Deliverable 5.1

Blueprint for Pathfinder: frameworks and pathways

Grant Agreement Number: 101093126

Funded by the European Union

UK Research and Innovation

The British associated partners of NextGen were funded by UK Research and Innovation (UKRI) under the UK government's Horizon Europe funding guarantee [grant agreements No. 10098097, No. 10104323]

Project funded by

Schweizerische Eidgenossenschaft
Confédération suisse
Confederazione Svizzera
Confederaziun svizra

Swiss Confederation

Federal Department of Economic Affairs, Education and Research EAER
State Secretariat for Education, Research and Innovation SERI

NextGen	
Project full title	Next Generation Tools for Genome-Centric Multimodal Data Integration In Personalised Cardiovascular Medicine
Call identifier	HORIZON-HLTH-2023-TOOL-05-04
Type of action	RIA
Start date	01/ 01/ 2024
End date	31/12/2027
Grant agreement no	101136962

Funding of associated partners
The Swiss associated partners of the NextGen project were funded by the Swiss State Secretariat for Education, Research and Innovation (SERI). The British associated partners of NextGen were funded by UK Research and Innovation (UKRI) under the UK government's Horizon Europe funding guarantee [grant agreements No 10098097, No 10104323]

D5.1 – Pathfinder frameworks and pathways Report	
Author(s)	Sander W. van der Laan, Aaron M. Lee, Francesca Mangili, Philippe Page, Fred Buining, Raja Appuswamy
Editor	Sander W. van der Laan
Participating partners	UMCU, QMUL, SUPSI, HIRO, HCF, EURECOM
Version	1
Status	Final
Deliverable date	M6
Dissemination Level	PU - Public
Official date	2024-06-30
Actual date	2024-06-28

Disclaimer

The NEXTGEN consortium consists of the following partners:

No	PARTNER ORGANISATION NAME	ABBREVIATION	COUNTRY
1	UNIVERSITAIR MEDISCH CENTRUM UTRECHT	UMCU	NL
2	HIRO MICRODATACENTERS B.V.	HIRO	NL
3	EURECOM GIE	EURE	FR
4	JOHANN WOLFGANG GOETHE-UNIVERSITAET FRANKFURT AM MAIN	GUF	DE
5	KAROLINSKA INSTITUTET	KI	SE
6	HUS-YHTYMA	HUS	FI
7	UNIVERSITY OF VIRGINIA	UVA	US
8	KLINIKUM RECHTS DER ISAR DER TECHNISCHEN UNIVERSITAT MUNCHEN	TUM-Med	DE
9	HL7 INTERNATIONAL FOUNDATION	HL7	BE
10	MYDATA GLOBAL RY	MYDTA	FI
11	DATAPOWER SRL	DPOW	IT
12	SOCIETE EUROPEENNE DE CARDIOLOGIE	ESC	FR
13	WELLSPAN HEALTH	WSPAN	US
14	LIKE HEALTHCARE RESEARCH GMBH	LIKE	DE
15	NEBS SRL	NEBS	BE
16	THE HUMAN COLOSSUS FOUNDATION	HCF	CH
17	SCUOLA UNIVERSITARIA PROFESSIONALE DELLA SVIZZERA ITALIANA	SUPSI	CH
18	DRUG INFORMATION ASSOCIATION	DIA	CH
19	DPO ASSOCIATES SARL	DPOA	CH
20	QUEEN MARY UNIVERSITY OF LONDON	QMUL	UK
21	EARLHAM INSTITUTE	ERLH	UK

Document Revision History

DATE	VERSION	DESCRIPTION	CONTRIBUTIONS
27/06/2024	1.0	Initial Blueprint of Pathfinder	UMCU
28/06/2024	2.0	Blueprint after corrections by Aaron M. Lee	UMCU, QMUL
28/06/2024	2.1	Blueprint after lay-out corrections Luca A. Remotti	Datapwer, UMCU

Authors

AUTHOR/EDITOR	ORGANISATION
Sander W. van der Laan	UMCU
Aaron M. Lee	QMUL
Francesca Mangili	SUPSI
Philippe Page	HCF
Fred Buining	HIRO
Raja Appuswamy	EURECOM

Reviewers

REVIEWER	ORGANISATION
Aaron M. Lee	QMUL
Philippe Page	HCF

List of terms and abbreviations

ABBREVIATION	DESCRIPTION
AI	Artificial intelligence
B1MG	Beyond 1 Million Genomes
DOA	Data Oriented Architecture
EDHS	European Health DataSpace
FAIR	Findability, Accessibility, Interoperability, Reuse
FL	Federated (machine) learning
GDI	(European) Genomic Data Infrastructure
GDPR	General Data Protection Regulation
GWAS	Genome-wide association study
HTA	Health Technology Assessment
KPI	Key Performance Indicator
ML	Machine learning
MMIO	Multi-model integration object
PGS	Polygenic (risk) score
TDA	Trusted Data Agent
VCF	Variant call format
WP	Work Package

1 Table of contents

2 Executive Summary

The goal of the **NextGen Pathfinder** is to create a small-scale-European Health Data Space (EHDS), namely a health specific ecosystem comprised of functionalities, rules, common standards and practices, infrastructures, and a governance framework which will enhance citizen trust in health technologies. The **Pathfinder** project and its component pilots serve as practical demonstrations of the efficacy of **NextGen** tools and project outputs. In this small-scale-EHDS data are shared in a federated manner: individual researchers can work with the different datasets, but they can never 'see' or 'touch' them nor do they ever leave the premises of the respective institutes. Put simply, **Pathfinder** will help researchers to connect efficiently and securely with different datasets across the globe and execute specific analyses in a federated manner.

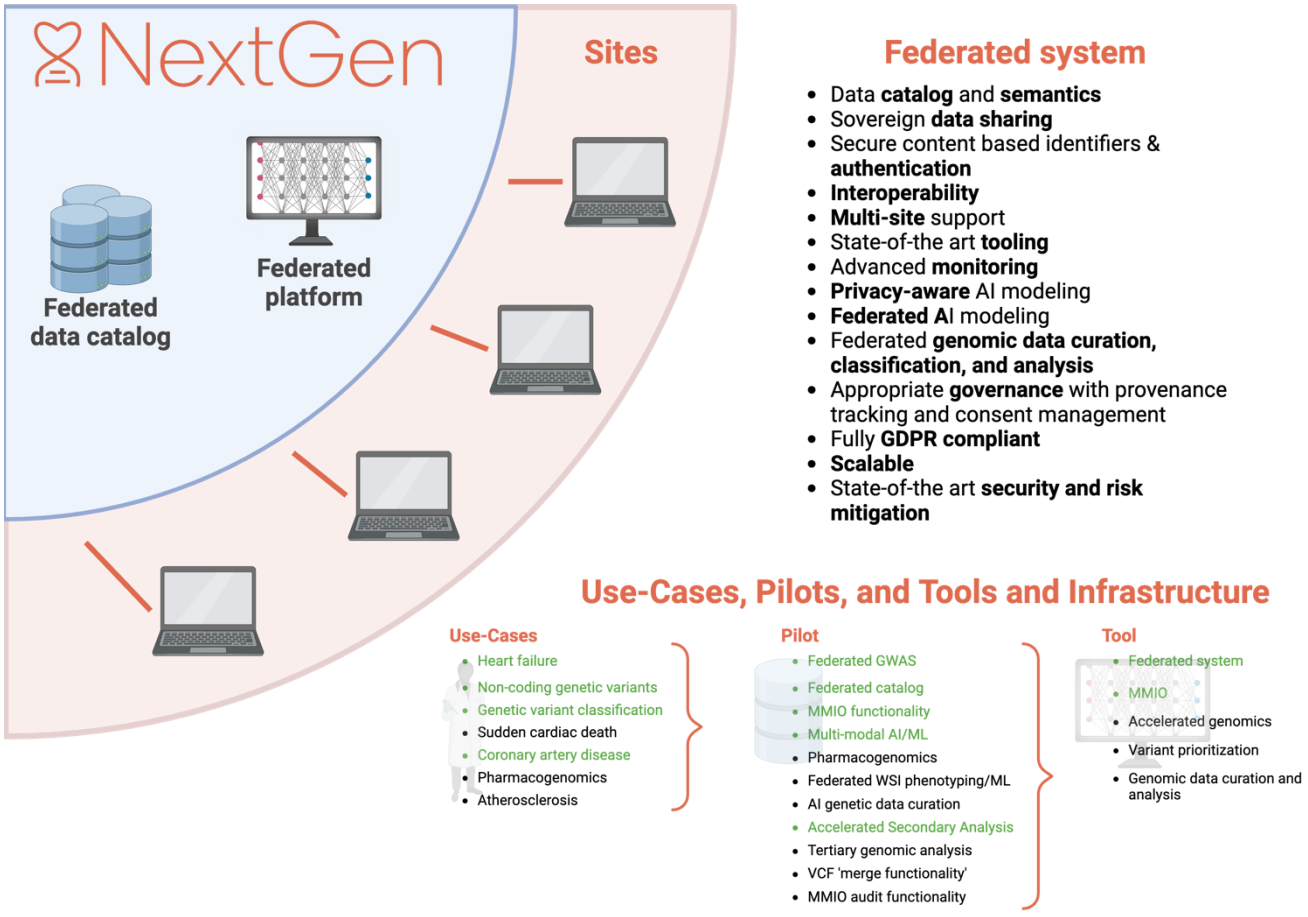

Figure 1: Graphical Abstract. *The **NextGen Pathfinder** is a platform, a small-scale-European Health Data Space, to execute health-specific analysis. An inventory of **Use-Cases** is mapped to specific **Pilots** and **Tools and Infrastructure** with a given set of **functionalities**. These enable a given user at a specific site (in Europe or the United States of America) to execute health-specific research. Although **NextGen Pathfinder** is focused on cardiovascular disease, it is expected to be a generic platform amenable to other health outcomes. In green are **Use-Cases**, **Pilots**, and **Tools and Infrastructure** that have crystalized (see document). Created with BioRender.com.*

3 What is NextGen Pathfinder?

3.1 Introduction

We will integrate best practices through the **NextGen Pathfinder** and **Pilot** projects. These **Pilots** will demonstrate advanced integration and workflow tools in a federated platform (where applicable), showcasing their effectiveness in removing technical and operational barriers. The **Pilot** projects will be incorporated into the **NextGen Pathfinder**, a multi-site "mini-European Health DataSpace (EHDS)" network that highlights NextGen's innovations in data management, data governance, cataloging, computing, advanced data integration, genomics, and interoperability capabilities. The **Pathfinder** will adopt best practices from evolving EU-wide initiatives such as the EHDS and 1+MG, ensuring it remains at the forefront of innovation and excellence.

3.2 Federation

3.2.1 What is federation?

A key discriminating factor for NextGen Pathfinder is that it will host several federated initiatives, a catalog, and both machine learning and genomic analytics. In federation, no individual level data is to be shared or ever seen by any of the scientists using the **Pathfinder** platform. All the required data will remain at participating partner sites. This implies, for example, that such a platform will have to be secure, anyone logging in should have the right credentials, and will have to have some form of cataloging the available data. However, the appropriate governance should also be in place to enable the platform to access site-specific systems, and ensure that individuals' rights and privacy, i.e. that of the research participants, are protected. Of course, the software and hardware driving the platform should be state-of-the-art to handle the flow of big data information – sometimes tens of millions of genetic variants across thousands of individuals will have to be analyzed somehow, without leaving the institution's premises, et cetera.
In other words, no data – other than results from analyses which are intrinsically not mappable to individuals – is centralized, rather it stays put at the partners' sites. This is also the reason why the governance aspect and the risk assessment, as well as the involvement of stakeholders across the board is so critical for the success of demonstrating NextGen **Pathfinder**.

3.2.2 Why federation?

It is simple: we are flooded by increasing large volumes of data used in analytics, and particularly personalized medicine which require multi-omic data. This would mean, that 1) copying data over would be time and resources inefficient, , 2) security risks and privacy requirements are multiplied in centralisation, and 3) version control and updates of datasets becomes impracticable, limiting the responsiveness of algorithm development. We need to build systems that are securely working under the appropriate governance structures that are properly risk assessed, secured, and privacy preserving. In this way, secure data access, rather than data sharing, becomes the dominant paradigm to solve problems and challenges in healthcare, scalable with data volumes, through the development of intuitive, user-friendly, advanced analytical models, including artificial intelligence (AI).

3.3 Use-Cases, Pilots, Tools and Infrastructure

Each pilot is derived from the research **Use-Cases** undertaken by **NextGen's** clinical and academic partners. Each pilot consists of a set of specific **Tools** and **Infrastructure** required for one or multiple **Use-Cases**. In essence, the **Pathfinder** will be constructed from a sequence of pilot demonstrations using various project tools to facilitate the **Use-Cases**. Initially, we will demonstrate how genomic tools can improve workflow efficiency within the **Pathfinder** framework. Following this, we will focus on showcasing multimodal data integration tools, highlighting their capability to seamlessly integrate different types of data within the

Pathfinder. Next, we will demonstrate the application of federated analytics, particularly in genomics and artificial intelligence (AI) and machine learning (ML), to illustrate how these advanced techniques can be leveraged within the **Pathfinder**. Finally, we will deploy the complete multisite **Pathfinder** project, integrating multiple innovations and tools to create a comprehensive and robust system.

Use-Cases are defined as discipline-specific tasks (such as development of clinical machine learning predictive models) revolving around a research question that showcases how certain **NextGen Tools and Infrastructure** will work. A list of **Use-Cases** per thematic area is given below, a more detailed description of one exemplar **Use-Case** is given in the *Appendix*.

Thematic area	Exemplar research
Heart failure	Diagnosis and outcome prediction (hospitalization, arrhythmia, thromboembolism and sudden death).
Non-coding genetic variants	Congenital cardiac disorders
Genetic variant classification	Identification of high-risk causes of cardiomyopathy.
Sudden cardiac death	Assessment and prevention algorithms.
Coronary heart disease	Risk scores computation and integration into predictive models.
Pharmacogenomics	Adverse drug reaction prediction.
Atherosclerosis	Severity and outcome prediction models through whole-slide images.

Pilots consist of a set of Tools and Infrastructure that demonstrate a specific (part of a) task needed to carry out the larger task(s) for one or multiple **Use-Cases**. For instance, a **Pilot** could demonstrate the execution of a federated genome-wide association study (GWAS)[1] across multiple sites (that is, without ever sharing any of the individual-level data). A list of **Pilots** is given in the *Appendix*. **Pilots** are expected to evolve in a stepwise manner and become increasingly sophisticated (if needed) over their development cycle. Below is a list of current **Pilots**, which will be evolved, as required, in an agile manner.

Pilot classes	Description
Federated Genomics	Demonstration of federated genomic computation applied to (for example) Genome-Wide Association Studies
Federated Catalog	Demonstration of the advanced semantic content discovery and cataloging functionality.
MMIO Functionality	Sequence of pilots that demonstrate adapter, provenance and governance functionality brought by MMIOs.
Multimodal machine learning	Demonstration of multimodal machine learning sequentially integrating MMIO and federated functionality.
Pharmacogenomics	Demonstration of a curation and analytical federated pipeline to determine how patients will respond to drugs
Federated WSI Phenotyping	AI prediction models for clinical outcome and tissue phenotyping
AI genetic data curation	Scalable genomic data curation and analysis
Tertiary genomic analysis	Improved clinical efficiency of variant prioritization
Accelerated Secondary Analysis	More effective and accessible tools for genomic data analysis

Tools and Infrastructure when incorporated in each **Pilot** demonstrate the better integration of clinical data, including genomics, for improved clinical outcomes. For instance, tools could be a specific software designed for a specific discipline, such as PLINK[1], SNPTEST[2], R including specific packages, Python including specific libraries, or ANNOVAR[3]. Infrastructure could be the appropriate governance required to execute a Use-case (and **Pilot**), but also the software and hardware design to execute the above example of executing a federated

GWAS. Below a summary of tools and the barriers addressed that currently impeded the implementation of a system such as **NextGen Pathfinder**.

Tool	Barriers addressed
Federated AI/ML	Insufficient local data volume and variety in the context of AI/ML compute.
Federated Genomics	Insufficient local data volume and variety in the context of genomics.
Federated Catalog	Compatible and available datasets not visible / discoverable.
MMIO	Multiple formats for underlying data.
MMIO	Auditability/security/governance constraints.
Accelerated genomics	Bottlenecks in secondary/tertiary genomic data processing limiting use.
Variant prioritization	Inefficiencies in the ranking of variants by significance.
Genomic data curation	Inadequate scaling of genomic data curation pipelines limiting use.

3.4Six major tasks

There are 6 interconnected tasks to be executed for the **Pathfinder**, listed below:

- **T5.1: Pathfinder Frameworks and Deployment Pathways (QMUL, UMCU, SUPSI, EURE: M1-M6)**
  To establish the Pathfinder frameworks and deployment pathways, a plan for the staged development of **Pilots** demonstrating **Pathfinder** functionality will be created; this document **Blueprint for Pathfinder: frameworks and pathways**. This plan will include the high-level and functional designs for each pilot, mapping out requirements and dependencies. These requirements and dependencies will cover aspects such as real-world and synthetic data, real-world and virtual data sites, technical functionality, tooling prototypes, and use-case deliverables. An agile **Pilot** deployment plan will be developed based on the estimated delivery of data management (*WP1*), platform (*WP2*), tools (*WP3*), and use cases (*WP4*) to ensure the complete and fully functional deployment of the **Pathfinder**.
- **T5.2a: Genomic Curation/Interpretation Pathfinder Functionality (SUSPI, EURE, ALL: M6-M42)**
  A **Pilot** for genomic data curation and interpretation will be developed, integrating AI-based genomic data curation and interpretation functionality into the **Pathfinder**. This will enhance the **Pathfinder's** capability to handle genomic data efficiently, ensuring that the data is curated and interpreted accurately and effectively. This task also involves establishing synthetic genomic data specifications (where required) along with optimal generative and validation approaches.
- **T5.2b: Genomic Acceleration Pathfinder Functionality (EURE, WSPAN, ALL: M6-M42)**
  A **Pilot** for secondary and tertiary genomic acceleration analysis will be developed, integrating accelerated secondary and tertiary analysis solutions into the **Pathfinder**. This integration will be benchmarked to demonstrate a significant reduction in time-to-insight, enhancing the overall efficiency of genomic analysis within the **Pathfinder**. This task also involves establishing synthetic genomic data specifications (where required) along with optimal generative and validation approaches.
- **T5.3: Multimodal Data Integration Pathfinder Functionality (HCF, SUPSI, QMUL, UMCU: M6-M24)**
  Three **Pilots** of **Use-Cases** requiring the multimodal data adaptor functionality will be created to demonstrate the capability of the Pathfinder in integrating various data modalities seamlessly. An inventory of key modalities and source format variations, including clinical data, device data, images, and omics, will be developed. This task also involves establishing synthetic genomic data specifications (where required) along with optimal generative and validation approaches.
- **T5.4: Federated Computation Pathfinder Functionality (SUPSI, QMUL, UMCU, EURE: M12-M36)**
  This task will identify **Pilots** from the above tasks that can benefit from applications of federated learning (FL). These **Pilots** will be extended, or new ones will be created, using synthetic or real-world data from use cases requiring federated machine learning and genomic data analysis. Specific applications will include genome-wide association studies (GWAS), polygenic risk scores (PRS),

clustering, dimensionality reduction, and supervised ML. The pilots will be used to assess the methods within the context of the **Pathfinder**, demonstrating the efficacy of federated computation.

- **T5.5: Governance and Catalog Functionality (HCF, HIRO: M12-M24)**
- This task involves establishing a metadata catalogue for the **Pathfinder** and mapping out governance and regulatory frameworks to be added. Successively complex prototypes for the **Pathfinder** will be developed, ensuring robust governance and comprehensive cataloguing of data. This will enhance the Pathfinder's ability to manage and govern data effectively, aligning with regulatory standards and best practices.
- **T5.6: Pathfinder Development (HCF, HIRO, SUPSI, EURE: M24-M40)**
  A plan for the staged development of **Pathfinder** components based on tasks T5.1 to T5.5 will be established. The complete multisite Pathfinder project will be deployed, incorporating multiple innovations and tools. This deployment will ensure that the **Pathfinder** is a comprehensive, fully functional system that integrates the best practices and innovations developed throughout the various tasks, delivering a robust and effective solution for data management, analysis, and governance.

3.5 Main deliverables

There are two main **Pilot** deliverables for the **Pathfinder**:

- **To implement Pilots.** This means implementing project tools needed for the **Use-Cases** thereby facilitating the execution of the research described in the given **Use-Case**. These **NextGen Pilots** will demonstrate a better integration of healthcare (research) data and how this will improve clinical outcomes.
- **Showcase a set of Pilots through the Pathfinder network**. This showcase should be developed using data from five demonstration biobank sites, including the Athero-Express biobank study (AE; UMCU, Utrecht, The Netherlands), Biobank of Karlinska Endarterectomies (BiKE; KI, Stockholm, Sweden), Helsinki Carotid Endarterectomy Study (HeCES; HUS, Helsinki, Finland), Coronary Artery Multiomics Analysis study (COMA; UVA, Charlottesville, VA, USA), and Munich Biobank (TUM, Munich, Germany).

3.6 Key Performance Indicators

These deliverables are measurable and verifiable. The *key performance indicators (KPIs)* are:

- 6 pilot demonstrations
- 5 sites included in **Pathfinder**
- at least 1 successful public **Pathfinder** demonstration
- regulatory, governance and data tooling demonstrated for 7 countries (SE, UK, CH, FI, USA, DE, NL)

3.7 Blueprint

This document serves as a **Blueprint for Pathfinder**. It is intended as a high-level, executive summary of the various components comprising the **Pathfinder**. It is meant to be concise and accessible to all partners, with as little as possible discipline specific jargon. It is also meant to be agile: as our work progresses, and the **Pathfinder** is developed, so will this **Blueprint**.

The goal of **Nextgen**, and by extension **Pathfinder**, is to inventory all the aspects that are required to make a federated health data analytics platform a reality in the near future. In other words, to find the answer to the question ‘How does one execute a GWAS or apply a machine learning model without ‘seeing’ the actual individual level data?’

The **Pathfinder** is realistic, we do not presume we will have fixed a platform that can be rolled out across institutions in or outside Europe. We have not asked or required the Pathfinder sites to share data or allow connections to their systems. However, we will demonstrate a simulated federated network at the minimum.

4 Staged development and agile deployment of Pilots

We establish a plan for staged development of **Pilots** and their functionalities, see Table below (a more detailed description is given in the next chapter: *High-level functional design*). Through iterative meetings with the different partners we inventoried **Use-Cases** and distilled **Pilots**, and **Tools and Infrastructure** from these. The Table lists 5 Pilots that have crystallized, see *Appendix* for an example of a **Pilot**, and 6 **Pilots** are in development (pending). In the next few weeks these will crystallize to the runup of the 2nd Consortium meeting after which the **Pathfinder** development will commence.

Pilot crystalized	Documented	Lead	Other Partners	2nd Assembly
Federated GWAS	yes	SUPSI	UMCU	Demo
Federated Catalog	yes	HCF	HIRO	Demo
MMIO Adaptor Functionalities	yes	HCF	QMUL	Demo
Accelerated Secondary Analysis	yes	EUR	WSPAN, QMUL	Demo
Multimodal ML for heart failure	yes	QMUL	HCF, SUPSI	No
Pilot in development				
Pharmacogenomics	Pending	UMCU	Pending	No
Federated WSI Phenotyping	Pending	GUF	UMCU	No
VCF "Merge functionality"	Pending	HCF	UMCU	No
AI genetic data curation	Pending	SUPSI	Pending	No
Tertiary genomic analysis	Pending	EURE	WSPAN, QMUL	No
MMIO Audit Functionalities	Mapping	HCF	Pending	No

We created an evolving GANNT chart for the management of **NextGen Pathfinder** (see **Figure 2**). This should be changing with time and continuing insights as the project progresses. However, hard deadlines are set by the due dates of the Tasks listed in the previous Chapter: *What is NextGen Pathfinder?*.

Name	Start Date	End Date	Duration
T5.1a: Blueprint of Pathfinder (Lead: UMCU)	Jan 01, 2024	Jun 28, 2024	130 days
T5.1b: Evolving Blueprint of Pathfinder (Lead: UMCU)	Jul 01, 2024	Dec 31, 2027	915 days
▼ Pathfinder Content Development	Jan 01, 2024	Sep 13, 2024	185 days
Use-Cases: inventory + description	Jan 01, 2024	Sep 13, 2024	185 days
Pilots: inventory + description	Jan 01, 2024	Sep 13, 2024	185 days
Tools and Infrastructure: inventory + description	Jan 01, 2024	Sep 13, 2024	185 days
T5.2a: Genomic Curation (Lead: SUPSI)	Jul 01, 2024	Jun 30, 2027	783 days
T5.2b: Genomic Acceleration (Lead: EURE)	Jul 01, 2024	Jun 30, 2027	783 days
T5.3: MMIO (Lead: HCF)	Jul 01, 2024	Dec 31, 2025	393 days
T5.4: Federated Computation (Lead: SUPSI)	Jan 01, 2025	Jun 30, 2026	390 days
T5.5: Governance and Catalog (Lead: HCF)	Jan 01, 2025	Dec 31, 2025	261 days
▼ T5.6: Pathfinder Development (Leads: HIRO)	Jan 01, 2024	Dec 31, 2027	1045 days
Design	Jan 01, 2024	Sep 13, 2024	185 days
Development	Sep 13, 2024	Jun 30, 2027	729 days
Finalizing	Jul 01, 2027	Dec 31, 2027	132 days

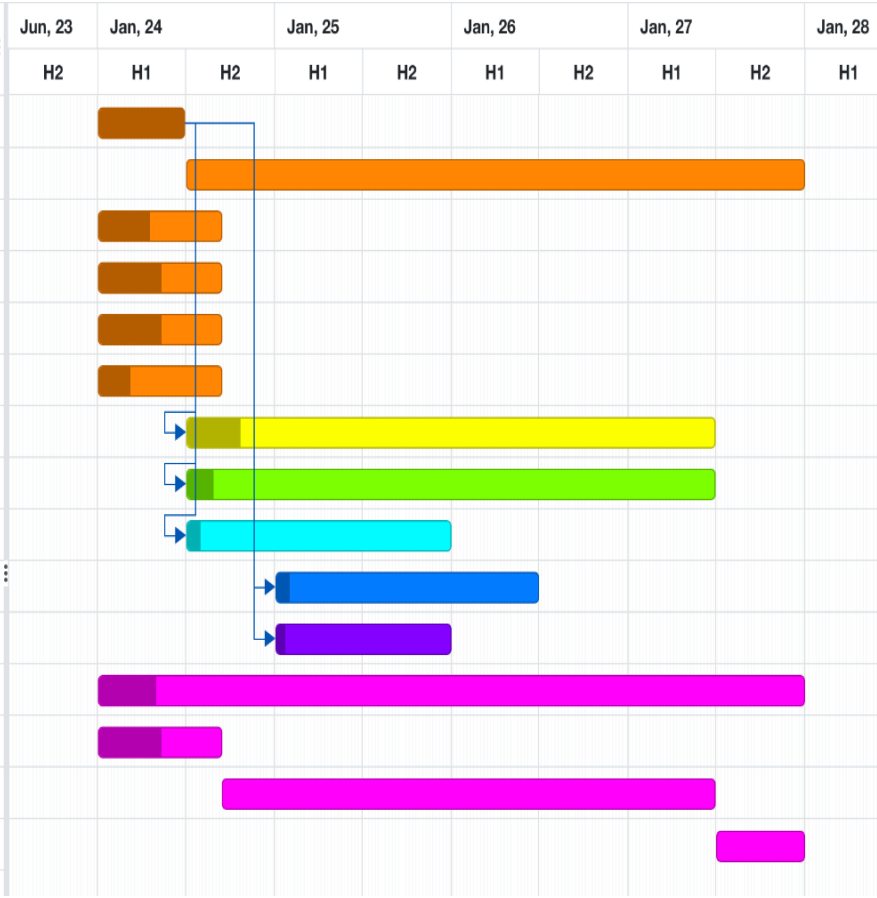

Figure 2: GANTT chart showing the agile development path for the NextGen Pathfinder. As the project evolves, so will this GANTT chart.

5 High-level functional design

Based on the discussions through *WP1*, *WP2*, and mostly *WP3* and *WP4* a few **Pilots**, and consequently **Tools and Infrastructure**, are distilled from the **Use-Cases**. We briefly touch upon each **Pilot** below to provide a high-level description of the functional design. A detailed description of an example **Pilot** is given in the Appendix.

5.1 Federated catalogs

NextGen semantic technology allows the true abstraction of meaning from format. **NextGen** cataloging supports discoverability of complex multimodal datasets within data ecosystems supporting specific research questions and allowing integration into federated computational pipelines.
Barriers addressed: Compatible and available datasets not visible / discoverable.

5.2 Multimodal integration objects (MMIOs)

Standardization of data formats alone is insufficient for research portability: it lags data generation (particularly for research using bespoke formats) and is not intrinsically multimodal. A given clinical research question requires a specified set of variables. Cross-border/federated portability requires this multimodal dataset be present at each site and ingestible despite heterogeneity of the underlying data formats and structures.

The **NextGen** multimodal integration object (MMIO) allows (i) conversion of the underlying constellation of multimodal formats into a user-defined harmonized AI/ML ready form, (ii) application of site-specific governance and regulatory requirements, and (iii) embedded authentication, audit, and integrity functionality. NextGen functionality allows true portability of multimodal, multi-omic research in a secure, auditable, and compliant manner independent of member-state location and underlying data composition.
Barriers addressed: Multiple formats for underlying data. Auditability/security/governance constraints.

5.3 Federated machine learning

Federated machine learning is an effective strategy for learning from distributed data without aggregation at a central site. Current approaches in federated learning either create a specialized application for each algorithm or use a distributed environment to share and run code among parties. The first approach is not easily generalizable and limits the possibility of end-users to experiment. The second is vulnerable to security threats such as malicious code execution. In **NextGen**, we decompose algorithms into predefined tasks and share tasks rather than code. Data owners will maintain complete control over their data, including the option to determine which data to share and when.
Barriers addressed: Insufficient local data volume and variety in context of AI/ML compute.

5.4 Federated genomics

Federated machine learning is now an established technique. However, the distributed computation of genome-specific algorithms (e.g. differential gene expression, polygenic risk scores (PGS), genome-wide association studies (GWAS), etc.) is considerably less advanced. In **NextGen**, we develop new federated implementations for genomic analysis where none currently exist. We will also extend new and existing methods (e.g. PGS) to privacy-preserving federated settings such as secure multiparty computation.
Barriers addressed: Insufficient local data volume and variety in context of genomics.

5.5 Accelerated genomics

The cost of whole exome and whole genome sequencing continues to fall, so that the bottleneck in the clinical adoption of genomics-based precision medicine has shifted from data generation to data analysis. Genomic data analysis is a computationally intensive process with multiple processing steps. With the amount of genomic data growing "exponentially" it becomes increasingly difficult to perform such an analysis in a timely and cost-efficient manner. In **NextGen** we develop several open-source hardware-independent approaches for accelerated genomic data analysis to enhance the integration of genomic data in multimodal contexts.
Barriers addressed: Bottlenecks in secondary/tertiary genomic data processing limiting use.

5.6 Variant prioritization

Genomic sequencing identifies variations in the genetic code. To develop diagnostic and treatment processes, variants need to be linked to diseases, and the "clinical validity" of a suggested gene-disease relationship is determined (variant annotation). This evidence-based process classifies relationships based on the level and quality of evidence. Genomic analysis produces variants lists from which gene-disease relationships are to be deduced by clinical scientists using established, but time-consuming, interpretation protocols. In **NextGen** we use machine learning to develop improved variant prioritization algorithms so that genes that are more likely to be causally related to the disease are ranked higher to reduce the manual processing time by an order of magnitude with the downstream benefit of shortening the time between presentation and diagnosis and improving patient outcomes.
Barriers addressed: Inefficiencies in the ranking of variants by significance.

5.7 Genomic data curation and analysis

Data curation in the genomics space requires a complex, integrated processing pipeline. Genetic association studies may encompass millions of genomic variants and stringent quality control is obligatory to establish reliable gene-disease relationships. Manual processes are not scalable and are time consuming. **NextGen** will develop extensible AI-guided genomic data curation pipelines, to complement other AI mediated data curation initiatives. Furthermore, the more accurate understanding of gene-disease relationships will allow the right genomic information to be included in multimodal datasets which will improve algorithmic precision and predictive power.
Barriers addressed: Inadequate scaling of genomic data curation pipelines limiting research useability.

6 Requirements & dependencies

We mapped out or are in the process of mapping out the necessary requirements and dependencies for each **Pilot** including:

- What real-world and/or synthetic *c.q.* virtual **data and sites** could be best used for development
- What the **federated catalog** should look like
- What the **technical functionality** and tooling prototypes should be

6.1 Data and sites

On a case-by-case basis we will map what data and partner sites may be required to develop a **Pilot**. We will aim to use real-world data where possible, this will also test our understanding of the required governance and aid in further improving Task 5.4.

6.2 Federated cataloging

To facilitate federated data curation and analysis, be it through AI or other means, we require a **federated catalog**. The federated catalog will provide high-level access to available meta-data and data of the connected sites and enables the user or tools to determine which datasets are usable for the given **Use-Case** or **Pilot**. Such a data catalog will hinge on complete, yet concise, and structured data. Thus, we mapped and continue to map throughout the different partner sites the following:

- What datasets are available?
- What are the data semantics?
- What are the different data formats used?
- How data can be accessed and what the 'rules of engagement' are in this context? In other words, what is the governance of the data?

With data semantics we mean to describe what the meaning and use of specific pieces of data are in data curation and analytics (or bioinformatics) that make use of these data. In other words, we describe and define, in an orderly fashion, what each component of a dataset is, and how it could be used. For instance, in modern-day complex genetics, the 4-letter DNA code is captured in the *variant call format*, VCF, v4.2 format (`.vcf`) described here: https://en.wikipedia.org/wiki/Variant_Call_Format. In data semantics it is described what is meant in each row and column of the file, and how this data can be used and with what tooling.
For a well-functioning federated catalog, it is critical to develop a good understanding of the national rules and regulation regarding the governance of the datasets at partner sites. The regulation, ethics and governance needed for **NextGen Pathfinder** and the federated tools to function are mapped in relation to *WP5* are mapped in collaboration with *WP6*.
This **federated catalog** is currently being mapped and the final product is expected to be delivered through Task 5.5.

The **Nextgen Platform data management requirements** were mapped, and the following are listed.

- Federated data catalog
    - Contains metadata that is managed by the federated catalog owner, i.e. the partner site, and allows control of the metadata to update, add, delete, searched. It also allows policies to be written and applied metadata and underlying actual data.
    - Policies enable authorized individuals and processes to interact with the metadata and provide them access to the underlying actual data by API, Connector, etc. Type and level of interaction with (meta)data depends on the type of authorization and policy. Interactions can be seeing the (meta)data, querying (meta)data, change of (meta)data, etc.
    - Authorization starts with data owners and their refined authorizations, including policies to others.

- **Sovereign Data Sharing Infrastructure:** the project will progress their understanding of a Sovereign Data Sharing Infrastructure and accordingly develop additional tools for reputable transactions, a data ecosystem that facilitates decentralization compliant with European values and is independent of closed platforms models and non-European tech players.
- **Content-Based Identifiers and Authentication:** **NextGen** will implement a system of content-based identifiers based on what the data is independently of where it is located. The control of these identifiers to create, access, transform or share data will rely on decentralized authentication technologies (i.e. KERI) and verifiable credentials. This results in novel search-and-match functionalities that can be linked to consent management and a reputation system to develop stakeholder recommendations.
- **Interoperability**: we aim to develop **NextGen** federated catalogs (*WP1*), integrating federated query system. This includes:
    - ***Domain specific*: Semantic Framework and Interoperability**: **NextGen** will develop a semantic framework (*WP1*) based on the refined eHealth European Interoperability Framework recommendations but adapted to the more decentralized and data-centric architecture demanded by the EHDS concept. The framework will implement all the adaptors required to meet the standards interoperability requirement of B1MG.
    - ***Domain specific* Semantic ontologies**: The traceability of the ontologies used throughout the data processing is provided by recording ontologies in the meta-data.

6.3 Functional Requirements

6.3.1 General requirements

The **Nextgen Platform general requirements** were mapped, and the following are listed.

- **Cardiovascular related applications and supportive tools including** data space infrastructure, DOA, and federated learning for:
  - Access to data(set) catalog containing relevant datasets.
  - Federated access to data sets relevant for the training of AI models.
  - Management the training of AI models including federated learning in a distributed federated infrastructure.
  - Application of trained AI models on patient data across the patient journey (health status);
  - Analysis causal inference.
  - Support of health policymakers.
  - Support of diagnosis by the healthcare professional.
  - Establishment of stakeholder recommendations.
- **NextGen platform data outputs:** data analysis outputs are structured to facilitate their usage as data input for further data usage, for example reporting, analysis, presentation, in healthcare and medical decision support systems.
- **Multi-site Support:** The dataspace should support operations across multiple sites, allowing collaboration and data sharing among locations.
- **Cross Platform Automated Orchestration:** A service for automated coordination of complex processes or systems and using a tool or platform to automate the execution for; model training, infrastructure management (storage, servers, networking), application deployment, and workflow automation).
- **Roles:** Data owner, Data producer, Data provider…Model owner, model user, etc. the definition of roles for **NextGen** will be customized to the minimal viable product that the pathfinder project develops.
- **Data cleaning:** **NextGen** should provide tools to execute data cleaning.
- **Data Imputation, transfer learning, and meta-analysis:** **NextGen** should implement tools for data imputation, transfer learning, and meta-analysis to overcome missing or lack of data.
- **Data products:** Data products are created from processed, shaped, cleansed, aggregated, and normalized raw data and meet agreed-upon quality standards for analytical consumption. These should be available for users, however NextGen will not provide a lasting repository of these.
- **Transformational Tools for Data Formats:** Transformational tools that take underlying multimodal datasets (specified for each research question with the variables and modalities needed) that may exist in multiple formats and standards on the different sites – and seamlessly transform these groups of inputs to predefined and common formats (ML ready if needed) to ensure cross-site portability of research. This is called the **multi-modal integration object (MMIO)** and is developed by *WP1* in collaboration with *WP2* and *WP4*.
- **Data formats:** **NextGen** should be able to support different formats of **data products** that are going to be processed by the **MMIO**.
- **Observability/Monitoring:** Auditable protocols for data quality should be defined.
  - **Data monitoring:**
    - Every partner shall be able to monitor and audit their data, to
      - Gain insights about the data
      - Know what happened to the data
  - **Data space monitoring:**
    - To monitor the data space
    - Includes logs
- **Data Hub/Lake/Lakehouse Connector**: Able to schedule and submit queries, in the end, user data hub/lake/lakehouse, receive a storage location (caching) of the query outcome, receive storage location of data product, fingerprinting the Data Hub/Lake/Lakehouse performance indicators (e.g. latency, the life cycle of caching storage), combining queries into a data product.

- connect with data hub/lake/lakehouse
- execute queries against them
- store the results in the cache
- combine the results

6.3.2 AI model management requirements

The **Nextgen Platform AI model management requirements** were mapped, and the following are listed.

- **Privacy-Aware ML Methods**: **NextGen** will work to guarantee awareness of the proposed ML methods concerning privacy threats by providing quantitative measures of privacy leakage (such as differential privacy) so that the maximum amount of shared data can be bounded, and algorithms chosen correctly to account also for this aspect.
- **Sandbox**: A safe, secure environment where a model and a sample dataset can be tested. The sandbox also creates a transparent view of the model (explainable AI).
- **Training builder (training and prediction pipelines)**: Secure environment where assets (data, models, infrastructure, services) can be chained together into a fully planned process, before being sent to the clearing house and cross-platform automated orchestration.
- **Trainer**: Secure GDPR compliant environment to execute the training of a model. A more detailed map towards the development of certification and a risk management framework for the GDPR compliance will follow in the coming months. While input data are not transferred through the platform, model parameters and other meta-data are transferred, these may pose a privacy and security risk which needs proper evaluation and risk mitigation.
- **Federated Machine Learning:** The platform should facilitate federated machine learning.
- **Genomics curation and analysis**: The platform should facilitate distributed genomic curation and analysis: data quality control and imputation, differential gene expression, genome-wide association studies, polygenic risk scores creation and analysis, clustering, and dimensionality reduction.
- **Accelerated Genomic Analysis**: Open-source software for secondary and tertiary genomic analysis acceleration. This should include:
  - **Benchmarking Capabilities**: Incorporation of benchmarking capabilities to assess the performance and efficiency of data analysis processes within the dataspace environment. (for Accelerate Genomic Analysis)
- **Model Validator**: Tools for researchers to annotate the models and datasets to be used by other researchers.
- **Model Deployer**: Tools to deploy a model into the existing infrastructure of the end customer, health care practitioner.

The **Nextgen Platform MarketPlace services requirements** were mapped, and the following are listed.

- **Marketplace**
  - It is a Virtual meeting place for the supply and demand of data and services. Marketplace has a visitor recommender system that draws its intelligence from interactions on the Marketplace by all stakeholders.
  - Catalog of all available data, services, and assets in the **NextGen** dataspace and other dataspaces should be searchable based on all information stored in the **MMIO**, and other published metadata provided by those other dataspaces. Connects to the Marketplace.

6.3.3 Security, authorization and policies requirements

The **Nextgen Platform security, authorization and policies requirements** were mapped, and the following are listed.

- **Security, authorization and policies**
  - **Trusted Data Agent (TDA)**: an agent to authenticate the MMIOs held by other stakeholders or published in the data space catalog and MarketPlace without a third-party authentication/ identification service.
  - **MMIO**: Unique authentication of data, asset, service, infrastructure. Harmonization of a multi-omic multimodal dataset by converting the underlying heterogenous constellation of data into

a user-defined AI/ML ready form consistent with data requirements (Trustworthy AI). The authenticity component of MMIOs is enforced by binding digital fingerprints (cryptographic hash of digital content that can represent a pointer for the location of a digital asset(s) or an attachment(s) of the asset(s) itself) for schemas (objects), records (events), policies (rules), and agreements (actions) to a central core as a transient integrated object for processing data at the edge. The cryptographic integrity of the MMIO allows governance mechanisms like consent to be enforced by machine actionable governance administration. The MMIO can be transferred between users without necessarily moving the data.

6.3.4 Governance requirements

The **Nextgen Platform governance requirements** were mapped, and the following are listed.

- **Governance**: Repository of the governance structure in **NextGen** including schemas that can be attached to the MMIO.
  - **Electronic consent management**
  - **Data provenance tracking**
  - **Date Ecosystem Governance Server (DEGS)**: Server with hardware root of trust) that carries the complete and most up-to-date set-up of the data ecosystem and overview of the distributed data ecosystem.
- **Clearing house**: Generic, cross-domain service receiving information about transactions, participants, and references to existing legal contracts, storing in a non-reputable, verifiable form, and making it available to the participants. Where needed billable data usage is settled before the escrow locker gets unlocked and the data service is consumed.
  The Clearing house also has a scientific and ethical body to assess any potential risks in the proposed research.
  - **Escrow Locker**: An arrangement to hold the assets of a transaction temporarily. The assets are kept in a third-party account and are only released when all terms of the agreement have been met. The use of an escrow account in a transaction adds a degree of safety for both parties.

6.3.5 Non-Functional Requirements

The **Nextgen Platform non-functional requirements** were mapped, and the following are listed.

- **GDPR compliant** access to sensitive personal data, including clinical and genomic data
- **Scalability** of data management resources
- **Data Protection:** schedule snapshots and backups of data flowing through the **NextGen Platform**

7 References

1. Chang, C. C. *et al.* Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience* **4**, 7 (2015).
2. Galesloot, T. E., van Steen, K., Kiemeney, L. A. L. M., Janss, L. L. & Vermeulen, S. H. A comparison of multivariate genome-wide association methods. *PLoS One* **9**, e95923 (2014).
3. Wang, K., Li, M. & Hakonarson, H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res.* **38**, e164 (2010).

8 Appendix

In this appendix we describe an exemplar **Use-Case** and **Pilot**. Note that use cases and pilots are developed in agile manner in Didact Gothic, both to optimise scientific validity and to best demonstrate the potential improvement in clinical outcomes through the use of **NextGen** tools.

8.1 Exemplar Use-Cases

8.1.1 Accelerated variant annotation

Below is a high-level description of an exemplar "Accelerated Variant Annotation" **Use-Case**.

Description of clinical research use-case
Development of accelerated variant annotation (tertiary bioinformatics) in those with disease phenotype and in healthy individuals (i.e. screening for Precision Health) to classify genetic variants into American College of Medical genetics and Genomics (ACMG) classification. Combining variant annotation with phenotype data to identify individuals with early subclinical disease using electronic health records (diagnostic codes), 12-lead ECG, cardiac magnetic resonance (CMR), to identify high-risk individuals and prevent sudden death, all-cause mortality and morbidity (heart failure, atrial fibrillation and stroke). For example, *FLNC* is a highly penetrant and lethal Mendelian cause of dilated cardiomyopathy and arrhythmogenic cardiomyopathy, where the presence of 12-lead ECG abnormalities such as T wave inversion, presence of late gadolinium enhancement on CMR, indicate disease, even when left ventricular ejection fraction is abnormal, with a high risk of sudden death.
What barriers in this use case will NextGen remove?
Reduction of bottlenecks in variant annotation phase by multiple innovations with a goal of >10x improvement. For prediction of subclinical phenotype, using EHR/genomic/ phenotype data, the construction of a multi-modal data object, specific to this particular research question (i.e. including the appropriate deep phenotypical multi-modal data) will address the enhanced data training need by ensuring potability of the algorithm to other locales as well as federated contexts.
What data will be needed, including synthetic data (testing, analysis, imputation)?
Whole exome sequencing (WES), whole genome sequencing (WGS), 12-ECG, imaging data, specific clinical information from EHR will be needed to develop the algorithms. To demonstrate the "adaptor layer" effect of using multi-modal data objects (which are agnostic to underlying data formats, when transformation layers have been defined), a variety of formats for underlying data will be needed (per modality).

8.1.2 Accelerated secondary genomic analysis

Below is a high-level description of an exemplar “Accelerated Secondary Analysis” **Use-Case**.

Description of clinical research use-case
Cardiac disorders including heart muscle, dyslipidaemias, arrhythmias and aortopathies, congenital heart disease can affect about 1% of living births, and up to 3% of the population depending on cardiac defects considered. Although chromosomal aberrations, copy number variants and de novo-mutations within protein coding genes have been associated with the outcome, up to 72% of congenital heart defect do not have an identified genetic basis. Mendelian cardiovascular disorders can be relatively common due to adult onset, after reproducing and passing variants on to offspring: dilated cardiomyopathy affects 1 in 250 and hypertrophic cardiomyopathy affects 1 in 500. It is very likely that a substantial proportion of those unexplained outcomes could be associated with damaging mutations within functional noncoding regions of the genomes associated with gene regulation (e.g. promoters, enhancers). ERLH is carrying out research on integrating genomic, transcriptomic, epigenomic data for the reconstruction of the regulatory networks, and the annotation of the noncoding functional elements involved in cardiac regulatory networks.
What barriers in this use case will NextGen remove?
This use case relies on a secondary data analysis pipeline to transform raw WES/WGS reads into actionable variants. The most popular and widely used pipeline is the GATK pipeline from Broad Institute. This pipeline is well known to be computationally intensive and extremely time consuming especially for genotyping large datasets. As a consequent the effective incorporation of genomic data into subsequent analytical pipelines is hindered. The secondary analysis pipeline developed here is a foundational infrastructure component that can democratize accelerated genomic data analysis and its application in several research areas.
What data will be needed, including synthetic data (testing, analysis, imputation)?
Publicly available WES or WGS data such as the 1000 human genome project. Other datasets will be used as available and in appropriately secure and GDPR compliant contexts.

8.2 Exemplar Pilot

Below is a detailed description of an exemplar "Federated GWAS" **Pilot**.

Genetic analyses	
Responsible partners	**UMCU, UVA**
Description of use case (clinical question/research/activity)	
Clinically relevant genomic analyses, such as GWAS, etc.	
Description of pilot (demonstration of project tools applied to the use case)	
Execute a GWAS, a meta-GWAS, a loss-of-function (LoF), polygenic score creation and analysis (PGS), or a quantitative trait locus (QTL) analysis using data from multiple sites without moving data.	

APPLICATION OF STANDARD TOOLS			
Tool	**Barriers addressed**	**Used?**	**Details**
Federated AI/ML	Insufficient local data volume and variety in context of AI/ML compute.	No	-
Federated genomics	Insufficient local data volume and variety in context of genomics.	Yes	GWAS+
Federated catalogue	Compatible and available datasets not visible / discoverable.	Yes	TBD
MMIO	Multiple formats for underlying data	No	-
MMIO	Auditability/security/governance constraints	No	-
Accelerated genomics	Bottlenecks in secondary/tertiary genomic data processing limiting use	No	-
Variant prioritisation	Inefficiencies in the ranking of variants by significance	No	-
Genomic data curation	Inadequate scaling of genomic data curation pipelines limiting use	No	-

DATA		
Dataset type	**Definition**	**Used?**
Local	Individual level data available at a specific site.	No
Test	Test data NOT derived from individual level patient data	Dummy for a dummy meta-GWAS > from MetaGWASToolKit: https://github.com/swvanderlaan/MetaGWASToolKit/tree/master/EXAMPLE/RAWDATA.
Synthetic	Test data derived from individual patient level data	No
"Open" public	Data that is publicly available or downloadable and which may have modest usage requirements or restrictions.	Welcome Trust Case-Control Consortium (WTCCC) to validate an actual GWAS in CAD on 9p21
"Closed" public	Data that is available on application/payment with a high level of restrictions	No?

ADDITIONAL/PROPOSED TOOLS

- GWAS QC: in-house developed collection of scripts that executes quality control.
- GWASToolKit: https://github.com/swvanderlaan/GWASToolKit
- LoFTK: https://github.com/CirculatoryHealth/LoFTK
- MetaGWASToolKit: https://github.com/swvanderlaan/MetaGWASToolKit
- QTLToolKit: https://github.com/swvanderlaan/QTLToolKit
- PGSToolKit: https://github.com/swvanderlaan/PGSToolKit

ACTIONS MAPPED TO WORK PACKAGES		
Work package tasks	**Specific actions for this pilot**	**Tasks**
WP1 – Data Management	Develop relevant data catalogue over sites	Data discovery (T1.2, T1.4)
WP2 – Infrastructure	Incorporate pilot needs into blueprint	Blueprint (T2.1), Services (T2.3)
WP3 – Tools	Development of federated genomic computation, and privacy assessment. Assessment of any additional data needs.	Federated genomics (T3.3), Privacy Assessment (T3.4), Test data (T3.5)
WP4 – Personalised Medicine	Application of project tool (federation) to the research use case.	Incorporation tools (T4.3)
WP5 – Pathfinder and Pilots	Translation of functionality developed in WP3 to this pilot and implementation in the Pathfinder.	Federated implementation (T5.4), Catalogue (T5.5), Pathfinder integration (T5.6)
WP6 – REG	Ensure that all development related to this pilot has appropriate REG coverage.	Framework/guidance (T6.2)
WP7 – Broader Engagement	Ensure outputs add to the existing ecosystem (novel, synergistic).	Gap analysis (T7.1)