Journal of Huntington's Disease xx (20xx) x–xx
DOI 10.3233/JHD-190345
IOS Press

Research Report

Predicting the Risk of Huntington's Disease with Multiple Longitudinal Biomarkers

Fan Li[a,b], Kan Li[c], Cai Li[a] and Sheng Luo[a,b,*] the PREDICT-HD and ENROLL-HD Investigators of the Huntington Study Group
[a]*Department of Biostatistics and Bioinformatics, Duke University School of Medicine, Durham, NC, USA*
[b]*Duke Clinical Research Institute, Durham, NC, USA*
[c]*Merck Research Lab, Merck & Co, North Wales, PA, USA*

Abstract.
Background: Huntington's disease (HD) has gradually become a public health threat, and there is a growing interest in developing prognostic models to predict the time for HD diagnosis.
Objective: This study aims to develop a novel prognostic model that leverages multiple longitudinal biomarkers to inform the risk of HD.
Methods: The multivariate functional principal component analysis was used to summarize the essential information from multiple longitudinal markers and to obtain a set of prognostic scores. The prognostic scores were used as predictors in a Cox model to predict the right-censored time to diagnosis. We used cross-validation to determine the best model in PREDICT-HD ($n = 1{,}039$) and ENROLL-HD ($n = 1{,}776$); external validation was carried out in ENROLL-HD.
Results: We considered six commonly measured longitudinal biomarkers in PREDICT-HD and ENROLL-HD (Total Motor Score, Symbol Digit Modalities Test, Stroop Word Test, Stroop Color Test, Stroop Interference Test, and Total Functional Capacity). The prognostic model utilizing these longitudinal biomarkers significantly improved the predictive performance over the model with baseline biomarker information. A new prognostic index was computed using the proposed model, and can be dynamically updated over time as new biomarker measurements become available.
Conclusion: Longitudinal measurements of commonly measured clinical biomarkers substantially improve the risk prediction of Huntington's disease diagnosis. Calculation of the prognostic index informs the patient's risk category and facilitates patient selection in future clinical trials.

Keywords: Cognitive disorders, cross validation, functional principal component analysis, Huntington's disease, motor diagnosis, risk prediction

INTRODUCTION

Huntington's disease (HD) is a long-term neurodegenerative disorder caused by a cytosine-adenine-guanine (CAG) repeat expansion in the Huntingtin (HTT) gene [1]. HD causes a progressive breakdown of brain cells, leading to uncontrolled movements, loss of cognitive function and mental illness. While the identification of the HTT gene allows clinicians to decide whether a patient will eventually progress to HD, there is a growing interest in developing prognostic models to predict the time for HD diagnosis. These prognostic models accounted for CAG repeat length, age, and their interaction as they were

*Correspondence to: Sheng Luo, Department of Biostatistics and Bioinformatics, Duke University School of Medicine, Durham, NC 27710, USA. Tel.: +1 919 668 8038; Fax: +1 919 668 7059; E-mail: sheng.luo@duke.edu.

ISSN 1879-6397/19/$35.00

found to be strongly predictive of the HD diagnosis [2, 3]. However, prediction based on CAG repeat length and age can be further improved by leveraging longitudinal biomarker information collected in HD studies.

The use of joint models greatly facilitated the development of prognostic models for HD diagnosis [4]. For instance, Paulsen et al. [5] used the Neurobiological Predictors of Huntington's Disease (PREDICT-HD) data to develop a joint model and evaluated the predictive performance of each marker for HD diagnosis. Such models are attractive because they simultaneously modeled the longitudinal trajectories and time-to-event, which improved risk prediction by identifying candidate markers with high prognostic values. Li et al. [6] illustrated how to make dynamic predictions with joint models as the longitudinal assessments accrue and provided a web-based calculator to assist clinical decisions. However, these models considered each longitudinal marker separately and did not incorporated the synergistic effects of multiple longitudinal markers. On the other hand, studies evaluating the combined prognostic values of multiple markers have mostly restricted to the baseline measurements [7–9], with the following two exceptions. Garcia et al. [10] exploited multiple outcomes in addition to motor diagnosis and improved the traditional single-outcome prediction model, while Long and Mills [11] developed a bivariate joint model which used the longitudinal trajectories of two biomarkers to improve risk prediction. To date, no prior studies have investigated the combined values of more than two longitudinal markers for predicting the risk of HD diagnosis.

Based on the state-of-the-art nonparametric functional data analysis (FDA) methodology, in this paper, we developed a new prognostic model for HD diagnosis by integrating the longitudinal information from multiple markers. The prognostic model effectively summarized the longitudinal trajectories of all markers and therefore facilitated clinical decision making based on all historical data. We developed and internally validated the model using the PREDICT-HD data, a publicly available data set well suited for our objective due to its large cohort size, the abundance in marker information, and the prospective study design. The model was then externally validated based on the ENROLL-HD data. With the new prognostic model, we also developed a dynamic prognostic index. Since the prognostic index could be updated over time as new marker measurements accumulate, it has the potential to monitor HD progression for at-risk patients and to assist targeted patient selection in future clinical trials.

MATERIALS AND METHODS

Study population

PREDICT-HD is a prospective cohort study evaluating predictors of time to first HD diagnosis. The study included 1,078 patients with mutation consistent with HD (greater than 35 CAG repeats in the HTT gene), but not yet diagnosed at study entry. Data were collected from 2002 to 2014 across six countries including USA, Canada, Germany, Australia, Spain and UK. Detailed information on study procedures, inclusion and exclusion criteria was published elsewhere [12, 13]. Written informed consent was obtained from all participants at recruitment, and the local institutional review board has approved the study.

ENROLL-HD is a worldwide, prospective observational study monitoring the dynamics of patients at-risk for or with HD. The study included HD families in North American, Europe, Latin America and Australasia. Initiated in 2011, the ENROLL-HD study recruited a total number of 11,906 participants by October 16, 2016. The study is ongoing and detailed information regarding study protocol is available at https://www.enroll-hd.org. We used a subset of ENROLL-HD for external validation. Specifically, we included participants who (i) were at least 18 years old at study entry, (ii) had mutation consistent with HD (defined by having greater than 35 CAG repeats in the HTT gene), (iii) did not have a motor diagnosis at baseline, and (iv) were not from Latin America (since PREDICT-HD did not include participants from Latin American countries). We obtained in a cohort of 2,000 participants who had similar characteristics to PREDICT-HD participants. All participants in ENROLL-HD were given a written informed consent, and the study was approved by the local institutional review board.

Both PREDICT-HD and ENROLL-HD collected a range of clinical data from participants. Here, our primary analysis focuses on examining the six strongest longitudinal predictors identified in Paulsen et al. [5], and available in both data sets: Total Motor Score (TMS), Symbol Digit Modalities Test (SDMT), Stroop Word Test (SWT), Stroop Color Test (SCT),

Stroop Interference Test (SIT) and Total Functional Capacity (TFC). The scoring and rating scale for these clinical markers are based on the Unified Huntington's Disease Rating Scale [14]. Specifically, the Total Motor Score ranges from 0 to 124, with a higher score indicating a greater degree of impaired motor functioning. The Symbol Digit Modalities Test measures working memory, complex scanning, and processing speed [15, 16], while the Stroop Color and Word Test consists of three trials (SWT, SCT and SIT), each of which generates a score and measures basic attention and processing speed in color identification and word reading [17]. Among them, SIT focuses on the ability to ignore an interference signal between words and colors. All four cognitive test scores were calculated using the number of correctly answered items in the administered task, and higher scores correspond to better cognitive functioning. The Total Functional Capacity is a 5-item clinician rating scale measuring the capacity of function in HD. The five items were summed to obtain a total score, ranging from 0 to 13 with higher scores corresponding to a more independent level of functioning. Since ENROLL-HD includes a few patients with more extreme baseline values of the six markers than PREDICT-HD, we excluded these ENROLL-HD patients whose baseline marker values are outside the range of the PREDICT-HD patients to ensure better comparability. In the secondary analysis, we also considered two imaging markers (Putman and Hippocampus volume) and a psychiatric marker (FrSBe executive subscale), which have been identified amongst the biomarkers of highest prognostic value [5]. However, these markers were only available in PREDICT-HD. Age at baseline, gender, CAG repeat length, and CAG-Age Product (CAP; [baseline age]$\times$[CAG-33.66] [3] were considered as other prognostic variables given their established association with motor diagnosis. Following the standard definition of the Unified Huntington Disease Rating Scale [14], we characterized the motor diagnosis of Huntington's disease as a rating of 4 on the diagnostic confidence level (DCL) during the follow-up period. Sample size and descriptive statistics of baseline variables for PREDICT-HD and ENROLL-HD data sets are shown in Table 1. Table 1 suggests that participants in two studies were mostly similar at baseline. However, ENROLL-HD patients had on average slightly higher TMS and lower cognitive test scores compared to PREDICT-HD, and may indicate higher baseline risk for progression.

Statistical methods

The multivariate functional principal component analysis (MFPCA) and Cox regression were used to jointly analyze the multiple longitudinal marker trajectories and survival outcomes. The MFPCA approach [18] assumes that the underlying trajectories from all markers follow a latent multidimensional stochastic process that accounts for the correlation among marker trajectories and summarizes the essential information from all trajectories into a set of multivariate functional principal component (MFPC) scores. The MFPC scores have a much lower dimension compared to the entire longitudinal history, and are assumed sufficient to characterize the overall trend and systematic patterns of the trajectories from multiple markers. The MFPCA approach makes minimal parametric assumptions; it accommodates missed visits by modeling the latent trajectory of each marker as a smooth process, and further accounts for measurement error by including a residual noise term [19]. We retained the set of MFPC scores derived from multiple longitudinal markers as a set of new prognostic variables, and used them to predict the risk of HD diagnosis in a survival model. The technical details on MFPCA and Cox models were available in the supplementary material.

We first investigated whether the use of longitudinal information from multiple markers increased the predictive accuracy compared to the use of baseline marker values. We primarily focused on the six predictors that are commonly measured in HD studies and available in both PREDICT-HD and ENROLL-HD, i.e., TMS, SMDT, SWT, SCT, SIT and TFC. Two pre-planned Cox models were considered: model 1 adjusted for only the measurements of the six markers at baseline, while model 2 adjusted for MFPC scores obtained from the six longitudinal markers. We kept the first 10 MFPC scores which accounted for at least 99% of the total variation in the six markers. We further adjusted for age, gender, length of CAG repeat expansion, and the CAG-Age Product in both models. The survival time in years was defined from study entry to the first HD diagnosis or censored at the last visit for subjects without diagnosis. In model 2, the set of MFPC scores were calculated for each subject using information obtained prior to the time of HD diagnosis or censoring. The two survival models were specified as

$$h_i(t) = h_0(t)\exp\{\gamma_1 baseline\ Age_i + \gamma_2 gender_i + \gamma_3 CAG_i + \gamma_4 CAP_i + \gamma_5 TMS_i$$

Table1
Summary statistics for demographic and clinical variables measured at study entry for PREDICT-HD and ENROLL-HD participants with baseline CAG mutation consistent with HD. *Table entries are mean (SD) or n (%). HD, Huntington's disease; CAG, length of Cytosine-adenine-guanine (CAG) repeat expansion in the Huntingtin (HTT) gene; CAP, CAG-Age Product; TMS, Total Motor Score; SDMT, Symbol Digit Modalities Test; SWT, Stroop Word Test; SCT, Stroop Color Test; SIT, Stroop Interference Test; TFC, Total Functional Capacity

	PREDICT-HD			ENROLL-HD		
	Progressed to HD during the study	Did not progressed to HD during the study	Combined	Progressed to to HD during the study	Did not progressed to HD during the study	Combined
Sample Size	223	816	1039	148	1628	1776
Women	145 (65%)*	518 (63.5%)	663 (63.8%)	77 (52%)	992 (60.9%)	1069 (60.2%)
Age (years)	42.95 (10.32)	38.93 (10.18)	39.80 (10.34)	47.24 (11.20)	40.23 (11.94)	40.82 (12.04)
CAG	43.57 (2.86)	42.21 (2.60)	42.50 (2.72)	43.20 (2.91)	42.38 (2.79)	42.45 (2.81)
CAP	405.01 (74.59)	318.84 (78.91)	337.34 (85.62)	425.47 (78.46)	332.50 (89.24)	340.24 (92.03)
Time in study (years)	4.25 (2.51)	4.38 (3.31)	4.35 (3.16)	1.30 (0.52)	0.82 (0.90)	0.86 (0.88)
TMS	8.61 (6.47)	3.80 (4.32)	4.83 (5.24)	9.68 (6.97)	3.23 (4.18)	3.76 (4.82)
SDMT	44.48 (10.71)	52.18 (11.10)	50.52 (11.45)	40.24 (11.55)	50.14 (11.81)	49.32 (12.10)
SWT	91.14 (16.71)	100.97 (17.24)	98.86 (17.59)	82.05 (19.32)	93.05 (18.13)	92.13 (18.48)
SCT	70.79 (13.48)	79.21 (13.74)	77.40 (14.11)	64.86 (15.19)	73.27 (14.48)	72.56 (14.72)
SIT	39.44 (9.24)	46.29 (10.31)	44.82 (10.47)	37.41 (10.47)	43.63 (10.95)	43.11 (11.04)
TFC	12.68 (0.83)	12.84 (0.68)	12.81 (0.72)	12.05 (1.46)	12.67 (0.95)	12.62 (1.01)

$$+\gamma_6 SMDT_i + \gamma_7 SWT_i + \gamma_8 SCT_i + \gamma_9 SIT_i + \gamma_{10} TFC_i\}, \quad (1)$$

$$h_i(t) = h_0(t)\exp\{\gamma_1 baseline\,Age_i + \gamma_2 gender_i + \gamma_3 CAG_i + \gamma_4 CAP_i + \sum_{p=1}^{10}\beta_p MFPCscore_{ip}\}, \quad (2)$$

where $h_0(t)$ is the baseline hazard, γ_i and β_p are regression coefficients.

The model performance was assessed using the integrated area under the time-dependent receiver operating characteristics curve (iAUC) [20] and the integrated Brier score (iBS) [21]. Higher iAUC indicates better predictive accuracy and lower iBS indicates better agreement between predicted and observed survival outcomes. Internal validation was performed based on the PREDICT-HD and ENROLL-HD data using repeated 10-fold cross-validation (CV) and we referred to this process as internal CV. The 10-fold CV was used to estimate iAUC and iBS, and we repeated the process for 100 times to account for the variability in randomly splitting the data. For external validation, we estimated model parameters from PREDICT-HD and assessed iAUC and iBS using ENROLL-HD data. The longitudinal marker trajectories in PREDICT-HD and ENROLL-HD were defined on the same time scale to ensure comparability and the latest follow-up time in PREDICT-HD was set as the upper bound. To further examine whether model 2 improved risk prediction over model 1, we tested for differences in time-dependent area under the receiver operating characteristics curves (AUCs) across models, and reported p-values at pre-specified time points [22]. The pre-specified time points were each whole year for PREDICT-HD, and each half year for ENROLL-HD; a finer scale was used for the latter due to its shorter follow-up. The reported p-values were adjusted to prevent an inflation of type I error due to repeated testing at multiple time points [23]. Details of this test and the adjustment for multiple comparison can be found in Blanche et al. [22]. Model 2 motivated a new prognostic index, PI_{HD}, which was defined as a linear combination of the prognostic factors:.

$$PI_{HD} = \hat{\gamma}_1 baselineAge_i + \hat{\gamma}_2 gender_i + \hat{\gamma}_3 CAG_i + \hat{\gamma}_4 CAP_i + \sum_{p=1}^{10}\hat{\beta}_p MFPCscore_{ip} \quad (3)$$

We estimated all regression coefficients from PREDICT-HD and computed the index for each patient in ENROLL-HD. Risk groups for HD progression were defined based on the quartiles of PI_{HD} obtained from PREDICT-HD. All analyses were performed by restricting subjects to those with complete baseline values of all six markers ($n = 1{,}039$ in PREDICT-HD and $n = 1{,}776$ in ENROLL-HD).

Since we have restricted the construction of the prognostic index to the six major markers considered, we further assessed the predictive utility of

additionally incorporating the longitudinal imaging and psychiatric measurements. For this objective, we sequentially included each of the following longitudinal markers—Putman volume and Hippocampus volume (imaging markers) and FrSBe executive subscale (psychiatric marker)—for calculating the MFPC scores and therefore estimating the iAUC and iBS corresponding to the new prognostic models. As these markers were only available for a subset of subjects in PREDICT-HD, we restricted the evaluation to a subset in PREDICT-HD with at least complete baseline information for all nine markers. All statistical analyses were performed using the *R* statistical software (version 3.4.0). The MFPCA step was performed using the *MFPCA* package [18, 24]; Cox models were fit using the *survival* package [25], and model performance measures were calculated using the *survAUC* and *pec* packages [26, 27].

RESULTS

Table 2 presents the iAUC and iBS results for model 1 and model 2 calculated using both data sets. From internal CV based on PREDICT-HD, the multivariable baseline marker model (model 1) indicates the iAUC = 0.856, while the prognostic model with multiple longitudinal markers (model 2) increased iAUC by about 6% to 0.911. Similar findings were confirmed from internal CV based on ENROLL-HD. From external validation, the performance statistics (iAUC) was slightly lower than the internal validation for both models, but still favored model 2 over model 1. Results for model comparisons were similar when using iBS as the performance statistic. Supplementary Tables 1 and 2 present the *p*-values for testing the null hypothesis that the time-dependent AUCs of model 1 and 2 are equal. After accounting for multiple comparisons [22], the adjusted *p*-values were significant at the 0.05 level for most of the time points considered, in both data sets. Such comparisons confirmed that the six longitudinal markers had substantially higher predictive values than their baseline assessments.

We obtained the prognostic index PI_{HD} for cohorts in PREDICT-HD and ENROLL-HD based on the model 2 estimated using PREDICT-HD. The quartiles of the estimated PI_{HD} were used to categorize each cohort into four risk groups. Figure 1 presents the Kaplan-Meier survival curves for these risk groups (survival is defined as the probability of not being diagnosed). The risk gradient was similar for PREDICT-HD and ENROLL-HD as the curves were in the same order from top to bottom even though PI_{HD} was estimated from PREDICT-HD. By design, PI_{HD} provided an equal classification of the four risk groups in PREDICT-HD, and it further resulted in approximately equal classification of the four risk groups in ENROLL-HD. However, the survival curves showed some differences between the two data sets. At year 3, for example, the high-risk group in ENROLL-HD was at a higher risk for HD diagnosis (survival probability = 0.212) than that in PREDICT-HD (survival probability = 0.628), suggesting that the ENROLL-HD included participants whose conditions deteriorated at a faster rate. For comparison purposes, we also created four risk groups based on PI_{HD} estimated from model 1, and present the Kaplan-Meier survival curves in Figure 2. Table 3 reports the survival probability estimates of various groups at year 3. Overall, the risk gradient was similar between model 1 and model 2. However, by accounting for the longitudinal patterns of clinical marker information, model 2 tends to better separate the healthier and sicker patients into the low and high-risk groups. For instance, in ENROLL-HD, the survival probabilities at year 3 are 0.964 and 0.995 among the low risk groups created based on model 1 and model 2, which suggests that model 1 includes a few less healthier patients in its low risk group than model 2, by basing the classification only on the baseline information. As model 2 leads to better predictive performance and provides more separated risk strata, we selected model 2 as a preferable model in the subsequent analysis.

The Cox regression coefficient estimates from model 2 were presented in Supplementary Table 3. The magnitudes of these coefficients determined the contribution from each prognostic factor towards the prognostic index. Larger value of the index suggested greater risk for HD progression, and the estimated quartiles of PI_{HD} (25%: –6.16; 50%: –5.46; 75%: –4.53) classify patients into risk strata and facilitate the interpretation of the survival curves. An attractive feature of the prognostic index is that it can be dynamically updated over time as new marker measurements accumulate. In Supplementary Tables 4 and 5, we illustrated how to update PI_{HD} for a selected patient from PREDICT-HD as her new marker measurements accrue. At baseline, the prognostic index was calculated as –5.81, indicating the patient has mid-low-risk for HD progression. When additional longitudinal information up to year 2, the prognostic index increased to –5.33, which classified

Table 2
Integrated AUC and Brier score for two prognostic models based on internal and external validations. Model 1: multivariable baseline marker including TMS, SMDT, SWT, SCT, SIT and TFC; Model 2: multiple longitudinal neurocognitive marker defined by 10 MFPC scores. All models control for baseline age, CAG, Age-CAG product and gender

	Model 1		Model 2	
Study	$iAUC_{INT}$	$iAUC_{Ext}$	$iAUC_{INT}$	$iAUC_{Ext}$
PREDICT-HD ($n = 1039$)	0.856		0.911	
ENROLL-HD ($n = 1776$)	0.864	0.803	0.888	0.856
	$Brier_{INT}$	$Brier_{EXT}$	$Brier_{INT}$	$Brier_{EXT}$
PREDICT-HD ($n = 1039$)	0.100		0.083	
ENROLL-HD ($n = 1776$)	0.073	0.112	0.066	0.111

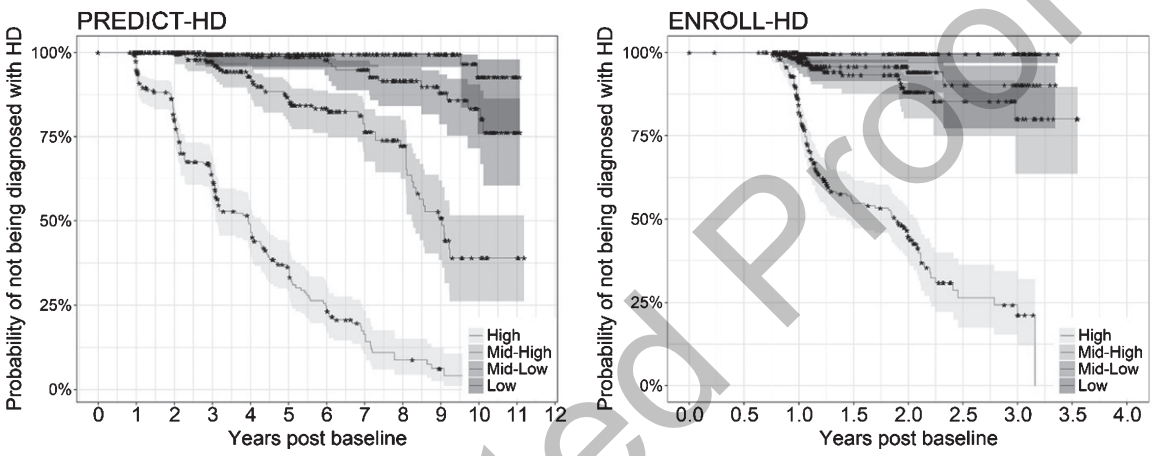

Fig. 1. Kaplan-Meier curves for four risk groups of motor diagnosis using the MFPCA model. PI_{HD} was computed based on the estimated parameters from model 2 using the PREDICT-HD data set, with 10 MFPC scores derived from longitudinal information of TMS, SMDT, SWT, SCT, SIT and TFC. Shaded regions represent the 95% confidence interval of the estimated survival probabilities. In PREDICT-HD, the percentages of each group is 25% (High Risk group), 25% (Mid-High Risk group), 25% (Mid-Low Risk group), 25% (Low Risk group). In ENROLL-HD, the percentages of each group is 24% (High Risk group), 28% (Mid-High Risk group), 26% (Mid-Low Risk group), 22% (Low Risk group).

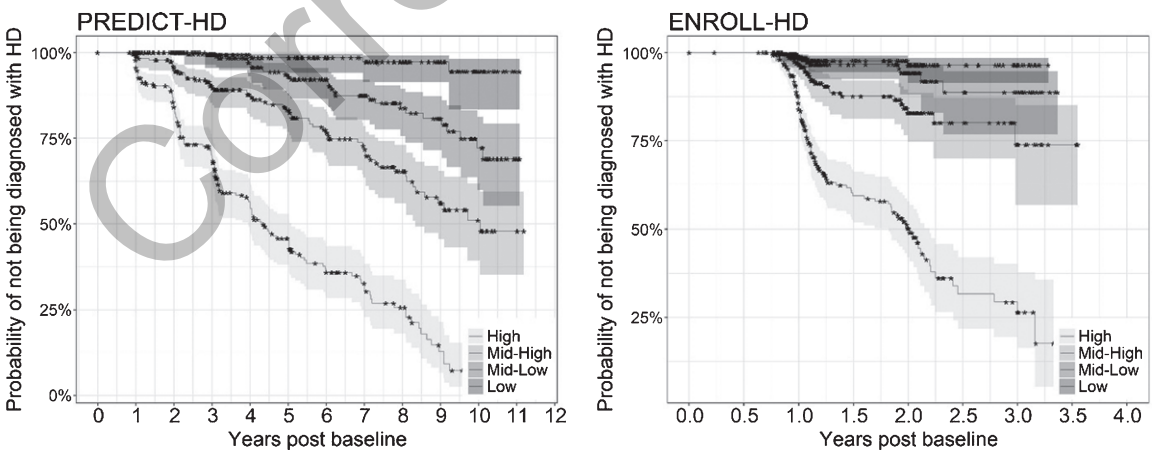

Fig. 2. Kaplan-Meier curves for four risk groups of motor diagnosis using the baseline model. PI_{HD} was computed based on the estimated parameters from model 1 using the PREDICT-HD data set, with baseline information of TMS, SMDT, SWT, SCT, SIT and TFC. In this case, $PI_{HD} = \hat{\gamma}_1 baselineAge_i + \hat{\gamma}_2 gender_i + \hat{\gamma}_3 CAG_i + \hat{\gamma}_4 CAP_i + \hat{\gamma}_5 TMS_i + \hat{\gamma}_6 SMDT_i + \hat{\gamma}_7 SWT_i + \hat{\gamma}_8 SCT_i + \hat{\gamma}_9 SIT_i + \hat{\gamma}_{10} TFC_i$ Shaded regions represent the 95% confidence interval of the estimated survival probabilities. In PREDICT-HD, the percentages of each group is 25% (High Risk group), 25% (Mid-High Risk group), 25% (Mid-Low Risk group), 25% (Low Risk group). In ENROLL-HD, the percentages of each group is 26% (High Risk group), 24% (Mid-High Risk group), 23% (Mid-Low Risk group), 27% (Low Risk group).

Table 3
Kaplan-Meier estimates and 95% confidence limits of survival probability at year 3 for four risk groups based on model 1 and model 2. The log-log transformation was used to obtain the 95% confidence limits

		Low Risk group	Mid-Low Risk group	Mid-High Risk group	High Risk group
Model 1	PREDICT-HD	0.993 (0.954, 0.999)	0.995 (0.965, 0.999)	0.903 (0.853, 0.937)	0.695 (0.627, 0.753)
	ENROLL-HD	0.964 (0.914, 0.985)	0.888 (0.769, 0.948)	0.740 (0.569, 0.851)	0.265 (0.163, 0.377)
		Low Risk group	Mid-Low Risk group	Mid-High Risk group	High Risk group
Model 2	PREDICT-HD	0.994 (0.958, 0.999)	0.995 (0.962, 0.999)	0.974 (0.938, 0.989)	0.628 (0.558, 0.689)
	ENROLL-HD	0.995 (0.968, 0.999)	0.902 (0.772, 0.960)	0.800 (0.636, 0.896)	0.212 (0.121, 0.320)

Table 4
Model comparisons with imaging and psychiatric biomarkers for PREDICT-HD ($n = 716$). Model testing was limited to subsets of participants who had complete observations for all baseline markers of interest (e.g., those in model 1d). *Base model with baseline variables from TMS, SMDT, SWT, SCT, SIT and TFC. From model 1b to model 1d, baseline imaging variables Putamen, Hippocampus and psychiatric variable FrSBe exec were successively added to the base model, i.e., model 1b included Putamen, model 1c included Putamen and Hippocampus, model 1d included Putamen, Hippocampus and FrSBe exec. †Base model with 10 MFPC scores derived from the longitudinal trajectory of TMS, SMDT, SWT, SCT, SIT and TFC. From model 2b to model 2d, longitudinal information of imaging variables Putamen, Hippocampus and psychiatric variable FrSBe exec were successively added to the MFPC score calculation, i.e., model 2b included Putamen, model 2c included Putamen and Hippocampus, model 2d included Putamen, Hippocampus and FrSBe exec

	Model 1a*	Model 1b	Model 1c	Model 1d
$iAUC_{INT}$	0.839	0.853	0.852	0.855
$Brier_{INT}$	0.101	0.097	0.097	0.097
	Model 2a†	Model 2b	Model 2c	Model 2d
$iAUC_{INT}$	0.915	0.915	0.914	0.909
$Brier_{INT}$	0.078	0.078	0.078	0.078

the patient to the mid-high-risk group for HD diagnosis. Medical intervention may be necessary at this stage as PI_{HD} showed that the patient was at an elevated risk to develop HD. Given any future time point, we could also approximate the baseline hazard $h_0(t)$ in model 2 using flexible spline functions to quantify the actual probability of HD diagnosis for a specific patient.

Table 4 presents the iAUC and iBS for models including imaging and psychiatric markers beyond the six clinical markers. Like model 2, model 2a estimated MFPC scores derived from the above six markers. From model 2b to model 2d, longitudinal information of imaging measurements and psychiatric measurement were successively added to the MFPC score calculation. Overall, models 2b through 2d had similar predictive performance compared to model 2a, and did not increase the value of iAUC or decrease the value of iBS. This implies that when the six clinical markers were already present in the model, including additional longitudinal imaging markers may not lead to improved predictive accuracy. Moreover, the inclusion of baseline imaging markers also did not substantially improve the predictive accuracy of the baseline model with six clinical markers (model 1a). However, adding the longitudinal markers substantially improves the model with only baseline information, regardless of the inclusion of imaging markers. Statistical tests comparing time-dependent AUCs at pre-specified time points revealed consistent results and were omitted for brevity. The results indicate that the relevant prognostic information from the imaging markers is likely reflected in the scoring patterns of the clinical variables already in the model. Overall, Table 4 confirmed that model 2a (model 2) was preferable for clinical decision making because the six clinical markers could be collected at minimal cost, while the imaging markers require specific equipment and are more expensive to collect.

DISCUSSION

In this study, we proposed a novel prognostic model to jointly analyze the trajectories of multiple longitudinal HD markers and time to HD diagnosis. The multivariate functional principal component analysis (MFPCA) was applied on multiple longitudinal marker data to derive a set of scores for each individual, and these derived scores were used as a

new set of prognostic factors in the Cox regression model. We demonstrated that including the longitudinal information from multiple markers, beyond the baseline measurements, significantly improved the predictive accuracy for HD diagnosis, as indicated both by internal CV and the statistical tests for comparing time-dependent AUCs. The model performance was also assessed by external validation, which has been recommended for developing prediction models in clinical studies [28, 29]. The improved predictive performance from external validation with ENROLL-HD data supported the general usefulness of the proposed model. Of note, our prognostic model shares the same spirit with the bivariate joint model developed in Long and Mills [11] in that both models exploited the synergistic effect of multiple marker trajectories. Although Long and Mills only examined TMS and SDMT in their study, an extension of their model could possibly accommodate the six longitudinal markers in the current study. Nevertheless, such a multivariate joint model relies on parametric assumptions and its computational burden may be much heavier than the proposed MFPCA model. By contrast, the MFPCA approach is a flexible nonparametric approach that does not impose any parametric forms of the longitudinal trajectories and is computationally efficient [18]. Although the derived MFPC scores may lack clinical interpretation as they are implicit functions of the multivariate marker trajectories, we performed additional analysis of the PREDICT-HD sample to understand the information from each marker that drives the most important MFPC scores and presented the results in Supplementary Figure 4 to Supplementary Figure 9. While the majority of the variability in the four cognitive tests (SMDT, SWT, SCT, SIT) lies in the direction close to the overall mean, the clinical score contrast between early, medium and later times also contribute substantial variability in TMS and TFC trajectories. This suggests that we could interpret the effect from the set of MFPC scores as the additional effect due to dynamic changes from the TMS and TFC values. The successful capture of this effect favored the application of MFPCA as the specification of such dynamic effect is not as straightforward in parametric models. Additionally, because the MFPC scores greatly reduce the dimension of longitudinal information and enhance the performance of the survival model for risk prediction, we used these scores in constructing the prognostic index.

There are several advantages of the proposed prognostic index. First, it is easy to calculate with commonly measured prognostic factors. In fact, our prognostic model included only the motor, cognitive and functional assessments which are routinely collected and it is widely applicable in the clinical setting. Imaging markers were more expensive to obtain as they may require specific equipment, and was not universally available in all HD studies (e.g., ENROLL-HD). We have shown that the addition of the imaging markers offered negligible improvement in predictive accuracy when the six clinical markers were already included in the model. For this reason, we excluded the imaging information in deriving the prognostic index. Second, the proposed prognostic index can be dynamically updated over time, as we illustrated in the Supplementary Material. Further, the value of the prognostic index can be used to predict when a patient will progress to HD. Third, we could use the prognostic index to identify target patients for clinical trials from participants in existing cohort studies. It is likely that a history of longitudinal marker observations has been recorded for these participants, which could inform the calculation of prognostic index. For example, if a clinical trial plans to recruit patients who are less susceptible to develop HD, the patients in the low-risk group with the smallest PI_{HD} should be considered. On the other hand, if a trial involves a medical intervention that targets patients who are more susceptible to develop HD, the investigators might recruit patients in the mid-high or high-risk categories. In the modern era of big data and as the routine collection of patient information becomes the norm (e.g., electronic medical records), our prognostic models have the potential to efficiently summarize essential information from historical marker trajectories and inform medical decisions on HD-specific interventions.

There are several limitations in our study. First, we have demonstrated the improved risk prediction of the MFPCA-Cox model relative to a simple baseline model adjusting for only linear effects from the six markers. To place our results in the context of current literature, we further examined the predictive utility of three existing HD prognostic models, developed in Long et al. [7, 9], and report the performance statistics in Supplementary Table 6. Based on the iAUC performance metric, both the internal and external cross validation favored the MFPCA-Cox model, consistent with our primary analysis. However, based on the iBS performance metric, the external validation seemed to favor two baseline models with nonlinear TMS effects and a TMS-CAG interaction. A possible explanation is that since ENROLL-HD had relatively

fewer longitudinal information, the alternative baseline specifications may avoid the risk of overfitting and lead to less variable predictions (henceforth the slightly smaller iBS). Further simulation studies are necessary to provide a more comprehensive assessment on the pros and cons of each model in the presence of limited longitudinal follow-up, such as the current ENROLL-HD sample. Second, we did not consider the combinations of all markers in constructing the prognostic model, and prioritized the ones that demonstrated greater importance in previous studies. A combination of clinical, imaging and genetic markers would potentially provide a more comprehensive prognostic index. However, our restriction to the non-imaging markers bears a pragmatic consideration as the imaging markers are not broadly available in all studies such as ENROLL-HD. Third, participants of the PREDICT-HD and ENROLL-HD may not represent the more general population at risk for HD and selection bias may exist. We considered ENROLL-HD as a data set for external validation because (i) the patients were largely similar to those in PREDICT-HD at baseline, (ii) ENROLL-HD collected a breath of clinical variables over time, and (iii) ENROLL-HD is an ongoing study with wide scientific interests. The limitation, however, is that ENROLL-HD has a much shorter follow-up and fewer events, which may preclude a more in-depth investigation. Future studies may consider other HD cohorts that collected multiple longitudinal marker data to validate the proposed prognostic model, similar to the validation study presented in Long and Mills [11]. Finally, although the proposed model easily accommodates information from multiple longitudinal markers and is computationally more efficient than the multivariate joint models [6], it may be subject to bias from informative censoring. For instance, there could be patients with worse prognosis who are more likely to withdraw from the study, and consequently less longitudinal marker information is observed due to informative censoring. However, our initial simulations comparing the MFPCA-Cox model with the joint modeling approach found limited difference on predictive performance in the presence of informative censoring [30]. Regardless, such methodological issues indicate the necessity of additional research to understand the relative merits of MFPCA-Cox model and the multivariate joint model.

In summary, our study developed a prognostic model for HD that effectively leveraged multiple longitudinal marker information to predict time to HD diagnosis. We demonstrated that including commonly collected longitudinal markers from the motor, cognitive, and functional domains substantially improved risk prediction, based on the PREDICT-HD and ENROLL-HD data sets. We further illustrated how the proposed prognostic index with multiple longitudinal marker information might guide targeted patient recruitment in future clinical trials.

ACKNOWLEDGMENTS

Sheng Luo's research was supported partly by the NINDS R01NS091307 and NIA R56AG062302.

CONFLICT OF INTEREST

The authors stated no conflict of interest.

SUPPLEMENTARY MATERIAL

The supplementary material is available in the electronic version of this article: http://dx.doi.org/10.3233/JHD-190345.

REFERENCES

[1] MacDonald ME, Ambrose CM, Duyao MP, Myers RH, Lin C, Srinidhi L, et al. A novel gene containing a trinucleotide repeat that is expanded and unstable on Huntington's disease chromosomes. Cell. 1993;72(6):971-83.

[2] Langbehn DR, Hayden MR, Paulsen JS, the PREDICT-HD Investigators of the Huntington Study Group. CAG-repeat length and the age of onset in Huntington disease (HD): A review and validation study of statistical approaches. Am J Med Genet B Neuropsychiatr Genet. 2010;153(2):397-408.

[3] Zhang Y, Long JD, Mills JA, Warner JH, Lu W, Paulsen JS, et al. Indexing disease progression at study entry with individuals at-risk for Huntington disease. Am J Med Genet B Neuropsychiatr Genet. 2011;156(7):751-63.

[4] I brahim JG, Chu H, Chen LM. Basic concepts and methods for joint models of longitudinal and survival data. J Clin Oncol. 2010;28(16):2796-801.

[5] 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(12):1193-201.

[6] Li K, Furr-Stimming E, Paulsen JS, Luo S. Dynamic prediction of motor diagnosis in Huntington's disease using a joint modeling approach. J Huntingtons Dis. 2017;6(2):127-37.

[7] Long JD, Paulsen JS, the PREDICT-HD Investigators and Coordinators of the Huntington Study Group. Multivariate prediction of motor diagnosis in Huntington's disease: 12 years of PREDICT-HD. Mov Disord. 2015;30(12):1664-72.

[8] Long JD, Mills JA, Leavitt BR, Durr A, Roos RA, Stout JC, et al. Survival end points for Huntington disease trials prior to a motor diagnosis. JAMA Neurol. 2017;74(11):1352-60.

[9] Long JD, Langbehn DR, Tabrizi SJ, Landwehrmeyer BG, Paulsen JS, Warner J, et al. Validation of a prognostic index

for Huntington's disease. Mov Disord. 2017;32(2):256-63.
[10] Garcia TP, Wang Y, Shoulson I, Paulsen JS, Marder K. Disease progression in Huntington disease: An analysis of multiple longitudinal outcomes. J Huntingtons Dis. 2018;7(4):337-44.
[11] Long JD, Mills JA. Joint modeling of multivariate longitudinal data and survival data in several observational studies of Huntington's disease. BMC Med Res Methodol. 2018;18(138):1-15.
[12] Paulsen JS, Hayden M, Stout JC, Langbehn DR, Aylward E, Ross CA, et al. Preparing for preventive clinical trials: The Predict-HD study. Arch Neurol. 2006;63(6):883-90.
[13] Paulsen JS, Langbehn DR, Stout JC, Aylward E, Ross CA, Nance M, et al. Detection of Huntington's disease decades before diagnosis: The Predict-HD study. J Neurol Neurosurg Psychiatry. 2008;79(8):874-80.
[14] Huntington Study Group. Unified Huntington's disease rating scale: Reliability and consistency. Mov Disord. 1996;11:136-42.
[15] Wechsler D. Manual for the Wechsler Adult Intelligence Scale-Revised (WAIS-R). San Antonio, TX: The Psychological Corporation; 1981.
[16] Lezak MD, Howieson DB, Loring DW, Fischer JS. Neuropsychological Assessment. Oxford University Press, USA; 2004.
[17] Stroop JR. Studies of interference in serial verbal reactions. J Exp Psychol. 1935;18(6):643-62.
[18] Happ C, Greven S. Multivariate functional principal component analysis for data observed on different (dimensional) domains. J Am Stat Assoc. 2018;113(552):649-59.
[19] Yao F, Müller H-G, Wang J-L. Functional data analysis for sparse longitudinal data. J Am Stat Assoc. 2005;100(470):577-90.
[20] Uno H, Cai T, Tian L, Wei LJ. Evaluating prediction rules for t-year survivors with censored regression models. J Am Stat Assoc. 2007;102(478):527-37.
[21] Gerds TA, Schumacher M. Consistent estimation of the expected Brier score in general survival models with right-censored event times. Biom J. 2006;48(6):1029-40.
[22] Blanche P, Dartigues J-F, Jacqmin-Gadda H. Estimating and comparing time-dependent areas under receiver operating characteristic curves for censored event times with competing risks. Stat Med. 2013;32(30):5381-97.
[23] Bretz F, Westfall P, Hothorn T. Multiple Comparisons Using R. Chapman and Hall/CRC; 2016.
[24] Happ C. Multivariate functional principal component analysis for data observed on different dimensional domains. R package version 13-2 [Internet]. Available from: https://github.com/ClaraHapp/MFPCA.
[25] Therneau T. A package for survival analysis in S. R Package version 238.
[26] Potapov S, Adler W, Schmid M. survAUC: Estimators of prediction accuracy for time-to-event data. R package version 10-5.
[27] Mogensen UB, Ishwaran H, Gerds TA. Evaluating random forests for survival analysis using prediction error curves. J Stat Softw. 2012;50(11):1-23.
[28] Royston P, Altman DG. External validation of a Cox prognostic model: Principles and methods. BMC Med Res Methodol. 2013;13(33):1-15.
[29] Debray TP, Vergouwe Y, Koffijberg H, Nieboer D, Steyerberg EW, Moons KG. A new framework to enhance the interpretation of external validation studies of clinical prediction models. J Clin Epidemiol. 2015;68(3):279-89.
[30] Li K, O'Brien R, Lutz M, Luo S, the Alzheimer's Disease Neuroimaging Initiative. A prognostic model of Alzheimer's disease relying on multiple longitudinal measures and time-to-event data. Alzheimers Dement. 2018;14(5):644-51.
[31] Xiao L, Li C, Checkley W, Crainiceanu C. Fast covariance estimation for sparse functional data. Statistics and Computing. 2018;28(3):511–522.