frontiers in Neurology

ORIGINAL RESEARCH
published: 20 May 2021
doi: 10.3389/fneur.2021.678484

Ranking the Predictive Power of Clinical and Biological Features Associated With Disease Progression in Huntington's Disease

Naghmeh Ghazaleh[1], *Richard Houghton*[1], *Giuseppe Palermo*[1], *Scott A. Schobel*[1], *Peter A. Wijeratne*[2,3] *and Jeffrey D. Long*[4,5]*

[1] F. Hoffmann-La Roche Ltd., Basel, Switzerland, [2] Department of Computer Science, Centre for Medical Imaging Computing, University College London, London, United Kingdom, [3] Department of Neurodegenerative Disease, Huntington's Disease Research Centre, Queen Square Institute of Neurology, University College London, London, United Kingdom, [4] Department of Psychiatry, University of Iowa, Iowa City, IA, United States, [5] Department of Biostatistics, University of Iowa, Iowa City, IA, United States

Huntington's disease (HD) is characterised by a triad of cognitive, behavioural, and motor symptoms which lead to functional decline and loss of independence. With potential disease-modifying therapies in development, there is interest in accurately measuring HD progression and characterising prognostic variables to improve efficiency of clinical trials. Using the large, prospective Enroll-HD cohort, we investigated the relative contribution and ranking of potential prognostic variables in patients with manifest HD. A random forest regression model was trained to predict change of clinical outcomes based on the variables, which were ranked based on their contribution to the prediction. The highest-ranked variables included novel predictors of progression—being accompanied at clinical visit, cognitive impairment, age at diagnosis and tetrabenazine or antipsychotics use—in addition to established predictors, cytosine adenine guanine (CAG) repeat length and CAG-age product. The novel prognostic variables improved the ability of the model to predict clinical outcomes and may be candidates for statistical control in HD clinical studies.

Keywords: Huntington's disease, disease progression, prognostic variables, machine learning, random forest

OPEN ACCESS

Edited by:
Emilia Mabel Gatto,
Sanatorio de la Trinidad Mitre, Argentina

Reviewed by:
Zhong Pei,
Sun Yat-Sen University, China
Yi-Ting Hsu,
China Medical University, Taiwan

***Correspondence:**
Jeffrey D. Long
jeffrey-long@uiowa.edu

Specialty section:
This article was submitted to Movement Disorders, a section of the journal Frontiers in Neurology

Received: *09 March 2021*
Accepted: *26 April 2021*
Published: *20 May 2021*

Citation:
Ghazaleh N, Houghton R, Palermo G, Schobel SA, Wijeratne PA and Long JD (2021) Ranking the Predictive Power of Clinical and Biological Features Associated With Disease Progression in Huntington's Disease. Front. Neurol. 12:678484. doi: 10.3389/fneur.2021.678484

INTRODUCTION

Huntington's disease (HD) is a rare, genetic, neurodegenerative disease caused by a cytosine adenine guanine (CAG) repeat expansion variant of the huntingtin gene (*HTT*) (1) and is characterised by a triad of cognitive, behavioural, and motor symptoms (2, 3). Disease onset, defined as the onset of motor signs and symptoms as measured by a Diagnostic Confidence Level of 4 (3, 4), typically occurs in the prime of life, between the ages of 30 and 50 years (2). HD is associated with increasing disability, worsening of function and loss of independence, leading to death within approximately 15 years of onset (2, 5). Motor and cognitive symptoms deteriorate steadily as the disease progresses (3, 6–9), while behavioural symptoms tend to be episodic (10).

With potential disease-modifying therapies for HD in clinical development (11), there is interest in measuring disease progression and characterising prognostic variables in order to improve the efficiency and accuracy of clinical trials (12). Prognostic variables can be used to identify a patient

TABLE 1 | Attrition table showing number of patients included after applying filters for each inclusion criterium.

Total initial population in the dataset: 15,301
1) Age at baseline = 25–65 years; diagnosis age >20 years: **6,432**
2) Manifest only (HDCAT = 3): **6,025**
3) DCL = 4: **5,694**
4) IS at baseline 100 ≥ IS > 70: **4,277**
5) At least 2 year's follow-up score for all four outcomes: **1,608**

Totals are for subjects with complete data on all outcome scores.
DCL, diagnostic confidence level; HDCAT, HD category; IS, Independence Scale.
The bold values indicate the number of patients included after applying each inclusion criterium.

TABLE 2 | Patient demographics.

Demographic	***N* = 1,608**
Age, years, mean (SD)	49.60 (9.32)
Male sex, *n* (%)	819 (50.9)
Region, *n* (%)	
Australasia	69 (4.3)
Europe	1,068 (66.4)
Latin America	9 (0.6)
North America	462 (28.7)
CAG repeat length, mean (SD)	43.93 (3.04)
CAP score, mean (%)	488.18 (82.35)
Shoulsen and Fahn Stage at baseline, *n* (%)	Stage I: 716 (44.53) Stage II: 731 (45.49) Stage III: 159 (9.89) Stage IIII: 2 (0.12)

CAG, cytosine adenine guanine; CAP, CAG-age product; SD, standard deviation.

population through an enrichment strategy to reduce interpatient variability in clinical trials or alternatively to enrich for faster progressors, and to eventually inform the optimum time to start treatment (12). Statistically controlling for prognostic baseline variables may also be important in non-randomised (e.g., open-label) studies as they could confound the relationship between treatment exposure and outcomes. Additionally, when testing hypotheses in randomised studies, the probability of detecting a treatment effect will usually increase by including prognostic variables as covariates in the analysis, as this would explain a significant amount of variability observed due to random error.

Large prospective cohort studies have shown that manifestations of progression, that is, clinical signs and symptoms of HD, as well as known biological predictors of progression such as CAG repeat length and CAG-age product (CAP) score, can predict clinical progression or motor onset (7, 13). However, no study has systematically ranked the importance of predictors of progression in a manifest HD population (i.e., after the onset of unequivocal motor symptoms).

Random forest (RF) regression models permit interrogation of large, complex clinical datasets to capture non-linear associations between multidimensional predictive variables and clinical outcomes with high predictive accuracy (14, 15). RF approaches are well-suited to classification and regression problems, such as identifying variables with predictive potential for disease progression from clinical datasets. Here, we use modern machine learning methods to examine a large number of HD variables to identify the most important predictors of progression on five clinical outcomes: total functional capacity (TFC), a measure of function; stroop word reading (SWR), a measure of attention and psychomotor processing speed; symbol digit modalities test (SDMT), a measure of executive function, visuo-spatial working memory, attention and processing speed; total motor score (TMS), a measure of motor function; and the composite unified HD rating scale (cUHDRS), an equally weighted composite outcome measure of the TFC, TMS, SDMT, and SWR that was developed based on an early manifest HD population (16). The large prospective Enroll-HD cohort (NCT01574053) is used to investigate the relative contribution and ranking of potential prognostic variables to predict clinical progression in a clinical trial-like manifest HD population.

RESULTS

The analysis included 1,608 individuals meeting typical criteria for clinical trials in manifest HD and with CAG repeats between 36 and 64 (filtering criteria shown in **Table 1**). Patient demographics are shown in **Table 2**.

The highest-ranked variables predictive of disease progression for each outcome are shown in **Figure 1** and the top 10 variables for each outcome shown in **Table 3**. CAP was found to be the most predictive variable for all outcomes and CAG repeat length was ranked as the second most important variable for all outcomes. Other prognostic variables associated with faster progression that ranked in the top 10 for at least three of the five outcomes were: age at diagnosis (all but SWR and TFC), being accompanied to clinic visits (for all outcomes), history of cognitive impairment (all but SWR), tetrabenazine use (for all outcomes) and antipsychotics use (all but TMS). The effect of these variables on disease progression trajectory as measured by the cUHDRS is shown in **Figure 2**.

The common variables among the top 10 most important features for all outcomes were: CAP score, CAG repeats, accompanied or unaccompanied at clinic visit, tetrabenazine use, antipsychotics use and having severe cognitive impairment.

Unadjusted R^2 measures were calculated for the RF models including CAG and CAP score only and compared with the model including all the features (**Table 4**). Using additional features with CAP and CAG can capture the variance of outcome by 17% more for cUHDRS and 15% more on average for the other outcomes. The slight improvement in model fit with CAP, CAG and age compared with the model built with the shared top 10 features could be due to the different cross-validation sets, and also the very high contributions of CAP and CAG to the model fit.

DISCUSSION

This analysis used real-world data from the large Enroll-HD registry and a machine learning algorithm to identify novel predictors of HD progression with significant impact on the slope of clinical decline observed over a 2-year follow-up period. The

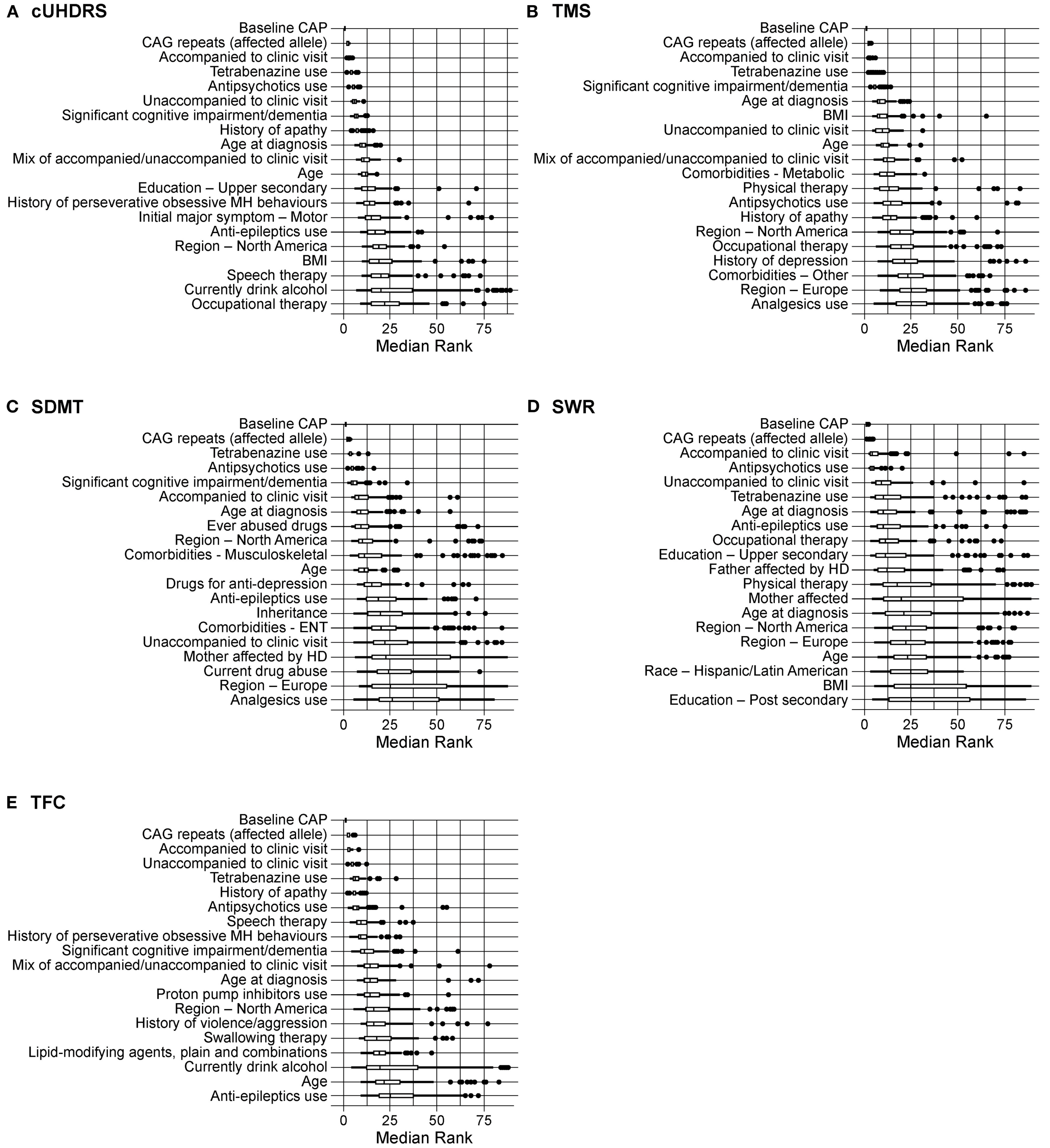

FIGURE 1 | Rankings of predictors of clinical progression as measured by **(A)** cUHDRS; **(B)** TMS; **(C)** SDMT; **(D)** SWR; **(E)** TFC. Boxplots are shown with the upper box edge representing the 75th quantile and the "whisker" extending to 1.5 times the IQR. A circle is an outlier, defined as a ranking that extends beyond a whisker. BMI, body mass index; CAG, cytosine adenine guanine; CAP, CAG-age product; cUHDRS, composite Unified HD Rating Scale; ENT, ear, nose, throat; IQR, interquartile range; MH, mental health; SDMT, symbol digit modalities test; SWR, stroop word reading; TFC, total functional capacity; TMS, total motor score.

two most important predictors identified were CAP score and CAG repeat length, in agreement with previous studies (7, 13). In addition, several strong predictors were identified that have either not been previously studied (being accompanied to a visit) or have had inconsistent effects in other studies (cognitive impairment, use of tetrabenazine or antipsychotics) (7, 17–19). The novel variables identified were predictive of progression over multiple clinical domains, measured by motor, cognitive and

TABLE 3 | Top 10 predictive variables for each outcome.

Rank	Outcome				
	cUHDRS	**TMS**	**TFC**	**SDMT**	**SWR**
1	Baseline CAP	Baseline CAP	Baseline CAP	Baseline CAP	Baseline CAP
2	CAG repeats (affected allele)	CAG repeats (affected allele)	CAG repeats (affected allele)	CAG repeats (affected allele)	CAG repeats (affected allele)
3	Accompanied to clinical visit	Accompanied to clinical visit	Accompanied to clinical visit	Tetrabenazine use	Accompanied to clinical visit
4	Tetrabenazine use	Tetrabenazine use	Unaccompanied to clinical visit	Antipsychotics use	Antipsychotics use
5	Antipsychotics use	Significant cognitive impairment or dementia	Tetrabenazine use	Significant cognitive impairment or dementia	Unaccompanied to clinical visit
6	Unaccompanied to clinical visit	Age at diagnosis	History of apathy	Accompanied to clinical visit	Tetrabenazine use
7	Significant cognitive impairment or dementia	BMI	Antipsychotics use	Age at diagnosis	Age at diagnosis
8	History of apathy	Unaccompanied to clinical visit	Speech therapy	History of drug abuse	Anti-epileptics use
9	Age at diagnosis	Age	History of perseverative obsessive MH behaviours	Region—North America	Occupational therapy
10	Mix of accompanied and unaccompanied to clinical visit	Mix of accompanied and unaccompanied to clinical visit	Significant cognitive impairment or dementia	Comorbidities—musculoskeletal	Education—upper secondary

BMI, body mass index; CAG, cytosine adenine guanine; CAP, CAG-age product; cUHDRS, composite Unified HD Rating Scale; MH, mental health; SDMT, symbol digit modalities test; SWR, stroop word reading; TFC, total functional capacity; TMS, total motor score.

functional endpoints, as well as the composite endpoint. Using all predictors in addition to known prognostic variables improved the ability of the model to predict clinical outcomes (see video abstract in the **Supplementary Materials**).

Some of the features tested, which may have been expected to rank highly as prognostic variables based on previous studies in premanifest HD (i.e., prior to the onset of unequivocal motor symptoms)—including smoking, alcohol intake and body mass index (BMI) (20–22)—were not found to be important predictors of progression. These results may not be directly comparable to the current study, which was carried out in a manifest HD population. It is also known that self-reporting of smoking and alcohol use is unreliable in the general population, as revealed by studies using advances in DNA methylation measurement to assess substance use status (23). We found BMI to be a weak discriminating factor among patients with different values of change in outcome. A further potential explanation for the disparity between our findings and previous studies could be that the prognostic value of these variables may be dependent on disease stage. In the current study, the population was relatively progressed, and it is possible that other variables associated with the disease could outweigh environmental variables.

It should be noted that we used an RF algorithm with the setting that prevents bias in the ranking based on the data structure. Whilst it is still possible that a feature can rank highly due to collinearity with another feature that is a strong predictor of the outcome, this could be prevented by calculating the conditional importance of the features, which is computationally very complex (24).

Additionally, the observed associations are based on observational data and are therefore not indicative of causal relationships, due to measured and unmeasured potential confounding factors. For example, being accompanied to clinic visits may affect the clinical outcome scores measured by virtue of the companion's additional report which informs the clinical rating. It may also be because healthier participants are able to continually attend visits alone, whereas those who are on worse clinical trajectories need additional emotional or practical assistance (e.g., driving) to complete visits. Similarly, antipsychotics may be used to treat motor symptoms in HD, and therefore may be expected to reduce TMS without influencing overall disease progression. Cognitive outcome measures may also be related to variables including dementia and severe cognitive impairment. RF approaches have good performance in modelling complex, multidimensional disease-specific datasets (like Enroll-HD) (25). In this application, an RF approach was used to find novel associations (e.g., identify variables with predictive potential for disease progression), and does not imply causality (i.e., the aetiological role of the variable during disease progression) (26). Nevertheless, by virtue of the strength of the associations observed, some of these features may be important to control for in analyses of observational studies and may have implications for companion participation in interventional trials.

The current study focused on clinical variables only and did not include imaging or fluid biomarkers, which previous studies have suggested may be predictive of disease progression (7, 27, 28). This limitation was due to the nature of the currently available HD databases. In this study, we used the

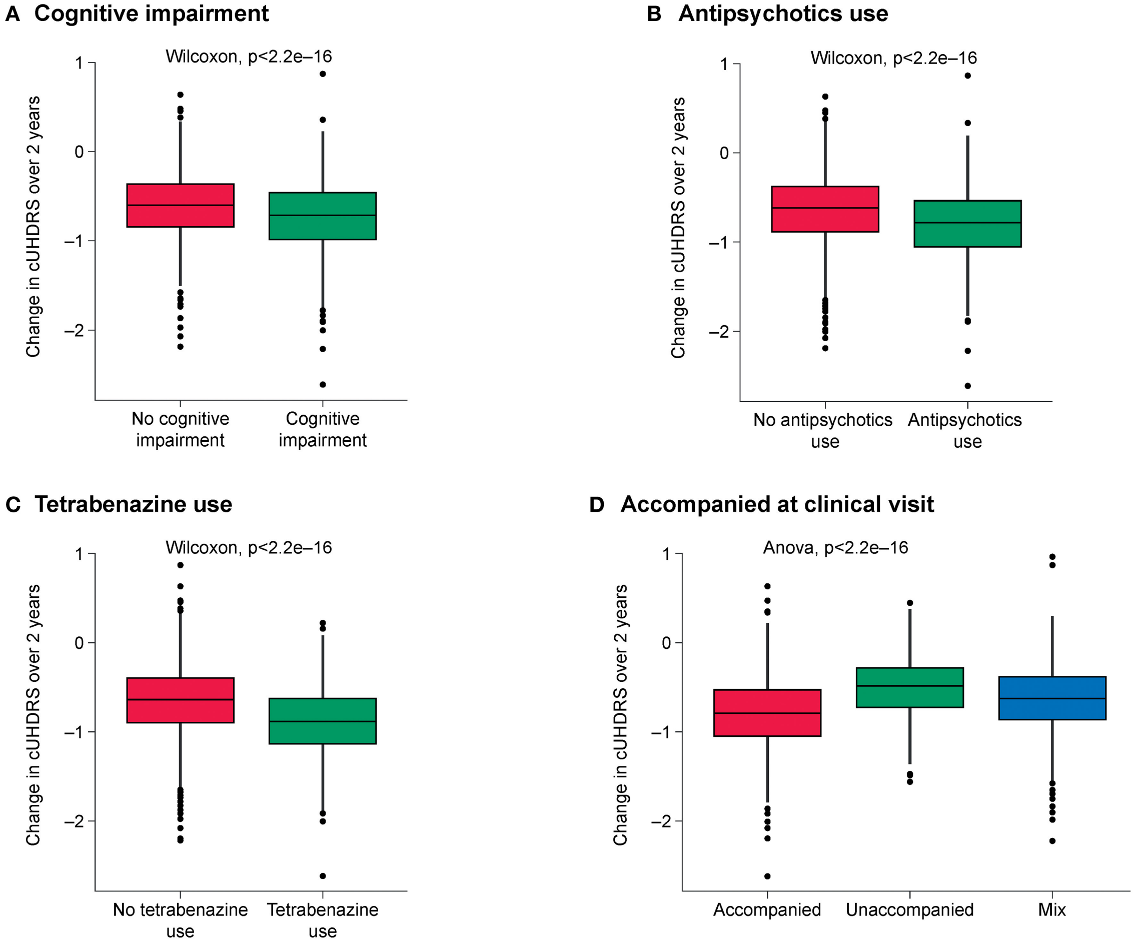

FIGURE 2 | Effect of the highest-ranked variables on clinical progression trajectory as measured by cUHDRS. **(A)** Cognitive impairment; **(B)** Antipsychotics use; **(C)** Tetrabenazine use; **(D)** Being accompanied at clinic visit. cUHDRS, composite Unified HD Rating Scale.

Enroll-HD database, which provides a sufficiently large sample size for the analysis but does not include imaging or biofluid data as part of the main study. Imaging databases such as PREDICT-HD (NCT00051324) are available, but do not provide the comprehensive range of clinical variables that is available in Enroll-HD, such as medication history. The available biofluid databases, such as the HD-CSF study (28, 29), are too small to be informative on the scale of the current analysis.

A further limitation is that the results described here are based on a selected cohort intended to reflect the inclusion criteria of ongoing clinical trials, and therefore may not be representative of the wider HD population, including younger patients (juvenile-onset HD), elderly patients (>65 years), late-stage patients (>Stage III) or premanifest patients. Further research is needed to determine the wider applicability of these results to these populations.

This study made use of a supervised RF regression model to identify putative and novel predictors of disease progression in HD. Identifying prognostic variables usually requires large sample sizes to optimise predictive accuracy, which may be a limitation for rare conditions such as HD. Since 2012, the Enroll-HD registry, which includes over 19,000 participants from 177 sites in 20 countries, has allowed a large, high-quality dataset to be available for researchers to advance the understanding of HD. The power of RF modelling is particularly relevant within the context of HD, where improved understanding of this multidomain disease and need for efficient trial design is evident.

To overcome known methodological limitations, RF approaches are being developed to harness the full potential of long-term registry data in clinical risk prediction and may, in future, accelerate disease risk and course prediction in HD. Given the dynamic nature of disease, the recently published RF Survival, Longitudinal and Multivariate model was developed to evaluate the temporal nature of variables (such as rate of change of variables) (30). Such approaches will further refine identification of clinically meaningful predictive variables not only for risk of disease progression as a static entity, but risk of disease progression over time and clinical course. Such temporal approaches will prove useful for future studies in HD, where disease course is highly variable.

In summary, the RF approach described here using the Enroll-HD dataset has identified novel prognostic variables which may

TABLE 4 | Comparison of model performance with all the features (1), with the discovered top-ranking features (2) and with only established prognostic features (3 and 4).

	Model	Outcome variable				
		cUHDRS	SDMT	TFC	SWR	TMS
1	R^2 with all the features	41%	28%	30%	26%	36%
2	R^2 with the shared top 10 features*	31%	18%	20%	14%	24%
3	R^2 with CAP, CAG and age at baseline	29%	19%	20%	16%	25%
4	R^2 with CAP and CAG	24%	14%	15%	12%	20%
	Difference between 4 and 1	17%	14%	15%	14%	16%

**Shared features among the individual top 10 important features of each outcome: CAP, CAG, accompanied or alone at visit, tetrabenazine use, antipsychotics use and cognitive impairment.*

CAG, cytosine adenine guanine; CAP, CAG-age product; cUHDRS, composite Unified HD Rating Scale; SDMT, symbol digit modalities test; SWR, stroop word reading; TFC, total functional capacity; TMS, total motor score.

TABLE 5 | List of candidate prognostic variables included in analyses.

Age
Age at diagnosis
Rater's judgement of initial major symptom—motor
Rater's judgement of initial major symptom—cognitive
Rater's judgement of initial major symptom—psychiatric
Rater's judgement of initial major symptom—oculomotor
Rater's judgement of initial major symptom—other
Rater's judgement of initial major symptom—mixed
Rater's judgement of initial major symptom—unknown or missing
CAG repeats (affected allele)
CAG repeats (unaffected allele)
Baseline CAP score—age*(affected CAG repeats−33.66)
Male sex
Female sex
Race—black
Race—Native American
Race—Asian
Race—Caucasian
Race—Hispanic/Latin American
Race—mixed
Race—other
Previous alcohol problems
Ever smoked
Currently smoke
Ever abused drugs
Current drug abuse
Currently drink alcohol
Father affected by HD
Mother affected by HD
Inheritance unknown
Education—bachelor's degree or higher
Education—post secondary but not university degree
Education—upper secondary or lower
Marital status—married
Marital status—separated
Marital status—single
Residence—rural
Residence—village
Residence—town
Residence—city
Residence—unknown
Region—Australasia
Region—Europe
Region—Latin America
Region—North America
BMI
Comorbidities—renal
Comorbidities—gynaecological
Comorbidities—reproductive
Comorbidities—dermatological
Comorbidities—musculoskeletal
Comorbidities—neurological
Comorbidities—metabolic
Comorbidities—psychiatric
Comorbidities—ENT
Comorbidities—gastrointestinal
Comorbidities—allergy: immunological
Comorbidities—pulmonary
Comorbidities—ophthalmological
Comorbidities—cardiovascular
Comorbidities—hepatobiliary
Comorbidities—haematological/lymphatic
Comorbidities—none
Comorbidities—other
Nutrition—vitamin
Nutrition—herbs
Nutrition—teas
Nutrition—other
Nutrition—homeopathic
Nutrition—aromatherapies
Non-pharmacological therapies—physical therapy
Non-pharmacological therapies—occupational therapy
Non-pharmacological therapies—psychotherapy
Non-pharmacological therapies—counselling
Non-pharmacological therapies—speech therapy
Non-pharmacological therapies—swallowing therapy
Non-pharmacological therapies—music therapy
Non-pharmacological therapies—relaxation therapy
Non-pharmacological therapies—acupuncture
Accompanied to clinic visit*
Unaccompanied to clinic visit*
Mix of accompanied and unaccompanied to clinic visit*
History of irritability
History of depression
History of violence/aggression
History of apathy
History of perseverative obsessive MH behaviours
History of psychosis (hallucinations MH or delusions)
Family history of psychotic illness in 1st degree relative
Significant cognitive impairment or dementia
History of motor symptoms MH compatible with HD
Drugs for anti-depression
Lipid-modifying agents, plain and combinations
Thyroid therapies
Antipsychotics use
Anxiolytics, hypnotics and sedatives use
Analgesics use
Tetrabenazine use
ACE inhibitors (plain and combination)
Anti-epileptics use
Protein pump inhibitors use
Dopaminergic therapies use

ACE, angiotensin-converting enzyme; CAG, cytosine adenine guanine; CAP, CAG-age product; ENT, ear, nose, throat; HD, Huntington's disease; MH, mental health.

**Accompanied/unaccompanied/mix of accompanied and unaccompanied to clinical visits were separated into a trichotomous variable. Accompanied means the patient did not come alone to all Enroll-HD study visits throughout follow-up (maximum three visits); Unaccompanied means they came alone to all visits; Mix of accompanied and unaccompanied means they sometimes came alone, but not always. This is the only candidate predictive variable based on some post-baseline data.*

be important candidates for statistical control in clinical trials and observational studies in HD.

METHODS

Data Source

Data from the Enroll-HD database were used for this study. Enroll-HD is a global platform designed to facilitate clinical research in HD. Core variables are collected annually from all research participants as part of this multicentre, longitudinal, observational study. Data are monitored for quality and accuracy using a risk-based monitoring approach. All sites are required to obtain and maintain local ethical approval. The study began recruiting in 2012, and as of data released in 2018, includes over 19,000 total participants and more than 8,000 patients with manifest HD. The second version of the fourth periodic dataset release (PDS4 version 2.0) was used, which has a data cut-off date of 31 October 2018 and was made available in August 2019.

Patient Population

The study population is purposely limited to individuals meeting typical criteria for clinical trials in manifest HD, using the filtering criteria shown in **Table 1**.

The primary population of interest was individuals with manifest HD aged 25–65 years, inclusive. Patients with juvenile-onset HD (age of first symptom onset at age <20 years) were excluded. Participants were required to have Independence Scale >70 at baseline and at least two subsequent annual visits with clinical information recorded. The rationale for these criteria is that the typical duration of clinical studies in this population is 2 years.

Analytical Approach

A total of 102 prognostic variables (**Table 5**) were considered for each participant, including demographics, clinical characteristics, comorbidities, symptoms, as well as pharmacological and non-pharmacological treatments at baseline. The predicted variables are the estimated change of outcome measures over time. Estimated linear change was calculated for five outcome measures of HD which have known sensitivity to detect clinical change (change from baseline was measured out to 2 years): TFC, SWR, SDMT, TMS, and cUHDRS. The slope was estimated based on a linear mixed model with fixed and random intercept and slope respectively, and the individual-specific slope was computed as the sum of the random and fixed slope. Follow-up assessments up to 2 years that fell within a ± 90-day window around planned annual visits were included.

An RF regression model with 1,000 trees was trained to rank the features on their ability to predict the estimated linear change of each clinical outcome. The model randomly selects a subset of 34 variables (one third of all available) for splitting at each node within each tree. The training was repeated 100 times, each time on a 75% random sample of the data. In each round, permutation importance of each feature for prediction of the outcome was calculated and used for the ranking of the features. The median of the rankings of these 100 models was used for the final ranking of the feature importance.

The R^2 measure (the percentage of the slope variance that is explained by the model) was calculated for the following models predicting each outcome—a model trained with CAP score and CAG only, a model trained with CAP score, CAG and age, a model trained with the above-mentioned shared top 10 ranked features and a model trained with all features.

The analysis was done using R version 3.5.2, with `lmer()` from the `lme4` package for the linear mixed-effects model, and `Cforest()` from the `Party` package for the RF regression model.

DATA AVAILABILITY STATEMENT

The data analysed in this study was obtained from Enroll-HD, https://enroll-hd.org/, the following licenses/restrictions apply: to access data you must be a researcher employed by a recognized academic institution, company or non-profit organisation and apply for an Enroll-HD access account. Requests to access these datasets should be directed to https://enroll-hd.org/for-researchers/become-a-qualified-researcher/.

CODE AVAILABILITY

All the codes and algorithms for the analysis are available upon request.

AUTHOR CONTRIBUTIONS

NG contributed to the conception and design of the study, and acquisition and interpretation of data for the work. SS contributed to conception of the study, and acquisition and interpretation of data for the work. GP and PW contributed to the design of the study. JL contributed to the conception and design of the study. RH contributed to conception and design of the study, and acquisition and analysis of data for the work. All authors drafted or substantively revised a significant portion of the manuscript or figures, and approved the final version for submission.

FUNDING

This study was funded by F. Hoffmann-La Roche Ltd. The authors thank Matt Gooding and Caroline Sproat of MediTech Media, UK for providing medical writing support, which was funded by F. Hoffmann-La Roche Basel Ltd, Switzerland in accordance with Good Publication Practise (GGP3) guidelines (http://www.ismpp.org/gpp3). PW was supported by a UKRI Medical Research Council Skills Development Fellowship (MR/T027770/1).

ACKNOWLEDGMENTS

Enroll-HD is a clinical research platform and longitudinal observational study for Huntington's disease (HD) families intended to accelerate progress towards therapeutics; it is sponsored by CHDI Foundation, a non-profit biomedical research organisation exclusively dedicated to collaboratively developing therapeutics for HD. Enroll-HD would not be possible without the vital contribution of the research participants and their families.

The authors thank all the people who participated in this study.

REFERENCES

1. Bates GP, Dorsey R, Gusella JF, Hayden MR, Kay C, Leavitt BR, et al. Huntington disease. *Nat Rev Dis Primers.* (2015) 1:15005. doi: 10.1038/nrdp.2015.5
2. Roos RA. Huntington's disease: a clinical review. *Orphanet J Rare Dis.* (2010) 5:40. doi: 10.1186/1750-1172-5-40
3. Ross CA, Aylward EH, Wild EJ, Langbehn DR, Long JD, Warner JH, et al. Huntington disease: natural history, biomarkers and prospects for therapeutics. *Nat Rev Neurol.* (2014) 10:204–16. doi: 10.1038/nrneurol.2014.24
4. Reilmann R, Leavitt BR, Ross CA. Diagnostic criteria for Huntington's disease based on natural history. *Mov Disord.* (2014) 29:1335–41. doi: 10.1002/mds.26011
5. Keum JW, Shin A, Gillis T, Mysore JS, Abu Elneel K, Lucente D, et al. The HTT CAG-expansion mutation determines age at death but not disease duration in Huntington disease. *Am J Hum Genet.* (2016) 98:287–98. doi: 10.1016/j.ajhg.2015.12.018
6. 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 Neurol.* (2014) 13:1193–201. doi: 10.1016/S1474-4422(14)70238-8
7. 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 Neurol.* (2013) 12:637–49. doi: 10.1016/S1474-4422(13)70088-7
8. Paulsen JS. Cognitive impairment in Huntington disease: diagnosis and treatment. Curr *Neurol Neurosci Rep.* (2011) 11:474–83. doi: 10.1007/s11910-011-0215-x
9. Paulsen JS, Long JD, Johnson HJ, Aylward EH, Ross CA, Williams JK, et al. Clinical and biomarker changes in premanifest Huntington disease show trial feasibility: a decade of the PREDICT-HD study. *Front Aging Neurosci.* (2014) 6:78. doi: 10.3389/fnagi.2014.00078
10. Rosenblatt A. Neuropsychiatry of Huntington's disease. *Dialogues Clin Neurosci.* (2007) 9:191–7. doi: 10.31887/DCNS.2007.9.2/arosenblatt
11. Wild EJ, Tabrizi SJ. Therapies targeting DNA and RNA in Huntington's disease. *Lancet Neurol.* (2017) 16:837–47. doi: 10.1016/S1474-4422(17)30280-6
12. Frost C, Mulick A, Scahill RI, Owen G, Aylward E, Leavitt BR, et al. Design optimization for clinical trials in early-stage manifest Huntington's disease. *Mov Disord.* (2017) 32:1610–9. doi: 10.1002/mds.27122
13. Langbehn DR, Stout JC, Gregory S, Mills JA, Durr A, Leavitt BR, et al. Association of CAG repeats with long-term progression in huntington disease. *JAMA Neurol.* (2019) 76:1375–85. doi: 10.1001/jamaneurol.2019.2368
14. Epifanio I. Intervention in prediction measure: a new approach to assessing variable importance for random forests. *BMC Bioinformatics.* (2017) 18:230. doi: 10.1186/s12859-017-1650-8
15. Rigatti SJ. Random forest. *J Insur Med.* (2017) 47:31–9. doi: 10.17849/insm-47-01-31-39.1
16. Schobel SA, Palermo G, Auinger P, Long JD, Ma S, Khwaja OS, et al. Motor, cognitive, and functional declines contribute to a single progressive factor in early HD. *Neurology.* (2017) 89:2495–502. doi: 10.1212/WNL.0000000000004743
17. Keogh R, Frost C, Owen G, Daniel RM, Langbehn DR, Leavitt B, et al. Medication use in early-HD participants in track-hd: an investigation of its effects on clinical performance. *PLoS Curr.* (2016) 8. doi: 10.1371/currents.hd.8060298fac1801b01ccea6acc00f97cb
18. Dorsey ER, Brocht AF, Nichols PE, Darwin KC, Anderson KE, Beck CA, et al. Depressed mood and suicidality in individuals exposed to tetrabenazine in a large Huntington disease observational study. *J Huntingtons Dis.* (2013) 2:509–15. doi: 10.3233/JHD-130071
19. Schultz JL, Killoran A, Nopoulos PC, Chabal CC, Moser DJ, Kamholz JA. Evaluating depression and suicidality in tetrabenazine users with Huntington disease. *Neurology.* (2018) 91:e202–7. doi: 10.1212/WNL.0000000000005817
20. Schultz JL, Kamholz JA, Moser DJ, Feely SM, Paulsen JS, Nopoulos PC. Substance abuse may hasten motor onset of Huntington disease: evaluating the Enroll-HD database. *Neurology.* (2017) 88:909–15. doi: 10.1212/WNL.0000000000003661
21. Schultz JL, Harshman LA, Langbehn DR, Nopoulos PC. Hypertension is associated with an earlier age of onset of huntington's disease. *Mov Disord.* (2020) 35:1558–64. doi: 10.1002/mds.28062
22. van der Burg JMM, Gardiner SL, Ludolph AC, Landwehrmeyer GB, Roos RAC, Aziz NA. Body weight is a robust predictor of clinical progression in Huntington disease. *Ann Neurol.* (2017) 82:479–83. doi: 10.1002/ana.25007
23. Philibert R, Dogan M, Noel A, Miller S, Krukow B, Papworth E, et al. Dose response and prediction characteristics of a methylation sensitive digital PCR assay for cigarette consumption in adults. *Front Genet.* (2018) 9:137. doi: 10.3389/fgene.2018.00137
24. Strobl C, Boulesteix AL, Zeileis A, Hothorn T. Bias in random forest variable importance measures: illustrations, sources and a solution. *BMC Bioinformatics.* (2007) 8:25. doi: 10.1186/1471-2105-8-25
25. Mariani MC, Tweneboah OK, Bhuiyan MAM. Supervised machine learning models applied to disease diagnosis and prognosis. *AIMS Public Health.* (2019) 6:405–23. doi: 10.3934/publichealth.2019.4.405
26. Obermeyer Z, Emanuel EJ. Predicting the future—big data, machine learning, and clinical medicine. *N Engl J Med.* (2016) 375:1216–9. doi: 10.1056/NEJMp1606181
27. Byrne LM, Rodrigues FB, Blennow K, Durr A, Leavitt BR, Roos RAC, et al. Neurofilament light protein in blood as a potential biomarker of neurodegeneration in Huntington's disease: a retrospective cohort analysis. *Lancet Neurol.* (2017) 16:601–9. doi: 10.1016/S1474-4422(17)30124-2
28. Rodrigues FB, Byrne LM, Tortelli R, Johnson EB, Wijeratne PA, Arridge M, et al. Mutant huntingtin and neurofilament light have distinct longitudinal dynamics in Huntington's disease. *Sci Transl Med.* (2020) 12:eabc2888. doi: 10.1126/scitranslmed.abc2888
29. Byrne LM, Rodrigues FB, Johnson EB, Wijeratne PA, De Vita E, Alexander DC, et al. Evaluation of mutant huntingtin and neurofilament proteins as potential markers in Huntington's disease. *Sci Transl Med.* (2018) 10:eaat7108. doi: 10.1126/scitranslmed.aat7108
30. Wongvibulsin S, Wu KC, Zeger SL. Clinical risk prediction with random forests for survival, longitudinal, and multivariate (RF-SLAM) data analysis. *BMC Med Res Methodol.* (2019) 20:1. doi: 10.1186/s12874-019-0863-0

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fneur.2021.678484/full#supplementary-material

Conflict of Interest: NG, RH, SS, and GP are employees of F. Hoffmann-La Roche Ltd. JL is a paid Advisory Board member for F. Hoffmann-La Roche Ltd and uniQure biopharma B.V, and a paid consultant for Vaccinex Inc, Wave Life Sciences USA Inc, Genentech Inc and Triplet Inc.

The remaining author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

The authors declare that this study received funding from F. Hoffmann-La Roche Ltd. The funder was involved in the study design, analysis, interpretation of data and the decision to submit it for publication.