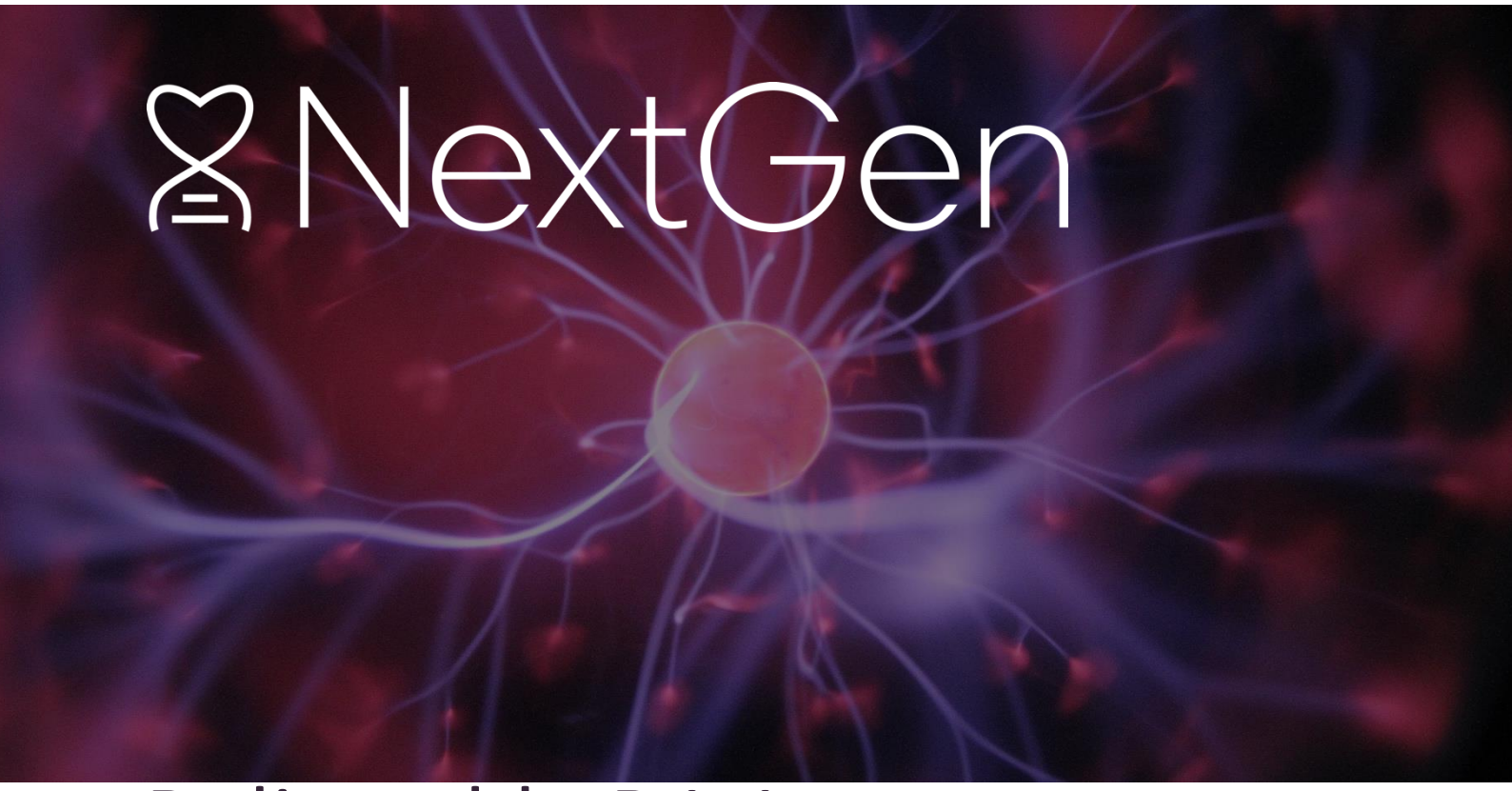

Deliverable D1.1

Data Management Plan 1.0

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

Autopoietic Cognitive Edge-cloud Services	
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]

D1.1 – Data Management Plan	
Author(s)	Ph.Page (HCF), P.Knowles (HCF), R.Benjamin(HIRO),
Editor	*See table below*
Participating partners	HCF, QMUL, HIRO
Version	1.0
Status	Final-DRAFT for Internal Review
Deliverable date	M6
Dissemination Level	PU - Public
Official date	June 30th 2024
Actual date	June 28th 2024

Disclaimer

The NEXTGEN consortium consists of the following partners:

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

Document Revision History

DATE	VERSION	DESCRIPTION
22/03/2024	0.1	Initial draft on EU Horizon DMP template
31/05/2024	0.4	Draft 1 for consortium review
06/26/2024	0.9	Final draft for consortium review
06/30/2024	1.0	**D1.1** Data Management Plan & Data Integration Framework
01/01/2026	2.0	**D1.2** Updated Data Management Plan
11/30/2028	3.0	**D1.5** Final Data Management Plan

Authors

EDITOR	ORGANISATION
Philippe Page	HCF
Aaron Lee	QMUL
Paul Knowles	HCF
Rafy Benjamin	HIRO

Reviewers/Contributors

REVIEWER	ORGANISATION
S.Haitjema	UMCU
J. van Setten	UMCU
S.van der Laan	UMCU
C.Barat	ESC
L.Remotti	DPOW
D.Malpetti	SUPSI
F.Mangili	SUPSI

A.Lee	QMUL
I.Hering	DPOA

List of terms and abbreviations

ABBREVIATION	DESCRIPTION
AI	Artificial Intelligence
B1MG	Beyond 1 Million Genomes
D&C	Dissemination and Communication
DOA	Data Oriented Architecture
EHR	Electronic Health Record
FAIR	Findability, Accessibility, Interoperability, Reuse
FL	Federated Learning
GA	Grant Agreement
GAL	Genomics Acceleration Library
GDI	(European) Genomic Data Infrastructure
GDPR	General Data Protection Regulation
GPU	Graphics Processing Unit
HTA	Health Technology Assessment
IDSA	International Data Space Association
IP	Intellectual Property
KPI	Key Performance Indicator
MDR	Medical Device Regulations
ML	Machine Learning
MMIO	Multimodal Integration Object
MVDE	Minimum Viable Data Ecosystem
OSG	Open Science Guidelines

ABBREVIATION	DESCRIPTION
PCA	Principal Component Analysis
PRS	Polygenetic Risk Scores
QC	Quality Control
REG	Regulation, Ethics and Governance Board
SCDY	Sudden Cardiac Death in Young
TDA	Trusted Data Agent
VC	Verifiable Credentials
XAI	Explainable Machine Learning

Table of contents

1 Data Summary

1.1 Overview of NextGen Data Usage

NextGen addresses the complex problem of integrating multiple types of data (multi-modal data), including genomic data, into research pathways that might include AI tools. NextGen develops tools to overcome barriers in data integration and access, to lead to an improved clinical outcome from healthcare research.

NextGen exploits on federated technology and decentralized data management techniques centrally aggregating data is outside the declared project scope (which is particularly relevant in the context of data subject to the GDPR).

Starting with datasets accessible by NextGen participants (see table 1.X) locally or through biobanks, the integration tooling listed in table 1.X will first be deployed locally while a concept of decentralized platform for research portability across different sites is developed around specific pilots defined during the project.

1.2 NextGen accessible datasets

in the following section, we address the critical issue of the accessibility of these diverse datasets, which often face constraints related to data silos, differing formats, and varying levels of accessibility due to privacy and proprietary concerns. NextGen focuses on developing robust tools to overcome these barriers and, in this phase, provides an overview of the data types expected.

1.2.1 Definition of Data Types:

The following table outlines the types of data covered by this DMP. It does not consider implications from the perspective of the regulatory (i.e. GDPR) compliance or governance perspective (REG), which are addressed later.

Dataset type	Description
Site or partner specific datasets	
Local Data (LD)	Individual level data available only at a specific Partner site.
Local Federated Data (LFD)	Data from local sites made accessible in federated applications.
Derived Data (LDD)	Data generated as a research output.
Data for development	
Test Data (TD)	Overarching term including Dummy/Synthetic/Proxy Data where such data is used for test purposes.
Dummy Data (TDD)	Test data generated WITHOUT using individual level patient data.
Synthetic Data (TSD)	Test data which may be generated using individual patient level data.
Proxy Data (TPD)	Test data derived from non-human models (such as livestock genomic/phenomic data).
Test Federated Data (TFD)	Test data for federated applications (may be TDD, TSD, TPD)

External/widely available datasets	
External Data	Data that is publicly available (Open Access Data or On Application Data).
Open Access Data	Data that is publicly available or downloadable and which may have modest usage requirements or restrictions.
On Application Data	Data that is available on application/payment with a higher level of restrictions.

1.2.2 Use, re-use and generation of data

Within NextGen, data is used in the following ways and further explained in a table:

- Clinical and academic research
- Testing of systems that process data
- Development of algorithms for enhanced data processing/manipulation

Data usage
Clinical & Academic Research
When carried out at Partner sites Local Data will be used in compliance with the GDPR. As such Local Data may also be Personal Data, it will not be generally made available to the Consortium and any discretionary data sharing between Partners will be subject to the GDPR and other governance measures including Data Sharing Agreements. External Data may also be used in compliance with the applicable requirement. Specific research applications may also use Proxy, Dummy or Synthetic Data, noting that Synthetic Data will be subject to the GDPR and will be subjet to a DPIA
System test Data
For the effective operation of the project, Dummy Data and Synthetic Data will be generated. These forms of data are highly configurable and can be used to test and develop basic functionality. Dummy Data is (by construction) outside of the GDPR which allows rapid development of technological components
Processing/Manipulation Algorithm development
The type of data used will depend on context. NextGen will carry out a number of activities on multi-modal data including genomic data for which Proxy Data is likely to be highly effective[1] and without privacy risk. Synthetic or external data (which may require assessment under the GDPR), can be used in research algorithm and is particularly advantageous as it can be built to demonstrate that the algorithms functions as intended on understood/constructed data before application to unknown/real data. Algorithm development will take place at local sites and whenever possible, be ported to different sites with the aim to create a decentralized platform enhancing research portability.

Generation of data: Outside of Dummy/Synthetic Data generated and used as detailed above, participating organisations may produce data as part of their scientific research. Use of such Derived Data will:

- Follow FAIR principles.
- Be subject to legal and regulatory constraints.

- To the extent possible be generated for multi-jurisdictional use

The following table provides further details on the data generated:

Data Generation
FAIR principles
NextGen decentralized data management includes by design the FAIR principles as detailed in section 2 of this DMP. In addition, NextGen aims to go beyond FAIR by introducing discovery mechanisms leading to data catalogues including governance meta-data.
Legal & Regulatory Constraints
NextGen operates under GDPR and other applicable laws and regulations in a specific location. To lower multi-jurisdictional barriers, NextGen adopts a Privacy-and-Consent by Design approach with Data Protection specialists involved in the design phase as well as a regular follow up of the evolving regulatory landscape impacting health data management, genomic data in particular.
Multi-jurisdiction data generation
The project will generate data in NextGen participant's locations within their jurisdictions. NextGen aims to facilitate research portability across the participants through privacy-enhanced federated machine learning techniques and federated catalogues. Therefore, NextGen will create, to the greatest possible extent, a secured multi-jurisdictional data space for personalised cardiovascular medicine. A risk-based approach and sequential technology developments through pilots demonstrating impact ensure the development of a specific risk management framework and governance measures before NextGen multi-jurisdictional exchanges occur.

Available data: NextGen academic and clinical partners have access to Local Data derived from multimodal datasets including routine clinical data, biochemistry, imaging data, device data, genetic/genomic/multiomic data and bio samples in developing tools for personalised medicine with desirable characteristics including:

- Multiple distinct populations to reduce bias, improve stability and representativeness.
- Large sample sizes to improve accuracy and precision.
- Longitudinal data to allow (e.g.) modelling of disease trajectories.
- Multimodal data to allow embedding of inferred/hidden biomarkers (shape, volume, etc)

The table below lists the expected data type to be used.

Local Data			
Biobank name (partner, country, PF/UC)	Design	Population	Data types
Coronary Artery Multiomics Analysis study (**UVA**, USA)	cohort-biobank	Heart transplantation patients, donors	WGS, Multiomic (bulk and single-nucleus), clinical, WSI
Athero-Express Biobank Study (**UMCU**, NL)	cohort-biobank	Arterial endarterectomy patients	GWAS, clinical, WSI
Helsinki Carotid Endarterectomy Study (**HUS,** FI)	prospective cohort	AE/CE patients	Clinical, imaging, GWAS, serum biomarkers, plaque (WSI, proteomics, transcriptomics, lipidomics)
Biobank of Karolinska Endarterectomies (KI, SE)	cohort-biobank	CE patients	Clinical, imaging, GWAS, serum biomarkers, genetic, histological imaging, plasma proteomics, plasma metabolomics
Munch vascular biobank (TUM, DE)	cohort-biobank	AE patients	Serum, plaque (WSI & evaluation, proteomics, transcriptomics), clinical, imaging
CE=carotid endarterectomy; AE=arterial endarterectomy; WSI=Whole Slide Images; GWAS=Genome-Wide Association Study;			

The datasets accessible by the participants at the onset of the project are listed here. We indicate the location to evidence the multi-jurisdiction dimension of the NextGen project.

Research datasets (External Data or Proxy Data)	
Partner: ERLH	Multiple locations
Data types: WES or WGS data, transcriptomics and epigenomics	
Database(s): GTEx, Encode, PCGC, Conservation, Human Gene Mutation Database	
Details: GTEx (RNA-Seq and WGS: 372 atrial appendage, 386 left ventricle); Encode ChIP Seq (16 individuals); PCGC – congenital heart disease: (phs001735): 4547 WGS, 3218 genotypes; Conservation: conservation scores across 242 mammals (phylOP, phastCons, CADD, LINSIGHT);Human Gene Mutation Database; ClinVar	
Partner: WSPAN	USA
Database: Gene Health Project at WellSpan Health	
Data types: WES, WGS, SNPs, clinical data, ECGs, imaging data	

Details: N=2.2M with EHR (Epic); WES and WGS=100k; 12-ECG extends to year 2000 in XML format, imaging extends to 1995 in DICOM	
Partner: KI	Sweden
Database: Biobank of Karolinska Endarterectomies	
Details: Described in **Error! Reference source not found.**	
Partner: UVA	USA
Database: Coronary Artery Multiomics Analysis study	
Details: Described in **Error! Reference source not found.**	
Partner: UMCU	Netherlands
Database: Lifelines biobank[2]	
Details: Extensive collection of medical data for 167,000 individuals.	
Partner: ALL (subject to approval)	UK
Database: UK Biobank[3]	
Details: Extensive collection of medical data for 500,000 individuals.	

1.2.3 Modalities and data Format used and generated

A summary of the data modalities and format is provided in the following table.

Type/Modality	Data formats	Expected use
EHR data	SPSS, OMOP, CDISC, FIHR	Yes
Post-mortem registry data	PDF, JSON, DICOM, CDISC	Probable
Genetic data (including genomic, proteomic, transcriptomic, epigenomic, etc)		
Whole exome sequence data (WES)	FASTQ, SAM, BAM, VCF, BGEN, PLINK	Yes
Whole genome sequence data (WGS)	FASTQ, SAM, BAM, VCF, BGEN, PLINK	Yes

[2] https://www.lifelines.nl/

[3] https://www.ukbiobank.ac.uk/

Type/Modality	Data formats	Expected use
Genome Wide Association Studies (GWAS)	VCF, PLINK binary format (bed/bim/fam), OXFORD-format (bgen and gen). Flat tabular format (csv, tsv) for summary data	Yes
DNA methylation	R .RDS (SummarizedExperiment), IDAT, CGmap, ATCGmap, VCF	Yes
Transcriptomic	FASTQ, SAM, BAM, VCF. .txt (counts), R .RDS (SummarizedExperiment)	Yes
Proteomic	.xlsx	Yes
Epigenomic	BED, BED Graph	Possible
Clinical imaging data		
Magnetic resonance (MR)	DICOM, NIFTI	No
Computed tomography (CT)	DICOM, NIFTI	No
Extracted images (MR, CT, other)	Standard image formats (PNG, TIFF, JPG, pixel array, etc)	No
Histopathological data		
Whole slide image data	Hamumatsu .ndpi; Aperio .TIF	Yes
Device data		
Electrocardiogram	XML, DICOM	No
Echocardiogram	DICOM	No
Derived/measured data		
Coronary artery calcium scores		No

1.3 Generated data from NextGen Tooling

NextGen develops tooling to

- Provide an accelerated, vendor agnostic secondary and tertiary genomic analysis;.
- Extend standard genomic methods to a secure federated version.
- Enable a secure semantic data harmonisation for multi-modal data integration.
- Multimodal data integration and research portability.
- Extension of secure federated analytics to genomic computation.
- More effective federated learning over distributed infrastructures.
- More effective and accessible tools for genomic data analysis.

- Improved clinical efficiency of variant prioritisation.
- Scalable genomic data curation.
- Improved data discoverability and data management.
- Pathfinder: fully modelled data ecosystem.

NextGen tools will be demonstrated through pilots defined around the research use cases of participating organisations. All project tools will be included in the NextGen risk assessment procedures.

2 FAIR data -Beyond Fair

The NextGen project follows FAIR principles and, in addition, provides researchers with data management tools for FAIR data usage for their analytics. For example, data discovery tools (e.g., federated catalogues) include cryptographic verification of authenticity, integrity, and data lineage.

2.1 Making data findable, including provisions for metadata

Persistent identifiers are at the core of new data-sharing technologies. The NextGen project leverages Self-Addressing identifiers (SAIDs) for privacy-preserving mechanisms to discover data. SAIDs are also used in the Multi-Modal Integration Object (MMIO) for cross-border/federated portability, which requires the identification of multi-modal datasets to be present at each site and ingestible despite the heterogeneity of the underlying data formats and structures.

The Pathfinder sites in the NextGen project are the source of rich metadata that NextGen will use to adhere to established FAIR principles. This metadata, which identifies data types and modalities as described in the Data Summary, ensures consistent and interoperable modality integration.

Findability through Controlled Vocabulary Terms and Semantic Attributes: the following standards are used in NextGen:

- DCAT (Data Catalog Vocabulary), Version 3 (W3C standard designed interoperability).
- Overlays Capture Architecture (OCA) for metadata architecture and universal semantic adaptor across data formats.

DCAT keywords enable users to locate datasets of interest more effectively by using search terms relevant to their needs or research areas. NextGen federated catalogue will be DCAT native but can also ingest other vocabularies and ontologies.

NextGen project harvest and index metadata using the following components:

- Serialisation using JSON, a widely recognised, machine-readable format, offers straightforward data manipulation and retrieval.
- Metadata structure using OCA. This standardisation ensures that metadata is consistent, detailed, and contextually rich, enhancing both its utility and discoverability.
- Semantic repositories to store structured metadata.

2.2 Making Data Accessible

2.2.1 Repositories

The metadata in the NextGen project is stored in OCA Repositories, a trusted repository for structural metadata and semantic data objects. The repository also supports the resolution of persistent identifiers to the corresponding digital objects. The OCA Repository is accessed via a REST API interface, operates under local governance, and can be configured by the organisation hosting it as public (open access) or private (limited access).

Data is stored with the NextGen participants or in data repositories (e.g., biobanks) where the participant has lawful access. Within NextGen, the participant's organisation retains full control of their data. They are required to meet any local requirements and comply with GDPR in addition to any local regulatory or compliance rules.

2.2.2 Data Access

Data generated by specific project outputs (such as research) may be made available where possible, subject to the appropriate logistical, regulatory and legal requirements. To the extent these requirements allow, NextGen aims to develop the Pathfinder Platform to facilitate standardised data access and analysis in a federated manner. NextGen considers compatibility with other EU initiatives like B1MG and GDI for potential integration in its design. Data Agreements are necessary for managing data access and usage. The NextGen project delivers a sustainability plan that will include provisions for data management beyond the project duration.

NextGen relies on each participating organisation to fulfil its local obligations in terms of data management. This includes robust identity and access management (IAM) processes in compliance with local regulations and NextGen grant agreements (e.g. GDPR compliance). Therefore, initially, NextGen does not require a data access committee.

Where data is shared between participants, the Regulation, Ethics and Governance Board plays a crucial role in ensuring, where applicable, that this is carried out appropriately, thereby upholding the project's ethical standards.

In general, it is important to reiterate that metadata will be openly available and licensed under a public domain dedication CC0, as per the Grant Agreement. This commitment to open data is a key aspect of the NextGen project. Nevertheless, metadata generated by specific project outputs (such as research) may only be made available where allowed by the appropriate logistical, regulatory and legal requirements including NextGen Grant Agreement.

To the extent allowed by these requirements, NextGen aims to develop the Pathfinder Platform to facilitate standardised metadata access and analysis in a federated manner. NextGen considers compatibility with other EU initiatives like B1MG and GDI for potential integration in its design.

NextGen develops open-source code for the core components. The NextGen project delivers a sustainability plan that includes provisions for data management beyond the project duration. Documentation will provide guidance on the tools and software necessary to use the data effectively.

2.3 Making Data Interoperable

The NextGen methodology approaches data integration from the perspective of overcoming the existing barriers to data usage. Therefore, interoperability has two dimensions: Technology and Governance. In other words, interoperability is not only about technologies facilitating data access; it is also about creating the right governance framework to ensure that the regulatory, compliance, economic, and ethical aspects are considered.

2.3.1 Data and metadata vocabularies, standards, formats or methodologies

NextGen ensures data interoperability to facilitate data exchange and reuse within and across disciplines. To achieve this, NextGen adheres to the following data and metadata vocabularies, standards, formats, and methodologies:

Vocabularies

- Data Privacy Vocabulary (DPV): The DPV standardises the description and management of data privacy aspects, ensuring consistent and clear communication of privacy-related information.
- Data Catalogue Vocabulary (DCAT) - Version 3.

Standard Schema Design

- Phenopackets v2.0: The Phenopacket schema enables the structured representation and exchange of phenotypic data.

Semantic Architecture

- Overlays Capture Architecture (OCA): The OCA is the semantic architecture that ensures semantic interoperability across the NextGen ecosystem. It will provide the necessary frameworks and standards to support data's consistent interpretation and integration.

Generic Objects for Interoperability

- Multimodal Integration Objects (MMIOs): MMIOs are digital object designed for multimodal data portability. They are agnostic to the numerous underlying standards present in source modalities. For example, an MMIO for a Cardiovascular Risk Prediction model would integrate cryptographic assured reference to imaging, phenotypic, wearable, and genomic data as well as embedding specific governance requirements for the usage of the dataset.

Commonly Used Ontologies and Controlled Terminologies: Project-specific ontologies, these include:

- Unified Code for Units of Measure (UCUM): UCUM is a code system encompassing all units of measure used in contemporary international science, engineering, and business. It provides a standardised approach to unit representation, ensuring consistency across diverse datasets.
- Logical Observation Identifiers Names and Codes (LOINC): LOINC is a database and universal standard for identifying medical laboratory observations. It ensures that laboratory and clinical observations are consistently and accurately represented, thus facilitating data sharing and interoperability.

- SNOMED CT (SNOMED Clinical Terms): SNOMED CT is a systematically organised collection of medical terms providing codes, terms, synonyms, and definitions used in clinical documentation and reporting. This controlled terminology enhances the clarity and precision of clinical data, supporting effective data exchange and analysis.

2.3.2 Project specific generation of uncommon ontologies or vocabularies

In the NextGen project, specific use case requirements may necessitate using uncommon or project-specific ontologies or vocabularies. Our approach to handling these situations includes the following commitments:

Mapping of Uncommonly Used Ontologies: When using uncommon or generating project-specific ontologies or vocabularies becomes unavoidable, we will provide mappings to more commonly used ontologies. This understanding ensures that our data remains interoperable with broader scientific and medical communities, facilitating seamless integration and exchange.

Open Publication of Ontologies and Vocabularies: NextGen will openly publish any generated project-specific ontologies or vocabularies. This transparency is not just about sharing but inviting others in the ecosystem to reuse, refine, or extend these resources. The openly available terminologies will enhance the interoperability and utility of data across various applications and research projects.

2.3.3 Project Output & FAIR impact

The following table shows the key project outputs materialising the FAIR principles

FAIR in project output	
Deliverable / KPI	FAIR Impact example
AI Models	Dataset reuse (Discovery, Authentication), Schema reuse (Integrity)
Publication of research results	Results traceability with persistent identifiers
Multimodal Integration tools	Adaptor functionality for mapping multiple standards. Re-use of data in different formats
Federated Catalogues	Data and metadata discovery across the NextGen platform
Pathfinder network	Re-use of tools and dataset across different locations *(to the extent of regulatory allowed)*

3 Increase Data Reuse: NextGen data documentation

Alongside the technical documentation and related user guides, the project will consistently issue regular reports on data usage in healthcare. These reports, issued at regular intervals, will provide an in-depth analysis of regulatory, ethical, governance, or economic aspects surrounding the data generated by NextGen.

The NextGen project, in its commitment to transparency and collaboration, also includes scientific publications of the results and strictly adheres to Open Science principles. The NextGen project also include scientific publications of the results and adheres to Open Science principles.

Will the data produced in the project be useable by third parties, in particular after the end of the project?
NextGen has a dedicated work package for communication and dissemination that will provide third parties visibility on the NextGen outcome and reusable output. In addition, a work package in NextGen focuses on the broader engagement

of stakeholders, including other EU initiatives and creating a stakeholder platform as a dedicated community space through which NextGen outputs (e.g. reports, blueprints, etc..) could be accessed. As these contacts evolve, NextGen will advertise and assess the possibility of data integration with third parties. NextGen will also produce a Sustainability plan.

4 Other Research outputs

From a data management perspective, the NextGen research has set out to deliver

Deploy multimodal integration tools enabling cross-site portability of research & development. Key Performance Indicators (KPI)

- Incorporation of multiple data formats in 5 multimodal artificial intelligence (AI) algorithms for improving cardiovascular health incorporating data format/governance requirements for 7 countries (SE,UK,CH,FI,USA,DE,NL).
- Whole genome sequence data for 2 use cases.
- 5 multimodal integration objects for specific use cases.

Develop a sustainable platform beyond the project duration. NextGen will develop a sustainability plan to ensure full access and further development after the project end. With the following KPI

- Platform developed across different stakeholder groups

All published research data will be open and available for shared use if agreements (concerning ownership, rights to use, IPR and non-disclosure), legislation and ethical principles allow it, to allow for reproducibility of research

Contribution to body of knowledge: Research carried out within NextGen will enrich the body of clinical knowledge and, although indirect, this is the most important route to the derivation of evidence-based changes to clinical practice.

Publications: NextGen is expected to have 20-50 open-access papers submitted by the end of the project. The papers will be a combination of opinion pieces, systematic reviews, and targeted scientific publications. Such papers will be focused the specific clinical outcomes of the use cases and on the relevance of NextGen tools.

Medical device candidates: NextGen has a specific task (Task T4.5) to identify project algorithms that can be translated into medical devices which is also identified as part of our exploitation plan (2.2.3.1).

Direct input into decision making bodies: Members of the Consortium (and the SSSHB) participate in clinical decision-making bodies allowing first-hand insight into ways to translate project outputs into clinical impact.

5 Allocation of resources

What will the costs be for making data or other research outputs FAIR in your project

The allocation of resource will be included in the DMP updates based on the documents on platform design (M18) and platform deployment (M36). In addition, a ”Health Economics plan” and a ”Sustainability and exploitation plan” are also produced.

6 Data Security

It is important to recall that NextGen focuses on federated mechanisms and decentralised data management techniques, which means that centrally aggregating the original data is outside the declared project scope. This decentralised approach is relevant for GDPR and data security as the processing will be done within the participant's secured environment. Data security initially remains within the participating organisations.

The data in NextGen will be jointly analysed using federated learning approaches, a method where data remains on the user's device and only model updates are shared. Implementing federated learning algorithms is a key aspect of our project, as they guarantee that the shared information does not compromise the security of personal data. These algorithms use differential privacy, multi-party computation, and homomorphic encryption techniques to ensure data security.

Federated learning will be developed initially on public, synthetic, or dummy data before the technologies have been properly risk-assessed through sequential pilot development.

NextGen develops a four-layer data security concept covering the different levels of the technology stack. The project includes a security plan, "Platform design and architecture blueprint" (deliverable D2.1 due M18), to document the data security framework.

1. **Architecture level**: This layer secures the architecture by making sure that the data never leaves the data owner's location while giving them the power to control their data usage.
2. **Infrastructure level**: The security layer at infrastructure level relies on Kubernetes. This provides a secure communication between the NextGen nodes using a private network.
3. **Governance level**: This layer secures NextGen governance represented by policies, contracts, data agreements. It provides secured authentication and integrity in the network..
4. **Applications**: This security layer supports the communication between the applications by not exposing the data.

The tools developed in NextGen and sequentially demonstrated and developed in pilots will, to the extent allowed by the regulatory, legal and ethical governance measures, will lead to a decentralised platform for an inter-site collaboration demonstrating clinically oriented utility of a network serving an ecosystem of several geographically diverse research institutions with separate clinical data and sample biobanks.

As a result, NextGen data security focuses on securing the tooling developed (i.e. accelerated genomic analytic pipeline, extend federated machine learning techniques, semantic interoperability) and a decentralised infrastructure for a decentralised network. As a result, the provisions for data security cover:

- Technology risks (data storage, transfer).
- Risks of new technologies for multi-modal data integration.
- Privacy risks.

A Regulatory, Ethics and Governance board is in place to ensure continuous oversight of the development of the risk framework from design to the implementation in pilots of governance measures.

7 Ethics

NextGen develops tools to facilitate data integration, including genomics in cardiovascular research. As a result, a risk-based approach will be taken when implementing the tools in specific pilots, as they may generate potential new privacy risks. Therefore, NextGen adopts a Privacy-by-Design-and-by-default approach to ensure ethical considerations are captured early in the design. In addition to the regulatory and compliance risk, the approach considers ethics early in the project's design phase.

Addressing ethics or legal issues that can have an impact on data sharing

NextGen regulatory and ethical oversight is done by the appointed REG Board. The REG Board must be consulted prior to any NextGen data sharing intent. The Board will also stay informed in the development of European Data Space Regulations and initiate contact with health regulators.

The Board relies on participating organisations to respect their regulations and ethics rules applicable in the NextGen Grant Agreement in their respective jurisdiction.

Informed consent for data sharing and long-term preservation of questionnaires dealing with personal data?

Centrally aggregating data is outside the declared project scope. NextGen takes a decentralised approach to data management and will not store Personal Data as these remain under the control of participating organisations.

Therefore, the management of consent for data usage will respect the local regulations of the participating organisation.

All NextGen research organisations have received a questionnaire including consent questions to map the NextGen regulatory landscape.

Reference to ethics in NextGen project Description of the actions (AoA)

NextGen DoA has the following references to ethics:

- **Project Reports:**
  D6.3 – Health economics report
  D6.4 – Responsible personalised medicine recommendations
  D6.5 – Governance framework and legal guidelines
- **Work Package 6,** "Regulatory, Ethics and Governance (REG)": Ensures continuous review of regulatory, ethical and governance topics raised by NextGEn developments.
- **Work Package 7,** "Broader Engagement": Enable contact with EHDS initiatives and build upon their REG framework where applicable.
- Ethics assessment provided in the DOA (Annex 1, Part B, Section 4).
- REG Board established.
- Workshop on Data Protection in M6 to ensure awareness of the privacy risk and applicable laws.

8 Data Integration Framework (DIF)

To capture the nuances of multi-modal data integration in multi-jurisdictional settings, NextGen adopts a "Data Oriented" approach in its design. This requires that the authenticity, integrity, security and lawfulness of data and metadata are ensured across their life cycle and independently of their location. The NextGen requirements and specifications will result in an internal (sensitive) document, the Data Integration Framework.

A public extract of the DIF document will be part of the updated version of the NextGen DMP (D1.2 due in M24)

9 Concluding Remarks and Next Steps

The NextGen project is a pioneering research and innovation initiative exploring data, tools, and solutions pertinent to personalised cardiovascular disease. Our research aims to navigate and contribute to an evolving scenario characterised by continuous advancements in medical and data sciences. Through this endeavour, we aim to foster significant knowledge build-up, which will drive the development of innovative solutions addressing various aspects of cardiovascular disease.

As we advance, we recognise that each solution may unearth new issues, sparking further research and solutions. This cycle of continuous improvement and discovery is at the heart of our project's philosophy, underscoring the dynamic nature of research and development. The sequential development of pilots derived from use cases with clinical impact ensures continuous production of output results.

The Data Management Plan (DMP) we have drafted is a living document that evolves in tandem with our research progress. It will regularly reflect new insights, methodologies, and technological advancements. This adaptive approach ensures that our data management strategies remain relevant and effective, supporting the overall objectives of the NextGen project.

We are committed to maintaining a robust and flexible data management framework that can accommodate the complexities and demands of cutting-edge cardiovascular research. As we move forward, collaboration with stakeholders and continual reassessment of our data management practices will be crucial. Together, we will strive to make meaningful contributions to the field of cardiovascular disease, ultimately improving patient outcomes and advancing medical knowledge towards more effective and personalised medicine.