# INTERVENE

## Data Coordination Center

Deliverable 1.1.

Dissemination level: Public

Part of:

WP 1:
Compliant and standardised data access for federated analyses

| Project summary | |
|---|---|
| **Project acronym:** | INTERVENE |
| **Project full title:** | International consortium for integrative genomics prediction |
| **Project Coordinator** | Institute for Molecular Medicine Finland FIMM, University of Helsinki; Prof. Samuli Ripatti and Dr. Andrea Ganna |
| **Project start date:** | 1.1.2021 |
| **Project end date:** | 31.12.2025 |
| **Project duration:** | 60 months |
| **Action type:** | RIA |
| **Call identifier:** | H2020-SC1-FA-DTS-2018-2020 (Trusted digital solutions and Cybersecurity in Health and Care) |
| **Grant number** | 101016775 |
| **Document descriptors** | |
| **Deliverable No.** | 1.1 |
| **Work package** | WP1 |
| **Deliverable lead** | CSC |
| **Contributors** | CSC, EMBL, UTARTU, HUS |
| **Dissemination level** | Public |
| **Expected delivery date** | 30/06/2021 |
| **Submission date** | 30/06/2021 |

**Change history log**

| Version | Changes made | Date | Prepared by | PC approved |
|---|---|---|---|---|
| **0.1** | **First draft** | **09/06/2021** | **CSC (K. Mattila)** | |
| **1.0** | **Revision after coordinator's review** | **28/06/2021** | **CSC (Lappalainen)** | |

## Project description

INTERVENE is an international and interdisciplinary consortium that seeks 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. The consortium consists of leading research and other organizations representing EU member states as well as Norway, the United Kingdom and the USA. As part of INTERVENE, some of the largest biobanks in Europe and beyond will be securely linked and harmonized in a GDPR-compliant repository with data from more than 1.7 million genomes. The purpose is to leverage the these vast, but still underused data resources to generate clinically actionable knowledge for improved understanding of diseases and treatment options tailored to individuals.

This deliverable is part of Work Package (WP) 1 of INTERVENE, focusing on building the technical infrastructure to support federated analysis based on AI algorithms for the personalised medicine use cases. It describes the establishment of a Data Coordination Center (DCC) that aligns project efforts with the global data standards and supports INTERVENE consortium member access to the data. In this work package we aim to create a data analysis environment that hosts the consortium data collected from over 1.7 million individuals with associated genetic data. It will also provide computational resources for validating new AI algorithms developed by the INTERVENE partners and the broader AI community.

## Data Coordination Center

The key task of INTERVENE Work Package 1 (WP1- Compliant and standardised data access for federated analyses) is to establish and maintain an INTERVENE Data Coordination Center (DCC). The DCC is a virtual service unit to support project partners on data interoperability, best practices on data management and general coordination across the project partners.

The establishment and operation of the DCC is the responsibility of CSC – IT center of Science. CSC is a company owned by the Finnish state and higher education institutions. It provides various information technology and data management services for research, including human genomes (and their phenotypic information) consented for research. CSC also includes a national ELIXIR Node providing a bridge to more generic FAIR data management solutions in collaboration with ELIXIR Europe and other relevant research infrastructures (such as BBMRI). EMBL-EBI provides additional data management competences and some of the largest data resources relevant for the INTERVENE project (such the EGA, ENA and Biosamples). Helsinki University Hospital (HUS) has in-depth experience in managing clinical information and processes supporting consent-based research. University of Tartu (UTARTU) is linked to the Estonian biobank and already supports large-scale electronic data management on biobank samples. The DCC is based on the existing experience and synergistic competences of all these partnering organisations allowing DCC to develop and grow during the INTERVENE project.

The INTERVENE DCC is established as part of the CSC sensitive data services outreach office as a one-stop-shop to coordinate project member requests. While the DCC does not include technical staff members it will coordinate the answer, guidelines and best practices across the project partners. For example, the DCC will support WP2 on the methods, standards and policies that will be established for the scientific use cases. It will notify project partners when new versions of the consortium data are made available (WP1), it will provide best practices to support computation (WP3) on these data (by for example providing guidelines on transporting algorithms within containers that run on different environments) and it will also provide support between WP1 and WP4 for enabling collaboration between IGS4EU platform and the main INTERVENE data analysis environment.

In practical level, the recommended standards will be described in the data management plan (DMP) and user documentation. DCC will also support the revision of the DMP compliant with the required level of data security and privacy that is aligned with the recent recommendations of the OECD Council onHealth Data Governance and ELIXIR's ELSI policy.

## Current status of DCC

### Resources

The DCC was established in May 2021. At the beginning DCC is using CSC resources but as the use cases evolve, interoperability between other computing services and biobanks will have to be built to ensure capacity and access to data that cannot be centralised. At the moment DCC provides INTERVENE project members access to following CSC services:

1. CSC Sensitive data services: SD Connect and SD Desktop
2. Cloud resources: c Pouta and Rahti
3. High performance computing environment: Puhti
4. Object storage service: Allas

INTERVENE members will have access to these resources by contacting the DCC and requesting resources to support project specific development, testing and data analysis.

When required the DCC will support researchers on designing and deploying their software tools to be run in the secure data analysis environment. The support will focus on providing documentation together with other WPs as the scientific use cases and access models become clearer. All the documentation will be collected to members' area in the INTERVENE web site.

## Support

The DCC has a support e-mail address (intervene-dcc@csc.fi) that is linked to the CSC ticketing system allowing tracking and supporting e.g. audition processes. The ticketing system is actively monitored by the INTERVENE DCC support team. Smaller issues, such as technical support requests, will be processed inside the ticketing system, while larger issues that require planning and collaboration can be moved to a Jira based task management system.

At the moment the DDC support team includes:

- Francesca Morello (CSC)
- Kimmo Mattila (CSC)
- Aoife McMahon (EMBL)
- Reedik Mägi (UTARTU)
- Minttu Marttila (HUS)

## Policies, standards and procedures

In addition to providing technical solutions, DCC also makes sure that the tools and methods developed and used in INTERVENE as well as the data sharing and distributing procedures will be kept compatible with Federated EGA and INTERVENE related projects like CINECA and B1MG.