NextGen

Genome-Centric Multimodal
Data Integration in Personalised
Cardiovascular Medicine.

Deliverable D4.1

Model validation methods and approaches

Grant Agreement Number: 101136962

Funded by the European Union

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

Project funded by

Schweizerische Eidgenossenschaft
Confédération suisse
Confederazione Svizzera
Confederaziun svizra

Swiss Confederation

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

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

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

D3.5 – Synthetic datasets for testing and piloting-1	
Author(s)	Marco Scutari
Editor	Francesca Mangili
Participating partners	SUPSI
Version	1.0
Status	Final
Deliverable date	M12
Dissemination Level	PU - Public
Official date	2025-06-13
Actual date	2025-06-13

Disclaimer

The NEXTGEN consortium consists of the following partners:

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

Document Revision History

DATE	VERSION	DESCRIPTION	CONTRIBUTIONS
13/06/2025	1.0	Complete draft.	SUPSI

Authors

AUTHOR/EDITOR	ORGANISATION
Marco Scutari	SUPSI
Francesca Mangili	SUPSI

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	Multi-model integration object
PGS	Polygenic (risk) score
VCF	Variant call format
WP	Work Package

Table of contents

Summary

This report aims to present an overview of the validation strategies relevant to the NextGen project. To ensure that these strategies are closely aligned with actual project activities, we have structured it around a series of interactive workshops scheduled for the first half of 2025. These workshops serve as a collaborative space to refine validation procedures, acting as a concrete mechanism to deliver project tasks and outputs, and offer practical, context-aware guidance to the Consortium activities. In particular, they strive to:

- Raise awareness of the requirements for Trustworthy AI/ML.
- Provide concrete mechanisms to incorporate requirements into models.
- Provide mechanisms (model card, dataset datasheets) to document compliance.
- Contribute to meeting relevant Regulation, Ethics and Governance requirements.

These workshops explicitly target applications relevant to NextGen, such as the analysis of clinical trials and genomic data. In doing so, they give context to abstract concepts from machine learning and software engineering, making them more approachable to clinical practitioners.

This report summarises the contents of the workshops on trustworthy Artificial Intelligence (AI)/ Machine Learning (ML) models from January to June 2025 (Section 1), plans for further workshops (Section 2), a summary of relevant metrics for model evaluation (Section 3), and the slide decks from the presentations given in the workshops (Section 4).

1 Trustworthy AI Seminars

The workshops covered the following topics that are central to trustworthy AI: fairness (January), robustness (February), accountability (March), the practical application of model cards to statistical genetics (May) and explainable AI (June). No workshop was held in April. A brief summary of each presentation's key points follows. We refer the reader to the slide decks of the individual presentations (Section 4) for a more technical and detailed treatment of the particular topics.

1.1 Fairness in AI/ML

Fairness in machine learning extends beyond model design to data collection and inherent biases. ML models distil information provided to them, incorporating biases present in the data. Data collection issues directly impact fairness through sampling methods, population representation, and quality control procedures that may create bias. Fairness can be understood through a causal framework, where the goal is blocking causal paths from sensitive attributes to outcomes.

Implementation strategies include:

- Pre-collection: Recording sensitive attributes to detect and correct biases.
- Pre-processing: Applying sampling weights to counter selection bias or adjusting data points.
- Post-processing: Using hold-out sets to assess and alter predictions for improved fairness.

An inherent trade-off exists between fairness and predictive performance that must be balanced according to application requirements.

1.2 Model Evaluation, Validation and Robustness

The ML model lifecycle begins with project scoping, which involves:

- Defining the specific problem to solve.
- Identifying optimisation targets with measurable metrics.
- Determining necessary data sources and access methods.

Preliminary analysis that follows project scoping should assess multiple dimensions:

- Robustness against noise, model misspecification, and adversarial attacks.
- Interpretability and explainability of model behaviour.
- Fairness to prevent discrimination based on sensitive attributes.
- Privacy and security protections.

A principle to follow is that "good data trumps good models"—addressing issues through principled data collection is preferable to complex modelling solutions. Due to their many interconnected components, modern ML systems require holistic design.

Having defined the scope of the model, what information to train it and quantifiable measures to assess whether it is working satisfactorily, we can:

- Evaluate its performance in terms of predictive accuracy and other metrics of statistical goodness of fit.
- Validate the model using (human) expert advice and new data collected in field experiments.
- Assess different failure models of the model to ensure its robustness.

1.3 Accountability

Accountability encompasses both human oversight capabilities and responsibility allocation in case of problems. Related concepts include:

- Transparency: User awareness of automated decisions and understanding of decision processes
- Explainability: Ability to explain why a specific decision was made
- Interpretability: Ability to explain how a decision-making process functions

For proper accountability, organizations should track:

- Data: Sources, combination methods, feature selection, and pre-processing
- Model Training: Training data, hyperparameter selection, and evaluation procedures
- Model Deployment: Runtime environments, access controls, and update protocols
- Inference Outputs: Accuracy measurements and temporal changes

This comprehensive tracking supports regulatory compliance with frameworks like the EU AI Act, Cyber Resilience Act, and GDPR.

While model and data cards provide useful documentation, they are insufficient alone. Complete experiment tracking requires configuration management platforms with version control for all system components.

1.4 Applied Example: Feline Cardiomyopathy Prediction Model

The CHAI Model Card presented in May's workshop demonstrates these principles in practice through a Feline Cardiomyopathy Prediction (FCP) model. This model predicts cardiomyopathy subtypes from targeted NGS data with:

- Overall accuracy: 91%
- Subtype-specific F1 scores: HCM 0.92, DCM 0.81, RCM 0.87, No disease 0.95
- ROC-AUC: 0.96

The model was designed with fairness considerations to ensure similar accuracy across different feline breeds, ages, and sexes. Development involved diverse stakeholders, including veterinary cardiologists, genomic researchers, and data security experts. All cat owners provided informed consent, and ongoing post-development monitoring includes regular audits and updates as new data becomes available.

The CHAI model card is designed for complex AI models deployed in large-scale applications, so it contains redundant sections for simpler statistical genetics models used in research.

1.5 Explainable Artificial Intelligence

Explainable Artificial Intelligence (XAI) addresses the "black box problem" of deep learning models by making their predictions more understandable and interpretable for humans. XAI encompasses local (single instance) and global (all possible inputs) techniques to counter different forms of opacity in AI systems, from simply explaining which features contributed to a decision to establishing causal reliability through formal models. XAI approaches include:

- Post-hoc methods: Feature attribution (saliency masks, heat maps, scoring lists).
- Surrogate models that approximate black box behavior.
- Intrinsic (by-design) explainability incorporating interpretable structures.
- Causal Reliability: Moving beyond correlations to causal models with interventions and counterfactuals.

1.6 Overarching Frameworks and Principles

All workshops started with a brief introduction highlighting the overarching frameworks and principles of trustworthy AI. For the frameworks, we have identified:

1. The European Commission's High-Level Expert Group on AI (HLEG) has seven key requirements: human agency and oversight; technical robustness and safety; privacy and data governance; transparency; diversity, non-discrimination and fairness; societal and environmental well-being; and accountability.
2. NIST AI Risk Management Framework (RMF): which characterizes trustworthy AI systems as: accountable and transparent; valid and reliable; safe, secure and resilient; fair with harmful bias managed; explainable and interpretable; privacy-enhanced.
3. The Coalition for Health in AI (CHAI)'s Blueprint for Trustworthy AI provides guidelines for responsible AI use in healthcare.

As for the regulations, we identify several regulations and laws, including:

1. The EU AI Act: Holds individuals/organizations developing, deploying, or operating AI systems responsible for their actions.
2. The EU Cyber Resilience Act: Requires Software/Hardware Bill of Materials (SBOM) for systems based on AI/ML models.
3. The General Data Protection Act (GDPR): requires appropriate technical and organizational measures to demonstrate fair, appropriate and transparent data use and their effectiveness.

The identified overarching principles of trustworthy AI are those listed earlier:

- Accountability: human oversight and correction capabilities; responsibility allocation in case of problems; comprehensive tracking of the entire ML model and data pipelines; auditability, responsibility, and redress.
- Fairness: bias sources and their propagation, including in data collection; historical/social/temporal biases; measurement, sampling, representation issues; systematic errors; lack of representative data.

- Technical robustness, from original robustness (standard model training) to generalization capacity (say, against data distribution shifts) and other risks (say, adversarial attacks). Addressed through resilience to attacks and security, fallback plans and general safety, accuracy, reliability, and reproducibility.
- Documentation and transparency: use of model cards and data cards to document key information; experiment tracking using configuration management platforms; traceability mechanisms throughout the AI lifecycle; clear communication about AI involvement and capabilities

1.7 Implementation guidance

- Good data trump good models: addressing issues through principled data collection is preferable to complex modelling solutions.
- Holistic design is essential due to the many interconnected components in modern AI/ML systems.
- Trustworthiness requirements are different at various stages of the development and uses of AI/ML models.
- There exists an inherent trade-off between fairness and predictive performance.
- These frameworks collectively emphasize a comprehensive lifecycle approach to AI development that prioritizes human values, regulatory compliance, and responsible implementation practices.

2 Plans for Further Workshops

The trustworthy AI workshop series is planned to continue until at least the end of 2025. The first seminars concentrated on introducing the key concepts referenced in trustworthy AI. Future workshops will be more applied, with NextGen partners demonstrating these concepts in the context of their respective use cases in the second half of the year.

3 Metrics for Model Evaluations

Trustworthy AI requires crucially evaluating the statistical performance of models. The expected operating conditions are established at the beginning of the model lifecycle and are measured against suitable operating standards with metrics measuring either goodness of fit or predictive accuracy. In the following, we summarise the key metrics that may be of interest in NextGen.

It is recommended that these metrics be evaluated on an independent cohort rather than relying solely on data-splitting approaches such as cross-validation.

3.1 General Machine Learning Metrics

AI/ML model applications in NextGen are evaluated using both general-purpose metrics and application-specific ones, which we will discuss separately in later subsections.

For AI/ML models with discrete outcomes, such as case/control, the general-purpose metrics that are most appropriate are:

- The F1 score.
- The area under the ROC curve (AUROC) measures the trade-off between the false positive and true positive rates.
- The area under the PRC curve (AUPRC) measures the trade-off between precision and recall.

For models with continuous outcomes, such as quantitative phenotypes in statistical genetics, the general-purpose metrics that are most appropriate are:

- The squared Pearson's correlation (R^2), which is equivalent to the amount of explained variance.
- The rank correlation, which measures ranking agreement.

Useful references for these metrics are https://doi.org/10.1038/s41598-022-09954-8 and https://psycnet.apa.org/doi/10.1037/a0028087.

3.2 Sequence Data

The analysis of sequence data in statistical genetics is traditionally based on linear and logistic regression. As a result, the metrics used to assess AI/ML models for sequence data are R^2, or equivalently the proportion of explained variance, for continuous phenotypes and the AUROC for binary phenotypes

such as case/control labels. The proportion of variance explained by sequence data is often called "heritability" in this context.

3.3 Single-Cell Genomics

A comprehensive list is provided in https://doi.org/10.1038/s41592-021-01336-8.

The presence of batch effect can be assessed with:

- K-nearest-neighbor batch effect test (kBET).
- K-nearest-neighbor (kNN) graph connectivity.
- The average silhouette width (ASW) across batches.
- The graph integration local inverse Simpson's Index (graph iLISI, extended from iLISI21).
- PCA regression.

The conservation of biological variance, avoiding the "regression towards the mean" phenomenon, can be assessed with:

- Normalised mutual information (NMI).
- Adjusted Rand index (ARI).
- Average silhouette width (ASW per cell type).
- Graph local inverse Simpson's Index (cLISI).
- Isolated label F1 score.
- Isolated label silhouette
- Cell-cycle (CC) conservation.
- HVG conservation.
- Trajectory conservation.

3.4 Medical Imaging

A comprehensive list is provided in https://doi.org/10.1016/j.bspc.2016.02.006.

Image quality assessment for clinical images, such as magnetic resonance imaging (MRI), computed tomography (CT), and ultrasonic images, can be grouped into:

- Measures of noise that are functions of the signal-to-noise ratio (SNR): Shannon's information content, SNR, PSNR, Contrast-to-Noise Ratio (CNR), Mean Square Error (MSE), and Root-Mean-Squared Error (RMSE).
- Subjective measures of discrepancy: Double-Stimulus Continuous-Quality Scale (DSCQS), Difference Mean Opinion Score (DMOS) and Perceptual Difference Model (PDM).
- Feature extraction and comparison, based on methods like kernel transforms and discrete cosine transforms.

4 Slides

4.1 January

Aaron Lee

Marco Scutari (Fairness)

4.2 February

Aaron Lee

Marco Scutari (Robustness)

4.3 March

Aaron Lee

Marco Scutari (Accountability)

4.4 April

There was no workshop in April.

4.5 May

Jessica van Setten and Jade Raffle

4.6 June

Alberto Termine