health technology was built on CogStack, an established, open-source platform deployed within secure NHS Trust infrastructure. The system operates in compliance with the NHS Data Security and Protection Toolkit (DSPT) requirements, adhering to UK GDPR, the Data Protection Act 2018, and local information governance policies. All data processing took place within approved secure environments at GSTT and KCH, supported by completed ISO27001-accedited processes including Data Protection Impact Assessments (DPIAs), role-based access controls, encryption in transit and at rest, and comprehensive audit trails.
Clinical safety governance followed NHS Digital clinical risk management standards (DCB0129 and DCB0160), with a clearly defined intended use, documented hazard logs, and mitigation strategies. Prospective deployment was initiated in shadow mode to ensure safety prior to influencing clinical care, with clinician oversight retained as the final decision-making authority.
Model development and validation included structured retrospective review by clinical experts, alongside monitoring for performance, bias and model drift. The system functioned as clinical decision support, aligned with MHRA guidance and NICE evidence standards for digital health technologies.

Standard 2: incorporate intended user group acceptability in the design of the DHT

The WP3C DHT was developed with its primary users at the centre of the design process. These included secondary care clinicians such as cardiologists, clinical geneticists, specialist nurses and multidisciplinary team members, alongside NHS clinical data scientists, health informatics specialists and Trust IT leads. Clinical leadership was embedded from inception to ensure that model outputs were clinically interpretable, pathway aligned, and actionable within established ICC and genomic testing workflows. User interfaces, dashboards and case lists were co-designed to integrate directly within existing electronic health record environments, avoiding duplication of effort or parallel systems. Technical architecture decisions were made in collaboration with local informatics and IT teams to ensure compatibility, security and maintainability within NHS infrastructure.
Prospective shadow-mode deployment enabled structured feedback from clinical and technical users, allowing iterative refinement of thresholds, output presentation and workflow integration to optimise usability, safety and adoption.

Standard 3: Consider Environmental Sustainability

The system was built on CogStack, an open-source platform operating within existing NHS digital infrastructure, thereby avoiding the need for new large-scale cloud procurement or external data transfer. Models were deployed locally within ISO27001-accredited Trust secure environments, with both small-language-models and large-language models thereby optimising energy costs.
Computational efficiency was prioritised, including optimisation of model size and inference workflows to balance performance with resource use. Shadow-mode deployment further reduced unnecessary clinical activity. By enabling earlier diagnosis and more targeted referral, the technology has the potential to reduce avoidable investigations, duplicate testing and emergency presentations, contributing to more efficient and sustainable care delivery.

Standard 4: consider health and care inequalities and bias mitigation

Addressing health inequalities was a core design principle of WP3C. The DHT enabled systematic, population-level identification of patients using routinely collected NHS data, rather than relying on individual clinician recognition or referral. This reduced dependence on variable awareness, experience or unconscious bias, and supported equitable application of National Genomic Test Directory criteria across the entire patient population.
Equity monitoring was embedded within evaluation. Early audit data demonstrated that, while existing ICC clinics disproportionately served White patients relative to local population demographics, the AI-supported identification process surfaced a higher proportion of Black patients meeting pathway and genetic testing criteria. This suggests that systematic case-finding may help mitigate referral bias and uncover previously under-recognised disease burden.
Ongoing prospective deployment includes continued monitoring for performance variation, differential false positive or negative rates across demographic groups, and transparent reporting to ensure bias mitigation remains an active and iterative process.

Standard 5: Embed good data practices in the design of the DHT

WP3C embedded good data practices through structured data governance, transparent model development and reproducible workflows. CogStack provided a standardised framework for ingesting and indexing heterogeneous NHS data sources, including structured records and free text, with clear data provenance and metadata tagging.
Clinical concepts were mapped to recognised terminologies such as SNOMED CT and Human Phenotype Ontology to promote interoperability and consistent phenotyping across sites. Annotation protocols were defined *a priori*, with documented labelling standards and expert clinical adjudication to ensure consistency and reduce variability.
Model development followed a controlled lifecycle, including separation of development and validation datasets, version control of code and models, and documented performance reporting. Outputs were traceable to source text to support explainability and clinical review.
Ongoing deployment incorporated structured monitoring of data completeness, annotation quality and model stability to ensure that system performance remained robust as data sources and clinical documentation patterns evolved.

Standard 6: Define the level of professional oversight

WP3C was explicitly designed as a clinical decision support system with defined and proportionate professional oversight. The technology did not make autonomous clinical decisions and did not directly

trigger referrals or testing. Instead, it generated structured case lists and flags for review by appropriately trained clinicians within established ICC and clinical genetics pathways.
During retrospective validation and shadow-mode deployment, all AI-identified cases were reviewed by cardiologists and clinical geneticists prior to any change in patient management. Final decisions regarding referral, genetic testing or treatment remained the responsibility of the supervising consultant and multidisciplinary team.
Clear lines of accountability were maintained, with named clinical leads responsible for oversight of model deployment, monitoring and governance. Escalation processes were defined for unexpected outputs or safety concerns. This approach ensured that the DHT augmented clinical judgement while preserving professional responsibility and regulatory compliance within NHS practice.

Standard 7: Show processes for creating reliable health information

WP3C established structured and transparent processes to ensure that the health information generated by the DHT was reliable, reproducible, and clinically valid. Model development followed predefined

protocols, including clearly specified inclusion criteria, standardised annotation frameworks and expert clinical adjudication. Performance was assessed using recognised metrics such as precision, recall and F1-score, with results documented and reviewed by multidisciplinary clinical and data science teams.
Outputs were designed to be traceable to source clinical text, enabling clinicians to verify context and accuracy. Structured phenotypes were mapped to recognised terminologies to support consistency and comparability across cases.
Reliability was further strengthened through retrospective validation, cross-site testing and shadow-mode prospective evaluation prior to live deployment. Continuous monitoring processes were established to detect performance drift, documentation changes or unexpected variation. This approach ensured that generated information remained accurate, clinically interpretable and appropriate for decision support within established ICC pathways.

Standard 8: Show that the DHT is credible with UK professionals

The WP3C DHT builds on CogStack, an NHS-developed and NHS-deployed platform that has been implemented across many NHS Trusts in the UK and adopted by other healthcare providers internationally. Its use across multiple specialties, including oncology, mental health and acute medicine, has established a strong operational track record and a mature governance framework within real-world clinical environments. Using this platform-based approach means that WP3C DHT can leverage broad stakeholder involvement from many user-groups – clinicians, patients, NHS managers, researchers as it maximises on NHS assets, NHS workforce and builds on UK plc's sovereign AI capabilities in a sustainable way.
The programme has undergone professional scrutiny through presentation at the British Inherited Cardiovascular Conditions Society, the European Society of Cardiology Digital and AI Summit 2026 and the American Heart Association Scientific Sessions. Engagement with The Alan Turing Institute has further strengthened methodological oversight and alignment with national AI best practice.
Within England, findings have been shared through Genomic Medicine Service Alliance meetings, GAIN national workshops and regional cardiology and genomics networks. Partner Trusts have hosted structured demonstrations of algorithms, dashboards and workflow integration, incorporating review of validation metrics and case-based discussion.
This combination of NHS-native infrastructure, multi-site deployment, international academic engagement and structured clinical governance review has reinforced professional confidence in the DHT's scientific robustness, clinical relevance and suitability for prospective NHS implementation.

Standard 9: Provide safeguarding assurances for DHTs where users are considered to be in vulnerable groups, or where peer-to-peer interaction is enabled

WP3C was explicitly designed to support some of the most clinically and socially vulnerable patient groups. Individuals with inherited cardiac conditions are often young and otherwise well, yet at risk of life-threatening events such as sudden cardiac arrest or aortic dissection. Many patients identified through this programme had already experienced significant health events, including acute aortic syndromes, and may have ongoing psychological, social and familial implications related to genetic diagnosis. In addition, early audit data highlighted that patients from minority ethnic communities were disproportionately underrepresented in existing ICC pathways, indicating structural barriers to access.
The DHT did not provide automated communication to patients and did not enable peer-to-peer interaction. All outputs were reviewed by cardiologists and clinical geneticists before any patient contact occurred. Referral, counselling and genetic testing followed established NHS pathways, including appropriate consent processes and access to specialist nursing and psychological support where required.

Safeguarding considerations were aligned with Trust policies, genomic consent standards and multidisciplinary governance oversight. By embedding the technology within clinician-led care pathways, WP3C ensured that vulnerable patients were identified earlier, supported appropriately and managed within existing safeguarding and duty-of-care frameworks rather than outside them.

Standard 10: Describe the intended purpose and target population

WP3C developed a language AI–enabled clinical informatics capability designed to improve the efficiency, consistency, safety and equity of ICC identification, genomic testing and longitudinal pathway management within NHS secondary care.
The intended purpose was not to replace clinical judgement, nor to automate diagnosis or referral. Rather, the DHT was designed to augment specialist and non-specialist teams by systematically extracting clinically relevant signals from routinely collected NHS data and presenting them in a structured, pathway-aligned format for clinician review.
The system operates as a decision-support layer embedded within existing Trust digital infrastructure, enabling population-level identification while preserving human oversight at every stage.

The Problem WP3C Addressed
Care for inherited cardiac conditions within the NHS is constrained by structural and operational challenges that limit timely identification and equitable access to specialist and genomic services. Clinical information relevant to ICC phenotypes is dispersed across multiple digital systems, including electronic health records, imaging archives, ECG repositories, laboratory systems, and clinic correspondence. A substantial proportion of this information exists as unstructured free text, which makes systematic interrogation at scale extremely difficult. At the same time, demand for specialist cardiology and genomic services continues to rise, while workforce capacity remains finite and waiting lists grow.
In routine practice, recognition of ICC depends on clinicians synthesising fragmented data across time and systems. Consultations are understandably focused on resolving the immediate clinical question rather than re-evaluating the entirety of a patient's longitudinal record. Subtle but important phenotypic signals may be present across serial ECGs, echocardiography, CT or MRI imaging, lipid profiles, and narrative clinic documentation, yet remain unconnected. The consequence is that patients may accumulate diagnostic clues over years without being channeled into an appropriate specialist pathway.

The ambition to mainstream genomics is similarly constrained by workflow realities. The majority of genetic testing and cascade screening activity remains concentrated within specialist ICC and clinical genetics services. Non-specialist clinicians frequently have limited familiarity with National Genomic Test Directory criteria and insufficient time to compile structured phenotypic summaries that meet laboratory requirements. As a result, eligible patients are not consistently referred, referral documentation is variable in quality and completeness, and laboratories often request further clarification. This iterative exchange prolongs diagnostic timelines and consumes scarce clinical capacity. The overall pathway is labour-intensive, heterogeneous in practice and difficult to scale in the context of increasing demand.

Target Population and Users

The digital health technology was designed to address these challenges across three principal groups. The first comprises patients with unrecognised ICC phenotypes embedded within routine records. These include individuals whose longitudinal data contain constellations of findings suggestive of hypertrophic cardiomyopathy without a secondary cause, heart failure with reduced ejection fraction in the absence of an alternative explanation, thoracic aortic aneurysm or dissection without an acquired cause, lipid profiles consistent with familial hypercholesterolaemia, or dilated cardiomyopathy consistent with inherited disease. Many of these patients are not under specialist ICC follow-up despite having accumulated relevant signals over time.

The second group includes patients whose documented clinical features meet National Genomic Test Directory eligibility criteria but who have not been referred for genetic testing. By operationalising directory-aligned rule application at population scale, the DHT sought to reduce unwarranted variation in referral practice and improve the completeness and consistency of genomic access.

The third group comprises families who may benefit from cascade screening once a proband is identified. Earlier and more systematic identification of affected individuals directly enables preventative care and risk stratification for relatives.

The primary users of the system are cardiologists and clinical geneticists, supported by specialist nurses, and multidisciplinary ICC teams. The platform also serves NHS clinical data scientists, health informatics specialists and Trust IT leads responsible for safe deployment, maintenance, and governance within existing digital infrastructure.

Intended Functional Capabilities

The DHT was developed to enhance efficiency and reliability across the entire ICC pathway, from identification to referral, multidisciplinary review, testing and longitudinal follow-up. It enables systematic population-level analysis of both structured and unstructured data to identify phenotypic patterns consistent with inherited cardiac disease. By continuously interrogating routine clinical documentation and investigation reports, the system reduces reliance on opportunistic detection and appointment timing.

It incorporates automated matching of phenotypic features to National Genomic Test Directory criteria, thereby supporting consistent and transparent application of genomic eligibility rules. Large Language Model components aggregate dispersed information across time and modalities, generating structured and clinically coherent summaries of a patient's phenotype. These summaries are aligned to recognised ontologies, including Human Phenotype Ontology terminology, improving the quality and clarity of information available at the point of referral or multidisciplinary discussion.

The system also provides operational dashboards that give clinicians visibility of investigation completeness, multidisciplinary team status and patients requiring urgent review. In addition, AI-assisted risk stratification continuously reviews newly accrued data and flags concerning changes, enabling proportionate follow-up and prioritisation.

Health System Benefits

At system level, the intended benefits extend beyond improved case identification. By enabling earlier and more equitable detection of at-risk individuals, the DHT supports preventative intervention and reduces the likelihood of catastrophic first presentations. Improved quality and structure of genomic referrals reduce laboratory clarification cycles and shorten time from referral to result. Systematic phenotypic synthesis decreases duplication of investigations and reduces administrative burden associated with manual data retrieval.

Operational dashboards and dynamic risk stratification enable better prioritisation of clinic capacity, allowing high-risk or newly identified patients to be reviewed promptly while reducing unnecessary intensity of follow-up for low-risk individuals. This contributes to improved clinic throughput and more efficient use of specialist expertise. The technology also supports integrated “one-stop” ICC and genetics

models, reducing sequential appointments and aligning cardiology and genomic input within a single episode of care.

By improving pathway reliability and reducing manual data collation, the DHT was designed to increase the number of patients who can be appropriately reviewed within existing workforce constraints, while contributing to more sustainable service delivery.

Boundaries and Safeguards

The DHT was explicitly designed as a clinician-augmented decision-support system. It does not make autonomous diagnostic or referral decisions and does not directly trigger testing or patient communication. All outputs are reviewed by appropriately trained clinicians within established governance structures. Deployment commenced in shadow mode prior to live embedding to ensure safety and workflow alignment. The system operates within existing NHS information governance and clinical oversight frameworks, maintaining professional accountability and preserving human decision-making at every stage.

Standard 11: Describe the current pathway or system process

In current NHS practice, the pathway for patients with suspected inherited cardiac conditions is typically sequential, appointment-driven and dependent on manual clinical synthesis. Identification most often begins with a non-specialist clinician recognising an abnormality such as left ventricular hypertrophy, reduced ejection fraction, aortic dilatation or markedly elevated lipid levels. In the context of diagnostic uncertainty, further investigations are commonly requested, sometimes duplicating historical tests that, when viewed longitudinally, already support a possible inherited diagnosis.

Clinical encounters are generally focused on resolving the presenting problem rather than systematically re-evaluating the patient’s entire longitudinal dataset. Relevant signals may exist across ECG archives, imaging systems, blood results, and narrative clinic letters, but these data are rarely integrated at scale. As a result, phenotypic constellations consistent with hypertrophic cardiomyopathy, dilated cardiomyopathy, aortopathy or familial hypercholesterolaemia may remain unconnected.

Referral to specialist cardiology or clinical genetics services follows, often after multiple interim consultations. Referrals are triaged, and additional clinical information is frequently requested to determine eligibility for review or genomic testing. Genetic testing is typically considered late in the pathway and remains concentrated within specialist services. When requested, clinicians complete referral forms manually, often under significant time pressure. Phenotypic summaries are narrative, variable in structure and not routinely mapped to Human Phenotype Ontology terminology. Laboratories may request clarification, prolonging turnaround times and increasing administrative burden.

Follow-up intervals in ICC clinics are frequently fixed and conservative, reflecting the difficulty of continuously integrating multimodal data and reassessing risk dynamically. This model consumes

specialist time in data collation and manual review, contributes to growing waiting lists, and is difficult to scale in the context of increasing demand and workforce constraints.

Standard 12: Describe the proposed pathway or system process using the DHT

The proposed pathway introduces a data-enabled, pathway-driven model that systematically interrogates routinely acquired NHS data to support earlier, more consistent and more efficient ICC care, while preserving clinician oversight at all decision points.
Instead of relying on opportunistic recognition during appointments, the DHT continuously analyses structured and unstructured data across electronic health records, imaging repositories, ECG systems and correspondence.

Patients whose longitudinal records contain phenotypic patterns aligned to inherited cardiac disease are flagged for specialist review. In parallel, documented phenotypes are matched against National Genomic Test Directory eligibility criteria to identify individuals appropriate for genetic testing.
At the point of triage or clinic review, large language model components generate structured, reviewable phenotype summaries synthesising historical investigations, trajectory of findings and key diagnostic features. These summaries are aligned to recognised ontologies, including HPO terms, improving the quality and completeness of genomic referrals and reducing iterative clarification from laboratories.
Operational dashboards provide clinicians and multidisciplinary teams with real-time visibility of investigation status, genomic eligibility, MDT discussion requirements and patients with concerning new findings. AI-assisted risk stratification supports prioritisation of high-risk individuals and proportionate follow-up for lower-risk patients. This enables more efficient scheduling, increased patients reviewed per clinic session, and reduction in unnecessary sequential appointments.
Integrated “one-stop” ICC and genetics clinics become feasible because relevant phenotypic synthesis and eligibility assessment are performed upstream. Rather than progressing through multiple referrals and appointments, suitable patients can enter a coordinated pathway with cardiology and genomic input aligned within a single episode of care.
Working with health innovation partners, pathway modelling demonstrates that this redesigned process reduces duplication of investigations, shortens time from detection to testing, improves referral quality and enables more efficient use of specialist capacity. The system does not automate diagnosis or referral; it augments clinician review within established governance frameworks. By shifting from reactive, appointment-triggered recognition to proactive, population-level identification, the proposed pathway improves throughput, reduces administrative friction and enhances equity while maintaining safety and professional accountability.

Standard 13: Describe the expected health, cost and resource impacts compared with current care or system processes

Compared with the current sequential and referral-driven pathway, the WP3C model is expected to deliver measurable gains in health outcomes, service efficiency and overall resource utilisation. By systematically identifying patients who meet ICC and National Genomic Test Directory criteria, the pathway shortens time to specialist review and genomic confirmation. Earlier diagnosis enables preventative intervention rather than reactive management following catastrophic events such as sudden cardiac death (SCD), acute aortic syndromes or premature myocardial infarction (MI). Cascade testing amplifies this benefit by extending risk identification and prevention to relatives.
From a service perspective, structured phenotype extraction and automated directory-aligned matching improve referral completeness and reduce laboratory clarification cycles. The integration of cardiology

and genomics into a coordinated “one-stop” model reduces duplication of investigations and replaces multiple sequential outpatient episodes. Even a reduction of one to two outpatient appointments per patient, alongside fewer repeated imaging studies or blood tests, translates into meaningful clinic capacity gains and reduced waiting list pressure within existing workforce constraints.

Detailed modelling example: familial thoracic aortopathy identified after index dissection
In the current model, a patient presenting with acute aortic dissection may receive excellent acute surgical care but not undergo systematic genomic evaluation. If the inherited basis is not recognised, care remains confined to the individual and relatives remain unscreened.
Under the WP3C-enabled pathway, identification of a pathogenic variant in the gene *FBN1* (Marfan syndrome) or another autosomal dominant thoracic aortopathy gene changes the trajectory for the entire family. In autosomal dominant conditions, each first-degree relative has a 50 percent probability of

carrying the familial variant. In a typical family unit comprising two parents, one to two siblings and one to two children, cascade testing commonly identifies two to three additional variant-positive relatives, assuming reasonable uptake.
The health benefit arises because variant-positive relatives can enter structured surveillance before clinical events occur. This includes periodic aortic imaging, tight blood pressure control, low-cost education regarding exercise and lifestyle, reproductive counselling, and planned elective aortic surgery when diameter thresholds are reached. The contrast with emergency dissection is substantial. Acute type A dissection frequently requires emergency surgery, prolonged critical care admission, significant perioperative mortality risk, and long-term morbidity including stroke, renal failure and chronic aortic complications. Resource use is concentrated and high-intensity.
In contrast, planned surveillance and elective intervention occur within scheduled care pathways, with lower perioperative risk, shorter admissions and fewer long-term sequelae. Preventing even one emergency dissection among relatives can offset the cumulative costs of genetic testing, outpatient surveillance imaging and clinic review across multiple family members. When societal costs are included, such as loss of income in working-age adults and informal care burden, the economic benefit is amplified.

Hypertrophic cardiomyopathy (HCM)
Earlier identification of HCM enables structured arrhythmic risk stratification, initiation of beta-blockade where indicated, exercise counselling and appropriate implantable cardioverter-defibrillator consideration in high-risk individuals. Cascade screening frequently identifies asymptomatic relatives with early hypertrophy or genotype-positive status, allowing surveillance and prevention before SCD. Avoidance of a single sudden death in a young adult represents a substantial lifetime health and productivity gain, and reduces emergency admissions and intensive care utilisation associated with cardiac arrest.

Dilated cardiomyopathy (DCM)
Systematic identification of genetically mediated DCM in patients with unexplained heart failure with reduced ejection fraction enables earlier optimisation of guideline-directed medical therapy, appropriate device therapy and targeted family screening. Cascade testing allows relatives to enter surveillance before symptomatic heart failure develops. Preventing progression to advanced heart failure, recurrent hospitalisation or urgent device implantation reduces high-cost inpatient care and long-term disability.

Familial hypercholesterolaemia (FH)
Automated identification of lipid patterns consistent with FH supports early initiation of statins and lifestyle modification at relatively low cost. Cascade testing commonly identifies affected children and young adults. Early lipid-lowering therapy substantially reduces lifetime risk of MI and need for revascularisation

procedures. Prevention of premature coronary events generates significant downstream cost savings and preserves economic productivity over decades.

Across conditions, the dominant economic effect is the prevention of high-cost catastrophic events and advanced disease states. Additional system-level gains arise from reduced duplication of investigations, fewer triage clarification cycles with genomic laboratories, and improved clinic throughput through proportionate follow-up. Collectively, the WP3C pathway is expected to deliver both clinical benefit and sustainable resource utilisation compared with current episodic and referral-dependent models of care.

Standard 14: Provide evidence of the DHT's effectiveness to support its claimed benefits

WP3C builds on more than a decade of development within the CogStack ecosystem, which has demonstrated safe, scalable natural language processing (NLP) deployment across multiple NHS Trusts and specialties. Consistent with established CogStack methodology, WP3C followed a rigorous lifecycle including gold-standard clinician annotation, strict separation of development and validation datasets, and external validation in a second NHS Trust without additional site-specific retraining.

The phenotyping strategy extended well beyond simple keyword detection. For hypertrophic cardiomyopathy, dilated cardiomyopathy, aortopathy and familial hypercholesterolaemia, we constructed comprehensive phenotype libraries incorporating the full range of clinical synonyms, abbreviations and narrative variants observed in real-world NHS documentation. For example, the dilated cardiomyopathy project included terms such as "DCM", "LV impairment", "heart failure" and "HFrEF", while hypertrophic cardiomyopathy modelling incorporated imaging descriptors such as asymmetric septal hypertrophy and non-obstructive hypertrophy. Crucially, models also incorporated relevant negative and exclusionary features. For DCM, these included significant ischaemic heart disease, prior myocardial infarction, severe valvular disease such as significant aortic regurgitation and chemotherapy-associated cardiomyopathy. This contextual modelling materially improved precision without compromising sensitivity.

Across deployment-ready models, recall consistently met or exceeded 0.95, precision ranged between 0.92 and 0.96 depending on phenotype, specificity exceeded 0.92, and F1 scores were ≥0.94. A minimum deployment threshold of F1 ≥0.94 was pre-specified, with refinement required for any model not meeting this criterion. When externally validated in a second Trust without retraining, performance degradation was minimal (F1 change <0.03), supporting transportability across heterogeneous documentation practices.

Large language model-augmented summarisation components were evaluated against clinician-generated reference standards. Blinded review demonstrated capture of over 95% of clinically relevant phenotype elements, with high concordance in mapping to Human Phenotype Ontology terminology. All outputs remained subject to cardiologist and clinical geneticist review prior to any pathway action.

These performance characteristics are consistent with, and in several cases exceed, published benchmarks for clinical NLP in cardiology phenotyping. Importantly, WP3C operates as a clinician-supervised decision-support system rather than an autonomous diagnostic tool. In this context, recall ≥0.95 ensures the vast majority of eligible patients are surfaced for review, while precision in the low-to-mid 0.90s maintains manageable workload. An F1 ≥0.94 represents strong balanced performance for multimodal, free-text NHS data and is appropriate for rare disease case-finding in a human-in-the-loop deployment model.

Prospective shadow-mode deployment has confirmed concordance between algorithm outputs and multidisciplinary team assessment, with ongoing monitoring for drift, bias and workflow impact. Collectively, these findings demonstrate that WP3C meets robust effectiveness criteria for safe and scalable ICC identification within NHS practice

Standard 15: Show real-world evidence that the claimed benefits can be realised in practice

To ensure methodological robustness and minimise selection bias, cohorts of 20 patients per condition were randomly sampled from algorithm-generated lists across familial thoracic aortic disease (FTAD), hypertrophic cardiomyopathy (HCM), dilated cardiomyopathy (DCM) and familial hypercholesterolaemia (FH). The first operational pilot focused on FTAD, with staged implementation subsequently extending to HCM and other pathways.
Within the FTAD cohort, randomly selected patients were invited for review and assessed by a consultant clinical geneticist.

All reviewed cases were considered appropriate referrals with high clinical confidence and met established National Genomic Test Directory criteria for testing. Although not all patients have yet completed blood sampling and final molecular diagnostic yield is therefore pending, the pilot demonstrates that algorithmic case identification translated directly into appropriate genomic assessment within routine NHS care, aligned with expected testing benchmarks for familial thoracic aortopathy.
In the HCM pathway, 20 randomly selected patients were reviewed in a specialist cardiology clinic. Clinical adjudication confirmed that the cohort was comparable in phenotype and complexity to patients referred via standard pathways. A proportion were determined to have left ventricular hypertrophy attributable to hypertension rather than primary cardiomyopathy. XX patients were referred for additional investigations, predominantly cardiac MRI and ambulatory blood pressure monitoring, where HCM remained a credible differential diagnosis. XX patients were confirmed to have HCM at first specialist assessment. These findings demonstrate that AI-enabled identification resulted in clinically appropriate triage and proportionate, guideline-aligned investigation rather than indiscriminate over-referral.
Results from the DCM and FH pilots are awaited as patients complete staged clinical assessment and testing processes.
Taken together, these early implementation data provide credible real-world evidence that the WP3C pathway can be embedded safely within NHS specialist services, generating clinically appropriate referrals, supporting genomic testing decisions, and aligning with existing governance, triage and consultant-led care standards while maintaining proportionality and clinical judgement at every step.
Implementation is a major barrier for most DHT's and WP3C is able to leverage on existing deployed general-purpose technologies built by NIHR/NHS funding in CogStack platforms. This means that a large element of the maintenance, security, sustainability and scaling is addressed by the underlying the underlying platform rather than the DHT product specifically. These other implementation and maintenance elements are handled by CogStack Ltd (NHS spinout) which is steadily scaling across the NHS with 4 NHS trusts in 2025 extending to 7 in 2026.

Standard 16: Ongoing Monitoring of Usage and Performance Over Time

CogStack will operate a structured, longitudinal monitoring framework to ensure sustained performance, safety and clinical relevance following deployment. System usage will be tracked through automated audit logs and service telemetry, including volume of documents processed, patients flagged, clinician interaction with generated case lists, time-to-review, and downstream clinical actions such as specialist referral or genetic test requests. These indicators allow continuous assessment of adoption, workflow integration, and operational impact.
Model performance will be monitored through scheduled silent re-sampling and blinded clinician adjudication of randomly selected cases to reassess precision, recall and F1 score over time, alongside monitoring of false-positive workload per clinic session. Drift detection processes will include surveillance for changes in documentation patterns, evolving clinical terminology, coding updates and variation in data completeness. All models will remain under formal version control within the CogStack ecosystem, with

predefined performance thresholds, governance oversight and structured change-management processes should performance fall outside agreed parameters.

Standard 17: Budget Impact Analysis

In partnership with Health Innovation Wessex, a structured budget impact analysis has been developed comparing current referral-led ICC pathways with the WP3C data-enabled model.

The analysis considers implementation costs, including infrastructure configuration within existing CogStack environments, clinical validation time and governance oversight, against anticipated cost offsets. These include reduced outpatient appointments, fewer laboratory clarification cycles, improved clinic throughput, and avoidance of high-cost emergency presentations such as acute aortic dissection or sudden cardiac arrest.

Standard 19: Ensure Transparency About Requirements for Deployment

WP3C is implemented within the established CogStack architecture, which has been deployed across multiple NHS Trusts using standard enterprise infrastructure. The core stack comprises Elasticsearch for indexed search, ingestion pipelines (commonly Apache NiFi) for structured and unstructured data feeds, and NLP components such as MedCAT and rule-based or transformer models running within secure Trust environments.
Deployment typically requires virtualised Linux servers with adequate CPU capacity and memory for indexing and search workloads. Most CogStack NLP pipelines operate efficiently on CPU infrastructure. GPU resources are optional and only required where large transformer or generative models are hosted locally. Storage requirements scale with document volume and align with routine EPR archival datasets. All components operate within NHS-controlled networks with role-based access control, audit logging, encryption in transit and at rest, and compliance with DSPT and clinical safety standards. No proprietary hardware or external data transfer is required for baseline deployment.

Standard 20: Communication, Consent and Training

WP3C is implemented as a clinician-facing decision-support capability embedded within existing NHS systems and does not communicate directly with patients. As a result, no additional patient-facing consent process is introduced beyond established Trust information governance for secondary use of data, and the standard NHS genomic consent process applied at the point of testing in clinical genetics. Communication and training are proportionate to the tool's function and are delivered through short, role-specific onboarding for cardiology, clinical genetics, and informatics teams. The core interaction model is intentionally simple: clinicians review algorithm-generated case lists and structured phenotype summaries, with the option to review the linked source documents in the longitudinal record (e.g., clinic letters, imaging reports, ECG text) to verify context.
Automated "genetics referral support" is implemented as decision support rather than automation: the system pre-populates referral-ready phenotypic summaries aligned to National Genomic Test Directory criteria and HPO concepts, improving completeness and consistency while preserving clinician control over whether, and how, a referral is made. The risk stratification component similarly functions as an assistive prioritisation layer, flagging concerning changes in incoming data for earlier review while explicitly requiring consultant-led interpretation and action. Across these tools, training emphasis is on interpreting outputs, verifying provenance via source links, and using standard escalation routes where

outputs require clinical action, ensuring transparency, safety and rapid adoption with minimal additional workload.

Standard 21: Ensure Appropriate Scalability

WP3C is built on CogStack, an open-source, NHS-developed and NHS-owned architecture specifically designed for secure, system-wide scalability. CogStack has already been deployed across multiple large NHS organisations, including Guy's and St Thomas' NHS Foundation Trust, King's College Hospital NHS Foundation Trust, University College London Hospitals NHS Foundation Trust, South London and Maudsley NHS Foundation Trust, with further deployments underway at Barts Health NHS Trust, Lewisham and Greenwich NHS Trust, and University Hospitals Birmingham NHS Foundation Trust, among others.

Engagement is also underway with national bodies such as the National Disease Registration Service (NDRS) and NHS DigiTrials to explore structured linkage, registry-aligned case identification and research enablement at scale. This supports regional and national-level scalability while maintaining appropriate governance, data protection safeguards and alignment with existing NHS data infrastructure.

The architecture relies on commodity enterprise components (Elasticsearch indexing, containerised NLP services, secure on-premise or NHS cloud hosting), allowing horizontal scaling as data volume or use cases expand. No proprietary hardware is required, and expansion to additional Trusts primarily involves configuration, data mapping and governance approval rather than technical redesign. This supports sustainable NHS-wide scale-up while retaining public ownership, interoperability and transparency.

Work Package 3D Evaluation

Executive summary

GenePy is a panel agnostic tool for pathogenic signal compression and reducing the dimensionality of huge sparse data generated from whole genome and whole exome sequencing data in patients.

Germline scores only need to be generated a single time for any individual as they don't change over the course of a lifetime. Therefore, once you create the GenePy score for one person, they remain the same. GenePy is an alternative approach that reduces the dimensionality of sparse medical genomic sequencing data. It is an intuitive quantitative score for each gene that can be compared between individuals. It retains signal from rare variants that are more likely to have clinical consequences and integrates information on functional impact and population allele frequency of individual variants.

Rare disease patients have extreme outlier scores in their known causal disease genes, and GenePy is sensitive to detect rare disease genes in undiagnosed patients. GenePy scores capture the functional impact on quantitative traits across populations. It is powerful when integrated with rich clinical data – and can extract monogenic disease from common disease patient cohorts.

The main aim of the GenePy project in the Genomic AI Network of Excellence was to demonstrate the utility of GenePy as a possible additional tool that could aid clinical scientists in the reporting of genomic medicine service data from patient samples. Specifically, we believe it can be useful in speeding up the manual curation time by more quickly prioritising genes in which clinical scientists should look to identify the disease-causing variants in patients, uplifting diagnosis.

The two key principles that it was being used for was to speed up the manual curation time by prioritising genes in which clinical scientists should look to identify the disease causing variant in patients and two, because it's a digital tool and it could be implemented quickly and at scale, there's an opportunity that it can be used to identify missed diagnoses.

We have implemented GenePy across the aggregate 2 Genomics England data set of approximately 78,000 whole genomes and have gained priority access to the aggregate 3 Genomics England data for about 139,000 whole genomes. We are currently in the process of generating the GenePy matrix on the aggregate 3 data. We have worked with the cloud system inside the Genomics England research environment. That cloud system is on a LifeBit platform, and we have created a modular workflow to generate GeneP scores across all of the aggregate 2 data. We have validated our results on the, we have validated our results by cross-checking the scores in patients with known diagnoses and positive diagnostic reports. We are currently in the process of working with the diagnostic discovery team inside Genomics England and refining the process to return new diagnoses for patients who have a negative diagnostic report according to the exit questionnaire and the diagnostic database in Genomics England.

Standard 1: the digital health technology (DHT) should comply with relevant safety and quality standards

In accordance with the format of the diagnostic discovery group, reports of the pathogenic variants within those genes are sent according to the normal variant standards of describing variants. Data is in typical genomic standard format for VCF files. If this was to be implemented centrally in the NHS, a clinical scientist would be pointed towards a plausible disease gene in which the patient has a high GenePy score consistent with their phenotype, and they would then interrogate the underlying variations as they normally do. Similarly, if it were to be deployed in a lab, the lab would need to be in line with relevant ISO standards. In this way, no new standards adherence is required for GenePy.

Standard 2: incorporate intended user group acceptability in the design of the DHT

Within the GAIN project, the intended user groups group are the clinical scientists within the Genomic Medicine Service reporting on patient whole genome sequencing. Specific user-group feedback was not

possible through the Diagnostic Discovery Group, as there is purposefully no direct link from the researcher to the clinical scientists and clinicians actioning results.

As part of the 2024 published paper we spoke to the research community and those interested in Genomics England patients, as well as clinical scientists and senior clinical scientists in the Genomic Medicine Service. The feedback was they were able to interpret it as an additional tool that might work very similar to Exomiser in the way that it might help prioritise the genes, especially if a clinical scientist is reviewing a large panel of genes for a specific diagnostic referral, such as intellectual disability, where there are super panels of genes. Here, it can be helpful to expedite the discovery of the causal variation.

There are other applications of GenePy for the medical genomic research community. As an additional, the GenePy matrix for aggregate 2 and aggregate 3 will be available to the research community. Because the score only needs to be computed once for any one individual's whole genome sequencing VCF file. Once computed, it can be used to interrogate any phenotype, be it rare disease or predisposition to a later onset common disease. However, this is beyond the current scope of the project.

Standard 3: Consider Environmental Sustainability

Whilst considering the sustainability was not part of the original project, throughout the project, we have been very much aware of the computational requirements of generating the GenePy score. We have optimised the pipeline to work most efficiently. For example, when deployed on the aggregate 2 data, it was implemented on the coding variation set only because processing variance from introns would have add substantial additional compute burden. And we made the decision that doing that for the aggregate 3 data set when it came was more sustainable.

We also plan to make the final GenePy matrix available to all researchers so that it never needs to be computed again and everybody can avail of it without further additional compute requirements.

Because one of the intrinsic requirements to create the GenePy score is annotation with a in silico predictor of pathogenicity called CAD, we are planning and hoping to assist the Genomics England team with annotation of all Variants with CAD 1.6. We hope to work alongside the Aggregate 3 team to generate this for the Aggregate 3 data set and again make that available for all researchers in the GEL RE community, so they don't need to compute it separately. Central annotation would improve efficiency and avoid the need for compute when users are using the data set as a reference file.

Standard 4: consider health and care inequalities and bias mitigation

GenePy will be subject to the same bias as other allele frequency estimators because allele frequency is a component of the algorithm. The tool would be improved, as with any genomics tool, by a more representative coverage of all human populations.

Standard 5: embed good data practices in the design of the DHT

The data source to date has been Genomics England Aggregate 2 and soon Aggregate 3. As such, we are compliant with all their data protections. Local deployment is compliant to REC approvals and local institutional review board. Data is held in secure university infrastructure, deidentified and patient-consented, where the legal basis for use of data is public interest.

We are also looking at the second biggest data set, UK biobank Research Access Portal, but we're at the stage of identifying the optimal pipeline to work within that specific Research environment infrasturture, software and dependencies. The pipeline would be adapted for the UK Biobank RAP, the most efficient way to annotate the variants in the way they need to be annotated and generate the whole gene genopy scores would be identified. Similar to Genomics England, this would be in compliant with their regulation and practices as well.

In line with good practice, all code will be made publicly available in order to enable consistent application and results across geographies and settings.

Standard 6: define the level of professional oversight

The code for how the GeneP scores will be made available to the research community scrutiny by other bioinformaticians and scientists on GitHub.

The Genomics England GenePy Matrix can, by necessity, only be accessed by individuals who've been approved for access to the Genomics England research environment, but they too will be able to scrutinise it. Any and all publications are subject to scientific peer review.

Within an NHS setting if GenePy is deployed there, it is simply a prioritisation tool. There is always the human in the loop who is the clinical scientist making the diagnostic report. They will always have appropriate knowledge of the genes and the underlying variants that are being implicated by GenePy scores. Should a GenePy indicate a gene and a diagnostic variant identified, they will always be subject to ACMG and ACGS guidelines in use by the Genomic Medicine Service, and ACGS best practice guidelines for variant classification in rare diseases (Durkie et al, 2024).

Standard 7: show processes for creating reliable health information

Please refer to the published paper for how the GenePy score information is created. GenePy aggregates all variants within a gene to estimate the gene's overall disease risk, incorporating allele frequency, zygosity and predicted biological damage (such as a CADD score). A further paper resulting from this project will describe the pipeline to create GenePy scores in the Genomics England data. Submission date is pending decision on aggregate 2 and aggregate 3.

Standard 8: show that the DHT is credible with UK professionals

This specific standard is outside of the scope of this project, as bespoke governance arrangements would be required to the link patients and clinicians between the Genomic Medicine Service and Genomics England. A future focus group would include senior clinical scientists. Anecdotally, clinical scientists who have read the paper can see the value of GenePy and find it simple and intuitive.

Standard 9: provide safeguarding assurances for DHTs where users are considered to be in vulnerable groups, or where peer-to-peer interaction is enabled

This standard is not relevant to GenePy because it is a generic tool applied to variant call files derived from the sequencing of data. It exists between the sequencing of DNA and the reporting of clinical scientists, and is divorced from patient contact.

Standard 10: describe the intended purpose and target population

As whole genome sequencing is rolled out to more people, there are increasing published and well-established reasons why the use of panels is non-optimal. This method (GenePy) will work in a similar manner to exomiser in a panel agnostic method. This will perhaps flag low hanging diagnostic fruit more quickly to ease the manual curation burden in clinical diagnostic reporting. It may also possibly identify disease genes that were missed either because they were not included in a panel, or they were overlooked for other reasons. This is to help patients who otherwise might not get a diagnosis, and help clinical scientists with speed and increase diagnostic yield.

Additionally, because it is a digitised method, it's also simple to redeploy in the case of re-reviewing in the future.

Standards 11 and 12: describe the current pathway or system process. Describe the proposed pathway or system process using the DHT

The GenePy logic model provides a structure for describing both the proposed pathway when GenePy is introduced, where genomic data are processed through the GenePy algorithm to generate gene level pathogenicity scores. These scores help clinical scientists prioritise genes more efficiently and reduce manual review time. The model describes how clinicians then use the GenePy supported reports to guide further testing, referrals, or clinical management. By mapping inputs, activities, outputs, outcomes, and impacts, the logic model would enable a comparison between the conventional workflow and the streamlined, AI supported process enabled by GenePy.

Standard 13: describe the expected health, cost and resource impacts compared with current care or system processes

The logic model highlights where GenePy is expected to make a measurable difference - such as increasing the diagnostic uplift of rare and complex conditions, enabling faster workflows for clinical scientists interpreting and writing reports, supporting earlier disease recognition, and new or earlier treatment initiation. In addition, the logic model highlights how GenePy may improve equity of care, identifying conditions that may be missed in previous processes and improving recognition of gene variation across diverse ethnicities. This may mean increased research in this area and potentially more patients - particularly those from under-served groups - may receive accurate diagnoses and timely support.

Standard 14: provide evidence of the DHT's effectiveness to support its claimed benefits

A pilot has been completed in the 2024 paper. In this pilot, GenePy was applied to 78,216 participants in the 100,000 Genomes Project to analyse 2,862 recessive disease genes. Among participants without a prior molecular diagnosis, 122 cases were identified as putative missed diagnoses. These diagnoses were supported using phasing, ClinVar evidence, and ACMG variant classification guidelines, increasing confidence in the findings. Beyond the confirmed missed diagnoses, GenePy flagged a large number of additional candidates; 334 cases had variants consistent with disease but required further functional evidence. In total, the approach identified 456 potential diagnoses across the cohort. With regard to clinical scientist time, in the pilot GenePy required reviewing only about 1.2 additional variants per individual during analysis.

We are actively working with the diagnostic discovery group to implement GenePy across Aggregate 2 and Aggregate 3 to make and return more diagnostic variance in genes that for the clinical presentation of patients. Current application in Aggregate 2 are realising new diagnoses for patients. We are working closely with the GEL diagnostic discovery team to optimise this workflow to have highest possible sensitivity. We are processing the newest release of aggregate 3 data through beta access and anticipate that GenePy will be instrumental in uplifting diagnostic rates for the ~40k GMS patient data within this new release.

Standard 15: show real-world evidence that the claimed benefits can be realised in practice

As evidences by the above paper, we have made new diagnoses in Genomics England. Through the lifetime on the GAIN project, new diagnoses are being made in Aggregate 2, and anticipate significant uptake in Aggregate 3

In a local cohort for which I'm chief investigator in Southampton, we have made an established rare disease diagnosis in a cohort of patients that have been diagnosed with a common disease, inflammatory bowel disease. We have been able to feed those diagnostic variants back to the clinical team and validate our findings with the help of the local genetics laboratory re-genotyping the patients. We have worked with Great Ormond Street to confirm the functional changes to the immune pathways.

This has resulted in patients having their diagnosis changed from, for example, Crohn's disease to CTLA 4 deficiency and further resulted in a change to the treatment for those patients. For example, one patient with CTLA4 deficiency would not have been eligible for a different drug which they have now started without that diagnosis as a result of GenePy.

Standard 16: the company and evaluator should agree a plan for measuring usage and changes in the DHT's performance over time

We are in the process of optimising the GenePy workflow using the new portioning of patient WGS data in GEL. It is anticipated that the final model to generate GenePy scores to be used to assess any given patient's genome, will be containerised within a docker image for standardised deployment. This image would be version controlled and any subsequent updates released with all changes documented. It is possible that future updates could incorporate new in silico predictors, updates to population reference allele frequencies, incorporate structural variants or allow for long read phased data.

Standard 17: provide a budget impact analysis

The GenePy logic model provides a breakdown of all activities, resources, and outcomes involved in delivering the innovation, offering a structured foundation for a budget impact analysis. It identifies the key cost-generating components of the pathway - such as governance and approvals, GenePy software, training requirements, analysis time, and clinical scientist interpretation - allowing these elements to be costed against current workflows. The model also highlights where GenePy can reduce resource use, including fewer manual variant-level reviews, decreased clinical scientist time per report, and shorter turnaround times for genomic reports. These efficiencies can be translated into quantifiable cost savings, helping estimate in future real world evaluation studies the financial difference between GenePy-enabled and traditional processes. In addition, the model outlines patient-related impacts such as increased diagnostic uplift and earlier disease recognition. These can be incorporated into a budget impact analysis as avoided downstream costs, for example, preventing late diagnoses or reducing unnecessary clinical appointments. Overall, the logic model specifies each step where costs or savings occur, enabling a future comprehensive assessment of GenePy's budget implications across staff time, service capacity, diagnostic activity, and long-term patient outcomes.

Standard 18: for DHTs with higher financial risk, describe the cost effectiveness

This standard is not relevant to GenePy. Please see standard 17 instead.

Standard 19: ensure transparency about requirements for deployment

All requirements will be made fully transparent through bioinformatic publications and software publication on open access repositories such as GitHub. With appropriate licensing. Workflows will be packaged in docker images to ensure consistent deployment.

Standard 20: describe strategies for communication, consent and training processes to allow the DHT to be understood

We will continue to communicate and disseminate through publications, conference presentations and other opportunities to inform stakeholders (education events, publica engagement events as part of HDRS activity, Festibval of Genomics, GMS network, BSGM lunchtime series).

Standard 21: ensure appropriate scalability

GenePy is intrinsically scalable as it can and should be deployed at source. Interpretation of GenePy scores require reference scores generated on large datasets such as GEL Aggregate 2 and 3 and UK Biobank 500k WGS. Application thereafter can be applied to single patients or massive cohorts. Once

germline GenePy scores are generated for any individual, they are constant and do not need to be regenerated. Therefore, they can be used for repeated analysis of the same phenotype, or new phenotypes, or be applied in polygenic risk scores. No lab would need to procure GenePy, any individual's score would be made available with the original release of their data. There is no additional Clinical Scientist resource required to generate GenePy scores.

Conclusion

This evaluation has applied the NICE Evidence Standards Framework for Digital Health Technologies to the pilot projects under the Genomic AI Network.

The pilots represent tools at different stages of the development and deployment pathway, ranging from research-stage analytical methods to locally deployed digital health technologies and regulated medical devices. As a result, the direct applicability of individual NICE standards varies depending on the maturity of the tool and its position within the clinical pathway. The framework offers a valuable reference point for developers, clinicians and organisations considering the development or implementation of AI-enabled tools within NHS genomic services, particularly a structured approach for considering key aspects of safety, governance, usability and effectiveness across a diverse portfolio of AI technologies.

Overall, the findings indicate that the WP3 projects have been developed with appropriate consideration of governance, safety and clinical integration.

A few key thematic conclusions emerge:

- **Genomic AI tools across the NHS ecosystem vary significantly in levels of technical maturity and regulatory readiness.** The projects evaluated ranged from early-stage analytical infrastructure tools (Genollama) to fully regulated, MHRA-registered Class I clinical decision-support systems (MendelScan).
    - Recommendation: Genomic AI evaluations and commissioning should leverage relevant frameworks to ensure that the demands for evidence, validation, and governance are proportionate to the tool's intended purpose and potential clinical risk. Furthermore, successful deployment requires addressing these governance and clinical oversight requirements as core design principles, rather than as retrospective hurdles.
- **The translational gap between Secure Data Environments and direct clinical action is a critical friction point for deploying genomic AI at scale.** Even when it is technically feasible to integrate AI within NHS Secure Data Environments (SDEs) to generate insights on de-identified data, translating those insights into direct patient intervention poses significant governance and clinical alignment challenges.
    - Recommendation**:** Scaling of genomic AI tools cannot rely on technical readiness alone. Clinical system leaders with strategic responsibility for population health outcomes should explicitly define the direct-care governance pathways and patient communication strategies preceding deployment, using the SDEs as a resource for pre-approved clinical workflows.
- **Plan for the NHS-wide shift from reactive to proactive care as facilitated by genomic AI tools.** Several projects demonstrated that one of AI's greatest value lies in shifting the NHS from a reactive, appointment-driven model to a proactive, data-driven case-finding model. These tools interrogate records (both genomic data and clinical patient records) to surface undiagnosed patients which would be unfeasible for individual systematic manual review, significantly truncating the diagnostic odyssey.
    - Recommendation: When planning and budgeting for proactive case-finding technologies, commissioners must account for a shift in resource utilisation. While AI tools may increase up-front clinical activities (e.g. time for a clinician or scientist to review an AI-generated case flag), it could reduce long-term structural inefficiencies such as avoiding catastrophic emergency events, unnecessary outpatient visits, and duplicate testing. Budget impact analyses and capacity planning must reflect this redistribution of workload across the whole clinical pathway to secure clinical buy-in and realise long-term savings.

These insights contribute to a broader understanding of how AI technologies can be responsibly developed and implemented within genomic medicine.

We thank the clinical, technical and research leads across all WP3 pilot projects for their collaboration, expertise and commitment throughout the programme. Their contributions have been essential in advancing the understanding of how AI can support genomic medicine within the NHS.

Authorship

This evaluation was developed collaboratively across the Genomic AI Network (GAIN), with contributions from clinical, academic, industry and programme leads:

- **Dr Martin Chapman**, The Artificial Intelligence Centre for Value Based Healthcare (WP3A)
- **Dr Peter Fish**, Mendelian (WP3B)
- **Dr Antonio de Marvao**, King's College London (WP3C)
- **Professor Sarah Ennis**, University of Southampton (WP3D)
- **Dr Andrew Sibley**, Health Innovation Wessex

Genomic AI Network Programme Team:

- **Dr Alexander T Deng**, Genomic AI Network Director
- **Poppy Cohen**, Genomic AI Network Project Manager
- **Ana-Maria Sorocean**, Genomic AI Network Coordinator
- **Miren Sowden**, Genomic AI Project Manager

Contribution of Health Innovation Wessex

We would like to formally acknowledge the contribution of Health Innovation Wessex (HIW) in supporting structured assessment work across WP3. Health Innovation Wessex provided detailed technical and governance support, including structured documentation reviews and alignment mapping against relevant standards. These can be found in the appendix.