



European
Genomic Data
Infrastructure

GDI Stakeholder Forum - Dataspaces

23rd November 2023



GDI project receives funding from the European Union's Digital Europe Programme under grant agreement number 101081813.



gdi.onemilliongenomes.eu/



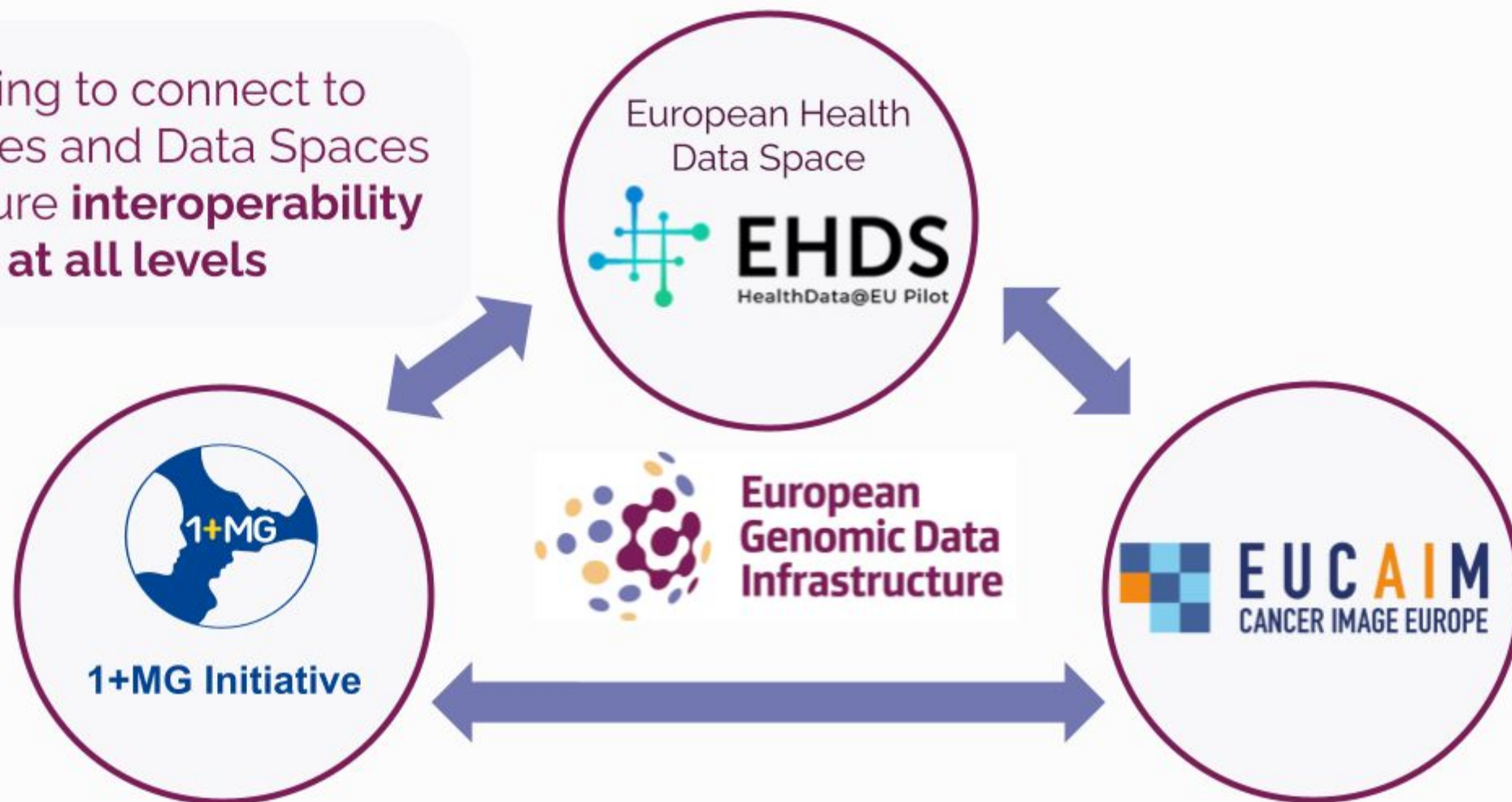
[@GDI_EUproject](https://twitter.com/GDI_EUproject)



[/company/gdi-euproject](https://www.linkedin.com/company/gdi-euproject)

Connecting to relevant EU Initiatives

Aiming to connect to initiatives and Data Spaces to ensure **interoperability at all levels**



European partnerships: Personalised Medicine, Rare Diseases; Digital Twins, Cancer mission



1+MG Initiative and GDI

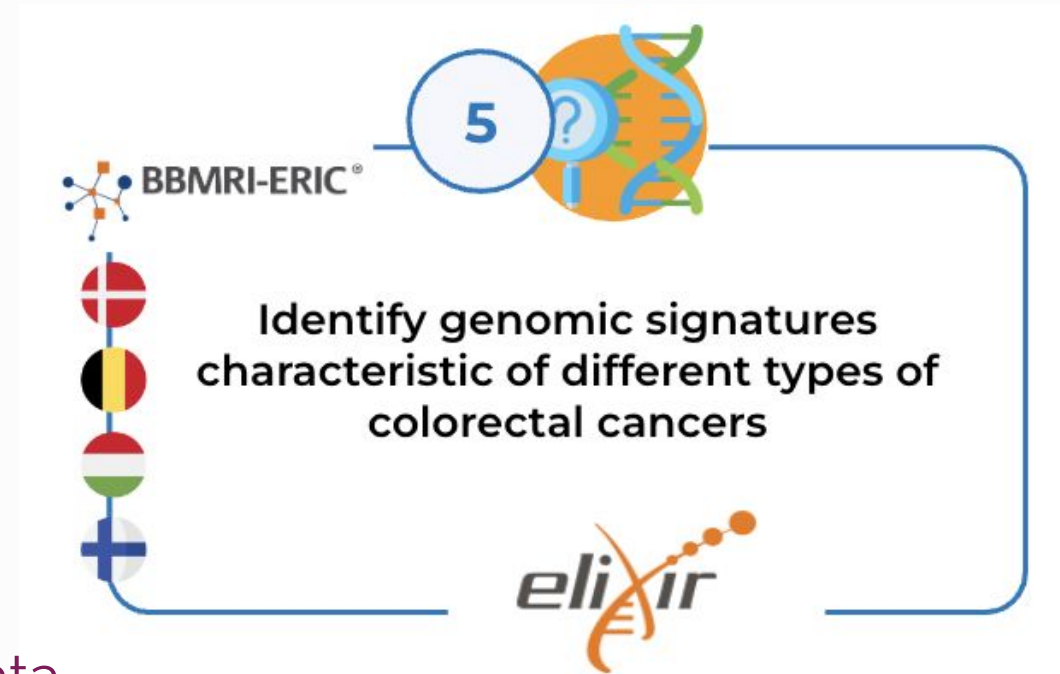
- GDI is realising the 1+MG Data Infrastructure required to provide access to high-quality genomic and health data across borders to make progress towards the practice of Personalised Medicine across Europe
- Building in the 1+MG Framework <https://framework.onemilliongenomes.eu/> representing the collective consensus of experts from 27 countries
 - ELSI
 - Data quality
 - Data models standards and ontologies
 - Technical infrastructure and open source reference implementation
 - Domain use cases: Cancer, Genome of Europe, infectious diseases, PGx, Rare diseases





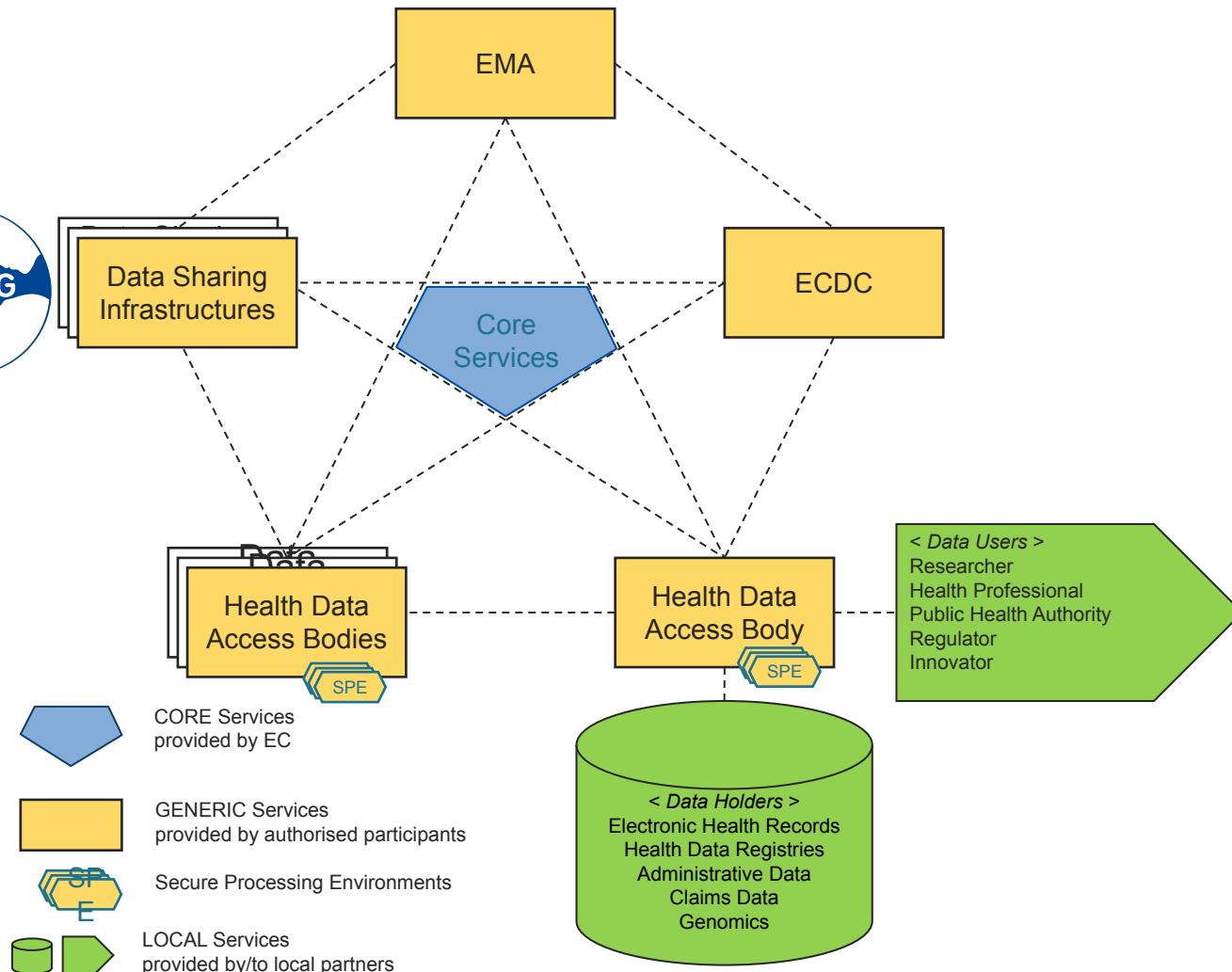
EHDS and GDI

- 1+MG CRC use cases included in Healthdata@EU pilot project
 - Accessing EHDS participants data
 - BE, DK, HU, NO
 - BBMRI-ERIC
 - WGS, Phenotypic and, socioeconomic data
- Demonstrate use of 1+MG/GDI standard to manage genomic and health data within the Healthdata@EU



HealthData@EU: vision and reference architecture

Slide developed by the European Commission (updated on 01/07/23)



- **New infrastructure for secondary uses of health data**
- Connecting health data access bodies and data sharing infrastructures
- Several health data access bodies are established, or in the process, across Member States



EHDS and GDI

- 1+MG use case on colorectal cancer leveraging data infrastructure participation in EHDS (ELIXIR & BBMRI)
- Common interests on crucial issues such as data discovery, data access, etc.
- GDI and Healthdata@EU pilot projects are complementary and synergetic projects for secondary use of health data. Close collaboration a key success factor for both projects
- **By delivering GDI we are also contributing to deliver EHDS**





EUCAIM and GDI

- Genomics, cancer image and phenotypic data
- Deployment and enhance of 1+MG/GDI standards and reference implementations
 - Beacon v2 for Image (DICOM)
 - Data discovery based on genomic, phenotypic or image metadata query
- Collaboration on enabling AI on image and genomic data
 - Developing federated learning approaches





Synergies and collaborations across projects and initiatives

Project / Initiative	Collaboration	Project / Initiative	Collaboration
GA4GH	Establishment of GDI as a driver project (on going). Observers in the GA.	DARWIN	Exploratory discussions took place
EHDS	Alignment at different levels with DG Sante e.g. data catalogue	HealthData@EU	1+MG CRC use case
DSSC	Sharing knowledge and experience on moving from design and testing to scale up	Australia Biocommons	Observer in GA
GAIA-X	Outreach to GAIA-X community	Genome Canada	Observer in GA
EUCAIM	Linking genomics and image data, expanding beacon discovery capacities, federated learning	HDRUK, Genomic England	Observer in GA
EUCAN	Support the management of Genomic Information	Swiss Institute of Bioinformatics	Observer in GA





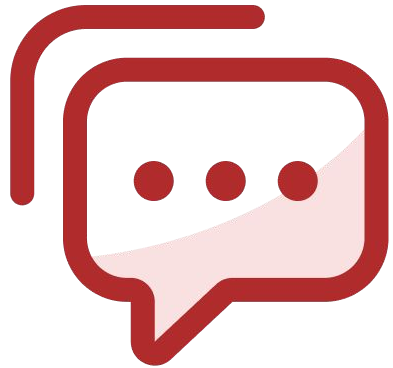
“ Thanks !

juan.arenas@elixir-europe.org





slido



Audience Q&A Session

① Click **Present with Slido** or install our [Chrome extension](#) to show live Q&A while presenting.





European
Genomic Data
Infrastructure

GDI Stakeholder Forum - Progress

23rd November 2023



GDI project receives funding from the European Union's Digital Europe Programme under grant agreement number 101081813.



gdi.onemilliongenomes.eu/



[@GDI_EUproject](https://twitter.com/GDI_EUproject)



[/company/gdi-euproject](https://www.linkedin.com/company/gdi-euproject)



1+MG data infrastructure

- Infrastructure has been demonstrated PoCs in rare disease and cancer (B1MG)
- Countries have committed to reach a target phase in European Genomic Data Infrastructure by 2026

Onboarding Addressing national challenges for deployment	Croatia, Cyprus, Hungary, Ireland, Malta, Romania
Deployment National node operational but not connected to EU level operations	Bulgaria, Latvia, Lithuania
Operational Fully deployed national node interconnected to EU services, able to provide access to data across borders following 1+MG Governance	Belgium, Czech Republic, Denmark, Estonia, Finland, France, Germany, Italy, Luxembourg, Portugal, Slovenia, Spain, Sweden, The Netherlands, Norway





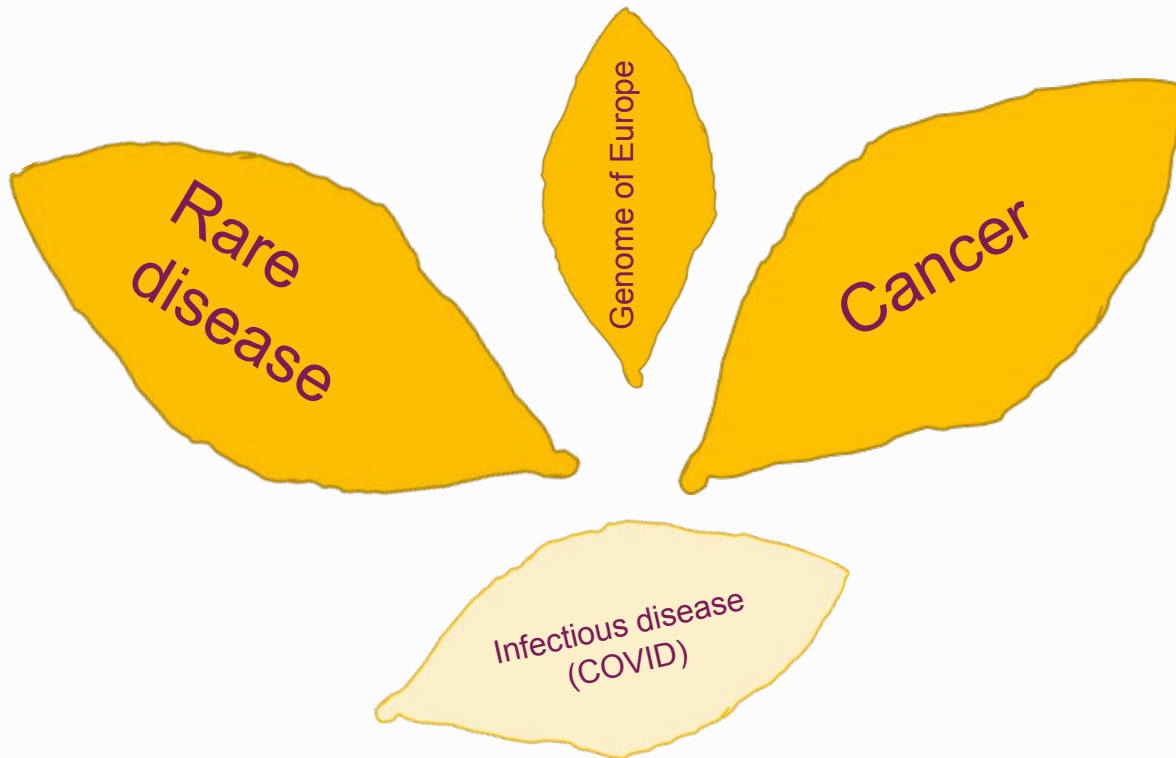
1+MG Long term sustainability

- 18 countries submitted the EoI for the establishment of the 1+MG EDIC
- 9 countries submitted the pre-notification (due to time constraints)
- EDIC WG established and supported from GDI Pillar I and MCP Accelerator
- Research data governance workflow endorsed by GDI Pillar I





1+MG minimal datasets: status



- Rare disease: 16 agreed upon Common Data Elements (CDEs)
- Cancer: 140 items (37 mandatory, 40 recommended and 63 optional) organised in eight conceptual domains
- GoE: 3 mandatory items
- Infectious disease: work in progress (24 -> 27 items)





Synergies and collaborations across projects and initiatives

Project / Initiative	Collaboration	Project / Initiative	Collaboration
GA4GH	Establishment of GDI as a driver project (on going). Observers in the GA.	DARWIN	Exploratory discussions took place
EHDS	Alignment at different levels with DG Sante e.g. data catalogue	HealthData@EU	1+MG CRC use case
DSSC	Sharing knowledge and experience on moving from design and testing to scale up	Australia Biocommons	Observer in GA
GAIA-X	Outreach to GAIA-X community	Genome Canada	Observer in GA
EUCAIM	Linking genomics and image data, expanding beacon discovery capacities, federated learning	HDRUK, Genomic England	Observer in GA
EUCAN	Support the management of Genomic Information	Swiss Institute of Bioinformatics	Observer in GA





Enhanced influence and value: GDI as a GA4GH Driver Project

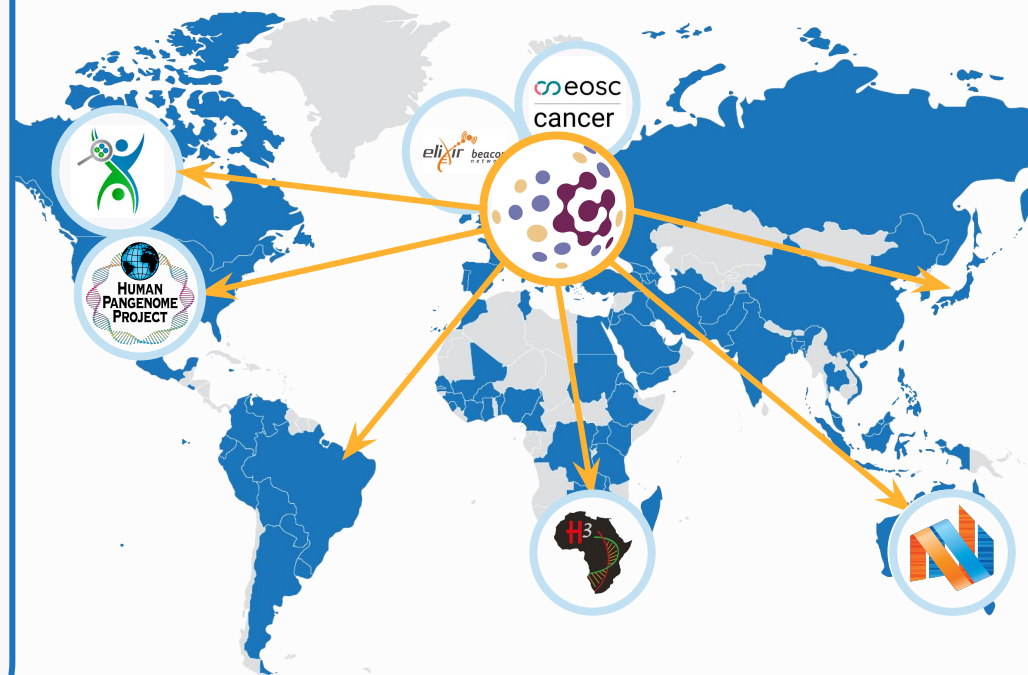


Global Alliance
for Genomics & Health

- **Increased global influence:** European experts collaboratively developing global standards
- **GDI leading the way** with the first end-to-end demo of international data sharing based on a framework GA4GH standards

THE GLOBAL REACH OF GA4GH

● GA4GH Driver Projects are active in 113 countries



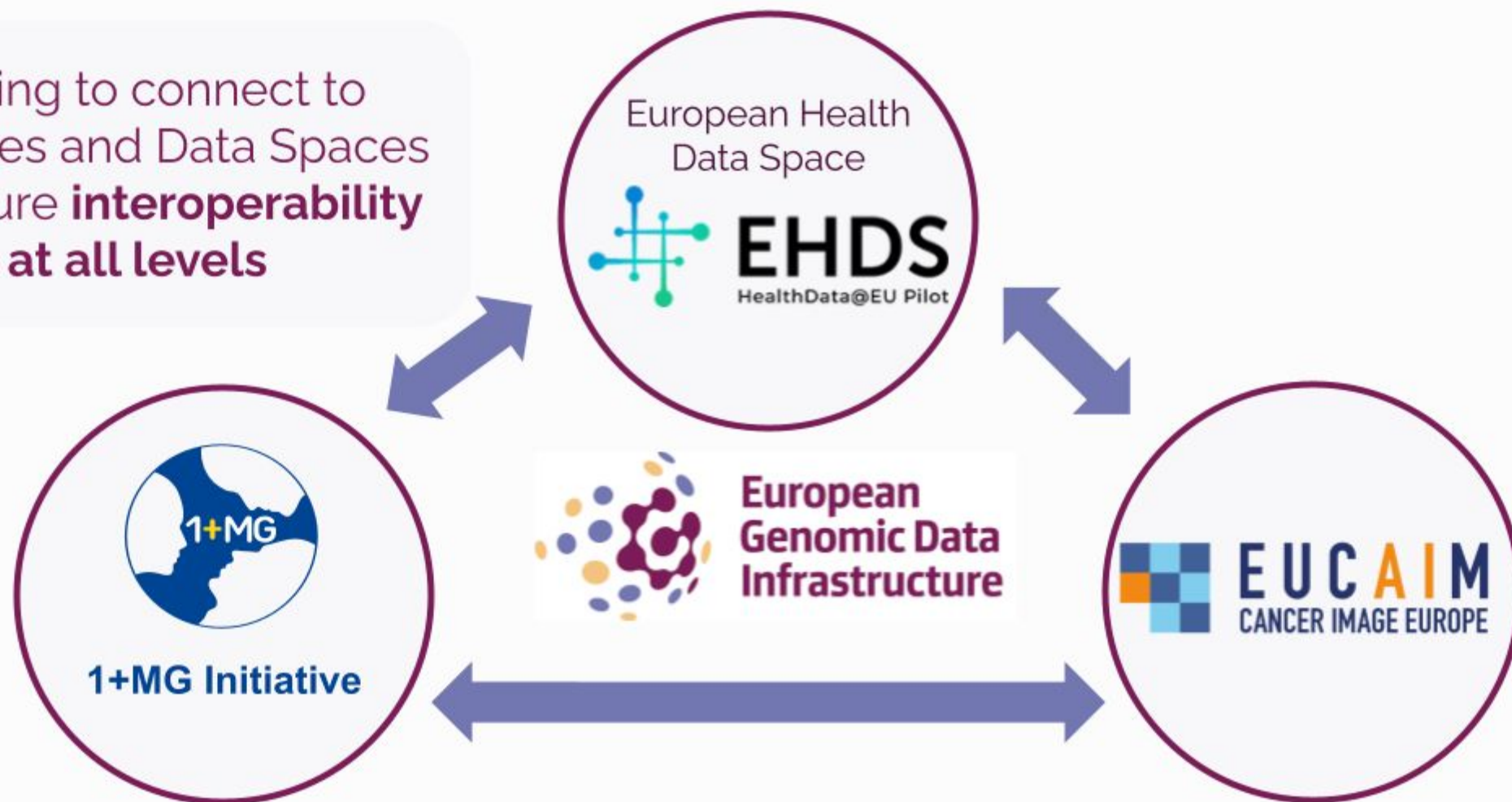
- **Global reference:** Canada & Australia seeking to align with GDI via Pillar II solutions built on GA4GH standards
- **Enhances the value** of our infrastructure by connecting Europe globally



GDI project receives funding from the European Union's Digital Europe Programme under grant agreement number 101081813.

Connecting to relevant EU Initiatives

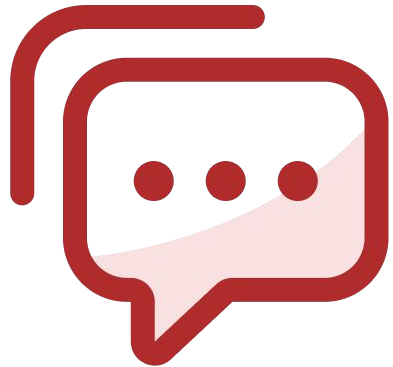
Aiming to connect to initiatives and Data Spaces to ensure **interoperability at all levels**



European partnerships: Personalised Medicine, Rare Diseases; Digital Twins, Cancer mission



slido



Audience Q&A Session

① Click **Present with Slido** or install our [Chrome extension](#) to show live Q&A while presenting.





European
Genomic Data
Infrastructure

GDI Stakeholder Forum

Giselle Kerry - ELIXIR Hub

23/11/2023



GDI project receives funding from the European Union's Digital Europe Programme under grant agreement number 101081813.



gdi.onemilliongenomes.eu/



[@GDI_EUproject](https://twitter.com/GDI_EUproject)



[/company/gdi-euproject](https://www.linkedin.com/company/gdi-euproject)



Gaps - Innovation and Data Technologies





Application and innovation solutions

- Provide **realistic disease scenarios** for the implementation and testing of **federated analysis and federated learning technologies**
- Contribute to interoperability of **genomics and health data**
- **Monitor the technical development of the project**, in the search for **gaps, missing elements and technologies** that required adaptation or could be ported to other environments as they constitute opportunities for innovation





Federated/distributed analysis

- **3 technological demonstrators** on federated/distributed analysis
 - Federated/distributed analysis of infectious diseases (pks) data
 - Data mobilization across analysis platform (Cancer use-case)
 - 5-safes RO-Crate compatible Trusted Research Environment (HUTCH + WfExS)
- Evaluate the alignment of the federated analysis technologies with the **data governance workflows** to identify gaps and define possible implementations scenarios.
- Work towards evaluation of technology readiness.





Solving gaps will require...

- Collaborative efforts (researchers, clinicians, policy makers and technology developers)
 - Continued innovation in both genomics and technology to unlock the full potential of human genomics - for improving both scientific knowledge and human healthcare
- We want industrial partners to contribute and/or benefit from the GDI implementations/project outputs





Interaction with Industry





GDI: Interfaces with Industry

Pillar I: Sustainability

- Business models
- Creating flexibility & structure for countries to use different innovative solutions

Make it last

Pillar II: Technical infrastructure

- Finding gaps in core infrastructure
- Open to innovative solutions: REMS

Make it work

Pillar III: Applications and innovation solutions

- Proposing Use Case-based challenges to the innovative industries
- Looking to the future: making sure the infrastructure can handle next-gen technology

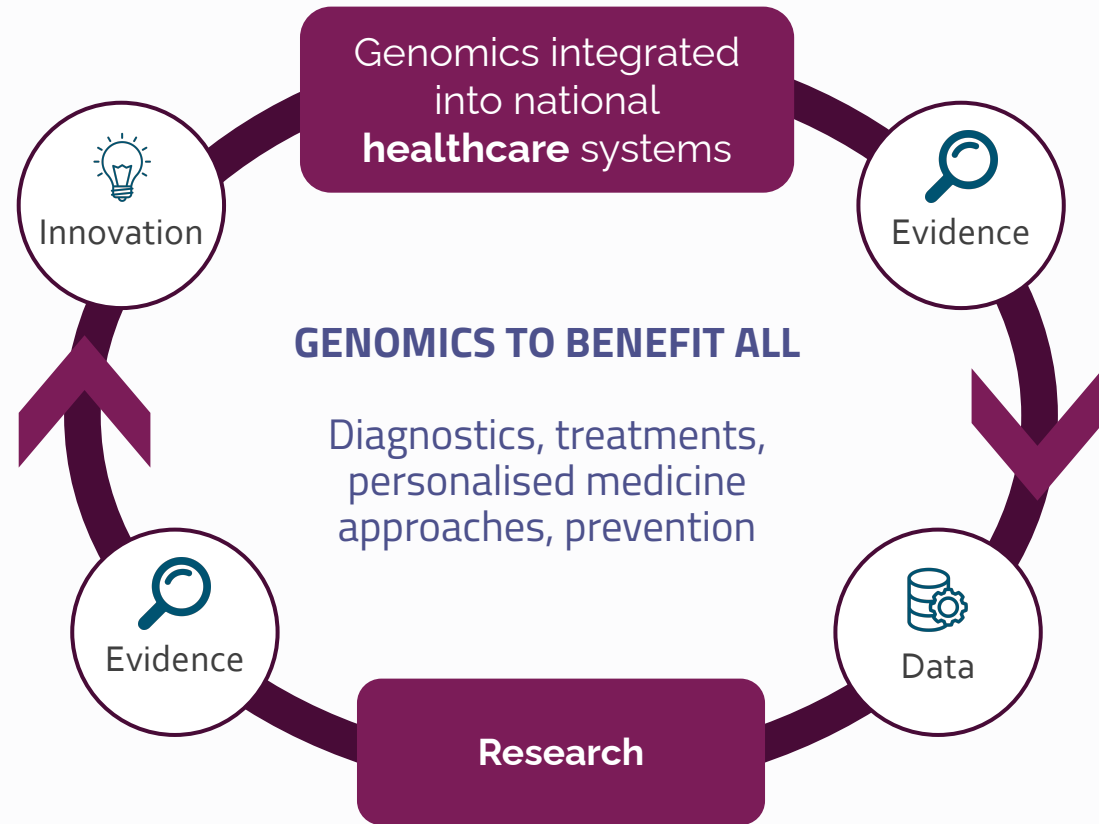
Make it useful

Influencing the genomic medicine ecosystem

GDI as a project works on part of the big picture

Industry required to fill the gaps in the genomic medicine 'infinity loop' to ensure it rotates effectively

Industry may also want to access data or contribute to generating data



Genomics Industry Landscape Exercise

Summary of current respondents of the industry mapping exercise in human genomics



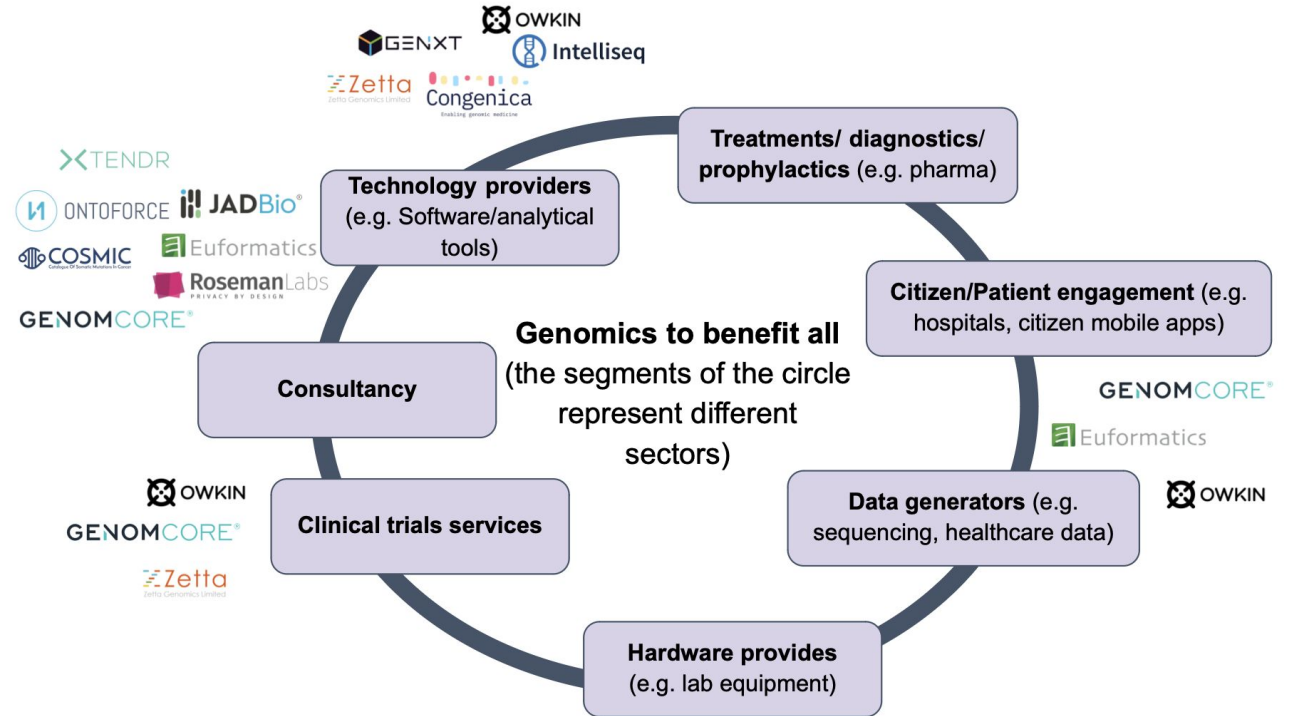
Add your company's services

Add your opinion

Agree to publish the information provided

Purpose

ELIXIR would like to map out the industrial ecosystem in genomics and the resources the companies offer.



Companies shared their views on what they think is missing from the ecosystem

Need for skills:

- Computational capabilities and skilled workforce to handle genomic data.
- Training opportunities for clinicians and other stakeholders to integrate genomics data successfully
- Training and funding on collaborative AI projects & ML analysis standards

Need for harmonized standards:

- Harmonized data sharing, legal, and ethical frameworks across geographies.
- Encouragement for the adoption of standards for data interoperability, sharing, and quality control.

Need for awareness:

- More communication with clinicians of actionable data.
- Public engagement for awareness of patient benefit and application of genomics in frontline care.

Need for best practices respecting privacy rules:

- Slow adoption of novel privacy-enhancing technologies.
 - Legal concerns, data privacy, data security - lack of trust and blocked collaborations, even if privacy assurances in place.
 - Furthering research into leading data privacy topics.
-

Genomics Industry Landscape Exercise

1
STEP



2
STEP



3
STEP

Add your company's
services

Add your opinion

Agree to publish the
information provided

Contact

Despoina Sousoni (ELIXIR Industry and Innovation officer;
despoina.sousoni@elixir-europe.org) if you have any questions.

Purpose

ELIXIR would like to map out the industrial ecosystem in genomics and the resources the companies offer.

How to get involved

Join the Miro board:

<https://miro.com/app/board/uXjVM-DkWmg=?moveToWidget=3458764565652264717&cot=14>





Breakout session





Breakout Rooms on Innovation Gaps

Topic: How can **industry** contribute in the genomics ecosystem to support the 1+Million Genomes (1+MG) initiative ambition to enable secure access to high-quality genomics and the corresponding clinical data across Europe for better research, personalised healthcare and health policy making

Discussion point 1

How can industry help to mobilise genomic and health data from healthcare/a clinical setting into the European Genomic Data Infrastructure for secondary use?

Discussion point 2

How industry can facilitate the analysis of data - federated analytics/learning to generate evidence

Discussion point 3

How can industry bridge the innovation gap between research results and a healthcare setting based on the data made available by 1+MG/GDI?





GDI project receives funding from the European Union's Digital Europe Programme under grant agreement number 101081813.



European
Genomic Data
Infrastructure

GDI Stakeholder Forum

23rd November 2023



GDI project receives funding from the European Union's Digital Europe Programme under grant agreement number 101081813.



gdi.onemilliongenomes.eu/



[@GDI_EUproject](https://twitter.com/GDI_EUproject)



[/company/gdi-euproject](https://www.linkedin.com/company/gdi-euproject/)



Welcome to the GDI Stakeholder Forum!

Thank you for joining the meeting
We will start promptly at 12:00 CET

Please ensure that you have named yourself correctly

Your full name (stakeholder type)

E.g. Nikki Coutts (GDI Coordination)

E.g. John Smith (SME)

To do this, click on the 'Participants' icon, hover over your name until the 'More' box appears, then click on 'Rename'





This meeting is being recorded FOR MINUTING PURPOSES ONLY and will be uploaded to the GDI project [folder](#)



We follow the [ELIXIR Code of Conduct](#) at all our events. For more info visit the ELIXIR Europe website



Please send in your questions for speakers through Slido Q&A: GDI-SF2023





What can you expect?

Project information, updates, and outcomes:
deploying European Genomic Data
Infrastructure

Discussion on how industry could provide
innovative solutions to some of the gaps and
challenges we face





Agenda

12:00 -17:00 (CET)

Introduction to the GDI project and achievements to date

Coffee break

**Building an ecosystem that boosts innovation in
genomics and healthcare research.**

Industry involvement, innovation, gaps and challenges

Lunch break

Innovation gaps and industry

Breakout session





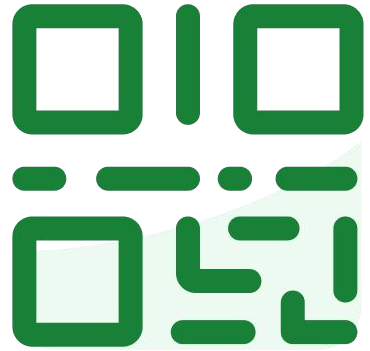
Breakout

How can industry contribute in the genomics ecosystem to support the 1+Million Genomes (1+MG) initiative ambition to enable secure access to high-quality genomics and the corresponding clinical data across Europe for better research, personalised healthcare and health policy making





slido



Join at slido.com
#GDI-SF23



Click **Present with Slido** or install our [Chrome extension](#) to display joining instructions for participants while presenting.



GDI project receives funding from the European Union's Digital Europe Programme under grant agreement number 101081813.



slido



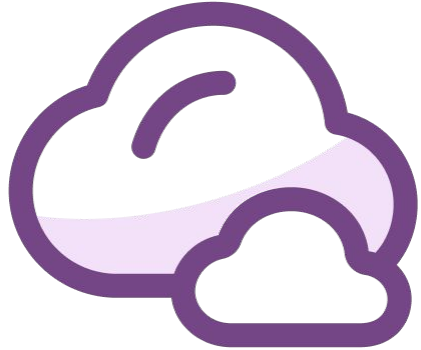
Which stakeholder group do you most associate yourself with?

ⓘ Click **Present with Slido** or install our [Chrome extension](#) to activate this poll while presenting.





slido



What is your field of expertise? (you may submit multiple answers)

ⓘ Click **Present with Slido** or install our [Chrome extension](#) to activate this poll while presenting.



GDI project receives funding from the European Union's Digital Europe Programme under grant agreement number 101081813.



slido



Which country do you represent?

ⓘ Click **Present with Slido** or install our [Chrome extension](#) to activate this poll while presenting.



GDI project receives funding from the European Union's Digital Europe Programme under grant agreement number 101081813.



slido



If you are from industry, which sector would you say you represent (e.g. SME, Pharma, etc)?

ⓘ Click **Present with Slido** or install our [Chrome extension](#) to activate this poll while presenting.



GDI project receives funding from the European Union's Digital Europe Programme under grant agreement number 101081813.



slido



Which project/initiative/company are you associated with?

ⓘ Click **Present with Slido** or install our [Chrome extension](#) to activate this poll while presenting.





slido



If you are associated with a project, please could you let us know which one?

ⓘ Click **Present with Slido** or install our [Chrome extension](#) to activate this poll while presenting.



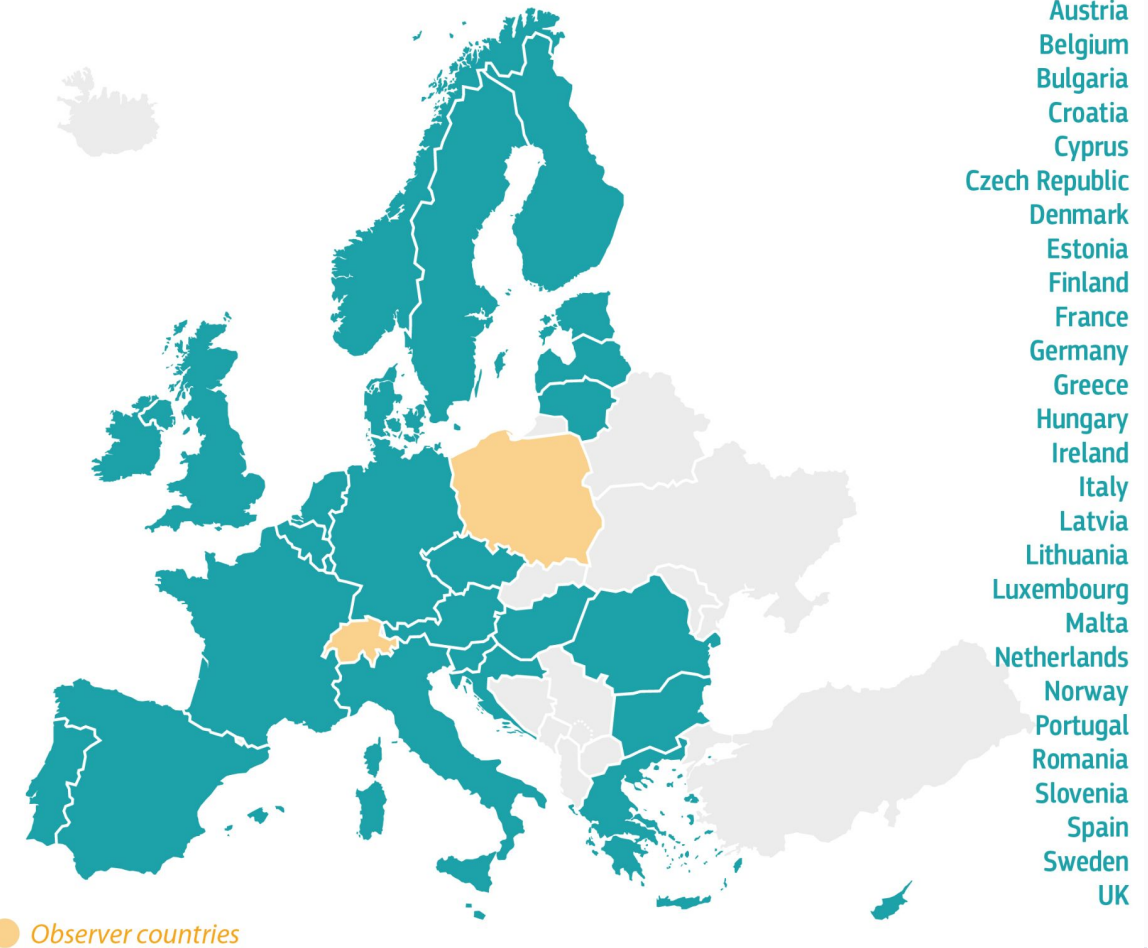


GDI project overview



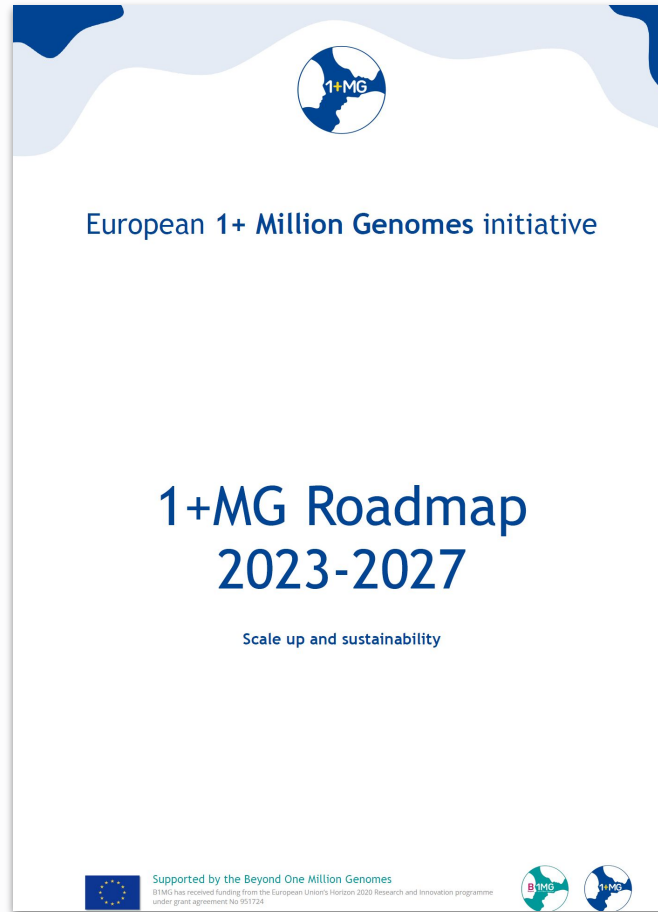
Cross-border access to genomic data, implementation of genomics-based health. Supported by 27 countries to date

Supports the 1+Million Genomes (1+MG) initiative ambition to enable secure access to high-quality genomics and the corresponding clinical data across Europe for better research, personalised healthcare and health policy making





1+MG Roadmap



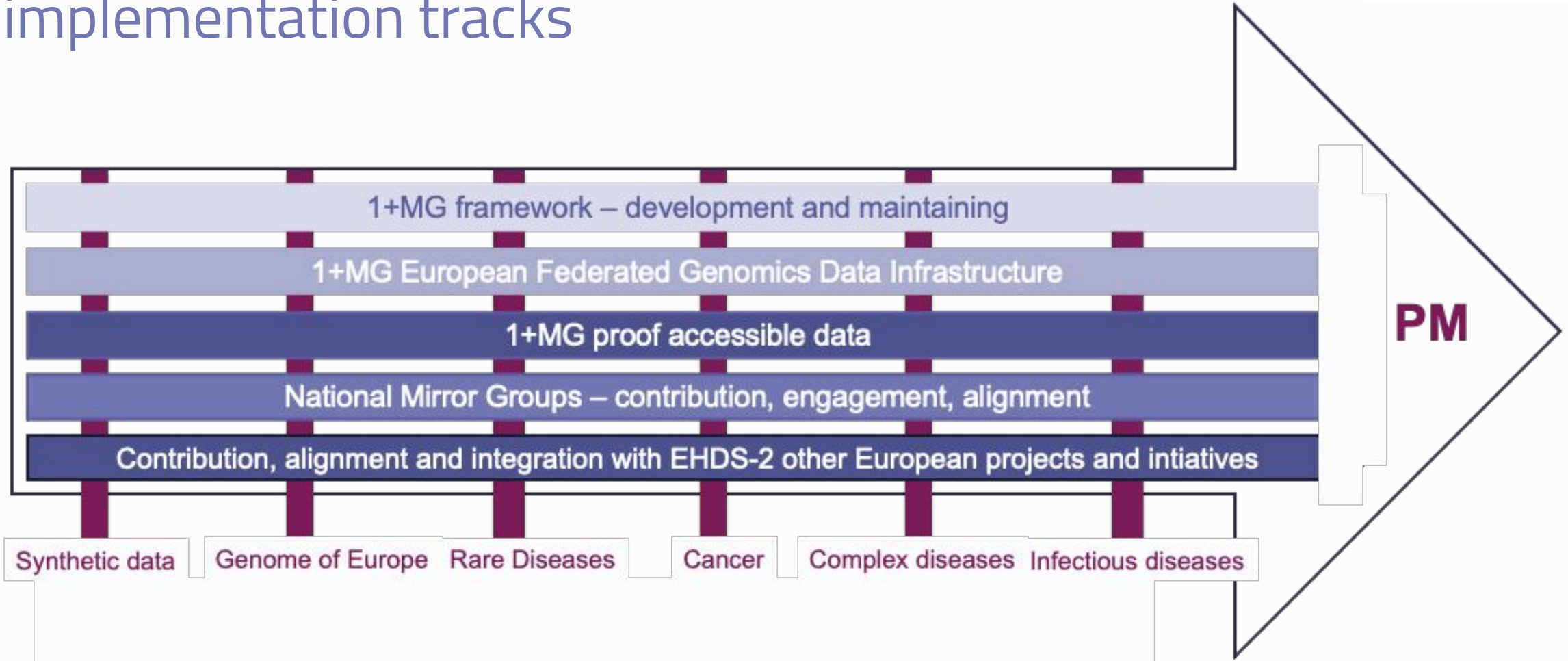
<https://ec.europa.eu/newsroom/dae/redirection/document/99974>





1+MG Roadmap 2023-2027

5 implementation tracks





Journey to the 1+MG Data Infrastructure



Cross-border access to genomic data, implementation of genomics-based health. Supported by 27 countries to date

2018

2020

2022

2023

2026

2027

Design & Testing

Scale-up & Sustainability



European Genomic Data Infrastructure



Population Genomics

Genome of Europe



GDI project receives funding from the European Union's Digital Europe Programme under grant agreement number 101081813.



GDI project overview

Scale-up & Sustainability



European Genomic Data Infrastructure

Establishes a federated, sustainable and secure infrastructure based on open community standards to access genomic and related phenotypic and clinical data across Europe

Builds on the Beyond 1 Million Genomes (B1MG) project outputs





GDI in numbers

€40M (50% co-funded)

45 Project Beneficiaries

6 Affiliated Entities

7 Associated Partners

58 Total Project Partners

20 European countries

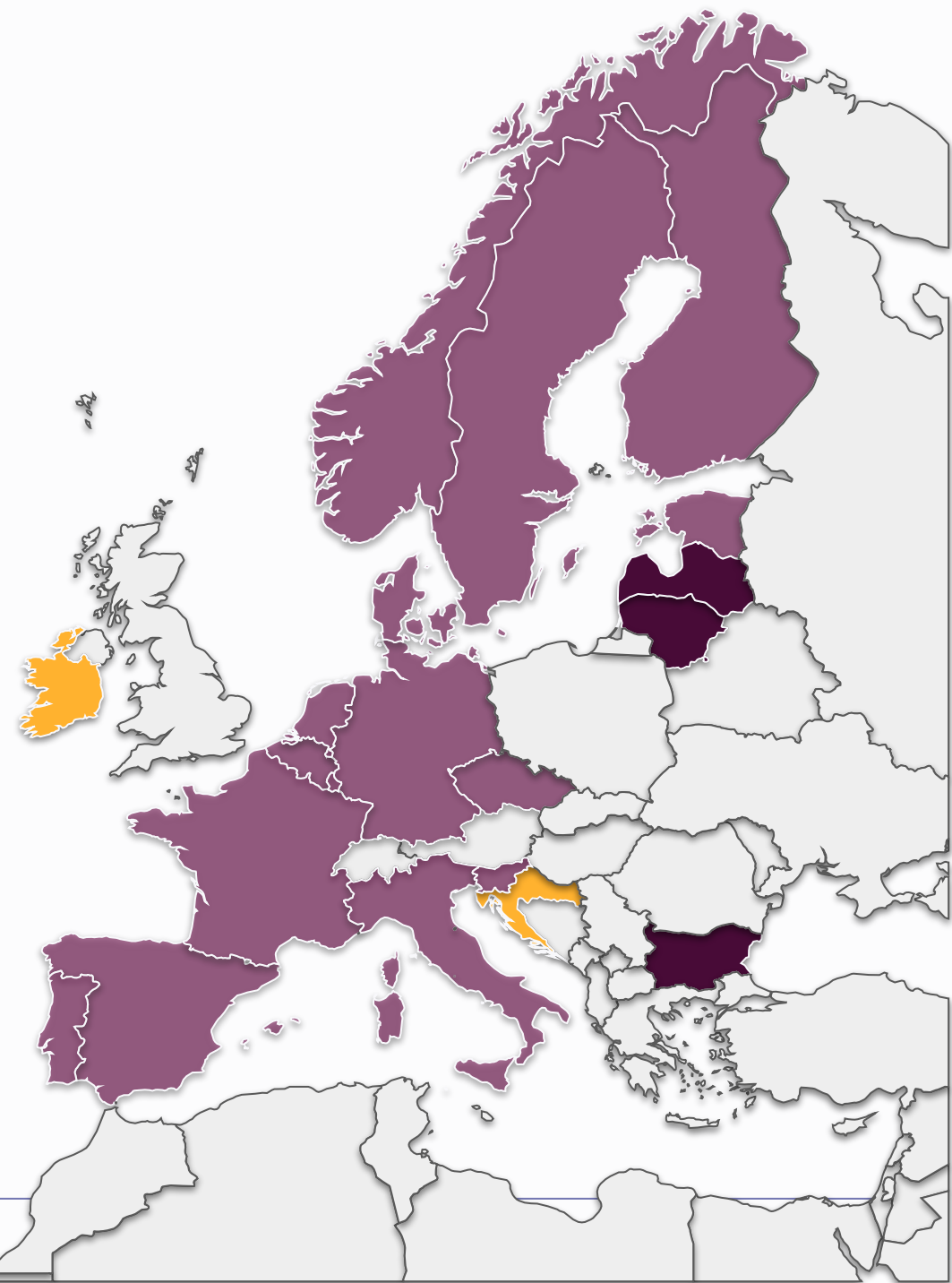
At least **6** countries technically operational by **2024**

At least **15** countries technically operational by **2026**

2 International Organisations*

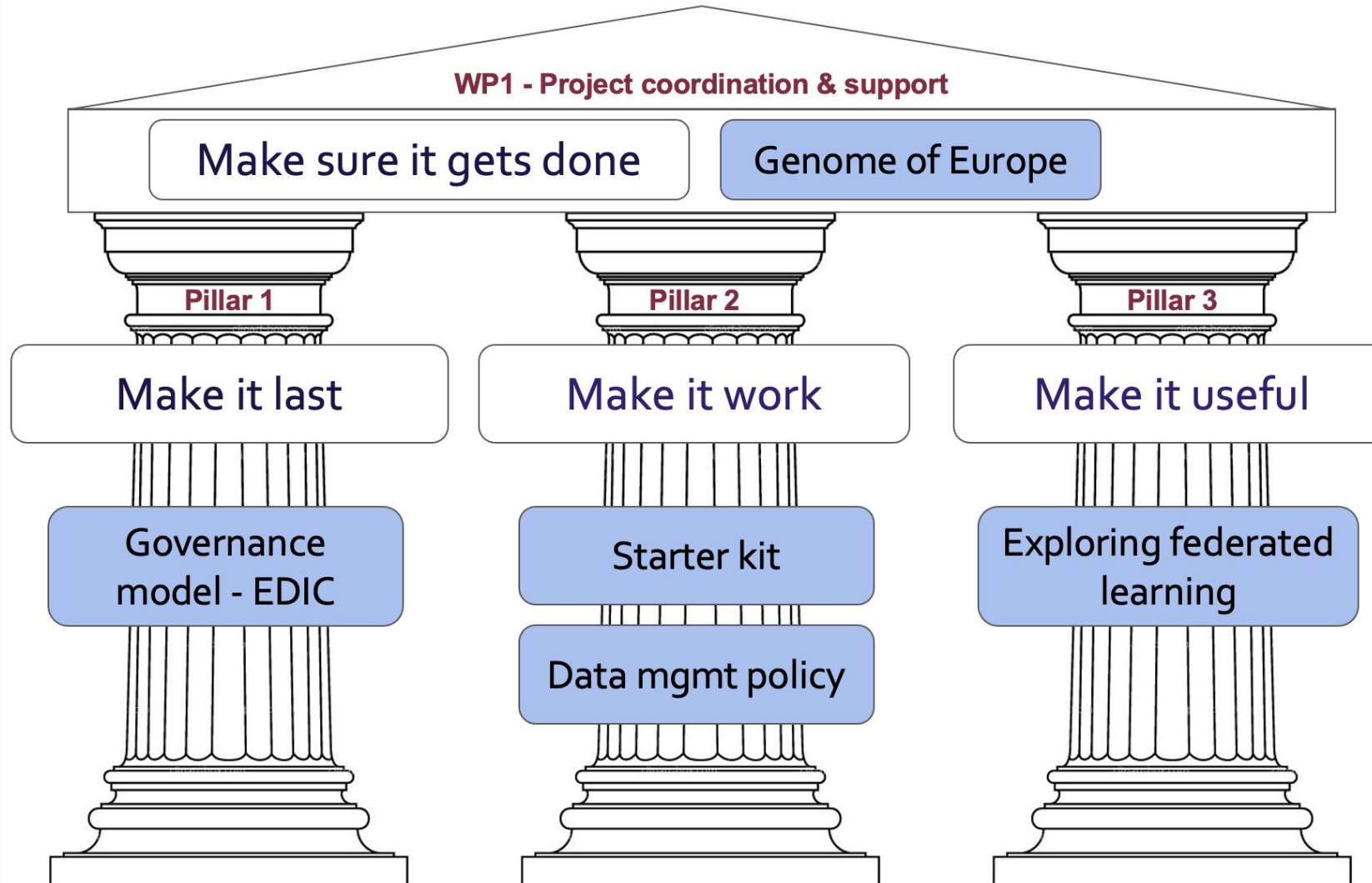
*EMBL and BBMRI-ERIC

Cyprus, Hungary, Malta and Romania joining GDI soon





GDI Project overview





1+MG Framework

Brings together the specific recommendations, guidelines and best practices that can be use to realise the 1+MG vision encompassing:



Data governance & ELSI



Implementation of Genomics into Healthcare



Data models and data quality



Use Cases



Technical Infrastructure



National implementation

Plus alignment with other key European projects such as EHDS
For all stakeholders to use



framework.onemilliongenomes.eu





1+MG Framework

<https://framework.onemilliongenomes.eu/>



The 1+MG Framework is a series of components based on the output of the 1+MG projects that provide guidance on ELSI, data quality, data standards, and technical infrastructure standards and APIs.

Core 1+MG Framework

Technical framework

- Sequencing guidelines**
Sequence data generation and quality requirements for WGS/WES data to be labelled as 1+MG compliant
- Data models, standards & ontologies**
1+MG minimal data models for different use cases and recommendations on ontologies and data standards
- Technical Infrastructure**
Stack of standards, open source references implementations, synthetic data and proof of concepts that can be used to establish a 1+MG node.

Implementation

- Data governance and ELSI**
Guidance and recommendations on how to address data governance and ELSI aspect to ensure data can be made available.
- Genomics into healthcare**
Assessment Maturity Level Model to guide healthcare systems on their journey to implement genomic medicine.
- National implementation**
Find pointers to country specific information resources and national research data management practices.

Applying the Framework

- EHDS Alignment**
Outlining how the elements of the 1+MG Framework align with the European Health Data Space User Journey and Data Life Cycle
- Use cases**
Learn about data management tasks that affect your domain or research community, and the solutions adopted to address them.

1+MG Framework Editorial Guidelines

The content is approved and maintained by working group leads or their deputies. Requests for changes or additions to the core content are open to all members of the consortium. National Implementations can be contributed by all member states. For more detail, please click here.

[Start contributing](#)

• About
• Contact
• Accessibility
• Privacy

B1MG has been received funding from European Union's Horizon 2020 Research and Innovation programme under grant agreement No. 951724. The development of the 1+MG Framework has been funded by B1MG.

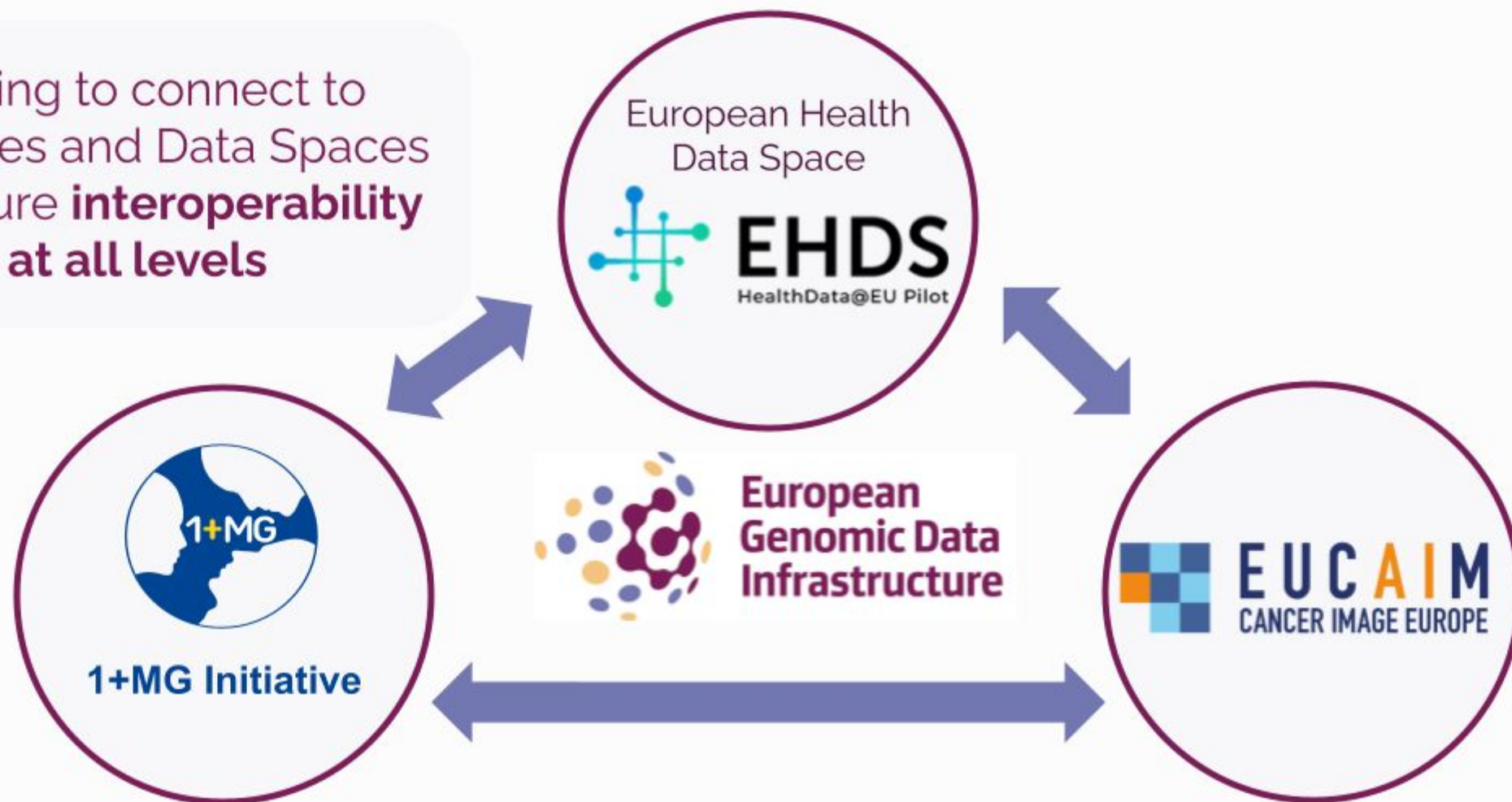
The 1+MG framework is licensed under a [Creative Commons Attribution 4.0 International License](#), except where otherwise noted.

Built with [FTI](#)



Connecting to relevant EU Initiatives

Aiming to connect to initiatives and Data Spaces to ensure **interoperability at all levels**



European partnerships: Personalised Medicine, Rare Diseases; Digital Twins, Cancer mission



European Genomic Data Infrastructure



@GDI_EUproject



/company/gdi-euproject



GDI website



GDI project receives funding from the European Union's Digital Europe Programme under grant agreement number 101081813.



GDI project receives funding from the European Union's Digital Europe Programme under grant agreement number 101081813.



Building an ecosystem that boosts innovation in genomics and healthcare research.

- Use cases and federated analysis

Salvador Capella-Gutierrez, BSC





European Genomic Data Infrastructure



GDI project receives funding from the European Union's Digital Europe Programme under grant agreement number 101081813.



gdi.onemilliongenomes.eu/



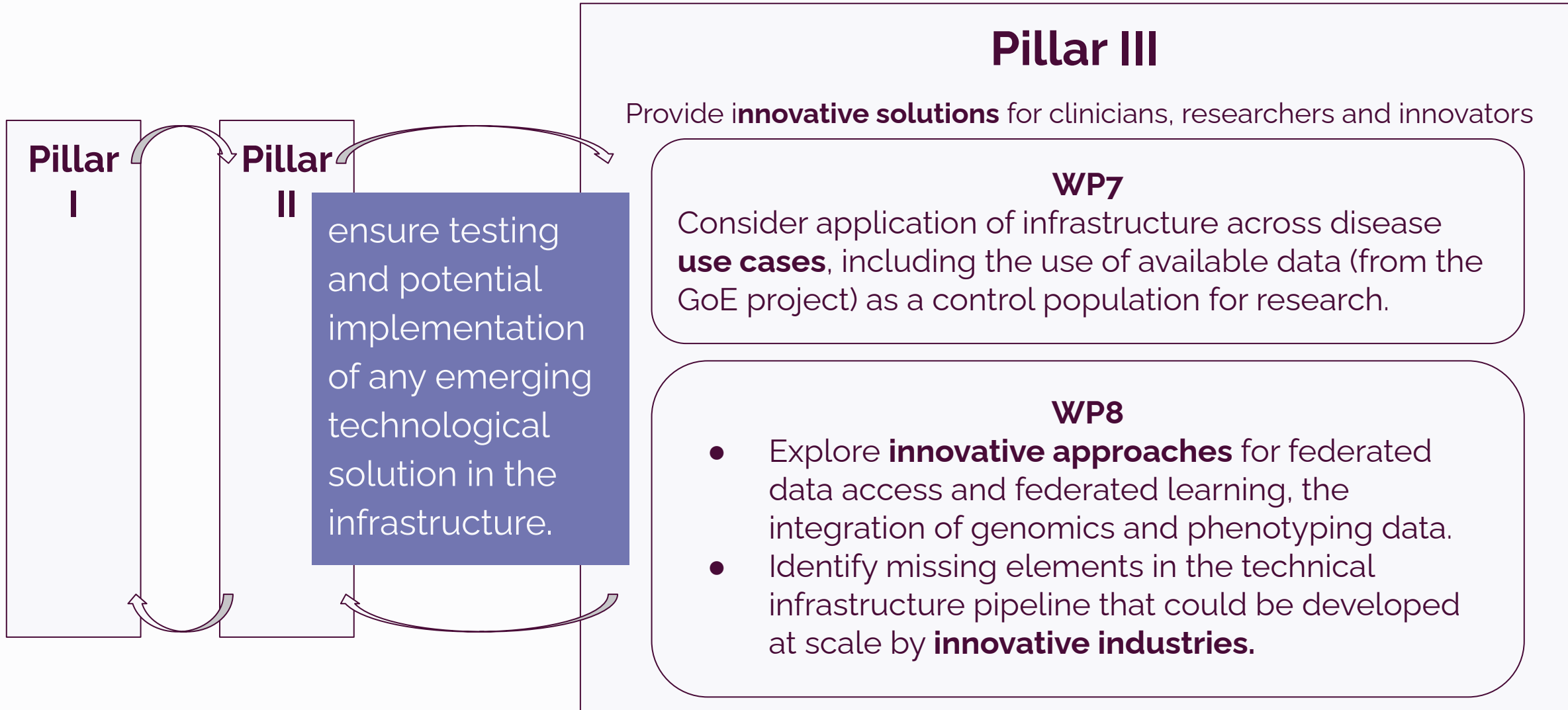
[@GDI_EUproject](https://twitter.com/GDI_EUproject)



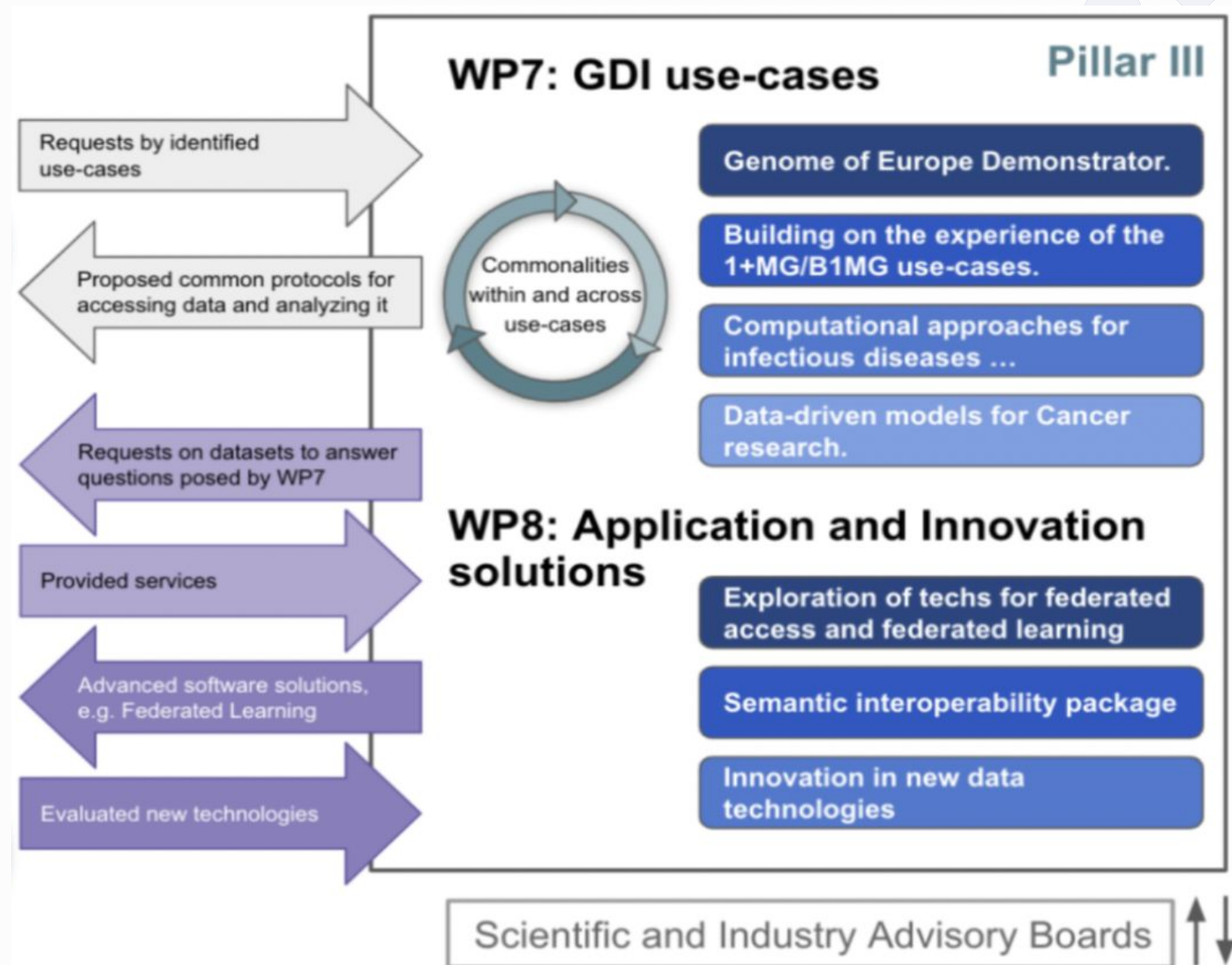
[/company/gdi-euproject](https://www.linkedin.com/company/gdi-euproject)



Introduction to Pillar III



Pillar III





<https://www.elespanol.com/>



Independent, nuclear use-cases ...

... able to run on common tracks



<https://www.siemens.com>



GDI project receives funding from the European Union's Digital Europe Programme under grant agreement number 101081813.



Use Cases (WP7)

- **Develop a selected set of disease-related use-cases for channelling questions of concern from European and member state needs, which**
 - i) represent transnational scenarios,
 - ii) are generalisable to other diseases, and
 - iii) are addressable via infrastructure developments.
- **Catalyse and consolidate the developments of the 1+MG and other European and national initiatives via the interactions with pillar II**, leading to the implementation and testing in the specific disease use-cases.
- **Provide guidance and instruments to EU projects and MSs** for the implementation of the technical developments across EU countries in a scalable, generalisable, and ELSI-compliant manner.





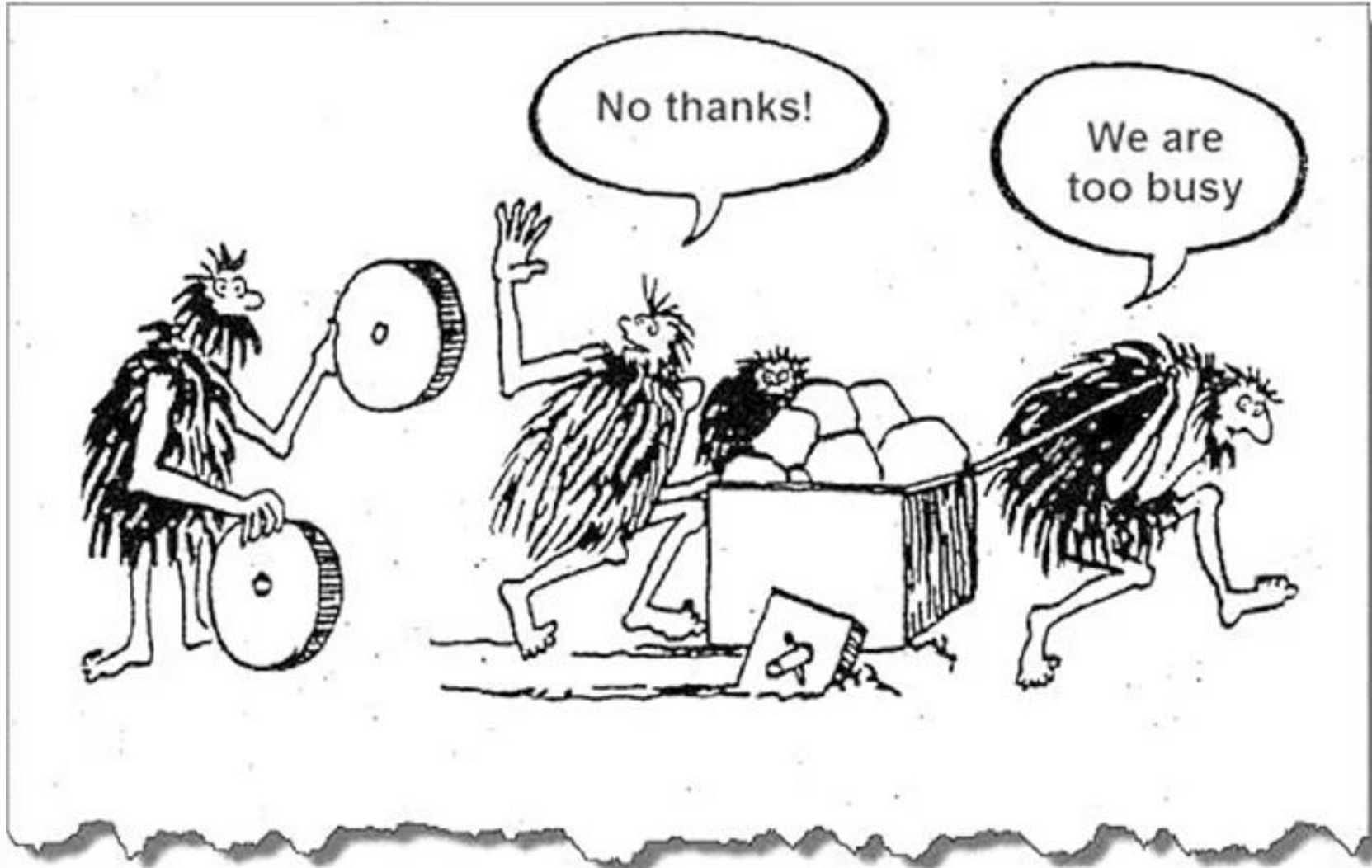
Application and innovation solutions (WP8)

- Provide realistic disease scenarios for the implementation and testing of **federated data and federated learning technologies** based on previous developments, in particular those concerning the use of Federated EGA technologies.
- Contribute to the consolidation and generalisation of the efforts of the 1+MG and related initiatives in the area of interoperability of **genomics and health data**.
- Become a European industry reference point, in particular SMEs, related to the **innovation in new technologies** for their possible implementation in the different health systems.





<https://www.creativityatwork.com/busy-innovate/>





WP7: Conceptual pipeline



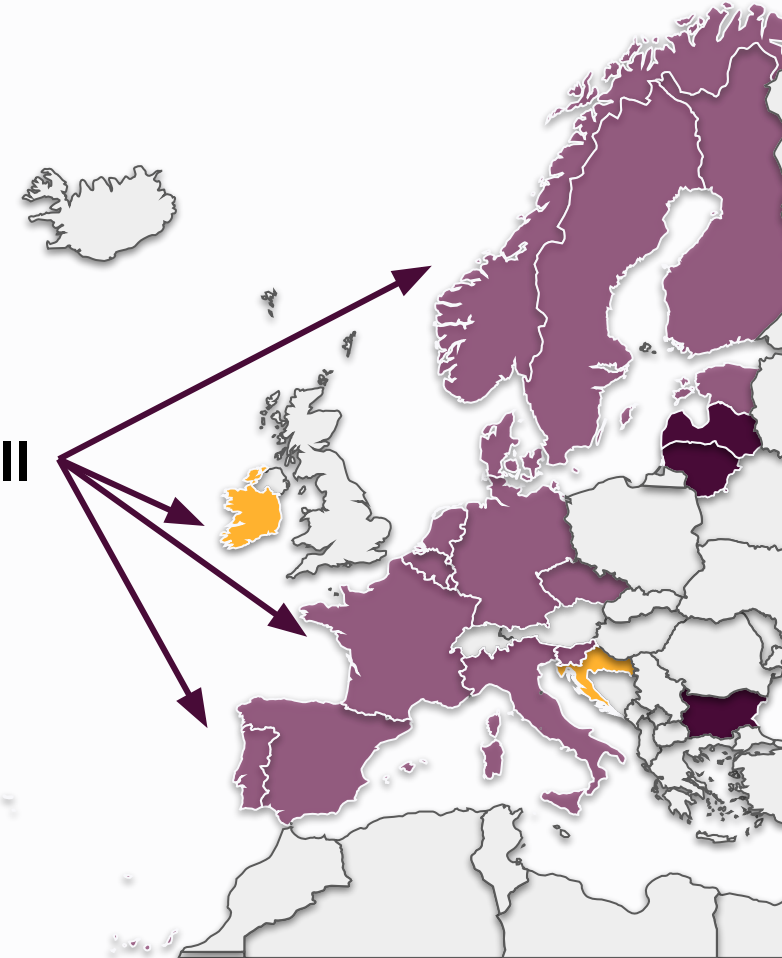
Different maturity level
per each use-case

Prototypical
Questions

Technological
Requirements

WP8

Pillar II





WP7: GDI Use Cases

Translation into technical requirements

For each of the questions, the following aspects are included:



Data reception (data preparation and inclusion)



Storage and interfaces (data storage and management)



Data discovery



Data access management tools



Data

processing

Also, the following questions are answered:

- Do you need to access data transnationally? Are you considering a federated system? Or something else?
- Are you already engaged with projects/efforts/initiatives beyond 1+MG/B1MG/GDI that are developing infrastructures for tackling your needs?
- Are you aware of any other use case requirements relevant for your use case?
- Are you using structured pheno-clinical data?





WP7: GDI Use Cases

Genome of Europe



- Bring together different aspects on the implementation of the resources and information on genomic data obtained by the 1+MG and the GoE projects.
- Three specific **research questions** relevant to the Genome of Europe demonstrator were collected through **ad hoc** meetings with 1+MG WG12 population genomics leadership involved in task 7.1.
 - i) Lookup of individual genetic variants
 - ii) Recalibration of polygenic risk scores
 - iii) Ancestry-specific imputation
- The Genome of Europe has been included in a **Digital Europe work-program**, with budget made available to collect genome sequencing data for this project. The program specifically includes the utilisation of GDI solutions to reach the goals of the Genome of Europe project.





WP7: GDI Use Cases

Building on the experiences of the 1+MG/B1MG use-cases



- Ensure all **lessons learnt** in the context of the 1+MG/B1MG are taken up in this project. Build the links with the use-cases on Rare diseases (WG8) and Common complex diseases (WG10).
- Take on board the experience and developments of the two **1+MG/B1MG PoC demonstrators** in rare diseases to implement to other use-cases. Concretely, this has started for the cancer use case.
- Two generic research questions on common complex diseases and pharmacogenomics were formulated for MS26.

- i) Why do people with certain disease-specific genes not develop the disease?
- ii) Why do some gene variants cause adverse side effects for medications?





WP7: GDI Use Cases

Infectious diseases with an initial focus in COVID-19



- Build on the **massive developments achieved during the COVID-19 pandemic**, including those of the 1+MG/B1MG use-case (WG11 leadership) and the BY-COVID project linked to the European COVID-19 data portal and the Infectious Diseases Toolkit (IDTk).
- Aim to build sustainable solutions that will **enhance and complement national projects**, making them sustainable beyond the COVID-19 pandemic and applicable to other infectious diseases.
- Two generic scientific questions on infectious diseases were formulated for MS26.
 - i) GWAS (validation): risk variants of severe COVID-19 (research)
 - ii) Variants that may guide prognosis and/or treatment (healthcare)





WP7: GDI Use Cases

Data-driven Cancer Research



- Contribute to elaborate feasible technical solutions to **facilitate the access** to genomic and phenotypic datasets provided by related projects (e.g **EOSC4Cancer**) and the technology and recommendations of the 1+MG/B1MG cancer use-case (WGg leadership).
 - The finalisation of a Minimal Cancer Dataset.
 - The generation of cancer use cases based on synthetic data, with the aim of testing the GDI technical infrastructure.
- Alignment with the **HealthData@EU and the cancer mission** to guarantee the flux of information and the sustainability of the access to medical data for secondary use in research.
- Joint work with **EUCAIM** on ensuring technical interoperability across domain and the use of similar approaches for common challenges, e.g. extending beacon for clinical imaging using OMOP extensions for radiology, extending the evaluation of federated learning frameworks to incorporate genomics data.





WP7: GDI Use Cases

Data-driven Cancer Research



- Three scientific questions were formulated for the cancer use-case for MS26.

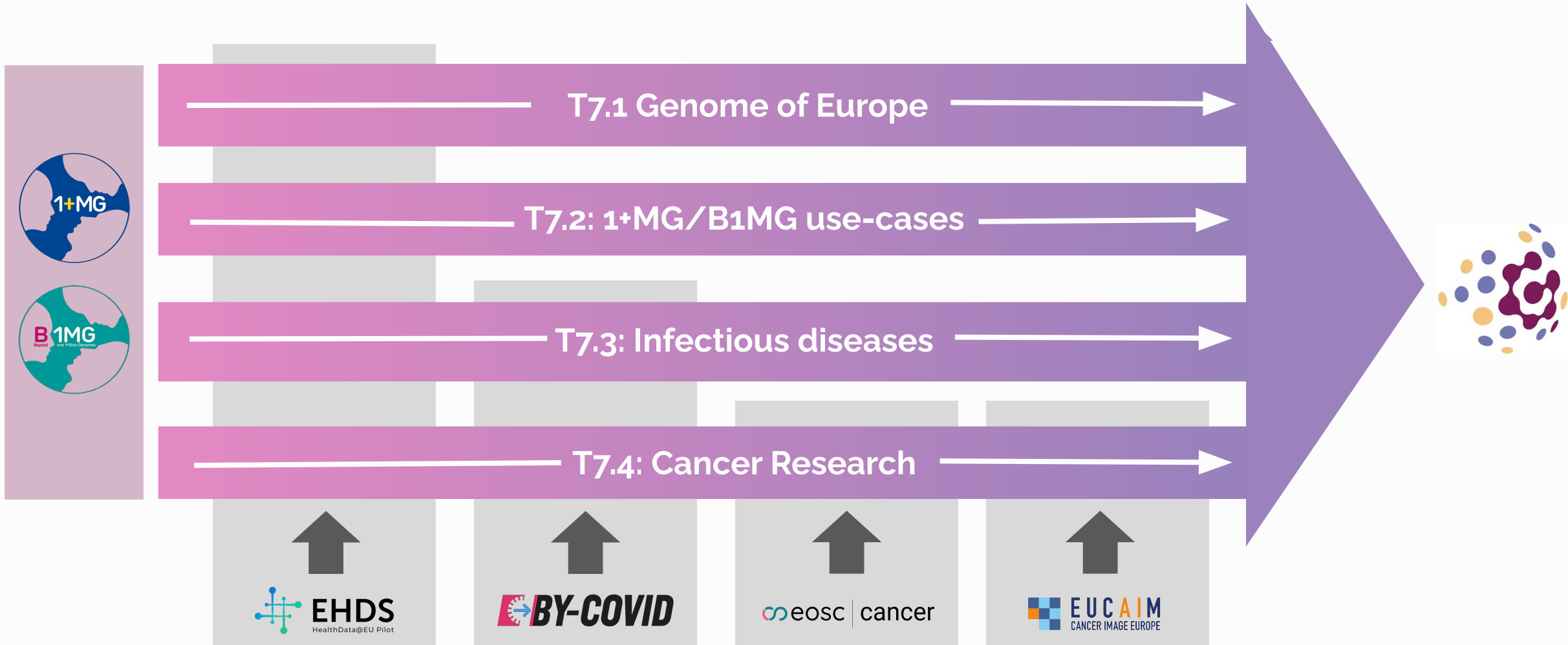
- i) for a patient diagnosed with NSCLC ROS1 fusion-positive cancer (T₀) was treated with standard of care TKI Crizotinib (T₁),
- ii) for a patient diagnosed with metastatic melanoma. Molecular standard analysis revealed a BRAF V600E point mutation (T₀) and
- iii) to determine MicroSatellite Instability score in ColoRectal Cancer.

- Adoption of cBioPortal as an open source solution for the visualization and interpretation of Cancer data.





WP7: GDI Use Cases

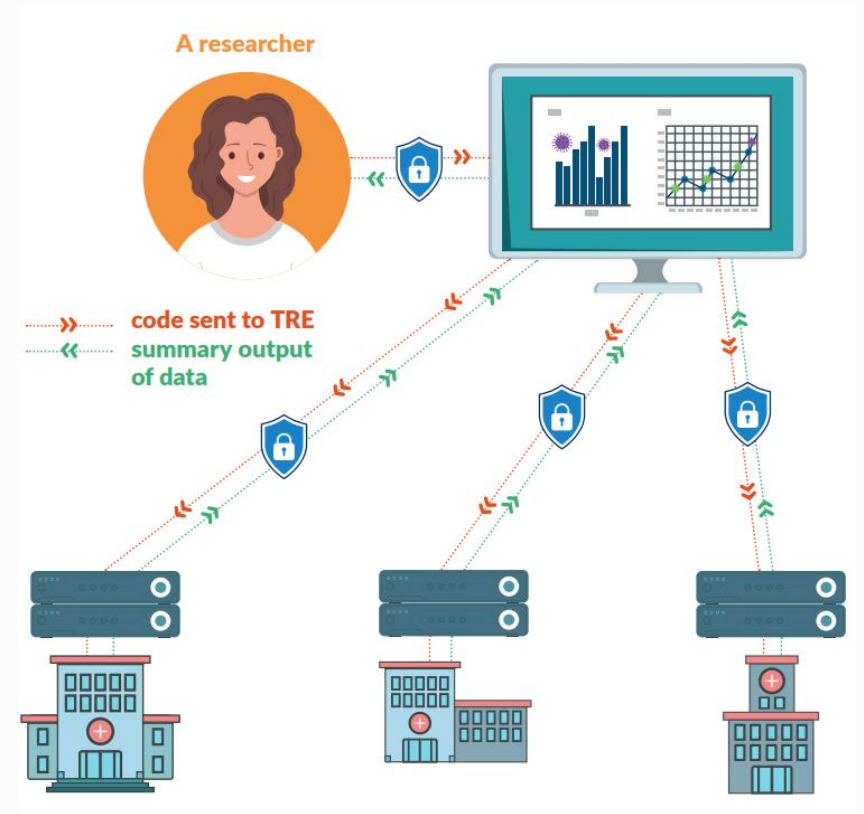




GDI approaches to federated analysis

Federated analysis allows researchers to send code to decentralized data sources rather than moving the data.

- **Data Stays Put**
 - No need to centralize data.
 - Each data source maintains control of its data locally.
- **Diverse Data Settings**
 - Ideal for studying data from various settings (e.g., hospitals nationwide).
 - Each setting creates and maintains its own data environment.
- **Remote Code Execution**
 - Researchers send their code to each location.
 - Local analysis is conducted, ensuring data privacy and security.
- **Collaborative Results**
 - Analysis results, subject to disclosure control, are combined for a comprehensive view.





Considerations

- Strong dependency on the data governance framework for GDI.
- Availability of data types: individual-level vs. aggregated datasets.
- Varying technology readiness levels.





Federated/distributed analysis of **infectious diseases** (pks) data

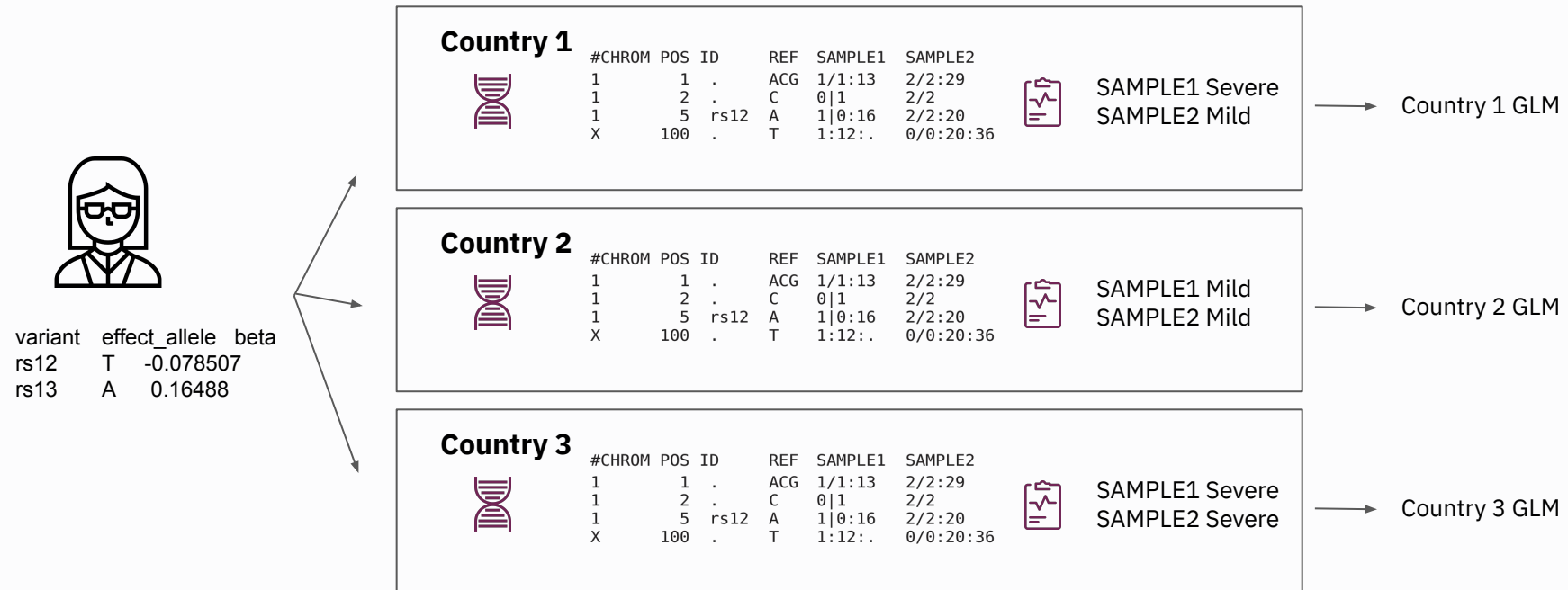
Description of the demonstrator

- Each country has its own group of sequenced patients and a table with the outcomes (Was this individual infected with SARS-CoV2? What was the severity of the disease?)
- A researcher in a country wants to test the association of specific genetic variants with prognosis (i.e., severity), taking into account possible confounders (e.g., sex, smoking status, ancestrality, whatever is available in the common data model).



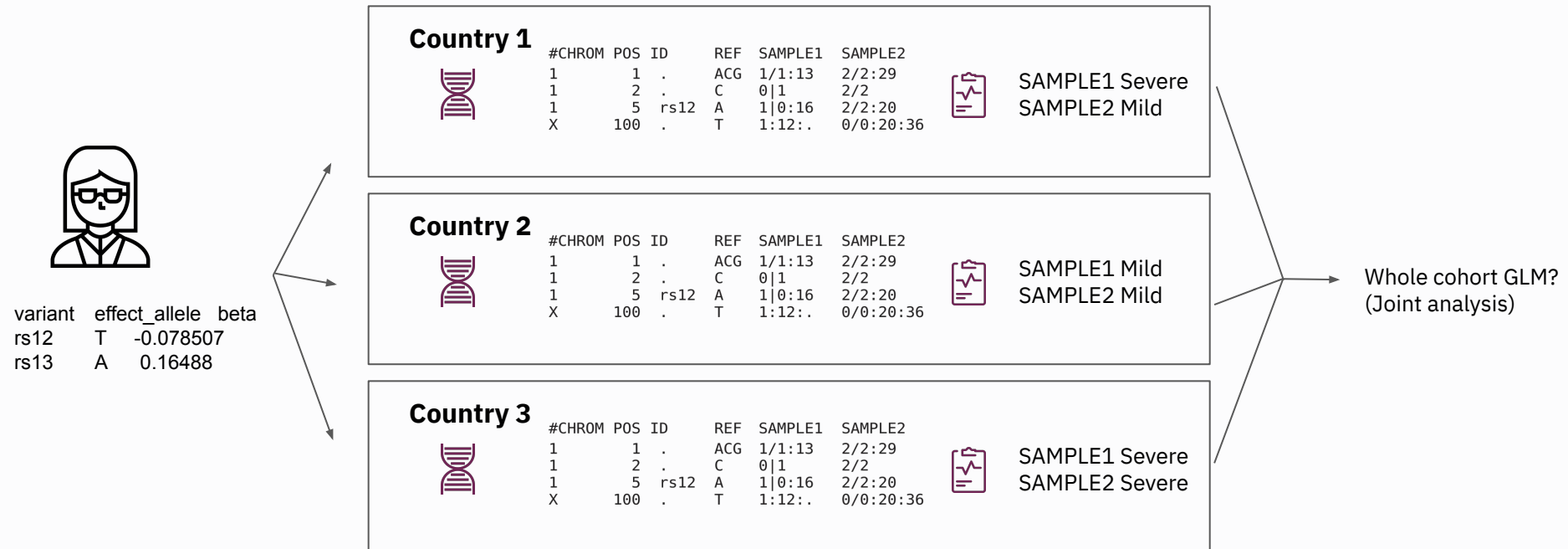


Federated/distributed analysis of **infectious diseases** (pks) data



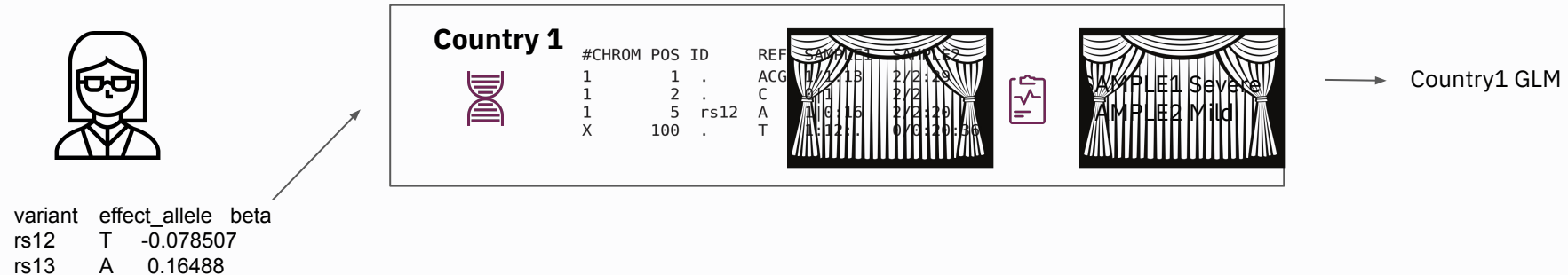


Federated/distributed analysis of **infectious diseases** (pks) data





Federated/distributed analysis of **infectious diseases** (pks) data



- Ideally, the actual scores are not returned, as it would reveal the risk of an individual versus a certain disease. But the researcher would have the summaries or fitting pooled GLM to assess relationships between phenotypes and PRS scores.
- The researcher does NOT upload the dataset to Galaxy and send it to whatever country to compute it. It sends the analysis request to a data that lives in that country and it does not see the raw contents of it. It is a privacy-preserving federated analysis.





Federated/distributed analysis of **infectious diseases** (pks) data

```
1 Belgian datasets:  
2  
3 https://usegalaxy.esgws.uno/restricted/country3_noduplicates.bed  
4 https://usegalaxy.esgws.uno/restricted/country3_noduplicates.bin  
5 https://usegalaxy.esgws.uno/restricted/country3_noduplicates.fam  
6  
7
```

1. Data is restricted to only the Galaxy instance

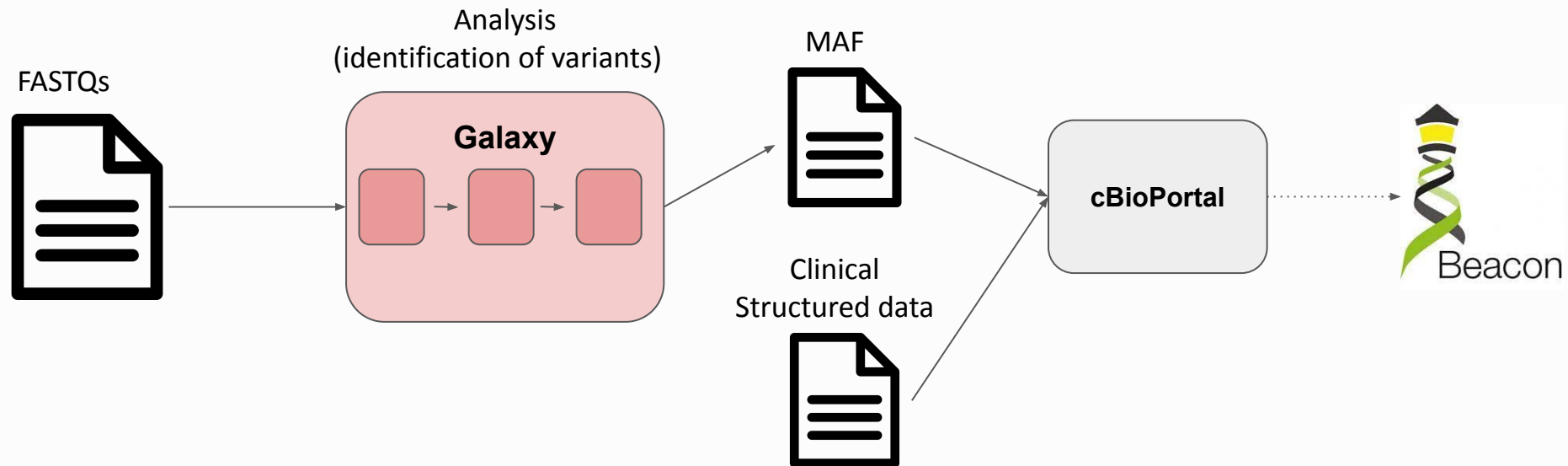




Data mobilization across analysis platform (Cancer use-case)

Description of the demonstrator

Workflow from FastQ to data analysis and visualisation in cBioPortal. Then, the data in cBioPortal, if it is well formatted, will be queried directly by Beacon (optional).





Data mobilization across analysis platform (Cancer use-case)



Data mobilization across analysis
platform (Cancer use-case)





5-safes RO-Crate compatible **Trusted Research Environment (HUTCH + WfExS)**

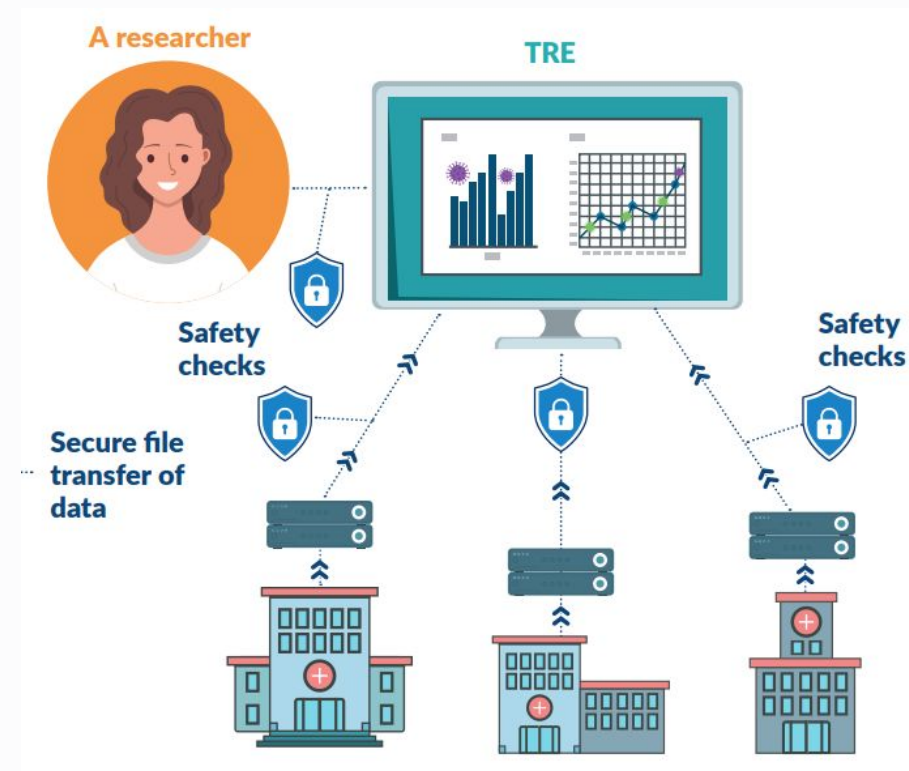
What is a Trusted Research Environment (TRE)?

A **Trusted Research Environment (TRE)** / Secure Processing Environment (SPE) / Secure Data Environment (SDE) is a highly secure computer system where data is stored.

Approved researchers can access this data remotely by using secure logins and passwords. These environments are designed to be safe, allowing only authorised individuals to access the data.

Data cannot be added or removed without proper permissions, ensuring transparency and accountability. Multiple sources of data can be combined in these environments to create a comprehensive dataset for research.

Researchers can apply to access and analyse the data using computer programs they develop but organisations must still follow legal requirements when sending data to a TRE.





5-safes RO-Crate compatible **Trusted Research Environment** (HUTCH + WfExS)

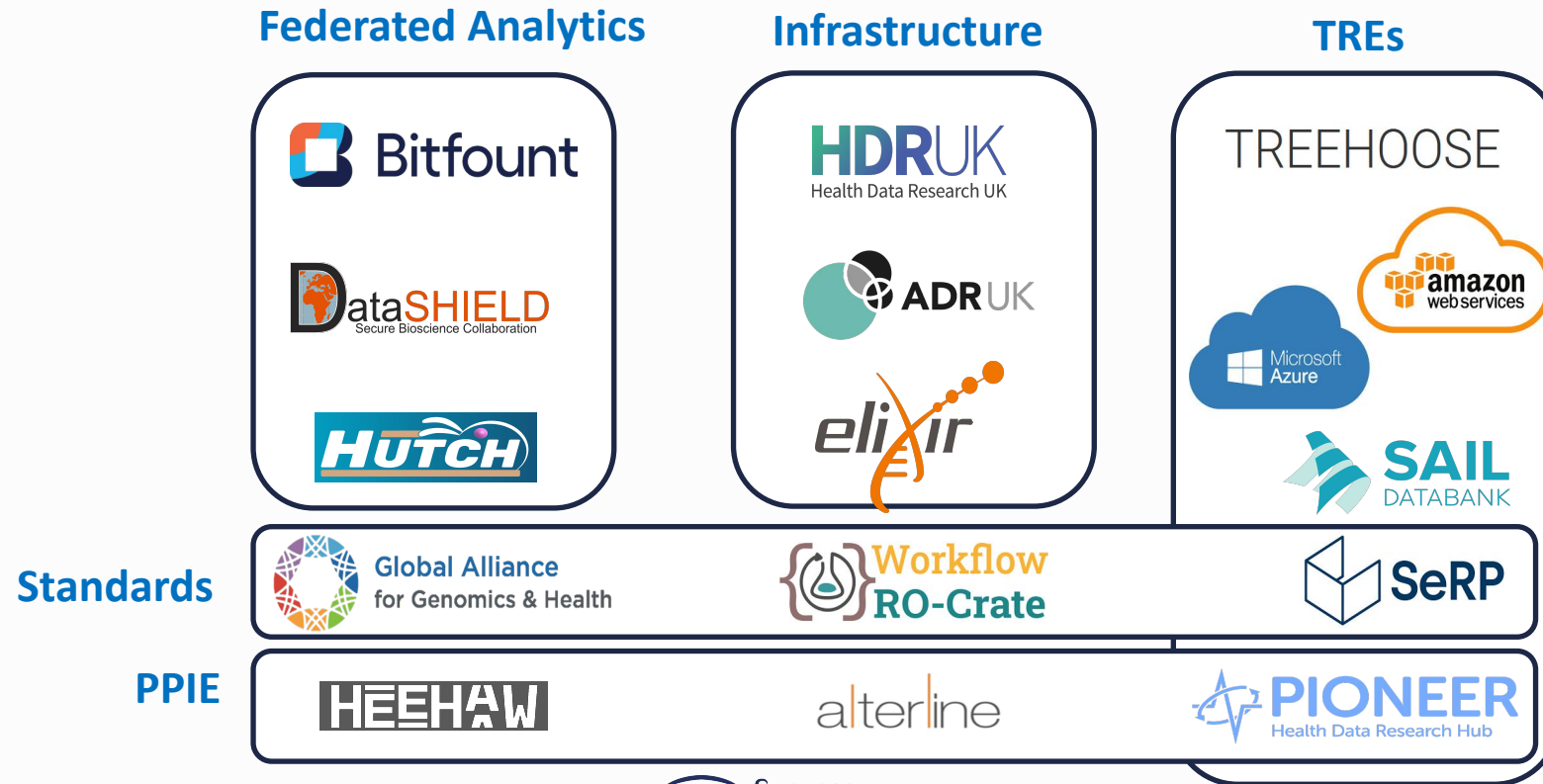
A more standardised approach

- How do we standardise the flow of information (metadata) for federated analysis within the Five Safes framework?
- Can we develop a mechanism of running federated analytics workflows for the queries in the TREs that is technology agnostic?



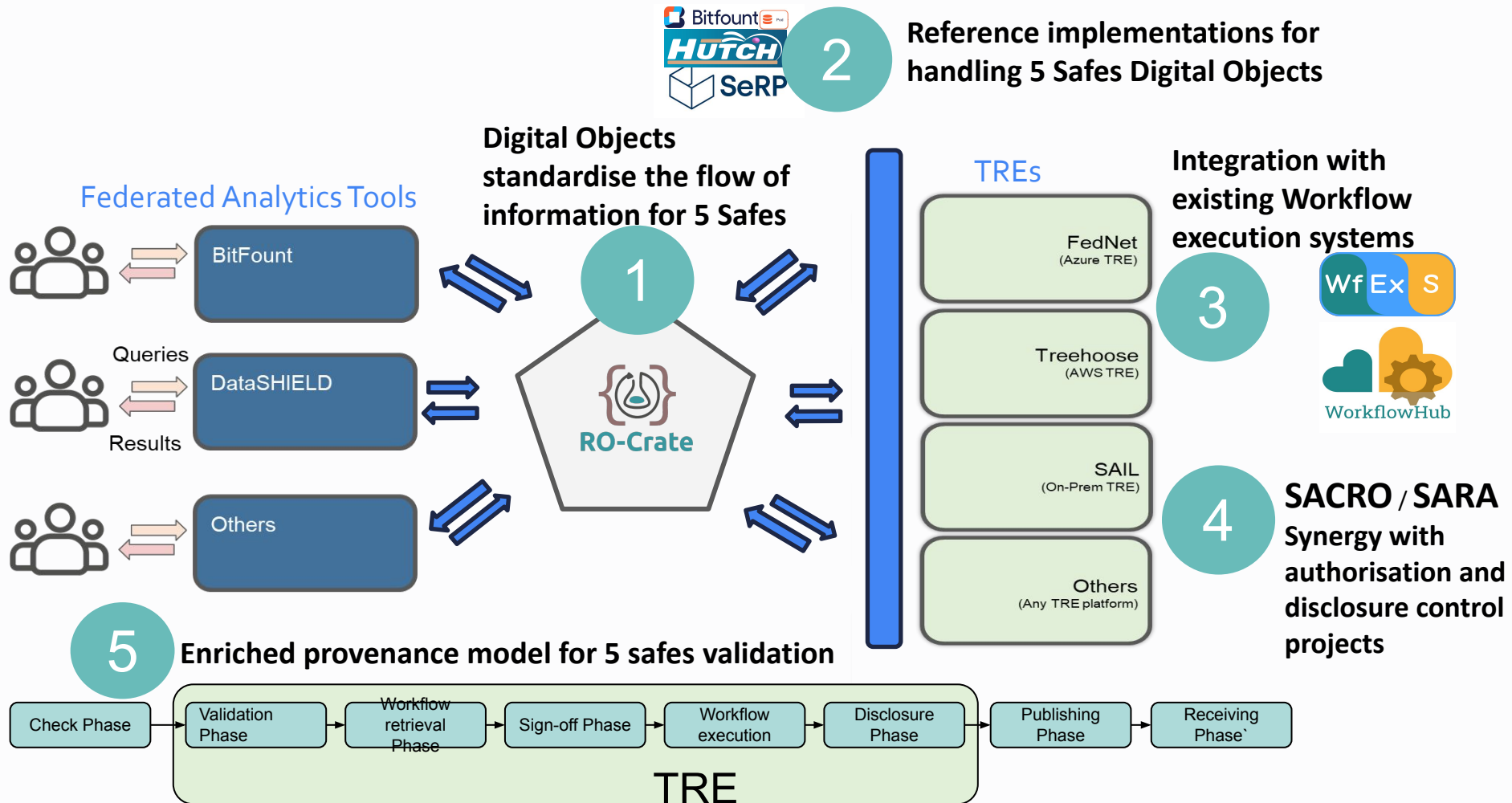


5-safes RO-Crate compatible **Trusted Research Environment** (HUTCH + WfExS)





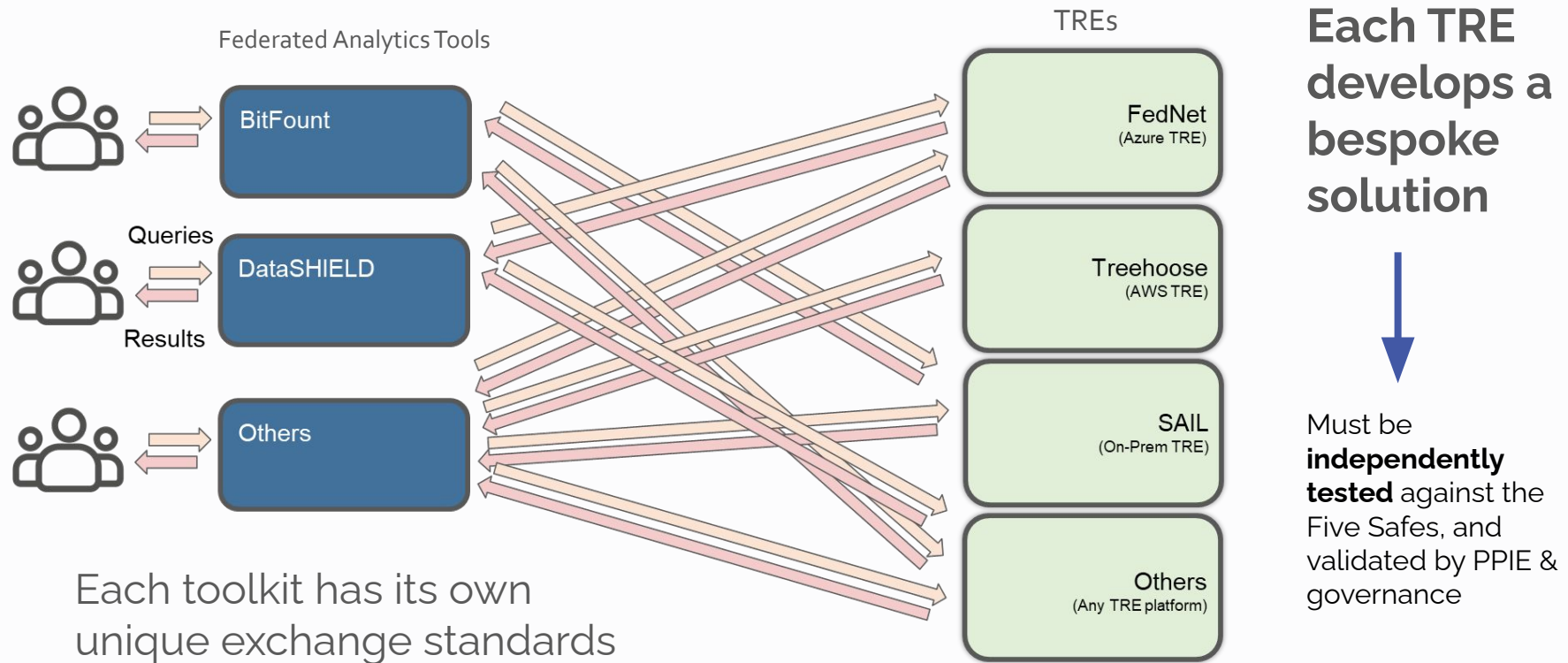
5-safes RO-Crate compatible Trusted Research Environment (HUTCH + WfExS)





5-safes RO-Crate compatible **Trusted Research Environment** (HUTCH + WfExS)

You can't always get all the data into one TRE but analyses across multiple TREs is a pain





What is next?

- Strengthen demonstrators exploring how to couple existing technologies with the GDI data governance framework.
- Elaborate a map on available technologies vs. data types associated with use-cases.

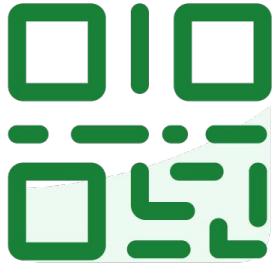




Thanks!



slido



Join at slido.com
#GDISF1

① Click **Present with Slido** or install our [Chrome extension](#) to display joining instructions for participants while presenting.



Breakout Room 1

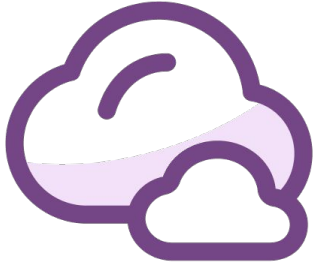
Topic: How can **industry** contribute in the genomics ecosystem to support the 1+Million Genomes (1+MG) initiative ambition to enable secure access to high-quality genomics and the corresponding clinical data across Europe for better research, personalised healthcare and health policy making

Discussion point 1

How can industry help to mobilise genomic and health data from healthcare/a clinical setting into the European Genomic Data Infrastructure for secondary use?



slido



Challenges faced when mobilising genomic and health data from healthcare/a clinical setting into the European Genomic Data Infrastructure for secondary use?.

① Click **Present with Slido** or install our [Chrome extension](#) to activate this poll while presenting.

slido



What are the solutions that industry could help facilitate?

① Click **Present with Slido** or install our [Chrome extension](#) to activate this poll while presenting.



Breakout Room 1

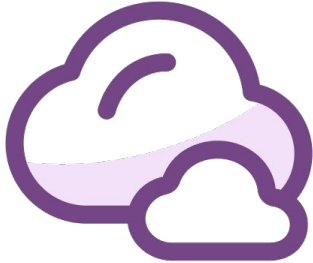
Topic: How can **industry** contribute in the genomics ecosystem to support the 1+Million Genomes (1+MG) initiative ambition to enable secure access to high-quality genomics and the corresponding clinical data across Europe for better research, personalised healthcare and health policy making

Discussion point 2

How industry can facilitate the analysis of data - federated analytics/learning to generate evidence



slido



What are the gaps that exist in currently available federated analytical tools & services?

① Click **Present with Slido** or install our [Chrome extension](#) to activate this poll while presenting.

slido



What are the solutions that industry could help facilitate?

① Click **Present with Slido** or install our [Chrome extension](#) to activate this poll while presenting.



Breakout Room 1

Topic: How can **industry** contribute in the genomics ecosystem to support the 1+Million Genomes (1+MG) initiative ambition to enable secure access to high-quality genomics and the corresponding clinical data across Europe for better research, personalised healthcare and health policy making

Discussion point 3

How can industry bridge the innovation gap between research results and a healthcare setting based on the data made available by 1+MG/GDI?



slido



What elements are required to facilitate innovation to improve the use of genomics in personalised healthcare?

① Click **Present with Slido** or install our [Chrome extension](#) to activate this poll while presenting.



Breakout Room 1 - Summary Slide

Topic: How can **industry** contribute in the genomics ecosystem to support the 1+Million Genomes (1+MG) initiative ambition to **enable secure access to high-quality genomics and the corresponding clinical data** across Europe for better research, personalised healthcare and health policy making

- How can industry help to mobilise genomic and health data from healthcare/a clinical for secondary use?
 - Challenges: GDPR/Consent, data storage, inconsistent workflows
 - Solutions: Sustainability of tools, supporting EC goals of encouraging innovation
- How industry can facilitate the analysis of data - federated analytics/learning to generate evidence
 - Challenges: Costing of computational capacity
 - Solutions: Cloud services providers (e.g. Google, Azure) have costing model in place
- What elements are required to facilitate innovation to improve the use of genomics in personalised healthcare?
 - Need clarity on where the gaps are in the infrastructure





Breakout Room 1 - Summary Slide



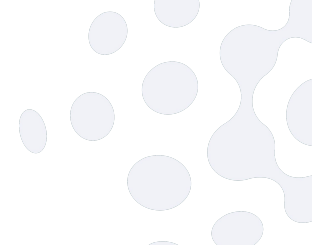
Challenges faced when mobilising genomic and health data from healthcare/a clinical setting into the European Genomic Data Infrastructure for secondary use?.

Wordcloud Poll 10 responses 5 participants





Breakout Room 1 - Summary Slide



What are the solutions that industry could help facilitate?

Open text poll  2 responses  2 participants



Anonymous

Lobbying local governments to encourage the EU level legislation

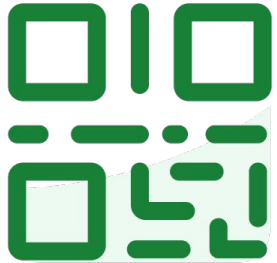


Anonymous

Sustainability of tools



slido



Join at slido.com
#GDISF2

① Click **Present with Slido** or install our [Chrome extension](#) to display joining instructions for participants while presenting.



Breakout Room 2

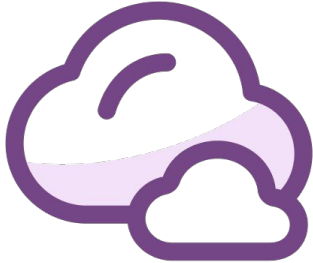
Topic: How can **industry** contribute in the genomics ecosystem to support the 1+Million Genomes (1+MG) initiative ambition to enable secure access to high-quality genomics and the corresponding clinical data across Europe for better research, personalised healthcare and health policy making

Discussion point 1

How can industry help to mobilise genomic and health data from healthcare/a clinical setting into the European Genomic Data Infrastructure for secondary use?



slido



Challenges faced when mobilising genomic and health data from healthcare/a clinical setting into the European Genomic Data Infrastructure for secondary use?.

① Click **Present with Slido** or install our [Chrome extension](#) to activate this poll while presenting.

slido



What are the solutions that industry could help facilitate?

① Click **Present with Slido** or install our [Chrome extension](#) to activate this poll while presenting.



Breakout Room 2

Topic: How can **industry** contribute in the genomics ecosystem to support the 1+Million Genomes (1+MG) initiative ambition to enable secure access to high-quality genomics and the corresponding clinical data across Europe for better research, personalised healthcare and health policy making

Discussion point 2

How industry can facilitate the analysis of data - federated analytics/learning to generate evidence



slido



What are the gaps that exist in currently available federated analytical tools & services?

① Click **Present with Slido** or install our [Chrome extension](#) to activate this poll while presenting.

slido



What are the solutions that industry could help facilitate?

① Click **Present with Slido** or install our [Chrome extension](#) to activate this poll while presenting.



Breakout Room 2

Topic: How can **industry** contribute in the genomics ecosystem to support the 1+Million Genomes (1+MG) initiative ambition to enable secure access to high-quality genomics and the corresponding clinical data across Europe for better research, personalised healthcare and health policy making

Discussion point 3

How can industry bridge the innovation gap between research results and a healthcare setting based on the data made available by 1+MG/GDI?



slido



What elements are required to facilitate innovation to improve the use of genomics in personalised healthcare?

① Click **Present with Slido** or install our [Chrome extension](#) to activate this poll while presenting.



Breakout Room 2 - Summary Slide

Topic: How can **industry** contribute in the genomics ecosystem to support the 1+Million Genomes (1+MG) initiative ambition to enable secure access to high-quality genomics and the corresponding clinical data across Europe for better research, personalised healthcare and health policy making

- **Discussion Point 1 On Data mobilisation:** Privacy - Public Trust (building citizens' trust on 2ndary data use) - Fight the political bias against certain pharmaceutical companies
 - Ensure company is motivated, with a clear goal on benefits for the company. Need for staff redirection- need for change culture, approach & follow a model on a greater scale for data mobilisation.
 - Benefit of looking to other country activities and models on data generated from pharma.
- **Discussion Point 2 How industry will facilitate data analysis/generation and how industry can demonstrate their ideas?** Raise awareness, precompetitive projects (like OpenTargets), professional solutions is a key: democratising the systems for max impact - Developing cheaper storage systems- commercialisation of research outputs but with transparency on investment and visibility on the market - EC encourages industry involvement to achieve sustainability, but this is Work in Progress.
- **Discussion Point 3 Elements required to facilitate innovation? Gaps? Industry:** To bridge the gap: Need to feed data already in GDI, back to healthcare professionals for access and use via an interface, to increase the user experience - Need to re-identify patients after data has been used; this is important for moving back to clinical practice.
 - Industry needs to be part of the overall research.
 - Need for a legal framework & solid regulations established by the EU, without GDPR to be neglected. Technology is currently ahead of Law
- **Level of Readiness for the Healthcare System? What do we need there?** Responsibility also on industries to increase awareness- a lot is happening via Research Projects and industry is not invited to knowledge contribution, but use more traditional ways that do not promote the current needs. Need to keep up with how the European landscape changes.



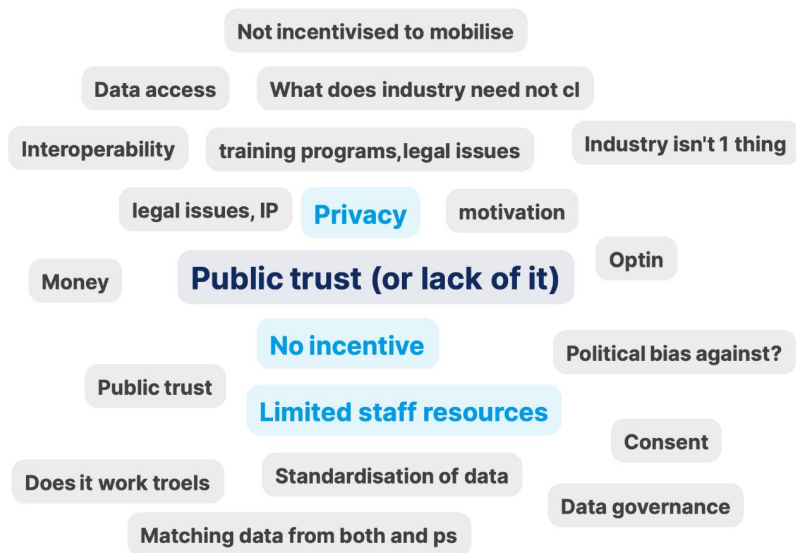


Point 1



Challenges faced when mobilising genomic and health data from healthcare/a clinical setting into the European Genomic Data Infrastructure for secondary use?.

Wordcloud Poll 37 responses 15 participants





Point 1



What are the solutions that industry could help facilitate?

Open text poll 11 responses 5 participants



Anonymous

Benefit sharing - super important.



Anonymous

Specify requirements for multicountry clinical trials? Build tools to identify subjects based on inclusion criteria?



Anonymous

Collaboration - so important. And we saw that through the pandemic.



Anonymous

Technology will surely never be the rate limiting factor in furthering genomics - surely it's more likely to be how people work and think. There are conscious and unconscious biases.



Anonymous

Industry means so many different things - great work you are doing to map that Despoina.



Anonymous

Democratising the systems for maximising impact



Anonymous

Is European politicians ready to engage industry in the delivery system of 1+MG? EHDS?



Anonymous

Eventually, implementing machine learning workflows in the infrastructure (e.g. hugging face do great bioinformatics work)



Anonymous

developing cheaper storage systems



Anonymous

Maximising user experience. Backend software is important but so is front end!



Anonymous

How can industry engage to demonstrate any of their ideas?





Point 2



What are the gaps that exist in currently available federated analytical tools & services?

Wordcloud Poll  1 response  1 participant

Priably legal (GDPR)





Point 3



What elements are required to facilitate innovation to improve the use of genomics in personalised healthcare?

Open text poll 4 responses 3 participants



Anonymous

Transparency in the analysis and explainability



Anonymous

Could you expose a practical exemple for this question ?



Anonymous

At the moment ehds is spilt strictly in 2 data spaces between primary and secondary.
Hence the legal barrier is definitive.



Anonymous

Isn't this a question only all those companies with these solutions can answer?





European
Genomic Data
Infrastructure

GDI Stakeholder Forum GDI Starter Kit

Dylan Spalding

23 November 2023



GDI project receives funding from the European Union's Digital Europe Programme under grant agreement number 101081813.



gdi.onemilliongenomes.eu/



[@GDI_EUproject](https://twitter.com/GDI_EUproject)



[/company/gdi-euproject](https://www.linkedin.com/company/gdi-euproject)



GDI Pillar II

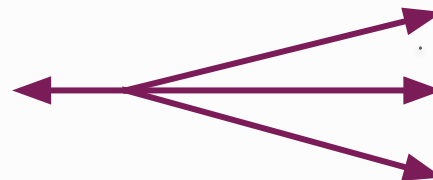


Design & Testing
2020-2023

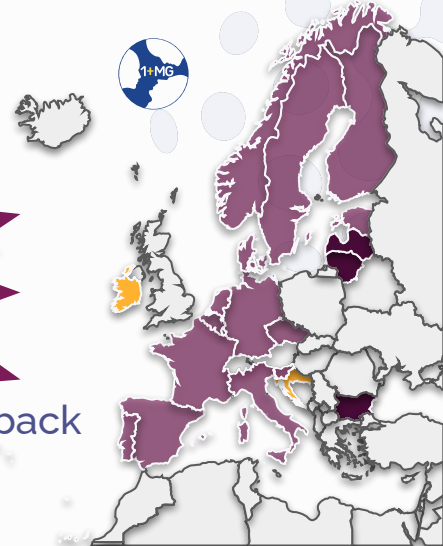


European
Genomic Data
Infrastructure

Scale up & Sustainability
2022-2026



Deployment & Feedback
2022-2026



Deployment of 1+MG national nodes

Operational

National nodes interconnected



Deployment

National node fully operational



Onboarding

Design and testing phase

Considerations & Feedback

- National Strategy
- Emerging technologies
- Pillar III Use Cases





Genomic Data Infrastructure

Countries in GDI:

- Will provide a node or Data Hub
- Each country manages their own data (e.g. regional hubs)
- Data hubs provide cross-border data analysis using a common framework of standards & APIs

Overall data infrastructure provides 5 main functionalities as defined by the 1+MG Scoping paper¹ in 2021



Data
discovery



Access
management tools



Data
processing



Data
reception



Storage and
interfaces

<https://zenodo.org/doi/10.5281/zenodo.6089582>

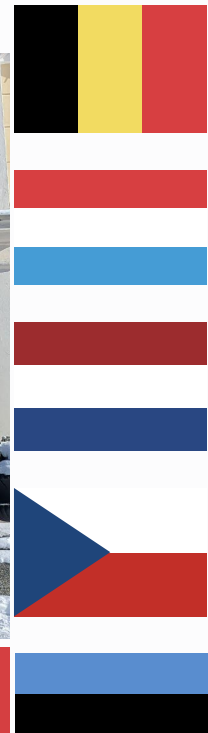




GDI Starter Kit



- Defined at a workshop in Sweden in March 2023
 - 65 experts
 - 18 countries
- Successful deployment and operation demo at ELIXIR AHM in June 2023





Data Discovery

Function:

- Public visibility and search of genomic data, requiring descriptive non-sensitive metadata like summary level descriptions of data
- Link to data access management where users can apply for detailed data access.

GDI Starter Kit Elements:

- Beacon: **Beacon, DUO, phenopackets**
- Beacon Network: **Beacon**
- User Portal: Data Catalogue



Data
discovery



Access
management
tools



Data
processing



Data
reception



Storage
and
interfaces





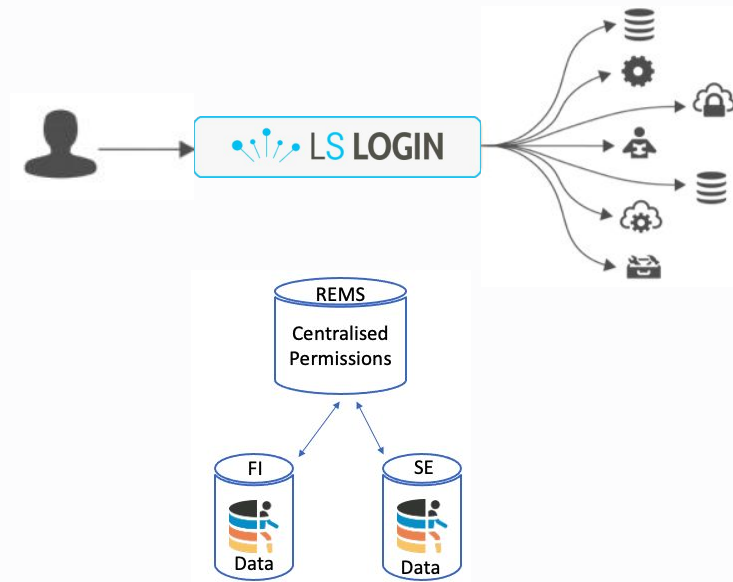
Data Access Management

Function:

- Management of data access according to Data Protection by Design & Default, i.e. facilitation and audit of secure access
- Tools manage:
 - data access applications for data use
 - data access authorisations from the data controllers
 - communication of access rights

GDI Starter Kit Elements:

- LifeScience AAI: **GA4GH Passports, AAI**
- REMS: **GA4GH Passports, AAI, DUO**
- User Portal: Access management: **AAI, DUO**



Data
discovery



Access
management
tools



Data
processing



Data
reception



Storage
and
interfaces





Storage & Interfaces

Function:

- Organisations store data and offer APIs that form the interoperable (standard-based) infrastructure backbone
- Aim is to leverage existing investments in e-infrastructure capacities that can provide data privacy and confidentiality

GDI Starter Kit Elements:

- Sensitive Data Archive: **Crypt4GH**
- Integration of user identity and data access: **DRS, Passport**



Data discovery



Access management tools



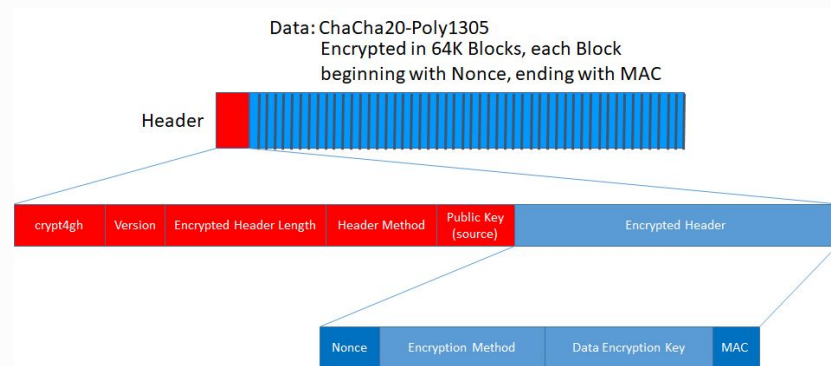
Data processing



Data reception



Storage and interfaces





Data Reception

Function:

- Uniform processes for quality control and standardisation
- Receiving (e.g. upload) or access (e.g. API) to data and metadata
 - Adhering to global standards and principles
 - Genotypic and phenotypic data
- Logical description of datasets

GDI Starter Kit Elements:

- Synthetic data: **phenopackets, VCF, BAM/CRAM**
- Transfer: **htsget**



Data
discovery



Access
management
tools



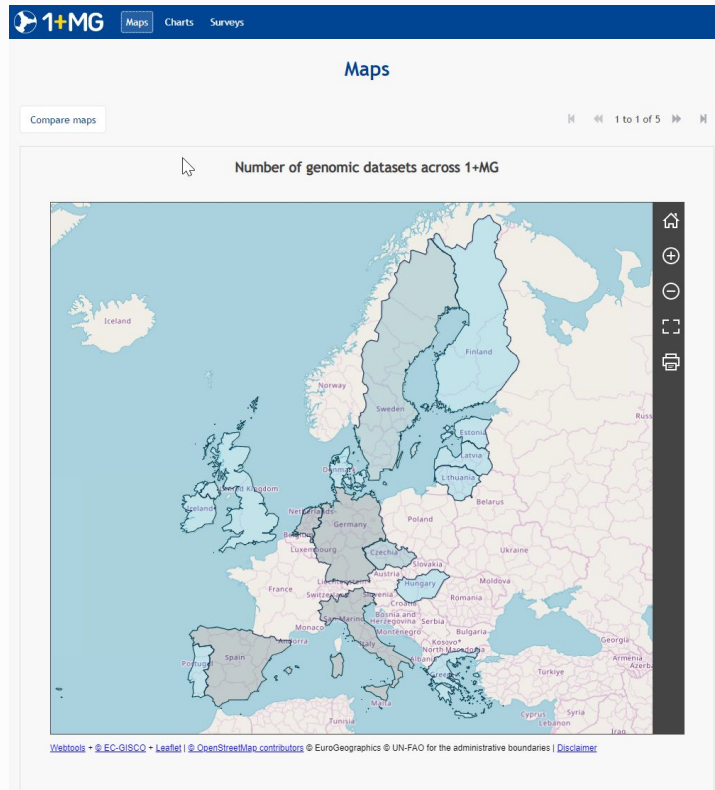
Data
processing



Data
reception



Storage
and
interfaces



Processing (compute)

Function:

- Local, high-performance and cloud computing with security standards appropriate for 1+MG to analyse data
- Processing happens locally or distributed in the infrastructure

GDI Starter Kit Elements:

- Containerised & Federated computation - **TES, WES, TRS**



LUMI datacenter Photo by Pekka Agarth



Data
discovery



Access
management
tools



Data
processing



Data
reception



Storage
and
interfaces





Generalised end to end user Story from 1+MG

1. User discovers phenotypes of interest aggregate data in 1+MG data
2. User logs in via LS AAI (registered level), and discovers both a genomic variant and treatment regime and/or phenotype of interest
3. User applies for data access to 1+MG data
4. Data Access Committee grants access to a virtual cohort
5. User executes analysis as a controlled access user on this virtual cohort across federated locations



WG8 – RD: Do you have any individuals with a mutation in the RYR1 gene and a similar phenotype to congenital myasthenic syndrome?

WG9 – Cancer: Do you have any individuals with a mutation in the PTEN gene, who have BRAF biomarkers and are being treated with vemurafenib?





GDI Infrastructure, user journey & standards

1+MG
5 FUNCTIONALITIES

Data
Discovery



Data Access
Management



Storage &
Interfaces



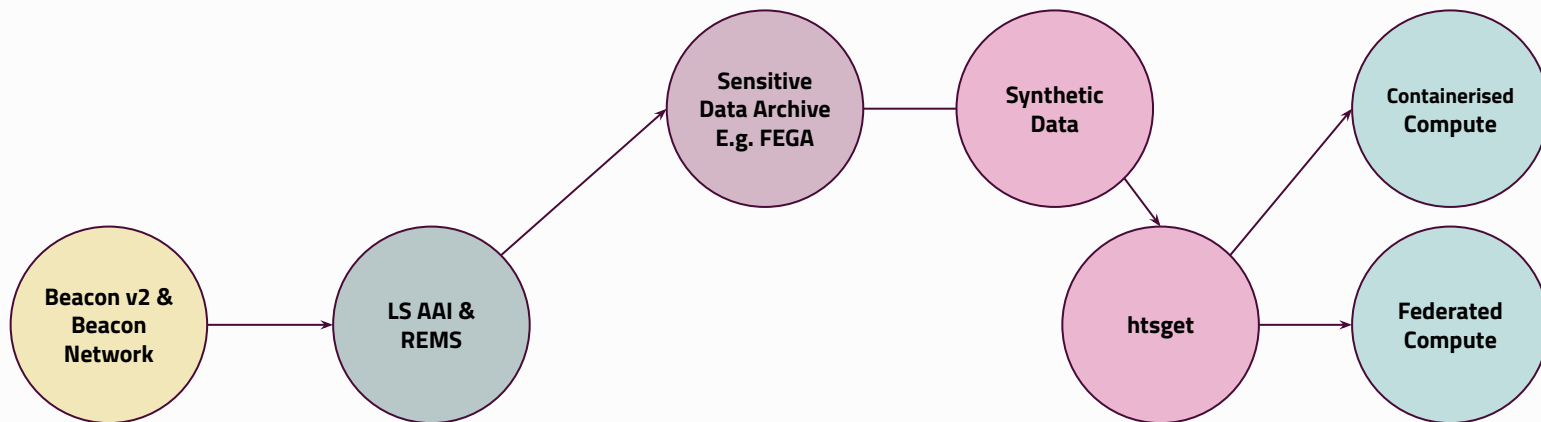
Data
Reception



Data
Processing



STARTER KIT



DUO
Phenopackets
Beacon

Passports
DUO
AAI

CRYPT4GH

Phenopackets
VCF
SAM/BAM/CRAM
htsget

TES
WES
TRS

ISO
STANDARDS





GDI Infrastructure, user journey & standards

1+MG

5 FUNCTIONALITIES

Data Discovery



Data Access Management



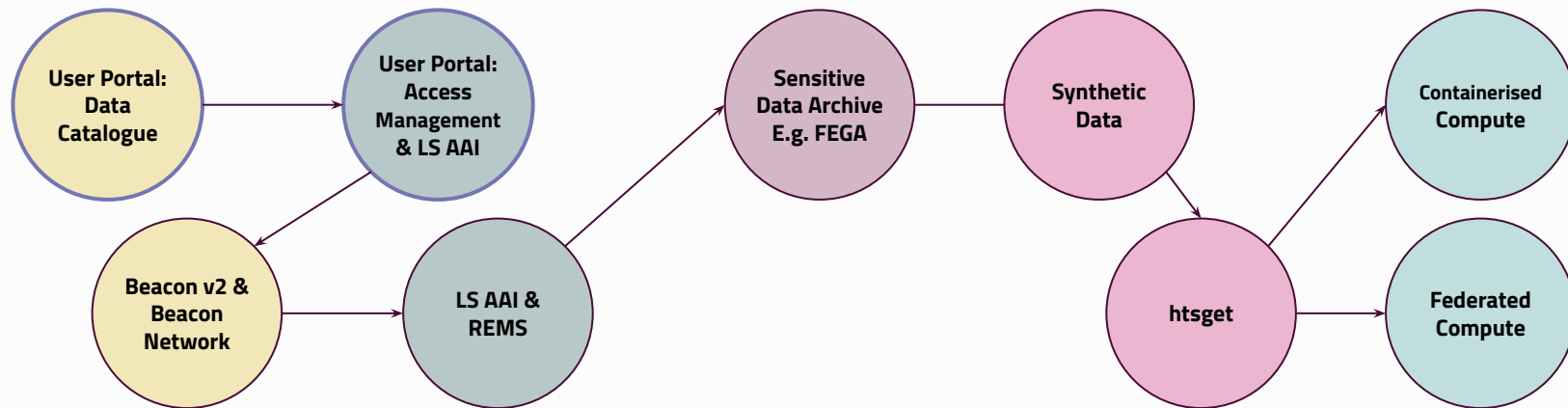
Storage & Interfaces



Data Reception



Data Processing



STARTER KIT

DUO
Phenopackets
Beacon
DCAT-AP

Passports
DUO
AAI

CRYPT4GH

Phenopackets
VCF
SAM/BAM/CRAM
htsget

TES
WES
TRS

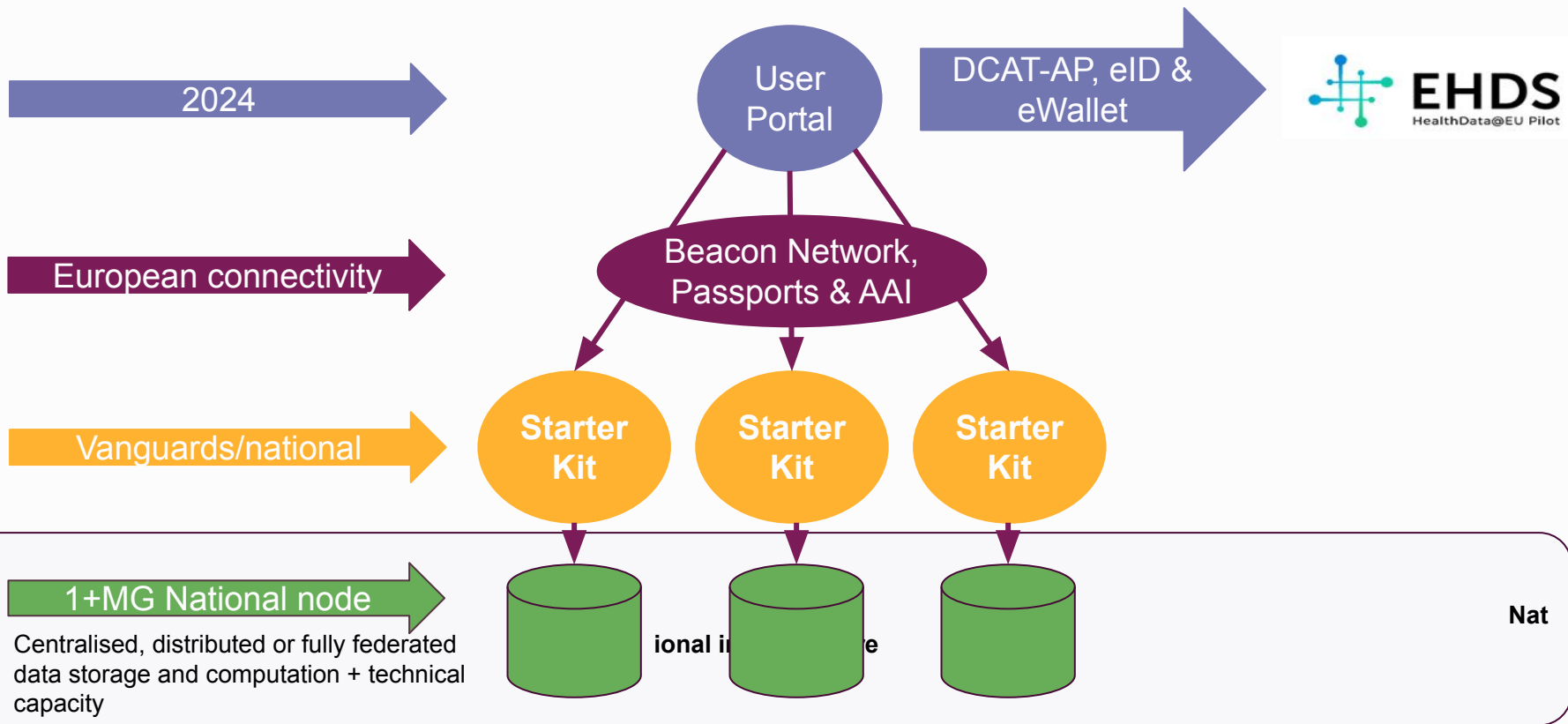


STANDARDS





GDI Starter Kit deployment





GDI Starter Kit = 1st demonstration software bundle released June 2023

Product	Outline	Prod	PO	Function
Sensitive Data Archive	Securely stores data	✓	Sweden	
LifeScience AAI	Provides a federated Identity	✓	Czechia	
REMS	Allows data access applications and decisions	✓	Finland	
Beacon	Genetic and phenotypic data discovery	✓	Spain	
Beacon Network	Federated network of Beacons	✓	Finland	
Synthetic Data	Artificial anonymous data		Finland	
htsget	Secure genetic data distribution		Sweden	
Containerised Computation	Computation via virtualised portable software packages	✓	Czechia	
Federated Computation	Interoperable distributed workflows			
Packaging and Deployment	Packaging and deployment of the starter kit		Spain	
User Portal – Data Catalogue	European level catalogue of data within nodes		Netherlands	
User Portal – Access management	European level application / access management		Luxembourg	

✓ - Product in production in another project (NB: Beacon V1 / ELIXIR Beacon Network)

Future - Due 01/2024

European level

<https://github.com/GenomicDataInfrastructure>



Funded by the European Union



Global Alliance for Genomics & Health



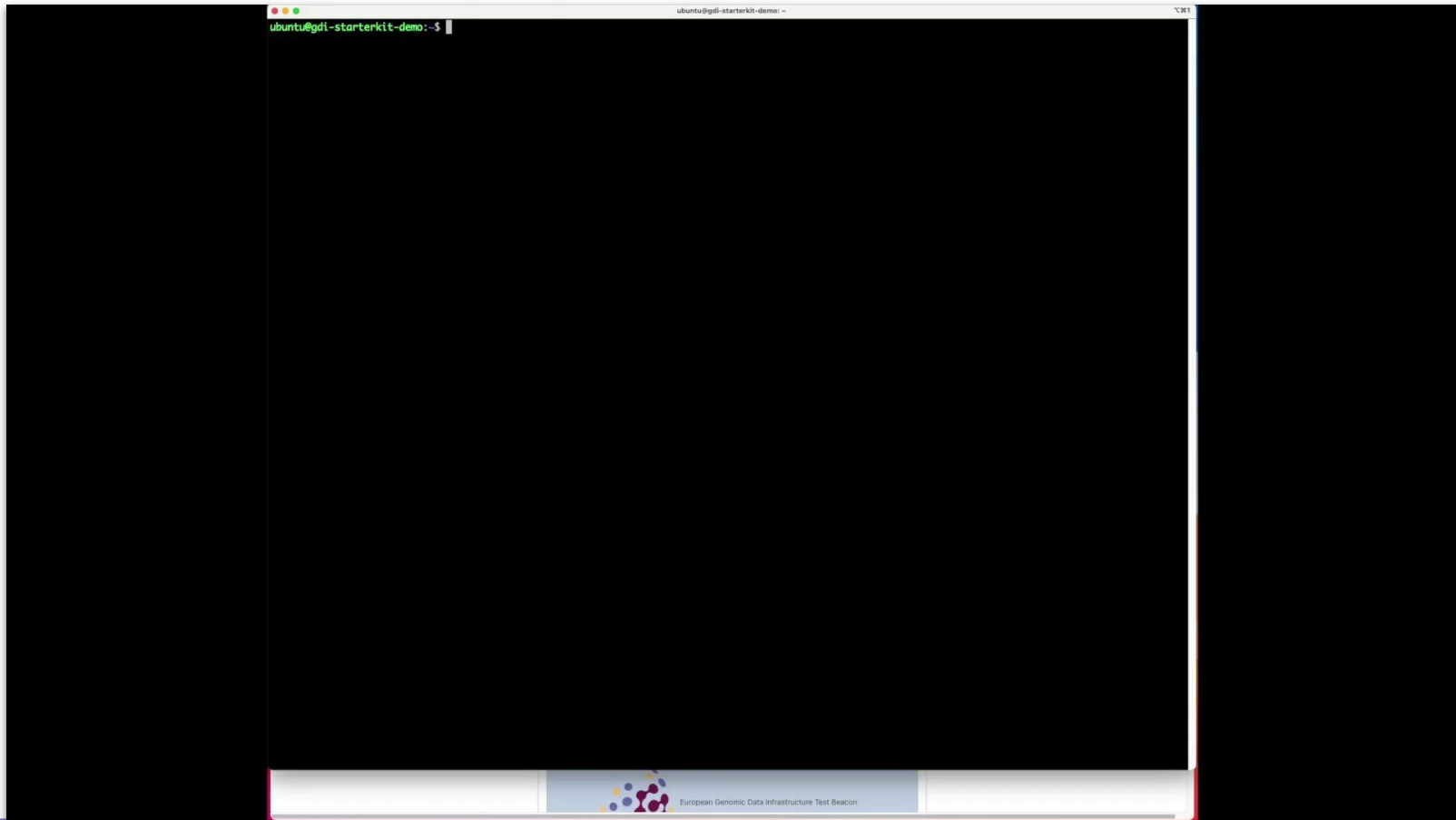
What's next

- Continuous development of Starter Kit products
 - Compliance with **Data Protection by Design and Default** principles
 - **Support** GDI Nodes on the **deployment and operations** of the Starter Kit
 - Not pushing starter kit software to nodes, they may use **compatible alternatives** for producing **functionalities** based on the chosen open **standards**
 - Support **diversity** of node requirements
- **Federated Analytics** / Learning (Pillar III) interaction
- Use cases **demonstrators** (Pillar III)
 - **Genome of Europe**
 - **Cancer**
 - **Infectious diseases**





Demo





Follow our progress

- github.com/GenomicDataInfrastructure
- gdi.onemilliongenomes.eu
- framework.onemilliongenomes.eu

Starter Kit

@Zenhub



European
Genomic Data
Infrastructure



<https://www.youtube.com/watch?v=6MtIJA4xXdU>





Acknowledgements



GDI project receives funding from the European Union's Digital Europe Programme under grant agreement number 101081813.





GDI Consortium



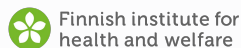
Instituto de Salud Carlos III



Instituto Nacional de Saúde
Doutor Ricardo Jorge



C S C



UNIVERSITETET
I OSLO



Funded by
the European Union



GDI Consortium (continued)



REPUBLIC OF ESTONIA
MINISTRY OF SOCIAL AFFAIRS



THE GOVERNMENT
OF THE GRAND DUCHY OF LUXEMBOURG
Ministry of Higher Education and Research



Funded by
the European Union



Generating synthetic data for health applications

Alberto Labarga



Telecommunications Engineer
Head of Biomedical Data Hub @BSC
More than 20 years crunching data

open data - open source – open science

I am Alberto

 alabarga

 alabarga

 /in/albertolabarga

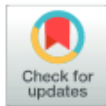


RESEARCH ARTICLE

Synthetic data in health care: A narrative review

Aldren Gonzales^{1*}, Guruprabha Guruswamy², Scott R. Smith¹

1 Office of the Assistant Secretary Planning and Evaluation, US Department of Health and Human Services, Washington, District of Columbia, United States of America, **2** Department of Health Administration and Policy, George Mason University, Virginia, United States of America

* aldren.gonzales@hhs.gov

Abstract

Data are central to research, public health, and in developing health information technology (IT) systems. Nevertheless, access to most data in health care is tightly controlled, which may limit innovation, development, and efficient implementation of new research, products, services, or systems. Using synthetic data is one of the many innovative ways that can allow organizations to share datasets with broader users. However, only a limited set of literature is available that explores its potentials and applications in health care. In this review paper, we examined existing literature to bridge the gap and highlight the utility of synthetic data in health care. We searched PubMed, Scopus, and Google Scholar to identify peer-reviewed articles, conference papers, reports, and thesis/dissertations articles related to the generation and use of synthetic datasets in health care. The review identified seven use cases of synthetic data in health care: a) simulation and prediction research, b) hypothesis, methods, and algorithm testing, c) epidemiology/public health research, d) health IT development, e) education and training, f) public release of datasets, and g) linking data. The review also identified readily and publicly accessible health care datasets, databases, and sandboxes containing synthetic data with varying degrees of utility for research, education, and software development. The review provided evidence that synthetic data are helpful in different aspects of health care and research. While the original real data remains the preferred choice, synthetic data hold possibilities in bridging data access gaps in research and evidence-based policymaking.

OPEN ACCESS

Citation: Gonzales A, Guruswamy G, Smith SR (2023) Synthetic data in health care: A narrative review. *PLOS Digit Health* 2(1): e0000082. <https://doi.org/10.1371/journal.pdig.0000082>

Editor: Alistair Johnson, SickKids: The Hospital for Sick Children, CANADA

Received: June 29, 2022

Accepted: December 6, 2022

Published: January 6, 2023

Peer Review History: PLOS recognizes the benefits of transparency in the peer review process; therefore, we enable the publication of all of the content of peer review and author responses alongside final, published articles. The editorial history of this article is available here: <https://doi.org/10.1371/journal.pdig.0000082>

Copyright: This is an open access article, free of all copyright, and may be freely reproduced, distributed, transmitted, modified, built upon, or otherwise used by anyone for any lawful purpose. The work is made available under the [Creative Commons CC0](https://creativecommons.org/licenses/by/4.0/) public domain dedication.

Data Availability Statement: All data are in the manuscript.

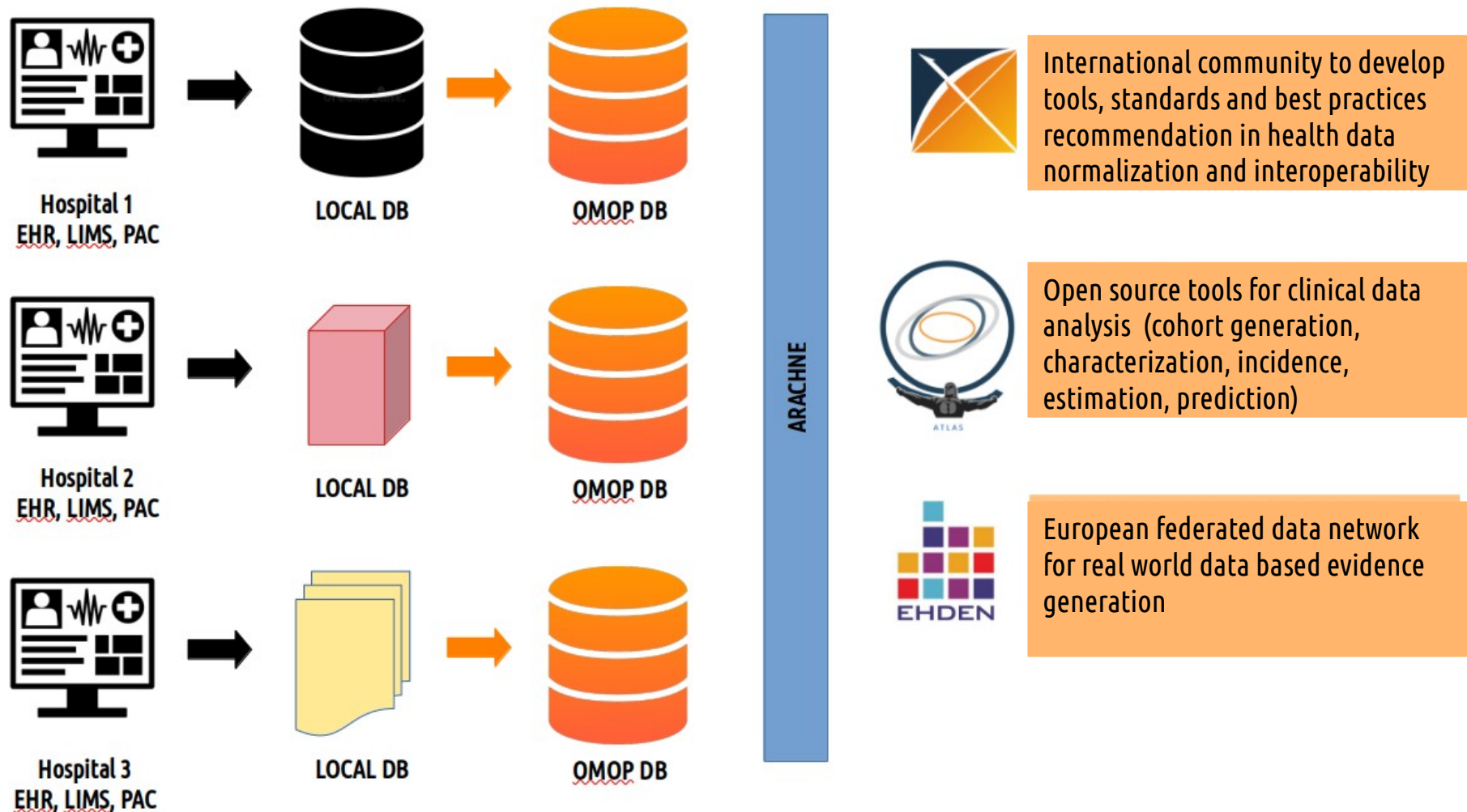
Funding: The authors received no specific funding for this work.

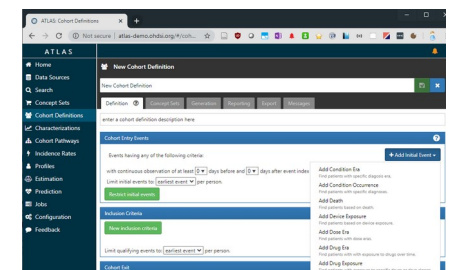
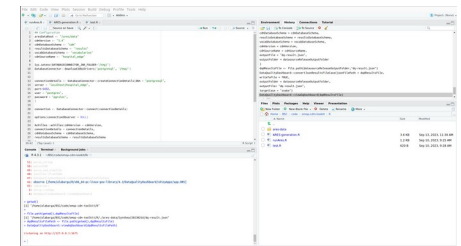
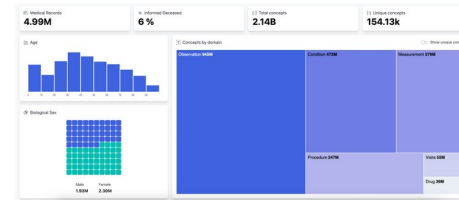
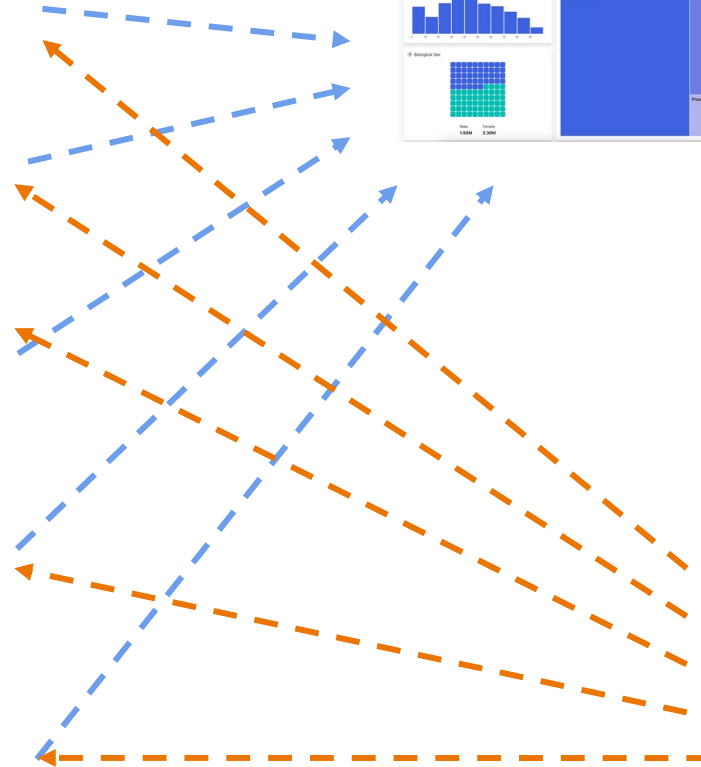
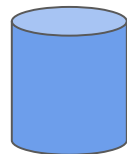
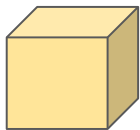
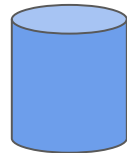
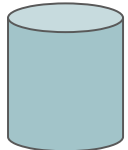
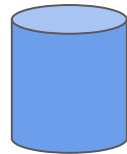
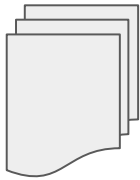
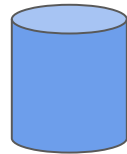
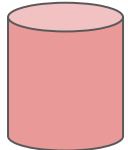
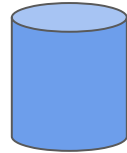
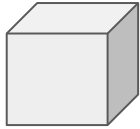
Author summary

Synthetic data or data that are artificially generated is gaining more attention in the recent years because of its potential in making timely health care data more accessible for analysis and technology development. In this paper, we explored how synthetic data are being used by reviewing published literature and by looking at known synthetic datasets that are available to the public. Based on the available literature, it was identified that synthetic data address three challenges in making health care data accessible: it protects the privacy of individuals in datasets, it allows increased and faster access of researchers to health care

Synthetic Health Records

OMOP-CDM





Research and Applications

Synthea: An approach, method, and software mechanism for generating synthetic patients and the synthetic electronic health care record

Jason Walonoski,¹ Mark Kramer,¹ Joseph Nichols,¹ Andre Quina,¹ Chris Moesel,¹ Dylan Hall,¹ Carlton Duffett,¹ Kudakwashe Dube,² Thomas Gallagher,³ and Scott McLachlan²

¹The MITRE Corporation, Bedford, MA, USA, ²HIKER Group, Massey University, Palmerston North, New Zealand, and ³Department of Applied Computing and Engineering Technology, University of Montana, Missoula, MT, USA

Corresponding Author: Jason Walonoski, 202 Burlington Road, Bedford, MA 01730, USA. E-mail: jwalonoski@mitre.org. Phone: +1-781-271-2021

Received 10 May 2017; Revised 15 June 2017; Accepted 5 July 2017

ABSTRACT

Objective: Our objective is to create a source of synthetic electronic health records that is readily available; suited to industrial, innovation, research, and educational uses; and free of legal, privacy, security, and intellectual property restrictions.

Materials and Methods: We developed Synthea, an open-source software package that simulates the lifespans of synthetic patients, modeling the 10 most frequent reasons for primary care encounters and the 10 chronic conditions with the highest morbidity in the United States.

Results: Synthea adheres to a previously developed conceptual framework, scales via open-source deployment on the Internet, and may be extended with additional disease and treatment modules developed by its user community. One million synthetic patient records are now freely available online, encoded in standard formats (eg, Health Level-7 [HL7] Fast Healthcare Interoperability Resources [FHIR] and Consolidated-Clinical Document Architecture), and accessible through an HL7 FHIR application program interface.

Discussion: Health care lags other industries in information technology, data exchange, and interoperability. The lack of freely distributable health records has long hindered innovation in health care. Approaches and tools are available to inexpensively generate synthetic health records at scale without accidental disclosure risk, lowering current barriers to entry for promising early-stage developments. By engaging a growing community of users, the synthetic data generated will become increasingly comprehensive, detailed, and realistic over time.

Conclusion: Synthetic patients can be simulated with models of disease progression and corresponding standards of care to produce risk-free realistic synthetic health care records at scale.

Key words: electronic health records, computer simulation, patient-specific modeling, clinical pathways, RS-EHR

BACKGROUND AND SIGNIFICANCE

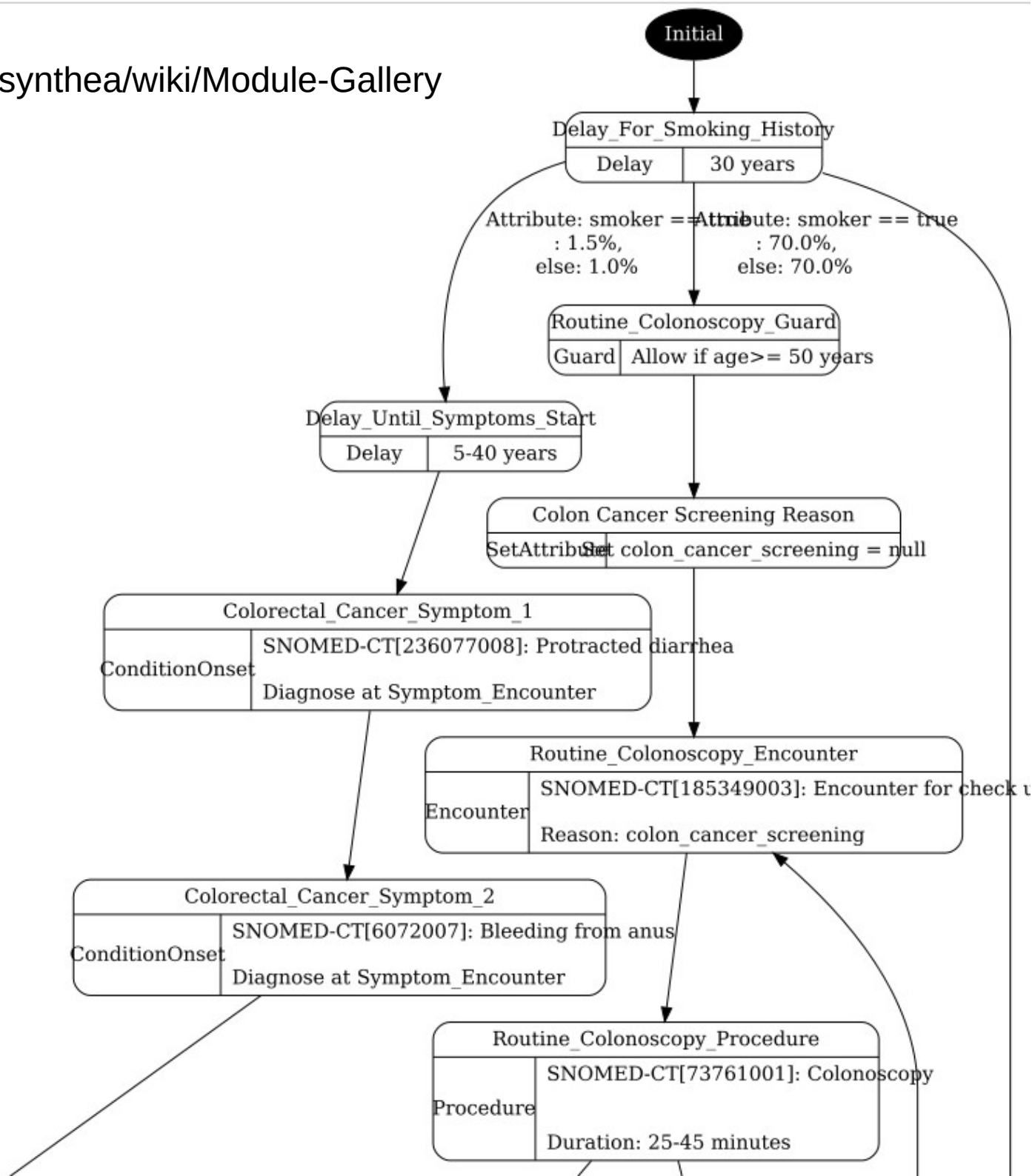
Health care lags other industries in information technology, data exchange, and interoperability. To close these gaps, developers require

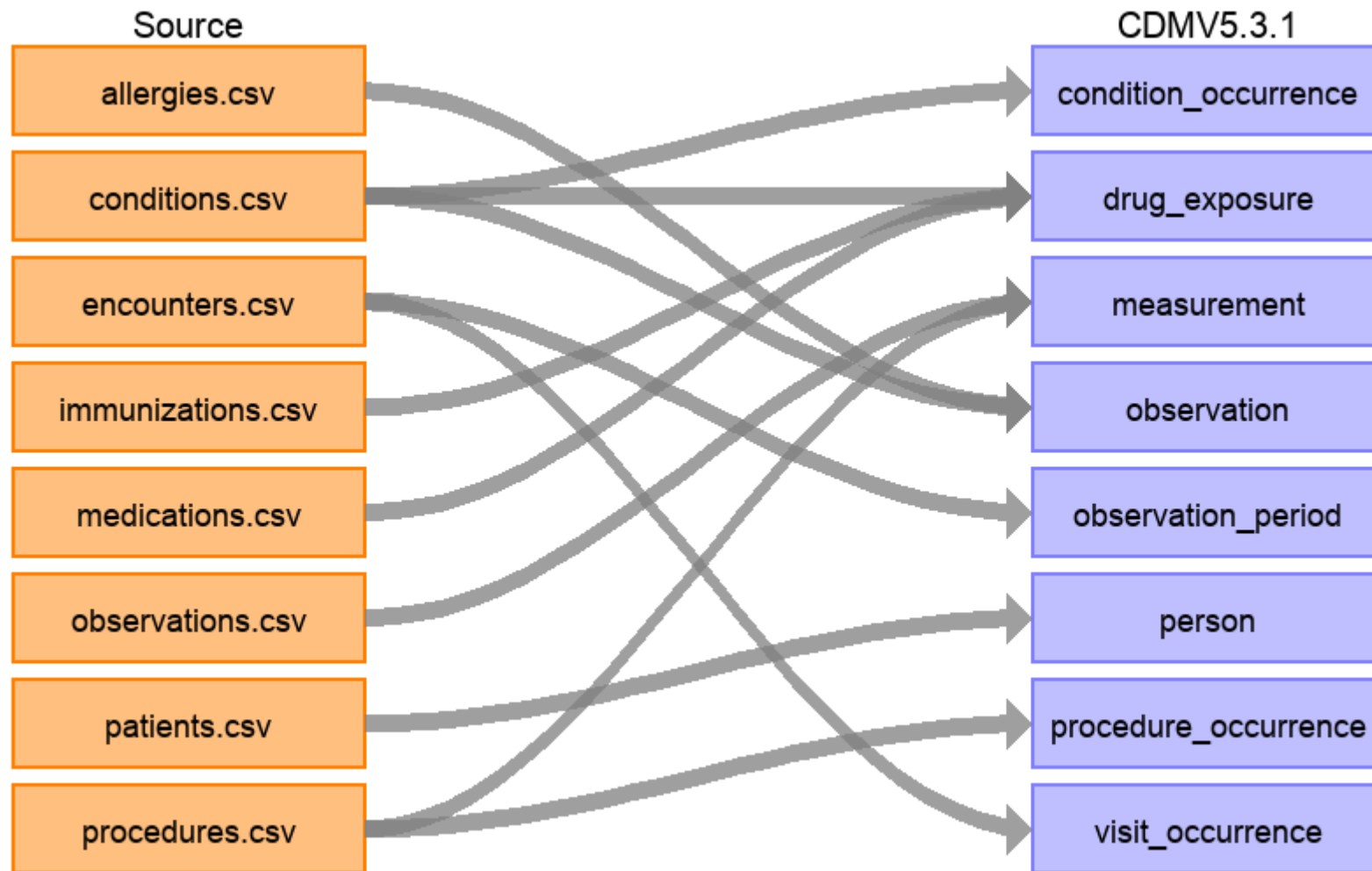
access to large repositories of high-quality health datasets for a range of secondary uses that have no clinical or medical implications, including software development, testing, and clinical training.^{1–4} However, access to real electronic health record (EHR) data

© The Author 2017. Published by Oxford University Press on behalf of the American Medical Informatics Association.

This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (<http://creativecommons.org/licenses/by-nc/4.0/>), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited. For commercial re-use, please contact journals.permissions@oup.com

230





pysynth

```

events:
- name: A
  description: Patient creation
  type: person
  values:
- name: person_id
  type: seq
- name: gender_concept_id
  type: list
  params:
- val: 8532
  prob: 0.45
- val: 8157
  prob: 0.55
- name: age
  type: dist
  params:
  mean: 70
  std: 0.45

```

```

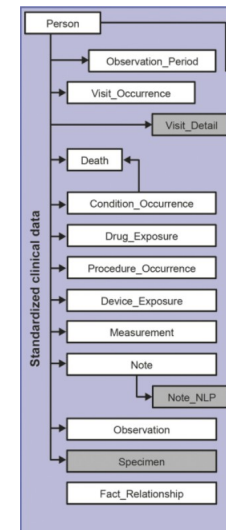
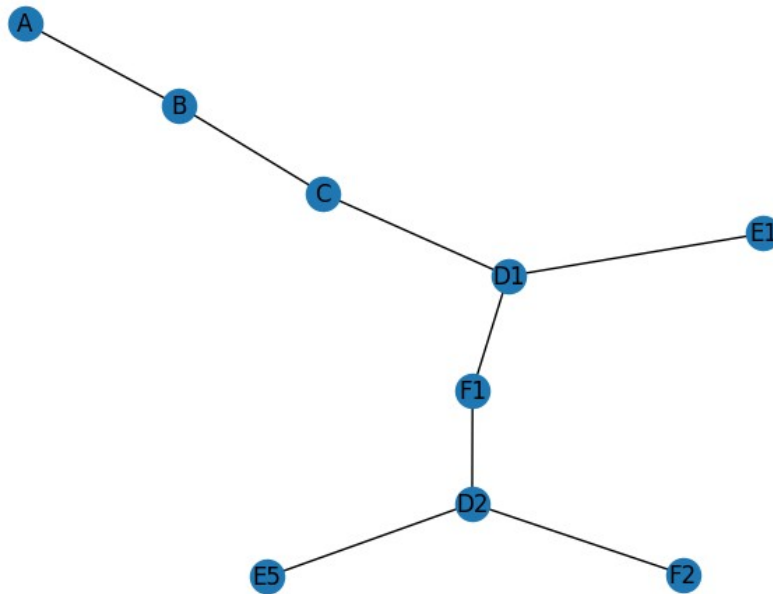
- name: B
  description: First visit
  type: visit_occurrence
  values:
- name: person_id
  type: parent
- name: visit_occurrence_id
  type: seq
- name: visit_datetime
  type: date_between
  params:
  start: '2019-01-01'
  end: '2021-12-31'
- name: visit_concept_id
  type: list
  params:
- val: 9203
  prob: 0.15
- val: 9201
  prob: 0.85

```

```

- name: C
  type: condition_occurrence
  description: Diagnosis
  values:
- name: condition_occurrence_id
  type: seq
- name: person_id
  type: parent
- name: visit_occurrence_id
  type: parent
- name: condition_start_datetime
  type: parent
  params:
  field: visit_datetime
- name: condition_concept_id
  type: list
  params:
- val: 432837
  prob: 0.15
- val: 4247719
  prob: 0.14

```



Generative Adversarial Nets

Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley,
Sherjil Ozair,[†] Aaron Courville, Yoshua Bengio[‡]
Département d'informatique et de recherche opérationnelle
Université de Montréal
Montréal, QC H3C 3J7

Abstract

We propose a new framework for estimating generative models via an adversarial process, in which we simultaneously train two models: a generative model G that captures the data distribution, and a discriminative model D that estimates the probability that a sample came from the training data rather than G . The training procedure for G is to maximize the probability of D making a mistake. This framework corresponds to a minimax two-player game. In the space of arbitrary functions G and D , a unique solution exists, with G recovering the training data distribution and D equal to $\frac{1}{2}$ everywhere. In the case where G and D are defined by multilayer perceptrons, the entire system can be trained with backpropagation. There is no need for any Markov chains or unrolled approximate inference networks during either training or generation of samples. Experiments demonstrate the potential of the framework through qualitative and quantitative evaluation of the generated samples.

1 Introduction

The promise of deep learning is to discover rich, hierarchical models [2] that represent probability distributions over the kinds of data encountered in artificial intelligence applications, such as natural images, audio waveforms containing speech, and symbols in natural language corpora. So far, the most striking successes in deep learning have involved discriminative models, usually those that map a high-dimensional, rich sensory input to a class label [14, 22]. These striking successes have primarily been based on the backpropagation and dropout algorithms, using piecewise linear units [19, 9, 10] which have a particularly well-behaved gradient. Deep *generative* models have had less of an impact, due to the difficulty of approximating many intractable probabilistic computations that arise in maximum likelihood estimation and related strategies, and due to difficulty of leveraging the benefits of piecewise linear units in the generative context. We propose a new generative model estimation procedure that sidesteps these difficulties.¹

In the proposed *adversarial nets* framework, the generative model is pitted against an adversary: a discriminative model that learns to determine whether a sample is from the model distribution or the data distribution. The generative model can be thought of as analogous to a team of counterfeiters, trying to produce fake currency and use it without detection, while the discriminative model is analogous to the police, trying to detect the counterfeit currency. Competition in this game drives both teams to improve their methods until the counterfeits are indistinguishable from the genuine articles.

^{*}Jean Pouget-Abadie is visiting Université de Montréal from Ecole Polytechnique.

[†]Sherjil Ozair is visiting Université de Montréal from Indian Institute of Technology Delhi

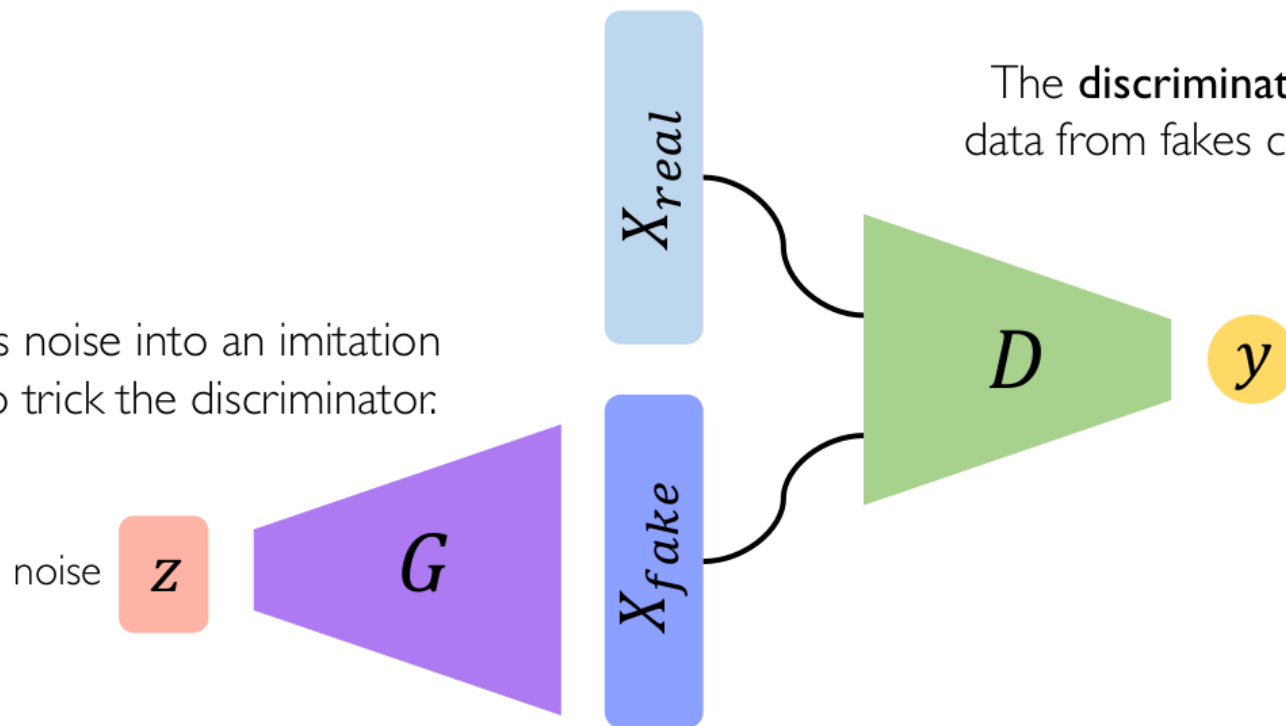
[‡]Yoshua Bengio is a CIFAR Senior Fellow.

¹All code and hyperparameters available at <http://www.github.com/goodfeli/adversarial>

GANs

Generative Adversarial Networks (GANs) are a way to make a generative model by having two neural networks compete with each other.

The **generator** turns noise into an imitation of the data to try to trick the discriminator.



The **discriminator** tries to identify real data from fakes created by the generator.

A review of Generative Adversarial Networks for Electronic Health Records: applications, evaluation measures and data sources

Ghadeer Ghosheh[†], Jin Li[†], and Tingting Zhu[†]

[†]Department of Engineering Sciences, University of Oxford
{ghadeer.ghosheh, jin.li, tingting.zhu}@eng.ox.ac.uk

ABSTRACT

Electronic Health Records (EHRs) are a valuable asset to facilitate clinical research and point of care applications; however, many challenges such as data privacy concerns impede its optimal utilization. Deep generative models, particularly, Generative Adversarial Networks (GANs) show great promise in generating synthetic EHR data by learning underlying data distributions while achieving excellent performance and addressing these challenges. This work aims to review the major developments in various applications of GANs for EHRs and provides an overview of the proposed methodologies. For this purpose, we combine perspectives from healthcare applications and machine learning techniques in terms of source datasets and the fidelity and privacy evaluation of the generated synthetic datasets. We also compile a list of the metrics and datasets used by the reviewed works, which can be utilized as benchmarks for future research in the field. We conclude by discussing challenges in GANs for EHRs development and proposing recommended practices. We hope that this work motivates novel research development directions in the intersection of healthcare and machine learning.

1 Introduction

Over the past decade, machine learning (ML) models have proven to have a high potential for supporting medical applications by using data collected in electronic health records (EHRs) [1, 2]. Hospitals and medical providers are increasingly adopting and deploying EHR systems. In the US alone, 84% of hospitals adopted EHR systems as of 2015, which is a 9-fold increase since 2008 [3]. The widespread recording of structured EHRs is paving the way for research opportunities in healthcare applications, such as patient-stratification [4], drug repurposing [5], public health surveillance [6], as well as the novel discovery of disease mechanisms and correlations as seen in the early COVID-19 applications [7]. EHRs also provide a valuable asset to develop data-driven and patient-specific clinical decision support systems (CDSS) for diagnostic, prognostic, healthcare cost containment and workflow improvement applications [8–10]. However, the full utilization of the wealth of the EHR data in such applications is impeded by several challenges, including data sharing and privacy concerns [11], where data protection guidelines and regulations such as the Health Insurance Portability and Accountability Act (HIPAA) [12] in the United States, and the General Data Protection Regulation (GDPR) [13] in Europe have detailed controlling measures that prevent direct access to much of the data for patient privacy purposes. Other data-specific challenges that make EHR processing burdensome include class imbalance [14], data missingness [15], noise [16], heterogeneity [17] and irregular sampling [18]. To mitigate these challenges, deep generative models have been proposed to generate synthetic data [19], notably variational autoencoders (VAE) [20], and Generative Adversarial Networks (GANs) [21].

In this paper, we review GANs as one of the most widely-used yet under-studied deep generative frameworks, specifically in the domain of EHR applications. There exist several reviews related to GANs evaluation [22], GANs applications for medical imaging [23], and for time-series signals [24] and observational health data [25]. However, in this review, we focus on GANs for structured EHRs, their applications, evaluation and challenges, which serves as a basis for a reading audience with diverse backgrounds. Furthermore, we provide a comprehensive and up-to-date review of the current works and group them based on their target application for healthcare applications, not only for generating synthetic samples, but also mitigating many of the data challenges of EHRs. To the best of our knowledge, this is the first work to discuss and categorise the wide range of evaluation metrics used for evaluating the quality of the synthetic EHRs data generated by GANs. We discuss several open-ended challenges and themes in the current works to motivate new research directions in both the computational and healthcare fields. Relevant literature was identified by searching Google Scholar using the keywords "GAN" AND "EHR", or "synthetic health data", and "GAN" AND "Health" up until January 2022. We then filtered out papers that used generative models other than GANs, and any duplicate papers.

The outline of the paper is as follows. In section 2, we briefly review the working principles and architecture of GANs and



OPEN **BEHRT: Transformer for Electronic Health Records**

Yikuan Li^{1,2}, Shishir Rao^{1,2,✉}, José Roberto Ayala Solares², Abdelaali Hassaine², Rema Ramakrishnan², Dexter Canoy², Yajie Zhu², Kazem Rahimi² & Gholamreza Salimi-Khorshidi²

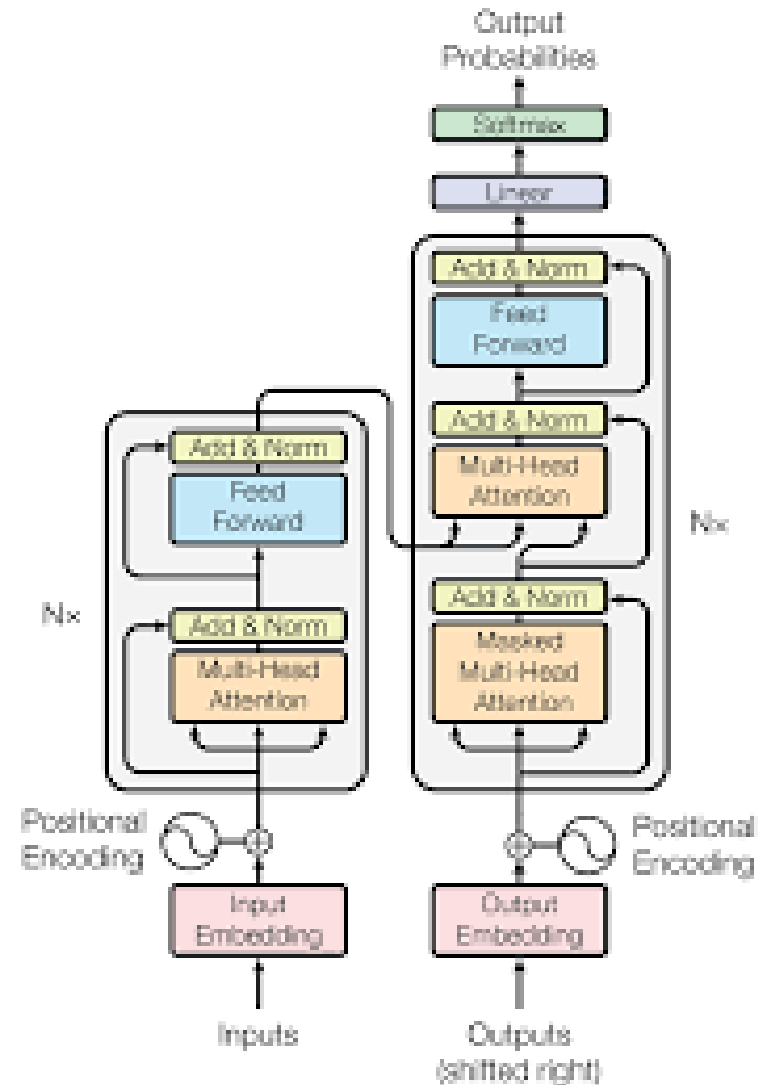
Today, despite decades of developments in medicine and the growing interest in precision healthcare, vast majority of diagnoses happen once patients begin to show noticeable signs of illness. Early indication and detection of diseases, however, can provide patients and carers with the chance of early intervention, better disease management, and efficient allocation of healthcare resources. The latest developments in machine learning (including deep learning) provides a great opportunity to address this unmet need. In this study, we introduce BEHRT: A deep neural sequence transduction model for electronic health records (EHR), capable of simultaneously predicting the likelihood of 301 conditions in one's future visits. When trained and evaluated on the data from nearly 1.6 million individuals, BEHRT shows a striking improvement of 8.0–13.2% (in terms of average precision scores for different tasks), over the existing state-of-the-art deep EHR models. In addition to its scalability and superior accuracy, BEHRT enables personalised interpretation of its predictions; its flexible architecture enables it to incorporate multiple heterogeneous concepts (e.g., diagnosis, medication, measurements, and more) to further improve the accuracy of its predictions; its (pre-)training results in disease and patient representations can be useful for future studies (i.e., transfer learning).

The field of precision healthcare aims to improve the provision of care through precise and personalised prediction, prevention, and intervention. In recent years, advances in deep learning (DL) - a subfield of machine learning (ML) - has led to great progress towards personalised predictions in cardiovascular medicine, radiology, neurology, dermatology, ophthalmology, and pathology. For instance, Ardila *et al.*¹ introduced a DL model that can predict the risk of lung cancer from a patient's tomography images with 94.4% accuracy; Poplin *et al.*² showed that DL can predict a range of cardiovascular risk factors from just a retinal fundus photograph, and more examples can be found in other works^{3,4}. A key contributing factor to this success, in addition to the developments in DL algorithms, was the massive influx of large multimodal biomedical data, including but not limited to, electronic health records (EHR)⁵.

The adoption of EHR systems has greatly increased in recent years; hospitals that have adopted EHR systems now exceeds 84% and 94% in the US and UK, respectively^{6,7}. As a result, EHR systems of a national (and/or a large) medical organisation now are likely to capture data from millions of individuals over many years. Each individual's EHR can link data from many sources (e.g., primary and hospital care) and hence contain "concepts" such as diagnoses, interventions, lab tests, clinical narratives, and more. Each instance of a concept can mean a single or multiple data points. Just a single hospitalisation, for instance, can generate thousands of data points for an individual, whereas a diagnosis can be a single data point (i.e., a disease code). This makes large-scale EHR a uniquely rich source of insight and an unrivalled data for training data-hungry ML models.

In traditional research on EHR data (including the ones using ML), individuals are represented by models as a vector of attributes, or "features"⁸. This approach relies on experts' ability to define the appropriate features and design the model's structure (i.e., answering questions such as "what are the key features for this prediction?" or, "which features should have interactions with one another?"). Recent developments in deep learning, however, provided us with models that can learn useful representations (e.g., of individuals or concepts) from raw or minimally-processed data, with minimal need for expert guidance⁹. This happens through a sequence of layers, each employing a large number of simple linear and nonlinear transformations to map their corresponding inputs to a representation; this progress across layers results in a final representation in which the data points form distinguishable patterns.

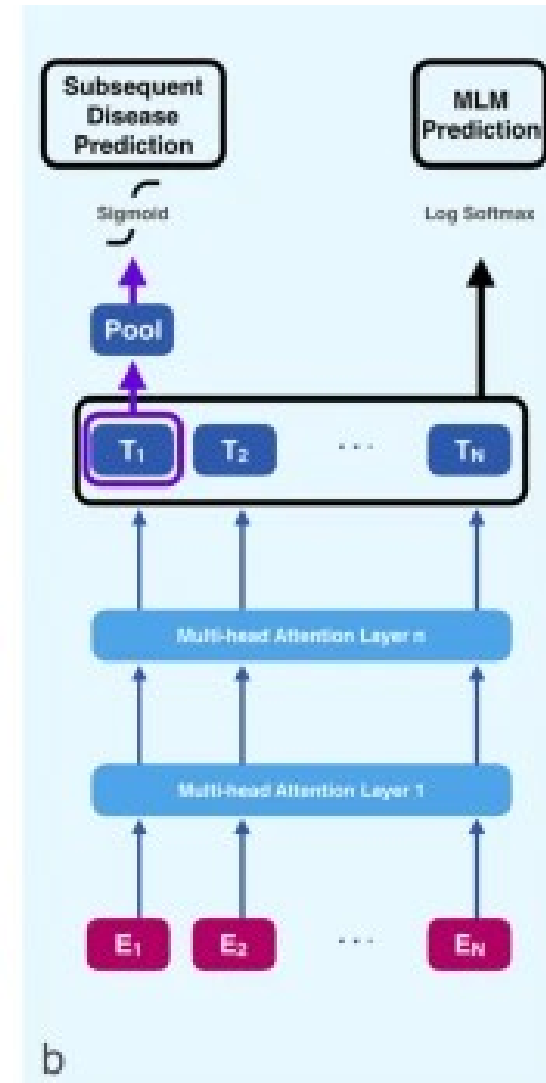
¹Deep Medicine, Oxford Martin School, University of Oxford, Oxford, United Kingdom. ²These authors contributed equally: Yikuan Li and Shishir Rao. ✉e-mail: shishir.rao@stctc.ox.ac.uk



Embedding Diagram and BEHRT Architecture



a



b

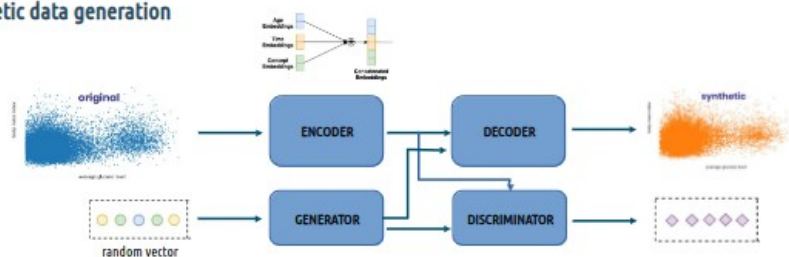
Generating Synthetic Data from OMOP-CDM Databases for Health Applications

A. Labarga^{1,2}, S. Aguiló-Castillo^{1,2}, S. Capella-Gutiérrez^{1,2}
¹Barcelona Supercomputing Center (BSC), Barcelona Spain.
²Spanish National Bioinformatics Institute (INB/ELIXIR-ES).



Analysis of Electronic Health Records (EHR) has a tremendous potential for enhancing patient care, quantitatively measuring performance of clinical practices, and facilitating clinical research. Statistical estimation and machine learning (ML) models trained on EHR data can be used to predict the probability of various diseases (such as diabetes), track patient wellness, and predict how patients respond to specific drugs. For such models, researchers and practitioners need access to EHR data. However, it can be challenging to leverage EHR data while ensuring data privacy and conforming to patient confidentiality regulations. Here we present an approach for generating synthetic health data from an OMOP-CDM. The goal of this study was to develop and evaluate a model for simulating longitudinal healthcare data that adequately captures clinical data temporal and conditional complexities.

Synthetic data generation

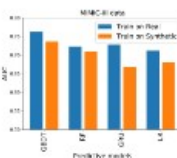
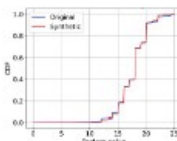


Generating synthetic data comes down to learning the joint probability distribution in an original, real dataset to generate a new dataset with the same distribution. Deep learning models such as **generative adversarial networks (GAN)** and **variational autoencoders (VAE)** are well suited for synthetic data generation but have problems capturing temporal and causal dependencies in the data or generating categorical variables common in clinical data.

We propose a novel generative modeling framework that combines GANs with a **bidirectional encoder representations from transformers (BERT)** architecture. We first train the encoder-decoder model using a reconstruction loss. Then, we use the trained encoder to transform the original inputs into latent space (encoder states). Lastly, we train the GAN framework using an adversarial loss in the latent space to incorporate temporal data across multiple clinical domains. We use a hybrid approach by augmenting the input to BERT using artificial time tokens, incorporating time, age, and concept embeddings, and introducing a new second learning objective for visit type.

Quality evaluation

FIDELITY	UTILITY	PRIVACY
How similar is this synthetic data as compared to the original training sets	How useful is this synthetic data for our downstream machine learning applications	Has any sensitive data been inadvertently synthesized by our model
Kullback-Leibler (KL) divergence, pairwise correlation difference	Accuracy, F1-score, ROC, and AUC-ROC	Membership inference, re-identification and attribute inference attacks



References

1. Murray Reet al. Design and validation of a data simulation model for longitudinal healthcare data. AMIA Annu Symp Proc. 2011
2. Pang et al. CEHR-BERT: Incorporating temporal information from structured EHR data to improve prediction tasks. Proc. of Machine Learning for Health, 158, 2021.
3. Yoon et al. EHR-Safe: Generating High-Fidelity and Privacy-Preserving Synthetic Electronic Health Records <https://doi.org/10.21203/rs.3.rs-2347130/v1>

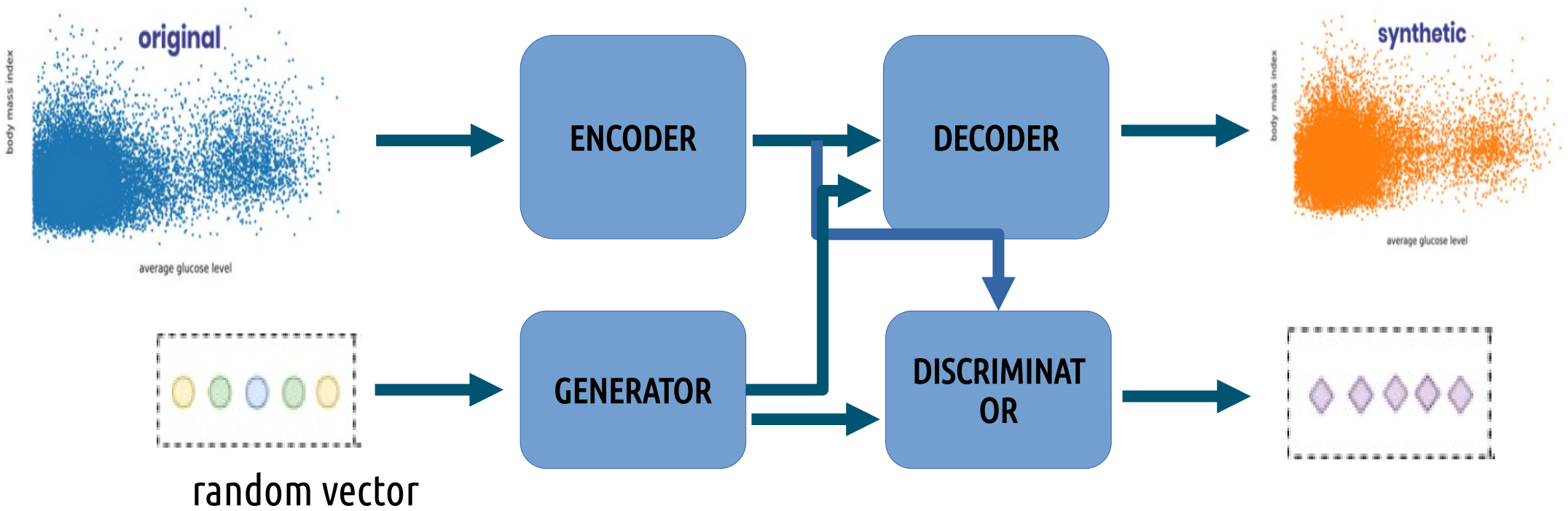
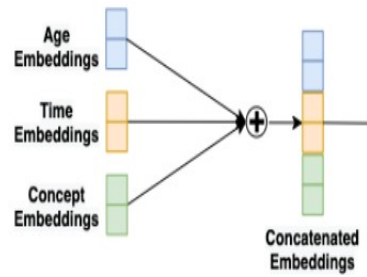
Contact

Alberto Labarga
 alberto.labarga@bsc.es

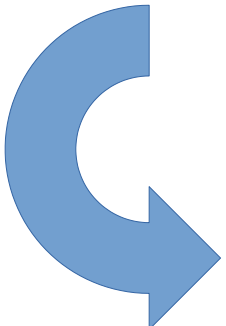
Barcelona Supercomputing Center (BSC)
 Plaça Euzkadi 161, 1.3. 08034 Barcelona, Spain

Funding






Synthetic notes



Acude con dolor abdominal derecho de 2 días de duración. Pauta de vacunación completa.



Finding *dolor abdominal.*

Spatial Concept *derecho.*

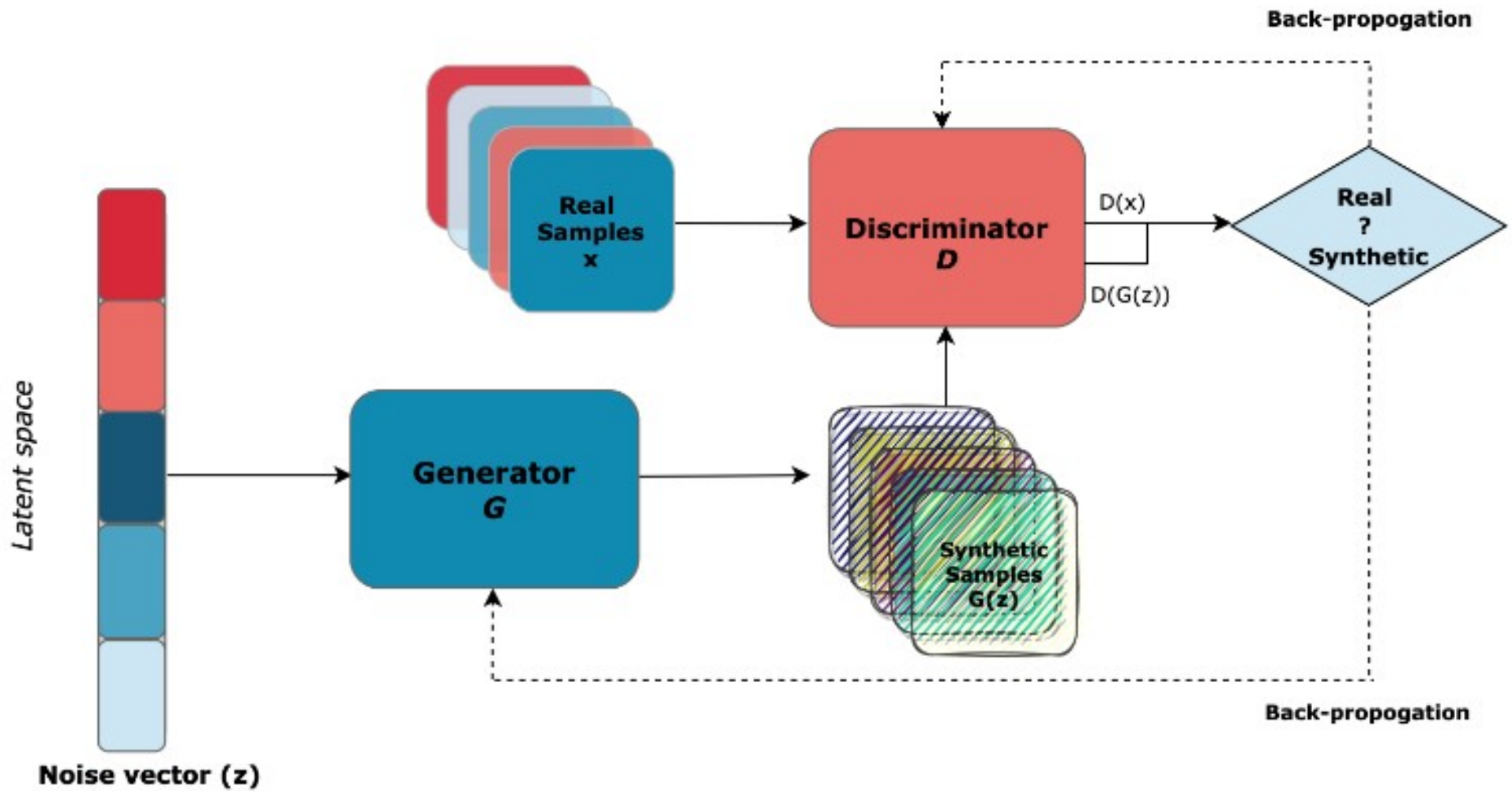
Temporal Concept *2 días, duración.*

Therapeutic or Preventive Procedure *vacunación.*

Qualitative Concept *completa.*

Synthetic images





Denoising Diffusion Probabilistic Models

Jonathan Ho
UC Berkeley
jonathanho@berkeley.edu

Ajay Jain
UC Berkeley
ajayj@berkeley.edu

Pieter Abbeel
UC Berkeley
pabbeel@cs.berkeley.edu

Abstract

We present high quality image synthesis results using diffusion probabilistic models, a class of latent variable models inspired by considerations from nonequilibrium thermodynamics. Our best results are obtained by training on a weighted variational bound designed according to a novel connection between diffusion probabilistic models and denoising score matching with Langevin dynamics, and our models naturally admit a progressive lossy decompression scheme that can be interpreted as a generalization of autoregressive decoding. On the unconditional CIFAR10 dataset, we obtain an Inception score of 9.46 and a state-of-the-art FID score of 3.17. On 256x256 LSUN, we obtain sample quality similar to ProgressiveGAN. Our implementation is available at <https://github.com/hojonathanho/diffusion>.

1 Introduction

Deep generative models of all kinds have recently exhibited high quality samples in a wide variety of data modalities. Generative adversarial networks (GANs), autoregressive models, flows, and variational autoencoders (VAEs) have synthesized striking image and audio samples [14, 27, 9, 58, 38, 23, 10, 32, 44, 57, 26, 33, 45], and there have been remarkable advances in energy-based modeling and score matching that have produced images comparable to those of GANs [11, 55].

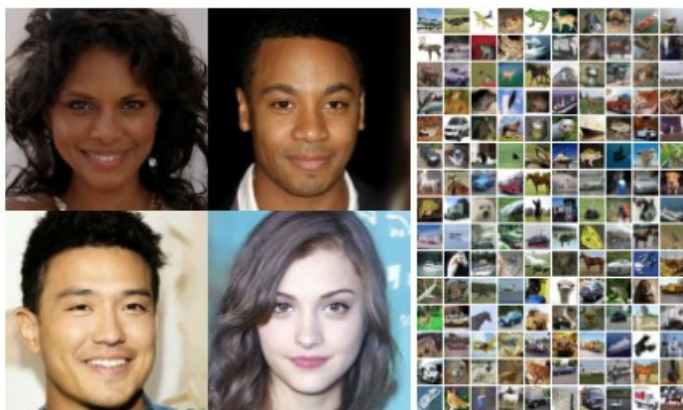
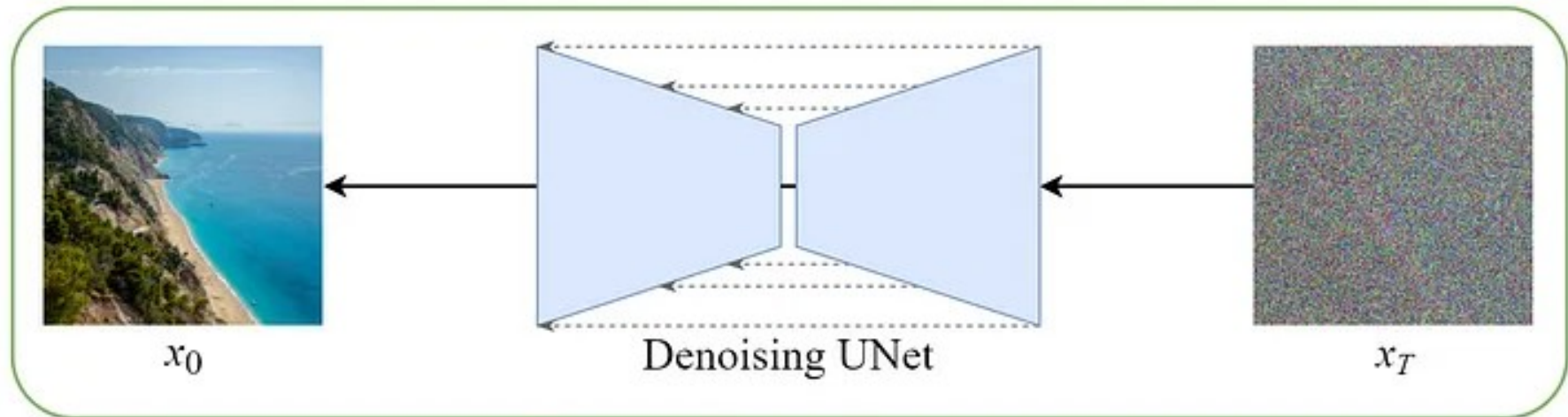
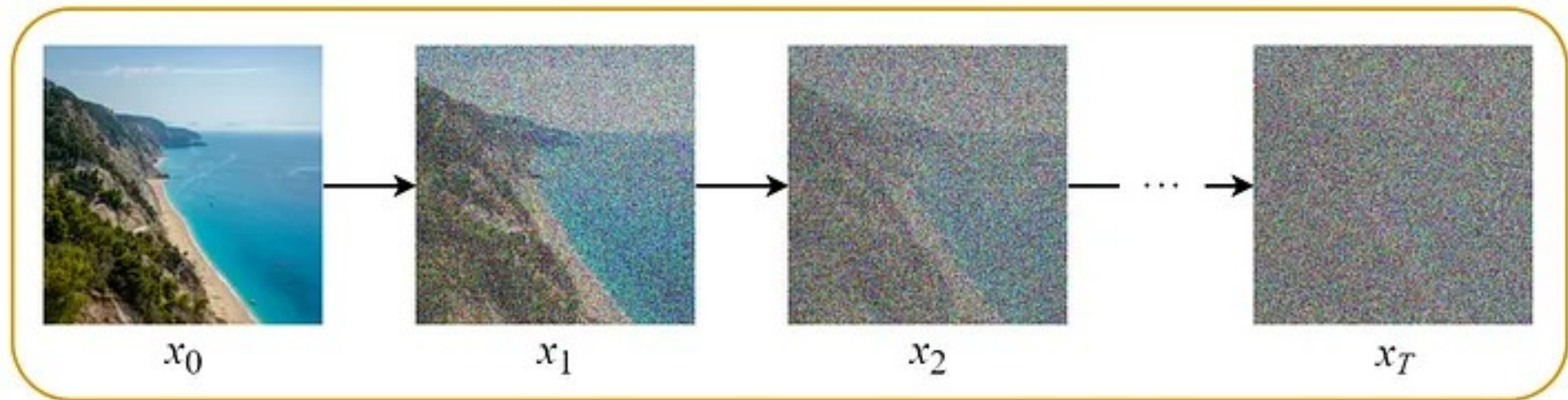
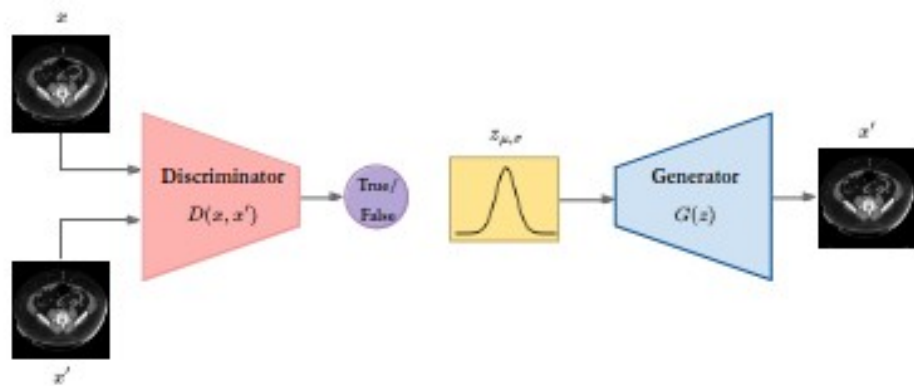


Figure 1: Generated samples on CelebA-HQ 256×256 (left) and unconditional CIFAR10 (right)

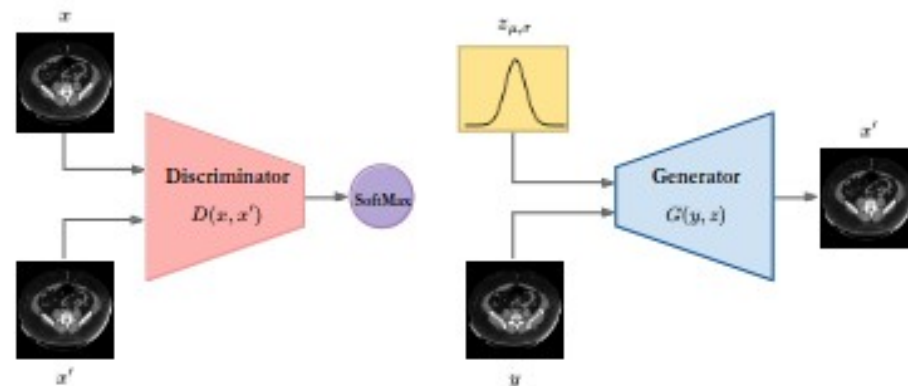
Forward Diffusion Process



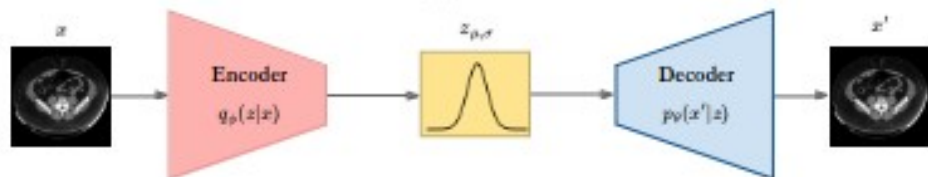
Reverse Diffusion Process



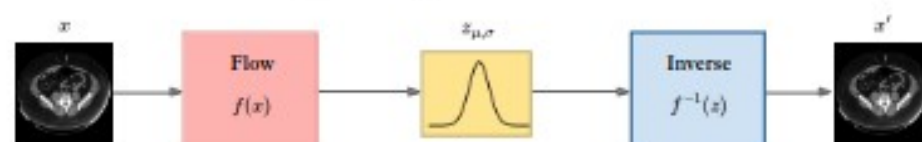
(a) GAN



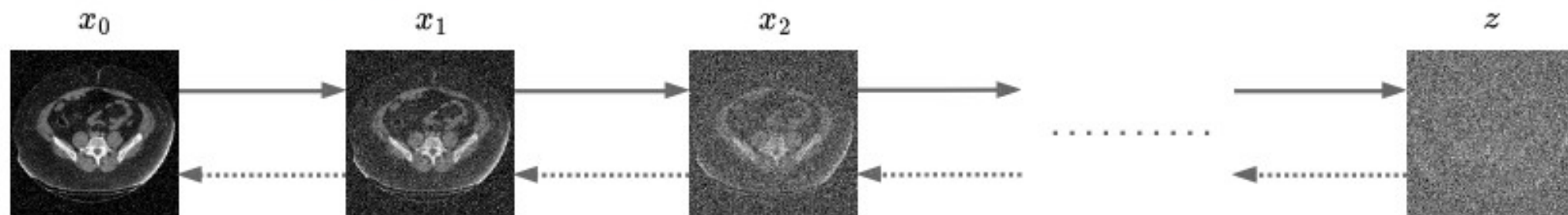
(b) Energy-based Models



(c) VAE



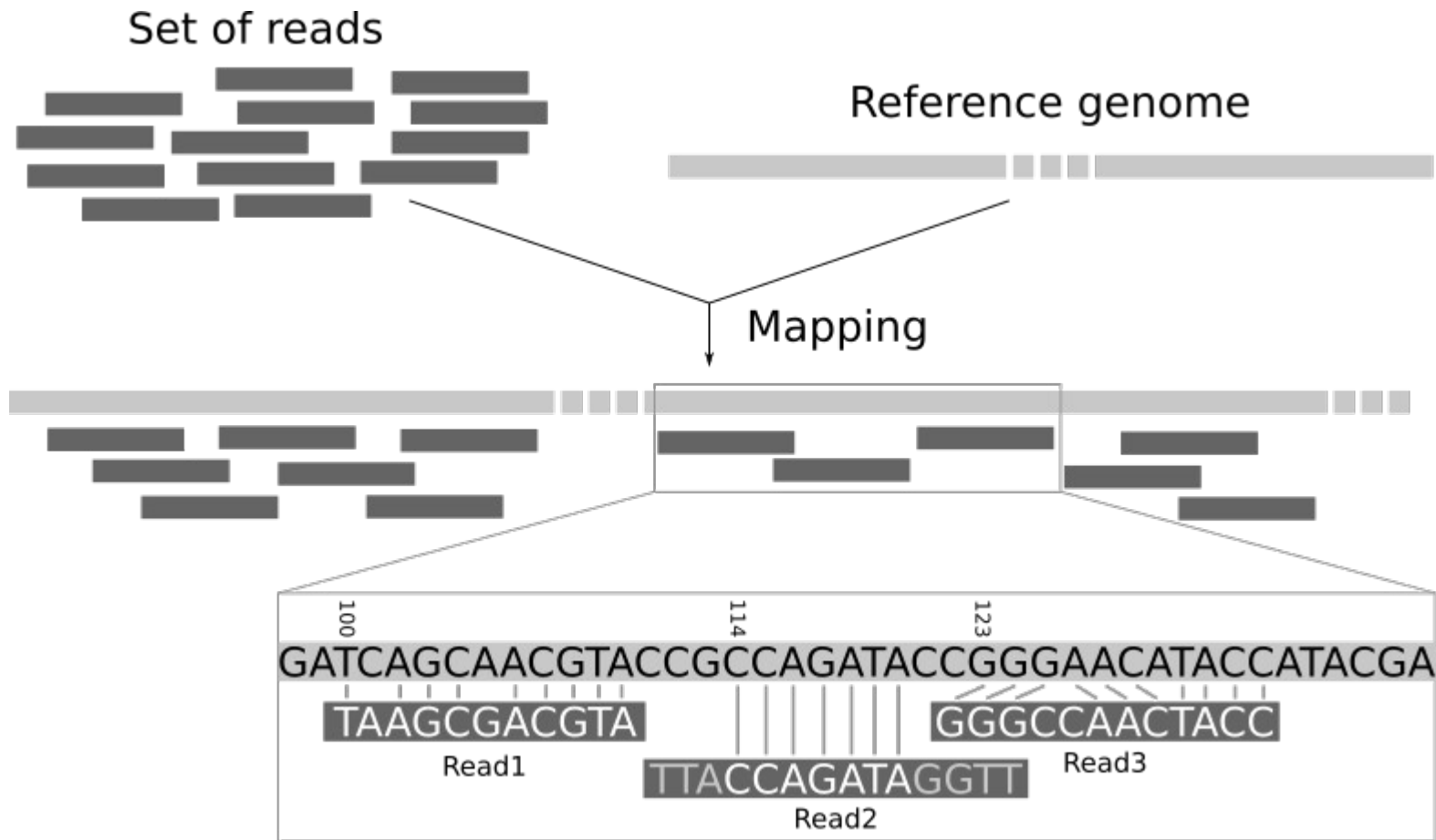
(d) Flow-based models

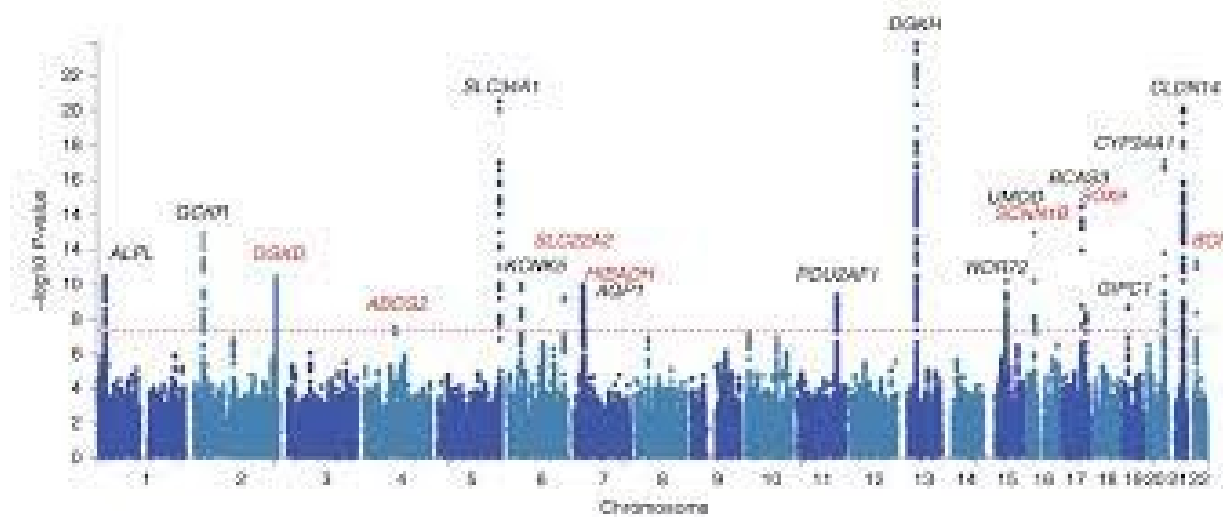
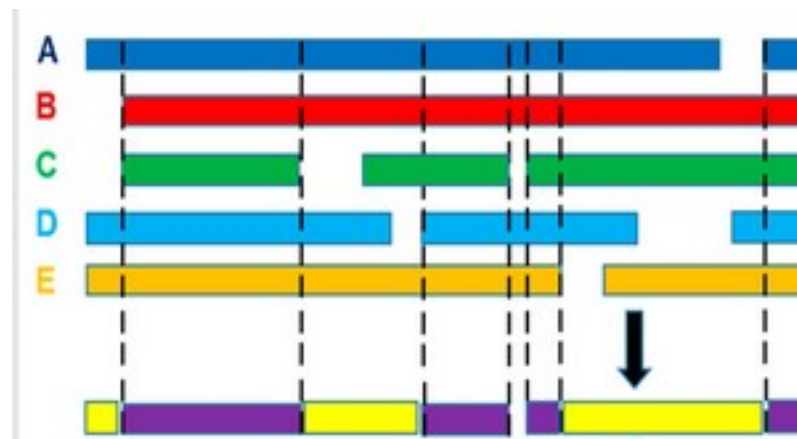
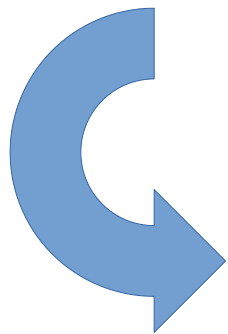


(e) Diffusion models

Synthetic genomic data

Characteristic	ART	DWGSIM	ISS	Mason	NEAT	wgsim
Accurate read count	+	+	- (over-sampled)	+	- (over-sampled)	o (slightly under-sampled)
Uniform sampling across the genome	- (sampled equal # reads per chromosome)	+	+	+	+	+
Genomic coverage	o (- MiSeq v.3; + all other models)	+	o (- MiSeq model; + all other models)	+	+	+
Edge effects	+	+	+	+	+	+
Fragment length	+	+	-	+	+	+
Mapping	+	+	o	+	+	+
"Golden" .bam ^a	+	-	-	+	+	-





Evaluation

FIDELITY

How similar is this synthetic data as compared to the original training sets

Kullback-Leibler (KL) divergence, pairwise correlation difference

UTILITY

How useful is this synthetic data for our downstream machine learning applications

Accuracy, F1-score, ROC, and AUC-ROC

PRIVACY

Has any sensitive data been inadvertently synthesized by our model

Membership inference, re-identification and attribute inference attacks



Infrastructure for synthetic health data

Núria Queralt-Rosinach¹, Basel Alshikhdeeb², Luca Bolzani², Muhammad Shoaib², Marcos Casado Barbero⁴, Sergi Aguiló-Castillo³, Davide Cirillo³, Leyla Jael Castro⁵, Ginger Tsueng⁶, Matúš Kaláš⁷, Magnus Palmblad⁸, Danielle Welter², Soumyabrata Ghosh², Venkata Pardhasaradhi Satagopam², and Rahuman S. Malik Sheriff⁴

¹ Human Genetics, Leiden University Medical Center, Leiden, Netherlands ² Luxembourg Center for Systems Biomedicine, University of Luxembourg, Luxembourg ³ Barcelona Supercomputing Center, Barcelona, Spain ⁴ European Molecular Biology Laboratory - European Bioinformatics Institute (EMBL-EBI), Hinxton CB10 1SD, UK ⁵ ZB MED Information Centre for Life Sciences, Cologne, Germany ⁶ Scripps Research Institute, La Jolla, CA 92037, US ⁷ Department of Informatics, University of Bergen, Norway ⁸ Center for Proteomics and Metabolomics, Leiden University Medical Center, Leiden, Netherlands

Introduction

Machine learning (ML) methods are becoming ever more prevalent across all domains of life sciences. However, a key component of effective ML is the availability of large datasets that are diverse and representative. In the context of health systems, with significant heterogeneity of clinical phenotypes and diversity of healthcare systems, there exists a necessity to develop and refine unbiased and fair ML models. Synthetic data are increasingly being used to protect the patient's right to privacy and overcome the paucity of annotated open-access medical data. Synthetic data and generative models can address these challenges while advancing the use of ML in healthcare and research.

Following up the efforts currently undertaken in the ELIXIR Health Data and the Machine Learning Focus Groups around the synthetic health-data landscape, this project focuses on the health data providers' need for a ready-to-use synthetic data platform assessed by health data experts, researchers, and ML specialists. Aligned with ELIXIR Health Data Focus Group's objectives, we aim at building an infrastructure for synthetic health data offering a containerised synthetic data generator based on the open-source libraries [Synthetic Data Vault \(SDV\)](#) and [ydata-synthetic](#) with state-of-the-art ML methods. This framework will enable users to generate synthetic data that have the same structure and statistical properties as the original dataset from a variety of sources (clinical, variational, or omics). Despite the capacity to generate their own datasets, a set of exemplary datasets will be publicly available in appropriate repositories and will include rich metadata descriptions according to the [DOME recommendations](#) and [GA4GH standards](#). [OpenEBench](#) will host a community of practice for comparing different approaches for synthetic data generation. Here, we present our proof of concept for the generation of synthetic health data and our proposed FAIR implementation of the generated synthetic datasets. The work was developed during and after the one-week-long BioHackathon Europe, by together 20 participants (10 new to the project), from different countries (NL, ES, LU, UK, GR, FL, DE, ...).

Infrastructure for synthetic health data

For test and stress development of new ML methods/tools, we need suitable data to properly demonstrate the method/tools application. However, ML developers without health data access are not able to see how the tool performs for its intended application. One way to enable this is to generate *synthetic data*, where new "fake" data is created from real data using a specifically designed generation model. Importantly, the generation model must maintain the

- Generation
- Evaluation
- Publication

BioHackathon series:
[BioHackathon Europe 2022](#)
Paris, France, 2022
[Project 15](#)

Submitted: 21 Jul 2023

License:
Authors retain copyright and
release the work under a Creative
Commons Attribution 4.0
International License (CC-BY).

Published by [BioHackrXiv.org](#)