INTERVENE

Report of available data resources and plan for data harmonization strategy

Deliverable 2.1

Dissemination level: public

Part of:

WP2:

Data harmonisation, integration, and evaluation of genetic scores

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.	2.1
Work package	WP2
Deliverable lead	UTARTU, EMBL, FIMM
Contributors	NTNU, HUS, QMUL, UNISI
Dissemination level	Public
Expected delivery date	31/10/2022
Submission date	31/10/2022

Change history log

Version	Changes made	Prepared by (name & organization)	Reviewed by (name and organization)	Date
0.1	First draft	Krisit Lall, UTARTU	Reedik Mägi, Triin Laisk - UTARTU	26.3.2021
1.0	Revision		Julius Anckar, FIMM	31.3.2021
2.0	Revision based on European Commission review: comprehensive restructuring, additional content added and reformatted	Kristi Läll &Triin Laisk (UTARTU), Aoife McMahon (EMBL-EBI)	Helen Parkinson (EMBL-EBI) Reedik Mägi (UTARTU) Julius Anckar (UH-FIMM)	Aug-Oct 2022

Contents

1. Introduction

A large part of the INTERVENE activities revolve around analysis of data originating from different biobanks using different methods, including artificial intelligence (AI) methods. It is crucial to have harmonized datasets for model training and validations when applying AI-based approaches, as different data structure and formatting may make the analysis impossible or yield false results, if the data is not cleaned and interoperable beforehand. The purpose of this deliverable is to lay the foundation for a data harmonization strategy for the INTERVENE project by mapping the data sources and standards of the biorepositories participating in INTERVENE Data either needs to be brought into a common format across all participating cohorts, or be made interoperable, which will allow at least partially automated analysis. Currently, each cohort's data variables are stored in different formats and are often based on different sources. For example, disease information can be obtained from questionnaires, linking to national registries, or derived from electronic health records through data mining techniques such as natural language processing. This deliverable will characterize existing data based on input from project partners and propose strategies to create harmonized phenotypes across participating cohorts.

2. Methods

2.1. Data collection methodology

To map the current data formats and standards of INTERVENE cohorts as a basis for a data harmonization strategy, we sent out two separate online surveys to 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 (EstBB), and Network for Italian Genomes (UNISI)). Each participant was asked to complete two surveys, the first on the availability,

coding, and amount of different data types in the cohorts. A second survey on the local legal and ethical requirements to access the data was also circulated, with the purpose of starting the process of obtaining ethical approvals and MTA agreements for the studies.

In the surveys we collected information on:

a) data access requirements
b) local legislation relevant for the project
c) potential sensitive content of data
d) technical aspects of the data (including data format, structuring of the data, data coding, data schema and Observational Medical Outcomes Partnership (OMOP) model availability)
e) availability of different 'omics' type of data
f) phenotypic data availability together with information regarding medical history and possible updates from different sources.

The full questionnaires are provided in the supplement (see Appendices 1 and 2). Questionnaires were based on PRECISE4Q (Predictive modelling for stroke) data transfer surveys[1]. This was selected as the basis for the questionnaires because as in the INTERVENE project, part of the H2020 funded PRECISE4Q project (Grant agreement No. 777107) was to study legal and technical aspects of data across participating cohorts to allow for harmonization of the data. Their goal was also to build a warehouse to host and provide pooled data from several data sources, which is also an interest to INTERVENE partners. By reuse of the prior work, we introduce consistency across Horizon 2020 funded projects.

[1] https://data.qmenta.com/p4q/data_transfer.html

3. Results

3.1. Mapping cohort data content and standards

The structured and summarised results of the INTERVENE data transfer survey are presented in Table 1 and give an overview of data format, language, accessibility, structure, coding, updating and downloading opportunities and other similar characteristics. A more comprehensive review of data access and analysis processes is available in D1.2 (Review of the consortium partner biobank and data collections, including access policies)[2].

Data features	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
Structured data	y	y	y	y	y	y	n	y
Data coded	y	y	y	y	y	y	y	y
ICD-10 codes	y	y	y	Yes, ICD-9 for older cases	y	y	y	y
Exact dates of ICD-10 codes	y	y	y	y	y	y	y	y
OMOP model	Under consideration	y	n	n	Under consideration	n	In process	In process
Images	y	n	n	y	n	n	n	y
Free text, semi structured data	y	n	y	y	n	n	y	y
Data updated with registries, health care services	y	y	y	y	Yes, upon request	y	y	y
Data accessibility	Internal analysis using provided scripts; Federated access	Internal analysis using provided scripts	Internal analysis using provided scripts, can be accessed via remote sandbox	Can be downloaded	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
Data egress	Rare cases	individual-level data cannot be downloaded, only aggregated data	n	y	n	Only export of summary stats is allowed	n	n

Table 1. Overall description of INTERVENE data repositories. Color scheme – blue(yes), orange (no). For more information, see the main text.

[2]Deliverable 2.1: https://www.interveneproject.eu/_files/ugd/17206f_4cb7eda889874e1bb905a798296c23dd.pdf

3.1.1. Data structure

We received and consolidated responses from eight INTERVENE biorepositories (EstBB, HUS, FinnGen, Genomics England, UKBB, NTNU, QMUL and UNISI). First, we explored the data structures of the cohorts. Phenotype data is available in Excel or similar spreadsheet-like format for five partners, in SQL or similar language database for three partners and in any desired format for one partner. Structured data exists for seven partners out of eight, free text and semi-structured data is available in three cohorts, and imaging data is available in three cohorts. Observational Medical Outcomes Partnership (OMOP) data model[3] allowing different databases to be incorporated into a common format – is already in use in one cohort. Two partners report that adoption of OMOP is underway, whereas four partners report that OMOP is theoretically possible, but not currently in process.

3.1.2. Data coding, schema and language

Data is already coded with clinical terminologies within all cohorts, the choice of these is often governed by the applications used to collect the data and/or country level decisions on which clinical coding system is used in healthcare or health data research. Clinical coding systems are often commercial products and this brings some challenges to harmonising between them. Many different coding systems are being used across the partners (ICD-10, ICD-9, The Anatomical Therapeutic Chemical classification (ATC) was reported in Finngen and in HUS, SNOMED in HUS and Genomics England and NOMESCO classification of surgical procedures (NCSP) in HUS), but ICD-10 codes are available in all the cohorts .

Data schema (ie description of the data structure) at least in some format is available for five partners out of eight. For four partners, data is entirely in English, for another three partially in English and for one, in Finnish. Exact dates for recruitment are also available in all cohorts. Data is also being regularly updated within all participating cohorts via linking with national databases, registries and health service data.

3.1.3. Available data (genotype, phenotype and other -omics)

All partners reported having genotype data, with largest data cohorts in UKBB, FinnGen and EstBB. UKBB and FinnGen have genotype data for nearly 500,000 individuals each, whereas EstBB has data for 200,000 individuals (Table 2). Genomics England has 100,000 individuals, NTNU 70,000 and they will add another 2000 by the end of 2021. QMUL has 34,000 individuals and more will be added soon. UNISI has 3000 individuals and HUS 22,000 individuals (note- this is overlapping with FinnGen data). The data is in different formats (*.bgen, *.vcf, plink format, etc). Data is almost always available in imputed format. Regarding different -omics datasets, metabolome data was reported available for a subset of individuals in four cohorts (EstBB (n=10,800), NTNU (n=8000 now, +9000 soon) and UKBB (n=121,000), UNISI (n=50)). Proteomics information is available for a small number of individuals in the EstBB(n=500) and in NTNU

[3] https://www.ohdsi.org/data-standardization/the-common-data-model/

(n=3500). Microbiome data is available in EstBB (n=2509), in HUS (n= unkown) and in NTNU (n=5500). Chromatin data was not reported available in any participating cohort, methylation and transcriptome data availability in other cohorts very limited.

Table 2. Data types generated by INTERVENE partners. Blue color (and letter "y") indicates that data is available, orange color (and letter "n") indicates that it is not.

Data types	EstBB	FinnGen	Genomics England	UKBB	NTNU	QMUL	UNISI	HUS
Genetic data (number of individuals)	~200k	500k	100k	500k	70k (+20k by 2022)	44k	3k	22k
Genetic data format	vcf	vcf	WGS	bgen	Vcf, sav	Bgen, pfile	WES	vcf
Metabolome data	NMR N=10800, clinical biochemistry N=2650	n	n	N~121k	N ~8k, +9k in 6 months	N~50	n	y
Microbiome data	Shotgun metagenomic sequencing, stool samples, N= 2509.	n	n	n	N=5500, targeted assay	n	n	y
Proteomics data	4 arrays, N=500	n	n	n	N=3500	n	n	n
Chromatin data	n	n	n	n	n	n	n	n
Methylation data	methylation array, N=450	n	n	n	Very limited	n	n	n
Transcriptome data	Purified CD4/CD8 t cells N=293 Whole Blood cell expression N=917	n	n	n	Very limited (lung cancer, colon cancer, psoriasis)	n	RNA-Seq n<10	n

Availability of phenotype data is summarised in Table 3. Five partners report that they have full information about prescribed drugs, two partners report that they have partial data, one partner reports that they are in process of obtaining the data. History of adverse effects of prescribed drugs are not available for three out of eight partners, partial data is available for three partners.

Table 3. Phenotype data description. Blue color (and letter "y") indicates that data is available, orange color (and letter "n") indicates that it is not.

Phenotype data	EstBB	FinnGen	Genomics England	UKBB	NTNU	QMUL	UNISI	HUS
Objective information	y	y	y	y	y	y	n	y
Education and work-related questions	y	y	y	y	y	n	n	n
Smoking habits	y	y	y	y	y	n	n	y

Alcohol consumption	y	n	n	y	y	n	n	n
Nutrition data	y	n	n	y	y	n	n	n
Sleeping habits	y	n	n	y	y	n	n	n
Physical activity	y	n	n	y	y	n	n	n
Reproductive health	y	n	n	y	y	n	n	y
Family history of diseases	y	y	y	y	y	y	y	n
Medical history of diseases	y	y	y	y	y	y	y	y
Pharmacological data	y	n	y	Limited	y	n	n	y
Prescribed drugs	y	y	y	Limited	y	In process for all	y	Limited
Adverse reactions to prescribed drugs	y	n	n	n	n	n	Limited	Limited

3.2. Data harmonization and standardization plan

3.2.1. Genetic Data

We have decided that all genotyped datasets will be converted to VCF v4.2 format (if possible) with necessary meta-data concerning their genotyping, quality control and imputation. The VCF format is chosen because it is a text file format widely used in the 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. VCF format specification is available here[4]. Polygenic risk score file format and reporting standardisation is addressed in Deliverable 2.2 (Plan of data standards for genetic risk scores)[5].

3.2.2. Phenotype Data

3.2.2.1. Selection of disease endpoints for study in the INTERVENE project

We first needed to decide which disease endpoints should be the subject of study, and therefore harmonisation, in the INTERVENE project. We have prepared a first draft of endpoints, to be finalised in Deliverable 2.3 (Identification of flagship diseases and secondary phenotypes) (Table 4). A full description of the process is described in D2.3, but in brief the list was compiled while focusing on broadly defined causes of death/injury with highest estimation of overall burden, expressed as the number of years lost due to ill-health, disability, or early death (DALY)[6] according to Global disease burden tool[7] and factoring in clinical expertise and research interests of partners. This list is meant to cover a wide range of medical conditions/disease and focuses mainly on cancers and cardiovascular diseases as they are the top two causes of death/injury in high socio-demographic index (SDI) countries in 2019 according to DALYs.

[4] https://samtools.github.io/hts-specs/VCFv4.2.pdf.
[5] Deliverable 2.2: https://www.interveneproject.eu/_files/ugd/17206f_d925b40d3176487a84e5b5e02daf9a30.pdf
[6] Vos et al, 2020. Global burden of 369 diseases and injuries in 204 countries and territories, 1990–2019: a systematic analysis for the Global Burden of Disease Study 2019. PMID 33069326
[7] https://vizhub.healthdata.org/gbd-results/

Finalisation of this list will be addressed in Deliverable 2.3 (Disease endpoint selection based on available resources and AI based methods).

Table 4. Draft disease endpoints selected for study in the INTERVENE project

Clinical Area	***Disease***
Neurology	Epilepsy, Focal epilepsy, Generalized epilepsy, Subarachnoid haemorrhage, Late onset Alzheimer's disease, Parkinson disease
Psychiatric	Depression, Bipolar affective disorders, Alcohol use disorder
Rheumatology	Seropositive rheumatoid arthritis, Psoriatic arthritis
Endocrinology	Hypothyroid (broad), Hypothyroidism (congenital or acquired), Obesity, Type 1 diabetes, Type 2 diabetes
Gastroenterology	Inflammatory bowel disease, Crohn's disease, Ulcerative colitis, Appendicitis, Hernia
Oncology	Melanoma of skin, Chronic lymphocytic leukaemia, Non-Hodgkin's Lymphoma, Malignant neoplasm of colon, Malignant neoplasm of bronchus and lung, Malignant neoplasm of breast, Uterine cancer, Thyroid cancer, Malignant neoplasm of prostate, All cancer
Pulmonology	COPD, Asthma, Sleep apnoea, Idiopathic pulmonary fibrosis
Cardiometabolic	Atrial fibrillation and flutter, Heart failure, Stroke (includes all strokes), Stroke (excluding subarachnoid haemorrhage), pulmonary embolism, Venous thromboembolism, Coronary artery disease/Cardiovascular disease (CAD/CVD) (major coronary heart disease event), Myocardial infarction, Abdominal/Thoracic aortic aneurysm, Hypertrophic obstructive cardiomyopathy, Chronic kidney disease
Musculoskeletal	Osteoarthritis, Coxarthrosis, Gonarthrosis, Osteoporosis, Gout
Ophthalmology	Primary open-angle glaucoma (POAG)
Infectious disease	COVID-19 (susceptibility and severity), Respiratory Syncytial Virus (RSV) severity
Other	Migraine, Low back pain

3.2.2.2. Definition of endpoints using ICD-10 codes

Since all surveyed cohorts employ ICD-10 coding we decided to use this standard coding as the basis of endpoint definition. Each disease endpoint requires a formal definition of inclusions and exclusions of ICD-10 codes to ensure we are considering comparable endpoints. To take advantage of and remain interoperable with ongoing phenotype standardisation efforts within the consortium, we chose to use the ICD-10 code definitions used by the FinnGen clinical endpoint library.

Within the FinnGen project, a comprehensive library of disease endpoint definitions covering the whole 21 chapters of diseases in ICD-10 were created in a hierarchical manner. The Finngen clinical endpoint library generally follows the ICD-10 hierarchy where chapters are divided into code blocks of similar diseases, containing three-character categories of usually single diseases, and four-character

subcategories, usually disease subtypes. Endpoint definitions were carefully curated by a large group of healthcare professionals and medical doctors. The full list of experts in this project can be found in Supplementary Table 1. A computer readable file with all the endpoints definitions is available[8], as well as a graphical representation of the code inclusion and exclusion workflow for each[9]. An example of an endpoint (venous thromboembolism, VTE) (Figure 1) highlights the process of definition whereby a specific combination of ICD code inclusions and exclusions defines the endpoint. FinnGen endpoint definitions are versioned, and we will use the Data Freeze (DF) 8 version[10].

[8] https://www.finngen.fi/en/researchers/clinical-endpoints

[9] https://risteys.finngen.fi

[10] https://www.finngen.fi/sites/default/files/inline-files/FINNGEN_ENDPOINTS_DF8_Final_2021-10-08_public.xlsx

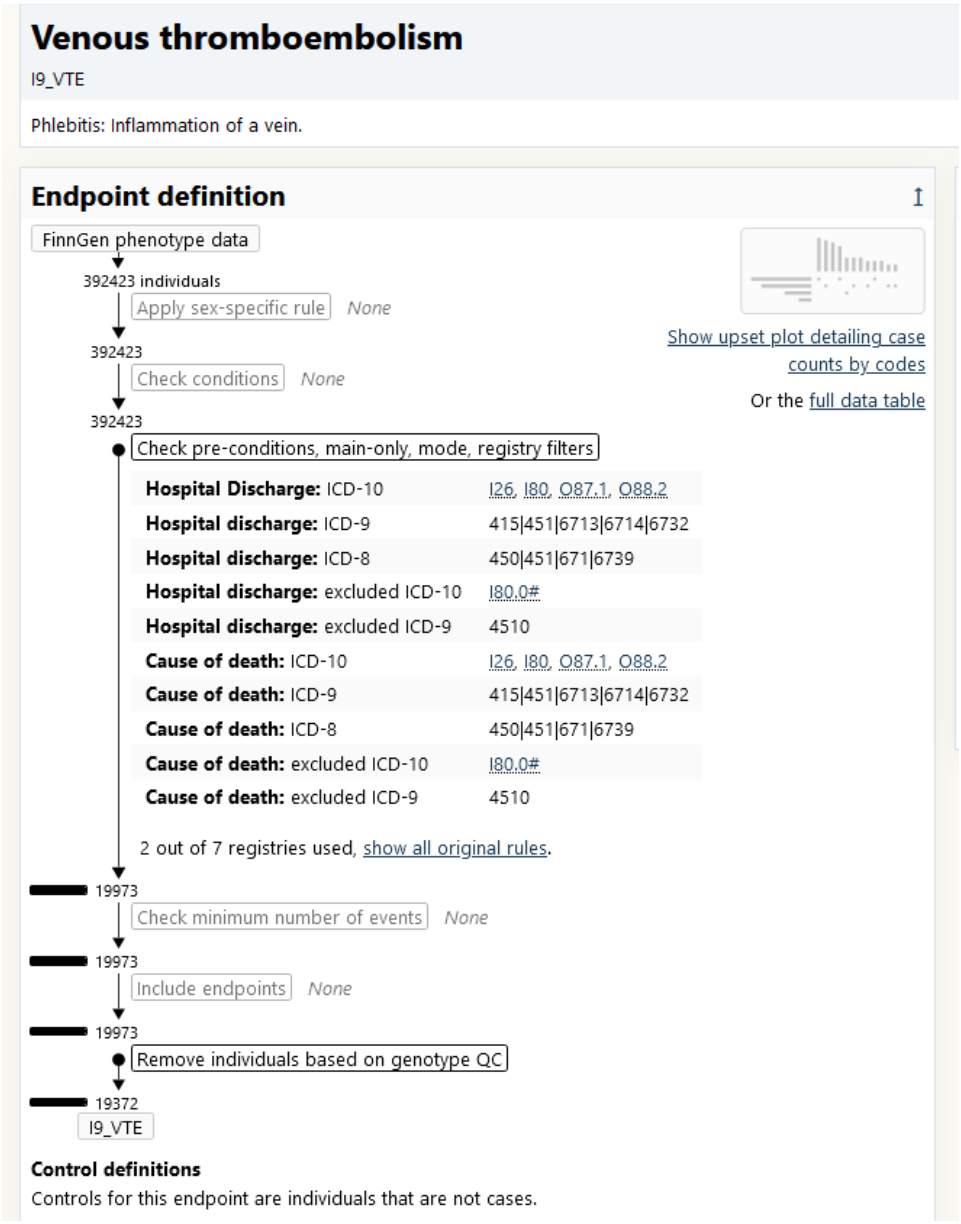

Figure 1. Example of ICD-10 endpoint definition (Venous thromboembolism) , with workflow executed in Finngen data. For this particular endpoint, no sex specific or other phenotype data base rules were applied, these may be applied for other endpoints. To define cases (and apply exclusion criterias for cases), two registries (Causes of Death and Hospital discharge information) were used In FinnGen project,whereas INTERVENE will use ICD codes regardless of their origin. For VTE, there are no lower level endpoints on which to base the definition, i.e. it is not a combination of other existing endpoints. No exclusion criterias for controls were applied.

3.2.2.3. Endpoint harmonization strategy for cohorts with non-ICD coding

As the INTERVENE consortium is open to include other biobanks in later stages of the analysis, a strategy must be developed to apply endpoint definitions to biobanks which do not use ICD-10 codes. For

instance, BioBank Japan, which has shown interest in joining risk score analyses efforts, does not use ICD-10 codes, however, they have disease names present in their data dictionary. For example in BioBank Japan disease names such as lung cancer are provided together with date of diagnosis and additional information such as site of tumor occurrence, date of treatment start, biopsy and histology related info[11]. Therefore it is possible to use disease names to create endpoints similar to clinical endpoints, but it is not straightforward to do so at scale or to use the exact same workflow and criteria as done for biobanks with ICD codes. We therefore investigated mapping strategies for cross dataset interoperability.

INTERVENE partner EMBL-EBI has a terminology mapping service that provides a set of precomputed mappings between terminologies in its Ontology Cross Reference Service (OxO)[12] . OxO acquires mappings from two sources: firstly, mappings between terms are declared in ontologies and these are acquired by OxO from the Ontology Lookup Service[13], secondly, mappings can be directly acquired from groups which have performed a mapping for a specific purpose. This could be curation of mappings for a specific data integration need or a declarative mapping between terminologies in clinical use such as Snomed-CT and ICD-10.

The critical parameters are provision of the list of terms to be mapped as a tab delimited list and choice of the ontology(ies) to map to. When selecting an ontology to map to the choice can be complex. Parameters such as coverage - how many of the terms can be mapped are important if all data is to be represented. Given the expertise within INTERVENE we expect that mappings generated using OxO would be manually reviewed prior to use in analysis.

Several terminologies exist that are likely to be a good match for cohort data and a mapping evaluation suggests that the Experimental Factor Ontology (Malone et al, 2010)[14] which is used by the NHGRI-EBI Genome Wide Association Studies (GWAS) Catalog[15] and Polygenic Score (PGS) Catalog[16] gives good coverage. UMLS already maps ICD-10 and Snomed CT, noting that the country editions of Snomed CT may introduce differences into the mappings. Once we receive a list of disease endpoints we will use OxO, test multiple ontologies and then will provide the mapping for review by experts who will also have access to contributing phenotype lists from the Japanese cohort. This process has been used in many projects and given the limited number of disease endpoints in INTERVENE is expected to be feasible and will bring the precision needed to ensure disease endpoints are comparable and documented. Once a validated set of mappings is generated, we will provide this from OxO for use by others ensuring our work is transferable to other groups and disseminated.

[11] Biobank Japan data dictionary: https://biobankjp.org/en/info/pdf/cohort_1st.pdf
[12] https://www.ebi.ac.uk/spot/oxo/
[13] https://www.ebi.ac.uk/ols/
[14] Malone et al, 2010. Modeling sample variables with an Experimental Factor Ontology. PMID 20200009
[15] https://www.ebi.ac.uk/gwas/
[16] http://www.pgscatalog.org

3.2.3. Cohort data dictionary harmonization strategy

Beyond genetic data and primary disease endpoints, cohorts collect a broad range of data items, as illustrated in Tables 1-3. It is an aim of INTERVENE to create integrative genetic risk scores, that is scores that integrate genetic risk with other data types e.g., biomarkers, clinical measurements, other -omics data. Therefore, for INTERVENE analysts to design experiments and create these scores they need to be able to understand which data types are available across the cohorts. While the surveys undertaken for this deliverable to extract this information from the cohorts and collate it here for use in INTERVENE are a start, the level and detail of information is not sufficient for experimental design. More broadly, the INTERVENE projects exists in a world of increasing use and interest in the data within cohorts. There's a global need to have interoperability between data generated in cohorts to validate findings across cohorts, for example, replication of GWAS findings and to ensure that all human variation is represented, providing equitable access to, and benefits from research data.

For these purposes interrogation of the cohort's data dictionary which comprises all the variables measured in a cohort is required, and we plan to use and extend existing work in this area to provide a layer of federated data discovery to INTERVENE cohorts. The H2020 funded Common Infrastructure for National Cohorts in Europe, Canada and Africa (CINECA) project[17] (grant agreement No.825775) has been working towards creating a semantic representation of a metadata model to standardise the diverse variables across cohorts[18]. The International 100K cohorts consortium (IHCC)[19] is leveraging this work to standardise cohort data representations and make them accessible in a cohort atlas[20]. Multiple INTERVENE partners are engaged in one or both of these projects (e.g. EMBL-EBI, CSC, UTARTU, QMUL, BBMRI-ERIC) providing opportunities to synergise across ongoing initiatives.

During the time INTERVENE has been running CINECA/IHCC have provided tools to achieve cohort meta-data harmonisation, such as the Genomics Knowledge Cohort Ontology (GECKO)[21] which provides a standard representation of terms and their attributes commonly used for genomics cohort description as well as individual-level data items. A series of tools is being developed to enable automated generation of harmonised data files based on a JSON schema mapping file. We expect to use GECKO and interactions with IHCC in future to bring cross cohort harmonisation strategies to cohorts to deliver a more sustainable approach to mapping than the current cohort-by-cohort and research question by research question model.

4. Discussion and next steps

We have gathered comprehensive information about the available biobanks participating in the INTERVENE project. Detailed information about data structure, coding, schema and language was collected and summarised. We have also aggregated information about omics data across partners

[17] https://www.cineca-project.eu
[18] CINECA Deliverable 3.2 Semantic and harmonisation best practice, https://doi.org/10.5281/zenodo.5055308
[19] https://ihccglobal.org/
[20] https://atlas.ihccglobal.org/
[21] https://github.com/IHCC-cohorts/GECKO, https://www.ebi.ac.uk/ols/ontologies/gecko

together with categories of phenotypic information available. This information provides important input for future analyses.

For instance, as can be seen From Table 3, objective information such as age at recruitment and sex is available in almost all biobanks - this means that these variables can be included in the risk models in a unified way. Similarly, smoking and education related information is available in some format in a majority of biobanks, whereas nutrition data, physical activity and reproductive health related information, sleeping habits and prescribed drugs are only available at some biobanks, making a widely comprehensive analysis using this type of information difficult; however, this information can be used in some smaller scale analyses involving only selected biobanks. Regarding omics data other than genetic data, the sample sizes are small across partners which limits use of non-genomic omics data in future large-scale analyses.

In methodological aspects, firstly, as exact dates for diagnoses (ICD-10 codes) are available in all biobanks, in addition to focusing on logistic regression models, we can additionally fit survival types of models, where age of diagnosis is used as a time scale. This allows us to fit absolute risk models while accounting for the length of follow-up which can be very different for individuals even in the same biobank given that recruitment can be done over time spanning several decades. It also allows us to compose absolute risk models using risk factor data collected at recruitment and estimate their predictive ability over time. Secondly, almost all biobanks report that downloading their data is not possible and analysis must be done locally - this means that building a warehouse which could host individual level data from partners is not feasible and analysis will be performed in-house using shared scripts and pipelines.

Based on our survey, we now know how the medical history is coded in different biobanks. Since ICD-9/10 codes are available in all participating cohorts, a clinical endpoint definition library has been selected as the basis to define a core set of endpoints in the future. The library of definitions is based on the ICD coding system, and therefore allows for a relatively simple way of harmonizing the data across different biobanks. However, ICD coding systems might not be available in all potential biobanks interested in joining our efforts later in the future. Therefore, a comprehensive plan to map between disease definitions has been developed.

The partners first considered using OMOP Common Data Model which allows for transforming data contained within several registries or databases into a common format and then perform systematic analyses using a library of standard analytic routines. However, OMOP was not available during our survey in most biobanks and therefore we chose the FinnGen clinical endpoint definition library instead, which could be directly applied with minimal effort in all participating biobanks.

It is not clear how similarly the harmonized endpoints will behave in polygenic risk score related analyses if different coding systems are used as the basis for definitions. This will require some investigation, for instance, we could look into possible heterogeneity of endpoint-PRS associations or compare pairwise genetic correlations calculated using genome wide association study results from all partners. This would allow us to characterise how well the mapping between different coding systems works. It is also possible that even though mapping tools exist, due to differences in guidelines of disease diagnosis or other factors we might not be able to account for, harmonization for all endpoints in an acceptable way might not be possible.

In conclusion, in this deliverable we mapped the data sources and standards of the biorepositories participating in INTERVENE together with ethical and legal issues related to accessing the data. Based on

the collected data, we developed a data harmonization strategy for the INTERVENE project which will provide input for future Tasks and Deliverables.

5. Appendix

Appendix 1. INTERVENE data transfer survey (general information and ethics)

Appendix 2. INTERVENE data transfer survey (data description)

Supplementary Table 1. Experts working on FInnGen clinical endpoints:

Neurology Group

Reetta Kälviäinen	Northern Savo Hospital District, Kuopio, Finland
Valtteri Julkunen	Northern Savo Hospital District, Kuopio, Finland
Hilkka Soininen	Northern Savo Hospital District, Kuopio, Finland
Anne Remes	Northern Ostrobothnia Hospital District, Oulu, Finland
Mikko Hiltunen	Northern Savo Hospital District, Kuopio, Finland
Jukka Peltola	Pirkanmaa Hospital District, Tampere, Finland
Minna Raivio	Hospital District of Helsinki and Uusimaa, Helsinki, Finland
Pentti Tienari	Hospital District of Helsinki and Uusimaa, Helsinki, Finland
Juha Rinne	Hospital District of Southwest Finland, Turku, Finland
Roosa Kallionpää	Hospital District of Southwest Finland, Turku, Finland
Juulia Partanen	Institute for Molecular Medicine Finland, HiLIFE, University of Helsinki, Finland

Gastroenterology Group

Martti Färkkilä	Hospital District of Helsinki and Uusimaa, Helsinki, Finland
Jukka Koskela	Hospital District of Helsinki and Uusimaa, Helsinki, Finland
Sampsa Pikkarainen	Hospital District of Helsinki and Uusimaa, Helsinki, Finland
Airi Jussila	Pirkanmaa Hospital District, Tampere, Finland
Katri Kaukinen	Pirkanmaa Hospital District, Tampere, Finland
Timo Blomster	Northern Ostrobothnia Hospital District, Oulu, Finland

Mikko Kiviniemi	Northern Savo Hospital District, Kuopio, Finland
Markku Voutilainen	Hospital District of Southwest Finland, Turku, Finland
Mark Daly	Institute for Molecular Medicine Finland, HiLIFE, University of Helsinki, Finland

Rheumatology Group

Kari Eklund	Hospital District of Helsinki and Uusimaa, Helsinki, Finland
Antti Palomäki	Hospital District of Southwest Finland, Turku, Finland
Pia Isomäki	Pirkanmaa Hospital District, Tampere, Finland
Laura Pirilä	Hospital District of Southwest Finland, Turku, Finland
Oili Kaipiainen-Seppänen	Northern Savo Hospital District, Kuopio, Finland
Johanna Huhtakangas	Northern Ostrobothnia Hospital District, Oulu, Finland
Nina Mars	Institute for Molecular Medicine Finland, HiLIFE, Helsinki, Finland

Pulmonology Group

Tarja Laitinen	Pirkanmaa Hospital District, Tampere, Finland
Margit Pelkonen	Northern Savo Hospital District, Kuopio, Finland
Paula Kauppi	Hospital District of Helsinki and Uusimaa, Helsinki, Finland
Terttu Harju	Northern Ostrobothnia Hospital District, Oulu, Finland
Hannu Kankaanranta	University of Gothenburg, Gothenburg, Sweden/ Seinäjoki Central Hospital, Seinäjoki, Finland/ Tampere University, Tampere, Finland
Riitta Lahesmaa	Hospital District of Southwest Finland, Turku, Finland

Cardiometabolic Diseases Group

Teemu Niiranen	Hospital District of Southwest Finland Turku, Finland
Felix Vaura	Hospital District of Southwest Finland Turku, Finland
Veikko Salomaa	The Finnish Institute of Health and Welfare Helsinki, Finland
Kaj Metsärinne	Hospital District of Southwest Finland, Turku, Finland
Mika Kähönen	Pirkanmaa Hospital District, Tampere, Finland

Daniel Gordin	Hospital District of Helsinki and Uusimaa, Helsinki, Finland
Juha Sinisalo	Hospital District of Helsinki and Uusimaa, Helsinki, Finland
Marja-Riitta Taskinen	Hospital District of Helsinki and Uusimaa, Helsinki, Finland
Tiinamaija Tuomi	Hospital District of Helsinki and Uusimaa, Helsinki, Finland
Jari Laukkanen	Central Finland Health Care District, Jyväskylä, Finland
Timo Hiltunen	Hospital District of Helsinki and Uusimaa, Helsinki, Finland
Jussi Hernesniemi	Pirkanmaa Hospital District, Tampere, Finland
Jenni Aittokallio	Hospital District of Southwest Finland, Turku, Finland
Amanda Elliott	Institute for Molecular Medicine Finland, HiLIFE, University of Helsinki, Finland / Broad Institute, Cambridge, MA, United States
Mary Pat Reeve	Institute for Molecular Medicine Finland, HiLIFE, University of Helsinki, Finland
Sanni Ruotsalainen	Institute for Molecular Medicine Finland, HiLIFE, University of Helsinki, Finland

Oncology Group

Tuomo Meretoja	Hospital District of Helsinki and Uusimaa, Helsinki, Finland
Heikki Joensuu	Hospital District of Helsinki and Uusimaa, Helsinki, Finland
Olli Carpén	Hospital District of Helsinki and Uusimaa, Helsinki, Finland
Johanna Mattson	Hospital District of Helsinki and Uusimaa, Helsinki, Finland
Eveliina Salminen	Hospital District of Helsinki and Uusimaa, Helsinki, Finland
Annika Auranen	Pirkanmaa Hospital District , Tampere, Finland
Peeter Karihtala	Northern Ostrobothnia Hospital District, Oulu, Finland
Päivi Auvinen	Northern Savo Hospital District, Kuopio, Finland
Klaus Elenius	Hospital District of Southwest Finland, Turku, Finland
Johanna Schleutker	Hospital District of Southwest Finland, Turku, Finland
Esa Pitkänen	Institute for Molecular Medicine Finland, HiLIFE, University of Helsinki, Finland
Nina Mars	Institute for Molecular Medicine Finland, HiLIFE, University of Helsinki, Finland
Mark Daly	Institute for Molecular Medicine Finland, HiLIFE, University of Helsinki, Finland

Opthalmology Group

Kai Kaarniranta	Northern Savo Hospital District, Kuopio, Finland
Joni A Turunen	Hospital District of Helsinki and Uusimaa, Helsinki, Finland
Terhi Ollila	Hospital District of Helsinki and Uusimaa, Helsinki, Finland
Hannu Uusitalo	Pirkanmaa Hospital District, Tampere, Finland
Juha Karjalainen	Institute for Molecular Medicine Finland, HiLIFE, University of Helsinki, Finland
Esa Pitkänen	Institute for Molecular Medicine Finland, HiLIFE, University of Helsinki, Finland

Dermatology Group

Kaisa Tasanen	Northern Ostrobothnia Hospital District, Oulu, Finland
Laura Huilaja	Northern Ostrobothnia Hospital District, Oulu, Finland
Katariina Hannula-Jouppi	Hospital District of Helsinki and Uusimaa, Helsinki, Finland
Teea Salmi	Pirkanmaa Hospital District, Tampere, Finland
Sirkku Peltonen	Hospital District of Southwest Finland, Turku, Finland
Leena Koulu	Hospital District of Southwest Finland, Turku, Finland

Odontology Group

Pirkko Pussinen	Hospital District of Helsinki and Uusimaa, Helsinki, Finland
Aino Salminen	Hospital District of Helsinki and Uusimaa, Helsinki, Finland
Tuula Salo	Hospital District of Helsinki and Uusimaa, Helsinki, Finland
David Rice	Hospital District of Helsinki and Uusimaa, Helsinki, Finland
Pekka Nieminen	Hospital District of Helsinki and Uusimaa, Helsinki, Finland
Ulla Palotie	Hospital District of Helsinki and Uusimaa, Helsinki, Finland
Maria Siponen	Northern Savo Hospital District, Kuopio, Finland
Liisa Suominen	Northern Savo Hospital District, Kuopio, Finland
Päivi Mäntylä	Northern Savo Hospital District, Kuopio, Finland
Ulvi Gursoy	Hospital District of Southwest Finland, Turku, Finland

Vuokko Anttonen	Northern Ostrobothnia Hospital District, Oulu, Finland
Kirsi Sipilä	Northern Ostrobothnia Hospital District, Oulu, Finland

Women's Health and Reproduction Group

Hannele Laivuori	Institute for Molecular Medicine Finland, HiLIFE, University of Helsinki, Finland
Venla Kurra	Pirkanmaa Hospital District, Tampere, Finland
Laura Kotaniemi-Talonen	Pirkanmaa Hospital District, Tampere, Finland
Oskari Heikinheimo	Hospital District of Helsinki and Uusimaa, Helsinki, Finland
Ilkka Kalliala	Hospital District of Helsinki and Uusimaa, Helsinki, Finland
Lauri Aaltonen	Hospital District of Helsinki and Uusimaa, Helsinki, Finland
Varpu Jokimaa	Hospital District of Southwest Finland, Turku, Finland
Johannes Kettunen	Northern Ostrobothnia Hospital District, Oulu, Finland
Marja Vääräsmäki	Northern Ostrobothnia Hospital District, Oulu, Finland
Outi Uimari	Northern Ostrobothnia Hospital District, Oulu, Finland
Laure Morin-Papunen	Northern Ostrobothnia Hospital District, Oulu, Finland
Maarit Niinimäki	Northern Ostrobothnia Hospital District, Oulu, Finland
Terhi Piltonen	Northern Ostrobothnia Hospital District, Oulu, Finland
Katja Kivinen	Institute for Molecular Medicine Finland, HiLIFE, University of Helsinki, Finland
Elisabeth Widen	Institute for Molecular Medicine Finland, HiLIFE, University of Helsinki, Finland
Taru Tukiainen	Institute for Molecular Medicine Finland, HiLIFE, University of Helsinki, Finland
Mary Pat Reeve	Institute for Molecular Medicine Finland, HiLIFE, University of Helsinki, Finland
Mark Daly	Institute for Molecular Medicine Finland, HiLIFE, University of Helsinki, Finland
Niko Välimäki	University of Helsinki, Helsinki, Finland
Eija Laakkonen	University of Jyväskylä, Jyväskylä, Finland
Jaakko Tyrmi	University of Oulu, Oulu, Finland / University of Tampere, Tampere, Finland
Heidi Silven	University of Oulu, Oulu, Finland
Eeva Slitz	University of Oulu, Oulu, Finland
Riikka Arffman	University of Oulu, Oulu, Finland

Susanna Savukoski	University of Oulu, Oulu, Finland
Triin Laisk	Estonian biobank, Tartu, Estonia
Natalia Pujol	Estonian biobank, Tartu, Estonia

Depression Group

Iiris Hovatta	University of Helsinki, Helsinki, Finland
Chia-Yen Chen	Biogen, Cambridge, MA, United States
Erkki Isometsä	Hospital District of Helsinki and Uusimaa, Helsinki, Finland
Kumar Veerapen	Broad Institute, Cambridge, MA, United States
Hanna Ollila	Institute for Molecular Medicine Finland, HiLIFE, University of Helsinki, Finland
Jaana Suvisaari	National institute of health and welfare (THL)
Thomas Damm Als	Aarhus University, Denmark

ENT (ear, nose and throath) Group

Antti Mäkitie	Hospital District of Helsinki and Uusimaa, Helsinki, Finland
Argyro Bizaki-Vallaskangas	Pirkanmaa Hospital District, Tampere, Finland
Sanna Toppila-Salmi	University of Helsinki, Helsinki, Finland
Tytti Willberg	Hospital District of Southwest Finland, Turku, Finland
Elmo Saarentaus	Institute for Molecular Medicine Finland, HiLIFE, University of Helsinki, Finland
Antti Aarnisalo	Hospital District of Helsinki and Uusimaa, Helsinki, Finland
Eveliina Salminen	Hospital District of Helsinki and Uusimaa, Helsinki, Finland
Elisa Rahikkala	Northern Ostrobothnia Hospital District, Oulu, Finland
Johannes Kettunen	Northern Ostrobothnia Hospital District, Oulu, Finland

Intervene data transfer survey (general information and ethics)

This survey aims to get an overview of the technical and legal aspects of the datasets to be used within the Intervene project. Please, answer one survey for each dataset at your institution to be used within the Intervene project.

* Required

1. Email address *

Instructions

Please, answer one survey for each dataset at your institution to be used within the Intervene project.
After responding, you will be sent a link which allows you to edit your response.

General information

2. Name

Enter your name here

3. Partner Center

Enter the name of your partner biobank/center here

Legal aspects of the data

4. Is a special permission or application form required to access/download the data?

   Mark only one oval.

   - ◯ Yes
   - ◯ No

5. Please describe in short the proccess of accessing your data. Also, indicate if some data is more restricted and needs any kind of special permission.

6. Is your data already anonymized?

   Mark only one oval.

   - ◯ Yes
   - ◯ No

7. Does any data location related jurisdiction apply?

   Mark only one oval.

   - ◯ Yes
   - ◯ No

8. Please explain details in the following field:

9. Can you enumerate any sensitive data elements? E.g. Full name, date of birth, etc.

Intervene data transfer survey (data description)

This survey aims to get an overview of the technical and legal aspects of the datasets to be used within the Intervene project. Please, answer one survey for each dataset at your institution to be used within the Intervene project.

* Required

1. Email address *

Instructions

Please, answer one survey for each dataset at your institution to be used within the Intervene project.
After responding, you will be sent a link which allows you to edit your response.

General information

2. Name

Enter your name here

3. Partner Center

Enter the name of your partner biobank/center here

Technical aspects

4. In what format do you have your phenotype data? Please select the options that apply.

Check all that apply.

- [ ] Excel (or other similar spredsheet-like formats)
- [ ] JSON file
- [ ] XML file
- [ ] SQL Database (MySQL, PostgreSQL, Oracle etc.)
- [ ] Other: _____

5. What is the degree of structuring of your data? Please select the options that apply.

Check all that apply.

- [ ] Structured
- [ ] Unstructured (free text)
- [ ] Semi-structured
- [ ] Images
- [ ] Other: _____

6. Is your data following OPOM data model?
https://www.ohdsi.org/data-standardization/the-common-data-model/

Mark only one oval.

- ◯ Yes
- ◯ No

7. If you are not following OPOM yet, if and when are you planning to implement it?

_____

8. Is your data already coded somehow?

Mark only one oval.

- Yes
- No

9. Please describe the coding system for your data.

10. Do you have a description of the data structure (i.e. data schema)?

Mark only one oval.

- Yes
- No

11. What is the data language used?

English, French, German, etc

12. What type of sources are aggregated in your data? Please select the options that apply.

Check all that apply.

- [ ] Answerset obtained at recruitment
- [ ] Registry information such as Cancer, Causes of Death, etc
- [ ] National databases, health services

Other: [ ] _____

13. Are sources of data updated regularly?

Mark only one oval.

- ( ) Yes
- ( ) No

14. Openness for sharing data with the Intervene consortium

Check all that apply.

- [ ] Full access
- [ ] Federated access
- [ ] Analysis done internally using provided scripts

15. How is the meta-data (e.g., data schema) and individual level data accessed and in which format and structure could (if possible) these be downloaded?

_____

_____

_____

_____

_____

16. What is approximately the size of the data?

500 GB, 1 TB, etc

Detailed information about the data types available

17. Do you have the following omics data?

Please select the options that apply

Check all that apply.

- [ ] Metabolome data
- [ ] Proteomics data
- [ ] Microbiome data
- [ ] Transcriptome data
- [ ] Methylation data
- [ ] Chromatin data

Please describe in detail (for how many individuals, in what format/structure is the data, how is it generated/measured)

Please provide a more detailed description for each omics data set you have.

18. Metabolome data

19. Proteomics data

20. Microbiome data

21. Transcriptome data

22. Methylation data

23. Chromatin data

24. Genomics data

Phenotype infomation

25. What types of phenotype data do you have? Please select the options that apply.

Check all that apply.

- [ ] Objective information (such as date of recruitment, weight, height, age at recruitment, sex, etc)
- [ ] Education and/or work related questions
- [ ] Questions about smoking habits
- [ ] Questions about alchol consumption
- [ ] Nutrition data (about frequency of different types of food intake)
- [ ] Reproductive health (such as number of children, for women - age of menarhe/menopause, number of pregnancies, etc)
- [ ] Physical activity
- [ ] Sleeping habits
- [ ] Family history of the diseases
- [ ] Medical history of the individuals
- [ ] Pharmacological data

26. Do you have exact dates of recruitment?

Mark only one oval.

- Yes
- No
- Other: _____

27. Do you have individuals medical history coded with ICD-10 codes?

Mark only one oval.

- Yes
- No
- Other: _____

28. Do you have exact dates of these ICD-10 codes?

Mark only one oval.

- Yes
- No
- Other: _____

29. Do you have individuals history of prescribed drugs?

Mark only one oval.

- Yes
- No
- Other: _____

30. Do you have exact dates of these prescriptions?

Mark only one oval.

- Yes
- No
- Other: _____

31. Do you have individuals history of adverse effects of prescribed drugs?

Mark only one oval.

- Yes
- No
- Other: _____

32. Which resources are used for linking with medical history (for diseases, prescription drugs, causes of death, different measurements such as blood/urine analyses, lipids, BMI, blood pressure etc) in your cohort and how often is the linkage carried out? Which information is linked? (ICD 10- codes, prescription drugs, dates of prescriptions/diagnosis, causes of death etc?)

_____

_____

_____

_____

_____