Understand and Interpret the Data

Contents

1. HD classification

The Enroll-HD dataset contains the variables *hdcat, hdcat_0 and hdcat_l.* These variables refer to the subject group (HD category) of the participant at different points in time.

Variables *hdcat_0* and *hdcat_l* are located in the *participation* data files and indicate the subject group at the time of the **baseline visit** evaluation (*hdcat_0*) and at the **most recent** evaluation (*hdcat_l*) included in the PDS.

The variable *hdcat* is included in the *enroll* and *registry* data files and denotes the subject group (HD category) of each participant at each study visit (note that in the *registry* file, *hdcat* is only available in R3, and not R2).

Values for *hdcat, hdcat_0* and *hdcat_l* are assigned by site staff, based on clinical signs and symptoms and genotyping performed as part of clinical care, independent of the Enroll-HD study (except for participants in the subject group 'genotype unknown' who are at risk without clinical symptoms whose gene status is unknown - see below for more details).

In the Enroll-HD periodic dataset (PDS) releases, there are four categorical response outcomes for *hdcat*, *hdcat_0,* and *hdcat_l*:

> **Pre-Manifest/-Motor-manifest HD** (*hdcat/hdcat_0/hdcat_l* = 2): Confirmed HD gene expansion carriers (HDGECs) without clinical features regarded as diagnostic of HD.
>
> **Manifest/Motor-manifest HD** (*hdcat/hdcat_0/hdcat_l* = 3): HD gene expansion carriers (HDGECs) with clinical features that are considered HD symptoms, in the opinion of the investigator.
>
> **Genotype Negative** (*hdcat/hdcat_0/hdcat_l* = 4): A first or second degree relative (i.e., related by blood) of a HD gene expansion carrier (HDGEC), who has undergone predictive testing for HD and is known *not* to carry the HD expansion.
>
> **Family Control** (*hdcat/hdcat_0/hdcat_l* = 5): Family members or individuals not related by blood to HD gene expansion carriers (HDGECs) (e.g., spouses, partners, caregivers).

For the purposes of PDS releases, participants classified in the Enroll-HD study as **Genotype Unknown** (*hdcat/hdcat_0/hdcat_l* = 1; i.e., first- or second-degree blood relatives of a known HDGEC, who have not undergone predictive testing for HD and therefore have an undetermined HD gene status) are reclassified as Manifest, Pre-manifest, or Genotype Negative based on the genetic research testing (a CAG determination made at a central Enroll-HD lab, indicated by *caghigh*) and based on the Diagnostic Confidence Level (*diagconf*)

reported by the investigator for the participant. The following rules are used in the reclassification of genotype unknowns:

Reclassify as Genotype Negative: research genotype larger CAG allele (*caghigh*) < 36

Reclassify as Pre-manifest: research genotype larger CAG allele (*caghigh*) ≥ 36 and Diagnostic Confidence Level from the UHDRS motor (*diagconf*) < 4

Reclassify as Manifest: research genotype larger CAG allele (*caghigh*) ≥ 36 and Diagnostic Confidence Level from the UHDRS motor (*diagconf*) = 4

Data on participants categorized as Genotype Unknown, not reclassified, may be obtained through **special request**, subject to Scientific Review Committee (SRC) approval. Please refer to the **Access Data and Biosamples** webpage on www.enroll-hd.org for information on how to request a specified dataset (SPS).

Note that investigators and participants are *blinded* to the results of central Enroll-HD genetic research testing (i.e., *caghigh*) and reclassification. In other words, all Enroll-HD participants are tested at a central lab, but these results are not communicated to the participant or to site staff. The only 'gene status' known by the participant and the respective site is based on *local* CAG testing, not Enroll-HD testing.

Community Controls (*hdcat/hdcat_0/hdcat_l* = 6) are excluded from the dataset.

Note that the *hdcat* variables are available for the studies Enroll-HD and REGISTRY 3 but are **not available** for **REGISTRY 2** and **Ad Hoc** since these studies did not use an HD classification system.

For further information on HD disease onset, diagnosis, staging and disease severity, and how these important concepts are captured in the Enroll-HD PDS, please refer to the following sections: ***HD Onset and Diagnosis Variables*** *and* ***HD Integrated Staging System***

2. Missing values

There are two overarching categories of missing data in the dataset: **system-defined** missing data (indicated by blank variable 'entries'), and **user-defined** missing data (indicated by specific codes, which indicate reason for the missing data).

System defined missing data

System defined missing data occurs where the electronic data capture (EDC) system dictates a missing variable field. These missing data values are indicated by **blank entries** in the dataset.

These instances arise if there is a dependency of specific question to the response to a '**parent**' question. For example, a 'no' response to the 'parent' question "Has the participant ever smoked?" (*hxtobab*) will result in a blank cell for response to the 'child' question "cigarettes per day?" *(hxtobcpd).*

In addition, **total scores** for assessments may also display as blank entries. This will occur where a **mandatory assessment item**, required for the calculation of the total score, is **missing**. Total scores are automatically generated by the 'system' if all necessary values are available.

Please note that blank entries may be converted to another value dependent on statistical software package (e.g., 'NA' in R).

User defined missing data

User defined missing data occur where a mandatory variable field, as determined by the EDC system, is **not completed**, or where the value entered into the EDC is **incorrect**. In these instances, data entry users are prompted to indicate why the value is missing, or why the value entered is not correct. These user-defined labels - 'exceptional values' – are listed below. Each one is represented in the dataset by a **specific code**:

- **Unknown** (entered by the site, only available for specific fields): 9999 (numeric); UNKNOWN (text)

Refers to mandatory values which are **occasionally unknown**. This exceptional value code may be selected as a response to the question, "Is/was your Mother affected by HD? (response: yes/no/unknown)", where a participant did not know their mother.

- **Missing** (value expected, but not entered): 9998 (numeric); MISSING (text); 9998-09-09 (date)

  Refers to mandatory values which could not be completed because **data collection was not performed**. This code may be used if a participant refuses to provide a response, if the collection of data was accidentally omitted, or if a value could not be obtained because required instrumentation was not available.

- **Not applicable** (value expected, but not entered): 9997 (numeric); NOTAPPL (text); 9997-09-09 (date)

  Refers to mandatory values which could not be completed because they **do not apply to the participant due to certain circumstances or characteristics**. For example, the question "Age at onset of symptoms in mother" for the mother who is still premanifest and does not have symptoms yet shall be answered as not applicable. The variable value is **purposefully** not entered. **Note the distinction between this value and system-defined missing values is that the *user*, as opposed to the EDC *system*, marks them as non-applicable**.

- **Wrong** (value was entered but declared as wrong by the site. Entered value excluded from dataset): 9996 (numeric); WRONG (text); 9996-09-09 (date)

  Refers to mandatory values which are entered into the EDC and then identified **to be wrong or highly questionable**. This may be because data were collected by the wrong person (e.g., assessment performed by untrained site member), faulty instrumentation (e.g., uncalibrated weighing scales), etc. Although these data are not technically missing, they are recoded as wrong using the codes indicated above for PDS releases.

3. Imputation

Imputation in the PDS is limited to the following instances:

HD-ISS variables

HD-ISS variables and input variables required for HD-ISS variable imputation (see ***HD Integrated Staging System***).

Date values

Imputation is performed for date variables in instances where an incomplete date has been provided (e.g., if the month and year are known, but not the day), according to the rules indicated in the section ***Date Values***. Note that because of these imputation (autocompletion) rules, events with clear temporal definition sometimes appear **out of sequence** or have the **same date** values. For example, the number of days between a medication start date and end date or comorbidity start date and end date may be zero or a negative number.

BMI

The BMI variable provided in the PDS (i.e., *bmi_imp*) is an imputed value for all visits except the Enroll-HD baseline visit. Imputed BMI values are based on the weight value observed at a specific visit, and *height as observed at Enroll-HD baseline*. This is to avoid fluctuations in BMI driven by unexpected/implausible variation in height, which are observed in Enroll-HD data, but cannot be observed by end-users as height (and weight) are not included in the PDS for identification risk purposes. BMI is set to system-defined missing (blank) for all visits where the participant is under the age of 18 years. Unimputed BMI (i.e., *bmi*), weight, and height, are available for all participants at all visits via Specified Dataset (SPS) request.

4. Date values

Transformation of date values

To minimize participant identification risk, the Enroll-HD PDS does not contain date values. Date values referring to visit dates are transformed to a **numeric value**, reflective of the number of days between the Enroll-HD **baseline visit date** and the date of interest. Date values that refer to date of birth or symptom onset are transformed into age values.

Note that date values are **negative** if the date refers to a point in time before Enroll-HD enrollment. This is typical for start dates of medications and comorbid conditions, and visit dates in other studies (e.g., Registry).

For example, date values for a participant with a baseline enrollment date of 2020-11-01 (YYYY-MM-DD) would read as follows:

Entered date	Representation in dataset
2020-11-01	0
2020-11-30	29
2020-10-31	-1

Imputation of date values/Autocompletion

Incomplete date values were **imputed** according to the following rules:

Day missing: YYYY-MM-***DD***(missing): YYYY-MM-15

Month and day missing: YYYY-***MM***(missing)-***DD***(missing): YYYY-07-01

For example, date values for a participant with a **baseline enrollment date** of 2020-11-01 (YYYY-MM-DD) would read as follows:

Entered date e.g. medication start	Imputed date	Representation in dataset
2020-11-01	N/A	0
2020-11	2020-11-15	14
2020	2020-07-01	-123

Note that because of these imputation rules, events with clear temporal definition sometimes appear **out of sequence** or have the **same date** values. For example, end dates may appear prior to, or on the same day as, start dates for comorbidities and pharmacotherapies. For example:

Entered start date	Imputed start date	Entered end date	Imputed end date	Date differential (days)
2020-11-01	N/A	2020-11	2020-11-*15*	14
2020-11	2020-11-*15*	2020-11	2020-11-*15*	0
2020	2020-*07-01*	2020-06-15	N/A	-16

An additional variable containing date value **precision** information (d, m, and y) can be obtained through **special request**, subject to Scientific Review Committee (SRC) approval. Please refer to the **Access Data and Biosamples** webpage on www.enroll-hd.org for information on how to request a specified dataset (SPS). The precision variable identifies the level of date completeness:

ymd – for a complete date (precision "days") i.e., YYYY-MM-DD

ym – if day information is missing (precision "months") i.e., YYYY-MM-***DD***(missing)

y – if day and month information is missing (precision "years") i.e., YYYY-***MM***(missing)-***DD***(missing)

5. Aggregated values

To minimize participant identification risk, **data aggregation techniques** are applied to specific variables for PDS releases. These variables, and the criteria/thresholds used for aggregation in the current PDS release, are described in the table below.

Note that aggregation **thresholds may differ between PDS releases**. Changes in the Enroll-HD cohort size and profile allow for such aggregation threshold adjustments while maintaining low identification risk thresholds.

Note that **numerical values with aggregated data have been converted to string values** (e.g., possible entry for *caghigh* = '>70'). Cells that contain **'>'** or **'<'** values should be **replaced by the end user with a numeric value for analysis**.

Deaggregated or suppressed data may be obtained through **special request**, subject to Scientific Review Committee (SRC) approval. Please refer to the **Access Data and Biosamples** webpage on www.enroll-hd.org for information on how to request a specified dataset (SPS).

Table: Aggregated variables and aggregation thresholds in PDS6.

Data file	Variable	Variable label	Criteria for aggregation
participation	*age_0*	Age at enrollment	<18
enroll, registry, adhoc	*age*	Age at visit	<18
profile	*caghigh*	Research larger CAG allele determined from DNA	>70
profile	*caglow*	Research smaller CAG allele determined from DNA	>28
profile	*race*	Ethnicity	Fewer than 100 cases per category*

* The following categories for ethnicity are aggregated into "Other (6)": "Native Hawaiian or Other Pacific Islander" (4), "Alaska Native/Inuit" (5), "African – South" (11), "African – North" (12), "Other" (6). The categories "Asian – West" (13) and "Asian – East" (14) are aggregated into the category "Asian" (16).

Table: Participants subject to aggregation thresholds in PDS6.

Data file	Variable	Label	Number of participants
participation	*age_0*	<18	38
enroll	*age*	<18	38
profile	*caghigh*	>70	38
profile	*caglow*	>28	332
profile	*race*	Other (6)*	390
profile	*race*	Asian (16)**	162

* Includes individuals from the following categories: "Native Hawaiian or Other Pacific Islander" (4; N=7), "Alaska Native/Inuit" (5; N=4), "African - South" (11; N=20), "African - North" (12; N=51), and "Other" (6; N = 308)
** Includes individuals from the following categories: "Asian – West" (13; N=94) and "Asian – East" (14; N=68)

6. Assessment score calculation

Assessment '**total scores**' are automatically calculated in the Enroll-HD EDC system.

If a **mandatory assessment item**, required for the generation of the **total score**, is **missing**, a blank data entry will be displayed (indicative of **system defined** missing data). Note that incomplete total scores (calculation of scores with the available values) or detailed items are also available for some assessments (motor, function).

Unified Huntington's Disease Rating Scale (UHDRS ®)

Enroll-HD PDS releases contains calculated composite UHDRS scores. Please refer to the following reference for further information:

> Huntington Study Group. Unified Huntington's Disease Rating Scale: Reliability and Consistency. Neuropsychiatry Movement Disorders 1996, Vol. II, No. 2, 136-142.

The motor component of the UHDRS® assesses domains such as chorea, dystonia, bradykinesia, and rigidity. The key disease variable generated by this assessment is total motor score (*motscore*), which ranges from 0 to 124.

The UHDRS® Motor/Diagnostic Confidence component, indicates rater's confidence in patient's motor onset, based on UHDRS motor assessment above (*diagconf*). This variable ranges from 0 (no abnormalities) to 4 (motor abnormalities that are unequivocal signs of HD; ≥99% confidence).

The total functional capacity (TFC) component of the UHDRS® consists of five items: occupation, finances, domestic chores, activities of daily living, and care level. The key disease variable generated by this assessment is total functional capacity score (*tfcscore*), which ranges from 13 (least severe) to 0 (most severe).

The functional assessment (FAS) component of the UHDRS® includes 25 yes/no questions about common daily tasks. The key variable generated by this assessment is functional assessment score (*fascore*), which ranges from 25 to 0.

The independence component of the UHDRS® assesses the participant's independence. The single independence scale score (*indepscl*) is a percentage ranging from 100% where no special care needed to 5% where participant is tube fed and needs total bed care.

Table. *UHDRS score calculation for multi-item component sections.*

UHDRS section	Variable	Score calculation
Motor	*motscore*	Sum of the values of all scores

Total Functional Capacity	*tfcscore*	Sum of the values of all scores
Functional Assessment	*fascore*	Sum of the values of all scores

Problem Behaviors Assessment – Short (PBA-s)

Enroll-HD contains calculated composite PBA-s scores. Please refer to the following reference for further information.

> Craufurd D, Thompson JC, Snowden JS. Behavioral changes in Huntington Disease. Neuropsychiatry Neuropsychol Behav Neurol. 2001 Oct-Dec;14(4):219-26.

This instrument measures frequency and severity of symptoms related to altered affect, thought content, and coping styles. It includes items that cover an extensive range of behaviors including: depressed mood, low self-esteem, anxiety, suicidal thoughts, aggressive behaviour, irritability, perseveration, compulsive behaviours, delusions, hallucinations, and apathy. Key disease variables are total score on each sub-scale (*depscore, irascore, psyscore, aptscore, exfscore*).

Table. *PBA-s sub-scale score calculations.*

PBA-s section	Variable	Score calculation
Depression	*depscore*	Addition of composite scores[§] for depressed mood + suicidal ideation + anxiety
Irritability/Aggression	*irascore*	Addition of composite scores[§] for irritability + angry or aggressive behavior
Psychosis	*psyscore*	Addition of composite scores[§] for delusions / paranoid thinking + hallucinations
Apathy	*aptscore*	Addition of composite score[§] for apathy
Executive function	*exfscore*	Addition of composite scores[§] for perseverative thinking or behavior + obsessive compulsive behaviors

[§] These composite scores are calculated by **multiplying severity by frequency** for each symptom, which are then **summed** to create a composite score. For example: Depression = (severity of depressed mood*frequency of depressed mood) + (severity of suicidal ideation*frequency of suicidal ideation) + (severity of anxiety*frequency of anxiety).

Mini Mental State Examination (MMSE) score

Enroll-HD contains calculated MMSE scores. Please refer to the following reference for further information:

> Folstein MF, Folstein SE, McHugh PR. "Mini-mental state". A practical method for grading the cognitive state of patients for the clinician. Psychiatr Res. 1975 Nov;12(3):189-98

The Mini Mental State Examination is an 11-question measure that tests five areas of cognitive function: orientation, registration, attention and calculation, recall, and language. The key variable generated is MMSE score (*mmsetotal*), calculated by summing the value of all assessment scores.

Hospital Anxiety Depression Scale / Snaith Irritability Scale (HADS-SIS) score calculation

The HADS-SIS assessment used in Enroll-HD is a combination of **two separate scales**, the Hospital anxiety and depression scale - **HADS** (Zigmond & Snaith, 1983) and the Snaith irritability scale -**SIS** (Snaith, 1978). It is important to recognize that the HADS-SIS is comprised of these two separate scales so that analyses can incorporate the respective subscales and items appropriately.

The HADS combined with the SIS offer a brief rating of depression, anxiety, and irritability (inward and outward) symptoms that reflect primarily mood rather than cognitive and somatic symptoms. Key variables are subscale total scores (*anxscore, depscore, irrscore, outscore, inwscore*).

The following reference provides further information on score calculation:

> Zigmond AS, Snaith RP. The hospital anxiety and depression scale. Acta Psychiatr Scand. 1983 Jun;67(6):361-70.
>
> Snaith, RP. A clinical scale for the self-assessment of irritability. Brit J Psychiat. 1978; 132: 164-171.

Short Form Health Survey - 12v2 (SF-12)

Enroll-HD contains calculated scores for SF-12 scales available in the ‘enroll’ data file.

The Short Form Health Survey-12 (SF-12) is extensively used in large population health surveys as a brief, reliable measure of overall health status. The 1-week recall version is used. Key variables are: group norm-based scores for physical functioning (*pf*), role-physical (*rp*), bodily pain (*bp*), general health (*gh*), vitality (*vt*), social functioning (*sf*), role-emotional (*re*), mental health (*mh*), all of which generate an overall physical component (*pcs*) and mental component (*mcs*).

The following reference provides further information on score calculation:

Ware JE, Kosinski M, and Keller SD. A 12-Item Short-Form Health Survey: Construction of scales and preliminary tests of reliability and validity. Medical Care, 1996; 34(3):220-233.

Short Form Health Survey – 36 v1/v2 (SF-36)

Enroll-HD PDS releases contains the total scores for the SF-36 scale (version 1 and version 2) available in the '*registry*' datafile. All sub-items for this scale are available upon SPS request.

The following reference provides further information on global score calculation:

Ware JE Jr, Sherbourne CD. The MOS 36-item short-form health survey (SF-36). I. Conceptual framework and item selection. Med Care. 1992 Jun; 30(6):473-83.

7. Derived variables (e.g., periodic dosage; drugs, pharmacotherapies, nutritional supplements)

Data on the use of drugs, pharmacotherapy, and nutritional supplements and/or periodic dosage are included in the Enroll-HD PDS.

The variable *cmdostot* is **derived** from raw measures of **dose** and **frequency of use**, i.e., multiplication of a drug dose by the number of intakes per day (for example: 25 mg taken 4 times per day equals 100 mg intake per day).

If a drug is taken "as needed" the frequency of use is often unknown and set to zero, thus the derived value is then zero (e.g., 25 mg taken at 0 times per day equals an intake of 0 mg per day). If the values for frequency of use are set to one of the exceptional values (e.g., 9998), the variable *cmdostot* is also set to that exceptional value e.g. 9998.

For combined drugs the dose is often not entered as a number, but rather as a string (eg. 25/100 mg). It is not possible to derive the total daily dose from this, and the value *cmdostot* remains blank.

A large number of unusually high values are also observed for *cmdostot.* These values may be correct, and attributable to small units of measurement listed for dose, or reflective of data entry errors. These values are *not* queried by the Enroll-HD team, and as such we highlight the need for careful review before use in analysis.

Another example for derived value is the variable '*packy*', indicative of an individual's cumulative lifetime exposure to tobacco in terms of pack/years. It is derived from daily intake (*tobcpd*) and years of smoking (*tobyos*) variables (*packy* = [*tobcpd*/20] * *tobyos*).

If one of these **input values** is **missing**, a system-defined missing value will be generated (see ***Missing Values***). If one of these values is **extremely low** or **zero** (as may be entered for *tobcpd* for occasional, non-daily smokers), the derived value may be zero, due to either a zero entry for input value, or rounding down of derived value of <0.05. Total number of participants with low *packy* values are identified in the ***Quality Control: Observations and Unusual Findings*** document provided along with the dataset.

The raw data values used to calculate dose can be obtained through **special request**, subject to Scientific Review Committee (SRC) approval. Please refer to the **Access Data and Biosamples** webpage on www.enroll-hd.org for information on how to request a specified dataset (SPS).

8. Data exclusion

For data to be included in the PDS, certain requirements must be met at both a participant- and visit-level basis.

If participant-level requirements are not met, that participant and all their associated data will be excluded from the PDS release. If participant-level requirements are met but visit level data requirements are not met for that participant for one or more visits, the participant will be included in the PDS, *but one or more of their study visits will be excluded*.

Participant level data requirements:

- Participant status is not 'quarantined'
- Existing value for *caghigh* (i.e., research CAG)
- Valid baseline value for *hdcat* (i.e., HD category at enrollment)
- Value for *hdcat* does not equal 6 (i.e., no community control)
- If participant status is 'withdrawn' or 'violator', an End form must be completed
- Central coding completed for pharmacotx (medications, indications), comorbidities and events
- Study visit status for baseline visit is 'completed' (i.e., onsite monitoring of baseline is complete)
- General data monitored at least once
- Participant is not on the 'excluded participants' list generated by the statistical monitoring team
  Reasons for exclusion include: exceeding identification risk threshold; *caghigh* is not consistent with *hdcat*

Visit level data requirements:

Enroll-HD (*enroll*) visits (applies to follow-up, unscheduled, phone contact):

- Study visit status is 'completed'
- Visit is not on the 'excluded visits' list. Reasons for exclusion include: duplicated visits/visits with the same visit date; visits not covered by a valid ICF

REGISTRY (*registry*) visits:

- Study visit status is 'completed', 'signed', or 'reviewing'
- Visit is not on the 'excluded visits' list

AdHoc (*adhoc*) visits:

- Study visit status is 'completed', 'signed', or 'reviewing'

- Visit is not on the ‘excluded visits’ list

9. Quality control: observations and unusual findings

Prior to each PDS release, an enriched set of remote data quality control checks are performed. These include custom checks for **unusual** or **implausible** values, and systematic checks of continuous variables for **extreme outlying values,** which are flagged using data-driven thresholds or pre-specified custom thresholds based on plausibility.

Unusual and implausible values are reviewed by the monitoring and/or medical monitoring teams and queried directly with sites where deemed appropriate by expert determination. In certain instances, however, these values **cannot be queried** and corrected (e.g., if the observation was recorded in a REGISTRY visit and then transferred into the Enroll-HD database) or are **queried and confirmed as correct by site staff**. In instances such as these, those unusual values are provided 'as is', and it is left to the analyst to determine whether to include or exclude these values or perform sensitivity analyses.

The *Quality Control: Observations and Unusual Findings* document lists all quality control checks that are performed for PDS releases, alongside frequency counts of how many identified values remain 'as is' for each individual check.

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.**

10. HD onset and diagnosis variables

HD **onset** and **diagnosis** are important clinical concepts and of critical interest to many researchers. In this section, we discuss these complex concepts in detail – and the nuances of the variables which capture them in the Enroll-HD study.

HD onset is complex. The timing of symptom onset, order of presentation, and consequent trajectory of symptoms in each domain - motor, cognitive, functional, or behavioural - are unique to each participant. Similarly, diagnosis of HD may be made for an individual in different ways at different times. In recognition of this, Enroll-HD collects data on a multitude of variables relating to timing of initial symptoms, disease onset, genetic testing and diagnosis. These are shown in Table 1.

Table 1. Key disease dates (symptoms, onset, diagnosis) captured in the Enroll-HD EDC. Note analogous age of onset variables are available in PDS releases.

Disease date domain	Variable	Variable Label	Form
Date of first symptoms	*sxsubj*	Date symptoms first noted by participant	HDCC
	sxfam	Date symptoms first noticed by family	HDCC
	sxrater	Rater's estimate of symptom onset	HDCC
	ccdepyr	Year of onset (depression)	HDCC
	ccirbyr	Year of onset (irritability)	HDCC
	ccvabyr	Year of onset (violent or aggressive behavior)	HDCC
	ccaptyr	Year of onset (apathy)	HDCC
	ccpobyr	Year of onset (perseverative/obsessive behavior)	HDCC
	ccpsyyr	Year of onset (psychosis)	HDCC
	cccogyr	Year of onset (cognitive impairment; first began impacting on daily life)	HDCC
	ccmtryr	Year of onset (motor symptoms)	HDCC
Date of diagnosis	*lbdtc*	Date of report (local CAG) NB: this variable is not available in dataset releases	CAG

	svstdtc & *diagconf*	Date of visit at which diagnostic confidence level (DCL) is updated from '1', '2', or '3' to '4.' NB: Indicates **disease onset, motor**	Variable items (Follow-up visit); Motor
	svstdtc & *hdcat*	Date of visit at which *hdcat* is updated from 'premanifest' to 'manifest.' NB: Indicates **disease onset, any domain**	Variable items (Follow-up visit);
	hddiagn	Date of clinical HD diagnosis (**based on symptoms in any domain**) NB: Indicates **disease onset communicated to participant**	HDCC

In Enroll-HD, dates relating to onset of **first symptoms** are captured as reported from several perspectives: the participant (*sxsubj*), their family *(sxfam)*, and the Enroll-HD clinician/rater (*sxrater*).

Note that since the EDC release in December 2017 the variable ***sxrater*** can only be completed when a **participant is considered manifest, as indicated by *hdcat*=3**. This rule does not apply for *sxsubj* and *sxfam* which can be completed regardless of *hdcat* value. Values collected before December 2017 may still be present for some participants not considered manifest.

Onset dates pertaining to **specific symptoms** in each domain are also captured (e.g., *cccogyr)*. These are completed from the clinician/rater's perspective, based on their best judgement. This considers participant and family reports, available history from medical records, as well as Enroll-HD assessment scores.

Note that of almost all '*cc….*' symptom onset variables, **only symptoms in the motor domain are required to be HD specific**. Psychiatric symptoms or cognitive symptoms indicated by these variables **may or may not be related to HD** and should be considered with this caveat in mind.

Given the exclusion of certain participant visits from the PDS (see ***Data Exclusion***) it is possible for the **age** listed for these **onset** variables **to be greater than the age at 'last' visit**.

The term '**clinical diagnosis**' is used to denote the unequivocal onset of symptoms or signs attributed to HD, which can occur at vastly different times for each individual gene-expanded carrier. In the Enroll-HD protocol, a clinician-

based judgement of disease "manifest" status, as indicated in Enroll-HD by participant category (i.e., *hdcat = 3)*, is based on symptoms in *any* of the disease domains (i.e., motor, cognitive, behavioral). To this point, note that some participants classified as manifest in Enroll-HD may have low values for UHDRS total motor score (e.g., *motscore* < 10 and a DCL of < 4). In these instances, the manifest categorization may be due to psychiatric or cognitive symptom onset as opposed to motor.

Enroll-HD captures the **date of clinical HD diagnosis** (*hddiagn*). This variable represents the date on which a participant is informed by a clinician that the disease is evident. However, this can be years after actual symptom onset if the participant has not been seen by a doctor. If the date of first diagnosis is unknown and cannot be identified, *hddiagn* can be missing, even if a clinician is confident in their diagnosis of symptomatic HD and has correspondingly marked *hdcat* as manifest. If the field date of clinical HD diagnosis is filled for participants classified as "pre-manifest", it is possible that the date of predictive genetic test result has erroneously been entered instead of date of clinical HD diagnosis. The other possibility is that the subject group was not entered correctly for the participant. The analyst should decide whether such values should be excluded from their analysis.

An alternative definition of disease onset, also termed "manifest" and widely used in the HD literature, concerns the transition from pre-symptomatic to symptomatic HD based on *motor symptoms only*; this is known as **motor onset** or **motor diagnosis**. This definition is based on a Diagnostic Confidence Level (DCL) score (i.e., *diagconf*) of 4, which indicates a clinician's confidence that, based on the UHDRS Motor assessment, motor signs unequivocally represent HD (≥ 99% confidence). Provided a participant was not classified as *hdcat* = 'manifest', or *diagconf* = '4', at study entry (i.e., at baseline visit), the date of the visit at which either of these variables are updated to the values above can be used to indicate date of clinical onset, as outlined respectively above.

Note that there may be **discrepancies** between **date of clinical diagnosis** (*hddiagn*) and **year of onset of motor symptoms** (*ccmtryr*). Estimation of onset of motor symptoms may be years **earlier** than date of clinical diagnosis, for example if a participant had not been seen by a doctor for a long period. Further, clinical **diagnosis** is also very distinct from first **symptoms**. Conversely, onset of motor symptoms may be years **later** than date of clinical diagnosis. This is plausible if the diagnosis was based on cognitive or psychiatric symptoms.

Genetic diagnosis of HD is performed by genetic test, which may be completed prior to symptom onset (known as a "**predictive test**"), or to confirm a clinical diagnosis (known as a "**diagnostic test**"). Diagnostic or predictive genetic testing is voluntary. Such genetic tests are completed at local labs for some individuals participating in

Enroll-HD (not all) and are performed independently of the Enroll-HD study. Separately, all Enroll-HD participants undergo research CAG repeat genotyping at a central research laboratory. These results are used solely for *research* as opposed to predictive or diagnostic purposes, and are never shared with participants, investigators, or sites. For the Enroll-HD study, an individual is an HD gene expansion carrier if they have 36 or more CAG repeats, although the literature states that CAG repeats between 36 and 39 (inclusive) are not fully penetrant. CAG values between 27 and 35 (inclusive) are considered intermediate alleles. All CAG repeat lengths of 40 and higher are fully penetrant. In symptomatic individuals *without family history of HD*, clinical diagnosis is confirmed by genetic testing; therefore, date of local genetic testing (i.e., *lbdtc*) may be used as "date of clinical diagnosis" in such individuals. In asymptomatic individuals, with family history, who undergo predictive testing, date of genetic testing may be used as "date of genetic diagnosis". *Note however that date of local genetic testing is not made available in Enroll-HD data releases*.

Finally, we highlight **CAP score** (CAG-age-product), which is a commonly used measure of cumulative exposure to mutant huntingtin. Multiple formulas for calculating CAP score exist. In the current Enroll-HD PDS release, we include CAP score (*capscore*) for applicable participants and visits, calculated per the definition provided by Warner et al.[1]

$$\text{CAP score} = \text{Age} \times (\text{CAG} - L)/K, \text{ where } L = 30 \text{ and } K = 6.49$$

This formula is standardized such that CAP = 100 at the expected age of onset.

Please note the following caveats associated with the CAP score values provided in the current Enroll-HD PDS release:

- CAG length is based on research CAG (i.e., *caghigh*)
- CAP score is calculated for each participant at each visit, provided CAG length (*caghigh*) is >/= 36
- Age is entered into the above formula as an **integer** (i.e., a whole number, without decimal points)
- Where **age** and/or **CAG values** are aggregated, **CAP score** will be **blank** (i.e., system defined missing)

CAP score values are provided for all visits in all studies included in the PDS (i.e., Enroll-HD, R2, R3, and Ad-hoc). These values are located in the following files*: enroll*, *registry*, *adhoc*.

[1] Warner, J. H., Long, J. D., Mills, J. A., Langbehn, D. R., Ware, J., Mohan, A., & Sampaio, C. (2022). Standardizing the CAP Score in Huntington's Disease by Predicting Age-at-Onset. Journal of Huntington's disease, 11(2), 153–171. https://doi.org/10.3233/JHD-210475

Quality Control of HD onset and diagnosis variables: *A specific set of custom multivariate quality control checks are performed on the HD onset variables prior to PDS releases in an effort to identify unusual or implausible values. These values are reviewed by the monitoring and/or medical monitoring teams and queried directly with sites where deemed appropriate by expert determination. In certain instances, however, these values cannot be queried and corrected (e.g., if the observation was recorded in a REGISTRY visit transferred into the Enroll-HD database) or are queried and confirmed as correct by site staff. In instances such as these, values may be provided 'as is', and it is left to the analyst to determine whether to include or exclude these values. These custom HD onset checks, alongside a summary of findings, are listed in the* ***Quality Control: Observations and Unusual Findings*** *document.*

11. HD Integrated Staging System (HD-ISS)

The current Enroll-HD PDS release includes imputed Huntington's disease Integrated Staging System[2] – or HD-ISS - variables.

Background

The Huntington's Disease Integrated Staging System (HD-ISS) is a four-stage evidence-based framework intended to facilitate clinical research. In Stage 0, individuals have the Huntington's disease genetic mutation (CAG ≥ 40) without any detectable pathological alterations. Stage 1 is marked by measurable underlying biomarker pathophysiology as indicated by striatal atrophy. Stage 2 indicates the appearance of HD signs and symptoms, and Stage 3 is evidenced by functional decline. Staging requires the collection of CAG length (Stage 0), caudate and putamen volume (Stage 1), and the UHDRS variables of total motor score and symbol digit modalities test (Stage 2), and total functional capacity and independence scale (Stage 3). Enroll-HD collects all the variables except brain volume. The missing imaging variables indicate that participants in Enroll-HD cannot be definitively staged. For this reason, we impute HD-ISS stage using machine learning methods that are known to generally provide excellent predictions in applied data analysis. In addition, we provide probabilities of classification of all the stages for a visit, which might be used to represent the confidence level of the imputed stage.

HD-ISS in the PDS

The Enroll-HD PDS includes five HD-ISS variables for each pwHD participant at each visit, as listed below. The values of each of these variables are the result of the imputation methods described below. These values are provided in the 'enroll' data file only; HD-ISS variables are not calculated for Registry and Adhoc visits.

"HDISS_stage_imp"	The imputed HD-ISS stage; specific to participant and timepoint
"HDISS_stage0_prob"	The probability of classification as HD-ISS Stage 0; specific to participant and timepoint
"HDISS_stage1_prob"	The probability of classification as HD-ISS Stage 1; specific to participant and timepoint
"HDISS_stage2_prob"	The probability of classification as HD-ISS Stage 2; specific to participant and timepoint
"HDISS_stage3_prob"	The probability of classification as HD-ISS Stage 3; specific to participant and timepoint

[2] Tabrizi SJ, Schobel S, Gantman EC, et al. A biological classification of Huntington's disease: The integrated staging system. The Lancet Neurology 2022; 21: 632–644.

Exclusions

It is not possible to impute HD-ISS stage for every participant and/or visit. Imputation is only calculated for participants of age 18 or older who have CAG in the range of 40 to 50 (inclusive). In these instances when an imputed stage cannot be calculated, the variable values will be blank (i.e., system defined missing).

Why is imputation of Stages 2 and 3 necessary in the Enroll-HD PDS?

The Enroll-HD dataset contains data on total motor score, symbol digit modality test, total functional capacity, and independence scale – i.e., the landmark variables required to assess entry into Stage 2 and Stage 3 respectively. Given that these data are known for Enroll-HD participants, it is reasonable to ask why must classification as Stage 2 or Stage 3 be imputed? In short, because information on **all** landmark variables is required to definitively stage individuals. The absence of imaging variables in Enroll-HD prohibits **definitive** classification of participants, although we are able to impute Stages 2 and 3 with a higher degree of confidence than we are Stages 0 or 1 because of the presence of the Stage 2 and 3 landmark variables in the dataset. This is reflected in the HD-ISS stage probability scores displayed alongside the imputed stage.

How is regression in stages possible across visits?

The staging system – by design – is cross-sectional, only applying to a single visit for a person. There is no constraint in the staging system that considers prior or subsequent values for stage. Regression of stage values across visits – and stage skipping - are observed phenomena in the Enroll-HD PDS, and 'ground truth' datasets. This is due to a whole host of factors, including measurement error, and real temporal fluctuations. This is not a unique issue to the staging system – and is often seen in scores for other clinical measures such as the TMS and DCL.

Missing Data Imputation

HD-ISS stage imputation for the periodic data set (PDS) was based on random forest[3] with chained equations[4] as applied in the "missForest" algorithm[5] (the "R" package "missRanger"[6] was used). Because Enroll-HD does not collect imaging variables, a database from studies that did collect imaging (IMAGE-HD, PREDICT-HD and TRACK-HD/ON) was

[3] Breiman L. Random forests. *Machine Learning* 2001; 45: 5–32.
[4] Buuren S van, Groothuis-Oudshoorn K. Mice: Multivariate imputation by chained equations in r. Journal of Statistical Software 2011; 45: 1–67.
[5] Stekhoven DJ, Buehlmann P. MissForest - nonparametric missing value imputation for mixed-type data. Bioinformatics 2012; 28: 112–118.
[6] Mayer M. missRanger: Fast imputation of missing valueshttps://CRAN.R-project.org/package=missRanger (2021).

used to train the algorithm. Chained equations constitute a conditional specification approach to imputation. The imputation is performed on a variable-by-variable basis with each incomplete variable acting as the outcome variable and using all other variables in the imputation model as predictors. Each conditional imputation model is variable-specific, using the appropriate methods for the data type of the variable, whether it be continuous, binary, multi-category, etc. Thus, in our application all variables with missing data were imputed, not just the HD-ISS. However, only the imputed HD-ISS stages are provided in the PDS distribution, along with post hoc probabilities of classification, as described below. The chained equation approach has been shown to work well in simulation studies. The main advantage of the method is that a specification of the joint multivariate distribution for all the variables is not required. The multivariate distribution may be difficult or impossible to specify when the variables are a mix of types, as we have in Enroll-HD.

The "missForest" imputation algorithm proceeds as follows. Suppose we have variable vectors x_1, x_2, x_3, each with dimension n by 1. Assume the first two variables have missing data, and say that x_1 has less missing than x_2. We start with x_1, and set it to be the outcome variable, y, which will be predicted by x_2 and x_3. The algorithm initiates by making a naive guess for the missing data in x_2, using the mean or mode (depending on the predictor variable type). Then a random forest is grown, consisting of a large number of random regression trees (1000 trees were used in this PDS release. The random forest is trained for the observed portion of y, and then the forest is used to predict the missing portion of y (those rows of x_2 and x_3 that correspond to the missing rows in y are “dropped down” the grown forest to compute predictions). After missing values on x_1 are imputed, we move on to setting x_2 to y, and a random forest is similarly used to predict the non-missing values using the newly imputed x_1 and the (non-imputed) x_3. The newly trained forest is used to impute the x_2 missing values. The process is repeated, and each time the imputed values are updated until a convergence criterion is reached.

Post Hoc Probabilities

Though a single stage is assigned in the imputation method, it is of interest to compute the probability of the HD-ISS stage candidates of 0, 1, 2, 3, for the vector of observed variables for a participant in a given study visit. For example, given the observed values of TMS = 10, SDMT = 40, TFC = 13, and IS = 100, we might want to know how probable a visit would be classified in any one of the stages. It is not possible to compute these probabilities directly from the "missForest” algorithm, so we used a separate random forest fitted on the imputed data. Specifically, the imputed HD-ISS stage was predicted using the landmark variables discussed above, which generated a probability of classification for each stage.