Data Quality Management and Participant Privacy

Contents

1. Data monitoring and quality

Each Enroll-HD PDS goes through stringent Quality Control (QC) 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., symbol digit modality test, TFC score), which are conducted every 6 months.

Prior to a PDS release, an enriched set of remote data QC checks (~400) are also performed. These include custom multivariate checks for unusual or implausible values, and systematic checks of continuous variables for extreme outlying values, flagged using data-driven thresholds, or pre-specified custom thresholds based on plausibility.

Values identified by the QC check battery are reviewed by the Data Monitoring and/or Medical Monitoring teams. Consequent actions (and outcomes) are listed below:

- Query issued, sent to study site > datum confirmed incorrect and updated by site | datum confirmed correct and not updated by site, retained 'as is' for PDS
- Query issued, sent to another source (e.g., BioRep) > datum confirmed incorrect and updated | participant added to "exclusion list"
- Query not issued > participant added to "exclusion list"
- Query not issued > visit added to "exclusion list"
- Query not issued > datum recoded with exceptional value code (e.g., WRONG/9996)
- Query not issued > datum retained 'as is' and included in the document "Quality Control: observations and unusual finding"

Issued queries may result in identified values being updated/corrected. Certain issued queries are checked and

confirmed as correct by site staff. In these cases, unless the datum is impossible (e.g., out of scale range), the value is left 'as is' in the dataset. All of these values should be listed in the PDS document "Quality Control: Observations and Unusual Findings". It is left to the analyst to determine whether to include or exclude these values or perform sensitivity analyses.

Certain specific QC findings cannot be queried or corrected (e.g., inconsistency in assigned *hdcat* and *caghigh*). These result in the participant – or a visit - being excluded from the dataset.

In other instances where identified values cannot be queried and corrected (e.g., if the observation was recorded in a REGISTRY visit transferred into the Enroll-HD database), the variable value may be recoded with an exceptional value data code (e.g., WRONG/9996).

In response to frequently asked questions received from users of the Enroll-HD PDS, we now include a section on 'HD onset and diagnosis' variables in the ***Understand and Interpret the Data*** document 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

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

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

1. **Informed consent and regulatory requirements:** Data and samples are only shared in accordance with **EU GDPR** with special consideration of any Personal Health Information (defined by the HIPAA Privacy Rule), and the participant's **informed consent**.

2. **Use agreements:** An Enroll-HD **Data Use Agreement** (DUA) and/or **Material Transfer Agreement** (MTA) must be signed, and the terms honored by any requester.

3. **Identification risk assessment and minimization:** Various methods are applied to data within each dataset to minimize identification risk, including variable suppression, transformation, and aggregation. 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.

To ensure the risk for participant identification is minimized, two methods are employed: 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. The HIPAA Privacy Rule outlines a list of 18 variables, such as geographic subdivisions smaller than state, and other characteristics that could uniquely identify the individual. To the extent that these data are collected in the study, the data are removed from the dataset or transformed to reduce identification risk (e.g., date of birth is converted to age).

The **Expert Determination** method requires 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 that may be potentially identifying. Many of these variables are suppressed, transformed, or aggregated in dataset releases. Full details of variable availability, transformation, and aggregation – pertaining to the Enroll-HD dataset releases - are provided in the Enroll-HD *Data Dictionary* and the document *Understand and Interpret the Data*.

In addition, the probability for individual 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. In Enroll-HD, these "key" variables are: HTT CAG size, age (baseline), race (aggregated), sex, educational level (ISCED; baseline), BMI (baseline) and region. These key variables were selected following thorough literature reviews and discussions with statistical experts. The R software package *sdcMicro* is used to determine individual participant identification risk based on the above variable set. Genotype unknown participants exceeding a 1% identification risk probability are excluded from the dataset, for all other participants a 3% identification risk threshold is applied.

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.**

Although some participants are excluded from a dataset due to exceeding acceptable risk thresholds, they may be included in future datasets; individual identification risk can change when additional data are added to a dataset.

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

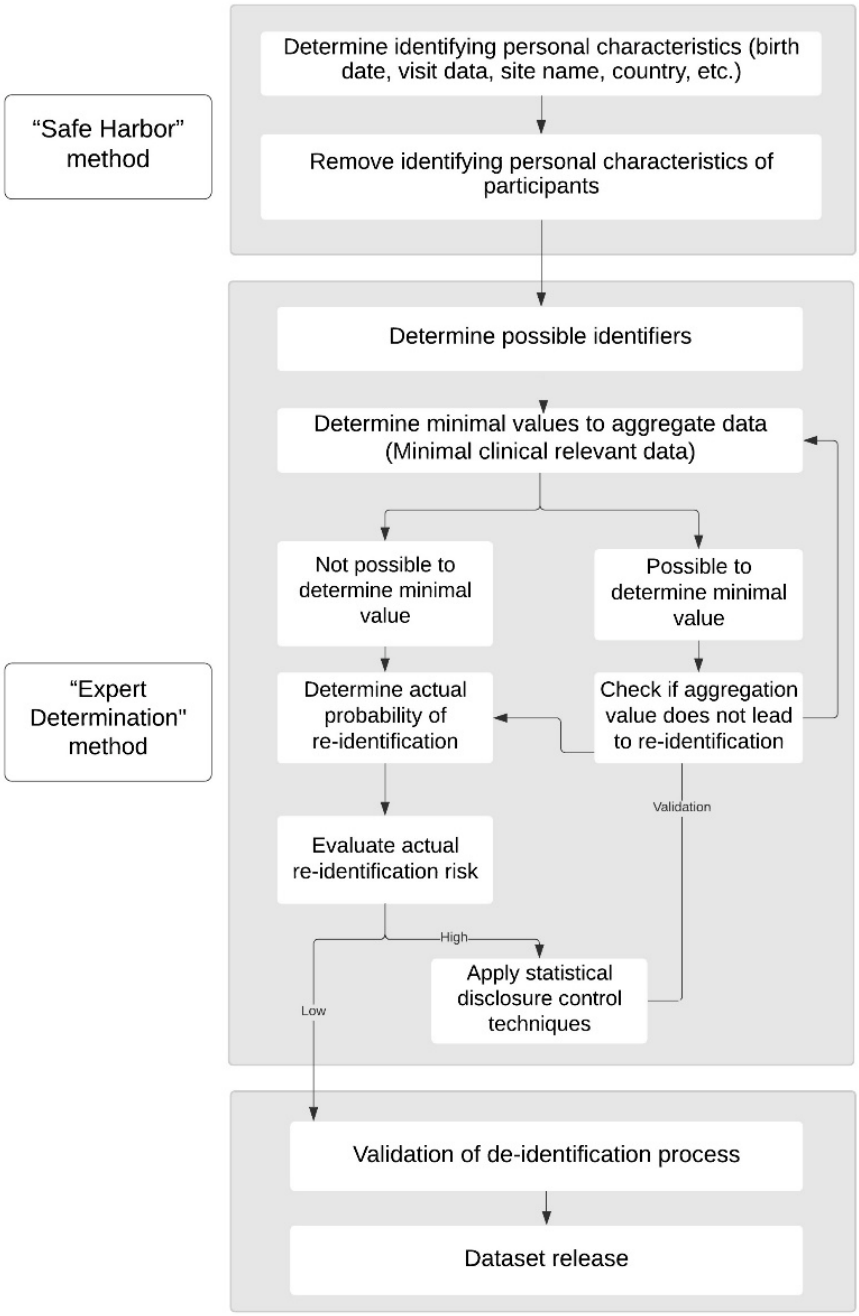

Figure 1. Enroll-HD identification risk management process for dataset releases.

[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.