Citation: CPT Pharmacometrics Syst. Pharmacol. (2016) 5, 437–445; doi:10.1002/psp4.12097

ORIGINAL ARTICLE

Modeling Variability in the Progression of Huntington's Disease A Novel Modeling Approach Applied to Structural Imaging Markers from TRACK-HD

JH Warner*and C Sampaio

We present a novel, general class of disease progression models for Huntington's disease (HD), a neurodegenerative disease caused by a cytosine-adenine-guanine (CAG) triplet repeat expansion on the huntingtin gene. Models are fit to a selection of structural imaging markers from the TRACK 36-month database. The models are of mixed effects type and should be useful in predicting any continuous marker of HD state as a function of age and CAG length (the genetic factor that drives HD pathology). The effects of age and CAG length are modeled using flexible regression splines. Variability not accounted for by age, CAG length, or covariates is modeled using terms that represent measurement error, population variability (random slopes/intercepts), and variability due to the dynamics of the disease process (random walk terms). A Kalman filter is used to estimate variances of the random walk terms.
CPT Pharmacometrics Syst. Pharmacol. (2016) **5**, 437–445; doi:10.1002/psp4.12097; published online 2 August 2016.

Study Highlights

WHAT IS THE CURRENT KNOWLEDGE ON THE TOPIC?

☑ Pathology in HD develops very slowly over many years. Considerable effort has been expended in collecting prospective data on subjects in various stages of HD. However, few (if any) subjects have been observed over the complete time course of the disease.

WHAT QUESTION DID THIS STUDY ADDRESS?

☑ There is a need to simulate a range of complete time courses by "patching together" shorter time courses obtained from many different individuals. The current study proposes a solution to this problem.

WHAT THIS STUDY ADDS TO OUR KNOWLEDGE

☑ This study demonstrates the feasibility of the proposed solution and takes a first step toward developing a more global view of disease progression in HD.

HOW THIS MIGHT CHANGE CLINICAL PHARMACOLOGY AND THERAPEUTICS

☑ The models presented here could be used to inform clinical trial simulations and to help in the search for naturally occurring factors that affect disease progression.

Huntington's disease (HD) is a neurodegenerative disease with autosomal dominant inheritance. HD is caused by a cytosine-adenine-guanine (CAG) triplet repeat expansion on the huntingtin gene that enters the pathological range when it reaches 36 repeats with longer CAG lengths associated with earlier onset.[1] Two recent studies have also found that longer CAG length is associated with faster progression of the disease, an effect which may decrease as the disease progresses.[2,3] Ref. 4 found that the interval between motor onset and death is independent of CAG length but does not dispute the claim that CAG-related progression of signs and symptoms occurs after disease onset. Ref. 5 documents CAG length and age-dependent effects in mouse models for HD.

In the present study, we present a general class of disease progression models for HD and fit them to a high quality dataset consisting of magnetic resonance imaging-based structural imaging markers from the TRACK 36-month database.[2] The models are of mixed effects type and should be useful in predicting any continuous marker of HD state as a function of age and CAG length controlling for the effects of normal aging. The models use natural regression splines to model the effects attributable to aging in healthy individuals and the effects of CAP score (short for CAG age product) in subjects who are gene-positive for HD.

The CAP score has been used extensively in the HD literature to model the effects of age and CAG length on various measures of HD state.[1,2,6–9] In what follows, we make use of the general form of the CAP score, as defined in Ref 1.

$$CAP = AGE \times (CAG - L)/K \quad (1)$$

where L and K are constants. L is an estimate of the lower limit of the CAG expansion at which phenotypic expression of the effects of mutant huntingtin could be observed, and K is a normalizing constant. When $L = 30$ and $K = 6.27$, CAP will be equal to 100 at the subject's expected age of onset of motor symptoms. CAP might be thought of as a measure of a subject's cumulative exposure to the toxic effects of mutant huntingtin. We have found that other values for L and K in the CAP score formula are preferable when modeling specific imaging

CHDI Foundation/CHDI Management, Princeton, New Jersey, USA. *Correspondence: JH Warner (john.warner@CHDIfoundation.org)
Received 18 March 2016; accepted 22 June 2016; published online on 2 August 2016. doi:10.1002/psp4.12097

variables. In what follows, we use CAP scores with $L = 30$ and $K = 6.27$ in global descriptive plots in which comparisons with each subject's expected age at motor onset is important. "Optimized" values of L and K are used when simulating trajectories and constructing scatter visual predictive checks.

We distinguish between the Base model, the Random Slope (RS) model, the Random Walk (RW) model, and the Full model. The Base model uses natural regression splines to model the effects attributable to aging in healthy individuals and the effects of the CAP score in subjects who are gene-positive for HD. The Base model also includes random intercepts, to account for between-subject variability, and traditional error terms designed to model measurement error. The RS model extends the Base model by adding random slopes to account for between-subject variability in the rate of progression. The RW model extends the Base model by adding an error term based on Brownian motion (or more formally the Weiner process) to provide an alternative model for variable progression rates. The Full model extends the Base model by adding both RS and RW error terms.

The RS and RW mechanisms are two ways of accounting for the observation that signs and symptoms from individuals with identical ages and CAG lengths often seem to progress at different rates. The RS model assumes that each subject possesses a unique deviation from the population average disease trajectory that persists throughout a significant portion of the patient's adult life. Such a model might hold if differences in disease progression were determined by the diversity of the patient population with respect to genetic, environmental, or life-style related factors. By contrast, the RW model assumes that the diversity of patient outcomes is the result of a random process acting over time in a uniform manner for all patients. The RW model might hold if progression was driven by a succession of environmental or life-style related shocks to which all patients are subject in the same manner. The RW model allows diverse patient outcomes to be obtained even in the absence of genetic, environmental, or life-style diversity. Keeping this in mind may protect one against embracing false causal conclusions.

The text book example of a RW generating diversity in an otherwise homogeneous population of individuals, is to consider the set of fortunes of N gamblers all starting with identical initial stakes and all playing identical games of chance (e.g., successive coin flips). It is well known that very substantial variability in the gambler's fortunes will be generated by such a process, even if the random processes affecting all gamblers are identical. We suggest that something similar might be occurring in slowly progressing diseases, such as HD.

The RW models required the development of the NONMEM code that implements a version of the Kalman filter similar to that introduced in ref. 4 and further developed in NONMEM 7.3[11] for the fitting of stochastic differential equation-based nonlinear mixed effect models. Our approach is novel to the HD field and, to the best of our knowledge, within the larger field of disease progression modeling. We adopt it because of our belief that RW mechanisms are plausible and underrecognized.

METHODS

General form of the model

For each imaging marker, separate Base, RS, RW, and Full models were fit to the population of HD participants and the population of healthy controls (both from TRACK-HD). Here, HD participants are defined as subjects who have tested positively for the HD gene expansion. HD participants may be divided into two groups: participants who have not yet been diagnosed with HD based on characteristic motor symptoms of the disease (called premanifest) and participants who have been so diagnosed (called manifest).

For healthy controls the Full model has the following form:

$$Y_{Cij} = S_C(AGE_{ij}, \beta_C) + d_{Ci} + \eta_{C1i} + \eta_{C2i}(AGE_{ij} - AGE0) + \gamma_{Cij}(t) + \epsilon_{Cij}. \quad (2)$$

The corresponding Full model for HD participants (including both manifest and premanifest participants) has the following form:

$$Y_{HDij} = S_C(AGE_{ij}, \beta_C) + S_{HD}(CAP_{ij}, \beta_{HD}) + d_{HDi} + \eta_{HD1i} + \eta_{HD2i}(AGE_{ij} - AGE0) + \gamma_{HDij}(t) + \epsilon_{HDij} \quad (3)$$

Where $Y_{Cij} = logit(Z_{Cij})$ is the logit transformation of the observation Z_{Cij} of the volume of a brain region from healthy control i at visit j. $Y_{HDij} = logit(Z_{HDij})$ is a similar observation made on an HD participant. Both Z_{Cij} and Z_{HDij} are expressed as percentages of the subject's intracranial volume. The logit transformation is defined as $logit(x) = log(p/(1-p))$, where $p = x/100$. Inverse logit transformations are applied to results from the models when estimates of Z_{Cij} and Z_{HDij} are needed. This procedure is guaranteed to produce estimates of Z_{Cij} and Z_{HDij} that are positive and bounded above by intracranial volume $\times$ 100. AGE_{ij} and CAP_{ij} are the age and CAP score of subject i at visit j[1] (i.e., $CAP_{ij} = AGE_{ij}(CAG_i - L)/K$). Where CAG_i is the CAG length of subject i. L and K are constants. For each imaging biomarker, separate models are fit with CAP scores defined with $L = 30$ and K=6.27 and with values of L and K that have been optimized for each biomarker. When $L = 30$ and $K = 6.27$ the CAP score will be equal to 100 at the expected age of onset of motor symptoms. The above connection with expected age at motor onset will hold only approximately for optimized choices of L and K. Details on the methods used to choose "optimal" values of L and K are given below.

- S_C and S_{HD} are natural cubic spline functions (as described in ref. 12) that estimate normal aging effects and the toxic effect of mutant huntingtin, respectively. S_C and S_{HD} are linear functions of the five dimensional parameter vectors β_C and β_{HD} but may be highly nonlinear functions of *Age* and *CAP*. Following recommendations on knot placement in ref. 12, the knots in S_C and S_{HD} are placed at the 5th, 27.5th, 50th, 72.5th, and 95th percentiles of the *AGE* and *CAP* score distributions, respectively. β_C is estimated in the population of healthy controls but held constant at its Base model value when ref. 3 is fit.

- d_{Ci} and d_{HDi} represent additive fixed effects in the control and HD populations. The present models include site effects (three dummy variables) and a gender effect.
- η_{C1}, η_{HD1} are random intercept terms (for healthy controls HD participants) η_{C2} and η_{HD2} are corresponding random slope terms. The $\eta_{HD} = (\eta_{HD1}, \eta_{HD2})$ and $\eta_C = (\eta_{C1}, \eta_{C2})$ are two-dimensional normal random variables with mean zero and covariance matrices Σ_{HD} and Σ_C. In tables, the SDs and correlations of these random variables will be designated by $\sigma(INT)$, $\sigma(SLOPE)$, and $cor(INT, SLOPE)$. The baseline constant $AGE0$ is taken to be 18.
- The ϵ_C and ϵ_{HD} are a normal residual error terms for the control and HD populations: they are assumed to have mean zero and SDs $sd(\epsilon_C)$ and $sd(\epsilon_{HD})$, respectively. These SDs will be designated as $\sigma(ERROR)$ in the tables.
- The $\gamma_C(t)$ and $\gamma_{HD}(t)$ are realizations of continuous time RWs (technically Wiener processes). These processes are assumed to be initialized to zero when subjects are 18 years of age. Their SDs are designated by $\sigma(W_C)$ and $\sigma(W_{HD})$ (or simply $\sigma(W)$ if the context makes the reference population clear. Following the definition of a Wiener process, t years after initialization $\gamma_C(t)$ will have an SD of $\sqrt{(t)}\sigma(W_C)$ and $\gamma_{HD}(t)$ will have an SD of $\sqrt{(t)}\sigma(W_{HD})$.
- The η, ϵ, and γ terms are assumed to be mutually independent. Note that the spline functions affect only the overall population trend. Random intercepts, RSs, and RWs are simply added on to this overall trend.

We refer to the model of Eq. 2 as the Full model for healthy controls. The submodel in which η_{C2} and $\gamma_C(t)$ both vanish is called the Base model. The submodel in which $\gamma_C(t)$ vanishes is called the RS model and the submodel in which η_{C2} vanishes is called the RW model. Similar naming conventions apply to the model of Eq. 3 and its submodels.

Model fitting

Final models were fit in NONMEM 7.3 using the first order conditional estimation method with interaction. Graphics, data analysis, and data management were performed in R version 3.2.0.[13] The Base and RS models are random effects models of the kind that NONMEM was designed to fit and their fitting is straightforward. The RW and Full models, however, require special attention in order to estimate the SDs of the random walks (γs). Inspection of Eqs. (2) and (3) above reveals them to be special cases of the Eqs. (1) and (2) from ref. 10. In particular, using the notation of ref. 10, the models of the current article reduce to a set of stochastic differential equations in which the measurement model is trivial (i.e., $f(x_{it}, \phi_i) = x_{ij}$) and system's dynamics are of a particularly simple form (i.e., the integral of $g(x_{it}, d_i, \phi_i)$ is given, depends on t, but does not depend on x_{it}). This leads to a simple Kalman filter algorithm, which is iterated at each step of the NONMEM optimization process and implemented through a user-defined $PRED script (see the online **Supplementary Material** for details) and NONMEM code.

Estimating optimal constants for the CAP score

In order to facilitate the estimation of nonlinear trends, the CAP score enters into the model of ref. 3 as an argument to the spline function S_{HD}. This complicates the process of fitting models that will optimize the choice of L for each imaging biomarker. Note that K is a normalizing constant that is useful in providing a consistent interpretation for the CAP score models but which does not affect the model fit.

To find optimal values of L, we set $K = 1$ and fit Base models to ref. 3 for each integer value of L between 21 and 40 inclusive. The Akaike information criterion (AIC) values for these models (subtracting off their minimum value) are plotted in **Figure 1** for each of the 10 imaging biomarkers. Optimal values of L (called L_{opt}) can be read from this graph. An appropriate normalizing constant K_{opt} can be computed from the formula:

$$K_{opt} = 6.27 \times \frac{43 - L_{opt}}{13} \tag{4}$$

The number 43 is a (somewhat arbitrary) centering constant, chosen because it is a very common value in the TRACK dataset and in data from other observational studies. Note that $L_{opt} = 30$ implies $k_{opt} = 6.27$, which agrees with the values used in ref. 1 and produces a CAP score that is equal to 100 at the expected age of motor onset. For any value of L_{opt} we may define a corresponding version of the CAP score as:

$$CAP(L_{opt}) = AGE \times (CAG - L)/K_{opt} \tag{5}$$

We have the following:

$$CAP(L_opt) = r \times CAP(30) \tag{6}$$

Where

$$r = \frac{1 + \frac{CAG-43}{43-L_{opt}}}{1 + \frac{CAG-43}{43-30}} \tag{7}$$

and $CAP(30)$ is the version of the CAP score with $L = 30$ and $K = 6.27$. If follows that $CAP(L) = 100 \times r$ at the expected age of onset of motor symptoms.

Model fitting in this section was done using version 3.1 of the NLME package in R.[14] Additional details are in the online Supplementary Material.

Track 36-month imaging data

All models were fit to 10 structural magnetic resonance imaging markers from the 36-month cutoff of the TRACK data.[2] The 36-month TRACK dataset contains annual observations on 366 subjects (45% men) equally distributed between four sites in Europe and North America. One third of the sample consisted of healthy controls. Of the HD participants, half were premanifest and half were in early stages of manifest disease. There are usually four observations for each subject for a total of 1,382. All markers represent volumes of anatomic features of the brain expressed as fractions of the total intracranial volume. The markers selected for this analysis include five markers processed by the Iowa center representing volumes of the thalamus, striatum, caudate nucleus, putamen, and whole brain. In addition, we consider five markers processed by the London center representing volumes of the caudate nucleus, whole brain, ventricle, cortical gray matter, and cortical white

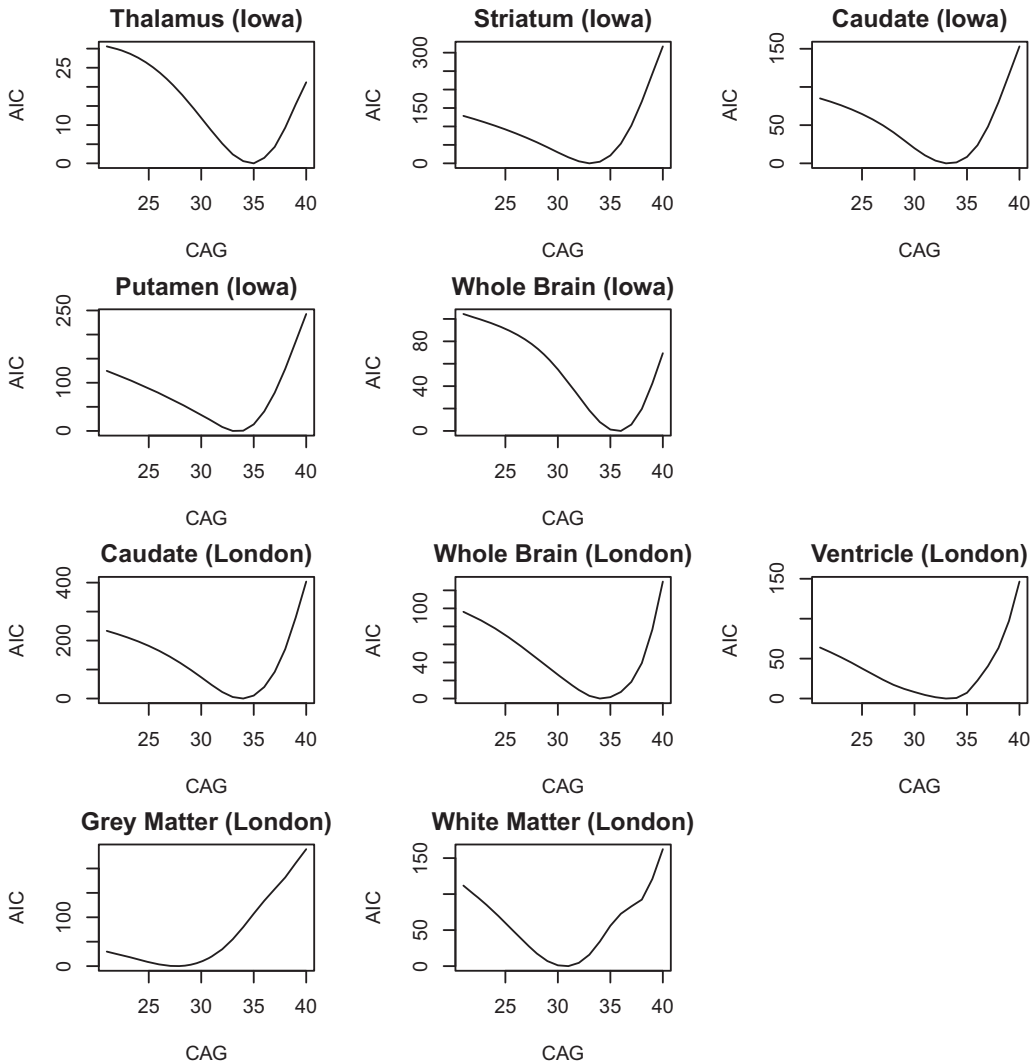

Figure 1 Plots of Akaike information criterion (AIC) for the Base model fitted with fixed values of L between 21 and 40 for all imaging biomarkers.

matter. All 10 measures are expressed as percentages of the participant's intracranial volume. The distinction between Iowa and London markers is important because the London markers were obtained using either the boundary shift integral or voxel-based morphometry (procedures that quantify annual changes in different locations of the brain within each individual). By contrast, the Iowa variables are simple snapshots of each brain at each time point.

RESULTS

Table 1 presents AIC statistics for comparing model fits of the Base model to the RS, RW, and Full models in the population of healthy controls. **Table 2** provides similar information for model fits in the HD population using optimal values of *L* from **Figure 1**. In both tables, lower values of AIC indicate better fit. AICs are normalized to equal zero for the Base model. **Table 3** compares models based on *CAP*(30) with models based on the $CAP(L_{opt})$: here, values of AIC less than zero indicate the superiority of the models based on $CAP(L_{opt})$. The AIC statistic,[15] is defined as $-2\log(lik)+2k$, where $-2\log(lik)$ is the objective function minimized by NONMEM and k is the number of model parameters. It seems from these tables that the Base model generally provides a better fit to the data for healthy controls than it does to the data for HD subjects. The largest improvements on the Base model, in both the healthy control and the HD populations, are for the cortical grey and cortical white matter variables: these improvements seem, most likely, to be accounted for by RS mechanisms particularly in the HD population. By contrast, in the HD population, RW mechanisms seem to be more important than RS mechanisms in explaining the volumes of the striatum, caudate (Iowa only), putamen, whole brain (London only), and the ventricle. Even in cases in which RS mechanisms outperform RW mechanisms (under the AIC criterion), the RW mechanisms

Table 1 AIC statistics for model comparisons in healthy controls

Marker	RS	RW	Full
Thalamus (Iowa)	0.4510	2.0000	−2.1320
Striatum (Iowa)	1.0030	−8.3020	−4.6940
Caudate (Iowa)	0.3450	−6.5080	−4.2410
Putamen (Iowa)	3.9810	−7.7520	−11.3690
Whole brain (Iowa)	−9.7910	2.0000	−9.0710
Caudate (London)	3.9050	−5.0190	−2.1610
Whole brain (London)	−13.5100	−5.2790	−21.5490
Ventricle (London)	−8.2620	−7.7760	−6.4010
Grey matter (London)	−116.2970	−107.6350	−118.0770
White matter (London)	−55.1570	−57.2220	−55.6260

AIC, Akaike information criterion; RS, Random Slope; RW, Random Walk.
Entries in rows 2–4 represent the AIC for the given model minus the AIC for the Base model. Positive values suggest the superiority of the Base model.

Table 3 AIC statistics for paired comparisons between models with optimal L and L = 30

Marker	Lopt	Base	RS	RW	Full
Thalamus (Iowa)	35.0000	−12.0110	−10.9660	−10.9930	−11.1560
Striatum (Iowa)	33.0000	−29.5850	−24.9830	−30.0490	−30.0750
Caudate (Iowa)	33.0000	−19.7540	−18.8480	−18.3180	−18.4100
Putamen (Iowa)	33.0000	−33.2110	−21.8700	−35.0190	−30.5340
Whole brain (Iowa)	36.0000	−55.3470	−45.9320	−50.3270	−50.2510
Caudate (London)	34.0000	−73.0850	−41.4110	−54.9780	−45.5380
Whole brain (London)	34.0000	−26.6410	−9.5470	−23.7830	−19.7720
Ventricle (London)	33.0000	−8.0800	5.4280	−6.6800	−3.6350
Grey matter (London)	28.0000	−9.4930	0.0930	−4.9570	−0.6880
White matter (London)	31.0000	−1.0510	3.9570	−2.2870	3.5070

AIC, Akaike information criterion; Lopt, optimal L; RS, Random Slope; RW, Random Walk.
Negative values indicate superiority of the model with optimal L.

tend to show substantial improvement over the Base model and the Full model tends to show an improvement over the RS model.

For the remainder of this section we focus on models for the Striatum (Iowa), which is regarded as the focal point for HD pathology. Complete details on all variables appear in the online **Supplementary Material**.

Table 2 AIC statistics for model comparisons in gene-positive HD participants

Marker	Lopt	RS	RW	Full
Thalamus (Iowa)	35.0000	3.2200	1.7730	4.3400
Striatum (Iowa)	33.0000	−7.9300	−36.8360	−33.2960
Caudate (Iowa)	33.0000	2.2410	−17.2990	−13.4510
Putamen (Iowa)	33.0000	−1.8020	−35.9590	−35.2970
Whole brain (Iowa)	36.0000	3.9890	−2.6570	−0.6560
Caudate (London)	34.0000	−108.3630	−92.1840	−121.5310
Whole brain (London)	34.0000	−28.0700	−55.6440	−53.8610
Ventricle (London)	33.0000	−77.0380	−117.9070	−114.2830
Grey matter (London)	28.0000	−258.0050	−195.1760	−261.3640
White matter (London)	31.0000	−360.7740	−266.8660	−361.5040

AIC, Akaike information criterion; HD, Huntington's disease; Lopt, optimal L; RS, Random Slope; RW, Random Walk.
All model fits use L_{opt}. Entries in rows 3–5 represent the AIC for the given model minus the AIC for the Base model. Positive values suggest the superiority of the Base model.

Figure 2 provides what we will call global descriptive plots for the striatum. Global descriptive plots are side-by-side plots with information on healthy controls on the left and information on HD subjects in the center and right. The left-hand panel plots the logit transformation of each marker controlling for covariates ($Y_C - d_C$) against AGE for healthy controls. The center and right-hand panels plot the logit transformation of the striatum controlling for covariates and normal aging ($Y_{HD} - S_C(AGE) - d_{HD}$) against CAP(30) and $CAP(L_{opt})$. The center and right-hand panels are very similar in appearance. However, the connection with age-at-motor onset is more direct in the center panel which is, therefore, recommended for routine use. Each subject's initial status (control, premanifest, and manifest) is indicated by color coding. The solid trend lines provide visual representations of $S_C(AGE)$ (left panel) and $S_{HD}(CAP)$ (center and right panels). In particular, $S_C(AGE)$ represents the effects of normal aging (controlling for covariates) and $S_{HD}(CAP)$ represents disease-related effects (controlling for covariates and normal aging). "Spaghetti" plots show individual observations. The spaghetti plots show that observations on each subject span a very limited portion of the complete time course of the disease. The solid lines are central to our approach to estimating the complete time courses. Both normal aging and HD pathology show a pronounced and well-known negative correlation with striatal volume. The plots in **Figure 2** are based on the Full model. Plots based on Base, RS, and RW models are very similar.

Figure 3 shows the results of 50 simulated trajectories of striatal volumes under the Base, RS, RW, and Full models. The simulations assume CAG lengths of 42 and plot trajectories for ages between 18 and 70 years. The simulated trajectories do not include measurement error. Trajectories for the Base models appear as smooth parallel curves.

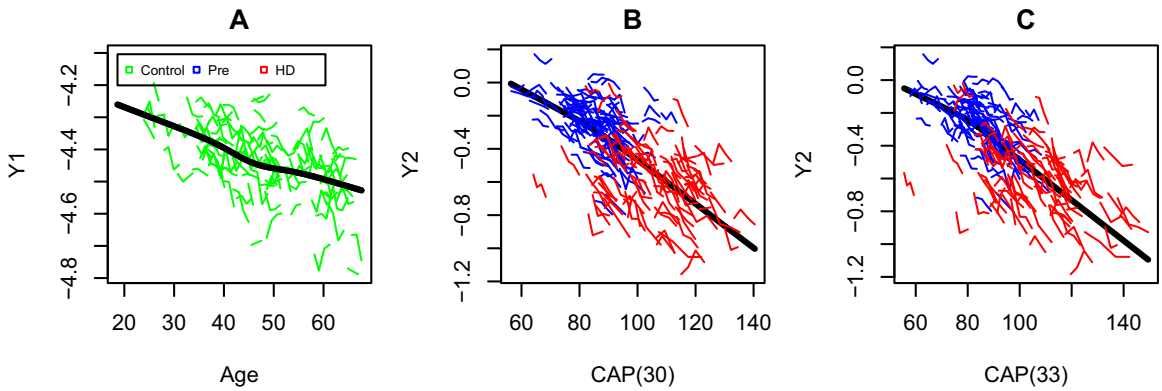

Figure 2 Global descriptive plot for $Y = \text{logit}\left(100 \times \frac{\text{Striatum(Iowa)}}{ICV}\right)$: Full Model. Panel (**a**): $Y1 = Y - d_C$ vs. AGE in Healthy Controls; Panel (**b**): $Y2 = Y - S_C(AGE) - d_{HD}$ vs. cytosine-adenine age product (CAP30) in Huntington's disease (HD) participants; Panel (**c**): $Y2 = Y - S_C(AGE) - d_{HD}$ vs. CAP(L_{opt}) in HD participants. Based on Eqs. 2 and 3.

Trajectories for RS models are nonparallel but smooth. Trajectories for RW and Full models are nonparallel and non-smooth. AIC values from **Table 2** indicate that the RW model has the best fit. Visually, **Figure 3** shows that the RW model displays the smallest deviation about the trend line. The RW model is also lacking in signs of age-related heteroscedasticity that are particularly apparent in the Base model.

Figure 4 presents scatter visual predictive checks for the Full models of striatal volume. Simulations were performed for healthy controls and gene-positive HD subjects with CAG lengths of 39, 40, 41, 42, 45, 48, 50, and 59. These CAG lengths account for about half of all HD participants. The restriction to subjects with the above CAG lengths was made so that all results for each imaging marker could be presented on a single graph. Results for excluded CAG lengths are similar. In addition, to facilitate plotting on a single graph, all plotted data values include corrections for each subject's covariates. Simulations are based on 1,000 replicates for healthy controls and 1,000 replicates for HD subjects at each CAG length. Solid lines represent 5%, 50%, and 95% quantiles of the distribution of predicted observations from the RW model. The implied 90% predictive interval completely captures all observations from 75% of healthy controls and 80% of gene-positive HD participants and captures some observations from 90% of healthy controls and 91% of gene-positive HD participants. Details for all 10 biomarkers are outlined in the online **Supplementary Material**.

DISCUSSION

Pathology in HD develops very slowly over many years. In the interest of understanding the dynamics of HD, considerable effort has been expended in collecting prospective data on individuals at various stages of the disease. At present, however, few subjects have been observed for the whole time course of the disease. As a consequence, if a complete disease progression model is wanted, there is a need to construct one by "patching together" many short time courses obtained from many different individuals and to adjust the resulting long time course estimate for covariates and changes that would be expected to occur in unaffected individuals. The current article aims to improve our understanding of HD by proposing a solution to this problem.

Our approach involves two novel methodological features: the use of RW error terms to model random aspects of disease dynamics and the use of the CAP score to account for effects related to the length of the mutant CAG expansion that lies at the root of HD.

Regarding the use of RW error terms, the framework is similar to that used in the stochastic differential equation-based approaches to the modeling of pharmacokinetic and pharmacodynamic.[10,16–21] In the present study, the above framework is applied to the modeling of disease progression. In the case of HD, and perhaps more generally in the study of disease progression, our mechanistic understanding is not sufficiently developed to determine a specific differential equation that drives the pathological process. For this reason, we replace the differential equations used in pharmacokinetic/pharmacodynamic modeling with flexible regression splines. The regression splines are applied to the subject's age at the time of observation for healthy controls and to the CAP score in gene-positive HD subjects. This has the effect of simplifying and streamlining the Kalman filter algorithm of ref. 10. Our Kalman filter algorithm is summarized in the online **Supplementary Material**.

Regression splines were used because: (1) some clear nonlinear trends appear to be present in the data; (2) the dynamics leading from excess CAG length and/or aging to observed pathology are too complex to be modeled mechanistically at the present time; and (3) we did not want to develop ad hoc empirical models for each of the dependent variables in our dataset. Regression splines have been previously used to model HD progression data in ref. 8.

The introduction of RW error terms into disease progression modeling reflects a shift from a paradigm in which each subject's disease progression is represented by a regression curve in two-dimensional space, to a paradigm in which each subject's progression is represented as a continuous time stochastic process measured with error at

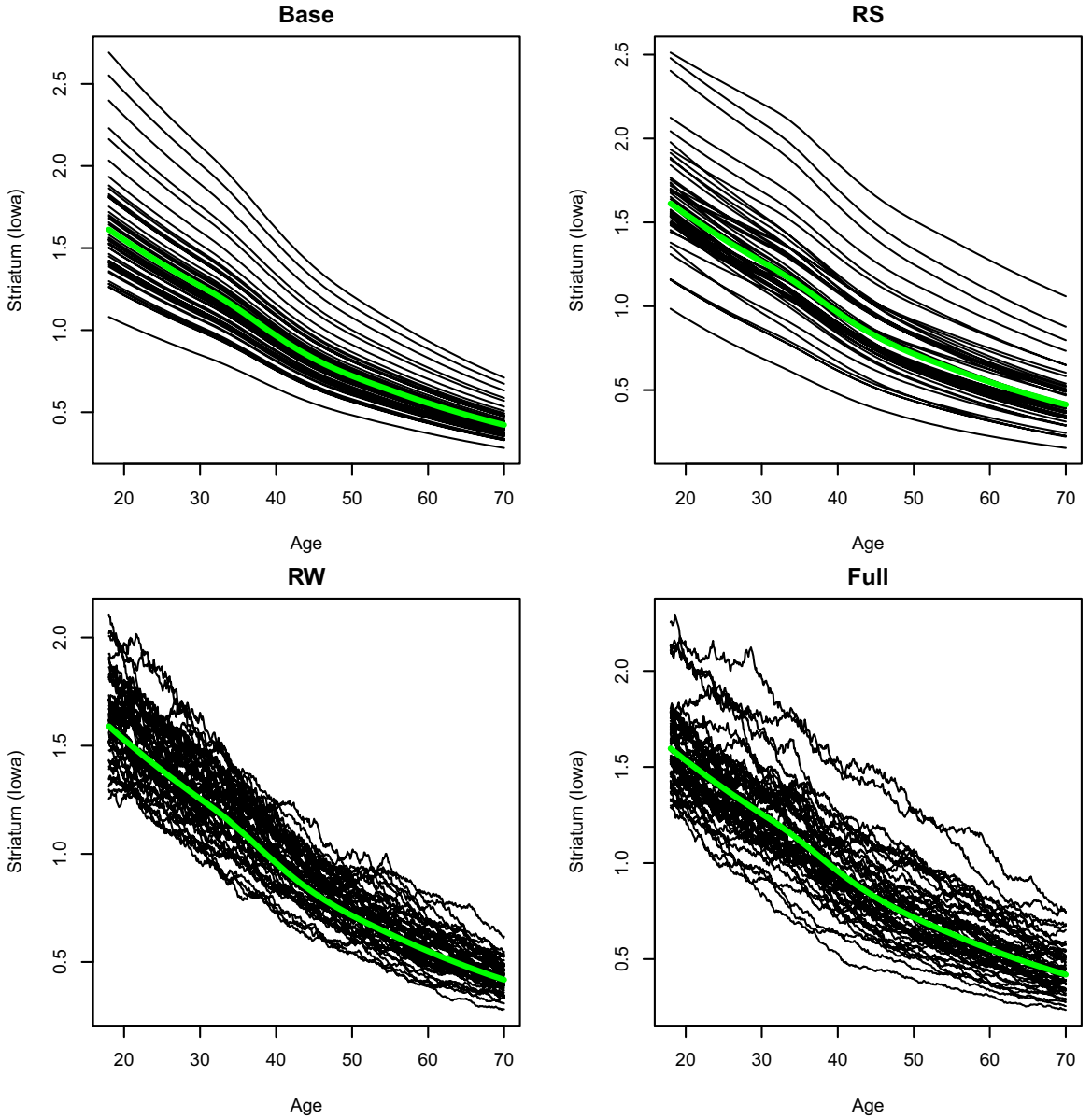

Figure 3 Simulated trajectories for striatum (Iowa): Huntington's disease (HD) participants trajectories are simulated under the Base, Random Slope (RS), Random Walk (RW), and Full models in the population of HD participants. The simulations are based on 50 replicates and assume cytosine-adenine-guanine (CAG) lengths of 42. Trajectories are simulated and plotted for ages starting at 18 years and ending at 70 years. Simulated trajectories are for a male equally likely to be selected from any of the four sites.

a discrete set of time points. Stochastic processes of this sort are commonly used in the modeling of many natural and manmade phenomena from the trajectories of space vehicles to the movement of stock prices. A readable account of relevant applications and theory is given in ref. 22. We feel that the stochastic process paradigm will, in time, demonstrate significant improvements over the simpler regression curve approach to disease progression modeling.

Future work will: (1) reproduce the work reported here on datasets that include longer time series of data for each subject. Longer time series should improve our ability to distinguish between RS and RW errors. (2) Apply the models developed here to other, noisier markers of the HD disease state. (3) Extend the models developed here to multivariate contexts where two or more markers are analyzed simultaneously. (4) Apply the modeling principles developed here to other neurodegenerative or trinucleotide repeat diseases. (5) Investigate alternatives to the AIC for models involving RW terms.

Although the use of RW error terms may be a general innovation that applies to a wide class of disease progression models, the CAP score encapsulates many of the features that are specific to HD. As the name implies, the CAP score is just a way of parameterizing the interactions between age and CAG length that figure in our prediction

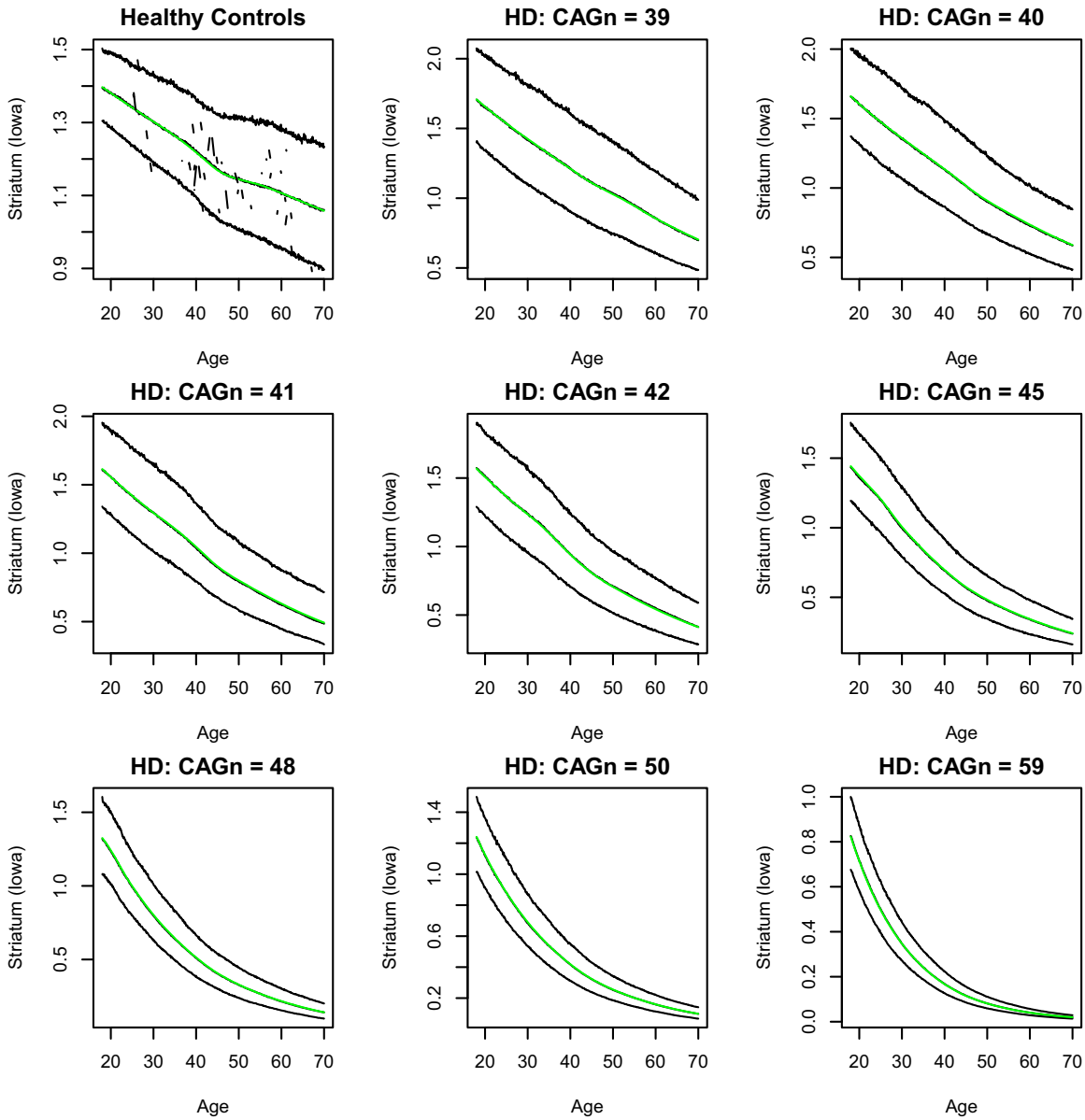

Figure 4 Scatter visual predictive checks for striatum (Iowa): full model simulations use 1,000 replicates for healthy controls and gene-positive Huntington's disease (HD) subjects with cytosine-adenine-guanine (CAG) lengths of 39, 40, 41, 42, 45, 48, 50, and 59. Simulated observations are for men from site 4. All observed data points are adjusted for site and gender. Solid black lines represent 5%, 50%, and 95% quantiles of the distribution of predicted observations. Solid green lines indicate model population predictions.

formula. Identical predictions could be obtained using more traditional parameterizations of the models in terms of main effects and interactions. From an interpretive point of view, however, the CAP score (1) suggests a connection with mechanistic models based on the cumulative toxicity of mutant huntingtin; (2) models the process by which increasing CAG length accelerates the progression of pathology in HD; (3) provides a connection with models that predict age at motor onset in HD; and (4) allows for easier comparison between progression models for different imaging biomarkers.

Point 2 above calls out for some additional discussion. Although the role of CAG length in accelerating the pathological process in HD is well established prior to motor onset, there remains some controversy regarding its role after motor onset. In particular, ref. 4 found that the interval between motor onset and death is independent of CAG length. Two interpretations of this finding are advanced by the authors. The first interpretation suggests that the role of CAG length as a driver of HD pathology diminishes as the disease progresses. The second interpretation posits two distinct CAG length dependent processes one leading to motor onset and another leading to death. Neither interpretation suggests that the role of CAG length terminates at motor onset. Indeed, there is evidence that the role of CAG length in disease progression is not so

terminated.[2,3] We are interested in understanding how the effect of CAG length on disease progression changes over time. This issue is of more than academic interest as it may have implications for the efficacy of gene silencing therapies in the later stages of the disease. This, and other issues, will be addressed in forthcoming articles.

Acknowledgments. We would like to thank Doug Langbehn, Yaning Wang, and other members of the Model-HD and CAP score working groups for helpful discussions on the CAP score model; Robert Bauer of ICON Development Solutions and the members of the NONMEM Users Group for help with the Kalman filter algorithm; and Sarah Tabrizi, the investigators of the TRACK project and the TRACK participants for the high quality data used in this analysis.

Conflict of Interest. The authors declared no conflict of interest.

Author Contributions. J.W. contributed model development, model fitting, data analysis, and writing of the manuscript. C.S. provided medical input regarding the accuracy of the text and the relevance of the models.

1. Ross, C.A. *et al.* Huntington disease: natural history, biomarkers and prospects for therapeutics. *Nat. Rev. Neurol.* **10**, 204–216 (2014).
2. Tabrizi, S.J. *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.* **12**, 637–649 (2013).
3. Rosenblatt, A., Kumar, B.V., Mo, A., Welsh, C.S., Margolis, R.L. & Ross, C.A. Age, CAG repeat length, and clinical progression in Huntington's disease. *Mov. Disord.* **27**, 272–276 (2012).
4. Keum, J.W. *et al.* The HTT CAG-expansion mutation determines age at death but not disease duration in Huntington disease. *Am. J. Hum. Genet.* **98**, 287–298 (2016).
5. Langfelder, P. *et al.* Integrated genomics and proteomics define huntingtin CAG length-dependent networks in mice. *Nat. Neurosci.* **19**, 623–633 (2016).
6. Penney, J.B. Jr, Vonsattel, J.P., Macdonald, M.E., Gusella, J.F. & Myers, R.H. CAG repeat number governs the development rate of pathology in Huntington's disease. *Ann. Neurol.* **41**, 689–692 (1997).
7. Zhang, Y. *et al.* Indexing disease progression at study entry with individuals at-risk for Huntington disease. *Am. J. Med. Genet. B Neuropsychiatr. Genet.* **156**, 751–763 (2011).
8. Paulsen, J.S. *et al.* Clinical and Biomarker Changes in Premanifest Huntington Disease Show Trial Feasibility: A Decade of the PREDICT-HD Study. *Front. Aging. Neurosci.* **6, 78**, 1–11 (2014).
9. Paulsen, J.S. *et al.* Prediction of manifest Huntington's disease with clinical and imaging measures: a prospective observational study. *Lancet Neurol.* **13**, 1193–1201 (2014).
10. Tornøe, C.W., Overgaard, R.V., Agersø, H., Nielsen, H.A., Madsen, H. & Jonsson, E.N. Stochastic differential equations in NONMEM: implementation, application, and comparison with ordinary differential equations. *Pharm. Res.* **22**, 1247–1258 (2005).
11. Bauer, R. Introduction to NONMEM 7.3.0 (2013). <http://www.cognigencorp.com/nonmem/current/2013-December/4868.html>.
12. Harrell, F.E. Jr. Regression Modeling Strategies: with applications to linear models, logistic regression, and survival analysis (Springer Series in Statistics). Vol. **3** (Springer, New York, NY, 2014).
13. R.D.C. Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. 2013 (2014). <www.r-project.org>.
14. Pinheiro, J., Bates, D., DebRoy, S., Sarkar, D. & R Core Team. nlme: Linear and Nonlinear Mixed Effects Models (2016). <http://CRAN.R-project.org/package=nlme>. R package version 3.1-124.
15. Akaike, H. A new look at the statistical model identification. *IEEE Trans. Automat. Contr.* **19**, 716–723 (1974).
16. Tornøe, C.W., Jacobsen, J.L. & Madsen, H. Grey-box pharmacokinetic/pharmacodynamic modelling of a euglycaemic clamp study. *J. Math. Biol.* **48**, 591–604 (2004).
17. Mortensen, S.B., Klim, S., Dammann, B., Kristensen, N.R., Madsen, H. & Overgaard, R.V. A matlab framework for estimation of NLME models using stochastic differential equations: applications for estimation of insulin secretion rates. *J. Pharmacokinet. Pharmacodyn.* **34**, 623–642 (2007).
18. Overgaard, R.V., Jonsson, N., Tornøe, C.W. & Madsen, H. Non-linear mixed-effects models with stochastic differential equations: implementation of an estimation algorithm. *J. Pharmacokinet. Pharmacodyn.* **32**, 85–107 (2005).
19. Klim, S., Mortensen, S.B., Kristensen, N.R., Overgaard, R.V. & Madsen, H. Population stochastic modelling (PSM)–an R package for mixed-effects models based on stochastic differential equations. *Comput. Methods Programs Biomed.* **94**, 279–289 (2009).
20. Møller, J.B., Overgaard, R.V., Madsen, H., Hansen, T., Pedersen, O. & Ingwersen, S.H. Predictive performance for population models using stochastic differential equations applied on data from an oral glucose tolerance test. *J. Pharmacokinet. Pharmacodyn.* **37**, 85–98 (2010).
21. Berglund, M., Sunnåker, M., Adiels, M., Jirstrand, M. & Wennberg, B. Investigations of a compartmental model for leucine kinetics using non-linear mixed effects models with ordinary and stochastic differential equations. *Math. Med. Biol.* **29**, 361–384 (2012).
22. Jazwinski, A.H. *Stochastic processes and filtering theory*. (Courier Corporation, North Chelmsford, MA, 2007).

Supplementary information accompanies this paper on the *CPT: Pharmacometrics & Systems Pharmacology* website (http://www.wileyonlinelibrary.com/psp4)