INTERVENE

Data Management Plan

Deliverable 7.2

Dissemination level: confidential

Part of:

WP 7:
Project management and coordination

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.	7.2
Work package	WP7
Deliverable lead	CSC, EMBL
Contributors	UH-FIMM
Dissemination level	Confidential
Expected delivery date	30/06/2021
Submission date	30/06/2021

Change history log

Version	Changes made	Date	Prepared by	PC approved
0.1	**First draft**	**24.05.2021**	**CSC, EMBL**	
0.2	**Major revisions, addition of data survey results, task PI approval**	**14.06.2021**	**CSC, EMBL, FIMM**	
1.0	**Technical revision**	**21.06.2021**	**FIMM**	**Yes**

Contents

1. EXECUTIVE SUMMARY 4
2. INTRODUCTION AND PURPOSE 5
2.1 Introduction to the DMP 5
2.2 Revisions and updates to the DMP 5
3. DATA SUMMARY AND USE OF EXISTING DATA 6
3.1 How will new data be collected or produced and/or how will existing data be reused? 6
3.3 Types of data to be utilised 8
3.4 Data exchange between partners 10
3.4.1 General principles 10
3.4.2 Processing and sharing personal data within INTERVENE 10
3.5 Mechanisms of data access 11
4. DATA SECURITY 12
5. LEGAL AND ETHICAL REQUIREMENTS 12
6. FAIR DATA (RE-USE AND LONG-TERM PRESERVATION BEYOND THE PROJECT) 13
6.1 Making data findable, including provisions for metadata 13
Version numbers 14
Metadata use 14
6.2 Data open access 14
Deposition of data 15
Machine readable licenses 16
Identification of persons accessing data 16
6.3 Data interoperability 16
Ontology standards and interoperability 17
6.4 Increase data re-use 17
Data licensing 17
Data availability for re-use 17
7. DATA MANAGEMENT RESPONSIBILITIES AND RESOURCES 18
Cost for FAIR compliance 18
Data management responsibilities 18
8. OTHER ISSUES 19
Use of other data management procedures 19
APPENDIX 1: DATA AGREEMENT 20
APPENDIX 2. DATA SOURCES 26
APPENDIX 3. INTERVENE WORK PACKAGE SPECIFIC DATA MANAGEMENT INFORMATION 28

1. Executive summary

The overarching purpose of INTERVENE is to build one of the largest pools of health data to date and integrate longitudinal and disease-relevant -omics data into genetic risk scores with improved potential for prediction, prognosis, and personalised treatments of complex and rare diseases. These data will provide a test bed to demonstrate the potential and benefits of powerful artificial intelligence (AI) technologies on the next generation of integrative genetic scores (IGS).

All data access to cohort datasets is determined by their existing data access committees (DACs) and the respective data processes for each dataset (data access requirements and procedures were surveyed and are outlined in Deliverable 1.2). In summary, except for INTERVENE partner datasets (datasets owned by partners or created in INTERVENE), the participating biobanks all restrict data access to their own platforms and this simplifies the DMP in that the activities will be subject to processes determined by the data sources. The DMP therefore describes the consequences of this for data management as well as addresses the management of data generated or owned by INTERVENE.

All project partners have relevant institutional/national and/or international experience in acquiring, storing, analysing, and sharing high complexity datasets and in the implementation of the FAIR principles for these resources.

This DMP describes the principles by which INTERVENE will manage data, identifies external dependencies on data consumed by INTERVENE and provides an ongoing and a 'by work package' analysis of data management needs for the project's data generation activities.

The DMP will be updated periodically and a DMP oversight group has been convened comprising representatives of FIMM, CSC and EMBL-EBI who have both the institutional expertise in data management and will work towards the deliverables which will implement the DMP throughout the project.

2. Introduction and purpose

2.1 Introduction to the DMP

This document constitutes deliverable D7.2 and comprises the first version of the Data Management Plan (DMP) for INTERVENE. The purpose of the document is to identify and describe the data management processes for data sets collected, processed, and generated throughout the INTERVENE lifecycle and stored beyond the conclusion of the project. In addition, the DMP describes central principles relating to open access of research and other results produced within the project. INTERVENE is a data-intense undertaking, leveraging personal genetic and health data from more than 1.7 million individuals across Europe and the USA. To that end, INTERVENE considers proper data management and control processes to be of utmost importance. The DMP will therefore be subject to regular scrutiny and updates to ensure that its contents are meeting the demands of the consortium, its composition, current ethical standards, as well as national or international legislation.

The DMP has been developed based on the core requirements for DMPs as described by the Science Europe Practical Guide to the International Alignment of Research Data Management and the Horizon 2020 online manual[1] and guidelines for data management. Possible updates to these guiding documents will be considered when revising the INTERVENE DMP.

2.2 Revisions and updates to the DMP

The DMP is defined as a dynamic, living document, intended to evolve with the maturation of the project and subjected to regular updates throughout the project's lifespan. The timeline for DMP updates will follow guidance and requirements provided by the EC. Specifically, the DMP will be subjected to periodic reviews in connection with periodic reports to the EC (months 12, 30, and 48). A final update of the DMP will be produced as a separate deliverable (D7.6) in connection with the final report to the EC at month 60.

Figure 1. Planned updates to the DMP

[1] https://ec.europa.eu/research/participants/docs/h2020-funding-guide/cross-cutting-issues/open-access-data-management/data-management_en.htm

In addition to the scheduled periodic reviews, the DMP will be updated to reflect possible substantial changes during the project, including, but not limited to

- Novel data and data types acquired or generated by the consortium
- Decisions to pursue and file patents
- Changes to the consortium composition

To ensure that the DMP continuously reflects the state and needs of the consortium, a DMP oversight group composed of members from CSC (Fuchs, Ahokas), EMBL-EBI (Parkinson, McMahon), and FIMM (Anckar) has been established and all have contributed to this deliverable. The oversight group will be chaired by the INTERVENE coordinating office (Anckar) and tasked with identifying changing needs to be proposed to the INTERVENE Managing Board for discussion. The DMP oversight group will meet periodically at 12-week intervals in addition to possible *ad hoc* meetings to oversee DMP revisions required by changes to the project, including those stated above. In addition, a presentation of the DMP will be included in the Managing Board meeting agenda quarterly. The DMP oversight group will be extended as needed as the project progresses. This includes ethics expertise provided by members of WP6 as well as expertise from INTERVENE's WP leads or members for questions and processes related to specific data types.

3. Data summary and use of existing data

3.1 How will new data be collected or produced and/or how will existing data be reused?

INTERVENE is primarily concerned with the secondary use of data and metadata collected to biobanks in the US and in Europe. INTERVENE draws upon the vast resources of genomic, other -omic, and health care data contained in biobanks or study repositories. The data contained in these repositories and used by INTERVENE has been collected under conditions fulfilling necessary local ethical and legal requirements before or during the project period. Utilization of this data by INTERVENE members builds on the sending of computational tasks to the biobanks and -repositories for local processing. Personally identifiable information will be pseudonymized and transferred to the research partners on an aggregate level with no individual-level data identifiable. The personal data contained in the biobanks will at all times remain under the control of the biobanks and for the majority of these will not leave their own infrastructure, meaning INTERVENE will comply with terms of mode of access mandated by these biobanks.

The individual data access policies are outlined in Deliverable 1.2 and this work has informed the construction of the DMP.

INTERVENE will generate new methods and algorithms for AI-based data analyses and the development of genetic risk scores. In addition, in WP6 and WP8, data from interviews, and surveys will be used to gauge the public's perception of genetic prediction tools in healthcare and to investigate the needs and requirements of INTERVENE stakeholder groups.

As part of WP5, to assess the real-world clinical performance and economic benefits of the genetic risk scores developed within INTERVENE, two clinical pilot studies will be performed. The clinical pilot projects will collect new individual level clinical, questionnaire and (gen)omics data, and separate ethics committee/Institutional Review Board approvals will be obtained for these collections. These personal data will not be shared outside the partners defined in the approvals. The content of the collected data will be defined and reported in D5.1 and D5.4.

DCC (Deliverable 1.1) provides services and resources to store and manage data that can be shared within the project. It provides the practical implementation of the DMP for data stored or made accessible from the DCC and this is described briefly below.

The DCC data made accessible to partners and WP4 competitors will be stored and distributed using the CSC SD-connect or Allas Object Storage services. These data will be stored in the original formats used by the original data providers. Sensitive data will be encrypted (with crypt4gh[2]) a GA4GH standard. The access control is based on CSC accounts and ELIXIR AAI but additional control layers can be added (e.g. REMS) if needed.

An itemised inventory of the data types collected or produced by the different work packages has been populated by work package leads (Appendix 3 & data sheet summary), noting that this is a living document available as a google document[3] and will be updated (as outlined in Section 1.2) and as more datasets appear these will be added. An example component of this document is shown in Figure 2 and is available in more detail in Appendix 3.

[2] https://www.ga4gh.org/news/crypt4gh-a-secure-method-for-sharing-human-genetic-data/

[3] https://www.google.com/url?q=https://docs.google.com/spreadsheets/d/1ZVpwOOC-stPUWTCMQvX50EaerhuwWHXSKmlGDX3M1Q/edit%23gid%3D319325758&sa=D&source=editors&ust=1623425478884000&usg=AOvVaw3cu1YGRhpR7RQs4kQU1WoX

WP	Data type, e.g. survey called X	Is the data collected, produced, or reused?	What data production/collection methods will be used?	If the data is reused, specify from where.	File formats of the data e.g. .txt	Volume (estimate of size ~)	How/Where will the data be stored during the project?
WP1	Sofware container	Produced/ Generated	Software repository	Data will be reused first during the software developmet process and then when the actual use cases are executed.	Sigularity, Docker	Less than 5 TiB	CSC Object Storage services or virtual machine volumes in cPouta.
WP1	Public reference data sets	Reused	Copying from other repositories	Data will be used as the training material in software development	Depends on the original reposotory	Less than 50 TiB	CSC Object Storage services or virtual machine volumes in cPouta.
WP1	Sensitive biobank data	Reused	Download processes on biobanks	Use cases can utilize this data with permission of the biobank	Crypt4gh encrypoted format. Inisde the ecrypition layer original biobank format will be used.	100 TiB	SD-connect / CSC
WP1	Survey of biobank data access and ethics policies	Collected	Survey	NA	.xlsx	negligible	EMBL-EBI
WP1							

Figure 2. An excerpt of the INTERVENE work package specific data requirements

3.3 Types of data to be utilised

The primary data employed by INTERVENE are previously collected molecular data (genomics and other -omics) and phenotypic data from across a range of biobanks and studies.

INTERVENE will also include data collected through:

- Surveys
- Questionnaires and interviews
- Registry data (observations)
- Clinical measurements
- Medical records in the form of electronic health records.
- Electronic health records
- Administrative records
- Intermediate statistics for meta-analysis

Tables 2 and 5.2 (Annex 1B) of the proposal also provide an overview of the data types.

With the cohort data variables stored in different formats and often based on different sources, data transference into a common format allowing large-scale automatic analysis and harmonizing definitions across data sources is necessitated. To map the current data formats and standards of INTERVENE cohorts as a basis for a data harmonization strategy, surveys to all biobanks/cohorts participating in the project have been issued as part of deliverable D2.1. An initial cataloging of data types, formats, and volumes contained within the INTERVENE biobanks and medical repositories is provided in Table 1.

To maintain a comprehensive catalog of the data types, formats, and volumes, regular data surveys will be employed within the consortium and used to update Appendix 3.

Table 1. Overall description of INTERVENE data repositories. Color scheme – green (yes), orange (no).

	EstBB	FinnGen	Genomics England	UKBB	NTNU	QMUL	UNISI	HUS
Data format	SQL	Excel	SQL	?	SQL, Text based	Excel	Excel	Excel, can provide in any format
Data language	Estonian, English	English, Finnish	English	English	English, Norwegian	English	English	Finnish
Data accessibility	Internal analysis using provided scripts; Federated access	Internal analysis using provided scripts	Analysis done internally using provided scripts, can be accessed via remote sandbox	Can be down-loaded	Only in cloud-based servers, no download or federated access	Internal analysis using provided scripts. Access can be granted to servers in UK for individual level data.	Internal analysis using provided scripts	Analysis done internally using provided scripts, Federated access
Structured data								
Images								
Free text, semi structured data								
OMOP model	Under consideration				Under consideration		In a year	In process
Data coded								
Data updated with registries, health care services					Yes, upon request			
Data downloadable	Rare cases	individual-level data cannot be downloaded, only aggregated data				Only export of summary stats is allowed		,
Data types	**EstBB**	**FinnGen**	**Genomics England**	**UKBB**	**NTNU**	**QMUL**	**UNISI**	**HUS**
Genetic data	~200k	500k	100k	500k	70k (+20k by 2022)	34k	3k	22k
Genetic data format	vcf	vcf	Wgs, format ?	bgen	Vcf, sav	Bgen, pfile	WES, format ?	vcf
Metabolome data	NMR for 10800, clinical biochemistry for 2650 ind.			N~121k	N ~8k, +9k in 6 months	N~50		
Microbiome data	Shotgun metagenomic sequencing carried out on stool samples, N= 2509.				5500, targeted assay with 50 bacteria			
Proteomics data	4 arrays, N=500				3500 samples, mostly CVD-related			
Chromatin data								
Methylation data	Illumina methylation array, n=450				Very limited			
Transcriptome data	Purified CD4 and CD8 t cells available for 293 individuals. Whole Blood cell expression maseq 488, illumina ht12v4 917 ind.				Very limited (lung cancer, colon cancer, psoriasis)		RNA-Seq n<10	
Phenotype data	**EstBB**	**FinnGen**	**Genomics England**	**UKBB**	**NTNU**	**QMUL**	**UNISI**	**HUS**
Objective information								
Education and work-related questions								
Smoking habits								
Alcohol consumtion								
Nutrition data								
Sleeping habits								
Physical activity								
Reproductive health								
Family history of diseases								
Medical history of diseases								
Pharmacological data				Limited				
ICD-10 codes				Yes, but ICD-9 for older cases				
Exact dates of ICD-10 codes								
Prescribed drugs				Limited		In process for all		Limited
Adverse reactions to prescribed drugs							Limited	Limited

.

3.4 Data exchange between partners

3.4.1 General principles

The members of INTERVENE have concluded an agreement under which access rights to data that is required for the performance of the own work of a party within INTERVENE is granted on a royalty-free basis. All requests for access rights must be made in writing to the data owner, for example through email, allowing data requests between the parties to be documented. Data access requests must demonstrate a need for the data, and comply with the original consent, for the performance of work conducted under INTERVENE. The granting of access rights may be made conditional on the acceptance of specific conditions aimed at ensuring that these rights will be used only for the intended purpose and that appropriate confidentiality obligations are in place. INTERVENE will adhere to these conditions.

3.4.2 Processing and sharing personal data within INTERVENE

INTERVENE aims to facilitate research projects that use data from several biobanks. Some of the biobanks allow researchers to download the data outside the biobank for analysis. In these cases the researcher must have a secure analysis environment that is trusted by the biobank. One of the tasks of INTERVENE (WP1 and WP4) is to provide this kind of analysis environment. Here the emerging platform provided by CSC and EMBL will play a key role in implementing the DMP.

However, the primary option for utilization of biobank data is that a researcher sends an analysis request to the biobank, upon which the biobank will perform the analysis in their local computing environment. In these cases, INTERVENE will harmonize how these analysis requests are created, submitted and processed so that the same request can be submitted to several biobanks.

At a technical level, software containerization and workflows are likely solutions to enable the harmonization required by cross-use of several biobanks. DCC will provide a repository that can be used to host and distribute software containers and other software components created during the project.

To clarify the roles and responsibilities of the parties with regard to the processing of personal data in the INTERVENE project, a personal data protection agreement will be concluded between all parties that process personal data in the project. The purpose of the agreement is to agree on the principles of disclosing and processing personal data in the project and avoid multiple individual agreements between different parties. The agreement includes an Appendix, in which transfers of personal data between the parties are documented. The agreement and appendix thereto are attached to this DMP as Appendices.

A summary of the agreement contents is given below:

- Each party must comply with applicable data protection laws, including laws and ethical guidelines regarding informed consent.
- Parties must give each other any information necessary for data protection compliance.
- Parties providing personal data to others must ensure they have the right to disclose the data.
- A recipient of personal data (from another party) may only process the data for the purposes of the project and must delete the data when no longer necessary for the project, unless the parties agree otherwise
- Transfer of data to recipients outside the EEA or to international organizations require safeguards specified in the GDPR.
- Personal data must be held confidential.
- A party that collects personal data is responsible for providing the required information about the project to/obtaining consent from data subjects, but the parties can also draft the documents in collaboration
- Requests from data subjects (such as information request or withdrawal of consent) must be communicated to all affected parties, and they are all responsible for taking the necessary steps
- Mechanisms for notification of data breaches.
- Contact details of the parties' data protection staff.

Additional data processing and data transfer agreements will be used when necessary. For example, an agreement between CSC/EMBL-EBI and data owners (biobank or other organizations) will be needed to store or process data within resources hosted at CSC and therefore additional work will be necessary where data is moved to CSC.

3.5 Mechanisms of data access

The overall coordination of data access to the biobank samples by INTERVENE members is performed through the Data Coordination Center (DCC), which will make the consortium data accessible by linking to external data or indexing the project's data holdings where these are held internally. The DCC will support INTERVENE's analysis and data access application scenarios from WP3 and WP4. These include provision of an analysis platform for deployment of methods deriving from WP3 and provision of access to synthetic and real data for competitions run by WP4. INTERVENE members will be able to access these resources by sending a request to the

DCC contact address, noting in some cases they will be referred to an external biobank as data cannot move locations.

Access to the data will require approval from the data custodians – biobanks and individual cohort owners that have the responsibility to host the physical sample collections and their digitized data products. To ensure secure and scalable data access processes INTERVENE deploys the most recent GA4GH standards (e.g. Researcher passport, Data use ontology) on the ELIXIR authentication and authorization services. Furthermore, the consortium data management processes will comply with security policies and we will also work closely on ELSI with WP6 to comply with the European GDPR and the national legislation.

Information on the data access application procedures and requirements for INTERVENE are currently being gathered from the partner repositories by means of a survey circulated to all partners. Information from this survey will be documented and will be outlined in Deliverable 1.2 (Review of the consortium partner biobank and data collections, including access policies). We anticipate multiple different data access scenarios, based on the rules and requirements of the different biorepositories, INTERVENE will respect the individual policies of each biorepository. Data access will be provided via the federated platform, but also via cohort specific architectures e.g. cloud platforms built on existing open science strategies and existing repositories.

4. Data security

The collaboration of the INTERVENE project with the B1MG project (Beyond 1M Genomes), funded under the EC's call for proposals 'SC1-HCC- 06-2020', through CSC as a shared partner, further allows us to define technical specifications and standards for the secure access and exchange of cross-border genomics, -omics and longitudinal health data (such as GA4GH Crypt4GH, htsget, Passport, DUO or phenopackets). Both CSC and EMBL-EBI provide secure Cloud services used for managing genomic and health data. The CSC ePouta Cloud service has ISO 27001 standard. Hence, the technological framework developed in INTERVENE is ensured to comply with national regulations and GDPR by evaluating the GDPR- compliance of the proposed data access and data-sharing procedures.

5. Legal and ethical requirements

Consent and long term preservation are addressed in task 6.1.2 Identifying gaps (M6-M12) and deliverable D6.3 (Report on Mapping ELSI and Identified Gaps). This task builds on T6.1A and assesses the gaps as regards to GDPR compliance especially in relation to the proposed data access and data-sharing procedures. This will be achieved by a partial data protection impact

assessment (DPIA) using the CNIL's PIA as a guiding document (esp. assessing risk assessment, identifying data controller, data provider, data workflow, data management plan). Ultimately, depending on the gaps, it will be assessed if a full DPIA is needed. The task will be achieved together with WP1, where the INTERVENE data infrastructure is designed and constructed.

Information on the informed consent and ethics approvals for INTERVENE are currently being gathered from the partner repositories by means of a survey circulted to all partners.

Information from this survey will be documented and contribute to Deliverables 6.1 (Report of the ethical statements or equivalent documentation from the participating biobanks), 6.2 and 9.3 (Report on the informed consent procedures in regard to data processing) and 9.2 (Protection of personal data requirement 2 – lawful basis for data processing & safeguards for rights of data subjects).

6. FAIR data (re-use and long-term preservation beyond the project)

6.1 Making data findable, including provisions for metadata

To enable findability of the data INTERVENE will first identify data sets and catalog these in the DCC (organised by data owner). Where dataset are consumed we will use existing identifier sets from the source organisations. Where the project generates data we will use existing identifier and meta data formats to maximise the FAIRNess of the data, for example, by using federated EGA (FEGA) for storage of genetic data we obtain access to stable identifiers and a metadata standard that corresponds to GA4GH standards. Where data can be made available from a public repository (PGS Catalog) we will use these and will evaluate FAIRness, though we expect that the repositories will already have made such evaluations.

To ensure that INTERVENE's research results are findable and compliant with EU funding acknowledgement provisions, published work will include DOIs and include bibliographic metadata, including the terms:

- "European Union", "Horizon 2020", and "Research and Innovation programme"
- "INTERVENE" & grant number "101016775"
- publication date, the length of the embargo period (if applicable) and a persistent identifier.

INTERVENE will provide search keywords that optimize possibilities for re-use and enable the querying of the data, and associated metadata across datasets made available by the partners.

Specific naming conventions will depend somewhat on the type of data to be named. The title should provide useful information about the file contents, together with date, and/or version number. For example, general files conform to the following: WPx_title_ddmmyyyy. ext.

Version numbers

Where existing repositories are used for data e.g. Genetic risk scores which will be deposited in the PGS Catalog then persistent identifiers can be used and for this example (PGS000XX), work is currently underway to facilitate transparent versioning. As not all data is generated yet future work will be to identify relevant identifier standards and use these. This will be the responsibility of each work package which generates data to define or use relevant identifier standards and to identify whether the data is stored internally to INTERVENE or whether a public repository is appropriate. Where external repositories are used and INTERVENE data is stored there, then the DCC will track data location and identifiers.

Metadata use

To maximize interoperability the following persistent and unique identifiers and controlled vocabularies will be considered. Some of these are implemented now, e.g. PGS Catalog implements the PGS Standards. The controlled vocabularies used to describe datasets will be documented and will be resolvable using persistent identifiers to the Ontology Lookup Service[4], an ELIXIR recommended interoperability resource or an alternative ontology access service. The documentation will be findable and accessible by anyone who uses the datasets.

6.2 Data open access

INTERVENE is committed to open access publishing and data sharing, supporting efforts by the European Commission to increase accessibility of scientific information. It is the responsibility of each partner to ensure that all peer-reviewed publications resulting from the project are accessible free-of-charge online through either the Green or Gold models of Open Access publishing. Raw (meta)data will be made available when privacy or commercial confidentiality do not preclude their publishing (see above).

[4] https://www.ebi.ac.uk › ols

Digital research data generated in INTERVENE is subject to provisions of the Grant Agreement (Section 29.3), stipulating that beneficiaries must

a) Deposit in a research data repository and take measures to make it possible for third parties to access, mine, exploit, reproduce and disseminate — free of charge for any user — the following:[5]
    i. the data, including associated metadata, needed to validate the results presented in scientific publications, as soon as possible;
    ii. other data, including associated metadata, as specified and within the deadlines laid down in the DMP;

b) Provide information — via the repository — about tools and instruments at the disposal of the beneficiaries and necessary for validating the results (and — where possible — provide the tools and instruments themselves).

These provisions do not change the obligation to protect results, confidentiality obligations, the security obligations, or the obligations to protect personal data as defined in the grant agreement. Decisions by data owners to protect the results on the basis of confidentiality and security obligations will be taken on a case-by-case basis and is governed by provisions in the INTERVENE consortium agreement.

Where public repositories can be used they will be and we will ensure these are evaluated for adherence to the DMP and the principles expressed herein. Where code can be made available we will do so and will licence it appropriately. Reasons not to make code available will include to protect IP (according to the CA) or for data security reasons. Where code is made available it will be licenced according to the FAIR Principles.

Deposition of data

- Machine-readable electronic copies of the published work and research data needed to validate the results will be deposited in repositories for scientific publications, such as ArXiv, bioRxiv, or Zenodo. Upon acceptance by the European Commission, all deliverables

[5] As an exception, parties do not have to ensure open access to specific parts of their research data under Point (a), if the achievement of the action's main objective would be jeopardised by making those specific parts of the research data openly accessible. In this case, the data management plan must contain the reasons for not giving access.

marked public will be made available on the INTERVENE website, in addition to publication in CORDIS.

- Code developed as part of WP3 and WP4 will be deposited in a dedicated INTERVENE github account.
- Genetic risk scores (scoring files) will be deposited in the PGS Catalog.
- Data of non-sensitive nature and immediately relevant for the entire consortium (e.g., data catalogues, deliverables) are made available on the INTERVENE intranet.

Much of the data consumed by INTERVENE has restrictions for use and these are discussed at length above. Where data is generated by the project we will strive to make the data aggregated and accessible for wide use within and in support of publications. Where this is not possible we will explore the use of an INTERVENE Data Access Committee but will do so using a controlled access repository (e.g. EGA) where possible.

Machine readable licenses

Where data is generated by INTERVENE we will conform to the FAIR principles and use machine readable licences.

Identification of persons accessing data

We will use the GA4GH passports standards and associate AAI implementation to identify users, this will be implemented in WP1/4.

6.3 Data interoperability

All genotyped datasets will be converted to VCF v4.2 format with necessary meta-data concerning their genotyping, quality control and imputation. VCF is a text file format widely used in community, which usually is stored in a compressed manner. It contains meta-information lines, a header line, data lines each containing information about a position in the genome and genotype information on samples for each position.

Genetic risk scores will be reported according to a standard described in Deliverable 2.2 (Standards for genetic risk scores), which is based on the state of the art community standard (PRS-RS) and designed to maximise interoperability with the field at large.

Ontology standards and interoperability

INTERVENE adheres to the recommendations and have aligned the goals with the EU '1+MillionGenomes' initiative, the International Common Diseases Alliance (https://www.icda.bio/) and have established a close collaboration with the Global Biobank Meta-analysis initiative (https://www.globalbiobankmeta.org/). The partners will use standards developed by GA4GH (https://www.ga4gh.org/) and federated data analysis solutions implemented by ELIXIR (https://elixir-europe.org/) and the CINECA consortium (https://www.cineca-project.eu/). In addition WP2 has meta data harmonisation outputs that are expected to address ontology interoperability.

6.4 Increase data re-use

Data licensing

Data licences and terms of use may be determined by the source biobank for data INTERVENE consumes- in this case we will adhere to these and for any derived data or data generated by the consortium e.g. PGS we will investigate use of permissive CC licences to maximise uptake. For example summary statistics stored in the NHGRI-EBI GWAS Catalog are made available using a CC-0 licence and the PGS Catalog is also investigating use of CC licences. A future action will be to look at the data generation components of the project as they proceed, for example, there are clinical trials in progress and the standards for licencing in this community may be different to maximise translational value.

Data availability for re-use

It is currently foreseen that the majority of data produced will be open access and disseminated through normal publishing practices using Green or Gold open access mechanisms. Data will be made available for re-use once it has undergone peer and editorial reviews and the final version of data is available. If generated data can reasonably be expected to be exploitable for commercial or industrial use, or if protection of generated data is otherwise possible, reasonable and justified, deviations to the availability policy will be taken on a case-by-case basis by the data owner.

INTERVENE will seek to ensure that possible embargo periods necessitated by protection of results or by journal policies are minimal, but appropriate for the embargo need. For publishing purposes, embargo periods permitted by the journal should be no more than 6 months.

We expect that INTERVENE data will be reusable by third parties but as data hasn't yet been generated this will require future work to determine whether restrictions apply. These will be documented in DMP updates.

7. Data management responsibilities and resources

Cost for FAIR compliance

INTERVENE does not yet have a detailed cost prediction for FAIR compliance. This will be addressed as data is generated. Costs for ensuring open access to research data have been included in the beneficiaries' individual budgets and it is expected that all costs pertaining to open access requirements can be fulfilled within this framework. Use of costs will be closely monitored throughout the project, and revised cost models will be prepared in connection with updates to the DMP.

Our strategy is to use public repositories and these provide long term preservation at minimal cost to the consortium. Where these are not available the task defaults to the project coordinator and the internal project repository can be used. We propose that data is retained for 10 years from the point of generation but may extend this as the project progresses.

Data management responsibilities

As a rule, the partners creating the research data are responsible for storing, hosting, and managing the data in accordance with the DMP, the grant agreement, as well as guidelines and requirements of the organization to which they belong. EMBL and CSC, together with UH-FIMM will together be responsible for the development and updating of the data management plan after input from the participating biobanks. The coordinator is responsible for all data management of common materials, such as deliverables and their updates, maintenance of contact details as well as meeting and evaluation reports. The coordinator also undertakes data management responsibility for all material uploaded on the project intranet, unless specifically agreed otherwise with the data generator.

The Data Coordination Center (DCC) will serve as the INTERVENE inventory for genomic data, providing a secure storage and computing environment for those datasets that can be used outside biobanks. Environments can be provided for both non-sensitive data and sensitive data with restricted access. DCC supports the INTERVENE project by providing technical support and resources for the project and coordinates and collects the methods, standards and policies that need to be established for the scientific use cases. DCC has a support email address that is linked to the ticketing system of CSC that is actively monitored.

8. Other issues

Use of other data management procedures

INTERVENE are using ELIXIR[6] strategies and processes for data management complemented by institutional expertise for specific data types, these are identified in Appendix 3 at a per work package granularity.

[6] https://elixir-europe.org/

Appendix 1: Data Agreement

PERSONAL DATA AGREEMENT

1 Parties

(list of Parties that process personal data in the Project)
hereinafter referred to jointly as the ”**Parties**” and individually a ”**Party**”.

2 The Purpose of the Data Agreement

2.1 This Personal Data Agreement (“Data Agreement”) supplements and forms an integral part of the INTERVENE Consortium Agreement (the “Consortium Agreement”) between the Parties.

2.2 The purpose of this Data Agreement is to agree on the roles and responsibilities of the Parties with regard to the Processing of Personal Data in the Project.

3 Definitions

3.1 The definitions used in the Consortium Agreement shall also apply to this Data Agreement.

3.2 “Controller” means the natural or legal person, public authority, agency or other body which, alone or jointly with others, determines the purposes and means of the Processing of Personal Data.

3.3 “Data Protection Laws” mean all applicable legislation concerning the protection of Personal Data that apply to the Parties (including but not limited to the General Data Protection Regulation (EU) 2016/679 ("GDPR") for Parties located in the European Union). For a Party that has the status of an intergovernmental organization, “Data Protection Laws” shall be understood as a reference to its own regulations on data protection.

3.4 Unless otherwise specified, any reference to Data Protection Laws in reference to a Party's obligations shall be understood as referring to the Data Protection Laws applicable to that given Party.

3.5 “Personal Data” means to information relating to an identified or identifiable natural person ("Data Subject"). An identifiable natural person is one who can be identified, directly or indirectly, in particular by reference to an identifier such as a name, an identification number, location data, an online identifier or to one or more factors specific to the physical, physiological, genetic, mental, economic, cultural or social identity of that natural person.

3.6 “Processing” means any operation or set of operations which is performed on Project Personal Data or on sets of Project Personal Data, whether or not by automated means, such as collection, recording, organisation, structuring, storage, adaptation or alteration, retrieval, consultation, use, disclosure by transmission, dissemination or otherwise making available, alignment or combination, restriction, erasure or destruction.

3.7 "Processor" means a natural or legal person, public authority, agency or other body which Processes Personal Data on behalf of a Controller;

3.8 "Project Personal Data" means any Personal Data processed, utilized or generated in the Project, including:

(a) Personal Data collected prior to and used in the Project, such as registry and patient specific data provided by Data Subjects, third parties or generated as part of other research projects; and

(b) Personal Data generated in the course of the Project.

3.9 "Provider" means a Party that provides Project Personal Data to another Party under this Data Agreement;

3.10 "Recipient" means a Party that receives Project Personal Data from another Party under this Data Agreement.

4 General Responsibilities of the Parties

4.1 Each Party shall act in accordance with the requirements of the Data Protection Laws applicable to it. Each Party is liable for ensuring that the Processing of Project Personal Data carried out by it is in accordance with the Data Protection Laws applicable to that Party, the Consortium Agreement and this Data Agreement.

4.2 Each Party shall:

(a) establish and document the legal bases on which it Processes the Project Personal Data as determined by the applicable Data Protection Laws, and if applicable, any special category data;

(b) take all appropriate technical and organizational measures to ensure that the Project Personal Data are Processed in accordance with the Data Protection Laws, the Consortium Agreement and this Data Agreement;

(c) and maintain a record of Processing activities as required by the Data Protection Laws.

4.3 A Party shall, at the request of another Party, provide any information about its Processing activities, as reasonably required by the other Party to comply with its obligations under the Data Protection Laws.

4.4 Each Party shall ensure that it complies with applicable laws and ethical principles regarding medical research, patient data and human samples and tissue, including but not limited to compliance with requirements related to informed consent.

5 Disclosure of Project Personal Data Between Parties

5.1 Provider may provide Project Personal Data necessary for the purposes of the Project to Recipient without separate agreements in the manner specified in this Section 5.

5.2 Provider is responsible for ensuring that it has the right to disclose the Project Personal Data to the Recipient under the Data Protection Laws to which Provider is subject. Recipient is responsible for ensuring that accepting disclosure of Project Personal Data from Provider is permitted under the Data Protection Laws to which Recipient is subject.

5.3 Recipient shall not Process Project Personal Data disclosed by Provider for any purposes other than the performance of the Project, and only to the extent necessary for the performance of the Project.

5.4 Disclosures of Project Personal Data between the Parties and the Processing activities undertaken by Recipient involving such data have been specified in Appendix 1 (Description of Data Disclosures), which may be updated from time to time. Appendix 1 may include additional criteria regarding the disclosure of specific Project Personal Data, with such additional criteria considered binding upon Recipient.

5.5 Provider may require the Recipient to conclude a separate agreement on the transfer of certain Project Personal Data if it considers that it cannot transfer the Project Personal Data under the terms of this Data Agreement, even with additional criteria specified in Appendix 1.

5.6 Project Personal Data shall be provided in pseudonymized, aggregated and/or coarsened format unless absolutely necessary for the performance of the Project. For the avoidance of doubt, Provider shall under no circumstances be obligated to disclose any additional information to Recipient that would permit re-identification of Data Subjects by Recipient.

5.7 Recipient shall not attempt to reidentify any Data Subjects.

5.8 Any transfers of Project Personal Data from a Provider located in the EU/EEA to a Recipient located outside the EEA or an international organization must comply with the requirements of Chapter V of the GDPR. The transfer instruments or safeguards shall be documented in Appendix 1.

5.9 Recipient shall securely delete all Project Personal Data received from Provider when the Project Personal Data is no longer necessary for the Project, unless otherwise agreed between the Provider and the Recipient.

6 Confidentiality

6.1 Each Party shall ensure that:

(a) only persons who need the Project Personal Data to carry out tasks related to the Project can access the Project Personal Data;

(b) a Party's personnel authorized to Process the Project Personal Data have committed themselves to confidentiality or are under an appropriate statutory obligation of confidentiality; and

(c) personnel who have access to the Project Personal Data are aware of their obligations concerning the Processing of the Project Personal Data and will Process the Project Personal Data in accordance with the Data Protection Laws, the Consortium Agreement and this Data Agreement.

6.2 Recipient shall hold all Project Personal Data received from Provider in strict confidence and not disclose it to any third party or another Party without the explicit approval of Provider.

6.3 Notwithstanding the confidentiality terms in the Consortium Agreement, the Recipient's confidentiality obligations regarding Project Personal Data shall continue in force in perpetuity, unless otherwise agreed between the Provider and the Recipient.

7 Data Processors

7.1 Each Party shall:

(a) impose on each of its Processors the obligations which Controllers are obliged to impose on Processors under the Data Protection Laws;

(b) monitor each of its Processor's compliance with those obligations and ensure that each Processor complies with those obligations; and

(c) be liable to the other Party for the acts and omissions of its Processors as though they were the Party's own acts and omissions.

8 Informing Data Subjects

8.1 The Parties shall inform Data Subjects and each other about the Processing activities conducted in the Project.

8.2 Each Party is responsible for providing the information required by the Data Protection Laws to Data Subjects and complying with applicable informed consent requirements regarding Project Personal Data that it collects or has collected. In addition to what is stated in Section 4.3, Parties may also collaborate to draft privacy notices, information and consent letters and other information meant for the Data Subjects.

9 Data Subject Rights

9.1 Each Party that receives a request concerning the use of the rights of a Data Subject or withdrawal of informed consent must, without undue delay, communicate the request or withdrawal to the other Parties who Process Project Personal Data concerning the Data Subject or that are otherwise affected by the request/withdrawal.

9.2 Provider may at any moment inform Recipient of a Data Subject withdrawing his/her consent or exercising his/her rights as a Data Subject under the Data Protection Laws. Recipient agrees to perform all actions reasonably necessary specified by Provider in order to fulfil the rights of the Data Subject as required by the Data Protection Laws to which Provider is subject. Additionally, Recipient is responsible for taking all necessary and appropriate measures required by the Data Protection Laws to which Recipient is subject in order to fulfill the rights of Data Subjects.

10 Data Breaches

10.1 If a Party becomes aware of a personal data breach concerning the Project Personal Data, it shall communicate the breach to the other Parties without undue delay. Each Party shall also notify the other Parties about any other problems or disturbances that may have an effect on the rights and freedoms of the Data Subjects.

10.2 Each Party affected by a personal data breach shall take measures in order to address the personal data breach and mitigate and rectify its effects.

10.3 The Parties affected by the personal data breach shall notify the personal data breach to the competent supervisory authority/authorities and the affected Data Subjects without delay in accordance with the Data Protection Laws. Each Party shall give the other affected Parties the information required for the notification.

11 Notices

11.1 Notices regarding Data Subject rights or personal data breaches shall be made to the contacts specified below in this Section 11.

11.2 Other notices shall be made in accordance with the Consortium Agreement.

Name of participating Institution (Name in short form):

Name of contact, Title of contact

Email:

Tel.:

[contact details of other Parties]

12 Term and Termination

12.1 This Data Agreement shall become effective retroactively from the effective date of the Consortium Agreement and remain in effect for the duration of the Consortium Agreement.

12.2 In the event of either a) termination or expiration of the Consortium Agreement, or b) if required by Provider after Recipient committed a breach of its obligations under this Data Agreement, Recipient shall return all Project Personal Data it has received from Provider and all copies thereof to the Provider, or, at Provider's choice, will destroy all copies of the same and certify to the Provider that it has done so, unless Recipient is prevented by the laws applicable to it or by Recipient's national authorities from destroying or returning all or part of such data, in which event the data will be kept confidential and will not be actively Processed for any purpose.

Appendices

Appendix 1: Description of Data Disclosures Signatures

The Parties have caused this Data Agreement to be duly signed by the undersigned authorised representatives in separate signature page**s.**

Appendix 2. Data sources

Shortlist of the data collected within the INTERVENE distributed repository (Grant Agreement Annex 1B)

Biobank /Medical repository	Existing samples	Disease coverage	Population coverage	Partner
	# samples	Type of diseases	Ancestry, ancestry, sex, age	
FinnGen	500,000	All	• European/Finnish • 56.7% female and 53.3% male • Average age of 52 (Std Dev 17.3)	**(1) UH-FIMM**
Helsinki Biobank	100,000	All	• European/Finnish • 63% female and 47% male • Average age of 54 (Std Dev 20.22)	**(13) HUS**
Estonia Biobank	200,000	All	• European/Estonian • 67.3% female and 38.7% male • Average age of 44 (Std Dev 16.0)	**(7) UTARTU**
The HUNT Study	100,000	All	• European/Norwegian • 52% female and 48% male • Average age of 52	**(5) NTNU**
NIG – Network for Italian Genomes	3,000	Rare diseases (mainly ID)	• European/Italian • 51.3% female and 48.7% male • Average age of 38 (Std Dev 22.48)	**(4) UNISI**
Partners biobank	115,000	All	• 85% European, 5% Black, 2% Asian • 56% female and 44% male • Average age of 56 (Std Dev 17.58)	**(15) MGH**
UK biobank	500,000	All	• 90% European, 2% Asian, 2% Black • 54% females and 46% males • Average age of 57 (Std Dev 8.09)	**N/A**
Genomics England	100,000	Rare diseases (70%) and cancer (30%)	• 68% European, 8% Asian, 2% Black • 53% females and 47% males • Average age of 45 (Std Dev 22.00)	**N/A**
Genes & Health	100,000	All	• Pakistani, British Pakistani, Bangladeshi, British Bangladeshi. • 54.7% female and 45,3% male • Average age: 45.1 (Std Dev 17.1)	**(19) QMUL**

Partner / data set	Existing samples	Disease coverage	Population coverage	Existing genomic dataset			Existing omics samples					Available phenotypic/registry data linking to the genomic data			Additional data / features
	# samples	Type of diseases	Ethnicity, ancestry, sex, age	GWAS	WES	WGS	Metabolomics	Metagenomics	Epigenetics	Proteomics	Transcriptomics	Health register	Registered data	Host institution	e.g. recall data
FIMM UH	500.000	all	European/Finnish, 56.7% females, average age 52 (stdev 17.3)	500.000	25.000	10.000	50.000		2.000		1.500	500.000	500.000	FIMM	Yes
Estonia Biobank (Tartu)	200.000	all	European/Estonian, 67.3% females, average age 44 (stdev 16.0)	150.000	2.500	2.600	11.000		700		1,700 (array + mRNA seq)	148.000		Estonian Biobank	
HUNT / NTNU	100.000	all	European/Norwegian 52% females, average age: 52 (stdev)	70.000		2.200	17.000			3.500		150.000		Norwegian University of Science and Technology	Yes (for genetic recall for a subset of the individuals)
NIG	3.000	rare diseases (mainly ID)	European/Italian, 51,3% females, average age: 38 (stdev 22.48)		3.000							3.000		Nine Institutions belonging to NIG	Yes
Genes & Health	100,000	all	Pakistani, British Pakistani, Bangladeshi, British Bangladeshi. 54.7% female, average age: 45.1 (stdev 17.1)	60.000	5,000							100,000		Queen Mary University of London	Yes (recall by genotype for some individuals
Partners biobank	115000	all	85% European, 56% females, average age: 56 (stdev: 17.58)	35.000								35.000		Partners Healthcare	Yes (algoritm-defined disease endpoints and super controls)
UK biobank	500.000	all	90% European, 54% females, average age: 57 (stdev 8.09)	500.000	50.000							500.000		Uk Biobank	
Genomics England	100.000	rare diseases (70%) and cancer (30%)	62% European, 53% females, average age: 45 (stdev 22.00)			100000						100.000		Genomics England	
HUS Helsinki Biobank	100.000	all	European/Finnish, 63% females, average age: 54 (stdev 20.22)	22,000 (NOTE: overlaps with FIMM UH cohort)								100.000		HUS	Yes

Appendix 3. INTERVENE work package specific data management information

We have constructed a living document that provides detail for each WP which is accessible to consortium members, a screenshot is shown below as an excerpt from the work package level detail.

https://www.google.com/url?q=https://docs.google.com/spreadsheets/d/1ZVpwOOC-_stPUWTCMQvX50EaerhuwWHXSKmlGDX3M1Q/edit%23gid%3D319325758&sa=D&source=editors&ust=1623425478884000&usg=AOvVaw3cu1YGRhpR7RQs4kQU1WoX

WP	Data type, e.g. survey called X	Is the data collected, produced, or reused?	What data production/collection methods will be used?	If the data is reused, specify from where.	File formats of the data e.g. .txt	Volume (estimate of size ~)	How/Where will the data be stored during the project?	Will data be shared within INTERVENE?	Will data be shared outside of INTERVENE?	What data and metadata vocabularies, standards or methodologies will you follow	Will the data be made available for reuse? Is an embargo period needed?	Where will the data be made available? (repository name e.g. EGA)	When will the data be made available (estimate)?	COMMENTS
WP1	Sofware container	Produced/ Generated	Sofware repository	Data will be reused first during the software development process and then when the actual use cases are executed.	Sigularity, Docker	Less than 5 TiB	CSC Object Storage services or virtual machine volumes in cPouta.	Yes	Mostly yes.		Yes. No embargo	INTERVENE web pages linking to CSC Object Storage.	2022	
WP1	Public reference data sets	Reused	Copying from other repositories	Data will be used as the training material in software development	Depends on the original repository	Less than 50 TiB	CSC Object Storage services or virtual machine volumes in cPouta.	yes	yes		Yes. No embargo	INTERVENE web pages linking to CSC Object Storage.		
WP1	Sensitive biobank data	Reused	Download processes on biobanks	Use cases can utilize this data with permission of the biobank	Crypt4gh encrypoted format. inside the ecryption layer original biobank format will be used	100 TiB	SD-connect / CSC	0	No			SD-connect	2022	
WP1	Survey of biobank data access and ethics policies	Collected	Survey	NA	.xlsx	negligible	EMBL-EBI	yes	yes, in the form of Deliverable 1.2				2021	Filled by Aoife McMahon
WP2	Public GWAS summary statistics and/or polygenic risk score metafiles	Reused	Downloading from repositories	Data will be used to develop new risk scores or repeat calculation of already published risk scores in biobanks	Depends on the original repository	Less than 200 GB	Data will be stored in University of Tartu servers for calculations done for Estonian Biobank	yes	probably		new polygenic scores are developed, the yes, they will be shared and probably embargo is needed until the scientific results are published			
WP2	Sensitive biobank data	Reused	Access in appropriate servers where data is stored	data can be accessed with permission of the biobank	Will depend on the biobank			No	No					
WP2	GWAS	Produced	Analysed in HUS		.gwas	10 GB	In HUS datalake	Yes	No			CSC		
WP2	Biobank sample data	Collected, reused	Collected in HUS	Helsinki Biobank repositories	.CSV	100 MB	In HUS datalake	If needed	No					
WP2	PRS	Produced	Analysed in HUS		.txt	100 MB	In HUS datalake	Yes	No			CSC		
WP2	Biobank clinical data	Reused	Collected from HUS EHR data	HUS EHR	.CSV	10 GB	In HUS datalake	If needed	No					
WP4	Polygenic risk scores (scoring files)	Reused	Data will be generated on the Intervene platform using scripts and or workflows and also remotely in biobank environments. Data generated will be indexed by the WP1 DCC	Scores developed by other INTERVENE work packages	.tsv, other?	Depends on the number of analyses and is not yet quantifiable as data is not yet available	Initially stored by score developers	yes	yes	PGS-RS, as described in Deliverable 2.2	yes	PGS Catalog	ongoing	specifying this data is reused as it is not generated by this WP, it is generated by other INTERVENE WPs.
WP4	Individual-level genetic scores	Produced/ Generated	Scores calculated on individual level genotype data	0	text	Depends on the number of analyses and is not yet quantifiable as data is not yet available	Available from the DCC	yes	aggregate analysis results	N/A	yes, if the data source permits.	Controlled access repository when supporting a publication if permitted by the data source	At publication	Dependencies on data sources for access rights, some require return of data
WP5	Questionnaire data	Collected					Pseudonymised data from the Finnish and Italian arms of the WP5 breast cancer pilot will be analysed in a secure computational environment, most likely at FIMM computing cluster or in a private cloud environment on server within EU.	No	No					Breast cancer study data
WP5	Personal health data	Reused		information collected from the participants' personal health record			same as above	No	No					Breast cancer study data
WP5	Genetic data	Produced/ Generated	Exome sequencing, GWAS array gentoyping				same as above	No	No					Breast cancer study data
WP5	Estonian Biobank (EstBB)	Reused	linking	from Estonian Biobank database	*.csv	500 MB	secured server of Tartu University	Not planned, can be done with approval of Estonian Commitee on Bioethics and Human Research	Not planned, can be done with approval of Estonian Commitee on Bioethics and Human Research	0	Original data can be reused by third parties after approval of Estonian Commitee on Bioethics and Human Research	Publicly not at all, aggregated data is avilable genomics.ut.ee		Estonian data protection practice is very strict. Original dataset is available after approval of Estonian Commitee on Bioethics and Human Research. Aggregated data is
WP5	Survey of study participants	Collected	face to face interview		*.csv	10MB						redcap.ut.ee		
	Family doctors	Produced/ Generated	examining and interviewing participants and their medical records, taking bloodsamples		*.csv	10MB						redcap.ut.ee		
WP6	Survey of biobank data access and ethics policies	Collected	Survey	NA	.xlsx	negligible	EMBL-EBI	yes	yes, in the form of Deliverables				2021	Filled by Aoife McMahon
WP7	Limited personal data of INTERVENE consortium participants and advisory board members (email, other contact information, date of birth)	Collected	Request through email or surveys		.xls	5MB	names and email addresses stored on University of Helsinki servers and accessible through the project INTRANET on wiki.helsinki.fi	Yes	No. Although the data may be accessible through public sources, it will not be made public due to potential misuse, e.g. by spam software.	N/A	N/A	N/A	M2	Continuously updated according to consortium composition. All data managed by the Coordinator
WP7	Survey of feedback, special preferences and considerations, including personalized nutrition in connection with annual meetings within consortium (INTERVENE consortium participants and advisory board members)	Collected	Survey		.xls	5MB	Data stored and managed by coordinator's project management office. Data deleted when no longer needed for INTERVENE	Partially - shared only within Meeting organizing group	No					
WP7	Transcripts or minutes from internal steering group meetings, workshops, etc	Produced/ Generated	Notes, collection of pre-produced material		Standard text formats	300 MB	Data stored and managed by coordinator's project management office.	Yes. Access may be restricted to e.g. Management Board or advisory boards	No					
WP8	Market Research stakeholder interviews													
WP8	Market Research e-surveys	Collected	Survey	-	word, excel	10MB		Yes	No		NA	INTERVENE intranet		
WP8	Market report	Produced/ Generated		-	word, excel	10MB		Yes	No		NA	INTERVENE intranet	M18	
WP8	IGS4EU Governance Interview	Collected	Interview	-	word, excel	10MB		Yes	No		NA	INTERVENE intranet		
WP9	General comment: WP9 will not collect, produce, or reuse data. Deliverables and outputs will restate data collected as part of WP6.													