PREDICT-HD

UNDERSTANDING THE DATA

Roland Zschiegner
Jeremy Bockholt
Jane Paulsen
3/15/2021

Principal Investigator:

Jane S Paulsen PhD
University of Iowa
Departments of Psychiatry, Neurology,
Psychology and Neurosciences
Iowa City IA, 52242
United States

Contents

HD-CLASSIFICATION:

The PREDICT-HD data set contains several variables that help to characterize the HD classification and HD status of a participant. The variables that help to characterize the subject are status, baseline_cap_grp, cap_grp, visit_diagnosis, and initial_dx. Each of these variables are found within the Demographics_Genetics data set and are explained in more detail below:

1. status – This variable characterizes the participant as either a case or control. Participants characterized as case have tested positive for the gene expansion and have a CAG (Cytosine, Adenine, and Guanine) repeat of 36 or greater. Controls are participants that tested negative for the gene expansion and have a CAG less than 36.

2. baseline_cap_grp – This variable categorizes the participant's stage at baseline. The groups are based on cap score which is calculated as CAP = Age X (CAG – 33.6600) [1]. The groups are:

   a. low = cap_score < 287

   b. med = cap_score between 287 – 367

   c. high = cap_score > 367

   d. cont = no cap_score is calculated.

3. cap_grp – This variable categorizes the participant's stage at the time of the visit. Groups are calculated as above.

4. visit_diagnosis – This variable indicates the first encountered Diagnostic Confidence Level (DCL) of 4 in the study.

5. initial_dx – This variable indicates diagnosis at the initial/baseline visit.

It should be noted that there are additional cap groups not included above. The additional notations are low_ex, med_ex, high_ex, and cont_ex. These notations indicate that the user should consider excluding the participant for analysis based on exclusionary criteria. Due to ethical considerations, some participants were given waivers to continue. This was done on a case by case basis.

In addition, it should be noted that there is inconsistency with DCL ratings on which the variables visit_diagnosis and initial_dx are based. For visit_diagnosis, there are 76 participants that have a DCL rating of 4 and a subsequent visit with a rating less then 4 and affects 162 rows of data. 2 of those participants have an exclusionary status which may explain the changes for those particular participants.

As for variable initial_dx, there are 30 participants that have a DCL 4 at visit 1. As mentioned previously, some participants were allowed to continue beyond the baseline visit for ethical considerations. Waivers were provided to allow the participant to continue. In addition, there are some DCL discrepancies. There are 6 participants with DCL significant changes at the next and in some cases subsequent visits.

MISSING VALUES AND IMPUTATION:

Missing values within the dataset are represented as blank entries within the data set. There are several reasons as to why a value may be missing within the data set. Due to the nature of the study, some measures were added and/or removed during the course of the study. In cases where the measure was removed, subsequent visits past the removal date would be a blank entry within the data set. Conversely, anytime a measure was added, visits prior to the addition of the measure would be blank as the measure would not have existed for those visits. Please refer to the Protocol document for more details on when measures were added and/or removed.

Total scores may also be missing. This is due to one or more of the values that are used to calculate the score is missing. If one or more values are missing to compute a score, the score itself is not computed. There are 5 exceptions to this rule. The Beck Depression Inventory II, Beck Hopelessness Scale, World Health Organization Disability Assessment Schedule, Symptom Checklist 90 Revised, and Total Functional Capacity are the exceptions. Imputation rules are applied to each of the tasks and are based on published scoring guides and/or community standard guidelines. Please review the section on each task later in this document.

DATE VALUES

TRANSFORMATION OF DATE VALUES

Date values within the PREDICT-HD data set are transformed to minimize participant identification risk where possible. However, it should be noted that some NIH funded repositories require a date of collection for submission. The date of collection is provided as an integer value, that reflects the number of days between the baseline visit and the date of the visit of interest.

Table 1 shows an example of conversion to number of days:

Table 1: number of days in the study

Entered Date	Representation in data set
2021-01-01	0
2021-01-30	30
2020-12-21	-10

Date values that represent the date of birth are given as age in years. Please note that some NIH repositories request the age in months.

Comorbidity and Concomitant medication start and stop dates are also represented as the number of days from baseline. In some cases, the participant did not remember the exact date and only provided the year. Dates with only the year provided were given a month of June and a day of 15. It should be noted that there are instances where the start and stop dates occur within the same month and therefore have the same number. In addition, when only year was provided, the imputation of month can also cause the start and stop dates to be the same. An additional variable called startdate_modifier is provided to indicate whether the comorbid condition or medication number of days from baseline provided occurred on the date provided or condition/medication occurred on an unspecified date such as when only the year was provided. For conditions/medications that occurred/taken prior to baseline, the number provided will be negative.

AGGREGATE VALUES

In an effort to minimize the risk of participant identification, specific variables are aggregated to provide a single value of when a threshold is reached. Table 2 shows the variables that are aggregated, description of the variable, and the criteria for aggregation.

Table 2 – Aggregated values

Data File	Variable	Variable Description	Criteria for Aggregation
Demographics_Genetics	RACE	Racial Category	fewer than 20 cases per category
Demographics_Genetics	HD_CAG_A2_10	1.0 Lab-reported CAG length (allele 2) from blood	> 28
Demographics_Genetics	HD_CAG_A2_10	2.0 Lab-reported CAG length (allele 2) from blood	> 28
Demographics_Genetics	HD_CAG_A2_SV	2.0 Lab-reported CAG length (allele 2) from saliva	>28

The following categories for race were aggregated into Other (8): American Indian/Alaska Native (1), Asian (2).

Table 3: Number of participants subject to aggregation

Data file	Variable	Label	Number of Participants
Demographics_Genetics	RACE	Other (8)	9
Demographics_Genetics	RACE	American Indian/Alaska Native (1)	3
Demographics_Genetics	RACE	Asian (2)	6

Demographics_Genetics	HD_CAG_A2_10	>28	11
Demographics_Genetics	HD_CAG_A2_10	>28	10
Demographics_Genetics	HD_CAG_A2_SV	>28	0

ASSESSMENT SCORE CALCULATION

Assessment scoring was generated within the PREDICT-HD data system where applicable. In some cases, the scores of computerized tasks were generated by the program and then imported into the PREDICT-HD data base.

In general, if assessment scoring requires all items for calculation and there is a missing item, the total score and any sub scores that use the missing item is left blank. The individual items are provided whenever possible so that the user may determine if the imputation is warranted. However, there are several scores that are calculated using imputed values. Table 4 shows the list of scores and their source.

Table 4: Imputed variables

Data File	Variable	Measure	Imputation rule
Psychiatric_P	bditotal	Beck Depression Inventory II	< 5 items missing
Psychiatric_P	bhs_tot	Beck Hopelessness Scale	< 4 items missing
Psychiatric_P	SCL90_PSDI_P	Symptom Checklist 90 Revised Participant	< 18 items missing
Psychiatric_P	tsclpsdi_p	Symptom Checklist 90 Revised Participant	< 18 items missing
Psychiatric_C	SCL90_PSDI_C	Symptom Checklist 90 Revised Companion	< 18 items missing
Psychiatric_C	tsclpsdi_c	Symptom Checklist 90 Revised Companion	< 18 items missing
Functional_P	TFC	Total Functional Capacity	When uhdrs72-74 are missing + 7
Functional_P	whodasp_total36	WHODAS 36 item participant	< 2 items missing and not from same domain
Functional_P	whodasp_total12	WHODAS 12 item participant	< 1 item missing
Functional_C	whodasc_total36	WHODAS 36 item companion	< 2 items missing and not from same domain
Functional_C	whodasc_total12	WHODAS 12 item companion	< 1 item missing

For more detailed information on scoring please refer to the CRF_Completion_Guidelines document.

UNIFIED HUNTINGTON'S DISEASE RATING SCALE (UHDRS)

The PREDICT_HD data provides elements from a Unified Huntington's Disease Rating Scale (UHDRS). Scores from several of the scales contained within the UHDRS are provided. Additional information about scoring for the UHDRS can be found here:

Unified Huntington's Disease Rating Scale: reliability and consistency. Huntington Study Group. Mov Disord. 1996 Mar;11(2):136-42. Doi: 10.1002/mds.870110204. PMID: 8684382.

The motor section of the UHDRS is comprised of 31 motor items (questions 1-15) comprising the Total Motor Score (TMS) which is the sum of all 31 items and range from 0-124. Within those 31 items there are several domains that are calculated: Oculo, Brady, Rigidity, Dystonia, and Chorea. Table 5 provides score calculations for all motor scores:

Table 5: Motor score calculations

Variable	Score Calculations	Score Range
oculo	Sum of ocular pursuit, saccade initiation, and saccade velocity	0-24
brady	Sum of dysarthria, tongue protrusion, finger taps, pronate/supinate hands, Luria, bradykinesia, gait, tandem, retropulsion	0-44
rigidity	Degree of resistance in movement in both arms (0=absent to 4=severe for each arm)	0-8
dystonia	Degree of sustained abnormal posturing/movement in trunk and extremities (0=absent to 4=prolonged for each of 5 areas)	0-20
chorea	Degree of rapid abnormal involuntary movements/contractions in face, trunk, and extremities (0=absent to 4=prolonged for each of 7 areas)	0-28
TOTAL_MOTOR_SCORE	Total Motor Score - Sum of 31 motor items (UHDRS1-UHDRS15)	0-124

The UHDRS Diagnostic Confidence Level (DCL) is scored as the rater's confidence level in the participant's motor onset. The range of the DCL is 0-4 with 0 indicating normal (no abnormalities) and 4 indicating motor abnormalities that are unequivocal signs of HD (greater than or equal to 99% confidence).

The Total Functional Capacity Scale is comprised of questions 70-74 of the UHDRS. The questions cover occupation, finances, domestic responsibilities, performing activities of daily activities, and level of care. The scoring range is 0-13 with higher scores indicating greater functionality. (See Table 6)

The Functional assessment (FAS) is comprised of 25 yes/no questions about a variety of daily tasks. The rater is asked to determine if the participant is able to perform the daily task based on their impression

of disability due to any cause, whether cognitive or physical. The scoring range is 25-0 where higher scores indicates greater functionality. (See Table 6)

The Independence Scale is intended to assess the ability of the participant to function independently in activities of daily living across the full spectrum of the disease. The scale attempts to quantify a subject's general level of function and is a percentage ranging from 100% to 5%. Higher percentages indicate greater functionality.

Table 6: TFC and FAS scoring

Variable	Score Calculations	Score Range
TFC	TFC Score (sum uhdrs70-uhdrs74)	0-13
FAS	FAS Score (sum uhdrs43-uhdrs67)	0-25

PROBLEM BEHAVIORS ASSESSMENT/UHDRS BEHAVIORAL SCALE (UHDRS-BEH)

The PREDICT-HD data set contains both the UHDRS Behavioral Scale (UHDRS-BEH) and Problem Behaviors Assessment (PBA). The UHDRS-BEH and PBA are very similar scales as the PBA was developed from the UHDRS-BEH. The scoring domains are Depression, Irritability, Psychosis, Apathy and Executive Function Score. The Executive Function (combined) score was developed for the PBA and both the UHDRS and PBA use nearly identical items for obsessive thinking and compulsive behaviora. It should be noted that the severity question for disorientation in the UHDRS-BEH is a yes/no question. This item was converted to a likert scale, 0-5, in the PBA. We mapped these items together in the PREDICT-HD dataset to facilitate longitudinal analyses. Individual investigators may choose to separate the UHDRS and PBA items and can do this. In addition, a UHDRS Psychiatric total sum is provided from the UHDRS-BEH. Table 7 shows the variable and score calculations

Table 7: PBA/UHDRS-BEH scoring

Variable	Score Calculations	Score Range
depscore	Depression score (Severity of Depressed Mood * Frequency of Depressed Mood) + (Severity of Suicidal Ideation * Frequency of Suicidal Ideation) + (Severity of Anxiety * Frequency of Anxiety)	0-48
irascore	Irritability Score (Severity of Irritability * Frequency of Irritability) + (Severity of Angry or Aggressive Behaviour * Frequency of Angry or Aggressive Behaviour)	0-32
psyscore	Psychosis Score (Severity of Delusions/Paranoid Thinking * Freqency of Delusions/Paranoid Thinking) + (Severity of Hallucinations * Freqency of Hallucinations)	0-32
aptscore	Apathy Score (Severity of Apathy * Freqency of Apathy)	0-16
exfscore	Executive Function Score (Severity of Perseverative Thinking or Behaviour * Freqency of Obsessive-Compulsive Behaviours) + (Severity of Perseverative Thinking or Behaviour * Freqency of Obsessive-Compulsive Behaviours)	0-32

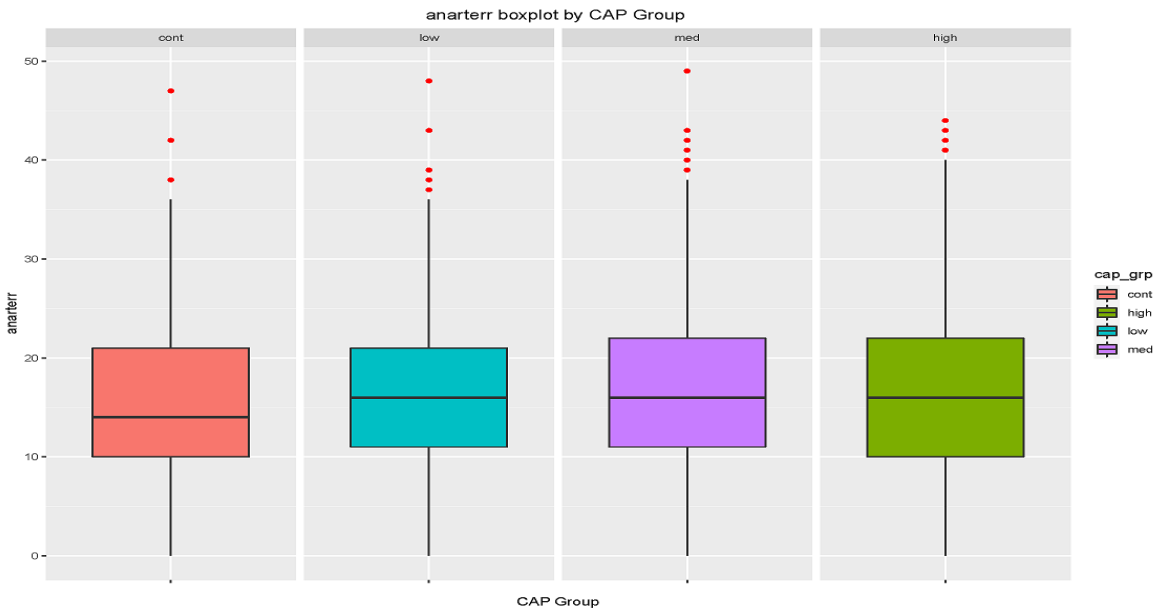
UHDRS_PSYCH_P	Companion UHDRS Psychiatric total - Sum of product of frequency and severity for UHDRS25a-UHDRS35b (UHDRS25a*UHDRS25b+...+UHDRS35a*UHDRS35b)	0-176
UHDRS_PSYCH_C	Companion UHDRS Psychiatric total - Sum of product of frequency and severity for UHDRS25a-UHDRS35b (UHDRS25a*UHDRS25b+...+UHDRS35a*UHDRS35b)	0-176

For additional information about scoring for individual measures, please see the data dictionary and /or Crf_Completion_Guidline.

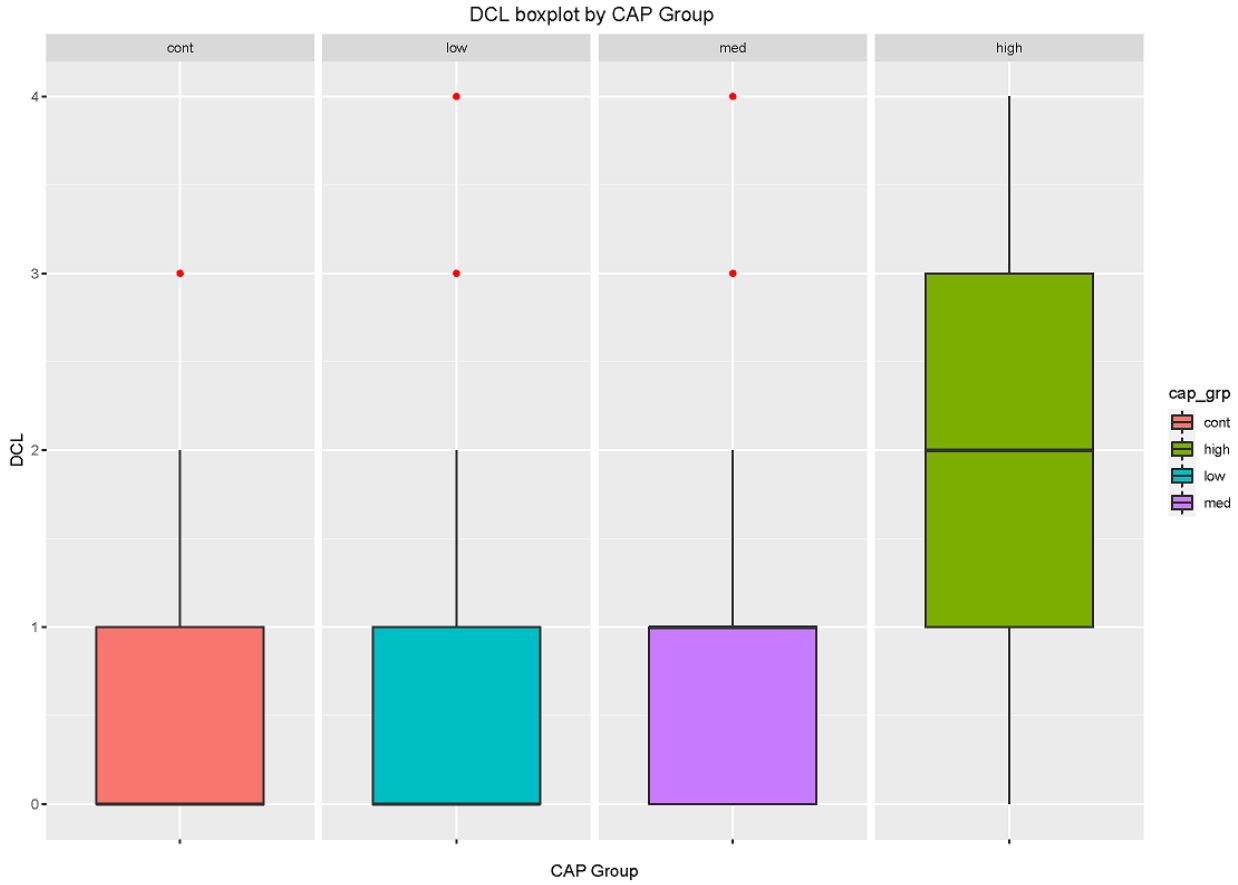
UNUSUAL FINDINGS

As part of the overall QA/QC before the release of the data sets, a series of custom checks are made for extreme outliers and unusual or improbable values. During the PREDICT-HD study, queries to study coordinators were made when items were questionable. Corrections were made when verified with site staff or when cognitive QA/QC rescoring occurred. However, there are some values that could not be queried and/or corrected due to time frame of discovery of value and date of collection as well as confirmation from the site that the value was correct. In these instances, the values are provided "as is" and it is up to the user to determine if the values should be excluded for analyses.

Examples of unusual or unexpected values can be found in several of the variables provided. The American National Adult Reading Test (ANART) has several scores in which the number of errors provided seems to reflect an accidental scoring error, where the participant's score was the number correct as opposed to the number incorrect. However, the Cognitive teams at Indiana during the 1.0 phase of the study and Brown University for the 2.0 phase, rescored and reviewed the scoring and determined the scores were correct. The following boxplot shows the ANART errors in relation to CAP group.



There are 9 controls that were given a Diagnostic Confidence Level of 3 during the course of the PREDICT-HD study. Queries were sent to the sites and the scores were confirmed. These controls should be considered outliers for any analysis and removed when possible. The boxplot below shows DCL by CAP group and reflects the DCL 3 for the handful of controls.



For more information, please also review the PREDICT-HD Data Set Overview and Data Dictionary documentation.