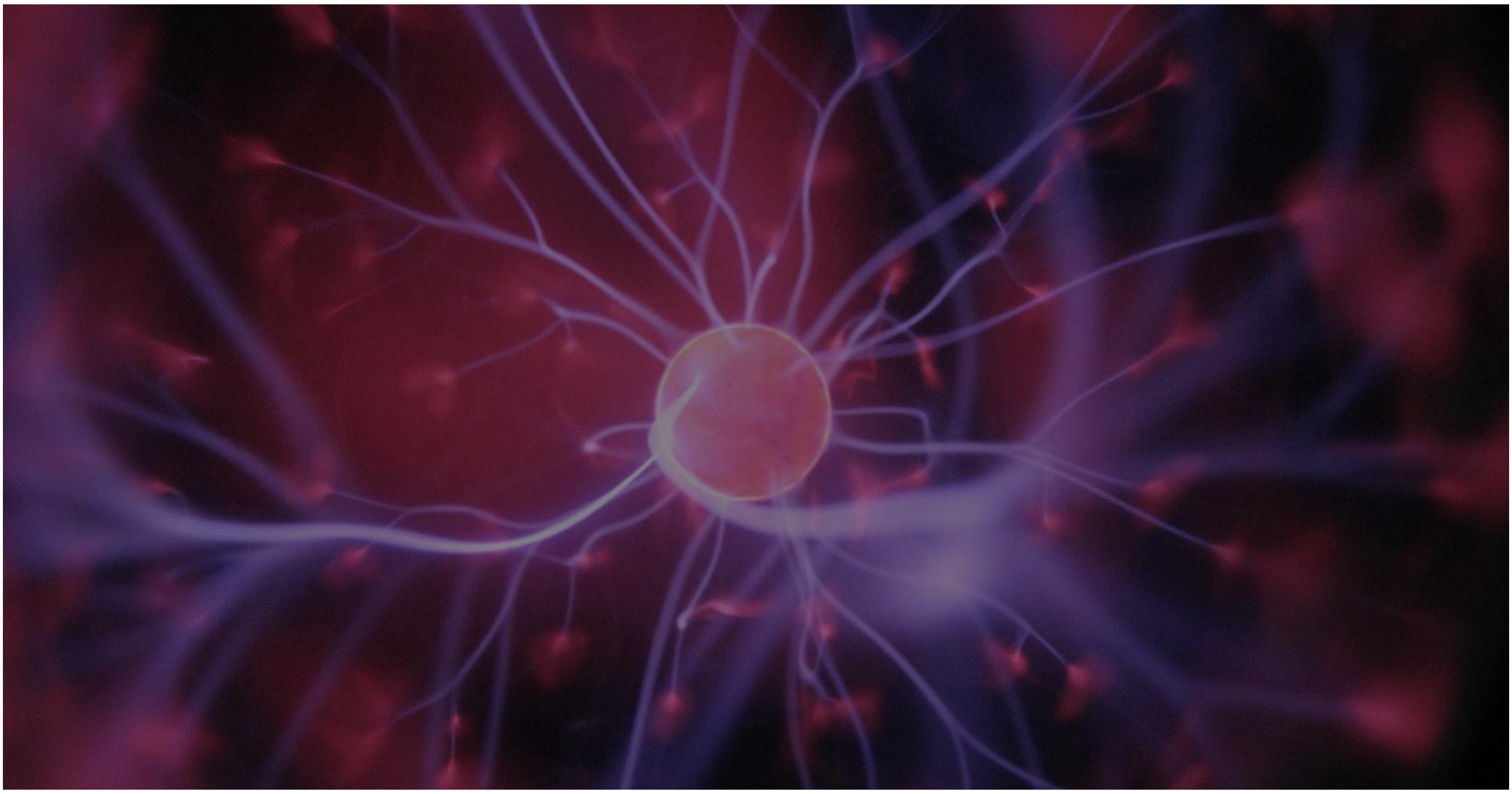
NextGen

Genome-Centric Multimodal Data Integration in Personalised Cardiovascular Medicine.



Deliverable 5.2
Pathfinder functionality requirements

Grant Agreement Number: 101136962

Funded by the European Union

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]

D5.2 – Pathfinder functional requirements Report	
Author(s)	Sander W. van der Laan, Aaron M. Lee, Francesca Mangili, Philippe Page, Fred Buining, Rafy Mehany, Raja Appuswamy, Jessica van Setten
Editor	Sander W. van der Laan
Participating partners	UMCU, QMUL, SUPSI, HIRO, HCF, EURECOM
Version	1.0
Status	Final
Deliverable date	M12
Dissemination Level	PU - Public
Official date	2024-12-31
Actual date	2024-12-16

Disclaimer

This document contains material, which is the copyright of certain NextGen contractors, and may not be reproduced or copied without permission. All NextGen consortium partners have agreed to the full publication of this document if not declared "Confidential". The commercial use of any information contained in this document may require a licence from the proprietor of that information. The reproduction of this document or of parts of it requires an agreement with the proprietor of that information., according to the provisions of the Grant Agreement nr. 101136962 and the Consortium Agreement.

Furthermore, a disclaimer will be included, indicating that: "Any communication or publication related to the action, whether produced collectively by the beneficiaries or individually in any format and through any medium, represents solely the views of the author, and the Commission bears no responsibility for any use that may be made of the information contained therein."

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
16/12/2024	1.0	Pathfinder functional requirements	UMCU

Authors

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

Reviewers

REVIEWER	ORGANISATION

List of terms and abbreviations

ABBREVIATION	DESCRIPTION
AI	Artificial intelligence
EDHS	European Health DataSpace
FL	Federated (machine) learning
GWAS	Genome-wide association study
KPI	Key Performance Indicator
ML	Machine learning
MMIO	Multimodal integration object
MVP	Minimal viable product
PGS	Polygenic (risk) score
WP	Work Package

Table of contents

1 Executive Summary

The goal of the NextGen Pathfinder is to create a small-scale-European Health Data Space (EHDS), namely a health specific ecosystem 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.

The purpose of this document is to describe the functional requirements, that is the things that are needed from the point-of-view of the users, that should be in the Pathfinder platform.

Note that these functional requirements may need further detailing in the future as we tackle each through the NextGen deliverables in the next years to come. Thus, like the Blueprint, this document is not set in stone, rather a 'living' document amenable to change due to continuing practical insights and ongoing discussion. Thus, like the Blueprint, mapping the Pathfinder functional requirements will be an iterative process.

2 Functional Requirements

This chapter describes the functional requirements in general and for specific subjects. For most, the NextGen Platform should align with (local) rules, regulations, and laws (protecting privacy and be sufficiently secure), as well as the EHDS (see European Health Data Space requirements), and finally must be user-friendly (build with the user in mind).

By using these functional requirements stipulated for the EHDS as guidance, the NextGen Platform aims to create a cohesive and secure environment that would align well with health research data across Europe in the (near) future, empowering different stakeholders (individual researchers, clinicians, and patients), and fostering innovation in healthcare research.

1.1 General requirements

The NextGen **Platform general requirements** were mapped, and the following are listed with additional detail provided in subsequent sections where required:

- **Requirements for MVP and development platform at TRL 3/4, that is experimental proof of concept and validation in a lab environment**:
    - **Federated catalogue functionality**: Enables the creation and management of distributed data catalogues, allowing users to find and access datasets across a federated system.
    - **Multimodal integration object functionality**: Provides capabilities to integrate and process diverse data types (for example, clinical, imaging, genomic) into unified analytical frameworks.
    - **Federated machine learning functionality**: Facilitates machine learning across distributed datasets without requiring central data pooling, ensuring privacy and compliance.
    - **Federated genomics functionality**: Supports distributed analysis of genomic data, ensuring data remains securely stored within its original location.
    - **Accelerated genomics functionality**: Optimizes the speed and scalability of genomic data processing and analysis for timely insights.
    - **Variant prioritization functionality**: Enables the identification and ranking of genetic variants based on their potential clinical or research relevance.
    - **Genomic data curation and analysis functionality**: Provides tools for annotating, curating, and analyzing genomic data to derive actionable insights.
    - **AI/ML model management functionality**: Includes tools for the lifecycle management of AI and machine learning models, from development to deployment and monitoring.
    - **Data-oriented architecture functionality**: Leverages a data-centric approach to structure and manage data for enhanced scalability, interoperability, and usability.
    - **Marketplace functionality**: Establishes a platform for users to discover, exchange, and utilize datasets, algorithms, and tools within a secure ecosystem.
    - **Security functionality**: Ensures robust security measures, including data encryption, access controls, and compliance with regulatory frameworks.
    - **Governance functionality**: Provides mechanisms for ensuring ethical, legal, and policy compliance across federated and collaborative data ecosystems.

    - **European Health Data Space core functionality**: Aligns with EU initiatives to enable secure, cross-border access and use of health data for research, diagnosis, and policymaking (see also European Health Data Space requirements).

Non-functional requirements: Addresses performance, reliability, scalability, and usability considerations to ensure the system meets user and operational expectations. See also section Non-Functional requirements.

1.2 Federated catalogues

NextGen semantic technology allows the true abstraction of meaning from format. **NextGen** cataloguing supports discoverability of complex multimodal datasets within data ecosystems supporting specific research questions and allowing integration into federated computational pipelines.

Overarching requirement: compatible and available datasets are visible/discoverable in a decentralized architecture with discovery supported for all medical data elements and modalities as and when required in **NextGen** use cases and pilots.

Specific requirements include:

- **Research stakeholder perspective:**
    - **Discovery supported for all medical data elements** (non-exhaustive examples provided in Appendix A: Example Cardiovascular Variables) as and when required in **NextGen** use cases and pilots.
    - **Discovery supported for all medical data modalities** (non-exhaustive examples provided in Appendix B: Example cardiovascular non-tabular modalities, typical formats and associated metadata) as and when required in **NextGen** use cases and pilots.
    - **Discovery of standard demographic descriptors** supported (age, sex, ethnicity, etc.), including a rudimentary form of summary statistics (minimal, maximum, range, average, median) similar to the UK Biobank Showcase (reference: https://biobank.ndph.ox.ac.uk/showcase/field.cgi?id=21022) per available dataset. *Note that this is not meant to be for the purpose of 'live tracking'*, for example by tracking circulating cholesterol levels in during a clinical trial and report these values back to a clinician or researcher. The cohorts included in and designing the NextGen Pathfinder use-cases are *retrospective* cohorts for research; non are clinical trials.
    - **Discovery of multimodal datasets** (specified as a combination of any of the above).
    - **Discovery mechanisms implemented** in a user-friendly and accessible manner. Follow examples such as the UK Biobank Showcase (https://biobank.ndph.ox.ac.uk/showcase/).
- **Other stakeholder requirements:**
    - **Discovery of dataset governance**, that is applicable rules, regulations, laws, restrictions. See also Governance requirements.

1.3 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:

1. conversion of the underlying constellation of multimodal formats into a user-defined harmonized AI/ML ready form,
2. application of site-specific governance and regulatory requirements, and
3. 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.

Overarching requirement: harmonization of multi-omic multimodal datasets with different underlying formats with embedded governance and provenance functionality.

Specific requirements include:

- **Research stakeholder perspective:**
  - **MMIO** adapter functionality **allows harmonization** of a multi-omic multimodal dataset by converting the underlying heterogenous constellation of data into a user-defined AI/ML ready form (as per definition in the Grant Agreement).
  - **MMIO** adaptor functionality **can be used in `Python`** (a common development environment for AI/ML algorithms).
  - **MMIO** adapter functionality can **support a minimally agreed subset of healthcare data** modalities each of which in a specified range of underlying formats (as relevant to cardiovascular medicine) to include standard tabular data, medical images, electrocardiograms and specific genomic data types (as per Appendix B: Example cardiovascular non-tabular modalities, typical formats and associated metadata).
  - **MMIO** adapter functionality to be **sufficient to enable demonstration** of relevant MMIO pilots.
- **Other stakeholder requirements:**
  - MMIO functionality to meet required key performance indicators (KPIs).

1.4 Federated machine learning

Federated machine learning (FL) 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.

Overarching requirement: design and implementation of a (privacy) secure, auditable, and scalable federated learning architecture.

Specific requirements include:

- **Research stakeholder perspective:**
    - Functionality **allows generation of virtual federated networks** (for testing and design purposes).
    - Functionality **allows getting or generation of non-identifiable datasets**, that is obtaining through downloads or uploads of 'real-world' datasets with the appropriate governance, or the generation of 'dummy' or 'fake' datasets, for the purpose testing, design and demonstration.
    - **Common ML development frameworks** (such as `PyTorch` and `TensorFlow`) and paradigms are supported.
    - **Ease of transition of code** from a non-federated to a federated paradigm.
    - **Ease of modification of core code** to add NextGen specific functionality, for example for specific (future) use-cases.
    - **All common ML architectures** are supported (including generative adversarial, transformer, and diffusion networks, among others).
- **Technical stakeholder perspective**: The federated learning framework selected must guarantee the following.
    - **Ease of Use.**
        - The framework should provide intuitive APIs, clear documentation, and user-friendly interfaces to streamline the setup, configuration, and deployment of federated learning models.
        - Tools for monitoring and managing federated learning workflows should be easily accessible.
    - **Privacy Features.** The framework must incorporate advanced privacy-preserving technologies, such as:
        - Differential Privacy to protect individual data contributions.
        - Secure Multi-party Computation (SMPC) or Homomorphic Encryption to ensure computations on encrypted data.
        - Federated Averaging mechanisms that allow aggregation without exposing individual datasets.
        - It goes without saying that compliance with GDPR and EHDS-specific data protection requirements is critical.
    - **Scalability.**
        - The framework must handle a large number of distributed nodes (that is multiple sites: hospitals or research centers) without significant degradation in performance.
        - Support for varying data sizes, compute capabilities, and network conditions at different nodes.

        - Scalability across both hardware (edge devices) and software layers.
    - **Community Support.**
        - Active and engaged community for troubleshooting, sharing best practices, and collaboration.
        - Availability of plug-ins, extensions, or integrations developed by the community, for example, through an API with GitHub or alike.
        - Open-source options with regular updates and maintenance.
    - **Production Readiness.**
        - Ability to transition models seamlessly from development to production environments.
        - Support for robust deployment tools, such as containerization, for example, through Docker.
        - Reliability in real-world healthcare settings with high availability and fault tolerance.
    - **Customizability.**
        - Flexibility to adjust algorithms, protocols, and configurations to fit specific healthcare use cases.
        - Modular architecture to allow integration with custom workflows, privacy mechanisms, or data formats.
    - **Learning Curve.**
        - Short learning curve to enable rapid adoption by IT teams, researchers, and data scientists.
        - Facilitate learning through for example training materials, tutorials, or clarifying examples.
        - Ability to integrate with widely used tools in machine learning, such as `TensorFlow`, `PyTorch`, or `Scikit-learn`.
- Other stakeholder requirements:
    - None.

1.5 Federated genomic data curation and analysis

Federated machine learning is now an established technique. However, the distributed computation of genome-specific algorithms (for example, 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 (for example PGS) to privacy-preserving federated settings such as secure multiparty computation.

Data curation, that is quality control (QC), in the genomics space requires a complex, integrated processing pipeline. Genomic association studies may encompass millions of genomic variants and thousands of genes, and stringent quality control is obligatory to establish reliable variant- or 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 variant- or 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: Insufficient local data volume and variety in context of genomics.

Overarching requirements: Management of multimodal datasets with different underlying formats with embedded integration of governance and provenance functionality. Facilitation of distributed genomic curation and analysis, that is: data quality control, differential gene expression, genome-wide association studies, polygenic risk scores creation and analysis, clustering, and dimensionality reduction.

Specific requirements include:

- **Research stakeholder perspective:**
    - **Federated tools for clustering and dimensionality reduction.** These tools are essential for the quality control (QC) and preprocessing phases of genome-wide association studies (GWAS), as well as bulk or single-cell RNA sequencing data analysis. In a federated setup, these tools must enable analysis across distributed data sources while preserving data privacy.
        - Privacy-preserving clustering:
            - Federated implementations of clustering algorithms such as k-means, hierarchical clustering, or density-based clustering with secure aggregation methods.
        - Federated dimensionality reduction:
            - Algorithms like PCA (Principal Component Analysis) or t-SNE (t-distributed Stochastic Neighbor Embedding) implemented with privacy-preserving protocols.
            - Support for linear and non-linear methods for reducing high-dimensional genomic data without requiring direct access to raw data.
        - 'Real-time' QC:
            - Federated tools for detecting outliers or batch effects across datasets in distributed nodes.
            - Static or interactive visualization of clustering or dimensionality reduction outputs to facilitate decision-making.
        - Compatibility:
            - Support for various genomic and single-cell RNA sequencing data formats, ensuring easy integration into existing workflows.
    - **Federated GWAS with a standardized and federated QC pipeline.** Federated GWAS enables large-scale genetic association studies by analyzing distributed datasets without transferring sensitive genomic data. A standardized QC pipeline ensures consistent preprocessing and high-quality data across all nodes.
        - Federated QC Pipeline:
            - Standardized preprocessing steps for GWAS, including data cleaning, normalization, and SNP filtering.
            - Implementation of QC metrics like Hardy-Weinberg equilibrium, minor allele frequency (MAF), and genotyping call rates, executed across distributed datasets without exposing raw data.
            - Privacy-preserving algorithms for removing low-quality variants or samples without centralizing sensitive information.
            - Preparation for downstream, on-site and node-specific imputation. *To be clear: imputation shall not be done through the NextGen platform.* Due to practical matters this shall be the prerogative of each node and

        must be done locally. However, the NextGen platform should efficiently facilitate this by providing the above tools to build this QC pipeline.

    - Scalable GWAS Framework:
        - Support for large-scale genomic data from thousands of participants across multiple sites, with efficient data encryption and aggregation protocols.
        - Distributed linear and logistic regression models tailored for GWAS, ensuring scalability and statistical robustness.
    - Interoperability:
        - Compatibility with existing GWAS software (for example, PLINK, Regenie) in a federated setup.
        - Integration with c.q. application of standardized ontologies (see Federated catalogues) and databases (for example, dbSNP, Ensembl).
    - Secure Aggregation:
        - Use of privacy-preserving technologies like secure multi-party computation (SMPC) or federated averaging to aggregate GWAS results, such as p-values or beta coefficients and other relevant statistics.
    - Transparency and Reproducibility:
        - Federated audit trails for QC steps and GWAS analysis, ensuring results can be traced back to their origins.
- **Federated machine learning for training pathogenic variant prioritization models or ML-based polygenic risk scores (PGS).**
    - Model Training in a Federated Environment:
        - Support for deep learning and other ML frameworks (for example, TensorFlow Federated, PySyft) to train pathogenic variant prioritization models and PGS models across multiple institutions.
        - Federated learning algorithms that handle class imbalance and sparse data typical of rare variant datasets.
    - Pathogenic Variant Prioritization:
        - Privacy-preserving feature selection methods for genomic annotations, allele frequencies, conservation scores, and functional impact predictions.
        - Training models (for example, random forests, gradient boosting, or neural networks) to classify variants as benign, likely pathogenic, or pathogenic based on distributed datasets.
    - Federated PGS Calculation:
        - Federated frameworks for PGS computation using genome-wide SNP data while ensuring secure exchange of summary statistics rather than raw genotypes.
        - Inclusion of stratified PGS models to address population-specific variations.
    - Advanced Privacy Mechanisms:
        - Use of secure aggregation for model weights, differential privacy for protecting individual contributions, and encryption for model parameter sharing.

      - Local validation of trained models to ensure no leakage of sensitive data during testing phases.
    - Scalability:
      - Handling large-scale datasets involving millions of variants and thousands of individuals, while minimizing computational and communication overhead.
    - Model Deployment:
      - Deployment-ready tools for using federated-trained models in real-world clinical or research settings.
      - Continuous learning capabilities to update models dynamically as new data becomes available at different nodes.
- Other stakeholder requirements:
  - None.

1.6 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.

Overarching requirement: Provision of mechanisms to overcome bottlenecks in secondary/tertiary genomic data processing.

Specific requirements include:

- **Research stakeholder perspective:**
  - **Novel algorithms for scaling of the GATK secondary analysis pipeline**, in particular the post-alignment data preparation stages, for example base quality score recalibration, variant calling, and read deduplication.
  - **Novel cross-architecture implementation of accelerated pipeline** that can exploit XPU (CPU and GPU), thus leveraging heterogeneous computing architectures to accelerate genomic data processing pipelines.
    - Focus on highly parallelizable tasks like sequence alignment and matrix operations.
    - Use portable frameworks to enable the development of pipelines that adapt dynamically to the underlying hardware.
    - Introduce intelligent scheduling to ensure maximum resource utilization.
  - **Data structure and I/O optimizations** to *Exomiser* that can reduce memory footprint and enable fast querying of internal variant annotation databases.
    - Use memory-efficient data structures and explore columnar storage formats (for example, `Parquet` or `Arrow`).

        - Leverage high-speed storage solutions, for example through caching, while preserving privacy and security.
    - **Incorporation of benchmarking capabilities** to assess the performance and efficiency of data analysis processes within the dataspace environment.
        - For example, include metrics for runtime, memory usage, I/O throughput, and accuracy of results.
        - Add static or interactive visualizations of (intermediate) results through a dashboard.
        - Implement benchmarking capabilities while preserving privacy and security through aggregation of performance data.
- Other stakeholder requirements:
    - None.

1.7 Variant prioritization

Genomic sequencing identifies variations in the genetic code and its validity occurs with variant annotation. 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. Furthermore, the classification of a variant (mutation) into benign or pathogenic categories (variant assertion) is vital to determine clinical actionability. 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.

Overarching requirement: Provision of functionalities for efficient ranking of variants by significance.

Specific requirements include:

- **Research stakeholder perspective:**
    - **Offer tools for off-platform post-GWAS analysis** such as fine mapping, colocalization analysis, functional genomics, functional enrichment, or others.
        - **Fine Mapping**: Provide input for tools to pinpoint causal variants with high confidence from significant GWAS loci.
        - **Colocalization Analysis**: Provide input to enable statistical methods to identify overlap between GWAS signals and eQTLs (expression quantitative trait loci) to infer shared genetic architecture.
        - **Functional Genomics**: Incorporate tools to enable integrate multi-omics data (e.g., RNA-seq, ATAC-seq, ChIP-seq) to annotate and prioritize variants based on biological relevance.
        - **Functional Enrichment Analysis**: Provide input for methods to identify overrepresented biological pathways or gene sets associated with prioritized variants (e.g., GO terms, KEGG pathways).

- **Improve models for pathogenic variant prioritization.** Develop machine learning models to predict which variants are likely to be pathogenic, improving on current scoring systems (for example, CADD, REVEL).
    - **Multi-Feature Integration**: Models should integrate diverse evidence types, including allele frequency, conservation scores, functional annotations, and experimental data.
    - **Rare Variant Prioritization**: Address limitations in analyzing rare variants, which are often implicated in monogenic diseases.
    - **Incorporation of Clinical Data**: Leverage phenotypic and clinical data (e.g., Human Phenotype Ontology terms) to improve variant classification accuracy.
    - **Scalability**: Ensure the models scale efficiently to analyze large variant datasets in real-time.
- **Provide explanation for ML-based pathogenic variant prioritization scores.** Provide interpretability for machine learning models to improve trust and usability for clinical scientists.
    - **Score Explanation**: Accompany ML-based prioritization scores with explanations highlighting which features (e.g., conservation, functional effect) contributed to the score.
    - **Visualization of Evidence**: Display evidence that influenced the prioritization, such as annotated variant features or pathogenicity predictions from different algorithms.
    - **Transparency**: Ensure model decisions can be validated through visual and tabular outputs, fostering user confidence.
    - **Customizable Parameters**: Allow researchers to adjust model parameters or weighting for specific evidence sources to fine-tune results.

- **Other stakeholder requirements:**
    - **Determine penetrance and phenotypic expression of the given variants.** Assess the likelihood that a specific genetic variant will lead to a clinical phenotype (penetrance) and characterize its phenotypic variability.
        - **Penetrance Estimation**: Use statistical and ML-based methods to estimate variant penetrance based on population studies, family-based data, and phenotype correlations.
        - **Phenotype Expression Analysis**: Link variants to observed phenotypic variability by integrating clinical records, patient registries, and phenotype databases (for example, ClinVar, OMIM).
        - **Longitudinal Data Integration**: Incorporate time-series or longitudinal datasets to assess age-dependent penetrance and disease progression patterns.
        - **Interactive Tools**: Provide visualization of penetrance estimates and phenotypic distributions for easier interpretation by clinical scientists.

1.8 European Health Data Space requirements

The EHDS is a data-sharing framework for health data in the European Union, aiming to give EU citizens better control over their health data. The NextGen "Broader Engagement & Exploitation" work package (WP7) specifically seeks to ensure project deliverables are synergistic with the EHDS and the NextGen Pathfinder will demonstrate EHDS functional specifications.

The EHDS is designed to enhance the accessibility, interoperability, and security of health data across the European Union. Its functional requirements encompass several key areas:

1. **Interoperability**: The EHDS must be capable of interacting seamlessly with various software applications and devices, regardless of the manufacturer. This includes the ability to transfer and receive personal electronic health data in a standardized, machine-readable format.
2. **Logging and Security**: The EHDS is required to maintain detailed logs of data access, capturing information such as the identity of healthcare providers accessing the data, categories of data accessed, and timestamps of access. This ensures transparency and accountability in data handling.
3. **Data Access for Secondary Use**: The EHDS establishes a framework that allows entities like researchers and policymakers to access health data for secondary purposes, including research and innovation. Applicants must justify the necessity for data access and demonstrate compliance with data protection regulations.
4. **Governance and Infrastructure**: The EHDS introduces a governance system at both national and EU levels. Additionally, cross-border digital infrastructures will be established to facilitate data sharing for both primary and secondary uses of health data.
5. **Certification and Compliance**: Manufacturers must ensure their products conform to essential requirements outlined in the EHDS, including interoperability, security, and logging capabilities. They are also obligated to provide technical documentation, accompany their systems with information sheets and user instructions, and affix a CE marking to indicate compliance.

Overarching requirement: The NextGen Pathfinder will demonstrate the core functionality and principles of the EHDS.

Specific requirements include:

- **Research stakeholder perspective:**
  - **Ensure matching with overall design and functional requirements of EHDS** (see above). Align research tools and workflows with EHDS interoperability, scalability, and data privacy standards to ensure seamless integration and future use.
- **Other stakeholder requirements:**
  - **HealthDCAT-AP / HealthDCCAT Editor (http://fair.healthdataportal.eu/editor2/):** Utilize the HealthDCAT-AP metadata editor to ensure FAIR data principles are met for datasets shared within the EHDS ecosystem.
  - **Article 56 Exploring Bias:** Implement mechanisms to analyze and mitigate algorithmic and dataset biases in accordance with Article 56 of the EHDS framework to ensure fairness and reliability.
  - **Quality and Utility Label:** Introduce quality and utility labels for datasets, enabling users to assess their relevance, accuracy, and compliance with EHDS standards.
  - **15 Data Categories in EHDS:** Support and structure data handling according to the EU-defined codification of 15 healthcare-related data categories to ensure compliance and interoperability.

- **Extensions for the NextGen Multimodal Universe:** Develop EHDS-compatible extensions that integrate genomic, clinical, and multimodal data types for advanced analytics in NextGen workflows.
-

1.9 Data oriented architecture requirements

We outline the data-oriented architectural principles and functionalities that form the foundation of the NextGen Platform, ensuring efficient, scalable, and interoperable data management. It highlights how data outputs are structured for downstream use, supports multi-site collaboration with federated computation, and incorporates automated orchestration for workflows like model training and infrastructure management. The inclusion of roles (for example, the 'data owner', 'model user' [that is the researcher]), data cleaning, imputation, and meta-analysis tools addresses challenges like missing data. The platform produces data products that meet analytical standards but does not act as a long-term repository. Transformational tools (e.g., MMIOs) standardize multimodal datasets for cross-site portability, while robust monitoring and observability ensure data quality and traceability. Additionally, integration with Data Hubs/Lakes/Lakehouses enables query execution, caching, and performance optimization, facilitating seamless data access and utilization across the NextGen ecosystem.

- **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 federated computation.
- **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 where possible for data imputation, transfer learning, and meta-analysis to overcome missing or lack of data. Note that in this case, we refer to imputation of for instance clinical variables as described in Appendix A: Example Cardiovascular Variables through packages/methods such as `MICE` (https://cran.r-project.org/web/packages/miceRanger/vignettes/miceAlgorithm.html).
- **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 **multimodal 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 (see** Appendix B: Example cardiovascular non-tabular modalities, typical formats and associated metadata**).**
- **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

1.10 Model management requirements

This section outlines the requirements for managing **AI/ML models** within the NextGen Platform, ensuring secure, privacy-aware, and scalable workflows for model development, validation, and deployment. It emphasizes compliance with **GDPR** and addresses key risks like privacy leakage and security during federated model training and sharing. Privacy-aware methods (for example, differential privacy) quantify and mitigate risks to data privacy, while **testing tools** allow synthetic data-based proof-of-concept demonstrations to partners (see also Federated machine learning). The platform incorporates a **Training Builder** for creating secure, end-to-end training and prediction pipelines and a **Trainer** for executing model training in GDPR-compliant environments. Tools for **federated machine learning orchestration** enable cross-site collaboration without centralizing data. Additionally, the platform provides a **Model Validator** to annotate models and datasets for reuse by researchers and a **Model Deployer** to integrate validated models into end-user infrastructures, such as healthcare systems, ensuring practical application and real-world impact.

The Nextgen **Platform AI/ML 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.
- **Testing tools**: Create the platform such that it can be deployed virtually to demonstrate and test using 'fake', synthetic data, to our partners and stakeholders as a proof-of-concept.
- **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.
- **Orchestration of federated Machine Learning** (as described in Federated machine learning).
- **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, that is: the researcher, IT professional, or clinician.

1.11 Marketplace requirements

The **Marketplace** is a core component of the NextGen Platform, serving as a virtual ecosystem where stakeholders can interact to exchange data, tools, and services. It facilitates seamless collaboration and efficient resource utilization, advancing research and innovation in genomic and clinical domains.

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.
  - Catalogue 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**.

Specific Requirements:

- **Virtual Meeting Place for Data and Services:**
  - The **Marketplace** functions as a centralized hub for connecting stakeholders, enabling the exchange of datasets, analytical services, and computational resources, thereby streamlining the supply and demand dynamics in the dataspace.
  - A **visitor recommender system**, leveraging stakeholder interaction data, provides personalized recommendations to enhance user engagement and collaboration opportunities.
- **Comprehensive Searchable Catalogue:**
  - The Marketplace must include a **searchable catalogue** of all data, services, and assets available within the NextGen dataspace and other interoperable dataspaces.
  - Search functionalities should be driven by metadata indexed in the **MMIO**, see below and Glossary) and augmented by metadata standards from external dataspaces to ensure broad accessibility and interoperability.
  - A seamless integration mechanism connects the catalogue with the **Marketplace** for real-time updates and efficient stakeholder navigation.
  - See additional requirements in Federated catalogues.

These requirements enable the **Marketplace** to act as an intuitive, interoperable platform for fostering collaboration and ensuring the efficient flow of data and resources within the NextGen Platform.

1.12 Security requirements

Security, authorization, and policies are critical components of the NextGen Platform to ensure data integrity, privacy, and trustworthiness. These measures enable seamless data exchange and collaboration within the platform while complying with ethical and legal frameworks such as GDPR and ensuring the secure use of AI/ML-ready datasets.

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** (see above) 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.
  - **Trusted Data Agent (TDA)**. The TDA acts as an independent, decentralized authentication mechanism for verifying MMIOs. Functionality:
    - Enables authentication without relying on third-party services, enhancing security and reducing vulnerabilities in the authentication process.
    - Facilitates trust among stakeholders by ensuring the authenticity of data and metadata shared across the dataspace catalog and **Marketplace**.
  - **MMIO**. The MMIO ensures secure, consistent, and trustworthy integration of diverse data assets into AI/ML-ready formats. Functionality:
    - **Authentication:** Unique identification of data, services, and infrastructure using cryptographic hashes (digital fingerprints) to guarantee the integrity and authenticity of objects, records, policies, and agreements.
    - **Harmonization:** Converts heterogeneous multimodal datasets into standardized formats that meet the requirements of AI and ML workflows (e.g., “Trustworthy AI”). For example, makes sure that variables as listed in Appendix A: Example Cardiovascular Variables are harmonized across the (selected) datasets in a given use-case.
    - **Data Governance:** Machine-actionable governance mechanisms allow for enforcement of policies such as consent, access restrictions, and data-sharing agreements at the edge, ensuring compliance with ethical and legal standards.
    - **Data Mobility without Data Movement:** Facilitates data access and processing by transferring MMIOs (metadata, schemas, and governance

        - components) instead of the actual data, ensuring data localization and reducing risks associated with data transfers.
        - **Processing at the Edge:** Transient integration of MMIOs allows data processing closer to the source, improving performance and minimizing data exposure.

1.13 Governance requirements

Governance (see also General requirements and Internal Report on "MS5: Overview of regulatory and legal concerns for all tools") ensures the **accountability, transparency, and trustworthiness** of operations within the NextGen Platform. It defines the structures, processes, and rules that govern the secure, ethical, and compliant use of data, services, and infrastructure across the platform.

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. Key components:
    - **Electronic Consent Management**: Ensures secure, transparent, and dynamic management of patient/research participant consent for data usage, aligned with legal and ethical requirements.
    - **Data Provenance Tracking**: Tracks the origin, modifications, and usage of data throughout its lifecycle, providing full traceability and accountability.
    - Data Ecosystem Governance Server (DEGS):
        - A server equipped with a hardware root of trust to maintain the authoritative, real-time configuration of the distributed data ecosystem.
        - Provides a global overview of data assets, their distribution, and governance status, ensuring transparency and consistency across stakeholders.
- **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** (see below) gets unlocked and the data service is consumed. Key components:
    - **Transaction and Contract Management**: Maintains immutable records of transactions, participants, and references to legal contracts in a **verifiable** and **non-repudiable** form.
    - **Billable Data Usage Settlement**: Settles financial terms for data access or usage **before unlocking the escrow locker**, ensuring fairness and compliance.
    - **Scientific and Ethical Oversight**: A dedicated body assesses proposed research for ethical, legal, and scientific risks, ensuring responsible data usage.
- **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. Key components:
    - **Assets** (data, resources, or services) are held temporarily and released only after all agreement conditions are satisfied.
    - **Enhances trust between parties** in a transaction by mitigating risks associated with data misuse or incomplete agreements.

1.14 Non-Functional requirements

The NextGen Platform **non-functional requirements** define the **operational characteristics** of the platform, ensuring it meets performance, scalability, security, and compliance standards. These are vital for both development (TRL 3/4) and production phases of the NextGen Platform. The non-functional requirements were mapped, and the following are listed.

- **Requirements for MVP and development platform at TRL 3/4**:
  - Defined **Risk Management Framework** (NIST Profiles and Tier):
    - Adopts the **NIST Cybersecurity Framework** to define clear risk management strategies and security baselines appropriate for development stages.
  - Scalability of Data Management Resources:
    - Ensures that the platform can handle increasing data volumes and computational demands efficiently as datasets and services scale during testing and development.
  - Energy Efficiency (as per KPI):
    - Optimizes resource utilization to minimize energy consumption, meeting predefined Key Performance Indicators (KPI) for sustainable operations.
- **Additional requirements** for production systems (outside of NextGen TRL 3/4):
  - **GDPR compliant** access to sensitive personal data, including clinical and genomic data.
    - Guarantees compliance with GDPR by implementing secure, restricted access to sensitive personal data, including clinical and genomic information.
  - **Data Protection:** schedule snapshots and backups of data flowing through the NextGen Platform:
    - Schedules regular snapshots and backups of data flowing through the platform to ensure recovery, continuity, and protection against data loss or corruption.

3 Appendix A: Example Cardiovascular Variables

1.1 Socio-demographics

- Age, years
- Sex (Male/Female)
- Gender
- Ethnicity
- Townsend deprivation index
- Educational level

1.2 Comorbidities

- Previous myocardial infarction
- Stroke
- Chronic obstructive pulmonary disease
- Asthma
- Atrial fibrillation
- Peripheral artery disease
- Hypertension
- Diabetes
- Hypercholesterolemia
- Chronic kidney disease

1.3 Physical measurements

- Height, m
- Waist hip ratio
- Waist height ratio
- Fat mass Index, kg/m2
- Fat free mass index, kg/m2
- DBP, mmHg
- SBP, mmHg
- Pulse rate, bpm

1.4 Lifestyle habits

- Smoking history
- Alcohol intake, frequency
- Time watching TV, hours/day
- Time using computer, hours/day
- Sleep duration, hours/day
- Physical activity -IPAQ score
- Dietary habits

1.5 Laboratory values

- Calcium, mmol/L
- Creatinine, umol/L
- Urea, mmol/L
- Urate, umol/L
- eGFR, mL/min/1.73m^2
- Aspartate aminotransferase, U/L
- Alanine aminotransferase, U/L
- Alkaline phosphatase, U/L
- Gamma glutamyl transferase, U/L
- Albumin, g/L
- Total bilirubin, umol/L
- White blood cell count, 10^9 cells/L
- Platelet count, 10^9 cells/L
- Mean corpuscular haemoglobin, pg
- Mean corpuscular haemoglobin concentration, g/dl
- Haemoglobin, g/dl
- Red blood cell distribution width, %
- HDL cholesterol, mmol/L
- LDL cholesterol, mmol/L
- Triglycerides, mmol/L
- Sodium in urine, mmol/L
- Potassium in urine, mmol/L
- Creatinine in urine, mmol/L
- C-reactive protein, mg/L
- Vitamin D, nmol/L
- Glycated haemoglobin, mmol/mol

1.6 Resting electrocardiogram values

- P duration, ms
- QRS duration, ms
- QT interval, ms
- QTc interval, ms
- PP interval, ms
- R axis, degrees
- T axis, degrees
- Ventricular rate, bpm

4 Appendix B: Example cardiovascular non-tabular modalities, typical formats and associated metadata

1.7 Tabular clinical data characteristics (text/numeric)

Typical data models: `CDISC`, `FHIR`, `OMOP`

Typical formats: `CSV`, `TSV` (tab-delimited text-file, counts), `JSON`, `SAV` (SPSS)

- General
- Demographic
- Biochemistry
- Medical conditions
- Surgical conditions
- Mortality data

1.8 Imaging data

Typical formats: `DICOM`, `NIFTI` (extracted image)

- Cardiac computed tomography (CCT)
- Chest X-Ray
- Cardiac Magnetic Resonance Imaging
- Echocardiogram

1.9 Device data

Typical formats: `DICOM`, `XML`

- Electrocardiogram

1.10 Genome sequencing

Typical formats: `FASTA`, `FASTQ`, `BAM`, `SAM`, `VCF`

Depth: e.g. 1x, 100x

- Whole exome sequence (WES)
- Whole genome sequence (WGS)
- Specific panels

1.11 Genome Wide Association Studies (GWAS)

Typical formats: `VCF`, PLINK binary format (`BED/BIM/FAM`), PLINK2 binary (`PGEN/PSAM`), OXFORD-format (`BGEN/GEN`).

Genomic build (e.g. hg19, hg38)

- Imputed (TOPMed, 1000G, etc)
- Array (Affymetrix SNP 5, Illumina GSA)

1.12 Epigenomic data

Typical formats: `IDAT`, `CSV`, `RDS`

- DNA methylation 450K Array
- DNA methylation EPIC Array
- DNA methylation seq
- ChIP-seq data
- NOMe-seq
- ATAC-seq
- histone modifications (e.g. H3K27me3, H3K4me3)
- non-coding RNA modifications (e.g. miRNA, siRNA)

1.13 Transcriptomic data

Typical formats: `FASTQ`, `BAM`, `SAM`, `TSV` (tab-delimited text-file, counts), `RDS` (`SummarizedExperiment`)

- RNA-seq
- Single-cells (sc) RNA-seq
- Single-nucleus (sn) RNA-seq
- Micro (mi)RNA-seq

1.14 Proteomics data

Typical formats: `mzML`, `mzXML`, `pepXML`, `protXML`, `gelML`, `sepML`, `FuGE`, `TSV` (tab-delimited text-file, counts)

- Mass spectrometry data
- Peptide identifications
- Protein abundance measures, e.g. from OLINK platform, or ELISA or LUMINEX assays
- Sample preparation metadata

1.15 Metabolomic data

Typical formats: `JSON`, `TSV` (tab-delimited text-file, counts), `PNG`

- Metabolite profiles
- NMR spectroscopic data
- LC-MS/MS data
- Compound abundance tables
- Spectral data networks

1.16 Whole slide image histological data

Typical formats: `TIF`, `NDPI`, `SVS`

Typical slide staining: e.g. HE (hematoxylin and eosin), SR (picrosirius), EVG (elastin/Elastin van Gieson)

- Carotid plaque
- Coronary plaque
- Myocardium tissue
- Endocardium tissue