

# UNDERSTANDING DATASET STRUCTURE

## Contents

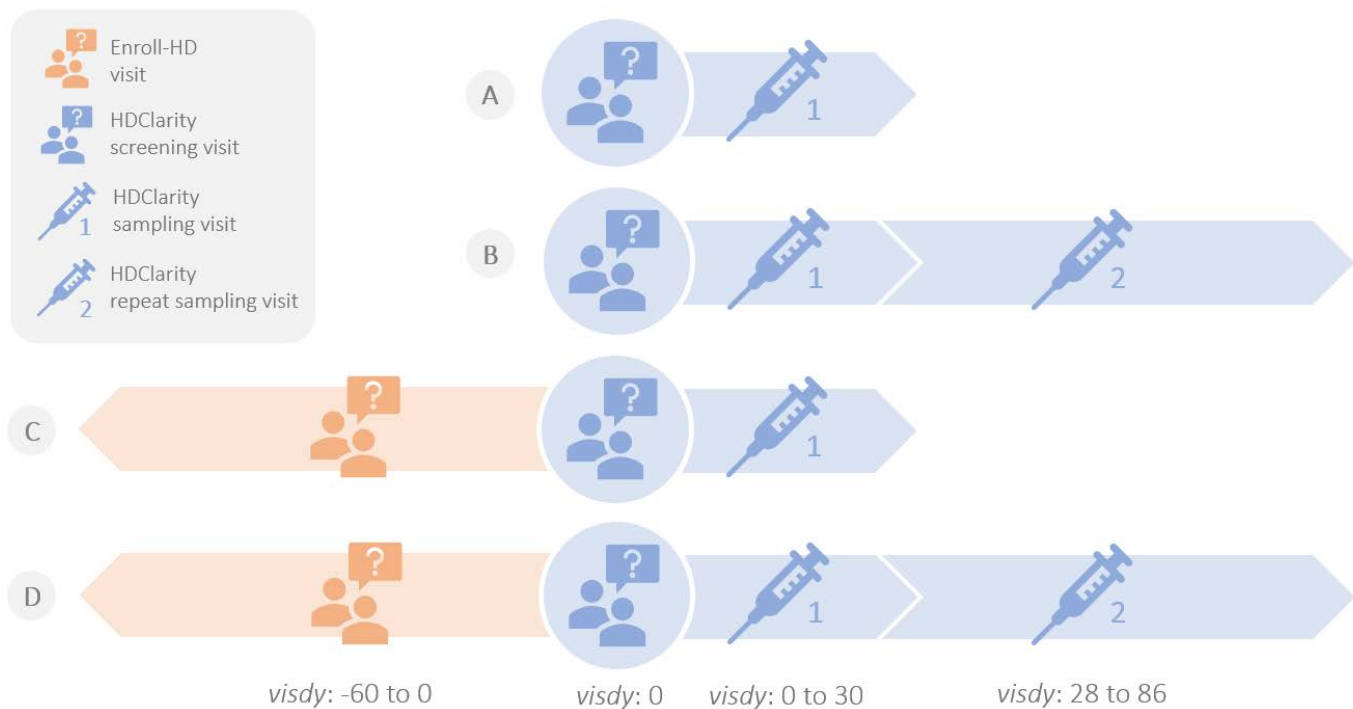
Participant identifiers.....	2
HDClarity “visit packages” .....	3
Dates.....	6
Structure of dataset.....	7
Studies within the HDClarity PDS.....	7
Data files within the HDClarity PDS.....	7
Entity relation diagram.....	10
Structure of variables.....	11
Visit data.....	14
Merging and aligning files.....	15
EXCEL.....	15
R.....	16
Linking HDClarity samples and clinical data.....	17

## Participant identifiers

All participants in HDClarity (CLR) have a unique participant ID (*usubjid*) in the format RYYYYYYYYY, where “Y” is a number between 0 and 9 or an “X”. The participant ID used in CLR is the same as the participant ID used in the Enroll-HD (ENR) study.

## HDClarity visit packages

HDClarity is an Enroll-HD platform study, which means all HDClarity participants are also enrolled in the Enroll-HD study. HDClarity “visit packages” comprise an HDClarity **screening** visit, one or two HDClarity **CSF sampling** visits, and – typically – an **Enroll-HD** visit. Clinical HD data (Motor, TFC, Function, Cognitive, PBA-s) and Variable form data (see Data Dictionary for complete variable lists by form) are gathered at the HDClarity screening visit if the screening visit took place more than 2 months after an Enroll-HD visit. If the HDClarity screening visit took place *less* than 2 months after an Enroll-HD visit, clinical HD and variable form data collected from this Enroll-HD visit are used. This is to minimize burden on participants. Data gathered at visits is outlined in ‘visit plan’ section of the Data Dictionary. All possible combinations of visits which may constitute a single HDClarity “visit package” are illustrated in **Figure 1**.



**Figure 1. HDClarity visit packages.** Each HDClarity “visit package” includes between two and four visits. A visit package contains, at minimum, an HDClarity screening visit, and an HDClarity sampling visit collected 0-30 days after screening (1A-1D). Some visit packages also include an HDClarity repeat sampling visit, conducted 4-8 weeks after the first sampling visit (1B, 1D). Most visit packages also include an Enroll-HD study visit (1C, 1D). Regardless of the number of visits included in the visit package, each package encompasses clinical data, screening data, and sampling data. Typically, participants attend an HDClarity screening visit within 2 months of an Enroll-HD study visit. In these instances, clinical HD data are leveraged from the Enroll-HD visit, and are not performed at the HDClarity screening visit (1C, 1D). If an Enroll-HD visit was **not** conducted within the 2 months prior to the HDClarity screening visit, then clinical HD assessments corresponding to the

core Enroll-HD assessments are performed at the HDClarity screening visit (1A, 1B). The *visdy* variable - provided alongside each visit - denotes the timing of the visit relative to the **first** HDClarity screening visit in number of days. Negative values for *visdy* indicate that the visit took place before the first HDClarity screening visit.

All HDClarity participants have an HDClarity screening visit and if eligible will be invited to a sampling visit within 30 days of the screening visit. Up to 20% of the HDClarity participants will be invited to return for an optional repeat sampling visit within 4-8 weeks of the original sampling visit. This means an HDClarity participant can have up to four visits, within a window of about 5 months, that are all considered part of the same HDClarity enrolment and we refer to the collective data as the “HDClarity visit package”.

All visits belonging to the same visit package for any given participant have the same *subjid* – a string variable consisting of the participant ID followed by a # and a number indicating the visit package. For example, a visit in the HDClarity dataset with a *subjid* suffixed “#1” indicates this visit was from the initial visit package at which the individual enrolled in the HDClarity study. The consequent visit package visits would be suffixed with “#2”, and so on (see Table 1).

*Table 1. Identifying the temporal order of visits and visit packages in HDClarity.*

usubjid	subjid	studyid	visit	seq	visdy
R001001004	R001001004#1	ENR	Follow up	1	-7
R001001004	R001001004#1	CLR	Screening	2	0
R001001004	R001001004#1	CLR	Sampling	3	10
R001001004	R001001004#1	CLR	Repeat Sampling	4	67
R001001004	R001001004#2	ENR	Follow up	5	354
R001001004	R001001005#2	CLR	Screening	6	354
R001001004	R001001005#2	CLR	Sampling	7	376

In Table 1, seven visits are listed for participant R001001004. Data from these seven visits are all considered part of the CLR study dataset. The **sequence** (*seq*) and **visit day** (*visdy*) variables clearly indicate the temporal order in which the visits took place. The number following the “#” in the *subjid* column delineates which visits belonged to first or second CLR visit package respectively. In this example, visit package 1 comprises an Enroll-HD visit that took place 7 days before the first HDClarity screening visit, an HDClarity screening visit, a sampling visit that took place 10 days after the screening visit, and a repeat sampling visit which took place 67 days after the screening visit. The participant came in again approximately a

year later for their next Enroll-HD visit (*visdy* = 354), completed their next HDClarity screening visit on the same day as the Enroll-HD visit (*visdy* = 354), and an HDClarity sampling visit an additional 22 days later (*visdy* = 376). The first HDClarity visit package therefore includes four visits, whereas the second HDClarity visit package only includes three visits.

## Dates

Date variables in datasets can increase the risk for participant re-identification. For that reason, all date variable values in the CLR dataset are changed to a numeric value representing the number of days since the participant’s **first ever** CLR screening visit (variable named *visdy*). This means the *visdy* variable is participant centric and represents the relative time interval between the very first CLR screening visit and any other visit for that specific participant.

Table 2 illustrates the visit schedule for three different CLR participants. Each participant is represented on a separate row. The time interval between the CLR screening visit and the CLR sampling visit is 21 days for both participants R001001001 and R123456789, but their visits took place in different years (2017 and 2018), which is not apparent from the *visdy* variable.

*Table 2. Conversion of dates to the “visdy” variable.*

Participant (usbjid)	ENR visit		CLR screening visit		CLR sampling visit		CLR repeat sampling visit	
	visdat	visdy	visdat	visdy	visdat	visdy	visdat	visdy
R001001001	10-Jan-17	0	10-Jan-17	0	31-Jan-17	21	28-Mar-17	77
R123456789	05-Jan-18	-5	10-Jan-18	0	31-Jan-18	21		
R222333444	05-Jan-19	-2	07-Jan-19	0	06-Feb-19	30		

Visits that took place before the first CLR screening visit will be represented with **negative** *visdy* variables. Both participants R123456789 and R222333444 shown in Table 2 had an Enroll-HD visit before their CLR screening visit (5 and 2 days prior, respectively).

HDClarity is a longitudinal study in which participants return for visits over consecutive years. **The visit day (*visdy*) is always anchored to the first ever CLR screening visit (day 0).**

**WARNING:** It is **not** possible to line up HDClarity data with Enroll-HD datasets (e.g. Enroll-HD PDS5) using the *visdy* variable. The *visdy* variable is a **study specific** variable calculated based on the baseline/screening date of a participant in each specific study.

In addition, while all HDClarity participants are also enrolled in Enroll-HD, not all participants in HDClarity periodic dataset (PDS) releases may be present in Enroll-HD PDS releases.

## Structure of dataset

The HDClarity periodic dataset (PDS) consists of nine data files. These files contain data items defined by variables. Variables are taken from the eCRFs from HDClarity and Enroll-HD visits. Some data values have been transformed or aggregated to decrease the risk of identification of participants (for further details, see document: Data Quality Management and Participant Privacy).

### Studies within the HDClarity PDS

All individuals in the HDClarity PDS are both HDClarity and Enroll-HD participants. Study specific protocols and eCRFs are housed under the [General Documents](#) section on the Enroll-HD webpage.

*Table 3: Studies within HDClarity PDS.*

Study Name	Acronym	Sponsor	Chronological order
HDClarity	CLR	UCL	HDClarity enrollment always follows enrollment in Enroll-HD
Enroll-HD	ENR	CHDI	All HDClarity participants are also in Enroll-HD. A significant volume of clinical data in HDClarity are leveraged from Enroll-HD

### Data files within the HDClarity PDS

HDClarity PDS is comprised of nine data files, each of which fall into one of three categories:

- Participant-based:** *profile, pharmacotx, nonpharmacotx, nutsuppl, comorbid*

These contain general study-independent information about the participant. This information is applicable to both studies. In the *'profile'* file there is one line per participant and the file includes variables that never – or very rarely – change such as gender, ethnicity, CAG repeat size, age at onset etc. The *'pharmacotx, nonpharmacotx, nutsuppl'* and *'comorbid'* files are continuous logs that are re-opened and updated at each visit.
- Study-based:** *participation, assessment*

These files contain study specific information about a participant within a study. The *'participation'* file includes one line per visit package per participant. Each line in *'participation'* file includes information on age and *hdcat* at the HDClarity screening visit in that visit package, the timing of the visits within that visit package and which biosamples were collected at each visit.

The assessment file contains information regarding the availability (yes or no) of relevant information, namely scales, or CSF collection, by visit.

- **Visit-based:** *visits, csfquality*

These files contain all visit-dependent information for the study, combined into one data file. The ‘*visits*’ file includes clinical data collected at each visit whereas the ‘*csfquality*’ file includes microscopic CSF erythrocyte and leukocyte counts and CSF hemoglobin measurements related to indicated sampling visits when available.

Each HDClarity PDS data file is described in Table 4.

**Table 4: HDClarity PDS data file descriptions.**

Data file	Type	Studies	Description
<i>profile</i>	participant	CLR, ENR	General and annually updated information including the forms: Demographics, HDCC (HD clinical characteristics), CAG.
<i>pharmacotx</i>	participant	ENR	Information about pharmacological therapies
<i>nutsuppl</i>	participant	ENR	Information about nutritional supplements
<i>nonpharmacotx</i>	participant	ENR	Information about nonpharmacological therapies
<i>comorbid</i>	participant	ENR	Information about comorbid conditions
<i>participation</i>	study	CLR, ENR	Provides information about HDClarity visit packages including the order of the visits, the timing of the visits, the types of biosamples collected, the age and hdcat at the screening visit within each visit package.
<i>assessment</i>	study	CLR, ENR	Visit-specific information about which assessments were performed at each visit
<i>visits</i>	visit	CLR, ENR	Data from the HDClarity and Enroll-HD studies
<i>csfquality</i>	visit	CLR	Information about CSF quality in HDClarity

For detailed information on each constituent *form*, please refer to the [CRFs](#).

The number of participants included in each PDS data *file* is illustrated in Figure 2.



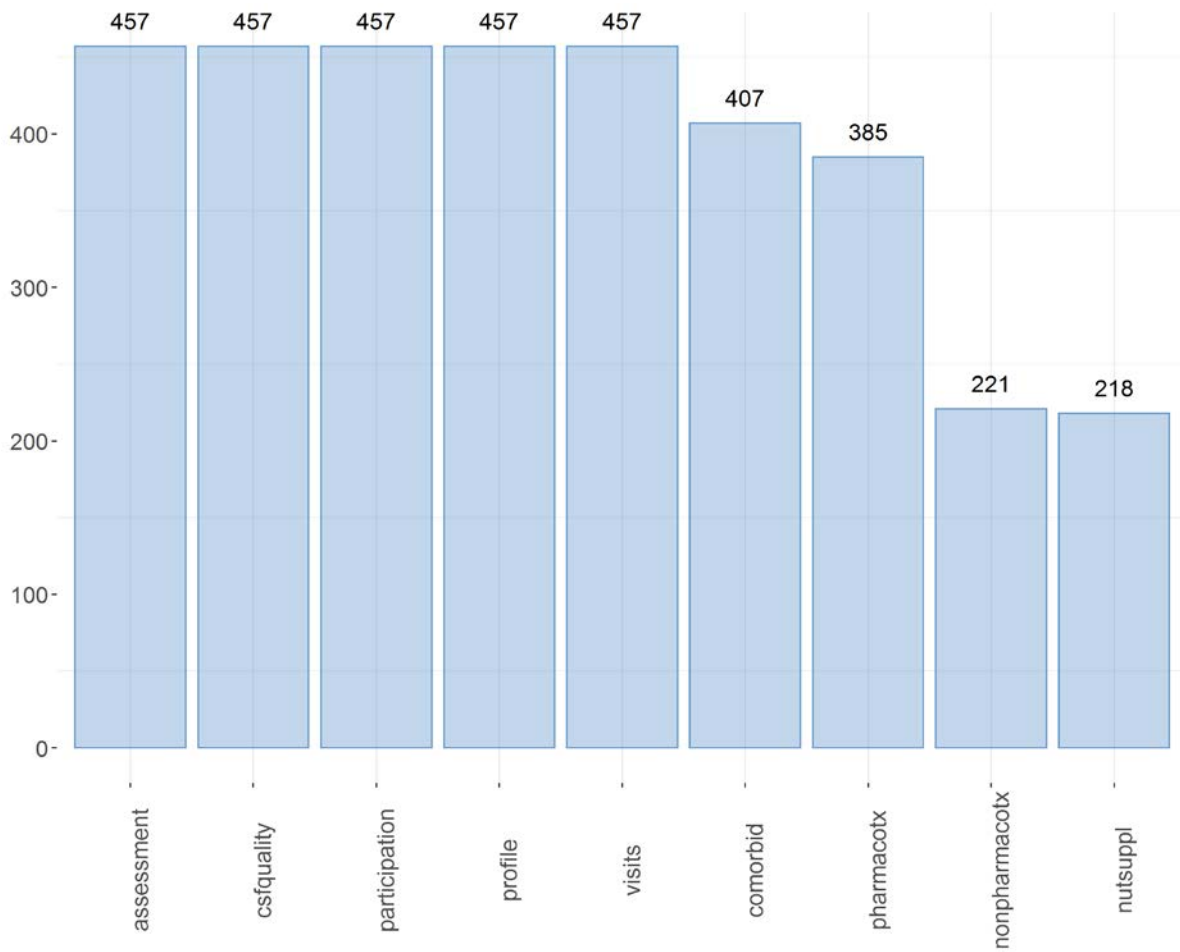


Figure 2. Number of participants included in each HDClarity PDS2 data file.

### Entity relation diagram

The HDClarity PDS data file **entity relation diagram** is presented in Figure 3. This illustrates the **relationship** between each of the component **data files**, along with their **key variables** (primary keys [PK] and foreign keys [FK]) which are required to combine the data files.

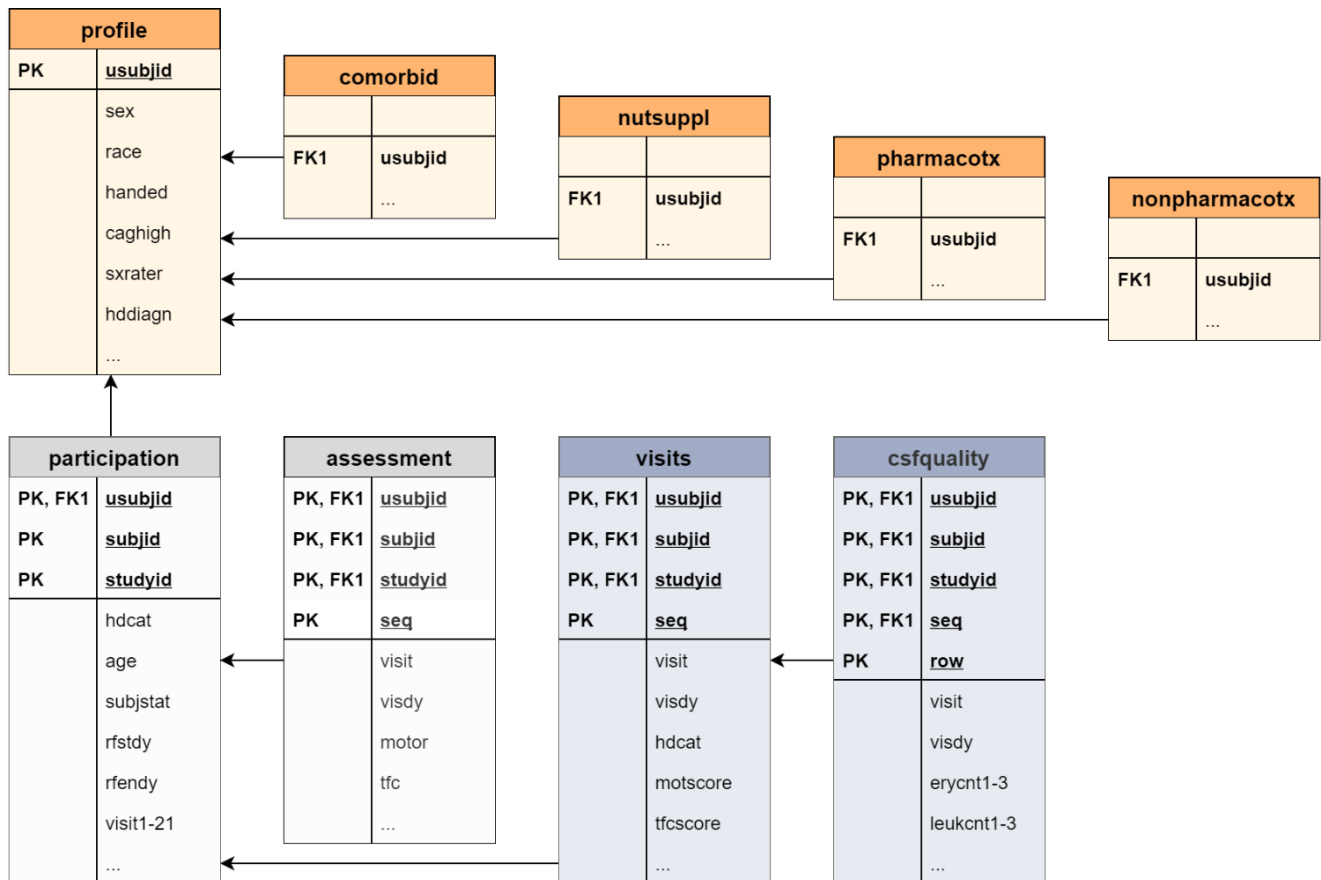


Figure 3: HDClarity PDS entity relationship diagram.

## Structure of variables

Each HDClarity PDS data file contains variables. The data dictionary, accessible via the [General Documents](#) tab, lists all variables by form with the following attributes:

**Table 5. Data dictionary column fields.**

Attribute	Description
Label	Description of variable
Domain	Variable domain abbreviation (CDISC SDTM compliant). Refer to Table 6 for definitions.
Category	Category (optional)
Variable	Internal variable name. Variable is defined in CDISC SDTM compliant naming convention or as close as possible.  <i>Boolean:</i> Represents the values 1 (yes) and 0 (no).  <i>Number:</i> Represents integer or floating-point data values.  <i>Text:</i> Represents alphanumeric string data values.
Data Type	<i>Date:</i> The date type is represented as the number of days relative to the date of the participant’s very first HDClarity screening visit date. Note that dates that have been specified in the original data as “incomplete” (e.g. without entry of a day) have been automatically completed by the following rule: use “15” as day if day is missing and use “1” as day and “7” as month if day and month is missing. After this completion, the number of days relating to the very first HDClarity screening visit date is calculated and provided in the data. The information about whether a date has been automatically completed is not included in the PDS but can be obtained via SPS request.  <i>Single choice:</i> Variable with assigned coding list where one item can be selected. The value within an export is taken from the underlying coding list which is defined as a parameter in the data dictionary tables.
Parameter	Parameter value of coded variables (optional).
Coding	Internal parameter value of coded variables (optional).
Unit	Unit of input field (optional).
Transformation	One important objective of the periodic dataset is to de-identify the HDClarity data in order to minimize the possibility to identify a participant. Therefore, some variables are transformed, recoded or outliers removed/cut. These transformations are described on a variable to variable basis. Details on variable transformations or calculations are also provided here.
Availability	All variables in the HDClarity dataset are listed in the Data Dictionary. This column allows the researcher to identify which variables are available in the PDS (“PDS”), which are available via special request (“available upon Scientific Review Committee approval”), and which are restricted (“not available”).

**Table 6: Variable domain definitions.**

Domain Abbreviation	Domain	Description of the domain
AE	Adverse Events	An event domain that contains data describing untoward medical occurrences in a participant which may not necessarily have a causal relationship with the study.
CM	Concomitant/Prior Medications	An interventions domain that contains concomitant and prior medications used by the subject, such as those given on an as-needed basis or condition-appropriate medications.
CO	Comments	A special-purpose domain that contains comments that may be collected alongside other data.
DM	Demographics	A special-purpose domain that includes a set of essential standard variables that describe each subject in the study.
DS	Disposition	An events domain that contains information encompassing and representing data related to subject disposition.
IE	Inclusion/Exclusion Criteria	A findings domain that contains those criteria that cause the subject to be in violation of the inclusion/exclusion criteria.
LB	Laboratory Test Results	A findings domain that contains laboratory test data such as hematology, clinical chemistry and urinalysis. This domain does not include microbiology or pharmacokinetic data, which are stored in separate domains.
MH	Medical History	An events domain that contains data that includes the subject's prior medical history at the start of the study
QS	Questionnaires	A findings domain that contains data for named, stand-alone instruments designed to provide an assessment of a concept. Questionnaires have a defined standard structure, format, and content; consist of conceptually related items that are typically scored; and have documented methods for administration and analysis.
SAE	Serious adverse event	An event domain that contains data describing serious untoward medical occurrences in a participant which may not necessarily have a causal relationship with the study.
SC	Subject Characteristics	A findings domain that contains subject-related data not collected in other domains.
SUPPAE	Supplemental Qualifiers for Adverse Events	A dataset used to represent non-standard variables and their association to parent records in the Adverse Events domain.
SUPPCM	Supplemental Qualifiers for Concomitant/Prior Medications	A dataset used to represent non-standard variables and their association to parent records in the Concomitant/Prior Medications domain.
SUPPLB	Supplemental Qualifiers for Laboratory Test Results	A dataset used to represent non-standard variables and their association to parent records in the Laboratory Test Results domain.
SUPPMH	Supplemental Qualifiers for Medical History	A dataset used to represent non-standard variables and their association to parent records in the Medical History domain.
SUPPQS	Supplemental Qualifiers for Questionnaires	A dataset used to represent non-standard variables and their association to parent records in the Questionnaires domain.

<b>SUPPSC</b>	Supplemental Qualifiers for Subject Characteristics	A dataset used to represent non-standard variables and their association to parent records in the Subject Characteristics domain.
<b>SUPPVS</b>	Supplemental Qualifiers for Vital Signs	A dataset used to represent non-standard variables and their association to parent records in the Vital Signs domain.
<b>SV</b>	Subject Visits	A special-purpose domain that contains the actual start and end data/time for each visit of each individual subject.
<b>VS</b>	Vital Signs	A findings domain that contains measurements including but not limited to blood pressure, temperature, respiration, body surface area, body mass index, height and weight.

## Visit data

The HDClarity PDS data file 'visits' contain HDClarity "visit packages" for all HDClarity participants that met the following criteria: a) Sampling visit successfully conducted (partial sampling visits included), b) informed consent form confirmed correct and complete, c) visit data monitored and closed.

For information on determining timing and temporal sequence of visits, please refer to '*HDClarity Visit Packages*' section above.

## Merging and aligning files

HDClarity PDS contains one **key variable**, *usubjid*, that is included in every data file. This allows the user to **merge** two or more data files, linking information for each participant across data files.

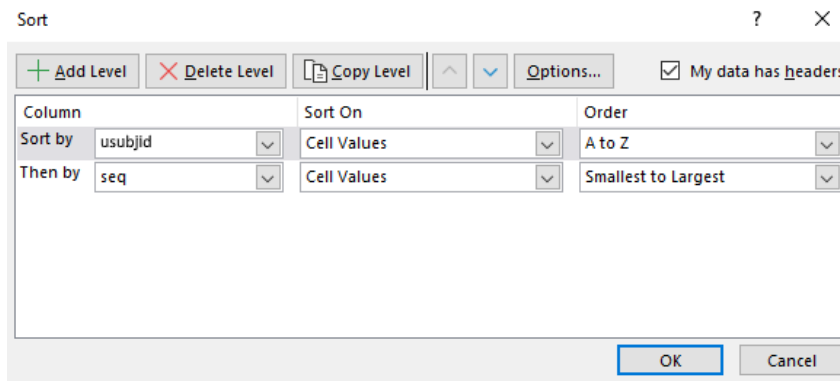
To merge longitudinal data available in visit-based data files, it is important to take the variable *seq* into consideration, as this variable provides information on the visit sequence. Use the variables visit day (*visdy*) and *studyid* to check visits align correctly.

Below we provide guidance on selecting entries/lines using Excel or R, respectively. The example described below shows how age of HD diagnosis (*hddiagn*) from the *profile* file can be merged to age at last visit (*age*) of each participant in the *visit* file.

**WARNING:** Merging data files in Excel can cause misalignment. Before analyzing the data, check that the resulting merged data file correctly lines up across appropriate fields. To avoid issues with merging data files, it is highly recommended that you use a reputable statistical software package.

### EXCEL

- Sort your *visit* file by *usubjid*, then add a level and sort by *seq* (Smallest to Largest);



- Create a new column with the name “*select*”;
- On the first row of this column type the formula “=IF ( A2=A3 , " " , " 1 " ) ”, where A corresponds to the column of *usubjid* and A2 to the first row/value of *usubjid*. Then press the **Enter** key and drag Auto fill to copy the formula to the range you need. This will create a column with the value “1” on the row with maximum *seq* for each participant;
- Filter for the variable ‘*select*’ with the value “1”;
- Create a new column with the name of the variable you want to merge in the ‘*visit*’ file (in this case *hddiagn*);

- In the new column, use a `VLOOKUP()` function to merge `'hddiagn'` from the `'profile'` file, using the variable `'usubjid'` as a linker. Then click the **Enter** key and drag Auto fill to copy the formula to the range you need.

## R

- Select the rows with highest value in the `seq` variable for each participant

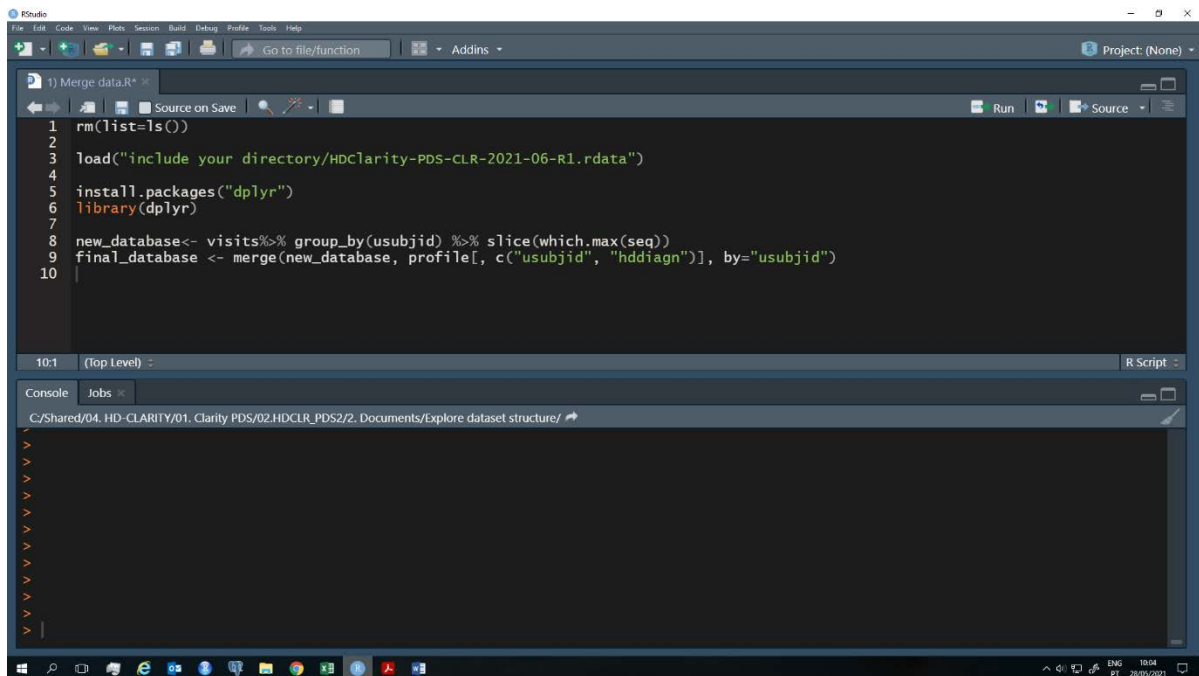
```
install.packages('dplyr')
```

```
library(dplyr)
```

```
new_database<- visits%>% group_by(usubjid) %>% slice(which.max(seq))
```

- Use the `merge()` function to merge the variable `hddiagn` from `profile` file. Example:

```
final_database <- merge(new_database, profile[, c("usubjid", "hddiagn")],  
by="usubjid")
```



```
1) Merge data.R
1 rm(list=ls())
2
3 load("include your directory/HDClarity-PDS-CLR-2021-06-R1.rdata")
4
5 install.packages("dplyr")
6 library(dplyr)
7
8 new_database<- visits%>% group_by(usubjid) %>% slice(which.max(seq))
9 final_database <- merge(new_database, profile[, c("usubjid", "hddiagn")], by="usubjid")
10
```

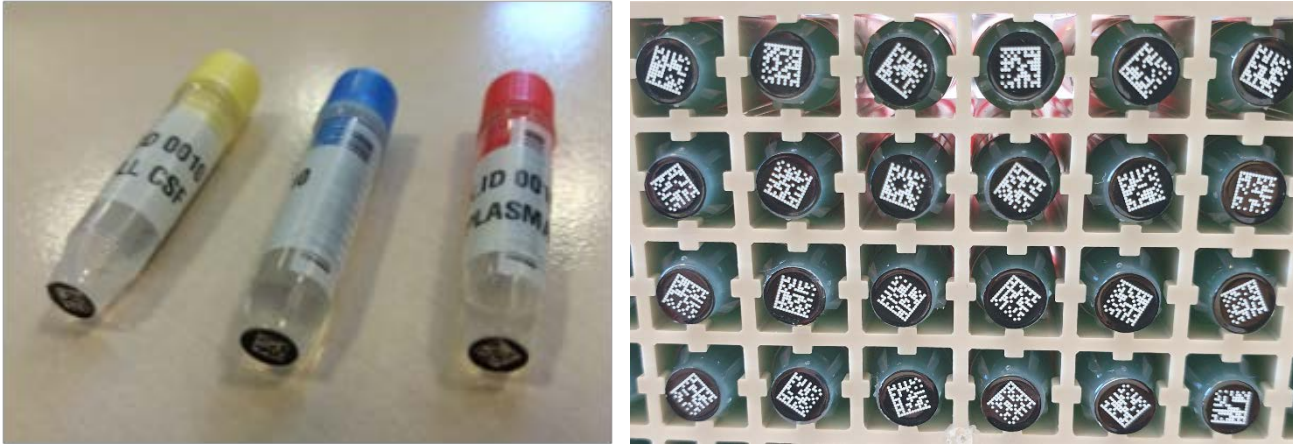
Note: Each user may use different codes to reach the same results, this is just an example. We recommend the user read and follow the guidelines available in:

R Core Team (2020). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL <https://www.R-project.org/>.



## Linking HDClarity samples and clinical data

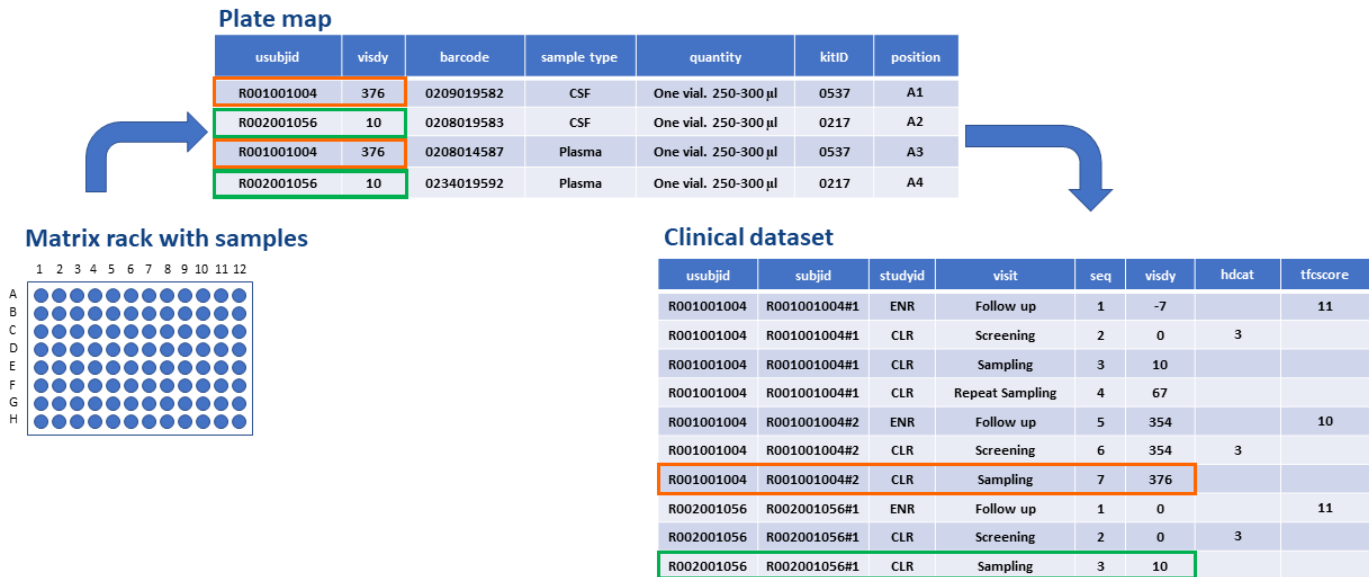
All samples collected in HDClarity are aliquoted into matrix vials that each have a unique **2D barcode** at the bottom and a sticker with the **KitID** on the side that can be read manually (see Figure 4).



*Figure 4. HDClarity biosample vials. Kit IDs are provided on stickers wrapped around the body of the vial (left) and 2D barcodes are provided on the vial base (right).*

The KitID can be used to unequivocally identify which participant and visit a sample came from since it is the same for **all samples** (CSF, plasma, serum, cells from CSF) collected at a **specific visit**, but is unique for each visit.

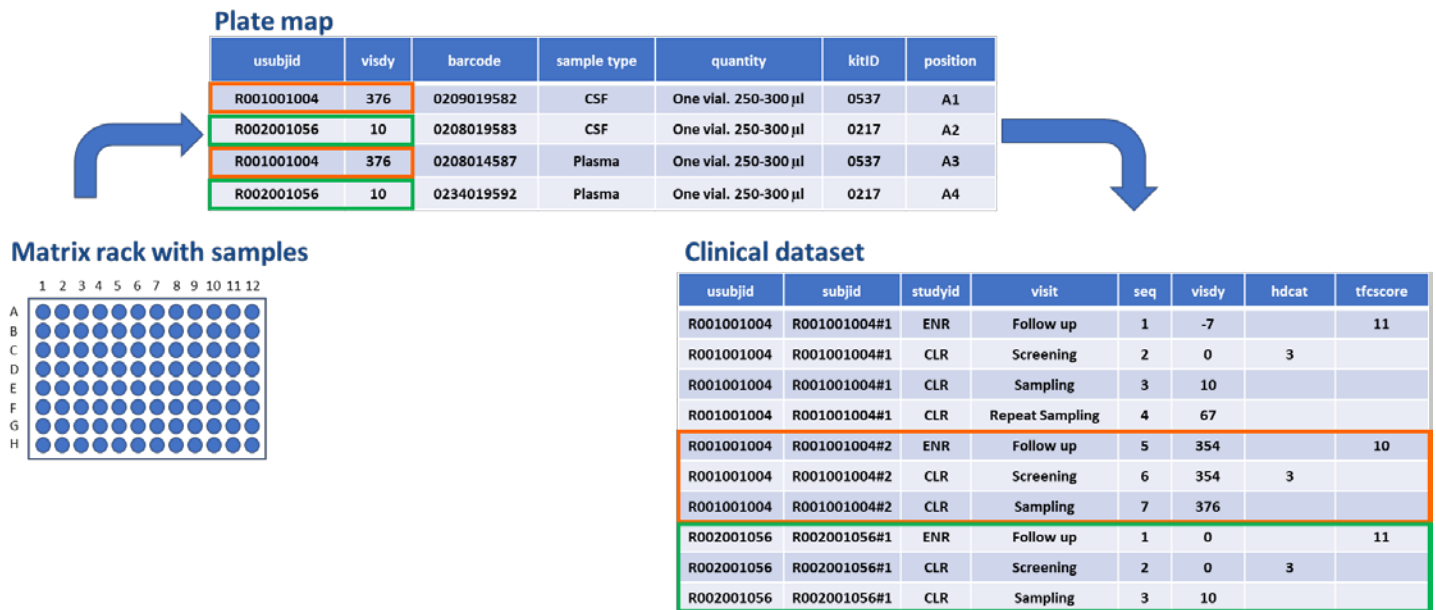
HDClarity samples are shipped in matrix racks (typically 8 x 12 format), which allows automatic registering of the complete rack of samples if requesters have access to a 2D barcode scanner. If a requester does not have access to a 2D barcode scanner, then the KitIDs can be used to manually register each sample.



**Figure 5. Plate maps.** The plate map allows the requester to link a sample in a specific position in a matrix rack to a sampling visit in the clinical dataset.

When HDClarity samples are shipped, a plate map (Figure 5) is provided to the requester that includes the participant ID (*usubjid*), the *visdy* variable value of the sampling visit at which the samples were collected, the sample type (CSF, plasma, serum, cells from CSF), the sample barcode (2D barcode at the bottom of the tube), the estimated volume in the tube (quantity), the KitID, and the shipping box position. The plate map allows the requester to link a sample stored in a specific position in a matrix rack to the clinical data. The following steps should be followed to link data and samples:

- 1) Identify the participant ID (*usubjid*) and visit day (*visdy*) for a given sample in the **plate map** (Figure 5).
- 2) Identify the HDClarity sampling visit that corresponds to the same participant ID and visit day in the visit file of the **clinical dataset** (Figure 5). Clinical information collected at the sampling visit (weight, height, TMS) will be listed on the same row, but since the vast majority of clinical variables are collected either at the HDClarity screening visit or Enroll-HD source visit, the *subjid* should be used to identify the HDClarity screening visit and Enroll-HD source visit that belong to the same “HDClarity visit package” since these will be the relevant visits to use (Figure 6). The clinical dataset in Figure 6 shows that participant R002001056 had a sampling visit at day 10, but the TFC score of “11” was assessed at the Enroll-HD visit that took place 10 days prior to the sampling visit (*visdy*=0).



*Figure 6. Linking biosamples and clinical data. The variable subjid is used to identify the CLR screening visit and ENR source visit that belong to the same HDClarity “visit package” as the sampling visit and thereby enables linkage to the clinical variables.*