NextGen

Genome-Centric Multimodal Data Integration in Personalised Cardiovascular Medicine.

Deliverable 3.6

Synthetic datasets for testing and piloting -2

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]

Author(s)	Aaron Lee, Marco Scutari
Editor	Francesca Mangili
Participating partners	SCUOLA UNIVERSITARIA PROFESSIONALE DELLA SVIZZERA ITALIANA, THE HUMAN COLOSSUS FOUNDATION, QUEEN MARY UNIVERSITY OF LONDON, EURECOM GIE, HL7 INTERNATIONAL FOUNDATION
Version	1.0
Status	Online version for update
Deliverable date	December 2025
Dissemination Level	Public
Official date	2025/12/18
Actual date	2025/12/01

Disclaimer

WP1 lead

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
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
22	ASSOCIACAO DO INSTITUTO SUPERIOR TECNICO PARA A INVESTIGACAO E O DESENVOLVIMENTO	IST-ID	PT

WP1 lead

Document Revision History

DATE	VERSION	DESCRIPTION	CONTRIBUTIONS
November 2025	Draft1	Online version for comment	Authors listed below

Authors

AUTHOR/EDITOR	ORGANISATION
Aaron Lee	QMUL
Marco Scutari	SUPSI
Philippe Page	HCF
Evangelia-Anna Markatou	HL7
Raja Appusawamy	EURECOM
Daniele Malpetti	SUPSI

Reviewers

REVIEWER	ORGANISATION
Francesca Mangili	SUPSI
Catarina Barata	IST-ID

WP1 lead

List of terms and abbreviations

ABBREVIATION	DESCRIPTION

Table of contents

WP1 lead

1 Introduction

The term "synthetic data" takes different meanings depending on how and why it is produced, for instance, to mimic salient characteristics of a relevant phenomenon for statistical analysis, or to stress software in specific ways to test its performance or correctness. The Royal Society and the Alan Turing Institute, in "Synthetic Data -- what, why and how?"[1], have proposed the definition: *"Synthetic data is data that has been generated using a purpose-built mathematical model or algorithm, with the aim of solving a (set of) data science task(s)."* In the ONS methodology working paper (series number 16 - Synthetic data pilot) a "synthetic dataset spectrum" is proposed: *"At one end we define a purely synthetic dataset suitable for code testing with no ambition to replicate underlying data patterns and at the other end of the spectrum sits a dataset created by clever augmentation of the original, which aims to replicate the patterns contained within the original data source. Clearly, the latter carries extremely high disclosure risk and needs to be approached cautiously."*

In NextGen, we use a broad definition and consider synthetic data as constructed to reflect some characteristics of a real or imagined dataset. Specifically, we will define a synthetic dataset to be a constructed entity that shares specific characteristics of actual (or imagined) reference data.

The characteristics referred to above include, but are not limited to:

- Technical characteristics of dataset representations (e.g. file size and format)
- Semantics - how information is represented
- Measurements and statistical properties related to the domain of the data

Defined in this manner, examples of synthetic data include both:

- "Technical synthetic data": Data which is generated without reference to an existing dataset (e.g. for developing or testing technical functionality),
- "Healthcare synthetic data": Data that is mathematically modelled on an existing source dataset to replicate some characteristics of its informational content (e.g. to remove personal information while retaining specific medical characteristics or summary statistics).

For the former, a simple data/model card that provides sufficient documentation to enable effective reuse, as technical synthetic data pose no privacy risks. The data/model card should describe the software used to generate the data, and report functional capabilities and limitations of the models used.

For the latter, documentation beyond data and model cards may be required because healthcare synthetic data is governed by specific regulations and ethical considerations. Among other sensitive issues, they are at a higher risk of privacy issues. For instance, the TEHDA2S "M7.2 Draft guideline on data minimisation, pseudonymisation, anonymisation and synthetic data"[2], the resistance to re-identification is also emphasised:

"The EDPB stated that the documentation of synthetic data generation should include the model's theoretical resistance to re-identification techniques (§58e) and meet the purpose and data

[1] arXiv:2205.03257,

[2] https://tehdas.eu/wp-content/uploads/2025/09/draft-guideline-on-data-minimisation-pseudonymisation-anonymisation-and-synthetic-data.pdf (accessed October 2025)

WP1 lead

minimisation principles (§64) (EDPB Opinion of the Board (Art. 64), Opinion 28/2024 on certain data protection aspects related to the processing of personal data in the context of AI models[]). In other words, synthetic data may fall under GDPR if individuals can still be re-identified with reasonable effort. Therefore, it is necessary to demonstrate the resistance to re-identification", M7.2 Draft guideline on data minimisation, pseudonymisation, anonymisation and synthetic data.

WP1 lead

2 Synthetic datasets used in NextGen

The synthetic datasets used in NextGen are described in the following table. Where possible, the generation code (rather than the actual datasets) will be made available for download before the end of the project. **Links presented here are for reference, but may not be persistent. At the end of the project, material that is intended to be openly available will be appropriately stored and referenced in Deliverable 3.7 "Synthetic datasets for testing and piloting-3" for Month 36.**

Dataset description (all fields are required)									
Name of dataset	Creator (Partner)	Description	Purpose	Technical or healthcare synthetic data or both (defined above)	Derived from individual level data (yes/no)?	Generation code intended to be accessible outside of NextGen	Access and Usage Restrictions	Method of access of generation code and documentation	Documentation (yes/no)
HAPMAP	Marco Scutari (SUPSI)	DNA sequence data, created from the reference data set of the same name. Generated with the HapGen software.	Testing random generation of genotype data from a panel. Testing phenotype generation from sequence data and annotated variants.	Healthcare synthetic data.	Yes.	Yes	Free for academic use only.	Code to generate the data from D3.5. Associated model card. (Both are included at the end of this report.)	Yes
Exomiser processed synthetic	Damian Smedley & Yasemin	Exomiser output from synthetic rare disease patients	Testing federated learning training for the machine learning	Technical synthetic data	Yes	The patient simulator used is already a publicly	None	Documentation on how to use PhEval to generate the	Yes

WP1 lead

Dataset description (all fields are required)									
Name of dataset	Creator (Partner)	Description	Purpose	Technical or healthcare synthetic data or both (defined above)	Derived from individual level data (yes/no)?	Generation code intended to be accessible outside of NextGen	Access and Usage Restrictions	Method of access of generation code and documentation	Documentation (yes/no)
patient data	Bridges (QMUL)		component of Exomiser.			available open source program (PhEval)		synthetic patient data and run through Exomiser will be provided	
Simulated reads	Raja Appuswamy (EURE)	Raw DNA sequencing reads in FASTQ format generated from HG38 reference genome using the read simulator tool that has been artificially modified to add a mutation relevant for cardiac disease.	The raw reads will be used to test the accelerated secondary analysis pipeline being developed by EURECOM and compare it with GATK with respect to both performance and accuracy.	Technical synthetic data	No	The read simulator used is already a publicly available open source program	None	https://github.com/BioInfoTools/BBMap	Yes
Synthetic Patients	Robert Mitwicki (HCF)	Sample cardiovascular variables/modalities relevant to the development of Pathfinder Minimum Viable Product	Test files to be used for demonstration and integration test of semantic and MMIO creation tooling	Technical synthetic data	No	The generation script is located at the HCF NextGen GitHub.	GitHub Access granted upon request to HCF	https://github.com/THCLab/NextGen/blob/master/README.md	Generation script has basic documentation

WP1 lead

Dataset description (all fields are required)									
Name of dataset	Creator (Partner)	Description	Purpose	Technical or healthcare synthetic data or both (defined above)	Derived from individual level data (yes/no)?	Generation code intended to be accessible outside of NextGen	Access and Usage Restrictions	Method of access of generation code and documentation	Documentation (yes/no)
Simulated Patient Data	Evangelia-Anna Markatou	Patient Data. In addition to medical data. It includes demographics and socioeconomic data.	Test Diversity Index	Technical synthetic data	No	Yes	None	Will be uploaded to GitHub upon completion	Will be available upon completion

WP1 lead

3 Next steps

We envisage that creating a catalogue of the synthetic data used in NextGen will be beneficial in the ongoing development of the platform and the associated software. After the project members have documented what synthetic data they use, they will be able to concentrate the testing of both newly developed software and ML/AI methods on this data. While this does not completely obviate the need for the end-to-end testing of the ML/AI pipelines developed within NextGen, it reduces the risk that inconsistencies between different steps (developed by different partners) within each pipeline will result in software bugs or statistically incorrect outputs. This document will be updated to form Deliverable 3.7 "Synthetic datasets for testing and piloting-3" for Month 36.

WP1 lead

WP1 lead

5 Appendix

This appendix contains a Model Card for HAPGEN 2 and a Data Card for the syntethic data described in Section "Synthetic Data Generation".

5.1 Model Card

Model Card - HAPGEN2	
Model Details	
Developer	Zhan Su, Jonathan Marchini, Peter Donnelly
Version	2.2.0
URL	https://mathgen.stats.ox.ac.uk/genetics_software/hapgen/hapgen2.html
Licence	https://mathgen.stats.ox.ac.uk/genetics_software/hapgen/LICENCE (forbids commercial use)
Type	Generates artificial data sets for case-control studies by applying an extension of the Li-Stephens algorithm to a panel of SNV data.
Reference	[1]
Intended Use	
– Intended for use in academic research on human genetics, commercial use is forbidden. – Potentially unsuitable for inbred species, polyploid species and recombinant species. – Standard panels HapMap 3 and 1000 Genomes are recommended but no longer provided.	
Factors	
– It can only generate SNV data for the SNVs present in the input panel, within a specified range of allele frequencies and without missing data. – It provides very limited options for assigning phenotypes to the generated genotypes. – The genotypes will have the same patterns of linkage disequilibrium and average allele frequencies as the panel provided to HAPGEN. Therefore, the population the input panel comes should be chosen to reflect the target population for the specific application (CEU, YRI, CHJB, JPT, etc.).	
Training Data	
– User-provided panel of genome-wide SNV data. The panel data used to generate	

the data is described in the Data Card below.

5.2 Data Card

Data Card – Input Panel	
Panel	HapMap 3 (release 2) haplotypes - NCBI Build 36 (dbSNP b126)
Population	CEU
Chromosome	18
Sample size	1301 individuals, 99454 SNVs
Phenotypes	None.
URL (alternative to HAPGEN's)	https://www.sanger.ac.uk/data/hapmap-3/
Licence	International HapMap Project Public Access License (copy: http://www.worldlii.org/int/other/PubRL/2003/4.html) This licence allows use of the data but not redistribution. The data must be donwloaded from the original source to reproduce the analysis described earlier in the report.
Notes	No medical or personal identifying information was obtained from individuals providing the samples. However, the samples are identified by the population from which they were collected.

Data Card – Generated Synthetic Data	
Panel	-
Population	CEU
Chromosome	18
Sample size	1000 individuals, the first 5000 SNVs in the chromosome.
Phenotypes	Binary phenotype with 1000 causative loci placed randomly in the chromosome, heritability ~60% and prevalence 25%.
URL (alternative to HAPGEN's)	https://www.sanger.ac.uk/data/hapmap-3/

Licence	HAPGEN's licence applies.

5.3 Computer Code

The following code calls HAPGEN2 to generate the SNVs.

```
#!/bin/bash -e

set -o pipefail

HAPGEN_URL=https://mathgen.stats.ox.ac.uk/genetics_software/hapgen/download/builds/x86_64/
v2.2.0/hapgen2_x86_64.tar.gz
HAPGEN=`basename $HAPGEN_URL`
OUTDIR=./output
PREFIX=hapmap
CASES=0
CONTROLS=1000

# download and extract HAPGEN 2.
wget -c $HAPGEN_URL
tar xvzf $HAPGEN hapgen2
# extract the HAPMAP panel.
gzip -d CEU.0908.chr18.hap.gz
# create a directory in which to save the intermediate files.
mkdir -p $OUTDIR
# generate the genotypes.
valgrind ./hapgen2 -h CEU.0908.chr18.hap \
   -l CEU.0908.chr18.legend \
   -m genetic_map_chr18_combined_b36.txt \
   -o $OUTDIR/$PREFIX \
   -dl 441 1 1.5 2.5 1124 0 3 4.5 1149 1 2.0 6.5 \
   -n $CONTROLS $CASES

# cut and format the data in 0/1/2 allele counts on columns
cut -d" " -f6- $OUTDIR/$PREFIX.controls.gen > $OUTDIR/formatted
split -l 500 $OUTDIR/formatted $OUTDIR/formatted2.

for FILE in `find $OUTDIR -type f -name "formatted2.*" | sort`
do
 awk '
 {
    for (i = 1; i <= NF; i++)  {
       a[NR, i] = $i;
    }
 }
 END {
    for(j = 1; j <= NF; j = j + 3) {
       str = ""
       for(i = 1; i <= NR; i++) {
          acount = a[i, j] * 0 + a[i, j + 1] * 1 + a[i, j + 2] * 2;
```

```
            str = str""acount" ";
          }
          print str;
      }
  }' $FILE > `sed -e "s/2/3/" <<< $FILE`
done

# merge the data back into a single file.
cut -d" " -f 2 $OUTDIR/$PREFIX.controls.gen | tr '\n' ' ' > $OUTDIR/$PREFIX.final.gen
echo >> $OUTDIR/$PREFIX.final.gen
paste -d" " $OUTDIR/formatted3* >> $OUTDIR/$PREFIX.final.gen
# compress and copy to the final location.
gzip -k -9 $OUTDIR/$PREFIX.final.gen
cp $OUTDIR/$PREFIX.final.gen.gz raw-$PREFIX.CEU.chr18.txt.gz
# clean up.
rm -rf output
```

The following code generates the PLINK2 files from the outputs of HAPGEN2.

```
# load phenotypes and genotypes.
pheno = readRDS("phenotypes.rds")
geno = readRDS("prepd-hapmap.rds")

geno = geno[, 10000 + 1:5000]

# FAM file.
contents = data.frame(
  FAMID = paste0("FAM", sprintf("%04d", seq(nrow(geno)))),
  WFID = rep(1, nrow(geno)),
  FA = rep(0, nrow(geno)),
  MO = rep(0, nrow(geno)),
  SEX = sample(c("F", "M"), nrow(geno), replace = TRUE),
  PHENO = as.integer(pheno$pheno)
)

contents$SEX = as.integer(factor(contents$SEX, levels = c("M", "F")))

write.table(contents, file = "generated.fam", quote = FALSE, row.names = FALSE,
  col.names = FALSE)

# PED file (FAM + calls).
for (g in seq(ncol(geno))) {

  calls = sapply(geno[, g], function(g) {

    if (g == 0)
      return(c(1, 1))
    else if (g == 2)
      return(c(2, 2))
    else
      return(sample(1:2))

  })

  contents = cbind(contents, t(calls))
```

```
}

write.table(contents, file = "generated.ped", quote = FALSE, row.names = FALSE,
  col.names = FALSE)

# MAP file.
legend = read.table("CEU.0908.chr18.legend", header = TRUE)

contents = data.frame(
  CHR = rep(18, ncol(geno)),
  ID = legend[seq(ncol(geno)), "ID"],
  MORGAN = rep(0, ncol(geno)),
  BP = legend[seq(ncol(geno)), "pos"]
)

write.table(contents, file = "generated.map", quote = FALSE, row.names = FALSE,
  col.names = FALSE)

# BIM file.
contents = cbind(contents, legend[seq(ncol(geno)), c("allele0", "allele1")])

write.table(contents, file = "generated.bim", quote = FALSE, row.names = FALSE,
  col.names = FALSE)

# finally, run: ./plink --file generated --make-bed --out binary
```

The following code generates the phenotypes from the SNVs.

```
# load the markers.
markers = readRDS("prepd-hapmap.rds")

generate.binary.phenotypes =
  function(markers, nqtl = 1000, size = 0.048, prevalence = 0.25) {

  # ... generate the noise term...
  noise = rnorm(nrow(markers), sd = 1)
  # ... sample the QTLs and the associated effect sizes...
  betas = rep(0, ncol(markers))
  betas[sample(ncol(markers), nqtl)] = rnorm(nqtl, sd = size)
  # ... and generate the continuous liabilities as intermediate phenotypes.
  liability = as.vector(as.matrix(markers) %*% betas + noise)
  # check the heritability of the phenotype.
  cat("@ estimated heritability is:",
      (var(liability) - var(noise)) / var(liability), "\n")
  # create the binary phenotypes from the liability to achieve the desired prevalence (a la LDAK).
  thr = quantile(liability, probs = 1 - prevalence)
  pheno = cut(liability, breaks = c(-Inf, thr, Inf), labels = c("CTRL", "CASE"))
  return(list(liability = liability, pheno = pheno, betas = betas))

}#GENERATE.BINARY.PHENOTYPES

pheno = generate.binary.phenotypes(markers)
```

```
saveRDS(pheno, file = "phenotypes.rds")
```

NextGen Data Processing

Responsibility

Responsibility	
Partner institution	SUPSI
Responsible person (s)	Marco Scutari
Comments (if relevant)	
Form completed by (name)	Marco Scutari
Date completed	12/03/2024

Description of the data used in processing

Brief description of the dataset/data used in the data processing	
Name of dataset/biobank/source	HapMap 3 (release 2) haplotypes - NCBI Build 36 (dbSNP b126)
Acronym (if available)	HapMap
Biobank details (*when source of data is a locally hosted biobank*)	
Origin of dataset (and link)	Sanger Institute (https://www.sanger.ac.uk/data/hapmap-3/)
Processing / proposed use (s)	Used a population panel to generate synthetic data with the HapGen software.
Characterisation of data	High risk data (directly identifiable personal data), pseudonymised data, anonymised data, synthetic data (etc).
Types/modalities of data (brief, outline)	Single Nucleotide Polymorphisms (SNPs). No medical or personal identifying information was obtained from individuals providing the samples. However, the samples are identified by the population from which they were collected.
Approximate number of patient level records (real or simulated)	1301 individuals.
Approximate “on disk” size/volume of data	~800MB

References/links	International HapMap Project Public Access License (copy: http://www.worldlii.org/int/other/PubRL/2003/4.html) This licence allows use of the data but not redistribution. The data must be downloaded from the original source to reproduce the analysis described earlier in the report.

Ethics and Data Protection

EU Ethics and Data Protection Tree (optional, recommended)
Please complete the Ethics and Data Protection Tree https://ec.europa.eu/assets/rtd/ethics-data-protection-decision-tree/index.html This is helpful to understand the EU position. It is a “yes/no” graphical decision tree.
[Optional but recommended: please add a PDF printout of the result of completion of this decision tree, or any relevant/helpful comments]

Description of Processing

Description of data processing activities within NextGen *How will you use the data in NextGen?*	
Please check all that apply and are related specifically to activities in NextGen (Personal Data is defined as per the GDPR)	Research & development Genomic research Imaging research Pathology research Multimodal research Algorithm or model development Generation of derived quantities **Construction of new datasets** Anonymised data **Synthetic data** Test data Other (please specify below) **System testing and development** Platform MMIOs **Genomic processing** Other (please specify below) Other data processing Storage/collection/moving Re-distribution of data or derived material Storage of Personal Data Collection of Personal data (Re)distribution of Personal data Other (please specify below)

Outline description of how you will use the data (processing) Please delete pre-filled guidance information for clarity (Optionally you may complete the Appendix to provide more specific information)	
Brief description	This data set is used as an input to the HapGen software to generate synthetic data for Deliverables 3.5, and to demo federated GWAS software splink in Deliverable 3.1.
Please indicate any additional considerations	This data set is a standard benchmark data set in the academic literature, and has been available to the public at large for more than a decade.
References	Deliverable 3.5.
Have you provided additional details in the Appendix (yes/no)	No.

Data generated in NextGen (e.g. synthetic data) Please complete a datasheet for any generated data	
Add link to datasheet(s)	The details of the synthetic data set generated from the HapMap data using HapGen are in the appendix.

Models produced in NextGen Please complete a model card for model architectures used/distributed in NextGen	
Add link to model card(s)	

Confirmation of Compliance

Confirmation that use of data is lawful and ethical	
Is a contract/license required for access/use?	No.
If yes, confirmation that this contract has been executed, or licence accepted, and will be complied with in full with respect to the activities carried out in NextGen	
Please confirm that the technical and organisational measures used in the data processing with respect to NextGen activities comply with all lawful requirements for use of this data	Confirmed.
Are there notable limitations on data processing?	No.
Is any of the data processing taking place outside of the EU?	No.
Please provide any other information/references relevant to use of this dataset (not otherwise provided)	Licence allows use of the data but not redistribution. The data must be downloaded from the original source to reproduce the analysis described earlier in the report.

Appendix: Further Details (Optional)

Construction of new datasets (complete if relevant / delete if not relevant) Please delete pre-filled guidance information for clarity	
What types of new datasets are being constructed for use in NextGen?	A small synthetic data comprising 1000 individuals of CEU descent, for 5000 SNVs in chromosome 18. Synthetic phenotype
Will these datasets be shared in NextGen?	No.
How will any risks associated with sharing this data in NextGen be addressed?	These synthetic data have no inherent risk.
Please indicate any additional considerations	HAPGEN's licence applies (forbids commercial use): https://mathgen.stats.ox.ac.uk/genetics_software/hapgen/LICENCE The HapGen software is available from: https://mathgen.stats.ox.ac.uk/genetics_software/hapgen/hapgen2.html