



Project acronym: EOSC4CANCER

Grant Agreement Number: 101058427

Project full title: A European-wide foundation to accelerate Data-driven Cancer Research

Call identifier: HORIZON-INFRA-2021-EOSC-01

D1.4 Synthetic cohort status

Version:	1.0
Status:	Final for publication
Dissemination Level:	Public
Due date of deliverable:	29.02.2024
Actual submission date:	29.02.2024
Work Package:	WP1
Lead partner for this deliverable:	BSC-CNS
Partner(s) contributing:	

Main author(s):

Salvador Capella	BSC-CNS
Sergi Aguiló	BSC-CNS
Miguel Vázquez	BSC-CNS
Alberto Labarga	BSC-CNS
Romina Royo	BSC-CNS

Abstract

This document provides the description of the work carried out to design and implement strategies for the generation of synthetic data that can be used as demonstrators of the technical developments within and beyond the project. The generated synthetic data can be freely available, hence it can be shared among all partners without the need of data transfer or sharing agreements. Cancer specific synthetic cohorts including different modalities (i.e., genomic data together with clinical data based on OMOP) will be created, delivered, and stored in publicly available repositories (e.g., EGA).

History of changes

Version	Date	Changes made	Author(s)
0.1	11.28.2023	FIRST DRAFT	Romina Royo (BSC-CNS)
0.5	01.18.2024	Including a description of genomic and clinical data generation	Miguel Vázquez (BSC-CNS), Alberto Labarga (BSC-CNS)
0.8	02.07.2024	Completing all sections	Sergi Aguiló (BSC-CNS), Romina Royo (BSC-CNS)
0.9	02.08.2024	Final review before sending to reviewers	Sergi Aguiló (BSC-CNS), Romina Royo (BSC-CNS)
1.0	02.22.2024	Addressed reviewers' comments	Sergi Aguiló (BSC-CNS), Romina Royo (BSC-CNS)

Statement of originality

This deliverable contains original unpublished work except where clearly indicated otherwise. Acknowledgement of previously published material and of the work of others has been made through appropriate citation, quotation or both.

TABLE OF CONTENTS

EXECUTIVE SUMMARY	4
1 Introduction	5
2 Description of work accomplished	6
2.1 Strategy overview	6
2.2 Data generation	8
2.2.1 Genomic data	8
2.2.2 Clinical data	10
3 Contribution towards technical work packages, use cases and the community	12
4 Summary of the synthetic cohorts	14
5 Conclusions	15
6 Next steps	16
7 Annex	17
7.1 Longitudinal genomic and clinical data	17
7.2 synth4bench	17
7.3 MIMIC-GENIE-CRC simulated patient profiles dataset	18
7.4 Synthetic tumour/normal WGS	19

EXECUTIVE SUMMARY

Access to human research data involves lengthy processes and agreements between data holders and researchers to ensure the ethical and responsible use of this sensitive data. These protocols can introduce significant delays to technical developments and deployment of early demonstrators of research use-cases. For this reason, we propose the generation and use of synthetic data, which can be freely available and shared among all participants, to push forward the technical achievements of the project.

This task will design strategies to generate synthetic data and develop a set of cancer-specific synthetic cohorts with multiple data modalities (clinical, molecular, etc.). These developments will be aligned with other projects like CINECA and B1MG that have developed synthetic cohorts modelled on their participating cohorts and experimental data types, and will respond to the needs of the EOSC4Cancer use-cases and their technical implementation. We will use our experience in ICGC-ARGO as the basis for the generation of genomic data, which will allow us to create longitudinal genomic samples at different stages of the disease. Clinical data will be generated from a state-of-the-art tool that generates patient data with OMOP CDM format, one of the most common models used by hospitals and further secondary use of data research nowadays.

These cohorts will be made available via the same discovery and data access interfaces as the real cohort data, in collaboration with WPs 2 and 4. Thus, they can serve as demonstrators for analysis and visualisation platforms, as well as federated analysis environments within WP3. Finally, the developed synthetic cohorts will be hosted and made publicly available to the whole cancer community via the participating data resources (e.g., EGA).

1 Introduction

In recent years, the field of health research has witnessed a transformative shift towards precision medicine, with the massive integration of genetic information, electronic health records and other clinical data. While this approach holds a promising goal for the near future medicine by increasing the accuracy of the treatments for each individual patient, it is constrained by the necessary protection measures surrounding sensitive patient data.

In trying to safeguard patient privacy, regulations such as General Data Protection Regulation (GDPR) have been created to mandate controlled access to this type of data, which occasionally leads to a lengthy and complex process until the accession approval. Consequently, the scientific community, especially the ones working with highly sensitive data, faces significant problems accessing diverse datasets, making it complex to continue with their research.

Synthetic data emerges as a solution to these challenges. Synthetic datasets, being artificially generated by mirroring real-world data patterns, offer an alternative for researchers to bypass the privacy and accession problems that they are facing. This will help the researchers with faster development and testing of protocols, fostering the creation of new technological advancements and demonstrators.

Moreover, the unrestricted nature of synthetic data allows for open-sharing and distribution. Researchers can utilise these datasets to establish robust infrastructures, refine analytical systems, and benchmark protocols, as it has been demonstrated during the course of EOSC4Cancer. Although synthetic data will never be a substitute for real data for achieving the highest level of scientific rigour, the data proves to be an ally in the pursuit of advancing research efficiently.

Hence, many projects are now generating synthetic cohorts used in their technical developments and prototypes. The working group on cancer from B1MG (WG9) has prepared synthetic data and open access data (based on cell lines) consisting of genomic data and clinical data that adheres to the Minimal Dataset for Cancer, a model they developed for the collection of cancer-related clinical information and genomics metadata. Similarly, GDI provides more than 2,500 synthetic genomics and phenotypic data (including cancer, rare diseases and population genomics) to enable nodes to rapidly deploy a functioning pilot system. CINECA (Common Infrastructure for National Cohorts in Europe, Canada, and Africa) has also created multiple cohorts to showcase and demonstrate federated research and clinical applications. The European Genome-phenome Archive (EGA), commonly used as an archive for genomic research, has a synthetic data landing page with several studies offering open-access synthetic data following the usual EGA protocols. We will also contribute to their synthetic data section by providing a new study for synthetic cancer data generated within the project.

In the following section, we will describe the work accomplished by task 1.2, with the generation of clinical and genomic synthetic data. The focus of this work has been on defining the strategies for the generation of synthetic data, mainly longitudinal genomic and clinical data, which mimics the usual trajectory of patients being diagnosed with cancer and having a follow-up that can show, for instance, a relapse after treatment with a resistant mutation or disease progression. Next, the contribution to other technical work packages and use cases will be explained. Finally, the conclusions section will highlight the pros and cons of synthetic data and the next steps sections will indicate the continuation and potential improvements of this work.

2 Description of work accomplished

2.1 Strategy overview

The complexity of cancer is compounded by the evolutionary nature of the disease, where the cancer cells undergo genetic changes over time, driving tumour heterogeneity and adaptability. So, understanding the temporal aspects of cancer progression is paramount, especially when treatments are applied.

The patient's cancer journey is characterised by distinct time points. The initial phases involve assessment of treatment efficacy, monitoring signs of remission, and looking for any resistant cell population. Longitudinal studies become crucial at these points, capturing the evolution of the disease over time.

Moreover, new mutations acquired within cancer cells add another level of complexity. These mutations can drive tumour evolution, alter the disease trajectory and impact treatment responsiveness. Also, investigating samples at relapse can provide insights into the mechanism of treatment resistance.

By simulating longitudinal studies in cancer research, we are mirroring the complex data frames generated by the mix of treatments, relapses, and any mutation appearing or expanding in the course of the disease. To do so, we have simulated clinical longitudinal data with the EHR of patients having a “typical” cancer journey, with their treatments, relapses and consequent mutations. Then, we generated the normal and multiple tumour samples by whole-genome sequencing according to their clinical history that reflects the genomic landscape at different time points shaped by treatment pressures and the evolution of their tumour. An illustrative example of this process can be seen in [Figure 1](#).

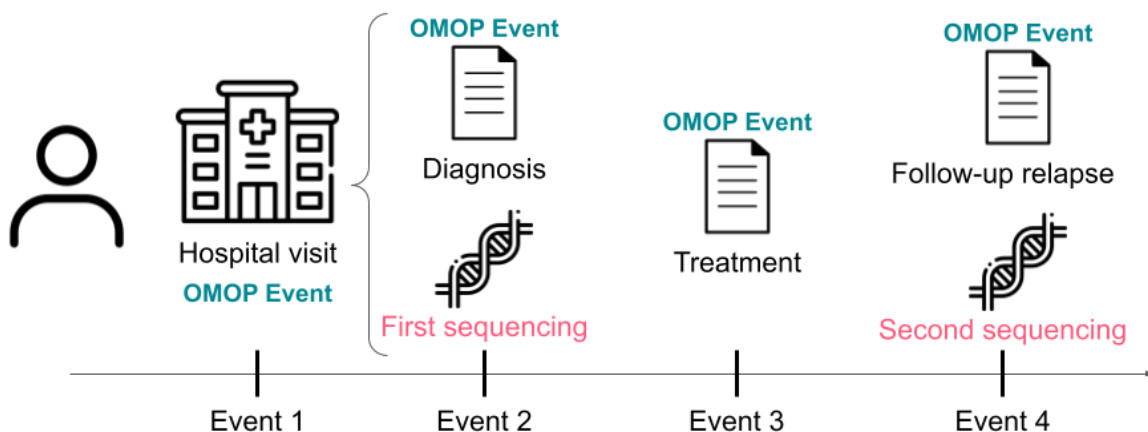


Figure 1. Simple schema of a patient's cancer journey in a hospital. In this example, a patient goes to a hospital for a visit. After the visit, they get a diagnostic, in this case cancer, and the first sequencing of the tumour/normal sample is performed. According to the tumour type and patient characteristics, the patient receives a particular treatment. Finally, the patient returns to the hospital where they are informed of a relapse. Consequently, a second sequencing is performed.

First, we generated a collection of genomic and clinical longitudinal data corresponding to one patient. This allowed us to develop and set up the procedures to simulate the data and verify their results in a controlled manner. More details about the developed methodologies can be found in the next section, '[Data generation](#)'.

For the first dataset, both the OMOP and the genomic generators received a table with pre-filled information about the patient, which was manually curated and based on real data. To mimic actual cancer data, we have used mutations found in colorectal cancer (CRC) patients from the Pancancer Analysis of Whole Genomes (PCAWG) initiative, an OMOP

model built on statistics of real CRC patients based on Navarra Hospital’s Cancer Unit, and the treatments together with their resistant mutations are based on literature. [Table 1](#) shows the events corresponding to the clinical history and the molecular data of this patient. Each row shows an event along the course of the disease, which indicates a genomic event (identification of a driver mutation, clonal evolution of the tumour, etc.) and/or a clinical event (information from a hospital visit). This information corresponds to the input data that both the Genomic generator and the OMOP generator will use to create the synthetic data, which will be validated in the end ([Figure 2](#)).

	Mutation burden	Clone	OMOP EVENT	SEQUENCING SAMPLE	Variant
Driver mutation TP53	low	1			chr17:g.7578406C>T
Driver mutation APC	low	2<-1			chr5:g.112174631C>T
cetuximab resistance mutation (BRAF V600E)	intermediate	3<-2			chr7:g.140453136A>T
Driver mutation FBXW7	high	4<-2			chr4:g.153247289G>A
Clone 4 expansion					
OMOP initial diagnosis, Tumor sequencing			YES	YES	
OMOP treatment with FOLFOX and cetuximab			YES		
Remission					
OMOP remission			YES		
Driver mutation RAD50	intermediate	5<-3			chr5:g.131924497A>G
Clone 5 expansion					
OMOP recurrence diagnosis, Tumor sequencing			YES	YES	
OMOP treatment with FOLFOX: 5-FU, leucovorin, oxaliplatin			YES		
OMOP progressive disease			YES		

Table 1. List of events for the generation of the first synthetic dataset corresponding to a patient with multiple timepoints. The events shown in the first column indicate the genomic and clinical history of this case. OMOP entries and genomic sequencing will be generated according to the ‘OMOP EVENT’ and ‘SAMPLE SEQUENCING’ columns. For the genomic data, the mutation indicated in the ‘Variant’ column will be used together with background mutations according to the ‘Mutation burden’ column. The subclonal architecture will be based on the clonal evolution indicated in the ‘Clone’ column, where each subclone is represented by a number, and the arrow indicates its ancestor subclone.

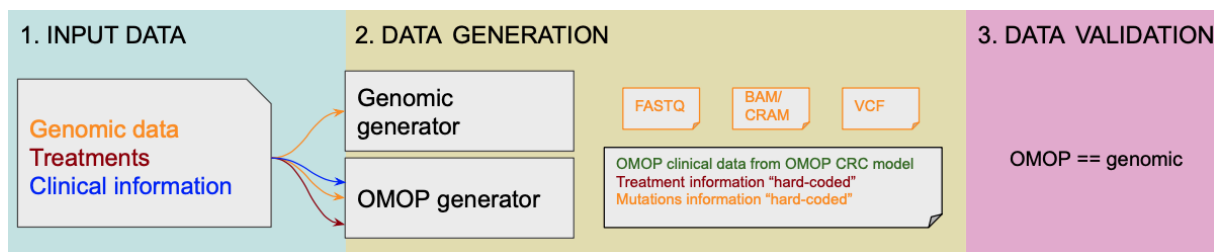


Figure 2. Schema of the methods used to generate the first dataset of synthetic data, including a patient with longitudinal clinical and genomic data. Input data is manually created and used by the Genomic and OMOP generators that, based on this, simulate raw genomic data (FASTQ format) and clinical data following the OMOP CDM. In this simulation, the treatment and mutation information was added to the previously generated OMOP data. The raw FASTQ files were aligned and mutations were called to obtain BAM and VCF files, respectively.

Next, we explored ways to generate the input data (semi-)automatically. The first approach, which we implemented and is presented here in this deliverable, uses the OMOP generator, to select driver mutations automatically based on their frequencies in a real cohort of CRC

patients (from the PanCancer Analysis of Whole Genomes project, PCAWG). Schema in [Figure 3](#).

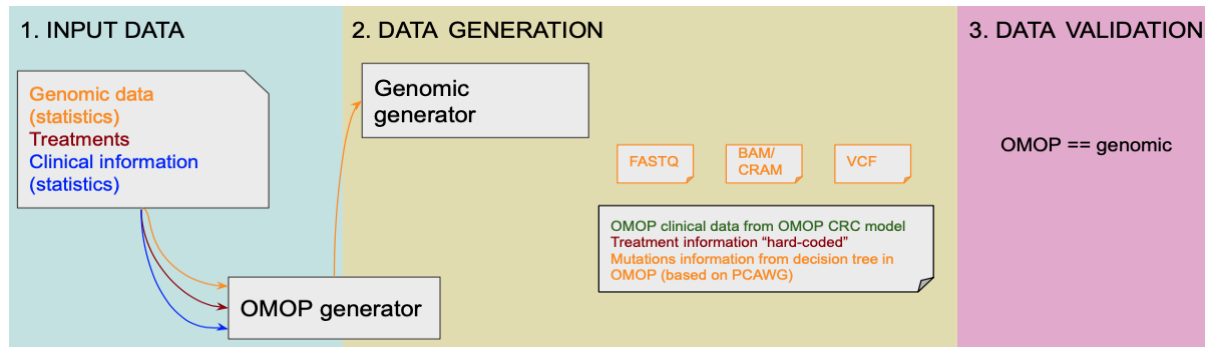


Figure 3. Schema of the second strategy for genomic and clinical data generation. Here, the driver mutations acquired along the course of the disease together with their clonal evolution is automatically generated in the OMOP generator.

With this, the OMOP generator can output equivalent tables with mutation information to be passed to the genomic generator. An example of these new tables can be seen in [Table 2](#).

Clone	Proportions	Mutations	Mutation Burden	Treatment resistance	Patient
1	0.27	NC_000017.10g.7577548C>T	low	0	186
1->2	0.27	NC_000001.10g.27106354C>T	low	0	186
2->3	0.06	NC_000012.11g.25378647T>A	high	0	186
3->4	0.13	NC_000012.11g.25380276T>A	high	0	186
1->5	0.24	NC_000007.13g.140453136A>T	high	1	186
5->6	1	NC_000017.10g.7578406C>T	low	0	186
...

Table 2. Table of mutations per patient produced by the OMOP generator and used as the input of the genomic generator. The example only shows the data for patient 186 (see the ‘patient’ column), but there is information for all the patients generated. The first two columns indicate the subclonal evolution of the genomic subclones: the ‘Clone’ column indicates the ancestry of the subclones represented by numbers and the arrow indicates the corresponding ancestor of each subclone, the ‘Proportions’ column indicates the fraction of the subclone within the tumour. The ‘mutations’ column indicates the driver alteration, the ‘burden’ column indicates the mutation burden of the subclone harbouring the corresponding driver mutation. And the ‘resistance’ column tells if the mutation is resistant to treatments according to the ICGC (International Cancer Genome Consortium) catalogue (0 indicates no resistance and 1 means resistance).

2.2 Data generation

2.2.1 Genomic data

The genomic data generator simulates whole-genome sequencing (WGS) based on the mutational information provided in the input table (see [2.1 Strategy overview](#)), which includes the definition of tumour subclones with their driver alterations and mutational burden,

together with their clonal evolution. As an illustrative example, the genes with driver alterations and their order of acquisition are shown in [Figure 4](#).

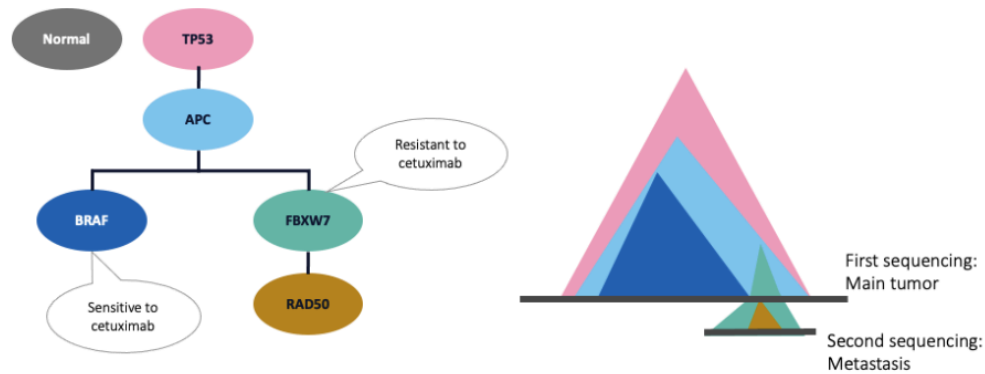


Figure 4. Clonal evolution showing driver genes. Each colour corresponds to a subclone. The genes with driver alterations and their order of acquisition is shown on the left, while the clonal evolution showing the proportion of each subclone in the simulated sequencings (horizontal black lines) is shown on the right.

The genomic data generation is performed by simulating reads from each of the clones separately and then mixing them according to their clonal proportions in each of the sequences. Normal tissue is also simulated. The final tumour and normal sets of reads can also include reads from each other according to tumour-in-normal and normal-in-tumour contamination values. A workflow is used to generate these clone reads and mix them into samples.

Each clone is simulated using NEAT GenReads¹, from the hg38 reference. To ensure proper phasing, each copy of the chromosome is simulated separately. A germline genotype is generated for each sample using 1000 Genomes, where each single-nucleotide polymorphism (SNP) is incorporated or not based on its allele frequency and placed into a random chromosome copy (currently, all SNPs are heterozygous, we are working on changing this to support homozygous SNPs as well). In addition to the germline genotype, each clone receives a set of somatic variants, which consists of a set of variants newly appearing on the clone as well as the variants from its clonal ancestry. The set of variants appearing on each clone consists of the driver variants specified in the patient description and several passenger mutations. Passenger mutations are drawn randomly from PCAWG, from a cohort specified by the patient's cancer type; first, a number of mutations are drawn collectively for the cohort's full list of somatic variants following a specified mutation burden (defaults to 15 mutations per MB), then these mutations are distributed across clones according to the clones specified relative mutation burden.

Structural variants are supported in the simulation as well. Currently, they are all introduced as passengers. They are formed by sampling random breakpoints in the specified PCAWG cohort and assigning them to three categories: deletion, insertion, or inverted insertion.

All the simulated reads are annotated with the clone and the chromosome copy and location from which they arose to facilitate debugging.

To cover the different needs of the end users, we also provide the alignment (BAM format) of the raw FASTQ files generated as previously described, as well as the mutations identified in them (VCF format). There are two workflows used to produce these data: the first one from FASTQ to variant calling²; the second for variant prediction/annotation and conversion to MAF format³. Both pipelines are openly available and can be run in Galaxy.

¹ <https://github.com/zstephens/neat-genreads>

² <https://usegalaxy.eu/published/workflow?id=2c3d05023c02113e>

³ <https://usegalaxy.eu/published/workflow?id=1da86d74f8535f4e>

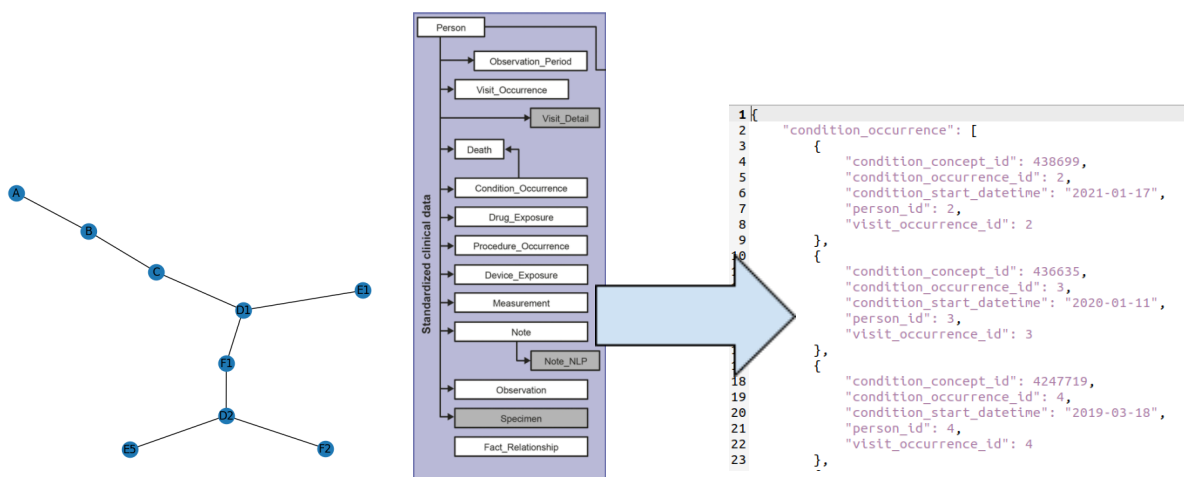
2.2.2 Clinical data

With the objective of generating a collection of synthetic electronic health records that is free of legal, privacy, security, and intellectual property restrictions, we developed Pysynth. This open-source software package simulates the lifespans of colorectal cancer patients in the OMOP-CDM format. The framework for the synthetic data generation process implemented by Pysynth is based on the use of PADARSER⁴, the Publicly Available Data Approach to the Realistic Synthetic EHR. The PADARSER framework is also used by Synthea⁵, another popular synthetic patient data generator. PADARSER relies on publicly available datasets and statistics to populate the synthetic EHR. It uses clinical guidelines or protocols in the form of care maps, and employs methods that guarantee inherent realistic properties in the resulting synthetic EHR, making them sufficient enough to replace real records for secondary uses that require realistic but not real EHRs.

Pysynth uses a configuration file in the YAML format that reflects the demographics and clinical statistics for the population of interest (sex, age, visits, conditions, treatments, etc.). In our case, these are based on publicly available health statistics about colorectal cancer in Spain. For the generation of genetic variants associated with the disease, we have used ICGC Data Portal (<https://dcc.icgc.org/>) to retrieve variant frequency and clinical significance for mutations with the primary site 'Colorectal'.

The configuration file defines an event graph where events correspond to the OMOP-CDM domains: person, visit_occurrence, procedure_occurrence, observation, and the transition probabilities between events. This graph is enacted, and the corresponding OMOP-CDM database entries are generated. The synthetic data can be saved as a SQLite database (Figure 5).

Pysynth code is available upon request at <https://gitlab.bsc.es/health-data/pysynth>.



⁴ Dube, K., Gallagher, T. (2014). Approach and Method for Generating Realistic Synthetic Electronic Healthcare Records for Secondary Use. In: Gibbons, J., MacCaull, W. (eds) Foundations of Health Information Engineering and Systems. FHIES 2013. Lecture Notes in Computer Science, vol 8315. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-53956-5_6

⁵ Jason Walonoski, Mark Kramer, Joseph Nichols, Andre Quina, Chris Moesel, Dylan Hall, Carlton Duffett, Kudakwashe Dube, Thomas Gallagher, Scott McLachlan, Synthea: An approach, method, and software mechanism for generating synthetic patients and the synthetic electronic health care record, Journal of the American Medical Informatics Association, Volume 25, Issue 3, March 2018, Pages 230–238, <https://doi.org/10.1093/jamia/ocx079>

Figure 5. Schema of the methodology to generate the clinical data. The data is generated with a decision tree, where it creates data of the patients depending on anterior events and the probability of the event taking place. For example, there is an event where it creates the age and sex of the patient. Depending on the values of the patient, the disease diagnostics will occur in a major or minor probability. Also, it defines the EHR with all the visits that this patient has undergone.

3 Contribution towards technical work packages, use cases and the community

The first set of synthetic genomic data served as a demonstrative basis of collaborative research between EOSC4Cancer and GDI, showcased in the Federated Analysis Workshop (11th of October at the Barcelona Supercomputing Center, Barcelona). The purpose of this joint use case is to exemplify data mobilisation across platforms. This involves the data flow from raw data, followed by their analysis, to the visualisation of the results. More specifically, the synthetic genomic dataset (in FASTQ format) created in EOSC4Cancer is stored in a Nextflow server, from where it is integrated into Galaxy, where the analysis is run. In this case, the analysis consisted of the alignment of the FASTQ files, quality control steps, and the variant calling phase, where mutations are identified. The result of this is a list of mutations that is returned in MAF format. Then, the resulting MAF files are imported into cBioPortal, where they can be shown in a graphical and user-friendly manner. This demonstrator shows the interoperability and compatibility of the generated data across different platforms. The schema can be seen in [Figure 6](#) and the video demonstrator presented in the workshop is available in the project Drive⁶.

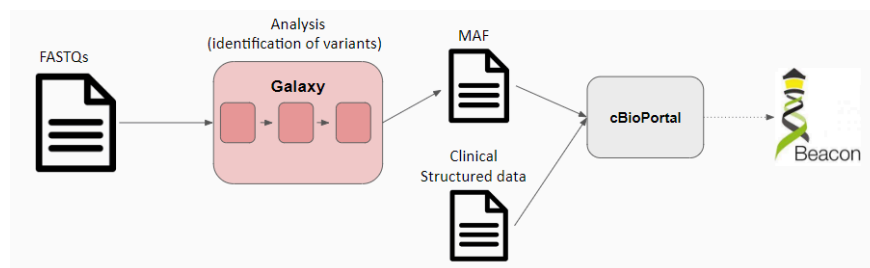


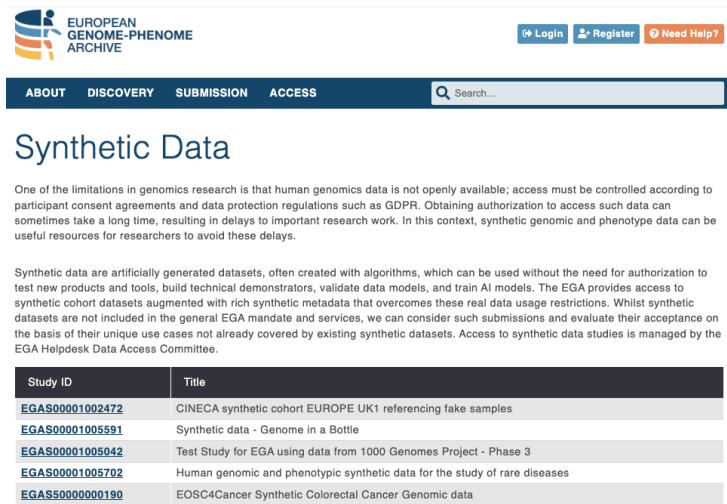
Figure 6. Schema of the demonstrator at the Federated Analysis workshop with all the steps from raw data to cBioPortal. The figure shows the data flow from the raw data, to Galaxy where analysis is run and the resulting MAF file is created, and then, together with the clinical data already formatted, the information is imported into cBioPortal. Finally, although this step was not done in the workshop, if the clinical data follows a data model it can be queried by Beacon.

To foster accessibility and utilisation, the synthetic datasets generated in the context of EOSC4Cancer will be publicly available in the European Genome-Pheno Archive (EGA), a repository highly used by the cancer research community to store and share private data. EGA has a dedicated section specifically for synthetic data⁷ (see [Figure 7](#) for an overview of all available datasets). The datasets generated in EOSC4Cancer will be cataloged in this section, ensuring that researchers within the cancer research have full access to these resources.

For now, the synthetic data generated in the project, clinical and genomic, are uploaded in a dataset within the EOSC4Cancer study ([EGAS50000000190](#)) in the Synthetic Data section in EGA.

⁶ <https://drive.google.com/file/d/1mFZGcpHvw29hpcvL9300-vCOzdo1xf6u/view?usp=sharing>

⁷ <https://ega-archive.org/synthetic-data/>



EUROPEAN
GENOME-PHENOME
ARCHIVE

Login Register Need Help?

ABOUT DISCOVERY SUBMISSION ACCESS

Synthetic Data

One of the limitations in genomics research is that human genomics data is not openly available; access must be controlled according to participant consent agreements and data protection regulations such as GDPR. Obtaining authorization to access such data can sometimes take a long time, resulting in delays to important research work. In this context, synthetic genomic and phenotype data can be useful resources for researchers to avoid these delays.

Synthetic data are artificially generated datasets, often created with algorithms, which can be used without the need for authorization to test new products and tools, build technical demonstrators, validate data models, and train AI models. The EGA provides access to synthetic cohort datasets augmented with rich synthetic metadata that overcomes these real data usage restrictions. Whilst synthetic datasets are not included in the general EGA mandate and services, we can consider such submissions and evaluate their acceptance on the basis of their unique use cases not already covered by existing synthetic datasets. Access to synthetic data studies is managed by the EGA Helpdesk Data Access Committee.

Study ID	Title
EGAS00001002472	CINECA synthetic cohort EUROPE UK1 referencing fake samples
EGAS00001005591	Synthetic data - Genome in a Bottle
EGAS00001005042	Test Study for EGA using data from 1000 Genomes Project - Phase 3
EGAS00001005702	Human genomic and phenotypic synthetic data for the study of rare diseases
EGAS00000000190	EOSC4Cancer Synthetic Colorectal Cancer Genomic data

Figure 7. Synthetic data section in EGA. It lists the different synthetic datasets that are available in the EGA. The data generated in the context of EO SC4Cancer is located in one of these studies.

4 Summary of the synthetic cohorts

Besides the longitudinal genomic and clinical data that we generated (see previous sections and [Annex 7.1](#)), there have been other efforts that produced more tailored datasets for specific tasks and/or other methodologies that can be used to generate synthetic data. We called on all EOSC4Cancer participants to gather their input on existing methods for generating cancer-related synthetic data and already existing datasets.

In the Annex, we provide a list of these datasets with further information. For each dataset, we answered a collection of questions to briefly describe the cohort, its purpose, the methods for data generation and quality assurance, and the links to the tools and/or data when available.

Some project members had done some work focused on the tools for generating the data rather than creating specific datasets. CERTH made an internal effort to generate synthetic genomics data ([Annex 7.2](#)) for the evaluation of somatic variant calling algorithms. At the same time, BBMRI has worked with methods to synthesise WSI data, thanks to the developments done at one of their BBMRI partners⁸ (Medical University Graz (MUG at BBMRI.at)).

On the other hand, UBx has been working on the generation of synthetic data for patients with molecular data and clinical trial matching ([Annex 7.3](#)). This effort has been specifically designed to meet their needs in T3.1, where they are developing an AI-based method to automatically propose clinical trials to patients according to their molecular profile. Similarly, UiO has generated tumour/normal pairs of WGS to serve as a demonstrator for task T3.2, where they will implement a joint genotyping connecting cBioPortal and Galaxy. Additionally, this dataset will also be re-used in GDI for technical demonstrations, and it will be added to the EOSC4Cancer study for synthetic data in central EGA, and possibly to the Norwegian federated EGA node.

⁸https://openaccess.thecvf.com/content/WACV2024/html/Harb_Diffusion-Based_Generation_of_Histopathological_Whole_Slide_Images_at_a_Gigapixel_WACV_2024_paper.html

5 Conclusions

As it can be seen in this deliverable, there is a major work in progress in the development of synthetic data in EOSC4Cancer, especially in regard to the clinical and genomic data. However, this work is still in progress and will continue throughout the project, responding to the needs of the technical developments as well as the research questions of the use cases. Moreover, the importance of the synthetic data is highlighted by complementary datasets generated by other project members, within and outside of EOSC4Cancer, contributing to the general view of the deliverable.

Synthetic data is now becoming a hot topic in our research, especially when dealing with private cancer data. Inside the project EOSC4Cancer, there have been active discussions regarding its purpose.

In summary, the discussions among project members underscore the nature of the challenges and potential solutions regarding synthetic data in cancer research. As the synthetic data may mimic real-world data, most of the time it is used to expedite hypothesis and algorithm development, thus having the same variables and values as the real data but without having any links to identify the patients. The other reflection is on the necessity of addressing data accessibility challenges. While we might be able to advance our work based on synthetic data (e.g., testing algorithms), in the end, we will need real data to solve scientific and clinical questions. Thus, we should also improve the accessibility of high-quality real datasets.

Indeed, the consensus among project members emphasises the importance of collaborative efforts to improve access to high-quality real-world data while recognising the potential of synthetic data as a temporary solution. Moreover, the time investment required for data preparation underscores the practical usefulness of synthetic data.

These discussions serve to illuminate the complexity of data accessibility challenges and highlight the diverse perspectives within the project regarding the role of synthetic data in advancing cancer research and AI. As we continue with the research, it is imperative to leverage both synthetic and real data, fostering collaboration and innovation to drive meaningful progress in the collective pursuit of understanding and combating cancer.

Overall, synthetic data offers a versatile and powerful tool for cancer research, enabling researchers to overcome data limitations, protect patient privacy, validate analytical methods, and advance our understanding of cancer biology and treatment. However, it cannot fully replicate the complexity and clinical relevance of real patient data. Therefore, real data remains indispensable for advancing our understanding of cancer biology, developing effective treatments, and improving patient outcomes.

Finally, to make the data available for the whole community, we have deposited the synthetic datasets in the EGA. EOSC4Cancer aims to create a centralised hub for researchers looking for synthetic data for their investigations in the context of cancer research. This leads to the dialogue around open data sharing and also empowers the cancer research community with a repository of high-quality datasets that can be freely distributed and used.

6 Next steps

In our previous work, we successfully developed tools for synthetic clinical and genomic data based on predefined parameters, encompassing patient clinical histories, treatments, and genomic profiles. Additionally, we have automatised the generation of clinical and genomic data, leveraging statistics from real patient datasets.

The next phase of the work involves automating the incorporation of treatment information, drawing from resources that integrate molecular profiling to specific treatments. In this way, driver alterations could prescribe the optimal treatment options for individual patients. To facilitate this, we will explore different available platforms, such as PanDrugs and IntOGen, which offer comprehensive insights into the associations between mutations and potential treatment modalities.

PanDrugs, for instance, provides a platform that prioritises anticancer drug treatments based on individual multi-omics data, helping with drug indication, approval status, gene-drug associations, and drug response profiles to yield personalised drug scores and therapeutic recommendations. Similarly, IntOGen offers a comprehensive framework for automatic knowledge extraction from mutation data, identifying cancer genes and describing their mechanism of action across various tumours.

We will continue in close contact with the use cases of the project, as well as the technical work packages, to ensure alignment with their evolving needs and objectives. As they evolve with their respective developments, we will refine our synthetic data generation processes and datasets to meet their requirements.

Furthermore, our engagement extends beyond the project, with collaborations with initiatives like GDI, where our synthetic datasets have already been utilised. We are committed to aligning our future work with these collaborative efforts, responding proactively, and contributing to our shared goals in cancer research.

7 Annex

7.1 Longitudinal genomic and clinical data

Dataset Name: Longitudinal genomic and clinical data

Short dataset description: Genomic and clinical data at different time points to model the evolution of the disease (e.g., diagnosis and after treatment relapse)

Was it generated within EOSC4Cancer? or somewhere else (if so, please, provide us the project/link to the project website)? EOSC4Cancer

Data type (genomics, imaging, proteomics, structured clinical data, etc): Genomics and clinical data

Do you follow any particular data model? Which one? Clinical data is generated using OMOP (and we will map it to the Minimal data set for cancer from B1MG in the future). For the genomic data, the raw data is generated in FASTQ format, and the alignment and variant calling are performed, producing BAM and VCF formats, respectively.

Do you impose any quality criteria? Which one? We will ensure the consistency of the genomic and clinical data (i.e., the driver mutations are both present in the OMOP records of the patient and the generated genomic data).

Is a synthetic dataset generated for 1) technical purposes, e.g. evaluating a given technology, 2) as a proxy to a one or more real dataset, 3) both, or 4) others (please, provide some explanation about it). It has been generated for technical purposes.

Has it been published or described somewhere? Can you share a link/DOI? The methods to generate this data have been described in this deliverable.

Is it publicly available? Can you share the link to it?
<https://ega-archive.org/studies/EGAS50000000190>

Is the methodology used to generate it publicly available in a repository? Can you share the link to it? [Synthea](#) for clinical data and [NEAT-genreads](#) for genomic data.

7.2 synth4bench

Dataset Name: synth4bench

Short dataset description: A framework for generating synthetic genomics data for the evaluation of somatic variant calling algorithms

Was it generated within EOSC4Cancer? or somewhere else (if so, please, provide us the project/link to the project website)? No, it is a CERTH-internal effort

Data type (genomics, imaging, proteomics, structured clinical data, etc): Genomics

Do you follow any particular data model? Which one?: NEAT (NExt-generation sequencing Analysis Toolkit) <https://github.com/ncsa/NEAT>

Do you impose any quality criteria? Which one?: Basic Bioinformatics metrics like quality of bases in reads, bases' distribution compared to reference genome etc.

Is a synthetic dataset generated for 1) technical purposes, e.g. evaluating a given technology, 2) as a proxy to a one or more real dataset, 3) both, or 4) others (please, provide

some explanation about it): Our datasets are created to explore NGS data intrinsic parameters and the effect they have on somatic variant callers.

Has it been published or described somewhere? Can you share a link/DOI?: 10.5281/zenodo.8432060, 10.7490/f1000research.1119575.1

Is it publicly available? Can you share the link to it?: Yes, 10.5281/zenodo.8095898

Is the methodology used to generate it publicly available in a repository? Can you share the link to it?: Yes, <https://github.com/BiodataAnalysisGroup/synth4bench>

7.3 MIMIC-GENIE-CRC simulated patient profiles dataset

Dataset Name: The name of the dataset is "MIMIC-GENIE-CRC simulated patient profiles dataset."

Short dataset description: The MIMIC-GENIE-CRC simulated dataset integrates genomic data from the AACR Project Genie (GENIE Cohort v15.0-public) and textual clinical information from the MIMIC-IV (Medical Information Mart for Intensive Care, 4th edition) dataset. The AACR Project Genie is a global collaborative effort to aggregate and share genomic and clinical data from cancer patients. The primary goal of AACR Project Genie is to create a comprehensive and diverse dataset that can be used for cancer research, with a particular emphasis on genomic information. On the other hand, The MIMIC (Medical Information Mart for Intensive Care) dataset is a publicly available database that contains de-identified health-related data collected from patients admitted to the Beth Israel Deaconess Medical Center (BIDMC) in Boston, Massachusetts. The dataset is widely used for healthcare informatics, machine learning, and data science research. The MIMIC-GENIE-CRC simulated dataset is a unique dataset that pseudo-randomly merges genetic variations data from AACR project Genie with ICU discharge notes from MIMIC-IV. In particular, we generate 484 hypothetical profiles, each representing a synthesised colorectal cancer patient, combining genomic insights with detailed clinical narratives that describe their present condition and clinical history. Although the matching has not been validated to ensure biological plausibility, careful consideration has been given to data compatibility by selecting patients whose condition follows the condition naming convention of the International Classification of Diseases, Tenth Revision (ICD-10). The goal of the simulated dataset is to aid in developing an AI-powered tool called TrialMatchAI, which recommends clinical trials for cancer patients based on their unique genomic and clinical profiles.

Was it generated within EOSC4Cancer? or somewhere else (if so, please, provide us the project/link to the project website)? It was generated within EOSC4Cancer and for a very specific step in developing the AI-powered patient-trial matching tool, TrialMatchAI.

Data type (genomics, imaging, proteomics, structured clinical data, etc.): The generated data includes genomics (mutations report from AACR Genie on cBioportal) and unstructured clinical notes (ICU discharge notes) of colorectal cancer patients.

Do you follow any particular data model? Which one? We do not follow a particular data model. The data are stored in their original form. In particular, genomic data follows the data schema of cBioportal, whereas the MIMIC-IV clinical notes are unstructured free texts.

Do you impose any quality criteria? Which one? Our data processing workflow implements a meticulous data filtration strategy aligned with the International Classification of Diseases, Tenth Revision (ICD-10) condition naming convention. Our focus centres on identifying patients whose condition field conforms to a specific regular expression pattern designed for

malignancies related to the colon, rectum, anus, hepatic flexure, cecum, appendix vermiformis, and large intestine.

The regular expression pattern used for selection is as follows:
`r'\b(?:malignant\s+(?=.*\b(?:colon|rect|rectum|rectal|anal|anus|hepatic flexure|cecum|appendix vermiformis|flexure|large intestine)\b)|.*\b(?:colon|rect|rectum|rectal|anal|anus|hepatic flexure|cecum|appendix vermiformis|flexure|large intestine)\b)\s+malignant\b'`

This approach ensures precision and standardisation in identifying relevant cases, promoting consistency with global healthcare coding standards defined by ICD-10.

Is a synthetic dataset generated for 1) technical purposes, e.g. evaluating a given technology, 2) as a proxy to a one or more real dataset, 3) both, or 4) others (please, provide some explanation about it): Creating the MIMIC-GENIE-CRC simulated patient profiles dataset is exclusively driven by a technical objective – the development of TrialMatchAI, an AI-powered tool engineered to identify pertinent clinical trials tailored to each patient's distinct genomic and clinical profiles. The overarching aim is to harness the dataset's richness to enhance the tool's capabilities in anticipating and processing critical types of cancer-related data when deployed in a clinical or research environment.

Has it been published or described somewhere? Can you share a link/DOI? No, the simulated dataset has not been published, nor is it intended to be published in the future. However, the parent datasets are available on online platforms with some limitations on access and usage. The AACR project Genie dataset is available on cBioportal (<https://genie.cbioportal.org/>) and requires authorization from the creators to access it. In contrast, the MIMIC-IV is available on PhysioNet (<https://physionet.org/content/mimic-iv-note/2.2/>) and requires passing free formal training courses before being granted access to download and use the dataset.

Is it publicly available? Can you share the link to it? Although we created the simulated dataset, we don't have legal authorization to share or transmit it publicly, even within the EOSC4cancer project. Interested parties must attain the required authorizations from the sources of the parent datasets before demanding access to our simulated dataset.

Is the methodology used to generate it publicly available in a repository? Can you share the link to it? Yes, the methodology used to create the dataset is publicly available as a Python Jupyter Notebook on the public repository of TrialMatchAI. Please note that the repository is still under development. The link to the code: <https://github.com/majdabd/TrialMatchAI/blob/main/src/SyntheticData.ipynb>

7.4 Synthetic tumour/normal WGS

Dataset Name: Synthetic tumour/normal WGS

Short dataset description: A synthetic tumour/normal cancer dataset for testing and demonstration.

Was it generated within EOSC4Cancer? or somewhere else (if so, please, provide us the project/link to the project website)? Yes, generated within EOSC4Cancer.

Data type (genomics, imaging, proteomics, structured clinical data, etc.): We generated 10 pairs (tumour and normal) synthetic whole genome sequencing samples (in a standard format of Illumina paired-end reads).

Do you follow any particular data model? Which one?: Yes, NEAT read simulator (version 3.0⁹) was utilised to synthesize these 10 pairs tumour and normal WGS samples. In the procedure of data generation, the simulated parameters (i.e., sequencing error statistics, read fragment length distribution and GC% coverage bias) were learned from data models provided by NEAT (see a description of these models in online documentation¹⁰).

For generation of normal WGS data per each sample, a germline variant profile of a real individual was down-sampled randomly, with inclusion of a subset of 50% germline variants, and then mixed together with an in-silico germline variant profile that was generated using a model of a mutation rate of 0.001, finally constituting a full germline profile for normal WGS data.

For the generation of tumor WGS data per each sample, a predefined somatic short variant profile (SNVs+Indels) learned from a real cancer patient was inserted into the dataset. Neither a copy number profile nor a structural variation profile was introduced in the generation of the tumor WGS data.

Do you impose any quality criteria? Which one?:

- 1) sequencing depth on the average: 110X for 10 tumour samples and 60X for normal samples.
- 2) read length: 147 bp for all datasets.
- 3) ploidy: 2 for all datasets.

Is a synthetic dataset generated for 1) technical purposes, e.g. evaluating a given technology, 2) as a proxy to one or more real datasets, 3) both, or 4) others (please, provide some explanation about it).: Datasets were generated for both technical and proxy purposes.

Has it been published or described somewhere? Can you share a link/DOI?: Aim for deposition in EGA/federate EGA, pending decision.

Is it publicly available? Can you share the link to it?: They will be publicly available from within EGA.

Is the methodology used to generate it publicly available in a repository? Can you share the link to it?: Yes, the tool NEAT (version 3.0) was utilised to generate these synthetic WGS samples, which is available on github¹¹.

⁹ <https://github.com/zstephens/neat-genreads>

¹⁰ <https://github.com/zstephens/neat-genreads/tree/master/models>

¹¹ <https://github.com/zstephens/neat-genreads>