INTERVENE

Deliverable 1.3 - Access to first INTERVENE consortium data sets available from EMBL- EBI and CSC Cloud services

Deliverable 1.3

Dissemination level: Public

Part of:

WP1:

Compliant and standardised data access for federated analyses

Project summary

Project acronym:	INTERVENE
Project full title:	International consortium for integrative genomics prediction
Project Coordinator	Institute for Molecular Medicine Finland FIMM, University of Helsinki; Prof. Samuli Ripatti and Dr. Andrea Ganna
Project start date:	1.1.2021
Project end date:	31.12.2025
Project duration:	60 months
Action type:	RIA
Call identifier:	H2020-SC1-FA-DTS-2018-2020 (Trusted digital solutions and Cybersecurity in Health and Care)
Grant number	101016775

Document descriptors

Deliverable No.	1.3
Work package	WP1
Deliverable lead	EMBL-EBI
Contributors	EMBL-EBI, CSC
Dissemination level	Public
Expected delivery date	30/09/2021
Submission date	30/09/2021

Change history log

Version	Changes made	Date	Prepared by	PC approved
0.1	Outline created	29/06/21	Helen Parkinson, EMBL-EBI	
0.2	Draft written	11/09/2021	Aoife McMahon, EMBL-EBI	
0.3	Addition	21/9/21	Kimmo Mattila, CSC	
0.4	Revision and extension	21/9/21	Helen Parkinson, EMBL-EBI	
0.5	Revision	22/9/21	Aoife McMahon, EMBL-EBI	
0.6	Review	29/9/21	Ilkka Lappalainen, Dylan Spalding, CSC	
1.0	Edits, technical formatting	30/0/21	Julius Anckar, FIMM	Yes

Contents

1. Introduction

This Deliverable relates to Task 1.2 - *Access to the INTERVENE consortium data* and addresses complementary objectives from WP1 and WP4 regarding data access supported by platform deployment. It facilitates access to sensitive data housed in federated secure data stores and provides it for consortium access within a secure cloud-based analysis environment. This is essential for delivering INTEVENE data analysis features such as the calculation of polygenic scores of INTERVENE cohorts and user supplied data, as well as the future artificial intelligence (AI) development activities and competition (WP3/4).

The two institutions, CSC and EMBL-EBI contribute to the two current federation nodes. These nodes are described in detail as the resources are necessarily different, have different features and underlying infrastructure. However, both are predicated on the use of Federated European Genome-phenome Archive[1] [2] (FEGA), the backend software architecture for federated analysis of accessible datasets.

The FEGA provides a network of connected resources to enable transnational discovery of and access to human data for research while also respecting jurisdictional data protection regulations. The EGA has central nodes, the founding node at EMBL-EBI and a community of federated nodes of which the CSC hosted node is one of the first[3] (Figure 1A). The benefits of leveraging FEGA for this deliverable are existing and sustainable infrastructure built and operated to Global Alliance for Genomics and Health (GA4GH) standards[4], collaboratively developed and access to existing strategies for data management and a developer community. The use of FEGA technology allows implementation of the GA4GH AAI[5] and Passport[6] standards, which uses authentication and authorisation infrastructure access tokens to transport a researcher's digital identity and permissions across organizations, tools, and environments via the GA4GH compatible ELIXIR Authentication and Authorisation Infrastructure[7], and then maps access to

[1] EGA website: https://ega-archive.org

[2] Lappalainen et al, 2015: https://www.nature.com/articles/ng.3312

[3] Federated EGA documentation: https://ega-archive.org/federated

[4] GA4GH: https://www.ga4gh.org

[5] GA4GH AAI: https://github.com/ga4gh/data-security/tree/master/AAI

[6] GA4GH Passport documentation and code: https://github.com/ga4gh-duri/ga4gh-duri.github.io/blob/master/researcher_ids/ga4gh_passport_v1.md

[7] https://elixir-europe.org/services/compute/aai

data across these (Figure 1B). Further, by reusing components and aligning with international standards the sustainability of the platform is enhanced (WP8). FEGA also underlies the B1MG[8] and CINECA[9] projects, providing access to pan-European cohorts similar in scale to INTERVENE; by use of shared technology and synthetic data we support sustainability by extending prior work.

As described in Deliverable 1.2 - *Review of the consortium partner biobank and data collections, including access policies* (Table 1; Appendix 4.1), most partner biobanks restrict data moving outside their local environments. Access to these data will be addressed in a Deliverable 1.4 - *Access to a biobank data set through implementation of local biobank APIs and cloud resources to support federated consortium access* in June, 2022.

This deliverable addresses access to data that can be moved to nodes of the federated platform (either CSC or EMBL-EBI) this includes UK Biobank, Partners Biobank, user supplied data and synthetic data.

As envisaged in the INTERVENE proposal, we are leveraging synthetic datasets to develop and test the integrity, scalability and sustainability of our analysis environments before sensitive data from the biobanks is used. This reduces risks of data breaches during development and removes lag periods waiting for access to sensitive data ensuring the delivery proceeds as planned.

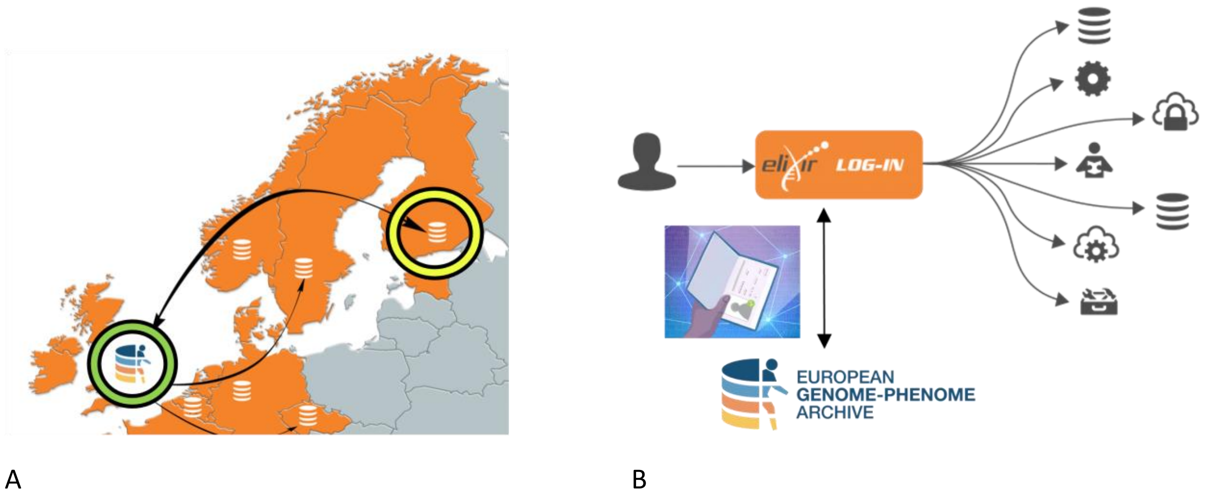

Figure 1. EGA infrastructure. Panel A Illustrates the federation of data access using geographically distant federated EGA (FEGA) instances. Central EGA (circled, green) and a federated node at (circled, yellow) are based at INTERVENE partner institutions, EMBL-EBI and CSC respectively. Panel B provides a schematic of federated identity passports in the context of FEGA.

[8] https://b1mg-project.eu/

[9] https://www.cineca-project.eu/

1.1. Objectives addressed

- WP1 Objective 1. To enable a federated analysis of cross-dataset and cross-institution datasets based on existing efforts in ELIXIR and in other H2020 funded projects
- WP1 Objective 2. To support the deployment of a multi-study analysis platform for testing and development of integrative genetic scores (in collaboration with WP2, WP3, and WP4).
- WP4 Objective 2. To create a platform (IGS4EU) hosted by EMBL-EBI and CSC that houses a series of new, high value capabilities.
- WP4 Objective 3. To enable secure upload of genetic data (in collaboration with WP1), performs the necessary calculations, and returns genetic scores post-analysis to the user.
- WP8 Objective. To support development of the ISG4EU platform and to build its sustainability and trustworthiness in the eyes of end-users

2. Progress to date

2.1. Use cases

We have enumerated use cases illustrating data access scenarios that we will facilitate. These use cases (Appendix 4.2) inform the development of the federated infrastructure and are used, or will be used, to test the implementation. User personas and stories will be added to and refined over the course of the project as they emerge, for example, from other work packages as the project progresses.

There are three primary user personas identified at present (Figure 2):

1. **INTERVENE partner analysts** - an employee of an INTERVENE partner institution. These named individuals will be added to the data access applications to the INTERVENE partner biobanks and will have a role in data analysis for the consortium.
2. **Data custodian** - an external user who is the data controller of the individual level data of a non-INTERVENE cohort and wishes to calculate the polygenic score risk profile of their cohort.
3. **Competition entrant** - an external user who has developed a PG scoring file, for example using an AI methodology, and wishes to evaluate its performance on initially synthetic and subsequently real data.

In the case of competition entrants, (WP4, Task 4.1.3), access to, and analyses of, sensitive data will be performed by INTERVENE analysts using the IGS4EU platform, aggregate results (in the form of scores or reports, developed with WP6) only are returned to users external to the Intervene consortium. Therefore, we have prioritized our implementation for Persona 1, the INTERVENE partner analysis. Handling access for INTERVENE analysts is a foundational requirement for the development of other use cases (Appendix 4.2).

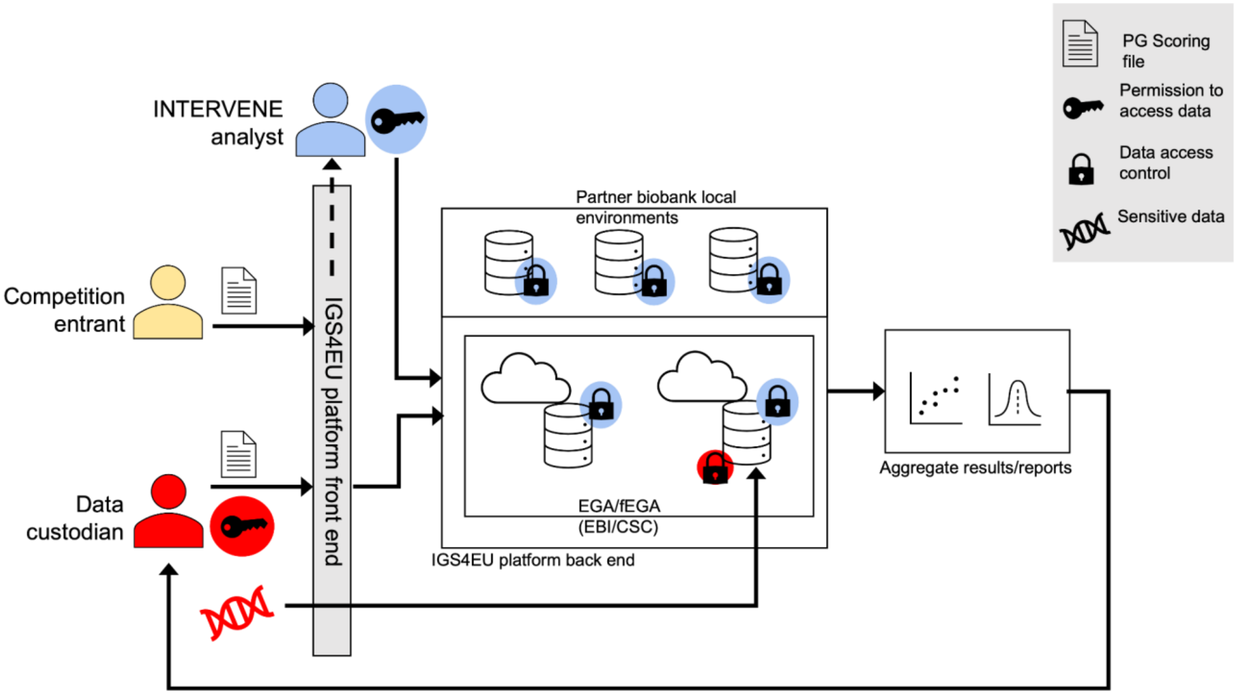

Figure 2. Schematic of user personas and the routes to data access. Intervene analysts (blue) will be included on relevant data access agreements with partner biobanks. Competition entrants/external researchers will not have direct access to sensitive data, but the platform/INTERVENE analysts will initiate analyses on their behalf. Data custodians will upload their data to platform-associated secure environments (EGA/CSC/FEGA) and will be able to initiate analyses only on their own data. Use cases are included in Appendix 4.2.

2.2. Data

As synthetic datasets developed by INTERVENE WP3 were not yet available, we chose to use synthetic data developed by the H2020-funded CINECA project (Common Infrastructure for National Cohorts in Europe, Canada, and Africa)[10]. Specifically, we have used “CINECA synthetic cohort EUROPE UK1”, a dataset consisting of 2504 samples which have genetic data based on 1000 Genomes and 76 synthetic subject attributes and phenotypic data derived from UKBiobank (EGA Dataset ID EGAD00001006673)[11].

These data were chosen as they represent appropriate test data sets for developing the platform and testing federation scenarios. This dataset is archived in EGA, and as such is subject to the same data access processes as controlled access datasets also held in the EGA, despite being synthetic, such as annotation using the GA4GH standard Data Use Ontology[12]. This allows us to test and demonstrate our processes and implementation of processes specific to the EGA.

Beyond its utility for access management, the data includes genotype, subject attribute and phenotypic data, which will be used for testing and benchmarking the pipeline to compute genetic scores (WP4, Task 4.2). Furthermore the data include a variety of errors introduced to mimic real data making it suitable as a test for quality control pipelines.

[10] https://www.cineca-project.eu/cineca-synthetic-datasets
[11] https://ega-archive.org/datasets/EGAD00001006673
[12] https://github.com/EBISPOT/DUO

2.3. EMBL-EBI infrastructure

Infrastructure to upload controlled data from the EGA to Embassy Cloud (using the EGA AAI for permission control)

The core of EBI federated node infrastructure has been developed and consists of an application (the IGS4EU file handler application) running on the EMBL EBI's private Cloud - the Embassy Cloud[13] - which downloads data directly from central EGA (Figure 3). Embassy Cloud workspace is an isolated, secure environment hosted by EMBL-EBI. This Infrastructure as a Service is hosted in a Tier 3+ secure data centre[14], and is logically outside the institute's local area network (LAN) and is therefore appropriate for sensitive human data.

An Embassy tenancy has been established for INTERVENE and has been configured as a Kubernetes cluster. A Kubernetes cluster is a set of nodes that run containerized applications, these are lightweight, flexible and easily portable and are therefore appropriate for our analysis use cases. Kubernetes containers are not restricted to a specific operating system and are able to run anywhere, making them suitable for developing pipelines that will also need to run in biobank local environments.

To download files from the EGA, the cluster is configured to interact with the EGA's (instantiated at EMBL-EBI) Data Distribution API[15] , a GA4GH htsget[16] standard compatible REST API providing secure and controlled access to the archive. As EGA data is encrypted at rest, the Data Distribution API decrypts the data on the fly and transmits this data securely over https. The IGS4EU file handler application requires valid EGA user credentials (obtained from the helpdesk once an EGA requester account is created) and the user must have access to the respective dataset, as specified by the EGA AAI (Authentication & Authorization server). These in turn are specified by the individual data access committee (DAC) associated with each dataset. If the user is authorized for access an access token is obtained from the EGA AAI and passed to the Data Distribution API and data is uploaded to the Embassy Cloud where it is stored securely. Only individual credentials can be used for authentication, since only individuals who are specified on data access agreements (and no other entities such as organisation or projects) can be granted access to datasets, this is an essential security model respecting the data access committee process.

Demonstration

The CINECA synthetic cohort EUROPE UK1 dataset is deposited in the EGA (EGAD00001006673) with the EGA Helpdesk functioning as a DAC to whom the applicant applies for access. The DAC then adds the dataset to the user's credentials via a GA4GH Controlled Access Grants Visa within a GA4GH Passport replicating a typical data access scenario. As a test of the implementation an individual INTERVENE analyst was granted access to the synthetic dataset and used the above IGS4EU file handler application to upload the data to the INTERVENE Embassy cloud. Successful upload of intact data to the Embassy cloud hosted infrastructure was verified by calculation of a file-specific identifier (md5 checksum) which matched the md5 checksum provided by the EGA Data distribution API.

[13] EMBL-EBI Embassy Cloud: https://www.embassycloud.org/about/

[14] Uptime tier classification system: https://journal.uptimeinstitute.com/explaining-uptime-institutes-tier-classification-system/

[15] EGA Data Distribution API documentation: https://github.com/EGA-archive/ega-data-api

[16] http://samtools.github.io/hts-specs/htsget.html

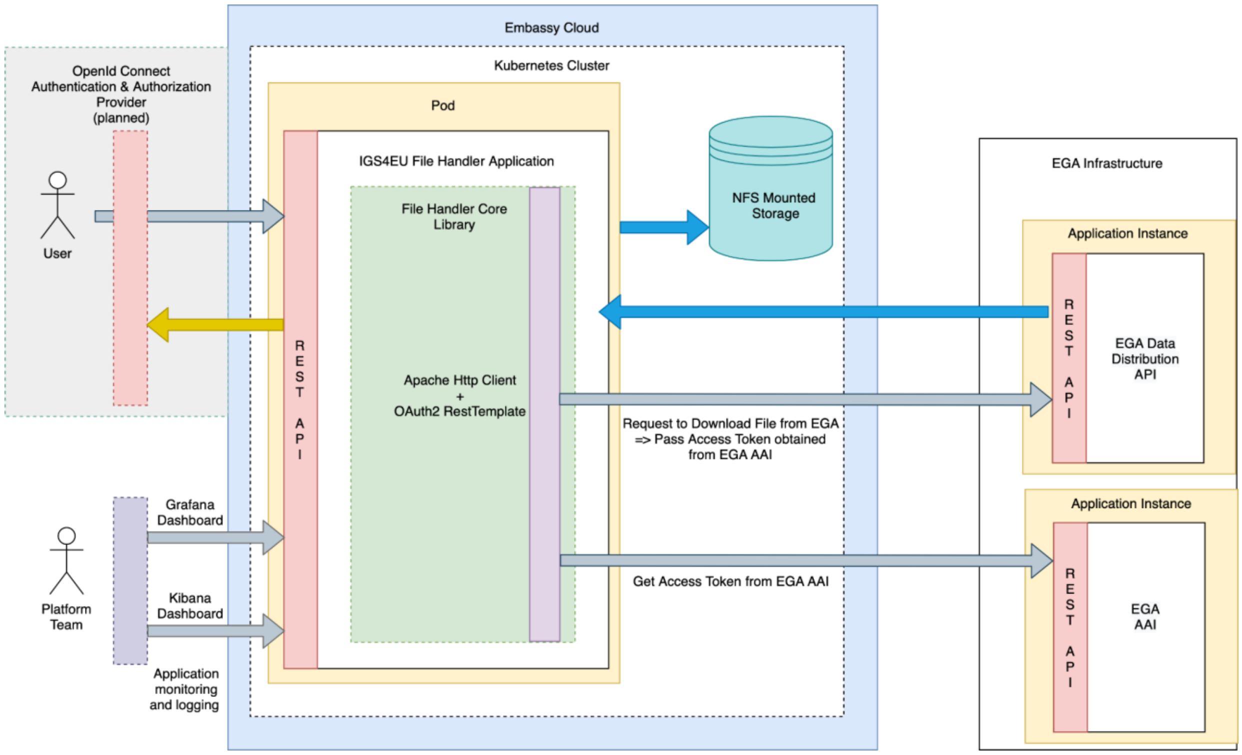
Embassy Cloud
Kubernetes Cluster
OpenId Connect Authentication & Authorization Provider (planned)
User
Pod
IGS4EU File Handler Application
File Handler Core Library
Apache Http Client + OAuth2 RestTemplate
R E S T A P I
NFS Mounted Storage
EGA Infrastructure
Application Instance
R E S T A P I
EGA Data Distribution API
Request to Download File from EGA => Pass Access Token obtained from EGA AAI
Application Instance
R E S T A P I
EGA AAI
Get Access Token from EGA AAI
Grafana Dashboard
Kibana Dashboard
Platform Team
Application monitoring and logging

Figure 3. EBI infrastructure including the IGS4EU file handler application. A schematic representation of the Kubernetes deployed Intervene infrastructure running on Embassy Cloud which authenticates and downloads files from EGA and stores files on NFS shared storage. Grey arrows indicate API calls. Coloured arrows indicate data flow, blue arrows represent potentially sensitive data, yellow arrows represent aggregate results.

2.4. CSC infrastructure

The first INTERVENE test data sets (described above) have been uploaded to the SD Connect[17], the sensitive data storage service at CSC provided for INTERVENE and other projects (Figure 4). SD Connect will be one of the storage platforms used for storing and distributing the data within the INTERVENE project. As these synthetic datasets do not contain sensitive data they are not currently stored in encrypted format. These data can however be processed using the same tools and procedures that will be used for sensitive data later.

When sensitive data are uploaded to the SD Connect service, it will be encrypted using the GA4GH standard Crypt4gh tool[18] with a service specific encryption key. After this the data can be used only by the other CSC sensitive data services such as the SD Desktop that provides remote desktop view to a project specific secure virtual machine isolated from the internet. Within this environment, collaborative projects may work with the sensitive datasets uploaded by the project members from one or more organisations without a risk that any data would be copied outside the protected environment. In the near future - CSC will pilot with the INTERVENE project members the SD Apply service that will allow INTERVENE project members to apply for access to the biobank data sets (within the project and outside

[17] SD Connect: https://docs.csc.fi/data/sensitive-data/sd_connect/

[18] Senf et al, 2021, Crypt4GH: a file format standard enabling native access to encrypted data, Bioinformatics, https://doi.org/10.1093/bioinformatics/btab087

of it). Hence, the SD services facilitate the research use cases for the INTERVENE analyst and data custodian personas.

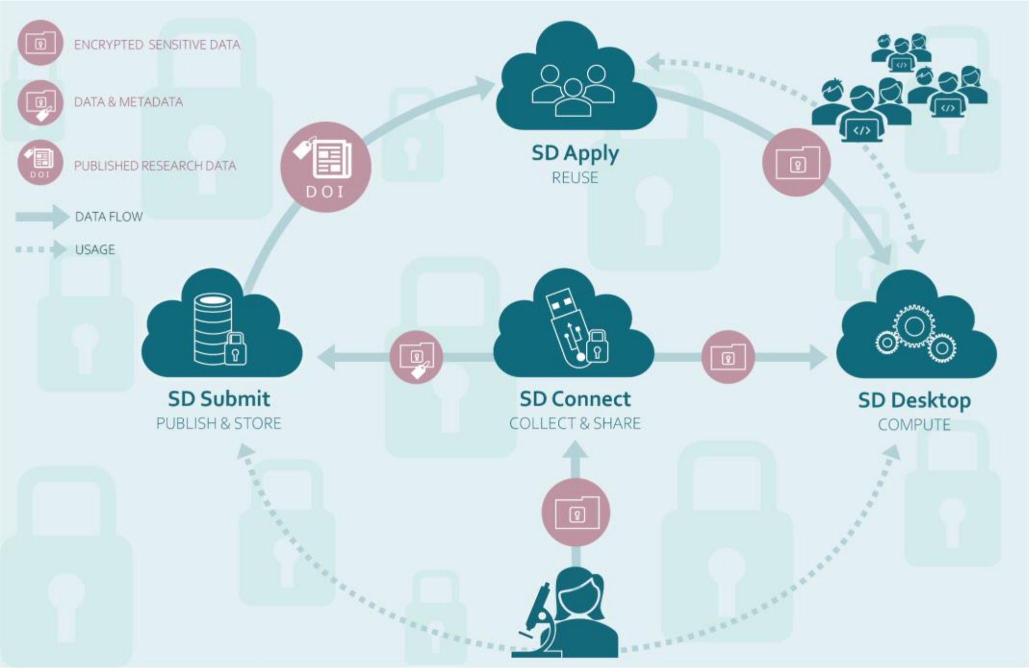

Figure 4. Components of the CSC sensitive data management environment. Public beta versions of SD Connect and SD Desktop are currently online. SD Submit (submission) and SD Apply (access application) components are planned to become available in 2022.

Demonstration of cross institutional access

The data upload process has been tested from CSC and EBI. At the moment, both the upload and data distribution processes are based on user accounts of CSC. INTERVENE project members can obtain CSC accounts that can be used to control the access of their own datasets and use datasets they have been granted access to. In cases where the data is considered as sensitive, the system level encryption restricts the availability of data to the CSC sensitive data environment only.

In conclusion, we have developed and demonstrated two modes of access for INTERVENE analysts (user persona 1, Figure 2) to access sensitive data from both EMBL-EBI and CSC Cloud services. The demonstration has been performed with synthetic, rather than real sensitive data, which is a necessary prerequisite for testing the integrity and interoperability of our systems at this phase of the project.

3. Future plans

3.1. CSC and EMBL-EBI infrastructure

Currently we are using separate native authentication and authorisation implementations for EBI and CSC. These differ in that the EBI system is predicated on the GA4GH compliant EGA data access management system, while access at CSC's SD Connect is exclusively determined by membership of a CSC 'project'.

In the future, data access management for data stored to SD Connect will be extended to Federated EGA compatible access management tools using GA4GH passports to link identities via the ELIXIR AAI and the Linked Identities Visa. Compatibility with Federated EGA is anticipated to be achieved during the next six months. This will enable more flexible ways for the access control of data sets stored in INTERVENE.

With these tools access control is no longer based on CSC accounts but other authentication and authorization methods can be used too. For example, authorization processes can be handled with a Resource Entitlement Management System[19] (REMS), an electronic tool that DACs can use to review the data access requests, and can function in the same way as the EGA AAI server. The EGA has an equivalent tool, and while the infrastructure isn't identical processes and functionality are comparable as they are based on global standards such as DUO, GA4GH AAI and Passports, and OAuth2.0[20].

For mature data collections and where appropriate, metadata, describing the data, will be created when the data is moved to a permanent storage platform like Federated EGA and SD Submit. This metadata can subsequently be made publicly available should it be required through metadata search tools of EGA and CSC FAIRdata services for external users maximising the FAIRness of the data. Use of FEGA at CSC aligns processes and operation with the EBI node, and facilitates the development of a unified IGS4EU platform log in.

The unified IGS4EU platform user authentication system will be developed using either ELIXIR AAI, a native access management system, or a combination of both, as appropriate. Once restrictions are in place, users will need to authenticate themselves on the common AAI system and then will be allowed access to IGS4EU Platform APIs at either location. Development of this feature will facilitate appropriate data access for all three user personas described above (Figure PERSONAS), as well as a common user experience.

3.2. Data

We plan to test our ability to run bioinformatic pipelines in our Cloud environments initially using tools and pipelines that have been developed (Deliverable 2.2) and are under development (WP3/4). Initially these will be run on the CINECA synthetic data, and then the first INTERVENE developed synthetic data which will shortly be available (Deliverable 3.4). These pipelines will be containerised for running on our environments, with a focus on portability and ability to run on other external biobank environments. We will perform benchmarking activities using the synthetic data and using UK Biobank metrics as a reference.

Next, we will move on to real sensitive data. The UK Biobank dataset is a highly accessed biobank dataset that can be moved to both CSC and EBI (see Table 1 from Deliverable 1.2). This would allow us to test our scaling and deploy and test the developed analysis pipelines. UK Biobank will shortly launch their own local research environment, the DNANexus Research Analysis Platform (RAP)[21], however the data we require (array-based genotyping data) will continue to be available for download. This will allow us to test the consistency and integrity of our pipelines by deploying them on the same data in both CSC/EBI and in the UK Biobank RAP. An application for access to UK Biobank data is underway, on which all relevant INTERVENE analysts are included. Facilitating access to data that may be moved to EBI and CSC Cloud

[19] Resource Entitlement Management System documentation: https://github.com/CSCfi/rems
[20] https://oauth.net/2/
[21]https://www.dnanexus.com/partnerships/ukbiobank

services also supports the use case of the data custodian (Figure 2) who uploads their own data to the IGS4EU platform.

Some data is restricted in that it cannot leave the originating biobank's local environment (described in Deliverable 1.4). Therefore, we will deliver portable analysis applications as containers which can be deployed in the federated nodes at EMBL-EBI or CSC or externally at biobank compute locations.

4. Appendix

4.1. Table 1 from Deliverable 1.2

	Biorepository	**Can data be moved outside the current repository?**	**Geographical restriction on where data may be moved to**	**Local analysis environment**
A. Data that may move	***UK Biobank***	Yes	No (data is already in EGA)	Under development
	Partners Biobank	Yes (if deidentified)	No	Yes
	Helsinki Biobank	Possibly to CSC, subject to legal considerations.	Yes	Yes
B. Data that cannot move	***FinnGen***	No	NA	Yes
	Network for Italian Genomes	No	NA	Yes
	Estonia Biobank	No	NA	Yes
	The HUNT Study	No	NA	Yes
	Genes & Health	No	NA	Yes
	Genomics England	No	NA	Yes

4.2. USE CASES

Persona 1 - Intervene analyst

Actor	As a	Intervene analyst
Narrative	I want to	access synthetic data
Goal	So that	I can apply a scoring file* to the synthetic data to calculate and download individual level PG scores.

* Scoring file may be self-supplied or accessed from the PGS Catalog

Actor	As a	Intervene analyst
Narrative	I want to	access real individual level genetic data**
Goal	So that	I can apply a scoring file to the real data to calculate and download individual level PG scores.

Actor	As a	Intervene analyst
Narrative	I want to	access real individual level genetic data** and apply multiple different scoring files
Goal	So that	I can compare performance/predictive ability of different scores in the same real genetic data

** real individual level genetic data hosted at EGA, CSC or hosted at a partner biobank

Persona 2 - Data Custodian

Actor	As a	External user (sensitive data owner)
Narrative	I want to	View relative performance metrics of existing scores

Goal	So that	I can select the most appropriate/best performing PG score to apply to my cohort

Actor	As a	External user (sensitive data owner)
Narrative	I want to	Upload my cohort's individual level genetic data to a secure environment
Goal	So that	I can calculate and download the individual level polygenic scores of the cohort

Actor	As a	External user (sensitive data owner)
Narrative	I want to	Apply PG scoring files to my cohort
Goal	So that	Assess the genetic risk profile of my cohort (download the individual level PG scores)

Persona 3 - Competition entrant

Actor	As a	External user (who has developed a scoring file by AI methods)
Narrative	I want to	Evaluate the predictive ability of my PG score on synthetic data (apply my scoring file to the synthetic data to calculate and download individual level PG scores.)
Goal	So that	I can compare it to performance of other PG scores (assess if it has performed well enough to merit application to real genetic data)

Actor	As a	External user (who has developed a scoring file by AI methods)
Narrative	I want to	Evaluate the predictive ability of my PG score on real genetic data (apply my scoring file to the real genetic data to calculate and download individual level PG scores.)

Goal	So that	I can compare it to the performance of other PG scores

Language disambiguation

- Scoring file = list of variants with effect weights
- Calculate a score = apply the weights in a scoring file to individual level genotype data
- Polygenic score = a single number which is the aggregate of the effects of many genetic variants of an individual