

The INCLUDE Data Hub: A collaborative effort by the INCLUDE Data Coordinating Center to enable data sharing and integrative analyses for Down syndrome research

Authors:

A. Resnick¹, J. Espinosa⁵, B. O'Connor⁴, J. DiGiovanna³, V. Ferretti², R. Carroll⁶, M. Haendel⁷, INCLUDE Data Coordinating Center;

1 Center. for Data Driven Discovery in Biomedicine, Children's Hosp. of Philadelphia, Philadelphia, PA

2 CHU Sainte-Justine Res. Ctr., Montreal, QC, Canada

3 Seven Bridges, Cambridge, MA

4 Sage Bionetworks, Seattle, WA

5 Linda Crnic Institute for Down Syndrome, University of Colorado Anschutz Medical Campus, Denver, CO

6 Department of Biomedical Informatics, Vanderbilt University Medical Center, Nashville, TN

7 Department of Biomedical Informatics, University of Colorado Anschutz Medical Campus, Denver, CO

Over the last few decades, the life expectancy for people with Down syndrome (DS) has significantly increased. Recent studies have shown that trisomy 21 (T21) causes a different disease spectrum across individuals, with a range of comorbidities and severities. To advance research in this area, the NIH launched the INCLUDE (INvestigation of Co-occurring conditions across the Lifespan to Understand Down syndromE) project. The INCLUDE project aims to assemble data from a large cohort of people with DS to accelerate research into mechanisms driving and modulating DS comorbidities.

Here we present an overview of the INCLUDE Data Coordinating Center (DCC). The DCC is composed of 3 cores led by experts around the world. The Administrative and Outreach Core (AOC) provides integrated administration, management, communications, and outreach to fulfill the mission of the DCC. The AOC leads diverse outreach activities to ensure significant contributions from a wide range of data-generating sources, widespread and effective use of the Data Hub, and connectivity with other data ecosystems. The Data Management Core (DMC) brings together experts in open science, data sharing, curation and harmonization. This group coordinates the ingest of research data from investigators, including clinical, genomic, transcriptomic, metabolomic, and other types of biological data. The DMC works to annotate, harmonize, and process data according to agreed-upon data models and workflows to enable interoperability and reusability of data by

researchers. The Data Portal Core (DPC) focuses on the development of the infrastructure and user interface surrounding the INCLUDE Data Hub, the project's outward-facing tool, to serve a diverse group of stakeholders in the DS community. Our datasets are standardized through a FHIR model to enable interoperability across multiple NIH cloud-based repositories. The Data Hub provides intuitive virtual cohort building of the datasets processed by the DMC. Through integration with the CAVATICA cloud-based analysis platform, users can also perform in-depth bioinformatic analyses.

Through the collaborative work of the DCC, we are creating a world-class resource for data sharing, access, and integrative analysis that would enable novel investigations into all DS comorbidities across the lifespan using diverse data modalities. Furthermore, this resource will empower physicians, educators, government officials, and advocates with the tools required to elicit evidence-based transformative action in the clinic, the classroom, the government, and society at large.