INTERVENE

Review of the consortium partner biobank and data collections, including access policies

Deliverable 1.2

Dissemination level: Public

Part of:

WP 1:

Compliant and standardised data access for federated analyses

Project summary Aims, process, completion, results, next actions	
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.	1.2
Work package	WP1
Deliverable lead	EMBL-EBI (Parkinson, McMahon, Freeberg, Keane)
Contributors	EMBL-EBI, CSC, BBMRI-ERIC
Dissemination level	Public
Expected delivery date	30/06/2021
Submission date	30/06/2021

Change history log

Version	Changes made	Date	Prepared by	PC approved
0.1	**Outline created**	**20-03-21**	**EMBL-EBI**	
1.0	**First draft**	**18-06-21**	**EMBL-EBI**	**Yes**

Contents

1. Introduction

This deliverable relates to Task 1.2, WP1 'Access to to the INTERVENE consortium data'. It addresses WP1 subobjectives 1, 3 and 4, and is a necessary precursor for delivery of the project's analysis platform and informs the function of the Data Coordination Centre (Deliverable 1.1).

This work builds on that of WP2, which has surveyed all biobanks/cohorts participating in INTERVENE (UK Biobank (UKBB), Genomics England, FinnGen, The Hunt Study (NTNU), Helsinki Biobank (HUS), Partners Biobank (MGH), The Genes & Health Study (QMUL), Estonian Biobank, and Network for Italian Genomes (UNISI)) to map the current data formats and standards of INTERVENE cohorts as a basis for a implementing a data harmonization strategy. These findings, including high level information on legal and ethical aspects of data access are summarised in Deliverable 2.1.

WP1 has extended this work and used comparable processes to acquire information on data access terms and conditions using a survey methodology of available datasets. Work in progress for WP4 is defining the use cases for INTERVENE data analysts (represented in WP3 and WP4) and the use cases require access to synthetic data and real data and therefore synthetic data availability is in scope. Synthetic datasets may also come with terms of use, particularly if these are obtained from external projects. This will be added in future when synthetic datasets are identified.

Subobjectives of WP1 addressed by this deliverable:

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
3. Provide access to the data with federated Authentication and Authorisation Infrastructure services (e.g. ELIXIR AAI) and ensure the interoperability between the computing environments.
4. To define and implement standards for data access across biobanks that comply with national and European regulation (in collaboration with **WP6**)

2. Data collection

To collect the information necessary for this deliverable we designed a survey for distribution to the biorepositories (biobanks) from which INTERVENE will access data. Questions on data access components were developed in consultation with data access experts from the European Genome-phenome Archive[1] team at EMBL-EBI and CSC team. Further questions on summary-level data were developed with WP3.

To avoid overwhelming partner repositories with multiple information requests from different working parties questions on ethical and legal aspects were included (required for Deliverable 6.1-Report of the ethical statements or equivalent documentation from the participating biobanks, D6.2/9.3 – Report on the informed consent procedures in regard to data processing, and D9.2 – Protection of personal data requirement 2 – lawful basis for data processing & safeguards for rights of data subjects).

The survey was created at EMBL-EBI using Survey Monkey and circulated by the INTERVENE project

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

management office. A GDPR privacy statement was included to inform survey respondents of how their personal data would be processed.

The survey contained the following sections and questions related to these areas[2]:

- **About survey**
  - An introduction to the purpose of the survey
- **General and contact information**
  - Respondant information
  - Separate contact details for specific domains e.g. scientific, ethical, legal etc.
- **Data access application process**
  - Documents required to apply for data access
  - Information on application review process (turn-around time, fees etc)
  - Data use conditions and restrictions
- **Data access and analysis**
  - How and where data may be analysed
- **Meta-data and summary level data**
  - Access to meta-data
  - Summary level meta-data (genetic ancestry, genotyping)
  - Links for publicly available summary level data
- **Ethics requirements**
  - Ethical statements
  - Legal basis for informed consent
  - Measures for the protection of personal data
  - Data subject rights
- **Data access and ethics documentation**
  - Links for publicly available documentation
  - Request for non-public documentation to be provided by email

[2] https://docs.google.com/spreadsheets/d/1DXrH6dUqDB63xEGwtqZadjPikUoF2EJet-qE8tiLg8c/edit?usp=sharing (de-identified version)

3. Results

The survey was circulated to all INTERVENE biobanks (FinnGen, Partners, NIG, Est BB, HUS, HUNT and Genes & Health, UK Biobank, and Genomics England). The majority of biobanks provided written answers to all or most of the questions. For UK Biobank or Genomics England, information was obtained from publicly available documentation. Answers to a selection of survey questions relevant to this deliverable are presented in Appendices 1-3. Some answers have been edited for clarity.

3.1 Data access application process

An overview of the process by which access to data is requested, including documents required, application review procedure and typical turnaround times were collected and are compiled in Appendix 1. This information will facilitate the planning of data access applications and will be clarified and built on over time.

To address Objective 4 (To define and implement standards for data access across biobanks that comply with national and European regulation) we surveyed the biorepositories regarding levels of regulation. We find that multiple biorepositories are subject to GDPR only, while others are subject to variants of GDPR or additional regional or national regulations. These have been specified in Appendix 1 (Question 1f). Further interrogation of the ELSI aspects of data access were gathered by this survey and are summarised in Deliverables D6.1, D6.2, and D9.2.

We aimed to assess if it is possible for one data access application to suffice for all INTERVENE partners. This may be possible for multiple biobanks (Appendix 1, Question 1e), but it is likely that each individual researcher will need to be named/registered on all applications.

3.2 Data access and analysis

Information on data mobility and analysis is presented in Appendix 2. The INTERVENE proposal posits two distinct data access models, one where individual level data can be deposited in central repositories (either central EGA or a federated instance of EGA at CSC) and a second where data cannot be moved and is analysed in local analysis environments. Results of this survey show that two biorepositories (Partners and UK Biobank) fall in the first category; these data can be moved with no geographical restriction as to where data may be moved. Further investigation is required to determine if data from a third, Helsinki Biobank, may be moved to CSC and will be investigated as part of WP1, Data Coordination activities (Table 1A).

Data from all other repositories must remain in their current locations (Table 1B). Of these, all have a local data analysis environment meaning that INTERVENE analyses can be executed on the data. However, responses from the biobanks indicate that the specifications and requirements on how these workflows may be set up will be diverse and will be subject to follow up analyses. For example, the UNISI resource has limitations on external workflow deployment. As anticipated, only aggregated data (results) may be exported from local data analysis environments, and in many cases is also subject to review by repository oversight mechanisms specific to each of the data sources.

Table 1. Summary of data mobility and local analysis environments.

	Biorepository	Can data be moved outside current repository?	Geographical restriction on where data may be moved	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

3.3 Fees

Some repositories charge for running analyses on their local environments (Appendix 2, Question 2f), while others do not. There is also variability in charging for data access applications (Appendix 1, Question 1d). The issue of charging for data access/computation will be addressed as a cross-consortium issue given the dependencies on external environments.

3.4 Commercial aspects

Multiple repositories specify non-commercial data use restrictions (e.g. NIG, EstBB, HUNT; Appendix 2, Question 2e), while others have different access application processes and fee structures for commercial versus non-commercial applicants (e.g. Genomics England; Appendix 1, Question 1a, 1d). Implications of these restrictions, in terms of future exploitation of project results (Task 8.6) will be further investigated in collaboration with WP8.

3.5 Meta-data (ancestry and genotyping)

To facilitate the work of WP3 in the development of realistic synthetic data (Deliverable 3.4) we have solicited high level information on population structure, genotyping and genome build (Appendix 3). The divergent use of the most recent (GRCh38) and previous (GRCh37) reference genome assemblies highlights the need for a harmonisation approach to genomic data, in which the analysis allows for different genome builds or a post-hoc genomic reference conversion is done. This will be factored into the design of the INTERVENE analysis pipelines and is a relatively common activity. The mandatory inclusion of the reference genome version in the INTERVENE scoring file standard (Deliverable 2.2) will facilitate interoperability.

3.6 Dissemination and contact

Full survey responses, excluding personal information to conform to GDPR, will be available on the INTERVENE intranet. This will include publicly available web links where further information may be found as well as documentation supplied directly to us. A further aim of the survey was to create a directory of specific contacts at each biobank, to whom specific follow-up enquiries may be made (e.g. regarding scientific, technical, ethical and legal issues). Follow-up queries to these contacts will be coordinated by the research coordinator to conform to GDPR requirements.

4. Conclusions

Our anticipated model of data access is supported by this review of data access policies. Two scenarios are apparent, 1) data cannot be moved from its existing environment but is accessible via an existing platform and 2) data can be moved to locations (EGA and/or CSC) where it may be used locally.

Dissemination of the survey responses to INTERVENE partners will facilitate and inform planning of future activities e.g. data access applications. It has highlighted multiple areas of divergence between biobanks (e.g. fee structures, commercial data use, local analysis environment requirements) which will be informative for the work of multiple work packages as ethical issues, design and delivery of the project's platform and analysis strategy..

5. Next Steps

Given the essentially binary nature of the data access conditions we have elucidated with the survey, we will design the data infrastructure and analysis infrastructure to allow for remote analyses and will provide portable analysis containers which we will also seek to deploy on our federated platform. We will benchmark these for performance and will work with the receiving platforms to ensure that these are aligned to each resource's analysis platform environment. This strategy will ensure that we are able to execute the project and also develop an environment to host local datasets.

We will also monitor the landscape as some resources, for example, UK BioBank, are also investing in a platform for data access and at some point we may also provide our analysis containers in this platform as a means of disseminating INTERVENE's products as well as enabling analyses. There are also cost implications for remote execution of analysis, as some of the resources may require payment for analyses and this will be considered on a case by case basis at a consortium level.

We will advertise the outcomes of this deliverable to the analysis working group (WP3/WP4) as these findings may increase their reliance on synthetic data and the datasets that are accessible from the project's platform. Further interrogation of the data content and structure of the biobanks will be required and is facilitated by the bank of specific contact details that we have produced via this deliverable. The technical limitations and specifications of some of the biobanks will also be explored in detail to ensure that the analysis strategy is aligned with the biobanks' data sharing models.

6. Appendices

Appendix 1. Data access application procedures (survey answers)

1a. What documents are required to apply for access to data?

	FinnGen (FIMM)	Network for Italian Genomes (UNISI)	Estonia Biobank (UTARTU)	Helsinki Biobank (HUS)	The HUNT Study (NTNU)	Genes & Health (QMUL)	Partners Biobank (MGH)	UK Biobank (UKB)	Genomics England (GeL)
Pre-application			Yes	Yes					
Research plan			Yes	Yes	Yes	Yes		Yes	Yes
Data handling plan			Yes	Yes	Yes				
Ethics committee approval			Yes	Yes	Yes				
Data access agreement			Yes			Yes	Yes		
Data Protection Impact Assessment (DPIA)		Yes		Yes	Yes				
Data handler/processor agreement			Yes						
Collaboration agreement	Yes	Yes	Yes		Yes		Yes		
Materials transfer agreement			Yes	Yes				Yes	
Other (please specify)	Data Protection Agreement Sandbox billing details (collaborating institution) FinnGen user identification agreement Findata confidential disclosure agreement Sandbox exam (specific user)	Appointment of Data Processor		Biobank sample and data access application		Genes & Health application form		Application form. Researchers must first register on the Access Management System (AMS), after which they will have access to the Data Showcase. Then they can select data items of interest and apply for access to that data.	Application form. (Answers apply to joining an existing GeCIP domain. GeCIP members are not eligible for data access until a detailed research plan been approved by ARC. There is a separate process for forming a GeCIP domain. Industry access is via a different route, the 'discovery forum'.)

1b. Who reviews data access applications?

	FinnGen (FIMM)	Network for Italian Genomes (UNISI)	Estonia Biobank (UTARTU)	Helsinki Biobank (HUS)	The HUNT Study (NTNU)	Genes & Health (QMUL)	Partners Biobank (MGH)	UK Biobank (UKB)	Genomics England (GeL)
Data access commitee (DAC)					Yes	Yes	Yes		Yes
Ethics commitee			Yes	Yes					
Institutional review board (IRB)			Yes						
Other (please specify)	FinnGen Administrators	Internal audit		Biobank internal decision				Access management team	GeL Staff One member of Scientific Advisory Committee Full Scientific Advisory Committee Access Review Committee (DAC)

1c. What is the typical turnaround for processing data access requests?

FinnGen (FIMM)	Network for Italian Genomes (UNISI)	Estonia Biobank (UTARTU)	Helsinki Biobank (HUS)	The HUNT Study (NTNU)	Genes & Health (QMUL)	Partners Biobank (MGH)	UK Biobank (UKB)	Genomics England (GeL)
1 - 3 months	NR	3 - 6 months	2 - 4 weeks	3 - 6 months	2 - 4 weeks	2 - 4 weeks	NR (Registration on Access Management System=10 days)	Estimate minimum 3 months. (Identity verification is performed by your Institution, not under the control of Genomics England. Access review committee meetings take place a minimum of three times per year, normally four.)

1d. Is there a fee associated with application?

FinnGen (FIMM)	Network for Italian Genomes (UNISI)	Estonia Biobank (UTARTU)	Helsinki Biobank (HUS)	The HUNT Study (NTNU)	Genes & Health (QMUL)	Partners Biobank (MGH)	UK Biobank (UKB)	Genomics England (GeL)
No	No	Other	Yes	Yes	No	No	Yes	No
		Fee for private companies Generally no fee for non-commercial use (depends on the application)	Academic fee is 300 euros (project cost) and ethics statement is free. Commercial fee is 3000 euros (project cost) and HUS ethical statement around 2000 euros. There might be additional smaller sample handling costs.				Access fee(s) for every application (due after approval). Additional fee for every collaborating institution.	

1e. Please describe if there is a procedure by which a single application could suffice for all or a subset of members of the INTERVENE project (grant agreement No 101016775)?

FinnGen (FIMM)	Network for Italian Genomes (UNISI)	Estonia Biobank (UTARTU)	Helsinki Biobank (HUS)	The HUNT Study (NTNU)	Genes & Health (QMUL)	Partners Biobank (MGH)	UK Biobank (UKB)	Genomics England (GeL)
NR	[Needs clarification]	Currently not an option. Even if it were each INTERVENE member who wants to access all datasets must also be on all necessary ethics approvals.	Project must have received a favourable statement from an ethics board. The ethics application can be applied for the complete project or for each subproject separately. If applied for the complete project, the subprojects should be stated clearly in the research plan and a complete list of all researchers involved in the project should be available. Helsinki Biobank can then deliver/grant access to the samples and data as described in the research plan, to all the researchers/organization listed in the ethics application. Theoretically one application to the biobank is enough. However, if the project is very large, including several subprojects and different sample batches, separate applications for the different subprojects are preferred. Nevertheless, the same ethics statement can be used for all applications.	NR	Yes, named on the same application to Genes & Health.	Yes, if there is a Mass General Brigham PI on the application	Yes, Applicant PI and Collaborator Leads are required to sign an MTA for each institute. Applicant PI and Collaborator Leads can add researchers to the application. All researchers must be registered on the AMS (Access management system).	NR

1f. Are you subject to regulation by GDPR or equivalent national/regional regulation?

FinnGen (FIMM)	Network for Italian Genomes (UNISI)	Estonia Biobank (UTARTU)	Helsinki Biobank (HUS)	The HUNT Study (NTNU)	Genes & Health (QMUL)	Partners Biobank (MGH)	UK Biobank (UKB)	Genomics England (GeL)
GDPR and additional regional regulations	GDPR	GDPR and additional regional regulations	GDPR and additional regional regulations	GDPR and additional regional regulations	GDPR	No	Equivalent national regulation	Yes
Finnish laws and Findata requirements of data secure data handling environments.		https://www.riigiteataja.ee/en/eli/523012019001/consolide	If samples or data are sent or processed in an institute outside of the EU, the receiving institute must comply to the EU Standard Contractual Clauses (Controller to Controller) for data protection (https://ec.europa.eu/info/law/law-topic/data-protection/international-dimension-data-protection/standard-contractual-clauses-scc_en)	The Norwegian health research act requires all health related projects to be pre-approved by the regional ethics committee			UK GDPR (adopted by HM Government of the GDPR into domestic law and to take into account the implications of the UK leaving the EU. The domestic law is set out in both the UK GDPR and the UK Data Protection Act 2018 (DPA).)	[needs clarification, possibly public information not updated to reflect Brexit changes, see UKB response]

Appendix 2. Data access and analysis (survey answers)

2a. Can data be moved outside the current repository or research environment?

FinnGen (FIMM)	Network for Italian Genomes (UNISI)	Estonia Biobank (UTARTU)	Helsinki Biobank (HUS)	The HUNT Study (NTNU)	Genes & Health (QMUL)	Partners Biobank (MGH)	UK Biobank (UKB)	Genomics England (GeL)
No	No	No	No	No	No	Yes, if deidentified	Yes	No

2b. Can the data be moved to a centralised repository in the UK (EGA) or Finland (CSC)?

FinnGen (FIMM)	Network for Italian Genomes (UNISI)	Estonia Biobank (UTARTU)	Helsinki Biobank (HUS)	The HUNT Study (NTNU)	Genes & Health (QMUL)	Partners Biobank (MGH)	UK Biobank (UKB)	Genomics England (GeL)
No	No	No	Subject to legal considerations, maybe CSC	No	No	Yes	Yes (already in EGA)	No

2c. If data cannot leave the repository - is there a local data analysis environment (e.g. sandbox, local server, cloud environment)?

If yes, can it run external workflows (specify if there are restrictions on what external workflows can be run or recommendations for workflow construction e.g. Docker/Singularity)

	FinnGen (FIMM)	Network for Italian Genomes (UNISI)	Estonia Biobank (UTARTU)	Helsinki Biobank (HUS)	The HUNT Study (NTNU)	Genes & Health (QMUL)	Partners Biobank (MGH)	UK Biobank (UKB)	Genomics England (GeL)
Local data analysis environment	Yes	Yes	Yes	Yes	Yes	Yes	Yes	No	Yes
					HUNT Cloud	UK Secure eResearch Platform (UKSERP) - soon to move to a Google Cloud-based platform.		Under development (DNAnexus)	
Can it run external workflows (please specify)	Yes, user cannot install programs themselves (so eg. their own dockers), but specific programs can be asked to be installed to the environment.	No	Yes, as long as our IT department can install it, should be fine	Needs to be further specified	Yes, you can run what you want	Yes, we have a local PC that can run external workflows on genetic data only.	Unknown	NR	Yes

2d. Can products of analyses (results) be exported from the environment?

FinnGen (FIMM)	Network for Italian Genomes (UNISI)	Estonia Biobank (UTARTU)	Helsinki Biobank (HUS)	The HUNT Study (NTNU)	Genes & Health (QMUL)	Partners Biobank (MGH)	UK Biobank (UKB)	Genomics England (GeL)
Yes, only non-identifiable aggregate results can be downloaded. Raw data download is not possible.	Yes, only aggregated data	Yes, data protection department has to approve and will be involved in exporting results	Yes, subject to national legistlation and according to the valid research permit	Yes	Yes, No individual level data exported, small number suppression used where appropriate. All exports reviewed by the PI.	Yes	NA	Yes, subject to review (the Airlock process)

2f. Are there costs involved in running or exporting results (e.g. data egress charges)?

	FinnGen (FIMM)	Network for Italian Genomes (UNISI)	Estonia Biobank (UTARTU)	Helsinki Biobank (HUS)	The HUNT Study (NTNU)	Genes & Health (QMUL)	Partners Biobank (MGH)	UK Biobank (UKB)	Genomics England (GeL)
Analysis charge	Yes, computing costs + monthly Sandbox storage cost.	No	Yes, https://hpc.ut.ee/en/about/prices/, however, small computations which don´t require much resources are usually not charged	Yes, the standard fee of the local analytics environment	Yes, storage/year and compute costs	Yes, HPC will have associated costs if/when a future Google Cloud based system is used for INTERVENE. At the moment the limited compute in UKSERP is free.	No	NA	No
Export charge	No	No	No, https://hpc.ut.ee/en/about/prices/	No	No	No	No	NR	No

2e. Do data use restrictions apply?

	FinnGen (FIMM)	Network for Italian Genomes (UNISI)	Estonia Biobank (UTARTU)	Helsinki Biobank (HUS)	The HUNT Study (NTNU)	Genes & Health (QMUL)	Partners Biobank (MGH)	UK Biobank (UKB)	Genomics England (GeL)
Non-commercial use only		Yes	Yes		Yes			NR	Yes
Other (please specify)	FinnGen Scientific Plan restricts the use of the data.			Access granted for specific purpose as defined in the research plan	Project specific, can include multiple diseases but needs to be described in the research protocol	Limited to work set out in the application - that may be both disease specific or non commercial, but doesn't need to be. Limited to approved users, results should be returned to the resource, results should be published. Subsets of health data *may* have geographical restrictions to the EEA.	No restrictions	Results should be returned to the resource.	Non-commercial if accessed via GeCIP domain. Industry access is via the 'discovery forum'.

Appendix 3. Ancestry and genotyping (survey answers)

3a. Please describe estimated genetic ancestry breakdown of the cohort using super populations (AFR, African; AMR, Ad Mixed American; EAS, East Asian; EUR, European; SAS, South Asian)

FinnGen (FIMM)	Network for Italian Genomes (UNISI)	Estonia Biobank (UTARTU)	Helsinki Biobank (HUS)	The HUNT Study (NTNU)	Genes & Health (QMUL)	Partners Biobank (MGH)	UK Biobank (UKB)	Genomics England (GeL)
EUR (Finnish)	NA	Mainly EUR	EUR	Mainly EUR	NR	EUR 60%, AFR 15%, AMR 10%, EAS and SAS 15%	EUR 95%, AFR 1.5%, AMR 0.2%, CSA 2%, EAS 0.6%, MID 0.4%	EUR 85%, SAS 7%, 1% AFR

3b. How was the cohort genotyped?

	FinnGen (FIMM)	Network for Italian Genomes (UNISI)	Estonia Biobank (UTARTU)	Helsinki Biobank (HUS)	The HUNT Study (NTNU)	Genes & Health (QMUL)	Partners Biobank (MGH)	UK Biobank (UKB)	Genomics England (GeL)
Genotyping technology (array name or sequencing coverage):	Multiple genotyping arrays used. Majority genotyped with Affymentrix FinnGen Custom chip	WES	Illumina Global Screening array	FinnGen ThermoFisher Axiom custom array v1/v2	Illumina Human core exome	NR	Illumina	UK BiLEVE Axiom array, UK Biobank Axiom array (https://biobank.ctsu.ox.ac.uk/crystal/label.cgi?id=263)	WGS
Imputation reference panel:	SiSU (v3) reference panel	NA	Estonian Whole Genome sequencing panel is used as reference for imputation	All GWAS data is imputed against a Finnish population specific whole genome sequence (WGS) backbone	HRC/TopMed	NR	U Michigan TopMed	Haplotype Reference Consortium and UK10K haplotype resources (https://biobank.ctsu.ox.ac.uk/crystal/label.cgi?id=100319)	NA
Genome build (GRCh37/hg19 or GRCh38/hg38):	GRCh38/hg38	GRCh37/hg19	GRCh37/hg19	GRCh38/hg38	GRCh37/hg19 and GRCh38/gh38	NR	GRCh38/hg38	GRCh37/hg19	GRCh37/hg19 and GRCh38/gh38