1. Data monitoring and quality

Each Enroll-HD PDS goes through stringent **Quality Control** (QC) and **de-identification** procedures prior to release. These are described in detail below.

Data quality control checks are implemented routinely at multiple levels, from **point of data entry**, through to **onsite** and **remote** data monitoring. Prior to a PDS release, data are also subject to an enriched, unique set of remote data review checks. All of these checks aim to maximize data integrity.

Onsite monitoring visits are carried out routinely at each Enroll-HD site to review source data, ensure compliance with study protocol, Good Clinical Practice, and applicable regulations, and retrain staff as needed.

Each month all data that has been signed off during the month is subject to **remote monitoring** procedures, where participant's data (core assessment and selected extended assessments) are subject to cross-sectional QC checks, which include checks for consistency, completeness and plausibility. Participant data are also subject to longitudinal (i.e., within subject) QC checks for a subset of variables (e.g., height, TFC score), which are conducted every 6 months.

Prior to a PDS release, an enriched set of remote data QC checks are also performed. These include comprehensive checks for **outliers**, custom checks for **unusual or implausible values,** and **missing data**. Outlying and implausible values are reviewed by the monitoring and/or medical monitoring teams and queried directly with sites where relevant. In certain instances, these values cannot be queried (e.g., observation recorded under Registry protocol), or are queried and confirmed as correct by site staff. **In these instances, values are provided 'as is'**, and it is left to the analyst to determine how best to evaluate the data. Outlying values and unusual entries are detailed in the *Unusual Findings* document.

In response to frequently asked questions received from users of the Enroll-HD PDS, we have added an additional section on *Age at Onset and Diagnosis Variables* to assist in interpretation of these variables, highlighting unexpected but plausible values and value combinations.

While substantial efforts are made to maximize data quality, researchers are encouraged to **visualize the data** and **perform their own QC checks prior to commencing analyses**.

2. Participant privacy and identification risk management

Overview

Enroll-HD takes great care to protect participant data and privacy.

Enroll-HD recoded participant level data and samples are provided to the research community following three overarching principles:

1. Data and samples are only shared in accordance with **EU GDPR** rules, **US HIPPA** rules, and the participant's **informed consent**.
2. An Enroll-HD **Data Use Agreement** (DUA) and/or **Material Transfer Agreement** (MTA) must be signed and the terms honored by any requester.
3. The **risk for participant identification** is assessed for all participants in Enroll-HD and steps are taken to reduce the risk of identification below a predetermined threshold before data release.

Data preparation for public dataset releases: identification risk management

Before any Enroll-HD data are put into the periodic (released) dataset, a thorough de-identification and quality control process is applied by the Statistics Team. Often data transformations are performed to ensure the data are considered safe to share.

To ensure the data are HIPPA-compliant and the risk for participant identification is low, the de-identification process uses two data safety methods: 1) the "**Safe Harbor**" method and 2) the "**Expert Determination**" Method.

The **Safe Harbor** method refers to the removal of specific identifiers that can directly identify a person in the dataset. As part of the HIPPA privacy rule, a universal list of 18 variables such as birth date, visit date, site name, and country have been determined as identifying characteristics and must be removed from the dataset.

The **Expert Determination** method refers to the obligation of the responsible organization to ask a qualified statistical expert to perform an analysis of the participant identification risk for all individuals in a dataset to ensure the risk is low. The methods used to make that determination and justification of the expert's opinion must be documented and retained. As part of the Expert Determination method, the Enroll-HD statistics team has identified additional variables, through comprehensive literature searches and discussions with HD experts, that can be highly identifying and are considered in identification risk analyses before Enroll-HD datasets are released. Some of these variables are transformed or aggregated for participants where the risk of identification is too high (see *Understand and Interpret the Data* for further details). Other variables are suppressed.

An overview of the de-identification process implemented for Enroll-HD public dataset releases is provided in Figure 1.

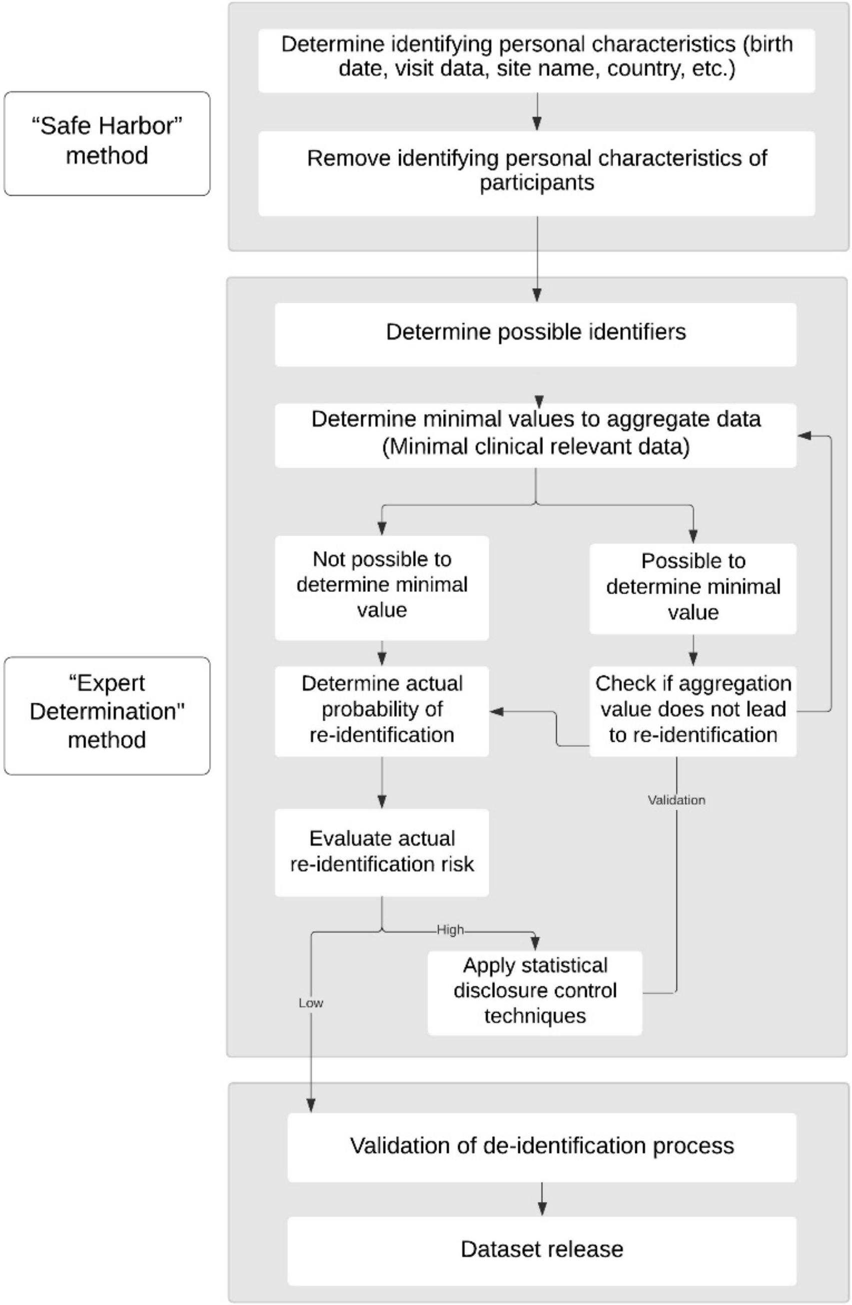
"Safe Harbor" method
Determine identifying personal characteristics (birth date, visit data, site name, country, etc.)
Remove identifying personal characteristics of participants
"Expert Determination" method
Determine possible identifiers
Determine minimal values to aggregate data (Minimal clinical relevant data)
Not possible to determine minimal value
Possible to determine minimal value
Determine actual probability of re-identification
Check if aggregation value does not lead to re-identification
Validation
Evaluate actual re-identification risk
High
Low
Apply statistical disclosure control techniques
Validation of de-identification process
Dataset release

Figure 1. Enroll-HD de-identification process for public dataset releases.

As part of the Expert Determination method, the **probability for participant identification** is calculated for each participant included in a dataset before release. For this purpose, a combination of potentially identifying variables (collectively referred to as a "key") are considered.

For Enroll-HD, these "key" variables are: HTT CAG size, age, race, sex, educational level (ISCED), and BMI. (Note that several potentially identifying variables are removed prior to this analysis, such as site and country.) These "key" variables were selected following thorough literature reviews and discussions with HD experts.

The individual risk of identification for a participant in Enroll-HD is defined as the **probability that a participant could be correctly** identified from the full Enroll-HD sample **by looking at a specific combination of the key variables**, i.e. the probability of a participant being unique for the combination of his characteristics. For example, a participant who is female, 42, has a Ph.D., and has a very high (and therefore rare) HTT CAG size might have a higher probability of being accurately identified than a 42-year-old female participant with a high school education and a HTT CAG size in the median range. The individual and global risk of identification is calculated for the Enroll-HD datasets by using the R software, version 3.0.3, package "sdcMicro."

The acceptable risk for identification of a participant in Enroll-HD is set to **3%**. For participants with a risk above 3%, SDC techniques (e.g. further aggregation of variables) must be applied and if these do not reduce the risk for identification below 3%, the participant is removed from the final released dataset. Given the delicate situation for participants that are from HD families but do not know if they carry the gene expansion, the risk threshold for these participants is set to **1%** and if the risk is higher, the participant is removed from the periodic data release.[i,ii] Although some participants are not included in a dataset, it does not mean they will never be included in any future datasets because the individual identification risk can change when additional data is collected.

One additional aspect of risk is assessed by the Statistics Team. Over time some requesters receive multiple Enroll-HD PDS and SPS datasets, bio-samples, and platform study datasets. All released Enroll-HD datasets and bio-samples include the same recoded ID for each participant, to allow the datasets to be easily linked into one extensive dataset. Since the specific combination of variables available determines the risk for identification, it is not enough to consider the risk for identification for each of the released datasets alone. In such circumstances, the **combined identification risk** across all datasets must be assessed prior to releasing additional data to a given researcher.

The identification risk assessment and the dataset preparation follow SOPs and guidelines maintained by the Enroll-HD team. These documents are updated periodically, and a record of distributed versions is maintained.

[i] Emam KE, Abdallah K. De-identifying Clinical Trials Data. *Applied Clinical Trials.* Mar 20 2015. http://www.appliedclinicaltrialsonline.com/de-identifying-clinical-trials-data. Accessed June 5, 2020.

[ii] Emam KE, Dankar FK, Vaillancourt R, Roffey T, Lysyk M. Evaluating the Risk of Re-identification of Patients from Hospital Prescription Records. *Can J Hosp Pharm*. 2009;62(4):307-319. doi:10.4212/cjhp.v62i4.812.