Applying the Huntington's Disease Integrated Staging System (HD-ISS) to Observational Studies

Authors:

Jeffrey D. Long, Departments of Psychiatry and Biostatistics, University of Iowa, jeffrey-long@uiowa.edu

Emily C. Gantman, CHDI Management/CHDI Foundation, Princeton, NJ, USA, emily.gantman@chdifoundation.org

James A. Mills, Department of Psychiatry, University of Iowa, jim-mills@uiowa.edu

Jatin G. Vaidya, Department of Psychiatry, University of Iowa, jatin-vaidya@uiowa.edu

Alexandra Mansbach, APS Consulting Services Ltd., Washington, DC USA, alexandra.mansbach@chdifoundation.org

Sarah J. Tabrizi, UCL Huntington's Disease Centre, Department of Neurodegenerative Diseases, UCL Queen Square Institute of Neurology, UK Dementia Research Institute, University College London, UK, s.tabrizi@ucl.ac.uk

Cristina Sampaio, CHDI Management/CHDI Foundation, Princeton, NJ, USA, cristina.sampaio@chdifoundation.org

Correspondence: Jeffrey D. Long, 500 Newton Road, Iowa City, IA 52246, jeffrey-long@uiowa.edu, 651-249-9558

Abstract

BACKGROUND: The Huntington's Disease Integrated Staging System (HD-ISS) has four stages that characterize disease progression over an individual's lifespan. Classification is based on CAG length as a marker of Huntington's disease (Stage 0), striatum atrophy as a biomarker of pathogenesis (Stage 1), motor or cognitive deficits as HD signs and symptoms (Stage 2), and functional decline (Stage 3). One issue in the implementation of the HD-ISS is that existing prospective studies may not collect all the data required to classify participants. For example, the largest active observational study, Enroll-HD, does not collect imaging. A second consideration is that the HD-ISS stages characterize periods of disease progression that may span several years and there is benefit in defining progression subgroups within a stage.

OBJECTIVES: Impute stages of the HD-ISS for Enroll-HD and other studies in which missing data precludes direct stage classification, and then define progression subgroups within stages.

METHODS: A machine learning algorithm was used to impute stages using participant age and HD-ISS landmark variables. Agreement of the imputed stages with the observed stages was evaluated using a variety of methods, including graphing and propensity score matching. The distributions of the progression indices were examined by stage, and descriptive statistics were used to define progression subgroups. Optimal cut-point analysis was performed to find values that maximally separated the distributions.

RESULTS: There was good overall agreement between the observed stages and the imputed stages. However, the algorithm tended to over-assign Stage 0 and under-assign Stage 1 for individuals who were early in progression. The medians of the progression indices increased with stage, but the distributions showed extensive overlap among stages.

CONCLUSIONS: There is evidence that the imputed stages can be treated similarly to the observed stages for the types of large-scale analyses typically supported by Enroll-HD and other studies. When imaging data are not available, imputation can be avoided by collapsing the first two stages using the categories of Stage ≤ 1, Stage 2, and Stage 3. Progression subgroups defined within a stage can help to identify groups of more homogeneous individuals. These results will facilitate the use of the HD-ISS in Enroll-HD to aid in the planning of interventional studies in HD.

Keywords: Huntington's disease, integrated staging system, missing data imputation, disease progression, Enroll-HD

Introduction

The Huntington's Disease Integrated Staging System (HD-ISS) [1] is an evidence-based framework intended to facilitate clinical research and interventional studies at points earlier in the disease course than previously considered. The HD-ISS characterizes disease progression from birth onward using four stages. 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 pathophysiology as indicated by striatal atrophy. Stage 2 indicates the appearance of HD signs and symptoms (motor or cognitive), and Stage 3 reflects functional decline.

Stage classification depends on the pattern of meeting threshold criteria for landmark variables. Each variable has a cut-off threshold; if either measured landmark variable surpasses the established cut-off threshold, then an individual meets the criteria for a stage. Final classification is based on the highest stage criteria reached, provided this is achieved in the order consistent with the HD-ISS (otherwise, classification is undefined).

The landmark for Stage 0 is CAG length, with a threshold of 40 or greater, which is based on current penetrance evidence [1]. The landmarks for Stage 1 are caudate and putamen volume, corrected for total intra-cranial volume. The landmarks for Stage 2 are the UHDRS [2] Total Motor Score (TMS) and Symbol Digit Modalities Test (SDMT; education is factored into the SDMT thresholds). Finally, the landmarks for Stage 3 are Total Functional Capacity (TFC) and the Independence Scale (IS).

For Stages 1-3, surpassing the threshold for either variable or both fulfills the criteria. These thresholds depend on age, but not on CAG length, having been derived from data of non-HD controls (CAG < 36). Tabled threshold values appear in the appendix of the original paper [1], and a web-based tool is available for staging based on user input (https://enroll-hd.org/HD-ISS-Calculator/).

The HD-ISS is intended to be applied cross-sectionally. The established landmark thresholds demarcate the extreme values derived from the non-HD control population, rather than signaling a within-person shift from previous visits. Therefore, stage criteria are based on the level of the landmark variables at a particular visit, and not based on change or the rate of change over visits.

To aid the use of the HD-ISS in clinical research, we address two challenges of implementation. First, precise staging requires that all the landmark variables be available. The largest active natural history study of HD, Enroll-HD [3], does not collect imaging. Because Enroll-HD is widely used to study natural history and to plan clinical trials, it would be beneficial to apply the HD-ISS to Enroll-HD.

Second, the HD-ISS defines stage boundaries, but it does not provide specific information regarding subgroups of progression within a stage. Individuals who recently entered a stage might be changing at a slower rate compared to individuals who will soon exit the stage. While the stages broadly categorize the phase of disease progression, identifying a progression subgroup within a stage can help researchers to define a more homogeneous subpopulation. This information can be used as the basis of prognostic enrichment strategies for improving interventional studies.

Both challenges are addressed in this paper. We impute HD-ISS stages for study visits that have missing landmark variables (for example, imaging in Enroll-HD) or that have patterns of variables that are incongruent with the system. Then, we map the progression indices, such as the CAG-age product (CAP) [5] and the HD prognostic index normed (PIN)[6], to the HD-ISS and define subgroups of progression within stages. The results are discussed in the context of clinical trial planning with an emphasis on using Enroll-HD data.

Materials and Methods

Participants. Four data sets were used in the analysis: Enroll-HD [3] (fifth periodic data set), IMAGE-HD [7], PREDICT-HD [8], and TRACK-HD/ON [9]. Study sites were required to obtain and uphold local ethics committee approvals and all participants gave written informed consent that included the distribution of coded data for research purposes. The Enroll-HD data set was obtained in December 2020, and the remaining data sets were obtained in February 2020. The data sets are publicly available (https://www.enroll-hd.org/).

Inclusion criteria for the analysis were CAG 40-50 and age 18 or older at the first visit. This resulted in a total sample size of $N = 15338$ with 41194 repeated visits. The number of participants (and visits) per study was 14190 (37513) for Enroll-HD, 70 (183) for IMAGE-HD, 915 (2353) for PREDICT-HD, and 273 (1145) for TRACK-HD/ON. The mean follow-up time was 3.62 years (SD = 1.68), with 31% of the overall sample having only one visit and 69% having 2-10 visits.

Measures. Imaging was performed in all the studies except Enroll-HD. Caudate and putamen volumes were obtained from segmentation using the recon-all pipeline of FreeSurfer version 6 (see the HD-ISS paper supplementary appendix [1] for more details). The volume of each structure was divided by total intra-cranial volume (ICV) to adjust for head size. In addition to the UHDRS variables of TMS, SDMT, TFC, and IS, the Diagnostic Confidence Level (DCL) and the Stroop Word Test (SWR) were used. DCL = 4 (the highest rating) is defined as clinical motor diagnosis in research contexts [10]. TFC and IS were not collected in IMAGE-HD. Education was treated as binary, with "low" being the UNESCO ISCED 1997 classification 0-3 and "high" being 4-6 [11].

Progression indices included the two versions of CAP from the HD literature. The first version [5] was computed as CAP = age × (CAG − 33.66). To provide context, CAP = 413 was associated with a 50% probability of clinical motor diagnosis (DCL = 4) for the data considered in this study (results not presented). The second version [12] was computed as CAP_{100} = age × (CAG − 30) ÷ 6.49. CAP_{100} is named as such because the expected age of clinical HD diagnosis as reported in Enroll-HD (`hddiagn`) is associated with a score of 100. We also considered PIN [6], computed as a weighted combination of TMS, SDMT, and CAP, PIN = (51 × TMS − 34 × SDMT + 7 × age × (CAG − 34) − 883) ÷ 1044. PIN = 0 indicates that if a hypothetical HD cohort started with this value at baseline, then 50% would be predicted to reach a rating of DCL = 4 within 10 years. PIN < 0 indicates it would take longer than 10 years (the cohort is farther from clinical motor diagnosis), and PIN > 0 indicates it would take less than 10 years (the cohort is closer to clinical motor diagnosis). For participants who had repeated visits, the progression scores were time-varying.

Statistical analysis. Missing data for each visit was singly imputed using the machine learning algorithm `MissForest` [13], which uses random forest [14] with chained equations [15]. Consistent with the cross-sectional nature of the HD-ISS, all visits were used to build the imputation model, and there was no explicit modeling of the repeated measurements nested within participants. `MissForest` imputes on a variable-by-variable basis with each incomplete variable acting as the outcome and using all other variables in the imputation model as predictors. Additional details are provided in the supplementary materials.

The stages compatible with the HD-ISS in PREDICT-HD and TRACK-HD/ON were not imputed, but rather treated as "ground truth" for training of the algorithm. Imputation was based on the landmark variables and education, age, DCL, SWR, and sex (though the last three had negligible predictive power). Because `MissForest` cycles through each incomplete variable, imputation was performed for all variables with missing values, not just the HD-ISS stage.

Graphical procedures were used to examine the imputation results, with an emphasis on the extent of agreement among the observed and imputed values [16,17]. In order to depict HD participants with a range of CAG lengths in the same graph, we mimicked previous approaches [8,18] in using CAP_{100} as the time metric for most graphing, which can be thought of as age-adjusted for CAG expansion.

To account for the pre-existing differences among the studies, a participant from PREDICT-HD or TRACK-HD/ON who had an observed stage was matched to a participant from Enroll-HD who had an imputed stage (1:1 matching). The goal was to balance several of the observed variables (with no missingness) common among the studies and then compare the observed and imputed stages among the matched samples. All the variables from the imputation analysis with complete data were used for the matching, which excluded the imaging variables. SWR was also excluded because of a relatively high rate of missingness. Exact matching was used for CAG length, and propensity score matching [19] was used for age, TMS, SDMT, TFC, DCL, education, and sex. A caliper was applied for age and TMS to help ensure similar means and variances among the groups. To evaluate the (dis)agreement of the observed and imputed stages, we arbitrarily created 100 bins of CAP_{100} and computed the proportion of individuals in a stage for each bin.

Finally, to define progression subgroups, the distributions of the progression indices (PIN, CAP_{100}, CAP) were examined by stage for the combined sample (observed and imputed data). To address the overlap of these distributions among stages, optimal cut-point analysis was conducted. A non-parametric method [20] was used to determine the optimal cut-point of the progression index that best separated two adjacent stages in terms of maximizing the product of sensitivity and specificity [21]. The area under the receiver-operator characteristic curve (AUC) was computed as an index of the optimal cut-point classification performance. (The optimal cut-points computed here should not be confused with the landmark cut-off thresholds for the HD-ISS conditions.)

The analysis was performed with the `R` computing platform [22] (version 4.1.3). The `ggplot2` [23] package was used for graphing, `missRanger` [24] for `MissForest` data imputation, `MatchIt` [25] for propensity score matching, `cutpointr` [20] for cut-point estimation, and graph smoothers were generalized additive (mixed) models (GAMs or GAMMs) estimated with `mgcv` [26].

Results

Table 1 shows baseline (first visit) descriptive statistics for key variables by study. The table indicates that on average, Enroll-HD had the oldest and most progressed HD participants, whereas PREDICT-HD had the youngest and least progressed. For example, Enroll-HD had $\overline{\text{PIN}} = 2.23$ and PREDICT-HD had $\overline{\text{PIN}} = -0.01$, with IMAGE-HD and TRACK-HD/ON in the middle, both with $\overline{\text{PIN}} = 0.93$. In addition, the PIN range was much wider for Enroll-HD (-2.68 to 9.86) than PREDICT-HD (-2.15 to 3.40), indicating more phenotypic diversity in Enroll-HD.

Table 2 shows the imputation results. Columns C0-C3 are indicators of whether the criteria for Stage 0-3 were met (1 if met, 0 otherwise). For example, the pattern 1, 1, 1, 0 indicates that the stage criteria up to and including Stage 2 were met (resulting in a Stage 2 classification). Counts in the table are for visits, and the patterns of the C0-C3 criteria indicators in each section are sorted by mean CAP_{100}. Stage Count and Stage Proportion show how the algorithm assigned the stage for each Stage Criteria pattern, with the following exception: patterns 1-4 are the observed indicator patterns that were compatible with the HD-ISS and therefore not imputed.

All the stages for patterns 5-25 were imputed. Patterns 5-8 were incompatible with the HD-ISS because one or more stage criteria were met out of order. Patterns 9-25 had inconclusive classification because one or more landmark variables was missing (indicated by a dot). Patterns 17-19 were the most frequent in Enroll-HD and could be consistent with staging (if the imaging criteria were met).

Figure 1 shows observed and imputed scores of four landmark variables as a function of CAP_{100} and imputation status (observed: 1882 visits, imputed: 39312 visits – with stage and caudate volume imputed for all visits, but only missing values imputed for the other variables). Putamen and IS were omitted because their results were very similar to caudate and TFC, respectively. The configuration of the imputed stages was similar to that of the observed stages in the sense that the stages tended to occur at similar CAP_{100} (e.g., Stage 0 was associated with small CAP_{100}, and Stage 3 was associated with large CAP_{100}). The imputed database had a wide range of progression and the GAM curves for the imputed data tended to decelerate (or plateau) for larger CAP_{100}. Overall, the imputed caudate volume (Panel A) was larger than the observed volume, which is illustrated by the GAM curves having different initial values. Both groups showed relatively orderly transitions from Stage 0 (left-most) through Stage 3 (right-most).

Propensity score matching resulted in the studies being much better balanced than without matching in terms of similarity of means and variances (see Table S1 in the supplementary material). However, slight differences remained, the largest being a mean difference in TMS. Figure 2 shows stage proportion as a function of CAP_{100} bin midpoint for the observed data and the matched imputed data. The GAM curves indicate very similar proportions in the range of $\text{CAP}_{100} \geq 80$ for all stages. However, there were sizable discrepancies for the range of $\text{CAP}_{100} <$ 80 with Stage 0 and 1. Panels in the upper row show that the `MissForest` algorithm tended to over-assign Stage 0 and under assign Stage 1 early in progression. The lower row indicates relatively minor under-assignment of Stage 2 early in progression and excellent correspondence with the observed data for all progression levels of Stage 3.

Figure 3 shows the longitudinal trends of key variables for Enroll-HD with baseline HD-ISS stage. The sample size (number of visits) for starting in Stage 0 to 3 was respectively, 2199 (22652), 1013 (10776), 1678 (17700), and 9292 (98824). The first column indicates that when starting in Stage 0, there tended to be progression through the stages over time. The last column shows that when starting in Stage 3, some regression to earlier stages did occur, but the vast majority persisted in Stage 3. The middle two panels show there was stage progression over time for a large majority, but some did regress.

Figure 4 shows the distributions of the progression indices for the combined data (all observed and imputed data). For all indices, the median increased with stage, but there was extensive overlap of distributions between stages. The overlap indicates that visits could be differently ordered by the HD-ISS and the progression indices. For example, the PIN distribution for Stage 2 shows that the lowest quarter of scores (below the bottom box edge) were smaller than the highest quarter of Stage 1 (above the upper box edge).

Table 3 shows key descriptive statistics for the progression indices by stage for the combined data. The overlap of the distributions reflected in the descriptive statistics motivated the optimal cut-point analysis to demarcate non-overlapping segments for each stage based on the progression indices. Results of the optimal cut-point analysis are shown in the last three columns of Table 3 (visualization is provided in Figure S1 of the supplementary material). By definition, Lwr and Upr provided limits that did not overlap among stages. The CAP_{100} cut-point was best at separating Stage 0 and 1 (AUC = 0.88), but PIN and CAP were close behind (AUCs = 0.86). The PIN cut-point was best at separating Stage 1 and 2 (AUC = 0.82), and the CAP scores had relatively poor performance (maximum AUC = 0.65). PIN performed well at separating Stage 2 and 3 (AUC = 0.88), and the CAPs performed less well (AUC $\approx$ 0.80).

Discussion

Our results show that by using a machine learning algorithm, the HD-ISS stage can be imputed when some of the landmark variables are missing. This finding is especially pertinent with Enroll-HD, for which imaging data are not currently collected. Imputation of all four stages is possible with Enroll-HD, but our results did show some discrepancies between imputed and observed values for early progression. The discrepancies might be accounted for by the pre-existing progression differences among the observed and imputed databases. There is evidence that the imputed stages can be treated similarly to the observed stages for the types of large-scale analysis supported by Enroll-HD, especially when the focus is on HD-ISS Stage > 1.

There are several practical implications for the application of the HD-ISS with the Enroll-HD database. First, the imputation results imply an alternative classification system that need not rely on imputation. Two of the most frequently occurring stage criteria indicator patterns of Enroll-HD had high consistency of stage assignment (Table 2 Pattern 18 and 19). When criteria for Stage 0 and 2 were met, but not Stage 3, the algorithm regularly assigned Stage 2 (84% of the time). When the criteria for Stage 0, 2, and 3 were met, the algorithm always assigned Stage 3 (100% within rounding). Therefore, for these observed patterns, if we were to assume that the Stage 1 criteria for brain atrophy were also met, we would often agree with the algorithm assignment. On the other hand, when the criteria for Stage 0 was met but was not for Stage 2 and 3 (Pattern 17), the algorithm assigned Stage 0 or 1 97% of the time. These results lead to the suggestion that if we are willing to collapse Stage 0 and 1, then the HD-ISS thresholds can be directly used to classify into the less precise categories of Stage $\leq$ 1, Stage 2, and Stage 3 without imputation. More simply stated: when Stage 1 criteria are missing, if the criteria for Stage 0, 2 and 3 are met, we classify in Stage 3; if the criteria for only Stage 0 and 2 are met, we classify in Stage 2; if the criterion for only Stage 0 is met, we classify in Stage $\leq$ 1. This approach grounds the classification in observed data, and it might suffice for much HD research that is currently focused on HD-ISS Stage 2 and 3. This also addresses the potential weakness of the imputation algorithm, which showed the greatest discrepancies in assigning Stage 0 and 1 for early progression.

A second practical advantage of our results is the definition of progression subgroups. The subgroups can help define treatment subpopulations to plan trials consistent with the HD-ISS. Subgroups based on PIN are most applicable to Stage 2, as the PIN combination contains the landmark variables for the stage. For Stage 3, subgroups might be defined by traditional TFC staging [34]. For example, early Stage 3 might be defined as TFC > 10, and early-to-mid Stage 3 defined as TFC > 6. Enrichment for Stage 1 (or Stage ≤ 1) is particularly challenging, as it is best to define subgroups using a biomarker. Imaging for enrichment is resource-intensive, and fluid biomarkers, such as neurofilament light chain, could possibly provide an alternative in the future.

An approach to clinical trial planning might involve the following steps. First, based on scientific considerations, the HD-ISS stage for participant recruitment is identified. Second, enrichment is used to define a more homogeneous subgroup within a stage. Third, the database is interrogated to estimate the untreated rate of change for a continuous endpoint (e.g., TMS), or the rate for a time-to-event endpoint (e.g., transition to Stage 3).

As an example of clinical trial planning for a continuous outcome, let us say the goal of a trial is to examine the effect of a treatment on TMS, that is, to slow its progression. Assume the treatment population is defined to be in Stage 2 at the start of the study and the goal is to exclude individuals who have recently entered the stage or are soon to exit. In this case, enrichment might focus on the middle of the Stage 2 PIN distribution, $0.47 < \text{PIN} < 1.84$. By applying the selection criteria to Enroll-HD and computing the relevant statistics, analytic formulas can be used to estimate the required sample size [27,28].

A similar strategy can be used for a time-to-event analysis in which a treatment is expected to, for example, delay the transition into Stage 3. In this design, suppose we want to recruit individuals whose first visit is in Stage 2, and follow them until they transition into Stage 3 or to the end of the study. The endpoint is time to any drop in the TFC or IS or both (i.e., entry into Stage 3). If we want to ensure that there is time for a measurable drug effect, the first visit should not be too close to Stage 3, and therefore selection might be below the median PIN in Stage 2, PIN < 1.09. After applying the selection criteria and computing statistics, equations for time-to-event (or survival) analysis can be used to estimate the required sample size [29,30].

A caveat for these trial planning scenarios is that Enroll-HD is not a treatment study, which means a placebo effect cannot be estimated. Placebo effects are caused by many factors [31], and there is evidence that they can be relatively strong in HD trials [32]. Data from completed HD trials can be used in planning to help account for placebo effects [33].

The above scenarios are just two examples, as enrichment can be used whenever there is a need to identify more homogeneous progression subgroups than are afforded by the HD-ISS stages themselves. Furthermore, subgroups need not be defined based on descriptive statistics as we have done in our analysis. Rather, different PIN or CAP_{100} ranges can be considered to optimize to the problem at hand.

It is best to use the progression indices within a single stage, as they may not properly indicate inter-stage progression. The HD-ISS stages provide classification into groups by disease status (no detectable pathology, displays of neurodegeneration, HD signs and symptoms, and functional deficits). Although the progression indices are powerful tools and are predictive of progression, they do not always agree with the HD-ISS regarding an individual's global clinical status. This is illustrated by the extensive overlap of the between-stage distributions. A sizable proportion of participants in Stage 3 (for example) have CAP_{100} (and PIN) values that are smaller than participants in Stage 2. The scenario is not unexpected because while over 50% of the variation in disease progression is accounted for by age and CAG, these variables are not the only determinant factors in explaining disease severity. The landmarks for Stage 3 that signal functional loss are not accounted for by CAP_{100} (or PIN). Therefore, the power of using these methods is in their combination. The overlap indicates that mixing the progression indices with the HD-ISS could result in individuals not being properly ordered in one sense or the other and is why we recommend use of PIN (or CAP_{100}) only within an individual stage. If compelled to consider the progression indices for use across stages, then proper ordering on both dimensions will be facilitated if we select the inner 50% of the distributions for each stage. This refinement will help to define progression groups that are ordered similarly for both the HD-ISS and the progression index at the group level.

There are a few caveats that deserve comment. The discrepancy between the imputed and observed brain volume is difficult to definitively resolve because of the progression differences among the databases. The training databases (PREDICT-HD, TRACK-HD/ON) for the imputation were different than the database on which the imputation was mainly applied (Enroll-HD for the most part). The study differences may not have been completely accounted for in the propensity score matching, which is never perfect. The slowing in the rate of loss for the imputed volume late in progression is biologically feasible (see Figure 1). It is logical that at some point, striatal loss becomes exhausted, leading to a slowing down in the rate of deterioration. A more sophisticated imaging processing approach using higher quality scans from PREDICT-HD did reveal the late slowing that was not apparent from our FreeSurfer version 6 results [35]. The Enroll-HD database had a much higher density of advanced progression visits than the observed database, which perhaps enables us to get a glimpse of the nature of striatal loss very late in the disease. However, the imputation is a type of extrapolation beyond the progression bounds of the observed volumes, thus the accuracy is unknown.

The `MissForest` algorithm is stochastic, meaning that when it is re-run with the same data the results will change. We did repeat the imputation (results not reported) and found that the overall stage classification proportions changed very little, by a maximum of approximately 0.1%. The imputed HD-ISS stage for each Enroll-HD visit is included in the recent Enroll-HD periodic data set (PDS6, released December 2022). Because of the stochastic stage assignment, each participant visit also has associated variables for the probability of assignment for each HD-ISS stage. This will help researchers understand the reliability of the imputed stage designations for proper applications.

The imputation algorithm treated visits from the same participant as independent, thus ignoring the nested nature of the data. Since the majority of participants in the analysis had repeated visits (69%), an argument can be made for using a longitudinal multiple imputation approach [36]. However, the HD-ISS is fundamentally a cross-sectional system and our intent was to be consistent with this design. If our approach needs to be defended, we would say that the HD-ISS stages will probably not be used as a primary outcome to be modeled longitudinally over time. Rather, we anticipate that the HD-ISS will be used as in our examples above, to anchor individuals within a stage based on a single visit, or to fix a stage transition event for a time-to-event analysis. The evidence from our analysis is that the stage imputation is adequate for these types of uses.

Single imputation was implemented rather than the multiple imputation that is recommended for general applied data analysis [37]. The `MissForest` algorithm does impute multiple values internally, but there is a final single value that is chosen by a type of popular vote among the random forests [13]. The primary reason for using single imputation in our study was to assess whether `MissForest` might be a successful approach. In the first trial analysis scenario discussed above, the single imputation might be sufficient because the HD-ISS was considered for selection of participants and not for use in the data to be collected over the trial. It is unclear if the added complexity of multiple imputation would improve the accuracy of the imputed HD-ISS stages for this preliminary step; additional research is needed. Should one want to use the HD-ISS in an analysis model, then the `MissForest` algorithm can be run several times to generate multiple imputed data sets. The analysis model can be fitted to the data sets and results combined using standard methods [37].

Finally, there are many machine learning methods that can be used for imputation [38]. The chained equations approach used here has been shown to provide good performance in a wide variety of scenarios [15], and random forest is known to provide good all-around prediction performance [14]. Whether there are benefits of using alternative machine learning approaches with or without chained equations is a topic for future research.

In summary, we have shown that despite missing brain imaging variables, observational studies such as Enroll-HD can be staged according to the HD-ISS, perhaps most reliably with a three-category staging scenario. Progression subgroups within a stage can be defined to hone the definition of treatment populations. Our hope is that this information will facilitate the use of the HD-ISS to aid in the planning of interventional studies in HD.

Acknowledgments

Data used in this work were generously provided by the research participants in Enroll-HD and made available by the CHDI Foundation Inc. IMAGE-HD data used in this work were generously provided by the participants in the IMAGE-HD study. TRACK-HD/ON data used in this work were generously provided by the participants in the TRACK-HD and TRACK-ON studies and made available by Dr. Sarah Tabrizi, principal investigator, University College London. PREDICT-HD data used in this work were generously provided by the participants in the PREDICT-HD study and made available by the PREDICT-HD investigators and coordinators of the Huntington study group, Jane S. Paulsen, principal investigator.

Conflict of Interest

Jeffrey D. Long reports personal compensation from Alynlam, Annexon, AskBio, Prilenia, PTC, Remix, Roche, Spark, Triplet, uniQure, Vertex, and Wave. His funding comes from CHDI and NIH.

Emily C. Gantman is an employee and receives salary from CHDI Management.

James A. Mills reports personal compensation from PTC and Triplet. His funding comes from CHDI and NIH.

Jatin G. Vaidya's funding comes from CHDI and NIH.

Alexandra Mansbach is a consultant to CHDI Management.

Sarah J. Tabrizi reports personal fees from F Hoffmann La Roche, Annexon, PTC Therapeutics, Takeda Pharmaceuticals, Vertex Pharmaceuticals, Alnylam Pharmaceuticals, Alphasights, Genentech, LoQus23 Therapeutics, Triplet Therapeutics, Novartis, Atalanta, Spark Therapeutics, Horama, University College Irvine, and Guidepoint; a patent application (2105484.6) and structural analogues licensed to Adrestia Therapeutics; funding from the CHDI Foundation, the UK Dementia Research Institute that receives its funding from DRI, the UK Medical Research Council, Alzheimer's Society, and Alzheimer's Research UK, and the Wellcome Trust (200181/Z/15/Z).

Cristina Sampaio is an employee and receives salary from CHDI Management, and has received consultancy honorariums (unrelated to Huntington's disease) from Pfizer, Kyowa Kirin, vTv Therapeutics, GW pharmaceuticals, Neuraly, Neuroderm, Green Valley Pharmaceuticals, and Pinteon Pharmaceuticals.

Supplementary Material

Contains a description of the `MissForest` algorithm, results of the propensity score matching, and a figure illustrating the optimal cut-point analysis.

References

[1] Tabrizi SJ, Schobel S, Gantman EC, Mansbach A, Borowsky B, Konstantinova P, et al. A biological classification of Huntington's disease: The integrated staging system. The Lancet Neurology 2022;21:632–44. https://doi.org/https://doi.org/10.1016/S1474-4422(22)00120-X.

[2] Huntington Study Group. Unified Huntington's Disease Rating Scale reliability and-consistency. Movement Disorders 1996;11:136–42.

[3] Landwehrmeyer BG, Fitter-Attas C, Giuliano J, et al. Data analytics from Enroll-HD, a global clinical research platform for Huntington's disease. Movement Disorder Clinical Practice 2016;4:212–24.

[4] Long JD, Mills JA. Joint modeling of multivariate longitudinal data and survival data in several observational studies of huntington's disease. Medical Research Methodology 2018;18:138–53.

[5] Zhang Y, Long JD, Mills JA, Warner JH, Lu W, Paulsen JS. Indexing disease progression at study entry with individuals at-risk for Huntington disease. American Journal of Medical Genetics Part B Neuropsychiatric Genetics 2011;156:751–63.

[6] Long JD, Langbehn DR, Tabrizi SJ, Landwehrmeyer BG, Paulsen JS, Warner J, et al. Validation of a prognostic index for Huntington's disease. Movement Disorders 2017;32:256–63.

[7] Poudel GR, Stout JC, Churchyard A, Chua P, Egan GF, Georgiou-Karistianis N. Longitudinal change in the white matter microstructure in Huntington's disease: The IMAGE-HD study. Neurobiology of Disease 2015;74:406–12.

[8] Paulsen JS, Long JD, Ross CA, Harrington DL, Erwin CJ, Williams JK, et al. Prediction of manifest Huntington's disease with clinical and imaging measures: A prospective observational study. Lancet Neurology 2014;13:1193–201.

[9] Tabrizi SJ, Scahill RI, Owen G, Durr A, Leavitt BR, Roos RA, et al. Predictors of phenotypic progression and disease onset in premanifest and early-stage Huntington's disease in the TRACK-HD study analysis of 36-month observational data. Lancet Neurology 2013;12:637–49.

[10] Liu D, Long JD, Zhang Y, Raymond LA, Marder K, Rosser A, et al. Motor onset and diagnosis in Huntington disease using the diagnostic confidence level. Journal of Neurology 2015;262:2691–8. https://doi.org/https://doi.org/10.1007/s00415-015-7900-7.

[11] United Nations Educational, Scientific, and Cultural Organization. ISCED 1997: International standard classification of education. New York: United Nations; 1997.

[12] Warner JH, Long JD, Mills JA, Langbehn DR, Ware J, Mohan A, et al. Standardizing the CAP score in Huntington's disease by predicting age-at-onset. Journal of Huntington's Disease 2022;11:153–71. https://doi.org/https://doi.org/10.3233/JHD-210475.

[13] Stekhoven DJ, Buehlmann P. MissForest - nonparametric missing value imputation for mixed-type data. Bioinformatics 2012;28:112–8.

[14] Breiman L. Random forests. Machine Learning 2001;45:5–32.

[15] Buuren S van, Groothuis-Oudshoorn K. Mice: Multivariate imputation by chained equations in r. Journal of Statistical Software 2011;45:1–67. https://doi.org/https://doi.org/10.18637/jss.v045.i03.

[16] Bondarenko I, Raghunathan T. Graphical and numerical diagnostic tools to assess suitability of multiple imputations and imputation models. Statistics in Medicine 2016;35:3007–20. https://doi.org/https://doi.org/10.1002/sim.6926.

[17] Nguyen CD, Carlin JB, Lee KJ. Model checking in multiple imputation: An overview and case study. Emerging Themes in Epidemiology 2017;14:1–2. https://doi.org/https://doi.org/10.1186/s12982-017-0062-6.

[18] Paulsen JS, Long JD. Onset of Huntington's disease: Can it be purely cognitive? Movement Disorders 2014;29:1342–50. https://doi.org/https://doi.org/10.1002/mds.25997.

[19] Rosenbaum PR, Rubin DB. The central role of the propensity score in observational studies for causal effects. Biometrika 1983;70:41–55. https://doi.org/https://doi.org/10.1093/biomet/70.1.41.

[20] Thiele C, Hirschfeld G. cutpointr: Improved estimation and validation of optimal cutpoints in R. Journal of Statistical Software 2021;98:1–27. https://doi.org/10.18637/jss.v098.i11.

[21] Liu X. Classification accuracy and cut point selection. Statistics in Medicine 2012;31:2676–86. https://doi.org/https://onlinelibrary.wiley.com/doi/10.1002/sim.4509.

[22] R Core Team. R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing; 2022.

[23] Wickham H. ggplot2 Elegant graphics for data analysis. New York: Springer-Verlag; 2009.

[24] Mayer M. missRanger: Fast imputation of missing values. 2021.

[25] Ho DE, Imai K, King G, Stuart EA. MatchIt: Nonparametric preprocessing for parametric causal inference. Journal of Statistical Software 2011;42:1–28. https://doi.org/10.18637/jss.v042.i08.

[26] Wood SN. Generalized additive models: An introduction with r. 2nd ed. Chapman; Hall/CRC; 2017.

[27] Zhao Y, Edland SD. Power formulas for mixed effects models with random slope and intercept comparing rate of change across groups. The International Journal of Biostatistics 2021;18:173–82. https://doi.org/doi:10.1515/ijb-2020-0107.

[28] Long JD, Paulsen JS, Marder K, Zhang Y, Kim J, Mills JA. Tracking motor impairments in the progression of Huntington's disease. Movement Disorders 2014;29:311–9.

[29] Long JD, Mills JA, Leavitt BR, Durr A, Roos RA, Stout JC, et al. Survival endpoints for Huntington's disease trials prior to a motor diagnosis. JAMA Neurology 2017;74:1–9.

[30] Machin D, Campbell MJ, Tan SB, Tan SH. Sample size tables for clinical studies. Hoboken, NJ: Wiley-Blackwell; 2009.

[31] Hafliadóttir SH, Juhl CB, Nielsen SM, Henriksen M, Harris IA, Bliddal H, et al. Placebo response and effect in randomized clinical trials: Meta-research with focus on contextual effects. Trials 2021;22:1–5.

[32] Reilmann R, McGarry A, Grachev ID, Savola JM, Borowsky B, Eyal E, et al. Safety and efficacy of pridopidine in patients with Huntington's disease (PRIDE-HD): A phase 2, randomised, placebo-controlled, multicentre, dose-ranging study. Lancet Neurology 2019;18:165–76. https://doi.org/https://doi.org/10.1016/S1474-4422(18)30391-0.

[33] Kieburtz K, Reilmann R, Olanow CW. Huntington's disease: Current and future therapeutic prospects. Movement Disorders 2018;33:1033–41.

[34] Shoulson I, Fahn S. Huntington disease clinical care and evaluation. Neurology 1979;29:1–3.

[35] Liu CF, Younes L, Hinkle JT, Tong X, Wang M, Phatak S, et al. Longitudinal imaging highlights preferential basal ganglia circuit atrophy in Huntington's disease 2023. Manuscript under review.

[36] Grund S, Lüdtke O, Robitzsch A. Multiple imputation of missing data for multilevel models: Simulations and recommendations. Organizational Research Methods 2018;21.

[37] Rubin DB. Multiple imputation for nonresponse in surveys. New York: John Wiley & Sons; 1987.

[38] Emmanuel T, Maupong T, Mpoeleng D. et al. A survey on missing data in machine learning. Journal of Big Data 2021;8:140-. https://doi.org/10.1186/s40537-021-00516-9

Tables

***Table 1**: Baseline (first-visit) descriptive statistics of key variables by study.*

Variable	Study	Missing Count	Proportion Complete	Mean	SD	Min	Q1	Q2	Q3	Max
Age	Enroll-HD	0	1.00	49.23	13.49	18.03	39.32	49.51	58.97	91.57
Age	IMAGE-HD	0	1.00	46.59	10.81	23.93	39.05	45.95	53.77	70.84
Age	PREDICT-HD	0	1.00	40.00	10.16	18.64	31.66	39.74	47.20	67.90
Age	TRACK-HD/ON	0	1.00	43.73	9.95	18.60	36.90	42.70	50.70	64.10
CAG	Enroll-HD	0	1.00	43.24	2.42	40.00	41.00	43.00	45.00	50.00
CAG	IMAGE-HD	0	1.00	42.86	2.18	40.00	41.00	42.50	44.00	50.00
CAG	PREDICT-HD	0	1.00	42.64	2.11	40.00	41.00	42.00	44.00	50.00
CAG	TRACK-HD/ON	0	1.00	43.37	2.26	40.00	42.00	43.00	45.00	50.00
Female	Enroll-HD	0	1.00	0.54	0.50	0.00	0.00	1.00	1.00	1.00
Female	IMAGE-HD	0	1.00	0.50	0.50	0.00	0.00	0.50	1.00	1.00
Female	PREDICT-HD	0	1.00	0.64	0.48	0.00	0.00	1.00	1.00	1.00
Female	TRACK-HD/ON	0	1.00	0.55	0.50	0.00	0.00	1.00	1.00	1.00
TMS	Enroll-HD	101	0.99	26.92	24.00	0.00	5.00	23.00	42.00	124.00
TMS	IMAGE-HD	0	1.00	10.46	12.87	0.00	0.25	6.00	16.75	60.00
TMS	PREDICT-HD	82	0.91	5.33	5.95	0.00	1.00	3.00	8.00	40.00
TMS	TRACK-HD/ON	0	1.00	11.61	12.51	0.00	2.00	5.00	21.00	52.00
SDMT	Enroll-HD	867	0.94	31.44	17.64	0.00	18.00	29.00	45.00	101.00
SDMT	IMAGE-HD	0	1.00	43.59	13.12	18.00	34.00	45.50	51.75	74.00
SDMT	PREDICT-HD	86	0.91	50.72	11.58	16.00	44.00	50.00	58.00	92.00
SDMT	TRACK-HD/ON	0	1.00	44.32	13.49	12.00	35.00	44.00	53.00	80.00
PIN	Enroll-HD	940	0.93	2.23	2.19	-2.68	0.35	2.31	3.78	9.86
PIN	IMAGE-HD	0	1.00	0.93	1.41	-1.63	-0.20	0.51	2.02	4.45
PIN	PREDICT-HD	93	0.90	-0.01	0.93	-2.15	-0.67	-0.07	0.54	3.40
PIN	TRACK-HD/ON	0	1.00	0.93	1.37	-1.39	-0.17	0.52	1.97	4.60
CAP$_{100}$	Enroll-HD	0	1.00	97.98	23.11	28.74	82.81	101.92	114.37	190.22
CAP$_{100}$	IMAGE-HD	0	1.00	90.33	17.15	40.57	79.49	90.66	103.75	129.11
CAP$_{100}$	PREDICT-HD	0	1.00	76.26	16.39	34.84	64.61	76.44	87.68	147.06
CAP$_{100}$	TRACK-HD/ON	0	1.00	87.80	14.90	53.81	78.12	85.61	97.69	128.74
CAP	Enroll-HD	0	1.00	455.71	115.78	118.24	379.01	469.48	533.96	952.15
CAP	IMAGE-HD	0	1.00	415.71	87.66	175.68	355.46	413.83	481.15	632.88
CAP	PREDICT-HD	0	1.00	348.55	80.65	149.81	294.14	351.07	401.04	721.56
CAP	TRACK-HD/ON	0	1.00	409.77	75.48	242.69	352.50	400.36	461.52	638.89

Note. SD = standard deviation, Min = minimum, Q1 = first quartile (25th percentile), Q2 = second quartile (50th percentile), Q3 = third quartile (75th percentile), Max = maximum

***Table 2**: Results of the imputation. Patterns 1-4 have observed HD-ISS stages, whereas 5-25 have imputed HD-ISS stages. Missing data is indicated by a dot (.).*

	Stage Criteria				Study Visit Count					Stage Count				Stage Proportion				Mean	
Pattern	C0[a]	C1[b]	C2[c]	C3[d]	Total	Enroll	IMAGE	PREDICT	TRACK	S0	S1	S2	S3	P0	P1	P2	P3	CAP$_{100}$	PIN
1	1	0	0	0	442	0	0	244	198	442	0	0	0	1	0	0	0	73.41	-0.52
2	1	1	0	0	534	0	0	356	178	0	534	0	0	0	1	0	0	82.31	-0.04
3	1	1	1	0	506	0	0	306	200	0	0	506	0	0	0	1	0	87.81	0.92
4	1	1	1	1	400	0	0	125	275	0	0	0	400	0	0	0	1	99.68	2.36
5	1	0	1	0	139	0	0	93	46	36	0	102	1	0.26	0	0.73	0.01	75.73	0.14
6	1	0	1	1	37	0	0	15	22	1	0	0	36	0.03	0	0	0.97	83.53	0.69
7	1	0	0	1	33	0	0	10	23	14	1	0	18	0.42	0.03	0	0.55	83.91	-0.03
8	1	1	0	1	49	0	0	32	17	0	6	0	43	0	0.12	0	0.88	85.38	0.11
9	1	0	.	0	2	0	0	2	0	2	0	0	0	1	0	0	0	58.97	-0.86
10	1	0	0	.	176	0	0	176	0	168	8	0	0	0.95	0.05	0	0	69.86	-0.57
11	1	0	.	.	115	0	42	73	0	89	4	20	2	0.77	0.03	0.17	0.02	71.63	-0.33
12	1	0	1	.	83	0	6	77	0	15	1	65	2	0.18	0.01	0.78	0.02	75.21	0.27
13	1	1	0	.	164	0	0	164	0	7	151	6	0	0.04	0.92	0.04	0	82.00	-0.06
14	1	1	.	0	5	0	0	3	2	0	4	1	0	0	0.80	0.20	0	85.68	0.35
15	1	1	.	.	186	0	47	137	2	2	67	88	29	0.01	0.36	0.47	0.16	89.58	0.92
16	1	1	1	.	368	0	85	283	0	0	6	272	90	0	0.02	0.74	0.24	96.20	1.70
17	1	.	0	0	7,408	7,298	0	46	64	4,682	2,539	145	42	0.63	0.34	0.02	0.01	72.24	-0.45
18	1	.	1	0	4,143	4,062	0	25	56	237	145	3,494	267	0.06	0.03	0.84	0.06	91.27	1.28
19	1	.	1	1	24,818	24,750	0	12	56	7	0	1	24,810	0	0	0	1	111.22	3.93
20	1	.	0	.	90	18	0	72	0	59	29	2	0	0.66	0.32	0.02	0	75.84	-0.46
21	1	.	.	0	92	92	0	0	0	54	26	11	1	0.59	0.28	0.12	0.01	76.03	-0.16
22	1	.	0	1	1,066	1,058	0	2	6	141	32	0	893	0.13	0.03	0	0.84	84.87	0.19
23	1	.	.	.	67	39	0	28	0	18	4	26	19	0.27	0.06	0.39	0.28	88.68	1.16
24	1	.	1	.	113	38	3	72	0	3	2	80	28	0.03	0.02	0.71	0.25	93.72	1.59
25	1	.	.	1	158	158	0	0	0	3	0	0	155	0.02	0	0	0.98	113.79	4.03

[a] C0 = 1 if CAG $\geq$ 40 and 0 otherwise (dot indicates missing; all participants had $40 \leq$ CAG ≤ 50)

[b] C1 = 1 if either putamen or caudate or both are below threshold

[c] C2 = 1 if either TMS or SDMT or both are beyond threshold

[d] C3 = 1 if either TFC or IS or both are below threshold

***Table 3**: Descriptive statistics and optimal cut-point analysis results of the progression indices for the HD-ISS stages. Results are based on the combined data (observed and imputed).*

Variable	Stage	Mean	SD	Min	Q1	Q2	Q3	Max	Lwr	Upr	AUC
PIN	0	-0.76	0.58	-2.91	-1.14	-0.74	-0.37	2.51	Min	-0.34	-
PIN	1	0.12	0.58	-2.07	-0.28	0.13	0.50	3.19	-0.34	0.60	0.86
PIN	2	1.20	1.02	-1.76	0.47	1.09	1.84	5.60	0.60	2.31	0.82
PIN	3	3.75	1.89	-1.75	2.47	3.66	5.00	10.21	2.31	Max	0.88
CAP_{100}	0	64.61	13.06	28.74	55.73	64.58	73.34	147.06	Min	74.27	-
CAP_{100}	1	85.20	11.84	43.64	77.19	85.03	92.85	153.87	74.27	89.52	0.88
CAP_{100}	2	91.02	15.72	33.00	80.94	91.43	101.89	171.18	89.52	102.34	0.62
CAP_{100}	3	110.23	16.45	30.14	100.66	110.89	120.49	200.88	102.34	Max	0.81
CAP	0	294.53	64.74	118.24	249.88	294.82	336.35	721.56	Min	337.61	-
CAP	1	386.31	58.72	183.04	347.44	383.51	422.51	806.27	337.61	410.75	0.86
CAP	2	421.66	78.47	140.37	370.24	422.94	474.86	871.77	410.75	470.64	0.65
CAP	3	514.35	88.47	124.01	460.03	513.94	567.21	1005.46	470.64	Max	0.79

Note. SD = standard deviation, Min = minimum, Q1 = first quartile (25th percentile), Q2 = second quartile (50th percentile), Q3 = third quartile (75th percentile), Max = maximum; Lwr = lower stage limit based on the cut-point analysis, Upr = upper stage limit based on the cut-point analysis, AUC is the area under the receiver-operator curve for the previous stage versus the current stage

Figures

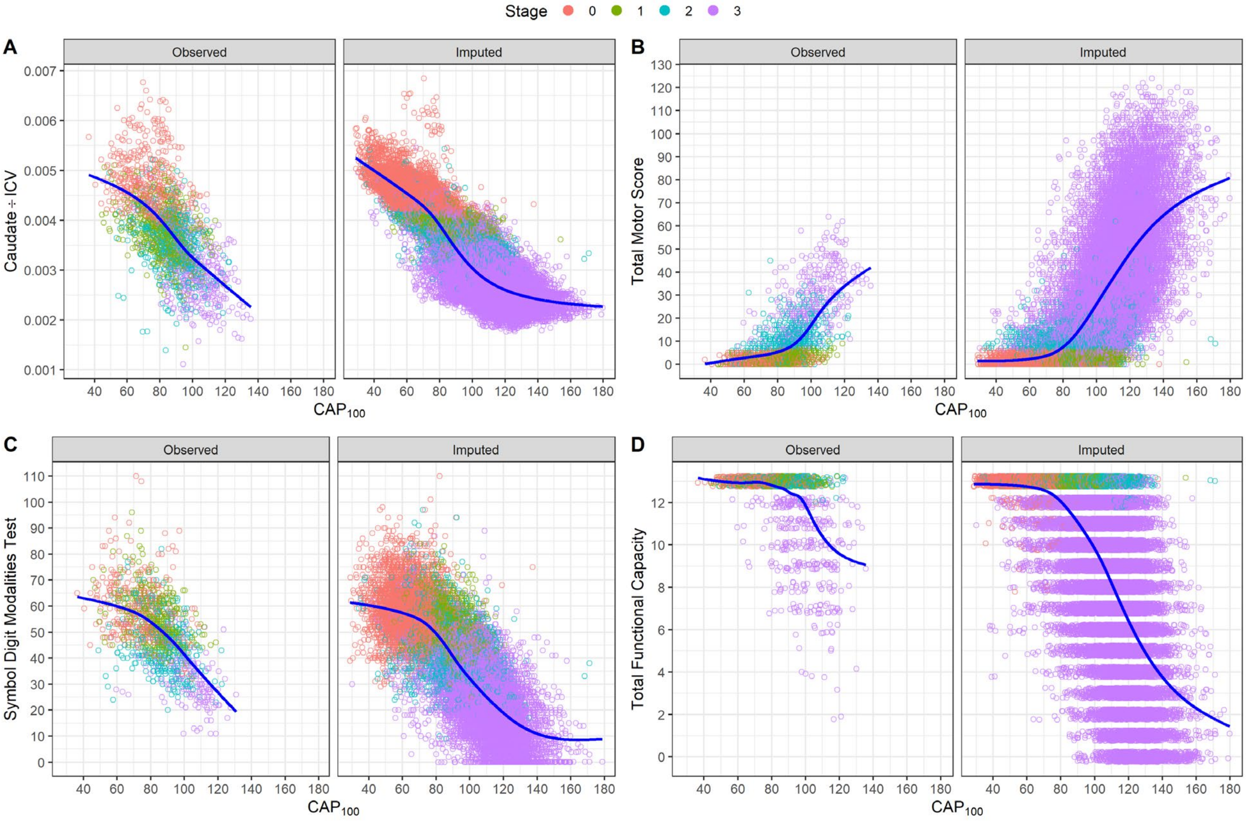

Figure 1. Observed and imputed scores of key variables by CAP_{100}, HD-ISS stage (color), and paneled by imputation status. Smooth curves are based on generalized additive models. Stage and caudate volume were fully imputed, but for the other variables only scores that were missing were imputed. ICV is total intra-cranial volume.

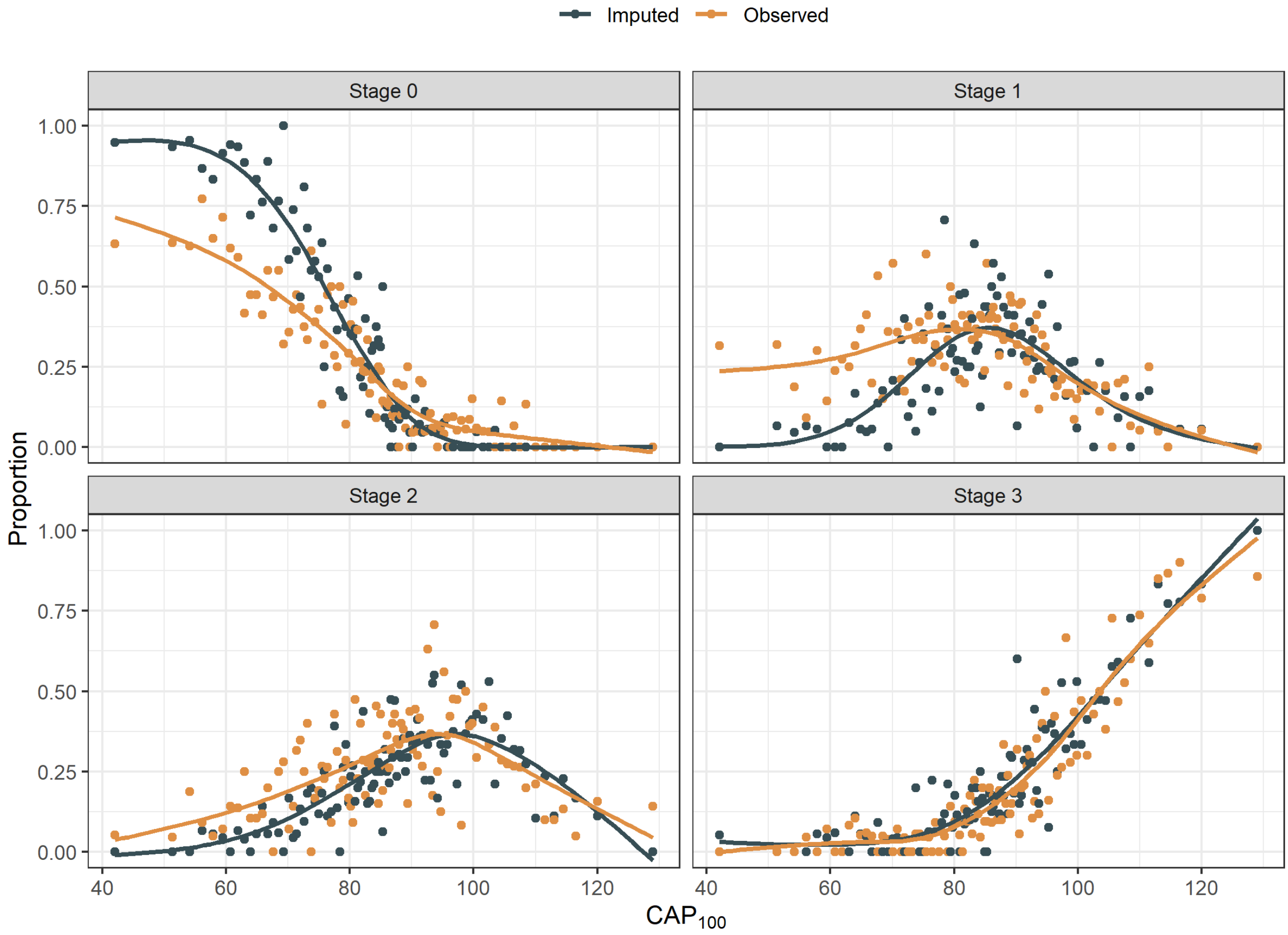

Figure 2. Propensity score matching results. Proportion of HD-ISS stage by CAP$_{100}$ and imputation status (observed, imputed). CAP$_{100}$ is *the midpoint of a bin range (100 bins total), and the smooth curves are based on generalized additive models.*

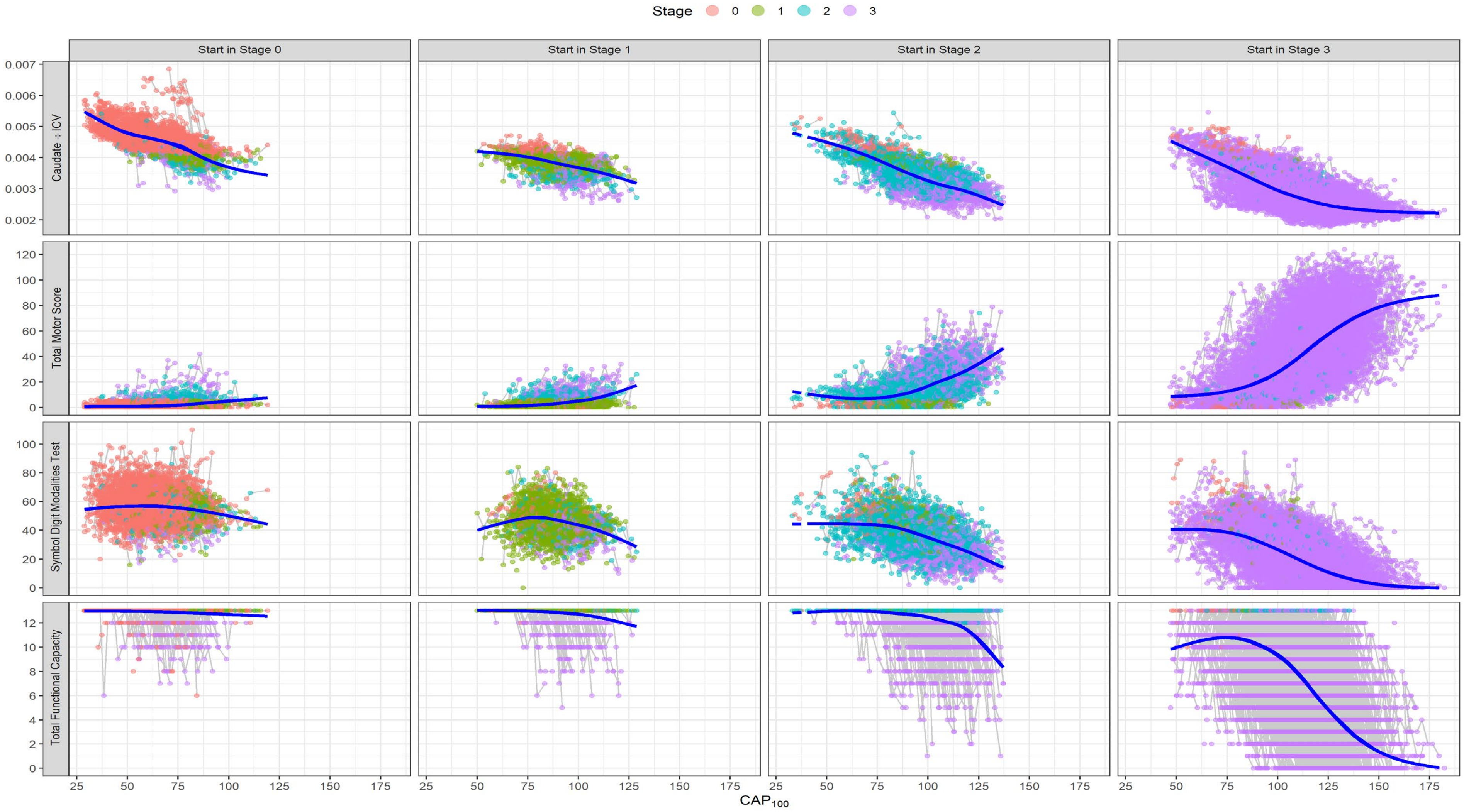

***Figure 3**. Enroll-HD longitudinal data of four variables (rows) for different HD-ISS starting stages (columns) with CAP$_{100}$ as the time metric (CAG-adjusted age). Stage and caudate volume (ICV = intracranial volume) were completely imputed, whereas other variables had partial imputation. Repeated visits of the same participant are connected by a thin line, and the smooth curves are based on group-level generalized additive mixed models.*

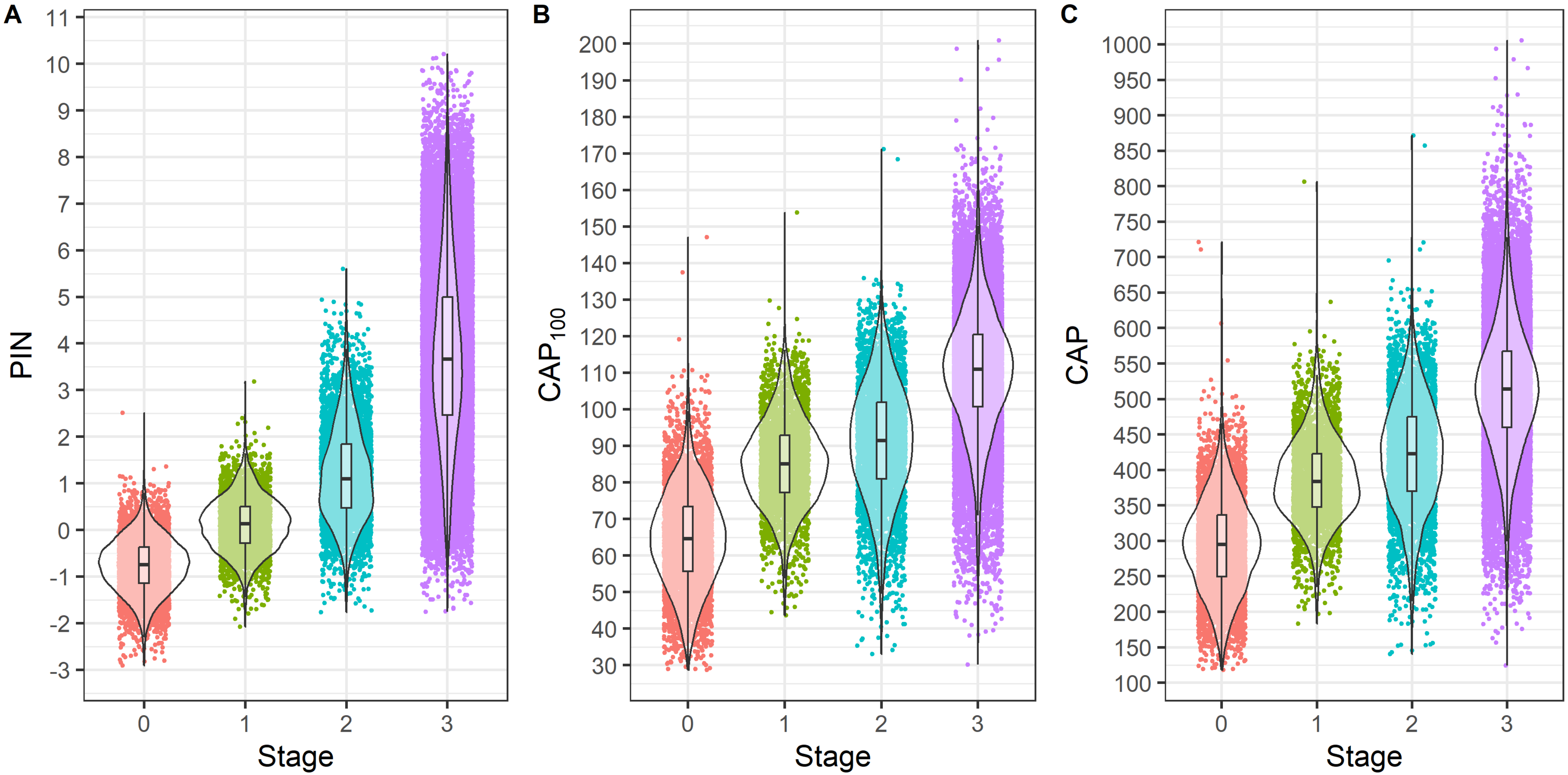

***Figure 4**. Distributions of progression indices as a function of HD-ISS stage for the combined data (observed and imputed). Jittered values (points) are shown with boxplots wrapped by violin plots. Panel A is for PIN, B depicts CAP_{100}, and C shows CAP.*

Supplementary Material

Overview of the `MissForest` algorithm

HD-ISS stage imputation was performed with the `MissForest`[13] algorithm that uses random forest[14] with chained equations[15]. Because Enroll-HD does not collect imaging variables, a database from studies that did collect imaging (PREDICT-HD and TRACK-HD/ON) was 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 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 stages. The chained equation approach has been shown to work well in simulation studies[15]. 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` algorithm proceeds as follows. Suppose we have variable vectors x_1, x_2, x_3, each with dimension $n \times 1$. 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. Note that for each row of y that has missing data, x_2 might also have missing data or not. 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[14] (1000 trees were used in our case). The RF 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 RF to compute predictions. After missing values on x_1 are imputed, we move on to setting x_2 to y, and a RF is similarly used to predict the non-missing values using the newly imputed x_1 and the (non-imputed) x_3. The newly trained RF 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.

Results of Propensity Score Matching

Balance before and after 1:1 matching of PREDICT-HD and TRACK-HD/ON to Enroll-HD is shown in Table S1.

***Table S1**: Balance statistics before and after propensity score matching.*

Matching	Variable	Mean Observed	Mean Imputed	Standardized Mean Difference	Variance Ratio
Before	Distance	0.1041	0.0497	1.0711	1.0190
Before	Age	43.7616	49.8036	-0.6274	0.5342
Before	CAG	42.9172	43.1415	-0.1023	0.8475
Before	TMS	9.8115	26.0703	-1.4082	0.2630
Before	SDMT	48.2360	31.5853	1.1952	0.5782
Before	TFC	12.3769	9.6763	1.7556	0.1925
Before	IS	97.2611	84.3836	1.8275	0.1779
Before	Educ	0.6668	0.5114	0.3297	NA
Before	Female	0.5830	0.5351	0.0972	NA
Before	DCL	1.7031	2.9310	-0.8086	0.8772
Before	CAP100	85.3139	98.4606	-0.8428	0.4889
After	Distance	0.1041	0.1031	0.0191	1.0849
After	Age	43.7616	43.7023	0.0062	0.9967
After	CAG	42.9172	42.9172	0.0000	1.0000
After	TMS	9.8115	8.9845	0.0716	0.9677
After	SDMT	48.2360	48.0737	0.0117	0.9206
After	TFC	12.3769	12.4106	-0.0219	1.1274
After	IS	97.2611	97.2931	-0.0045	1.0644
After	Educ	0.6668	0.6700	-0.0068	NA
After	Female	0.5830	0.5659	0.0347	NA
After	DCL	1.7031	1.6957	0.0049	0.8062
After	CAP100	85.3139	85.1962	0.0075	0.9965

Optimal Cut-Point Analysis

Related to Table 3 in the text, Figure S1 shows the densities of the progression indices by HD-ISS stage with the optimal cut-points indicated by vertical dashed lines.

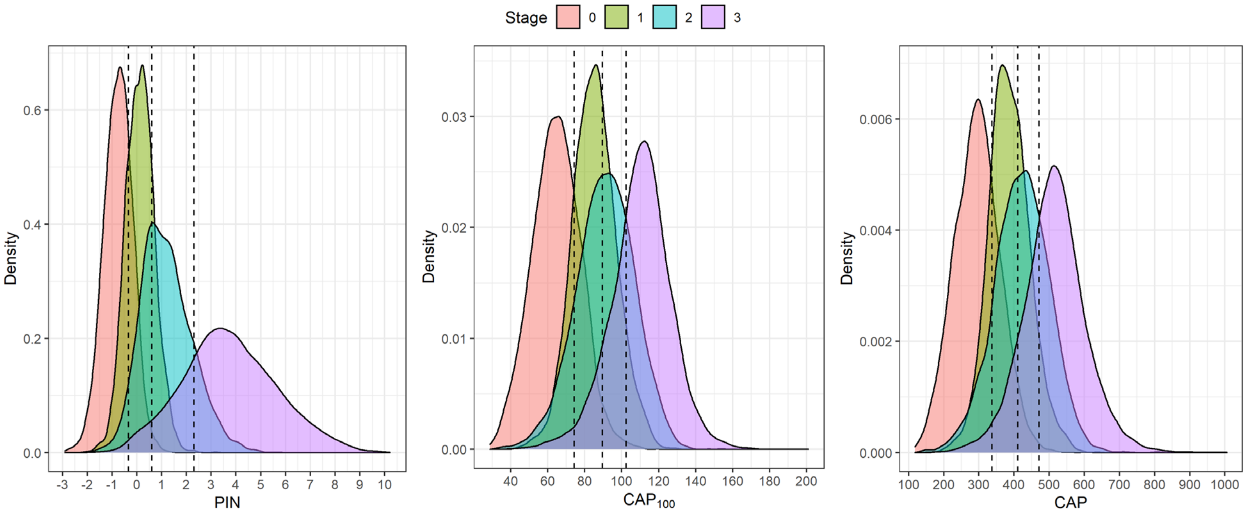

***Figure S1**. HD-ISS stages: densities and optimal cut-points for CAP, CAP$_{100}$, and PIN. Observed and imputed stages are combined.*