National Child Development Study

Proteomics

User Guide (Version 1)

March 2024

CENTRE FOR LONGITUDINAL STUDIES

Economic and Social Research Council

Contact

Data queries: help@ukdataservice.ac.uk

Questions and feedback about this user guide: clsdata@ucl.ac.uk.

Authors

Matt Brown, Dylan Williams, George Ploubidis, Vilma Agalioti-Sgompou, Aida Sanchez-Galvez, Nureen Hanisah Mohamad Zaki.

How to cite this guide

Matt Brown, Dylan Williams, George Ploubidis, Vilma Agalioti-Sgompou, Aida Sanchez-Galvez, Nureen Hanisah Mohamad Zaki (2024) *National Child Development Study: Proteomics User Guide (Version 1)*. UCL Centre for Longitudinal Studies

You should also acknowledge CLS following the guidance from https://cls.ucl.ac.uk/data-access-training/citing-our-data/

This guide was published in March 2024 by the Centre for Longitudinal Studies (CLS).

The Centre for Longitudinal Studies (CLS)

Centre for Longitudinal Studies

UCL Social Research Institute

University College London

20 Bedford Way, London WC1H 0AL

The UCL Centre for Longitudinal Studies (CLS) is an Economic and Social Research Council (ESRC) Resource Centre based at the UCL Social Research Institute, University College London. It manages four internationally renowned cohort studies: the 1958 National Child Development Study, the 1970 British Cohort Study, Next Steps, and the Millennium Cohort Study. For more information, visit www.cls.ucl.ac.uk.

This document is available in alternative formats. Please contact the Centre for Longitudinal Studies: tel: +44 (0)20 7612 6875
email: clsdata@ucl.ac.uk

Contents

About the National Child Development Study

The National Child Development Study (NCDS) is a longitudinal birth cohort study, following a nationally representative sample of over 17,000 people born in Britain in a single week in March 1958.

Cohort members have been surveyed throughout their lives, since birth, creating an incredibly rich resource for a wide range of research. The study data show the very long roots of childhood, how past experiences can reverberate through the years, and the interplay between the different facets of people's lives.

NCDS has equipped policymakers with robust evidence in areas as diverse as smoking in pregnancy, educational inequalities, adult basic skills, and social mobility. Today, with the cohort now in their sixties, the study is casting light on how people experience retirement and ageing in the 21st century.

1. Introduction

A biomedical sweep of NCDS was conducted in 2002/3 when participants were aged 44/45. It sought to obtain objective measures of ill-health and biomedical risk factors.

The broad aims were to explore the impact of developmental, environmental and lifestyle factors on ill-health, and physiological and psychological function among adults in early middle age; to investigate the effect of such factors on geographical and socioeconomic health inequalities; and to make possible comparisons between these adults in early middle age and members of the MRC funded 1946 birth cohort at the same age.

The research was also intended to address a wide range of specific hypotheses relating to anthropometry, cardiovascular, respiratory and allergic diseases, visual and hearing impairment, and mental ill-health.

Non-fasting blood samples were collected and (with consent) immortalised cell lines were created, and DNA extracted and stored for medical research purposes. A wide variety of assays of the blood were conducted. Blood samples continue to be stored in the CLS Human Tissue Bank at University of Bristol. Applications to make use of

the stored samples for research can be made via the CLS Data Access Committee (https://cls.ucl.ac.uk/data-access-training/data-access/accessing-data-directly-from-cls/).

In 2021, funding was provided by the Medical Research Council to conduct proteomics analyses to be conducted on the blood samples collected in the 2002 biomedical study. Proteomics is the study of the proteins found in blood and other biological samples and how these proteins can affect health and disease.

2. Project description

2.1 Plasma samples

A total of 9,377 study members participated in the NCDS biomedical study in 2002/3. All participants were aged 44/45 at the time of participation. 8,753 consented to provide a blood sample and 8,509 consented that these blood samples were stored for future analysis.

The proteomics analysis was conducted by Olink (https://olink.com/) using samples of EDTA plasma. The EDTA tubes were collected by nurses in the homes of participants and then posted to the laboratory where they were centrifuged on receipt and have subsequently been stored in 0.5ml aliquots at -80C. EDTA plasma samples were available for 8,005 participants.

One aliquot per participant was shipped to Olink from the University of Bristol.

2.2. Proteomics panels

Olink offers a variety of 'panels' to measure different proteins. Decisions regarding which panels to use were taken in consultation with the Systematic and Combined AnaLysis of Olink Proteins (SCALLOP) Consortium. Decisions were taken with the aim of maximising utility to the scientific community.

The chosen panels were Target 96 Organ Damage Panel and Target 96 CVD II Panel for all samples, and the Explore 3072 Panel for a subset of samples.

All 8005 NCDS samples were analysed using the Target 96 Organ Damage Panel (https://olink.com/products-services/target/cardiometabolic-panel/), and the Target

96 CVD II Panel (https://olink.com/products-services/target/biological-process/). Each of these panels measures 96 proteins.

In addition, a sub-set of 176 samples were analysed using the Explore 3072 Panel, which measures 3,072 proteins (https://olink.com/products-services/explore/). This is a combination of 8 separate Explore 384 panels: Inflammation, Oncology, Cardiometabolic, Neurology, Inflammation II, Oncology II, Cardiometabolic II and Neurology II. The 176 individuals on whose blood this additional analysis was conducted were randomly selected from a sub-sample of >1,000 participants where DNA methylation analysis had previously been conducted (https://cls-genetics.github.io/docs/NCDS.html).

The results of the analyses are provided in the data in Olink's Normalized Protein Expression (NPX) units, quantifying the relative concentration of a specific protein on a log2 scale. NPX data allow users to identify differences in the levels of the same protein across the set of samples analysed in the same run, but cannot be used to compare concentrations of different proteins of the same panel, nor to compare values of a given protein to NPX values in other datasets produced in separate experiments, i.e.from other studies (https://olink.com/faq/what-is-npx/).

2.3 Quality Control

Olink aliquoted, plated and performed the assays, following some sample management controls:

- To reduce the possibility of plate effects Olink identified 21 'super reference samples' with a large volume (minimum of 120 µL). These were repeated 3 times across the entire sample set to reference normalize the analyses if needed.
- Olink randomly selected 10 plates and re-randomised the samples into these plates, adding required reference samples according to the final plate manifests. This was repeated in 10 plate batches until the project was completed.
- Olink processed the data, performed their standard Quality Control analysis, calculated NPX values, and produced a Certificate of Analysis (CoA).

Target 96 Panel

For the Target 96 panels, all assays included built-in quality control based on four internal controls that are spiked into all samples and external controls. The quality control (QC) was performed in two steps:

1. Each sample plate is evaluated on the standard deviation of the internal controls. This should be below 0.2 NPX. Only data from sample plates that pass this QC step are reported.

2. The quality of each sample is assessed by evaluating the deviation from the median value of the controls for each individual sample. Samples that deviate less than 0.3 NPX from the median pass the QC. Data from all samples is included in the data output file. Samples that did not pass the QC are indicated in columns named "QC Warning". Data points from samples that do not pass QC should be treated with caution.

Following is the summary of QC for both Target 96 panels:

Olink panel	**No. of samples that passed QC / Total no. of samples**	**Passed samples (%)**
Target 96 Cardiovascular II	7940 / 8050	99
Target 96 Organ Damage	7909 / 8050	98

Explore 3072 Panel

For the Explore 3072 panel, three internal controls were added to each sample:

Extension Control for the generation of the NPX values, and Incubation control and Amplification control to monitor the quality of assay performance and quality of individual samples (https://olink.com/faq/what-are-internal-controls-and-how-are-they-used-in-the-data-analysis/).

In addition, three external controls are included in each run: Plate Control for data normalization, Sample Control to assess potential variation between runs and plates healthy pooled plasma, and Negative Control to calculate Limit of Detection for each assay and to assess potential contamination of assays. The following parameters are evaluated in the Quality Control (QC):

1. The average matched counts for each sample. To pass QC, there should be at least 500 counts, otherwise the sample receives a QC warning status.

2. The deviation from the median value of the Incubation and Amplification Controls for each individual sample. To pass QC, the deviation should not exceed +/-0.3 NPX for either of the internal controls, otherwise the sample receives a QC warning status.

3. The deviation of the median value of the Negative Controls from a predefined value set for each assay. To pass QC, the deviation of the median of the Negative Controls must be ≤5 standard deviations from the set predefined value, otherwise the assay will receive a warning status.

Several proteins are not included in the final data as the assays did not meet Olink's batch release quality control criteria. These are KNG1 (Explore 384 Inflammation II assay), SMAD1 and ARHGAP25 (Explore 384 Oncology).

Following is the summary of QC for all assays in this panel:

Olink panel	**No. of samples that passed QC / Total no. of samples**	**Passed samples (%)**
Explore 384 Cardiometabolic	159 / 176	90
Explore 384 Cardiometabolic II	157 / 176	89
Explore 384 Inflammation	155 / 176	88
Explore 384 Inflammation II	164 / 176	93
Explore 384 Neurology	145 / 176	82
Explore 384 Neurology II	165 / 176	94
Explore 384 Oncology	146 / 176	83
Explore 384 Oncology II	155 / 176	88

3. Research Data

3.1 Licensing and data access

The NCDS proteomics datasets are available as special safeguarded data from the UK Data Service (UKDS), subject to the UKDS Special Licence.

All users of the data need to be registered with the UKDS. Details of how to do this are available at https://www.ukdataservice.ac.uk/get-data/how-to-access/registration.

The data can be accessed by downloading the UKDS Special Licence application form. Once the form has been reviewed by UKDS and approved by the CLS Data Access Committee the data will be available to download.

3.2 Datasets

Name of the dataset	Content summary
ncds_proteomics_target96_v1.*	NPX protein expression values of 184 proteins for 8005 participants across the Target 96 Organ Damage Panel (96 proteins) and the Target 96 CVD II Panel (96 proteins)
ncds_proteomics_explore_v1.*	NPX protein expression values of 2941 proteins for 176 participants measured using the Explore 3072 Panel

3.3 Data documentation

Full validation data information and the protein assay lists for the Target 96 CVD II Panel, the Target 96 Organ Damage Panel, and the Explore 3072 Panel are accessible via Olink's Document Download Center (https://olink.com/resources-support/document-download-center/).

3.4 Data structure and variable description

Original data prior to curation: The original data was delivered in an Excel spreadsheet with information on the protein name, Uniprot ID, protein Limit of Detection (LoD), Missing Data frequency, NPX value, QC status, Assay status and the Normalization method. The samples were identified with anonymised numeric barcodes.

Super reference samples: in the Target 96 panel, the barcodes from the 21 'super reference samples' were suffixed ('_1', '_2', '_3') as Olink requires unique IDs for

generating results (Section 2.3). As part of the data curation, these suffixed barcodes were split into two variables:

- 'barcode'.
- 'multiple_index' (e.g. 'XYZ_1' becomes 'XYZ' and 'multiple_index' 1).

For normal samples, i.e. those with no repeat measurements, barcodes do not have a suffix so 'multiple_index' is set to 0. All barcodes were then replaced by the cohort member identifier NCDSID.

Limit of Detection (LOD): The LOD is the lowest measurable level of an individual protein, defined as 3 times the standard deviation over background (https://olink.com/faq/what-does-lod-mean/). For target 96 analytes, LOD values for each protein are provided in the proteins' variable labels in the datasets. For Explore data, LOD values are issued in standalone variables.

Normalization: the normalization method used by Olink is Intensity Normalization v.2, which adjusts the data so that the median NPX for a protein on each plate is equal to the overall median (https://olink.com/content/uploads/2023/02/olink-npx-signature-target-96-user-manual.pdf & https://olink.com/faq/data-normalization).

Target 96 Dataset

The final Target 96 dataset contains one row per NCDS participant and per 'multiple_index', with each row containing 184 proteins. This means that if a sample is one of the 21 'super reference samples', then there could be several rows with the same NCDSID.

The order in which variables appear in this dataset is:

Variable Name	Variable Label
NCDSID	NCDS Research ID
multiple_index	Index *if* multiple samples (max:3)
BMP6 (example)	Assay | Uniprot ID (Panel name) [LoD value]
CV2PLATEID	Plate ID (Target 96 Cardiovascular II)
ODPLATEID	Plate ID (Target 96 Organ Damage)
CV2QCWARN	QC Warning (Target 96 Cardiovascular II)
ODQCWARN	QC Warning (Target 96 Organ Damage)

The variable names such as ‘BMP6’ refer to individual proteins as described in the original file olink-cvd-ii-validation-data-v2.1.pdf and olink-organ-damage-validation-data-v2.0.pdf. Where applicable, the ‘-’ from a protein symbol is removed from the variable name.

Two proteins, PGF and KIM1, are suffixed with either ‘_OD’ or ‘_CV2’ to indicate that the proteins are from the Target 96 Organ Damage Panel or the Target 96 CVD II Panel, respectively. The variable label for each protein is concatenated using information as described in the original file from Olink.

Value labels are as provided by Olink.

Explore 3072 dataset

The final Explore 3072 dataset is in a long format, meaning that there is one row per NCDS participant (NCSID) and protein, each with the NPX value and QC information.

The order in which variables appear in this dataset is:

Variable Name	**Variable Label**
NCDSID	NCDS Research ID
Uniprot	UniProt (EXPLORE)
Assay	Assay (EXPLORE)
MissingFreq	Missing frequency: no. samples on the plate falling below LOD
Panel	Panel (EXPLORE)
PlateID	PlateID (EXPLORE)
QC_Warning	QC warning if internal controls deviate from expected QC levels
LOD	Limit of detection: background
NPX	Value: normalised protein expression
Normalization	Normalization method used to generate the data
Assay_Warning	Assay warning if LOD higher than validation

The variable names and value labels are as provided by Olink in the original data.

3.5 Identifiers

For NCDS, the data are identified with the same research IDs ('NCDSID') used for the rest of cohort data available at the UKDS. This enables the data to be easily merged with one another.

Target 96 Dataset

The flag 'multiple_index' indicates that this was a super reference sample, that is, that there are multiple measurements per NCDSID across all proteins. While the majority of NCDSIDs has only one measurement, a maximum of 3 measurements is possible.

- For NCDSIDs with only one measurement, the 'multiple_index' value label is '0'.
- For NCDSIDs with more than one measurement the multiple_index' values are '1', '2' or '3'.

When merging this dataset to other NCDS datasets, the user should take into account that there are multiple measurements for some samples, so this data structure should be considered as 'many'.

Explore Dataset

The flag 'Assay' indicates the protein measured for each NCDSID. Each NCDSID/Assay combination is a unique row. When merging this dataset to other NCDS datasets, the user should consider this data structure and consider it as 'many'.

3.6 Missing values, data errors and inconsistencies

Target 96 Dataset

Where there are missing values for the protein expression measurements, this is provided as is by Olink in the original data.

Explore Dataset

There are no missing values in this dataset.

The final number of cases in these datasets are subject to the latest data-sharing consents and may vary from the number of samples analysed per dataset in Section 2.