# UK Biobank Pharma Proteomics Project

## Quality control of Olink NPX dataset • November Data Release

**Benjamin B. Sun, Kyle Ferber** and **Tinchi Lin** (Biogen); **Christopher D. Whelan** (Janssen)

## Overview

This document provides a summary of the quality control protocol implemented by the UK Biobank Pharma Proteomics Project (UKB-PPP) for normalized protein expression (.NPX) data generated using the antibody-based Olink Explore™ Proximity Extension Assay (PEA) in blood plasma samples derived from approximately 54,000 UK Biobank participants. This protocol was developed and approved by scientists across the thirteen participating biopharmaceutical companies, including Amgen, Alnylam, AstraZeneca, Biogen, Bristol Myers Squibb, Calico, Genentech, GlaxoSmithKline, Janssen (Johnson & Johnson), Novo Nordisk, Pfizer, Regeneron, and Takeda.

The protocol follows a six-step process, including:

1. Selection of samples by the UKB-PPP consortium
2. Importing data and removing Olink control samples
3. Removing individuals who have withdrawn consent or whose data were not processed
4. Outlier sample detection and removal
5. Removing data with QC warnings or assay warnings
6. Removing likely sample swaps

At the end of each step of this quality control protocol, we provide a simple table summarizing the following attributes of the data, including:

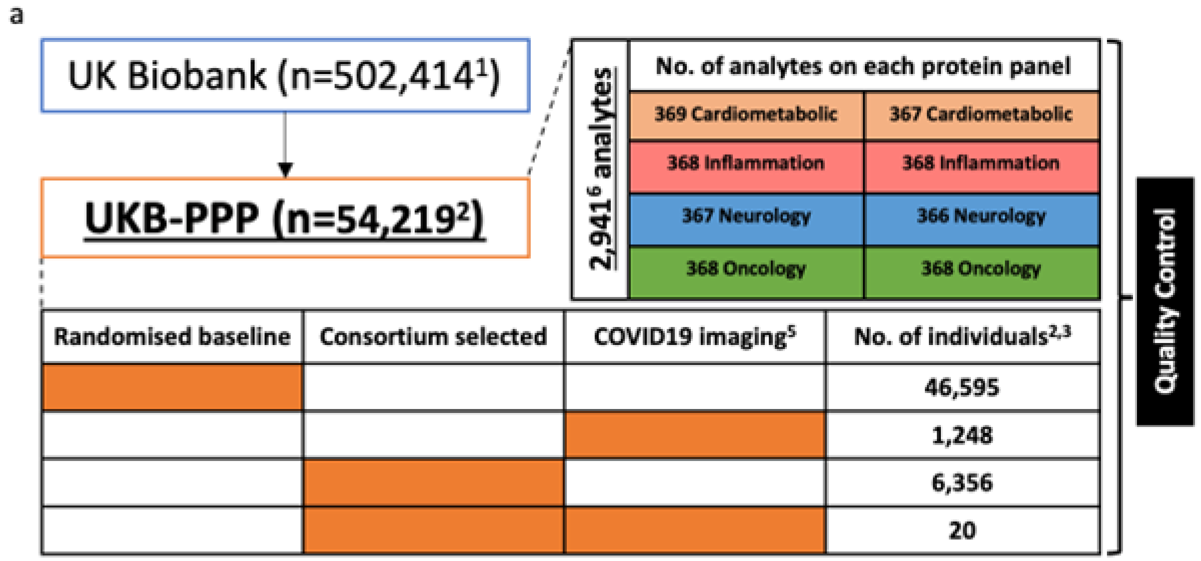
- The number of rows (‘n.row’)
- The number of unique samples (‘n.sample’)
- The number of unique individuals (‘n.ind’)
- The number of unique OlinkIDs, representing unique proteins (‘n.olinkid’)
- The number of rows with missing NPX values (‘n.npx.na’)

We performed all quality control using R 4.2.0 and have included a companion HTML report generated using RMarkdown. The code implemented for each step of this protocol is provided in the HTML report.

# 1. Selection of samples by the UKB-PPP consortium

The UK Biobank Pharma Proteomics Project (UKB-PPP) consortium conducted proteomic profiling on blood plasma samples collected from 54,219 UKB participants using the Olink™ Explore 3072 Proximity Extension Assay (PEA), which measures 2,941 protein analytes capturing 2,923 unique proteins. A full description of the sub-cohort composition is provided in Sun et al. (BioRxiv, 2022: https://www.biorxiv.org/content/10.1101/2022.06.17.496443v1.full).

Briefly, this sub-cohort included a randomized subset of 46,595 UK Biobank participants whose samples were collected at their baseline visit ("randomised baseline"), 1,268 participants who participated in the COVID-19 repeat imaging study at multiple visits, and 6,376 participants pre-selected by the thirteen participating consortium members whose samples were collected at baseline visits ("consortium-selected"; Figure 1, below).



| No. of analytes on each protein panel | |
|---|---|
| 369 Cardiometabolic | 367 Cardiometabolic |
| 368 Inflammation | 368 Inflammation |
| 367 Neurology | 366 Neurology |
| 368 Oncology | 368 Oncology |

| Randomised baseline | Consortium selected | COVID19 imaging$^{5}$ | No. of individuals$^{2,3}$ |
|---|---|---|---|
| 🟧 | | | 46,595 |
| | | 🟧 | 1,248 |
| | 🟧 | | 6,356 |
| | 🟧 | 🟧 | 20 |

**Figure 1**. UKB-PPP sub-cohort composition, adapted from Sun et al., 2022

For the 6,376 pre-selected participants, each consortium member selected baseline blood samples from approximately 500 participants each based on their characteristics of interest, e.g., disease status (Alzheimer's disease cases) or genetic background (e.g., participants of African ancestry). The 13 consortium members provided lists of EIDs for approximately 1,500 participants of their choice. UK Biobank analyses subsequently consolidated these lists, identified and removed duplicate EIDs or EIDs with low sample volumes, and selected blood plasma samples for the top 500 EIDs for each consortium member.

## 2. Importing data and removing Olink control samples

We began by importing the Olink™ Explore 1536 dataset (hereby referred to as the 'PHASE 1 DATASET') and the expanded Olink™ Explore 3072 dataset (hereby referred to as the 'PHASE 2 DATASET') – combining both datasets together and removing control samples provided by Olink. The QC details is concerned with the combined datasets, including all the 2,941 proteins.

Before removing the control samples, the dataset included 2,941 protein measurements from 58,876 samples comprising 54,221 individuals across 672 plates, summarized in Table 1:

**Table 1.** Dataset characteristics before removing control samples

| **n.row** | n.sample | n.ind | n.plate | n.olinkid | n.npx.na |
|---:|---:|---:|---:|---:|---:|
| **176609988** | 58776 | 54221 | 672 | 2941 | 1181786 |

Characteristics of the control samples are summarized in the table below:

**Table 2.** Dataset characteristics before removing control samples

| **n.row** | n.sample | n.ind | n.plate | n.olinkid | n.npx.na |
|---:|---:|---:|---:|---:|---:|
| **3952704** | 72 | 1 | 672 | 2941 | 26834 |

After removing the control samples, the dataset contained 2,941 protein measurements from 58,704 samples comprising 54,220 individuals across 672 plates, summarized in Table 3:

**Table 3.** Dataset characteristics after removing control samples

| **n.row** | n.sample | n.ind | n.plate | n.olinkid | n.npx.na |
|---:|---:|---:|---:|---:|---:|
| **172657284** | 58704 | 54220 | 672 | 2941 | 1154952 |

## 3. Removing individuals who have withdrawn from UK Biobank, or whose data were not processed

We excluded individuals who have withdrawn from the UK Biobank study since the UKB-PPP was completed, or whose data were not processed by Olink. These samples comprised 0.741% of the dataset, in total.

After the withdrawn and unprocessed cases were removed, the dataset contained 2,941 protein measurements from 58,353 samples comprising 54,219 individuals across 667 plates, summarized in Table 4:

**Table 4.** Dataset characteristics following removal of individuals who withdrew consent or whose data were not processed.

| n.row | n.sample | n.ind | n.plate | n.olinkid | n.npx.na |
|---:|---:|---:|---:|---:|---:|
| **171377949** | 58353 | 54219 | 667 | 2941 | 3681 |

We then excluded rows in the dataset that had missing NPX values (“NPX = NA”), comprising 0.0021% of data.

Following the exclusion of data points missing NPX values, the dataset contained 2,941 protein measurements from 58,353 samples comprising 54,219 individuals across 667 plates, summarized in Table 5:

**Table 5.** Dataset characteristics following removal of missing NPX values (“NPX = NA”).

| n.row | n.sample | n.ind | n.plate | n.olinkid | n.npx.na |
|---:|---:|---:|---:|---:|---:|
| **171374268** | 58353 | 54219 | 667 | 2941 | 0 |

# 4. Outlier sample detection and removal

First, we performed **PCA and IQR-Median Analysis** using the scripts provided in the RMarkdown file (see the companion HTML document).

**Figure 2**. PCA of Olink proteins in UK Biobank.

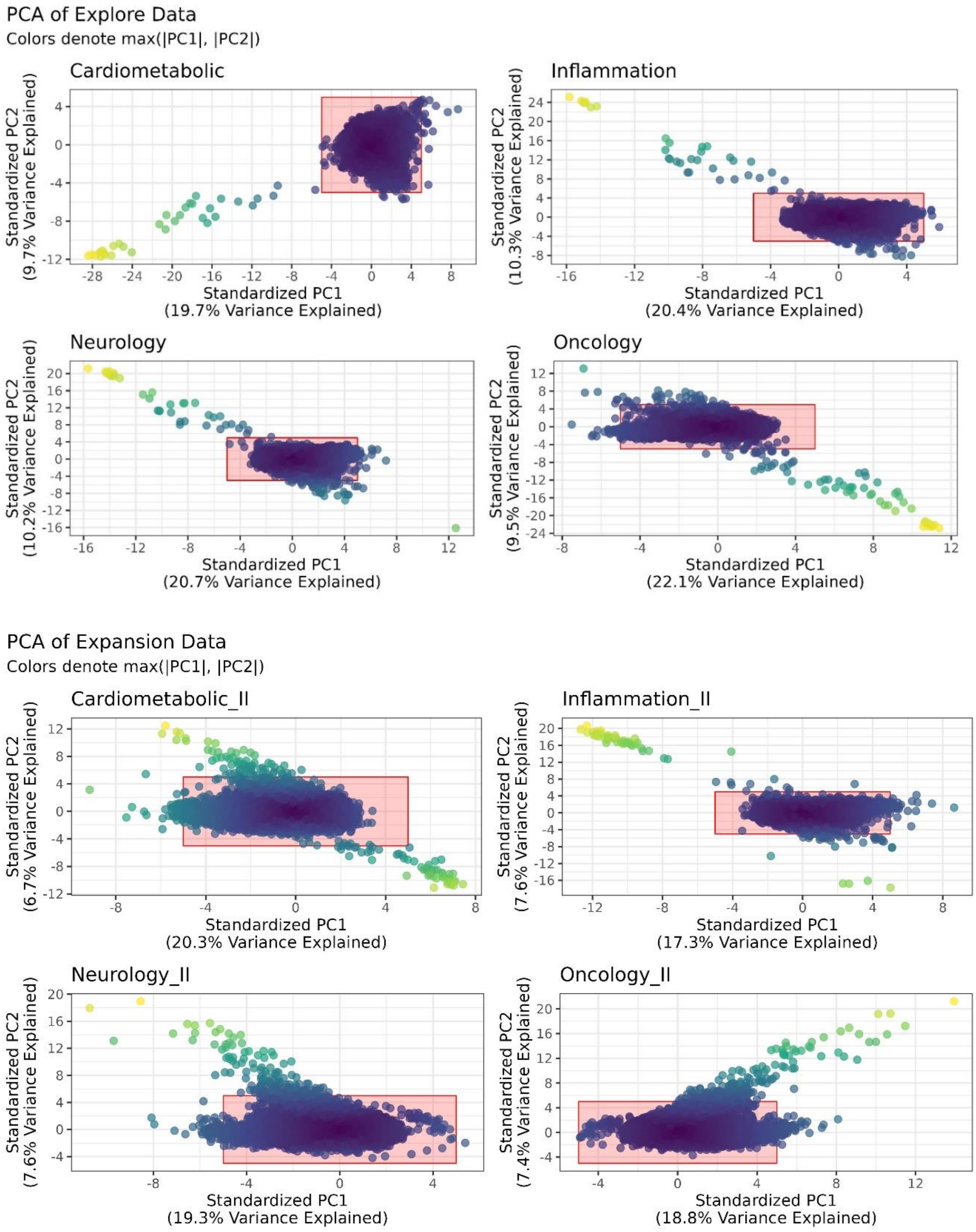
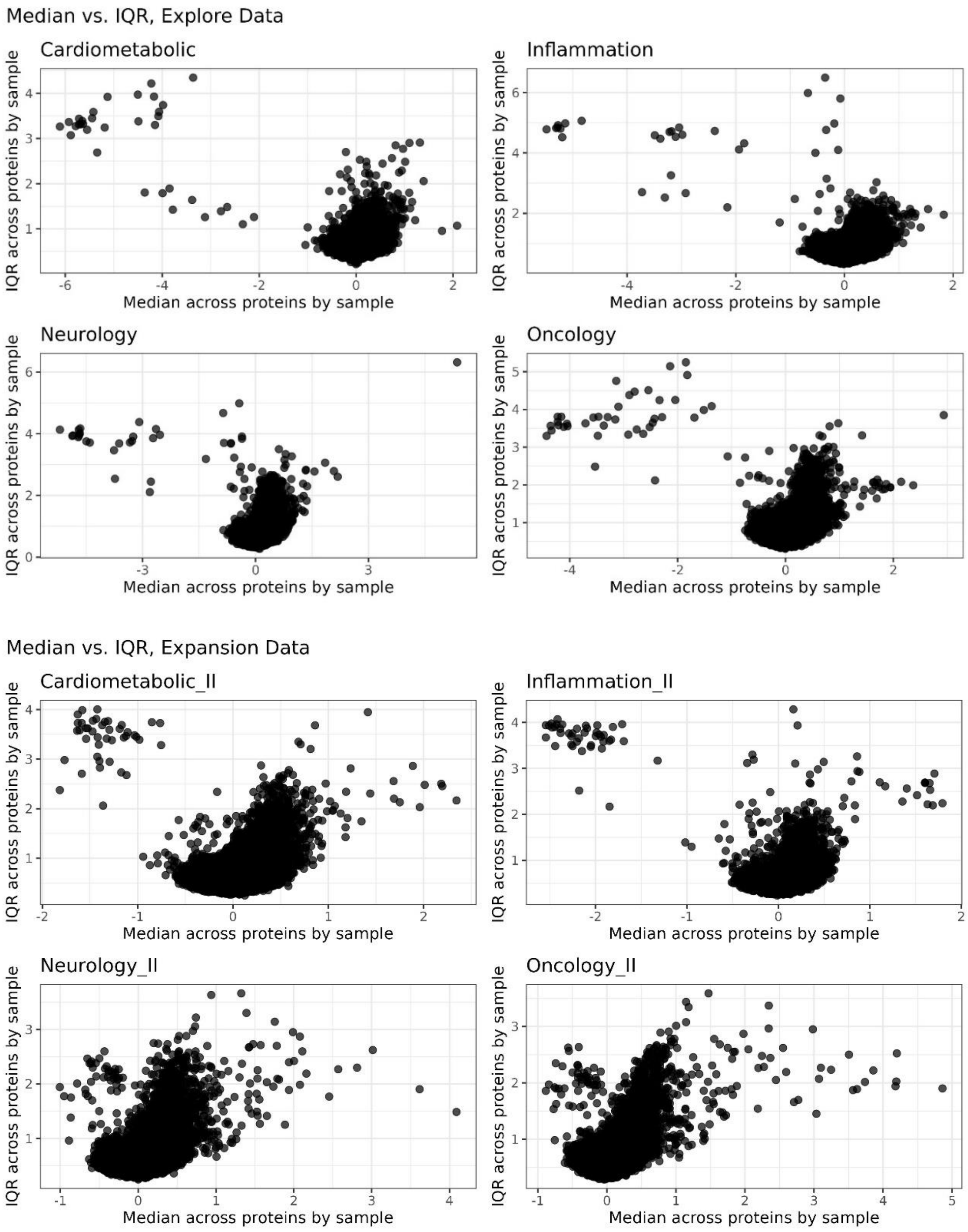


**Figure 3**. Median vs IQR of Olink proteins in UK Biobank.



We then removed (1) samples with a PC1 or PC2 value more than 5 standard deviations from the mean (those highlighted in the plots above), and (2) samples with median concentrations (NPX values) across proteins that were more than 5 standard deviations from the mean median, or those with IQR(NPX) across proteins that were great than 5 standard deviations from the mean IQR. Note that the removal of sample is panel-specific; it is likely that the cardiometabolic panel of a sample is identified as an outlier but the oncology panel not. In this case, only the cardiometabolic panel is removed.

After the outliers were removed, the dataset contained 2,941 protein measurements from 58,285 samples comprising 54,155 individuals across 667 plates, summarized in Table 6:

**Table 6.** Dataset characteristics following removal of outliers

| n.row | n.sample | n.ind | n.plate | n.olinkid | n.npx.na |
|---:|---:|---:|---:|---:|---:|
| **170249812** | 58285 | 54155 | 667 | 2941 | 0 |

## 5. Removing data with QC Warnings or Assay Warnings

We removed all data points with a QC warning or an Assay warning label – automatically assigned by Olink during data generation. We also removed data points with a warning that was manually added by Olink scientists during data generation – i.e., those labelled as "Manual_Warn". Note that a subset of a subject's data can be removed due to a warning, while the remainder of the data can still pass QC and remain in the dataset.

**Figure 4**. Proportion of data with a QC warning, by sample.

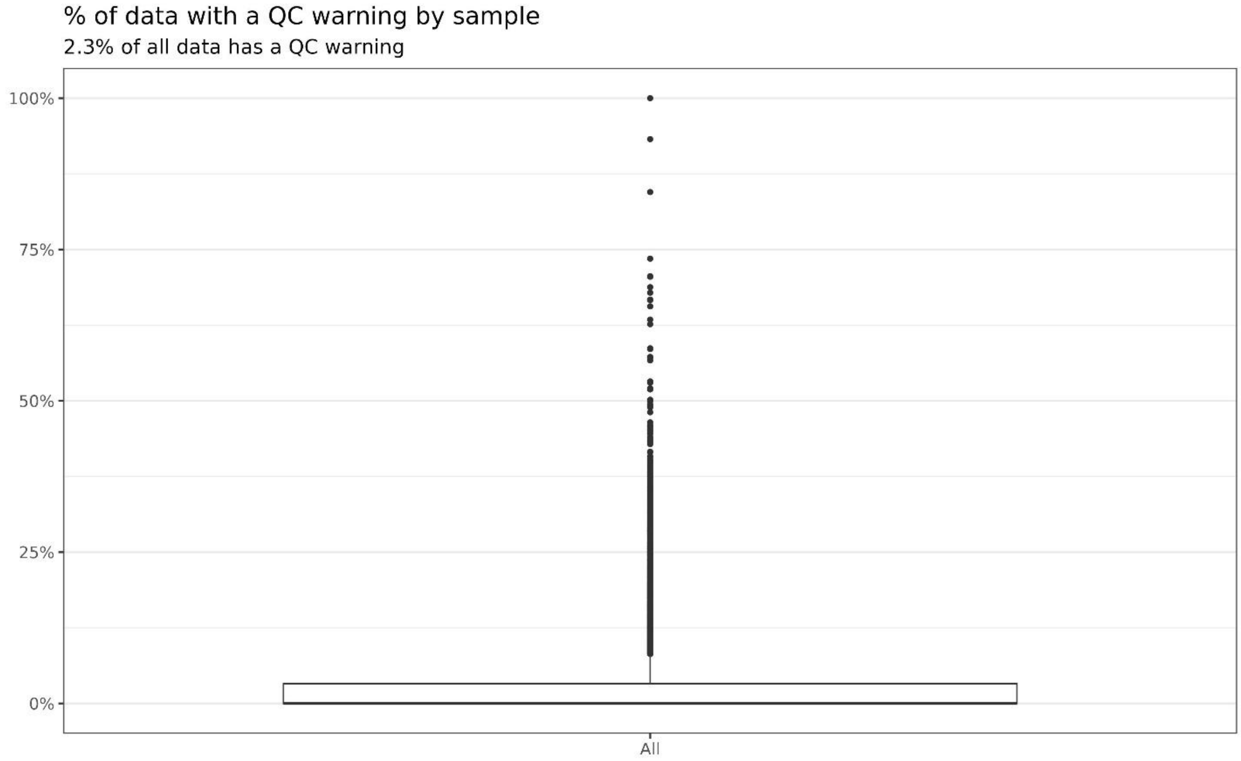
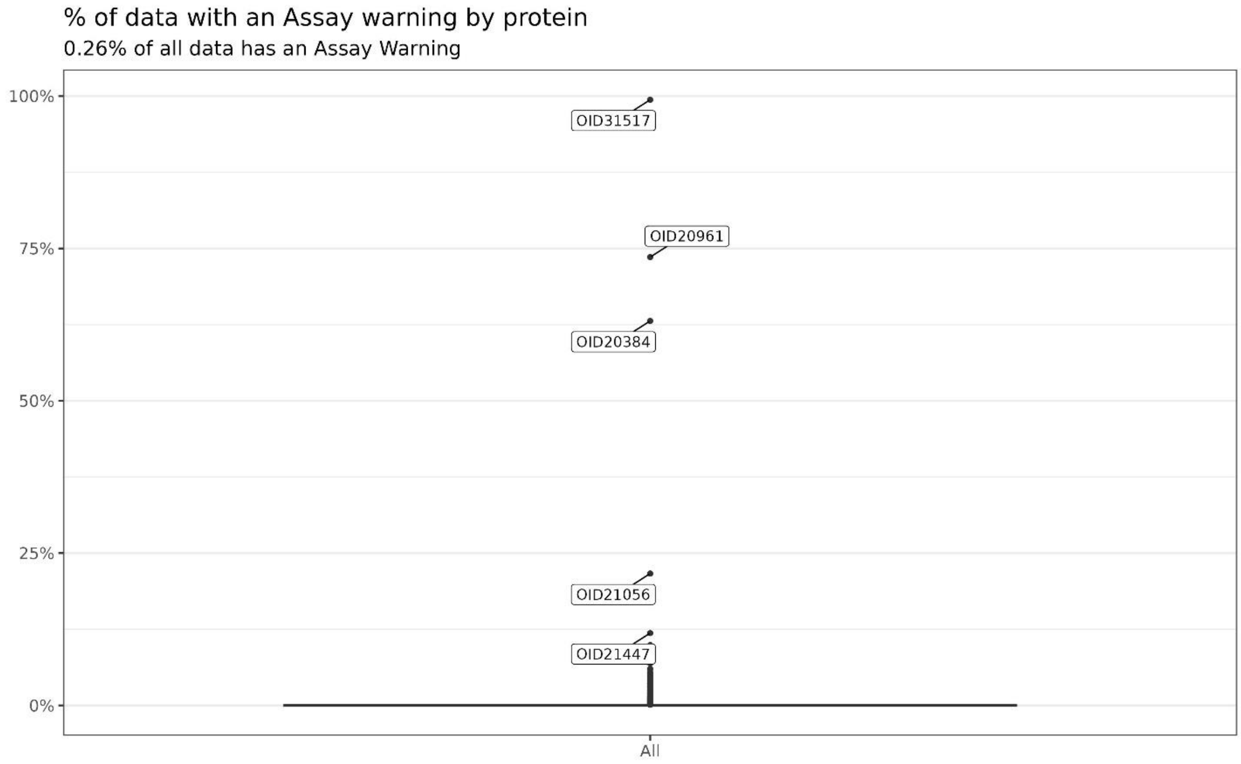


**Figure 5**. Proportion of data with an assay warning, by sample.



After QC or assay warnings were removed from the data, the dataset contained 2,941 protein measurements from 58,284 samples comprising 54,115 individuals across 667 plates, summarized in Table 7.

**Table 7.** Dataset characteristics following removal of 'QC warning' and 'assay warning' flags

| n.row | n.sample | n.ind | n.plate | n.olinkid | n.npx.na |
|---:|---:|---:|---:|---:|---:|
| **165891662** | 58284 | 54115 | 667 | 2941 | 0 |

# 6. Removing likely sample swaps

We flagged samples with likely “sex swaps” indicative of manual sample handling errors or gender mislabeling issues and excluded them from subsequent analyses. Specifically, we removed the following data points from the PHASE 1 dataset:

- Whole plate:
  - For **PlateID_explore** ending in 443, 343, 269 and 344, all data were removed
  - For **PlateID_explore** ending in 124, removed the cardiometabolic panel only
  - For **PlateID_explore** ending in 509, removed the neurology panel only
- Part of the plate:
  - For **PlateID_explore** ending in 030, 346, 413, 262, 488, 246, and 483, a small collection of problematic samples (ranging between 7-8 samples per plate) were removed.

We also removed the following data points from the PHASE 2 dataset:

- Whole plate: **PlateID_expansion** ending in 625, 463, 013, 209, 219, 232, 251, 203, 102, 024, 094, 052, and 106.
- Part of the plate: For **PlateID_expansion** ending in 105, 360, 551 and 030, only problematic samples (ranging between 7- 11) were removed

For the consortium version data, to err on the side of caution, if the aforementioned plates or samples contained 4 or more panels flagged as potential sample swaps, we removed the remaining panels of those plates or samples. These additional sample removals were labeled as "extra" in the **flag** column of the dataset. Removing these samples reduced the size of the dataset by about 1.1%. Please note that in the dataset provided to UK Biobank for approved researchers, we did not remove these additional samples; instead, we leave the decision on whether to retain or remove entire panels or plates based on a minority of potential sample swaps to the discretion of individual analysts.

After removing those flagged as likely sex swaps (but not those labeled as “extra” in the **flag** column), the dataset was reduced to 2,941 protein measurements from 58,227 samples comprising 54,151 individuals across 667 plates, summarized in Table 8:

**Table 8**. Dataset characteristics following removal of likely sample swaps

| **n.row** | n.sample | n.ind | n.plate | n.olinkid | n.npx.na |
|---:|---:|---:|---:|---:|---:|
| **163593426** | 58277 | 54151 | 667 | 2941 | 0 |

The November data release consists of all the Olink data except batch 7 data of Cardiometabolic_II, Inflammation_II, Neurology_II and Oncology_II panels, as shown in Table 9:

Table 9. Olink data by panel and batch

| PANEL | Batch 0-6 | Batch 7 |
|---|---|---|
| Cardiometabolic | √ | √ |
| Inflammation | √ | √ |
| Neurology | √ | √ |
| Oncology | √ | √ |
| Cardiometabolic_II | √ | x |
| Inflammation_II | √ | x |
| Neurology_II | √ | x |
| Oncology_II | √ | x |

After removing batch 7 data of Cardiometabolic_II, Inflammation_II, Neurology_II and Oncology_II panels, the UKB-version data have 2941 protein measurements from 58138 samples comprising, 54054 individuals across 666 plates:

**Table 10**. Dataset characteristics UKB-version data

| **n.row** | n.sample | n.ind | n.plate | n.olinkid | n.npx.na |
|---:|---:|---:|---:|---:|---:|
| **148907040** | 58318 | 54054 | 666 | 2941 | 0 |