NextGen

Genome-Centric Multimodal Data Integration in Personalised Cardiovascular Medicine.

Newsletter N°1

- The NextGen Project: Introduction
- Inside the Engine: NextGen for Researchers and Clinicians

A technical overview of the federated infrastructure, genomic tooling and clinical AI that NextGen is building for cardiovascular medicine

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

The Swiss associated partners of the NextGen project were funded by the Swiss State Secretariat for Education, Research and Innovation (SERI)

Funded by the European Union

This project has received funding from the European Union's Horizon Research and Innovation Programme under Grant Agreement No 101136962

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 numbers 10104323 & 10098097]

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]

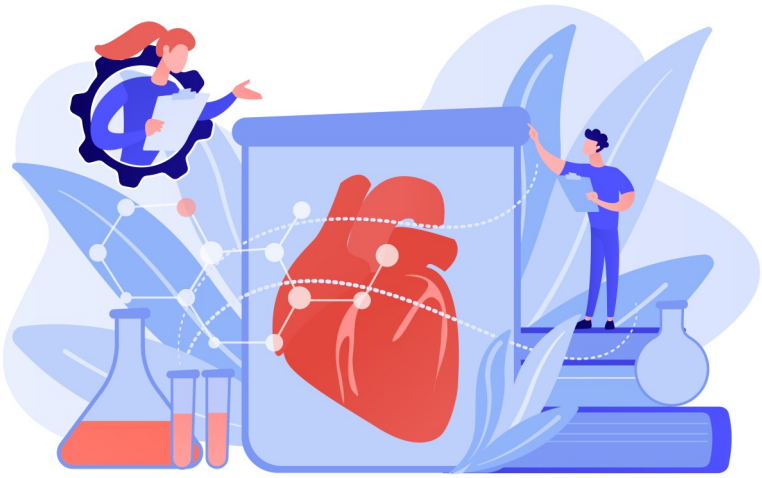
The NextGen Project: Introduction

NextGen — **Next Generation Tools for Genome-Centric Multimodal Data Integration in Personalised Cardiovascular Medicine** — is a four-year research and innovation action funded under Horizon Europe, running from **January 2024 to December 2027**. Coordinated by UMC Utrecht, it brings together 21 organisations across eleven countries. The project **responds to cardiovascular disease**, which causes roughly 3.9 million deaths in Europe each year, and to the fact that the health data needed to study it remains fragmented across institutions, formats and jurisdictions.

Rather than pooling records centrally, NextGen **develops a federated architecture in which analytical tools travel to where data resides**, returning results without patient-level records leaving their home institution. Its work spans data infrastructure, a semantic interoperability layer, genomic processing and clinical AI models, integrated and tested through a demonstration platform, the Pathfinder, alongside dedicated ethics and governance oversight.

At present **the demonstrable outcome is the platform's operational viability on representative dummy cardiovascular data**; clinical benefits remain a prospective, more distant goal.

In this first newsletter we present the technical overview of the NextGen design and implementation from the perspective of Researchers in AI and Clinicians.



Inside the Engine: NextGen for Researchers and Clinicians

A technical overview of the federated infrastructure, genomic tooling and clinical AI that NextGen is building for cardiovascular medicine.

A problem of access to data, not availability

Cardiovascular disease remains the leading cause of death in Europe, accounting for approximately 3.9 million deaths each year — more than cancer, more than respiratory illness, and any other single cause. It is also the largest single driver of healthcare expenditure across EU Member States. **Progress in diagnostics, risk assessment and therapy depends not on the scarcity of data but on the ability to use it**.

Contemporary cardiology continuously generates clinical records, laboratory results and imaging; biobanks hold tissue samples and genetic profiles; and genome sequencing, once very expensive, is now affordable enough that **interpretation is the constraint, rather than acquisition**.

This wealth of information is fragmented: the data relevant to a single patient is rarely held in one location and almost never within a single analytical pipeline. **It is divided by geography, institution, language, technical format and legal jurisdiction. NextGen sets out to close this gap**.

Federation as an architectural principle

NextGen adopts **a federated architecture where analytical tools are deployed to the locations where data already resides**, perform their computations in place, and return results without the underlying patient-level records ever leaving their home institution.

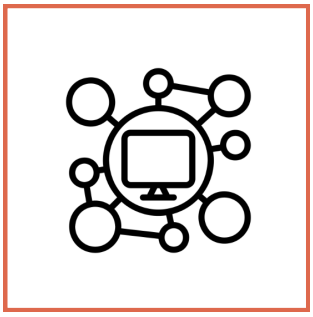
Data remains under the control of the institution that holds it; no raw record is retrieved. Researchers interact with structured queries and model outputs that travel across a network while the data stays local. They are allowed to interrogate data they cannot directly access, within legal frameworks designed to protect the individuals who own their sensitive data.

Federation **reduces the privacy risks** associated with centralising sensitive health information, **but it does not eliminate risk** and the infrastructure work explicitly addresses residual vulnerabilities such as statistical inference attacks, in which repeated queries might allow inferences about individual records.

The semantic layer

A federated network is only useful if the data it connects can be compared. In fact, the same information can and is recorded in different ways, with different metadata and using different structures and entries. Placed side by side, a computer cannot tell that these entries refer to the same underlying concept.

The semantic layer resolves this difficulty. It functions as a translation layer that sits above the data, describing in machine-interpretable form, what each field means, how its values should be understood, and how it relates to comparable fields recorded elsewhere. In effect it provides a shared vocabulary that allows datasets from different institutions to be analysed jointly without requiring any of them to modify their existing systems or adopt a single common format.



Within NextGen this is **delivered through a data management design built on Overlays Capture Architecture**, which permits data elements to be described and translated incrementally rather than through a single imposed standard.

Integration objects and discoverability

Alongside semantic interoperability, **the project is developing the Multimodal Integration Object**: a technical structure that combines semantic and governance information from genomic sequences, clinical measurements, imaging files and wearable-device outputs into a single, traceable package. **Each object carries a digital fingerprint together with a record of the data's origin and the governance rules applicable to its use**, so that when a workflow processes integrated data, provenance and permissions travel with it. A researcher working across genomic and imaging data from multiple sites can, in principle, audit precisely which records contributed to a given result and under which conditions.

A further pillar concerns discoverability. The data management activity has developed a distributed catalogue that renders datasets searchable across institutions without exposing the underlying records. It provides structured, semantically enriched descriptions — the modalities available, the populations represented, and the governance conditions that apply — while actual access remains subject to a separate, governed process. **The design supports discovery without conflating it with disclosure**.

The science engine: multimodal integration

The scientific objective rests on multimodal integration: **combining qualitatively distinct sources to construct a more comprehensive and more actionable representation of disease**. “Multimodal” here spans clinical and health records (electronic health records, together with psychosocial and behavioural data); omics data (whole-genome and whole-exome sequencing, proteomics, transcriptomics, metabolomics and epigenomics); and imaging and device data (MRI, CT scans, histological whole-slide images, and automatically generated outputs such as electrocardiograms).

The clinical rationale is familiar to cardiologists. For a patient with heart failure, an echocardiogram, a sequenced blood sample, several years of clinical records, wearable rhythm monitoring and repeated biomarker results each tell part of the story. A genetic variant associated with cardiomyopathy may be interpreted differently when examined alongside imaging evidence of structural change; a risk model for sudden cardiac death may prove substantially more accurate when genomic data is incorporated in addition to conventional clinical variables. **The principal constraint so far has been the absence of tools** able to work across these data types in a consistent, governed and scalable manner.

Genomic processing and clinical AI

NextGen’s analytical work is **organised into two interconnected areas**. The first is **genomic processing** — computational tools that analyse genomic data more rapidly and more accurately, and across distributed sites. A central line of work is an open-source alternative to the secondary genomic analysis pipeline that connects raw sequence data to medically relevant variants. Historically reliant on proprietary tools, this stage is computationally intensive; the NextGen implementation already demonstrates a substantial speed advantage while preserving accuracy, easing the bottleneck between sequencing and results and reducing the considerable environmental cost of genomic computation.

The second area is **clinical AI**: machine learning models that learn from multimodal data and produce outputs relevant to diagnosis, risk stratification, monitoring and treatment decisions. The project is developing models for atrial fibrillation, cardiomyopathy, congenital cardiac disorders, coronary heart disease, heart failure, myocardial infarction risk, pharmacogenomic responses to cardiovascular drugs, and sudden cardiac death. The breadth of the portfolio reflects the heterogeneity of cardiovascular disease itself: different biological mechanisms, different data signatures, and different points at which earlier or more precise information could meaningfully alter outcomes.

Federated methods in practice

The federated method portfolio is concrete. Federated genome-wide association studies build upon the secure-GWAS software published in Nature Biotechnology, adapted, integrated into the platform and validated. Federated versions of selected tools from the scVI-tools package enable the analysis of single-cell omics data — dimensionality reduction, batch correction and cell population characterisation — through probabilistic models shown to train and deploy effectively in a federated setting. **For bulk transcriptomics, a federated implementation of PLIER (Pathway-Level Information ExtractoR) performs dimensionality reduction on gene expression data while preserving, and potentially enhancing, interpretability.** These developments are orchestrated through Flower, a widely used open-source framework that manages communication and training across distributed clients and enables seamless integration of the tools into the platform.

Combining modalities is not straightforward. Images are intrinsically spatial; some data is tabular; some takes the form of time series; and some consists of long vectors of counts, such as those produced by RNA

The Pathfinder as integration test

These components are integrated into **the Pathfinder, a demonstration network of federated nodes** — each node a participating institution operating its own governed environment, connected to the others through shared standards, a common catalogue layer and interoperable analytical services. Its federated learning workflows allow algorithms to be trained across multiple sites without that training data leaving them, and federated genomic methods distribute analyses such as genome-wide association studies across partners.

Both the platform and the data management layer are **developed in alignment with the European Health Data Space, and compatibility with FHIR, OMOP and CDISC** — the dominant formats in clinical and genomic research infrastructure — is an explicit design objective. Compatibility with GDPR is treated as a design input, with legal and data-protection review embedded from the outset.

The Pathfinder is also the project's integration test. Across five research biobanks in several countries, pilots exercise the full system: multimodal machine learning for heart failure; federated genomic analysis of an inherited cardiomyopathy, where genetic signals too sparse at any single site can be detected by aggregating statistical power across sites without aggregating personal data; the federated catalogue and its discovery workflow; and accelerated secondary genomic analysis benchmarked against real sequencing workloads.

Maturity, and the road ahead

A note on maturity matters for clinical readers. **The AI tools are being built to technology readiness level three or four** — validated under controlled laboratory conditions, with some early-stage testing against relevant datasets, but well short of clinical deployment. The path from research validation to clinical use entails regulatory processes, prospective testing in patient populations, assessment against clinical governance standards, and scrutiny under frameworks such as the EU AI Act and medical device regulation.

NextGen encompasses both the technical development of these tools and the construction of the framework needed to evaluate which of them may eventually become candidates for that journey — including, in due course, candidacy as Software as a Medical Device — while remaining deliberate in distinguishing what it is currently building from what clinical deployment would subsequently require. **The concrete result at this stage is a demonstration that the federated infrastructure, the data management layer, the genomic tools and the analytical methods function together as a coherent system, established using synthetic data structured to represent real cardiovascular datasets.**

For any information needed, please contact **Luca Alessandro REMOTTI**, DataPower, luca.remotti@data-power.net.

Thanks for reading! Don't hesitate to share this newsletter and to follow us on our socials!

Website - NextGentools

Linkedin - NextGen