PREDICT-HD

DATA SET OVERVIEW

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

DATA SETS:

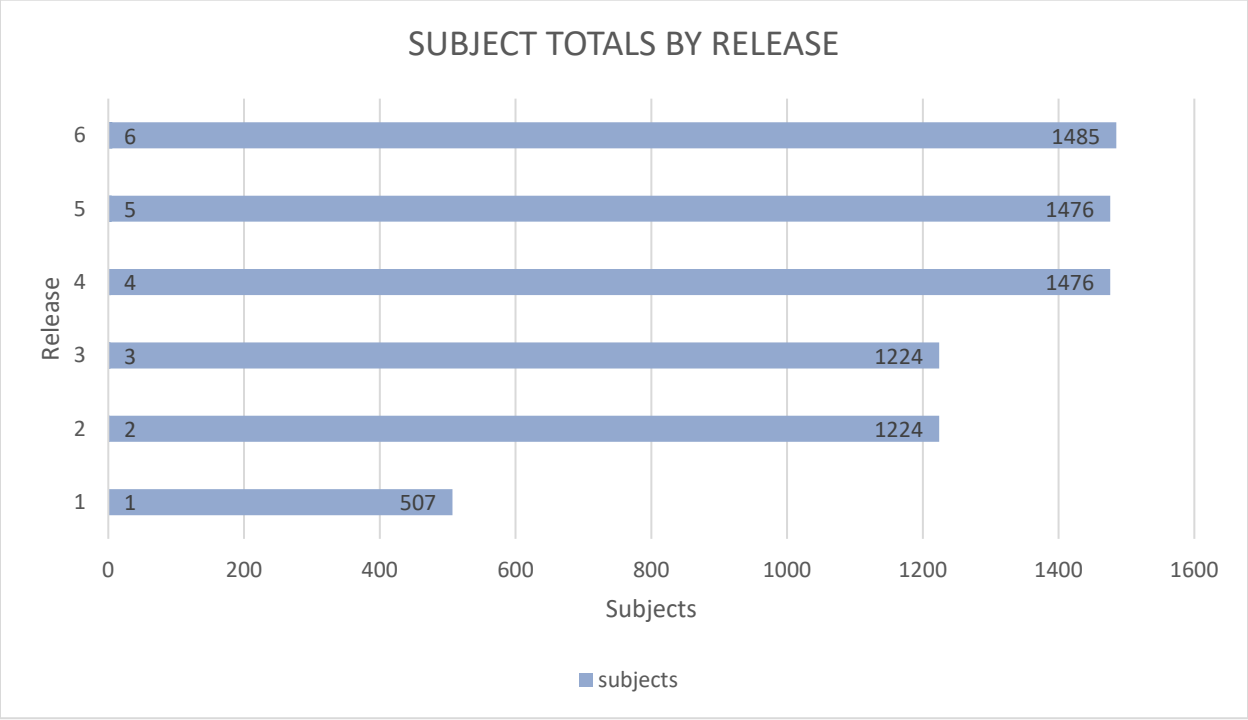
The data sets contained in Release 7 are sourced from data collected during the PREDICT-HD study. The PREDICT-HD study was a multinational, longitudinal, observational study aimed at identifying biological and refined clinical markers of pre-HD in humans, and then validating the optimal marker(s) and clinical end points for use in preventive clinical trials. The study ran from 2001 through 2014 and included 33 sites across the US, Canada, UK, Germany, Spain and Australia. Additional information was collected from 2014 through 2017 to help support NIH-funded ancillary grants attached to the parent PREDICT-HD study.

PARTICIPANTS:

The data sets contain 1485 participants that agreed to share their data for research. Over the course of the study, one participant opted to have all of their data and samples removed from the study. Four participants had their samples removed but allowed their data to be used. Those participants may have been included in earlier released data sets prior to their request for opting out. Once the study received the request to remove their data, it was removed and their samples were destroyed.

Sample sizes for prior data releases are as follows:

FIGURE 1 – SAMPLES SIZES



VISITS:

Release 7 contains data for 7700 visits. The visits comprise both baseline and follow up visits across the 1.0 and 2.0 studies. To delineate whether the visit occurred in the 1.0 study or the 2.0 study the visit number begins with a 1 or 2 followed by the number. As an example, participant XXXXXX had 4 visits in the 1.0 study and enrolled in the 2.0 study and was seen 3 times with no missed visits within the timeframe. The visit sequence would be 101, 102, 103, 104, 205, 206, 207 where visit 205

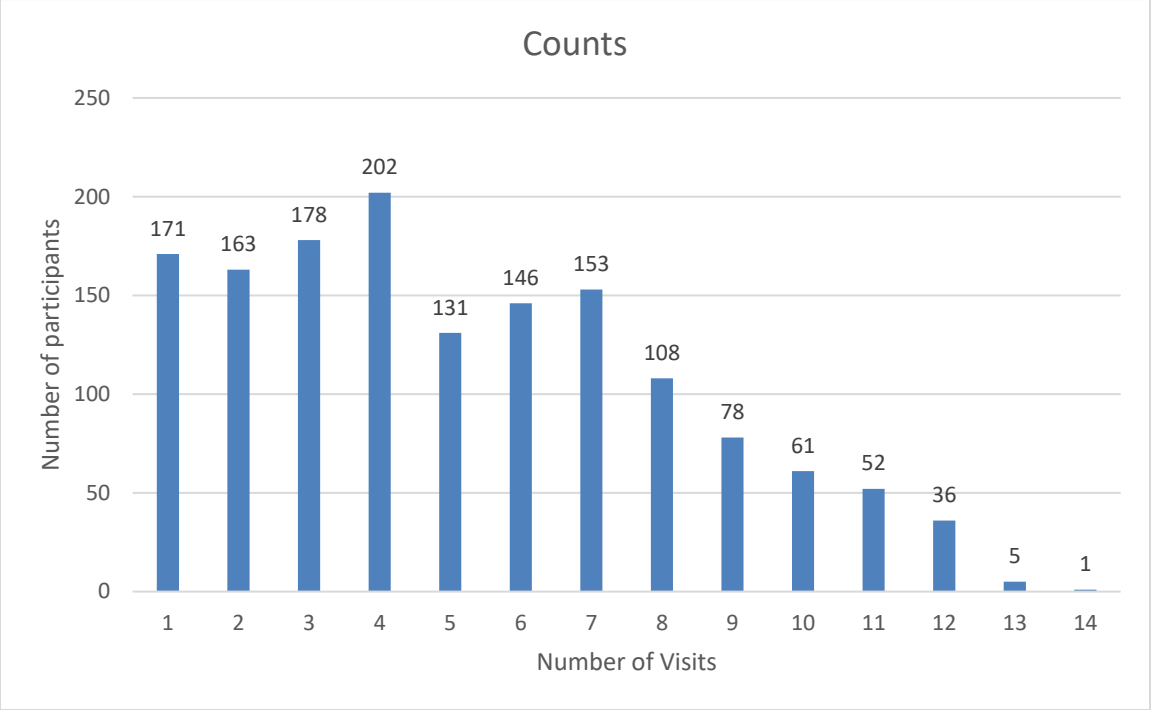
denotes enrollment into the 2.0 study. It should be noted that some participants have one or more missed visits due to scheduling issues. Those are denoted by a skip in visit number. For example, using the previous example, if the 4th visit was missed, the sequence would be the following: 101, 102, 103, 105, 206, 207, 208.

TABLE 1: VISIT EXAMPLES

Participant	Visit Type	Visit 1	Visit 2	Visit 3	Visit 4	Visit 5	Visit 6	Visit 7
XXXXXX	No missed visits	101	102	103	104	205	206	207
XXXXXX	Missed visit	101	102	103	105	206	207	208

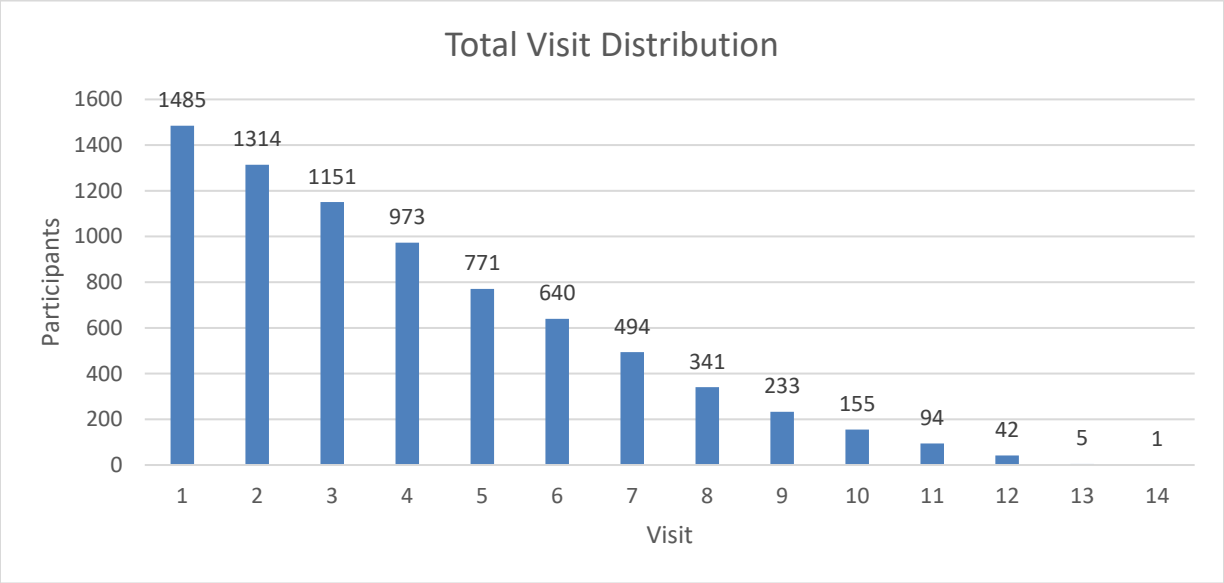
During the course of the study, there were 991 participants that had 6 visits or less while 494 had 7 or more. The distribution of the number of participants vs the number of visits made are listed in figure 2.

FIGURE 2 – DISTRIBUTION PARTICIPANT VS NUMBER OF VISITS (N = 1485)



The maximum participant's counts per visit is listed in figure 3.

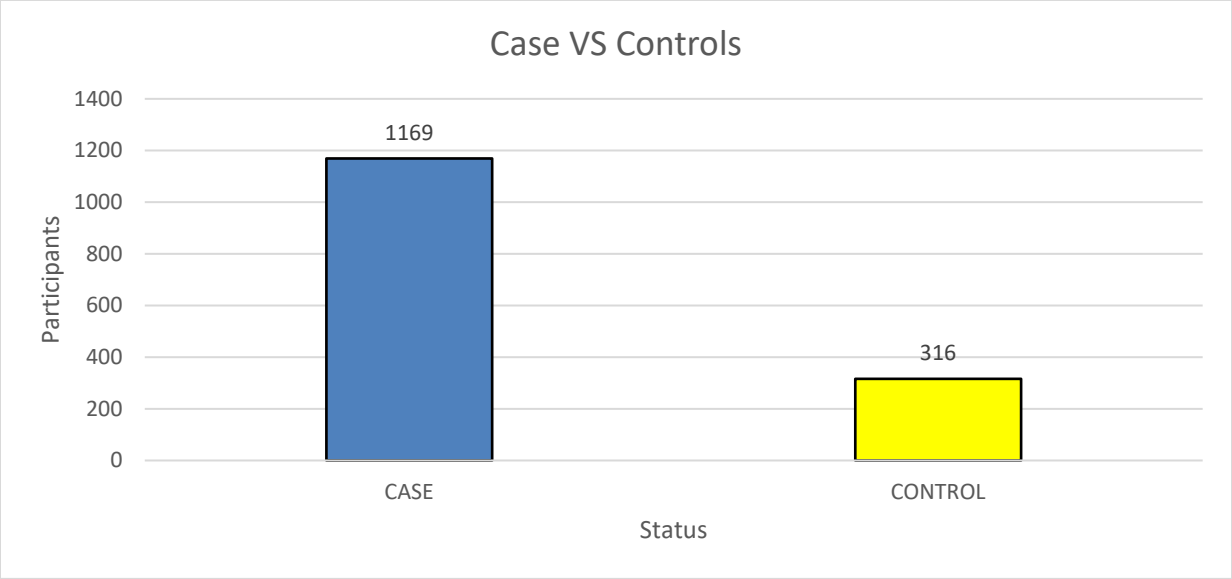
FIGURE 3 – MAXIMUM PARTICIPANT COUNTS PER VISIT



SAMPLE CHARACTERISTICS:

The participants within the PREDICT-HD dataset are generally characterized as case and control (see figure 4). A participant characterized as case is considered gene positive and has cytosine-adenine-guanine (CAG) expansion greater than or equal to 36. Individuals who are considered controls have a CAG of less than 36. It should be noted that individuals who have CAG repeat lengths of greater than or equal to 36 and less than 40 may or may not develop symptoms of HD over their lifetime.

FIGURE 4 SHOWS THE CASE VS CONTROLS



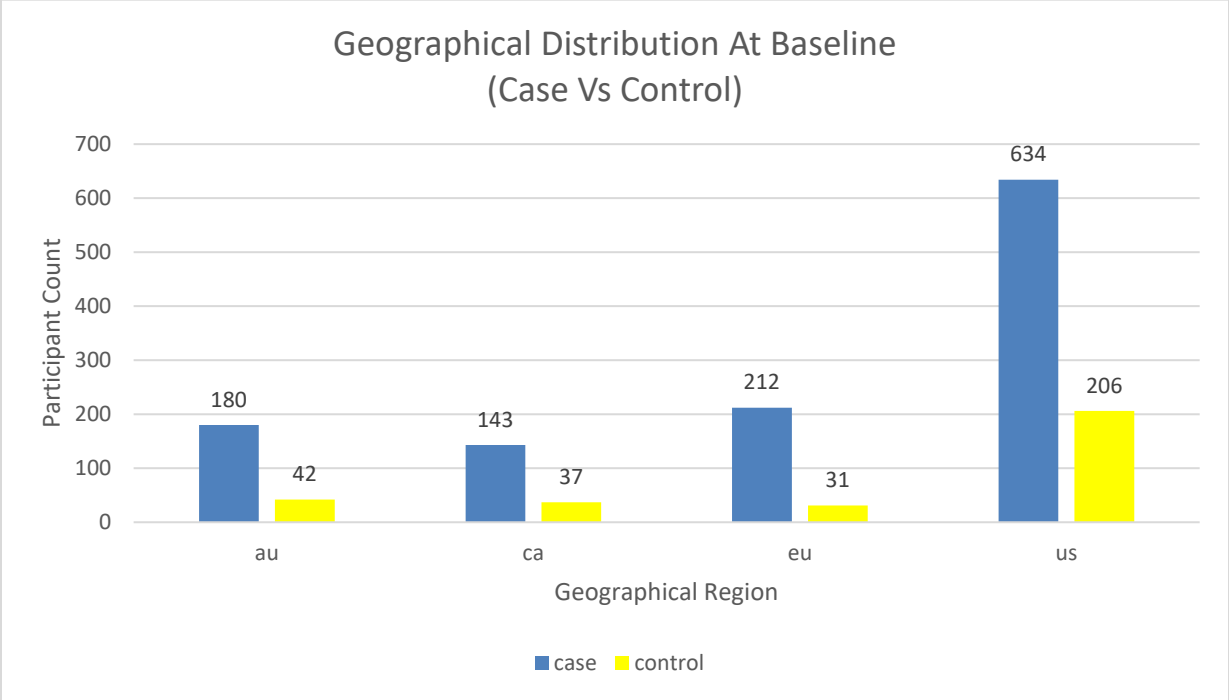
GEOGRAPHICAL DISTRIBUTION

The numbers of participants by geographic area at baseline are listed in figure 5. It should be noted that participants did change sites within geographic areas as well as between geographic areas. Users should take note of the site number provided when analyzing data. The list of geographical areas and the descriptions are listed in the table below in table 2.

TABLE 2: GEOGRAPHICAL AREAS AND DESCRIPTIONS

Geographical Area	Description
au	Australia
ca	Canada
eu	Europe (UK, Germany, Spain)
us	United States

FIGURE 5 – NUMBER OF PARTICIPANTS PER GEOGRAPHICAL AREA AT BASELINE (n = 1485)



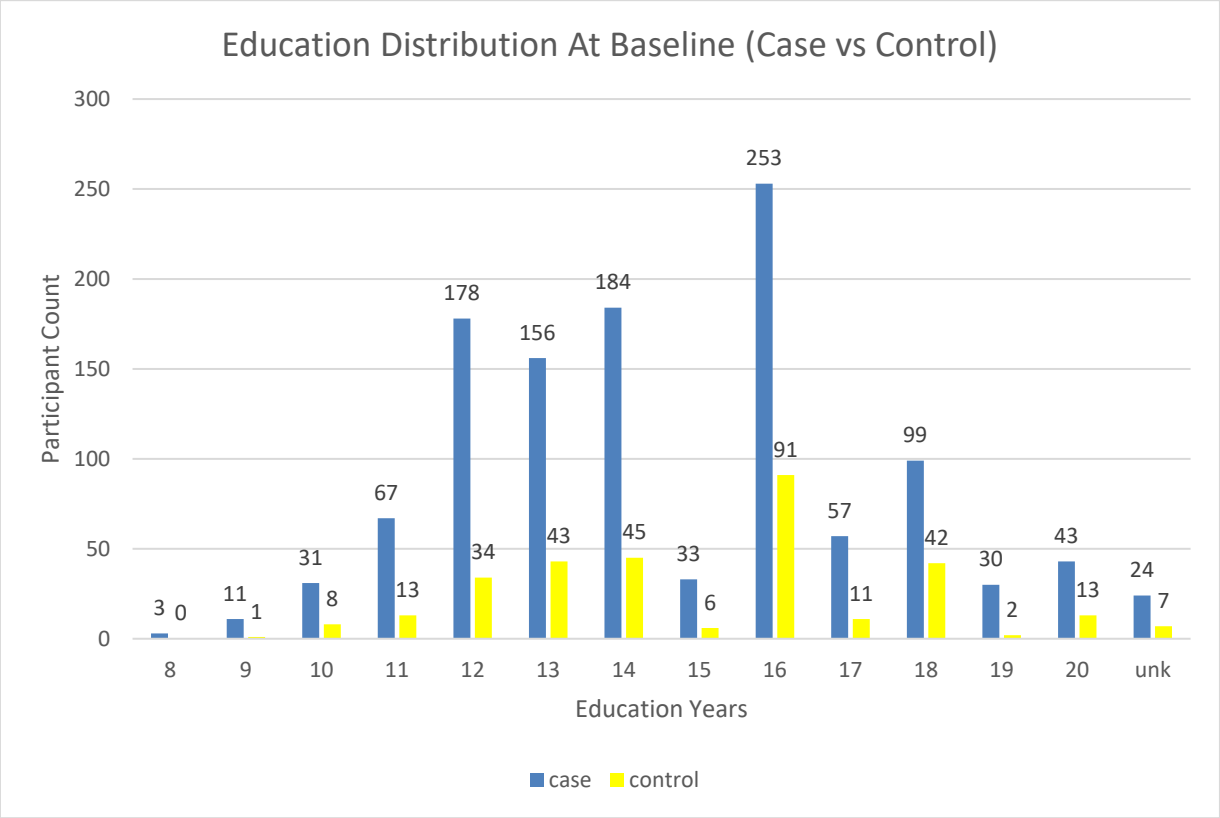
EDUCATION

Education in the PREDICT-HD study was categorized by number of years. Table 3 shows the number of educational years and degree level equivalent if the participant did not know the actual number of years. Please note that the Inclusion criteria for education is at least 8 years of education and it is up to the user to determine whether to use the minimum value of 8 when the number of years is missing within the data set. Figure 6 shows the educational distribution at baseline of participants in the study.

TABLE 3: DEGREE EQUIVALENT FOR NUMBER OF YEARS OF EDUCATION

Number of years	Degree description
11 years	GED
13 years	Additional education post-high school
14 years	2-yr Associates degree
16 years	4-yr College degree
18 years	Master's degree
19 years	JD
20 years	PhD and MD

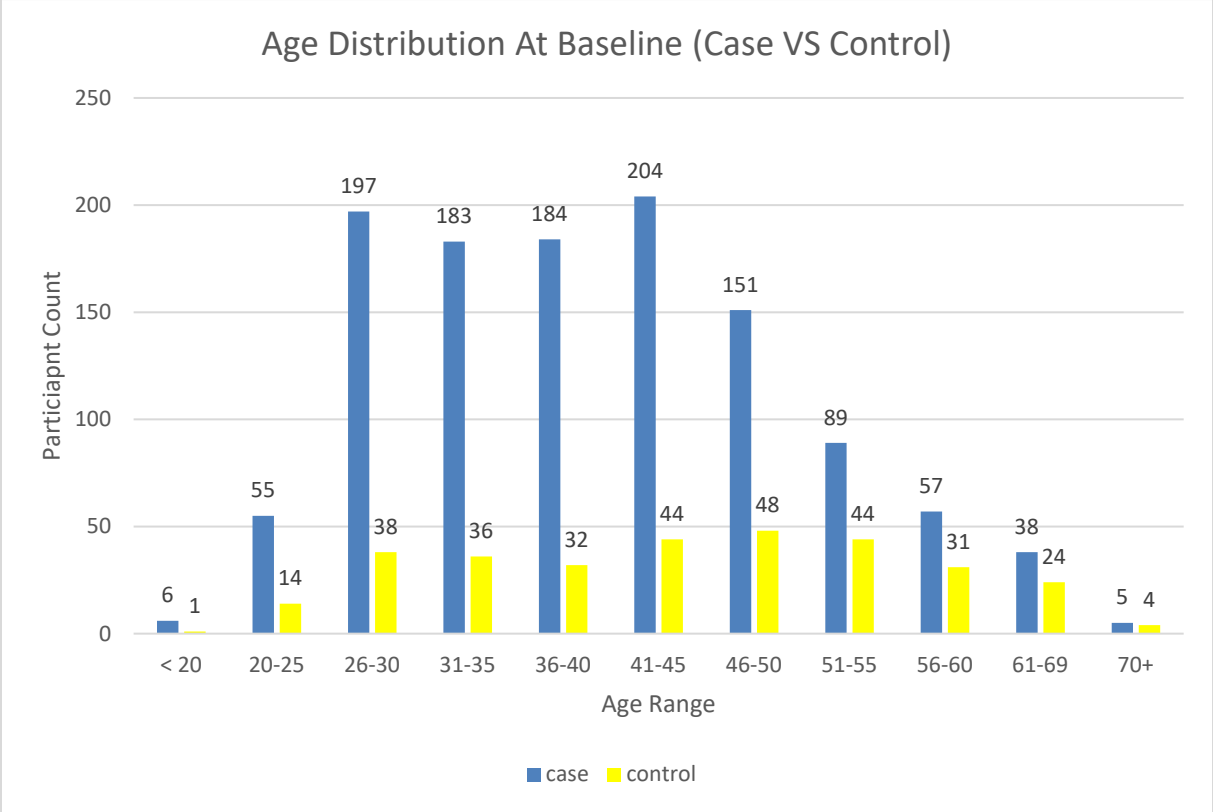
FIGURE 6: EDUCATION DISTRIBUTION AT BASELINE (n=1485 total)



AGE

Distribution of age at baseline is represented in figure 7. 311 participants were age 30 or under at baseline. 435 participants were between ages 31 and 40 at baseline and 739 participants were age 41 and up. This includes both cases and controls. Distribution of age by case and control is represented in the figures 8 and 9

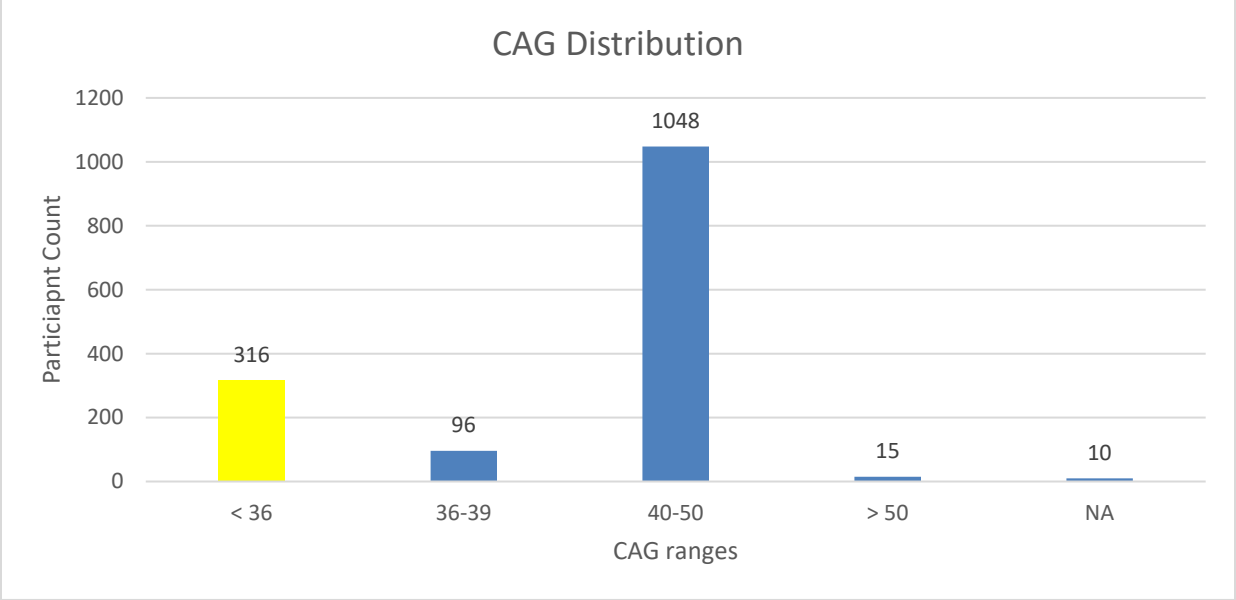
FIGURE 7 – AGE DISTRIBUTION AT BASELINE ALL SUBJECTS (n=1485)



CAG

The CAG distribution is represented in Figure 8. It should be noted that there were 10 participants in which blood samples were collected for lab verification. However, the samples were non-viable on receipt and a lab verified CAG could not be obtained. Out of the 10 participants missing CAG, 4 participants were lost to contact after the baseline visit and 6 participants had subsequent return visits. Additional attempts to collect a blood samples were made, but were unsuccessful. Self-reported CAG values were queried but trusted values were unable to be obtained.

FIGURE 8 – CAG DISTRIBUTION (n=1485)

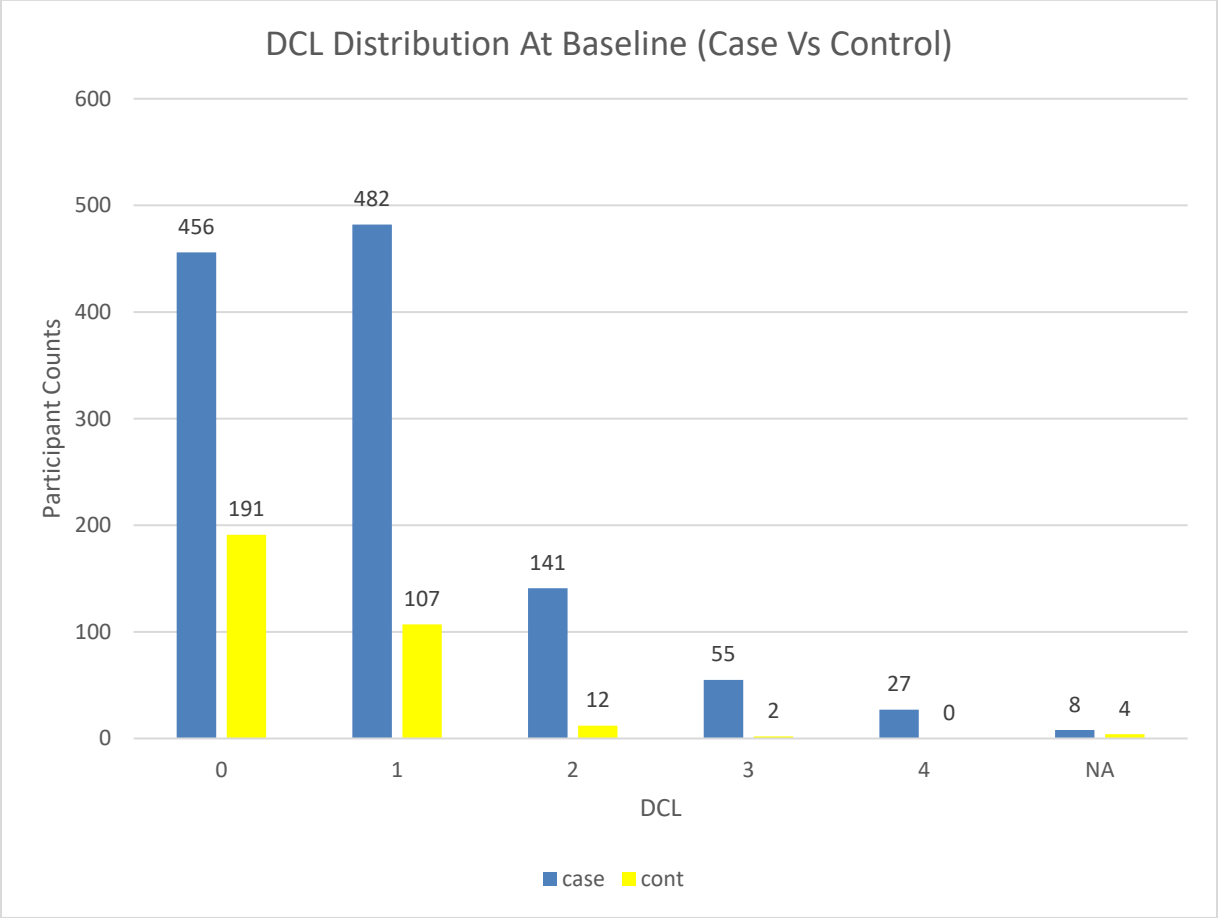


DIAGNOSTIC CONFIDENCE LEVEL (DCL)

Baseline distribution of Diagnostic Confidence Level (DCL) of the Unified Huntington's Disease Rating Scale is represented in Figure 9. It should be noted that participants that had a DCL of 4 at baseline were allowed to continue in the study due to ethical considerations. It should also be noted that motor raters were supposed to be blinded to the gene status and/or CAG repeat number while doing the motor rating. As a result, there may have been an underlying condition that may have affected the motor rating. During the 1.0 study, comorbid conditions were not collected at subsequent visits beyond baseline. Therefore, an underlying condition that was not reported at baseline could have affected the motor ratings. The 2.0 study corrected this omission and additional data collection of comorbid conditions were added. The values are left in the data set and should be noted when analyzing the data.

. .

FIGURE 9 – DCL DISTRIBUTION AT BASELINE (n=1485)



TOTAL FUNCTIONAL CAPACITY (TFC)

Total Functional Capacity (TFC) at baseline is represented in Figure 10. It should be noted that similar to DCL, a participant with diminished functional capacity was allowed to continue in the study for ethical reasons. In addition, underlying conditions could also contribute to diminished capacity. For example, there was a control participant at baseline that had diminished capacity due to several health conditions. The user should be aware that the value was left in the data as it is valid but unexpected.

FIGURE 10 – TFC DISTRIBUTION AT BASELINE (n=1485)