Journal of
Neurology, Neurosurgery
& Psychiatry

Modelling the natural history of Huntington's disease progression

Journal:	*Journal of Neurology, Neurosurgery, and Psychiatry*
Manuscript ID:	jnnp-2014-308153.R1
Article Type:	Research paper
Date Submitted by the Author:	26-Jul-2014
Complete List of Authors:	Kuan, Wei-Li; University of Cambridge, Department of Clinical Neuroscience Kasis, Andreas; University of Cambridge, Department of Engineering Yuan, Ye; University of Cambridge, Department of Engineering Mason, Sarah; University of Cambridge, Department of Clinical Neuroscience Lazar, Alpar; University of Cambridge, Department of Clinical Neuroscience Barker, Roger; University of Cambridge, Department of Clinical Neuroscience Gonçalves, Jorge; University of Cambridge, Department of Engineering; University of Luxembourg, Luxembourg Centre for Systems Biomedicine
Keywords:	HUNTINGTON'S, MOVEMENT DISORDERS, CLINICAL NEUROLOGY, STATISTICS
Specialty:	Movement disorders

Modelling the natural history of Huntington's disease progression

Kuan WL[1,*], Kasis A[2,*], Yuan Y[2], Mason SL[1], Lazar AS[1], Barker RA[1,†], Gonçalves J[2,3,†]

[*]These first authors contributed equally to this work
[†]These senior authors contributed equally to this work. To whom correspondence should be addressed to jmg77@cam.ac.uk, +44 (0)1223332770
[1]Department of Clinical Neurosciences, John van Geest Centre for Brain Repair, University of Cambridge, Forvie Site, Robinson Way, Cambridge, CB2 0PY, UK
[2]Department of Engineering, University of Cambridge, Cambridge CB2 1PZ, UK
[3]Luxembourg Centre for Systems Biomedicine, University of Luxembourg, 7 Avenue des Hauts-Fourneaux, L-4362 Esch-sur-Alzette, Luxembourg

Keywords: Huntington's disease, Unified Huntington's Disease Rating Scale (UHDRS), disease modelling, longitudinal study

Word count for the main text: 3500

Number of references: 18

ABSTRACT

Background The lack of reliable biomarkers to track disease progression is a major problem in clinical research of chronic neurological disorders. Using Huntington's disease (HD) as an example, we describe a novel approach to model HD and show that the progression of a neurological disorder can be predicted for individual patients.

Methods Starting with an initial cohort of 343 HD patients that we have followed since 1995, we used data from 68 patients that satisfied our filtering criteria to model disease progression, based on the Unified Huntington's Disease Rating Scale (UHDRS), a measure that is routinely used in HD clinics worldwide.

Results Our model was validated using: a) extrapolating our equation to model the age of disease onset, b) testing it on a second patient dataset by loosening our filtering criteria, c) cross-validating with a repeated random sub-sampling approach, and d) holdout validating with the latest clinical assessment data from the same cohort of patients. With UHDRS scores from the past four clinical visits (over a minimum span of two years), our model predicts disease progression of individual patients over the next two years with an accuracy of 89-91%. We have also provided evidence that patients with similar baseline clinical profiles can exhibit very different trajectories of disease progression.

Conclusion This new model therefore has important implications for HD research, most obviously in the development of potential disease-modifying therapies. We believe that similar approach can also be adapted to model disease progression in other chronic neurological disorders.

INTRODUCTION

The lack of reliable biomarkers to track disease progression is a major problem in clinical research of chronic neurological disorders.[1] This problem has gained prominence as the development of disease-modifying therapies starts to enter the clinic,[2] especially as some of these novel therapeutic agents or therapies involve direct delivery into the brain, and as such randomised controlled trials (RCT) are not always possible.[3] Furthermore RCTs in neurological disorders with a low prevalence, such as Huntington's disease (HD), can be further complicated by virtue of difficulties in patient recruitment.

HD is a genetic neurodegenerative disorder that affects 2.71 per 100,000 persons worldwide.[4] The pathology of HD is caused by an expansion in a trinucleotide CAG repeat in exon 1 of the Huntingtin gene, and the length of this repeat predicts disease onset in patients.[5,6] Models predicting disease onset enable researchers to study HD before the start of overt disease features and by so doing the possibility of delivering novel therapies at disease onset.[7] Whilst useful, we are still poor at modelling disease progression once the condition has started. We therefore sought to do this using our extensive database of 343 patients that we have followed longitudinally since 1995. We propose a model that tracks and predicts the natural progression of manifest HD based on the motor and functional components of the Unified Huntington's Disease Rating Scale (UHDRS), which are routinely used in HD clinics.

We have found that patients with similar initial clinical profiles can have very different patterns of disease progression, which renders the use of conventional regression analysis (that estimates a common slope among groups of patients) inappropriate. In contrast, our novel approach enables researchers to predict disease progression of individual patients for the next two years, based on assessments from the past four clinical visits (with a minimum span over two years). We have interrogated the quality of our prediction by: a) extrapolating our equation to model the age of disease onset, b) testing it on a second patient dataset by loosening our filtering criteria, c) cross-validating with a repeated random sub-sampling approach, as well as d) holdout validating with the latest clinical assessment data from the same cohort of patients, which was unavailable at the time of our original modelling.

We believe that our model will benefit clinicians and researchers in studying HD, especially for those developing potential disease-modifying therapies. Furthermore, our results enable researchers to reassess their existing data based on different profiles of patients' predicted disease progression. Similar approach can also be adapted to model disease progression in other chronic disorders.

MATERIALS AND METHODS

Patient recruitment and assessments

Data was collected from participants who attended the HD clinic at the John van Geest Centre for Brain Repair, UK, between 1995- 2013, either as part of their routine clinical care, or through participation in related studies. This study was approved by the Cambridge University Hospital NHS Foundation Trust, in accordance with the ethical standards laid down in the 1964 Declaration of Helsinki and its later amendments. All participants consented to their data being shared between research studies in an anonymised form. Motor and functional impairments were assessed using the UHDRS total motor score, functional assessment, and functional capacity scales, conducted by an experienced rater. The UHDRS total motor score ranges from 0 (no motor features detected) to a maximum score of 124. Manifest disease was defined as a total motor score ≥5, as done previously.[8] Demographic information was collected on patients including their CAG repeat size (where available), age and gender.

Modelling methods

A detailed description can be found in the supplementary file. In short, the initial data were filtered using the following three criteria: a) patients beyond the prodromal stage of disease with a ≥15 on their Generalised Index (GI) score (explained below) at their last clinical visit (up to 2012); b) patients with at least five clinical visits; and c) patients not showing a large negative validity score (table S1). The validity score was created to avoid potential complications from medications, based on an assumption that patients were not expected to improve with time, which is consistent with a recent finding.[9] The validity of patients that showed improvement between two consecutive visits was penalised and their data is more prone to exclusion. Most patients were filtered out using criteria a) and b) above and at the end of the filtering process 68 patients were eligible for modelling (figure S2).

Clinical data from the UHDRS was then transformed into GI scores, by deducting chorea and dystonia that has higher interrater variability,[10] which in our experience could fluctuate over short periods of time, and did not correlate over time in our data (figure S1). The GI is normalised to 100, which represents an average of the motor and functional components of the UHDRS, and lies between 0 (no features) and 100 (all features). The optimum function to data from individual patients was fitted to best describe his/her GI progression, using linear ($GI=B_0+B_1*Age$), quadratic ($GI=B_0+B_1*Age+B_2*Age^2$), and exponential ($GI=\exp[B_0+B_1*Age]$) models. The fitness of each model was quantitated as described in the supplementary file and previously.[11] We then searched the optimum coefficient for individual patients (B_1 for linear, B_2 for quadratic, both indicates the rate of disease progression) within a range (table S2), defined by the maximum and minimum value from the patient data (table S3).

Prediction and model validation

A detailed description can be found in the supplementary file. In short, the first N samples of each patient were used to classify and create a model that would describe the disease progression pattern for that patient (figure S3). The optimum coefficient was then generated as described above, while the other parameters were derived from the first n samples from that particular patient. Prediction is conducted in a moving-horizon sense;

as soon as the $(n+1)^{th}$ sample is available, the above algorithm would re-classify and refine the model for that patient. Our predictions were validated using four different methods. Firstly, we extrapolated our model to predict the age of disease onset and compared that with a benchmark model built using large cohorts of patients.[5] Secondly, we included data from the 31 new patients that were previously excluded for failing the validity score criteria. Thirdly, we took a repeated random sub-sampling approach by partitioning our 68 patients into training groups of 50 patients, with testing groups of 18 patients. The procedure was repeated 40 times to eliminate selection bias. Finally, 23 out of our 58 patients had revisited our clinic in 2013, such that their latest clinical assessment data was unavailable at the time of modelling, and we could therefore use these data to validate the predicted versus actual GI score of disease progression.

Data analysis

Normality of data was verified using either the Kolmogorov-Smirnov (>50 samples) or the Shapiro-Wilk (≤50 samples) tests. For univariate and multivariate analyses we used parametric (e.g. ANOVA, mixed models ANOVA) and nonparametric methods (Friedman test, Kruskal-Wallis and Mann-Whitney test) depending on the distribution of residuals. For mixed model analyses of variance with multiple levels of repeated measures data transformation was performed to obtain normality if required (such as square root transformation). Univariate analyses were corrected for multiple comparisons (Bonferroni) in order to avoid type I statistical error. Matlab (version 7.9) were used for data modelling. SAS (version 9.1) and SPSS (version 21) were used for statistical analysis.

RESULTS

The initial dataset contained a total number of 343 patients that we had followed since 1995. The majority of the patient data failed to meet our filtering criteria were those having less than five clinical assessments, followed by those in the premanifest or prodromal stage.

Disease progression from the final 58 patients that we were able to model, denoted by the GI, representing 85.3% of all eligible patients (n=68), are demonstrated (figure 1). Among these patients, 41 of them exhibited disease progression that could be described in a linear equation (GI = B_0+B_1*Age, left panels), while disease progression from the other 17 patients could be described in a quadratic equation (GI = B_0+B_1*Age+B_2*Age^2, right panels).

	All n=343	Filtered n=68	Linear n=41	Quadratic n=17	Test 2 n=31	P
Age at last visit (up to 2012)[a]	52.2±1.0	55.1±1.5	55.7±1.9	53.7±3.3	55.7±2.1	0.267
Male (%)[b]	44.3 (n=152)	39.7 (n=27)	31.7 (n=13)	58.8 (n=10)	45.2 (n=14)	0.243
CAG repeat[b]	43 (4)	44 (3)	44 (3)	44 (4)	43 (3)	0.018[c]
Years of follow-up[b]	5.1 (5.0)	6.0 (4.0)	6.6 (3.5)	6.0 (4.0)	5.7 (7.4)	0.009[c]
UHDRS motor score at last visit[b]	17 (30)	36 (36)[d]	45 (34)[d]	30 (26)[e]	18 (17)	<0.001
UHDRS functional score at last visit[b]	28 (24)	12 (19)[d]	8.5 (17)[d]	12 (20)[e]	25 (10)	<0.001

Table 1 [a]Data from normally distributed samples was expressed as mean ±SEM; [b] data from nonparametric samples was expressed as median (interquartile range, IQR); [c]Post-hoc Mann Whitney test with Bonferroni correction revealed no true difference; [d]Post-hoc analysis revealed significant differences (P<0.005, after Bonferroni adjustment) from the full cohort (n=343) and from the "test two" patients (relaxed validity score, n=31); [e]Post-hoc analysis revealed a tendency for a difference (P<0.05, after Bonferroni adjustment) between the full and included patients.

We then compared the demographic information between different subgroups of patients, including data sets from a further 31 patients used in validation test two by relaxing the validity score. Clinically the subgroups could not be distinguished from one another in terms of the age of patients at their last clinical assessment (up to 2012, $F_{4,493}$=1.306, P=0.267), gender distribution (H(4)=5.460, P=0.243), CAG repeat size (H(4)=11.953, P=0.018), and the average years of follow-up (H(4)=13.543, P=0.009) (table 1, post-hoc analysis with the latter two revealed no real difference). There were significant difference between patient subgroups in their UHDRS total motor (H(4)=55.216, P<0.001) and functional (H(4)=61.724, P<0.001) assessments as measured at the patient's last clinical assessment (table 1). Post-hoc analysis revealed that the data from all patients eligible for modelling (n=68), as well as patients exhibiting a linear (n=41), or a quadratic disease

progression (*n*=17) were either significantly, or had a strong tendency to be more impaired in their UHDRS total motor and functional assessments, compared with the whole cohort (*n*=343) or those patients who failed the validity score requirement that were used for validation test two (*n*=31). Such a difference could be attributed to the presence of the validity score, which assumed gradual deterioration of HD features over time and was consistent with previous findings.[9,12] However there were no inter-group differences between all patients eligible for modelling, and those demonstrating linear or quadratic disease progressions. Overall this indicates that patients sharing similar clinical profiles could exhibit very different patterns of disease progression.

As there is no comparable model to predict HD progression, we started validating our approach by extrapolating our model to predict the age of disease onset. Comparison was made with the most popular existing model to predict disease onset, constructed using large, independent cohorts of patients.[5] This helps deal with a major limitation of our model to predict disease progression, namely that we were only studying a relatively small cohort of patients from the east of England. To do this, data was selected from the 41 patients with linear disease progression and the age of disease onset (T_0) was defined as $T_0=-\frac{B0}{B1}$. Our predictions were modelled by using an approach similar to Langbehn's, as well as by evaluating the maximum fitness criterion. Our resulting equation (T_0=22.24+exp[9.844-0.156*CAG]), has very similar coefficients to Langbehn's (T_0=21.54+exp[9.556-0.146*CAG]) (figure 2). This indicates that, despite the fact that our model was built using a smaller cohort, our approach is comparable to the benchmark disease onset model constructed using 2913 individuals.

Our present modelling approach was based on the assumption that patients were not expected to improve over time, consistent with a recent observation.[9] Therefore patients demonstrating large improvements over consecutive clinical assessments were penalised, making their data sets prone to exclusion from the filtering stage. We then revisited our assumption, by including data from 31 patients who were previously excluded due to them having an excessive validity score penalty. The results were analysed in terms of the percentage change between the predicted and actual GI score during the latest clinical visit, which ranged between <6 months, 6-12 months, 12-18 months, and 18-24 months from the last visit used for modelling. Furthermore, we sought to investigate whether the number of prior clinical assessments used for modelling affected the accuracy of prediction. The prior clinical assessments were grouped in categories ranging from 3 to ≥8 prior visits, although our classification algorithm requires at least four prior clinical assessments to properly assign individual patients to their respective type of disease progression. By removing the validity score this represents the maximum level of prediction error we would expect from our modelling approach.

We used mixed model analyses of variance to estimate fixed effects of time of prediction and number of prior assessments on prediction error. The mean prediction error was 8.4% (±5.3%). The multivariate analyses yielded a strong significant main effect of time of prediction ($F_{3,73}$=6.97, P=0.0003), with no interaction between the two factors. In line with expectations that increasing time elapsed from the last clinical assessment would increase the prediction error, from an average of 5.9% (±4.2%) when predicting <6

month time period, to an average of 11.7% (±7.2%) at 18-24 month time prediction period (figure 3). Number of prior assessments used for the prediction did not yield a significant effect in the model ($F_{5,17}$=1.70; P=0.19), despite of a tendency for an average decrease of 7% in prediction error when ≥8 prior visits were used as compared to 3. This was also in accordance with expectations. Our results indicates that the accuracy of prediction using our model is dependent on the time elapsed between last clinical assessments and is not affected by the amount of prior data used for that prediction. It has to be noted that the significant decrease in prediction accuracy of almost 6% in average, from the shortest to the longest prediction time period used in our prediction model when validity score were removed, is still relatively small as compared to the overall accuracy of 91% (±8%).

We then attempted to cross-validate our approach by randomly partitioning the 68 patients into "training" and "testing" groups, in order to avoid overfitting of our model. Patient data from a group of 50 random patients (training) were used to obtain the new optimum parameters (B_1 for linear, B_2 for quadratic), while the remaining 18 patients (testing) were used to evaluate the predictive power of these newly derived equations using the same statistical models as presented above. This process was repeated 40 times to avoid random selection bias. The mean level of error across the 40 random shuffling was 8.6% (±1.2%) between predicted vs. actual GI. When the predicted clinical assessment was conducted within 6 months of the last visit, the prediction error was on average 6.8% (±1%) and increased to only 12.4% (±2.8%) when the prediction time was increased to 18-24 month. Nevertheless, this effect was highly significant across the 40 trials as the median P value was 0.0002 (IQR: 0.004) (figure. 4). In conformity with the previous test, the accuracy of prediction did not change when the number of prior visits was increased from 3 to ≥8, yielding a median P value of 0.38 (IQR: 0.6) across the 40 random trials. Similarly there was no interaction between the two factors.

We finally performed a holdout validation, by assessing our model against the latest clinical assessment data from patients who had come back to clinic in 2013 that were previously unavailable during model training. This holdout dataset consisted of 23 out of the original 58 patients we used to construct our model, with a total of 53 new clinical assessment data sets. Using a series of non-parametric analysis we could not observe any statistical differences between the predicted and the actual GI score in these patients (figure 5A). The median prediction error was 11.2 (IQR: 17.1). Using Spearman's coefficient we found a highly significant correlation between the predicted and actual GI score (r_s=0.91, P<0.001, figure 5B). Similar to the previous validation tests, we then sought to analyse if there were any differences in the quality of prediction with the duration between the last (up to 2012) and present (2013) clinical assessments. However, we found no significant effect of time elapsed since last clinical assessment on the accuracy of the prediction (figure 5C).

Although the size of CAG repeat length was not used to calculate the GI score, it is the major determinant of the age of disease onset, and we were therefore interested to examine if the number of CAG repeats affects the rate of disease progression (figure S4, table S4). For this, we used data from the 41 linearly progressing patients and calculated

the optimum B_1 (rate) for each of the motor and functional components of individual patients (figure S5).

CAG repeats	Motor (actual)	Functional (actual)	Motor (normalised)	Functional (normalised)
40-44	2.79	7.63	0.79	0.94
45-49	3.97	8.04	1.13	1.00
>49	6.88	11.63	1.96	1.44

Table 2 Mean values of B_1 for motor and functional indices for different ranges of CAG repeat sizes. Both the actual and normalised values are reported.

We divided patients into three subgroups according to their CAG repeat size, and noted that patients with longer CAG repeats had more rapid disease progression compared to patients with shorter CAG repeats (table 2). This is consistent with what had been reported previously.[13,14] Furthermore, when the CAG repeat size increases, UHDRS total motor score deteriorates at a quicker rate as opposed to the functional components (figure 6). We have also compared medications among different subgroups and did not observe any effects on disease progression profile (figure S6).

DISCUSSION

In the present study we describe a novel modelling approach that can be used to track, as well as to predict, HD progression in manifest patients. The primary strength of our UHDRS-based model is that it is derived from measures routinely assessed in HD clinics worldwide, and its quality has been scrutinised using four validation methods. With at least four prior clinical assessments over a minimum span of two years, we can faithfully predict HD progression for individual patients over the next two years. Patients with similar clinical profiles (age, CAG repeat length, UHDRS) can also exhibit very different profiles of disease progression, and we can model this along with providing evidence that patients with longer CAG repeat size have a quicker rate of disease progression. Further studies will however be required to determine the underlying causes for the latter two observations.

Over the past few years much effort has gone into uncovering potential biomarkers to track HD progression. For example, the level of striatal brain-derived neurotrophic factor (BDNF) was shown to be substantially reduced in HD patients.[15] Therefore the level of plasma BDNF in 398 HD patients was studied, before concluding that neither the serum level of BDNF protein nor mRNA could be reliably matched to stages of HD severity.[16] On the other hand, Weiss and colleagues[17] have demonstrated that the level of mutant Huntingtin (mtHtt) aggregation in the peripheral immune cells was significantly increased, when comparing premanifest to manifest HD patients. In addition, there was also a significant correlation between disease burden scores of individual patients with the level of mtHtt aggregation in the peripheral immune cells[17], although there was considerable intra-individual variability on the level of aggregation between samples from the same participant. Furthermore, Tabrizi and colleagues[18] have also systemically evaluated the utility of a range of biomarkers in large cohorts of patients (TRACK-HD). They demonstrated that the rate of changes in the motor and functional components of the UHDRS were associated with disease progression.[18] However, all these studies took a categorical approach by grouping patients in accordance to their disease stages for analysis. In contrast, patients in our study were tracked longitudinally as disease progressed and deteriorated, while the degree of GI changes was analysed on an individual basis. We believe that such an approach could better represent the heterogeneity of disease progression in individual patients, as we have observed in our cohort. Similar longitudinal strategies have also been employed in two very recent reports.[7,9] In the Dorsey study both the motor and functional components of UHDRS, as well as several cognitive measures, were found to consistently deteriorate in HD patients followed for three years.[9] In the Tang study, the authors used functional imaging tools to demonstrate their potential as biomarkers to track preclinical HD, as the metabolic activity of the neural network was linearly associated with disease progression in premanifest HD patients.[7]

There are though limitations to our study. Most notably all our patients were recruited from a single centre, and their generalizability remains to be demonstrated. We have however demonstrated that the age of disease onset derived from our cohort was very similar to that described by an international, multi-centre study using larger cohorts of patients.[5] Furthermore, our approach enables disease modelling and progression to be analysed on an individual basis, while the optimum coefficients can be re-defined for specific cohorts of patients. Such flexibility could facilitate the translation of our

approach to other research centres. Another problem is that all patients in our study were examined by a single clinician that removes issues to do with interrater variability, although it had previously been shown that there is a high correlation coefficient for the UHDRS total motor score between clinicians.[10] Finally to address the possibility of overfitting our model, we have cross-validated our modelling and prediction using a repeated random sub-sampling approach, as well as performing a holdout validation using data from the patients' latest clinical assessments, which took place between 2013 and were unavailable at the time of model training.

In conclusion, using HD as an example we have developed a model to track the natural history of disease progression in manifest patients. With data from the previous four clinical visits based on the conventional UHDRS assessment, we can predict disease progression that is statistically not different from the actual progression over the next 24 months. We believe that our model will be an extremely valuable tool, both in terms of enabling researchers to reassess their existing data according to patients' different types of predicted disease progression, as well as facilitating the development of novel disease modifying therapies in the future. We also believe that similar approaches can be adapted to model clinical progression of other chronic neurodegenerative disorders.

ACKNOWLEDGEMENTS

This study was supported by the Cotswold Trust, the Rosetrees Trust, donations to the Huntington's disease clinic in the John van Geest Centre for Brain Repair, and NIHR award of the Biomedical Research Centre - Cambridge University NHS Foundation Trust. This project was also supported by EPSRC through projects EP/I03210X/1 and EP/G066477/1.

We declare no competing financial interests.

REFERENCE

1. O'Keeffe GC, Michell AW, Barker RA. Biomarkers in Huntington's and Parkinson's Disease. *Ann N Y Acad Sci* 2009; 1180: 97-110.
2. Sah DW, Aronin N. Oligonucleotide therapeutic approaches for Huntington disease. *J Clin Invest* 2011; 121: 500-7.
3. Galpern WR, Corrigan-Curay J, Lang AE, *et al.* Sham neurosurgical procedures in clinical trials for neurodegenerative diseases: scientific and ethical considerations. *Lancet Neurol* 2012; 11: 643-50.
4. Pringsheim T, Wiltshire K, Day L, *et al.* The incidence and prevalence of Huntington's disease: a systematic review and meta-analysis. *Mov Disord* 2012; 27: 1083-91.
5. Langbehn DR, Brinkman RR, Falush D, *et al.* A new model for prediction of the age of onset and penetrance for Huntington's disease based on CAG length. *Clin Genet* 2004; 65: 267-77.
6. Langbehn DR, Hayden MR, Paulsen JS, *et al.* 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; 153B: 397-408.
7. Tang CC, Feigin A, Ma Y, *et al.* Metabolic network as a progression biomarker of premanifest Huntington's disease. *J Clin Invest* 2013; 123: 4076-88.
8. Begeti F, Tan AY, Cummins GA, *et al.* The Addenbrooke's Cognitive Examination-Revised accurately detects cognitive decline in Huntington's disease. *J Neurol* 2013; 260: 2777-85.
9. Dorsey ER, Beck CA, Darwin K, *et al.* Natural history of Huntington's disease. *JAMA Neurol* 2013; 70: 1520-30.
10. Huntington study group. The Unified Huntington's Disease Rating Scale: Reliability and Consistency. *Mov Disord* 1996; 11: 136-42.
11. Carignano A, Yuan Y, Dalchau N, *et al.* Understand and prediction biological network using linear system identification, in systems and synthetic biology: a systemic approach. *Springer Science*; 2012. p. 7.
12. Siesling S, van Vugt JP, Zwinderman KA, *et al.* Unified Huntington's disease rating scale: a follow up. *Mov Disord* 1998; 13: 915-9.
13. Ravina B, et al. The relationship between CAG repeat length and clinical progression in Huntington's disease. *Mov Disord* 2008; 23: 1223-7.
14. Rosenblatt A, Romer M, Constantinescu R, *et al.* Age, CAG repeat length, and clinical progression in Huntington's disease. *Mov Disord* 2012; 27: 272-6.
15. Zuccato C, Cattaneo E. Brain-derived neurotrophic factor in neurodegenerative diseases. *Nat Rev Neurol* 2009; 5: 311-22.
16. Zuccato C, Marullo M, Vitali B, *et al.* Brain-derived neurotrophic factor in patients with Huntington's disease. *PLoS One* 2011; 6: e22966.
17. Weiss A, Träger U, Wild EJ, *et al.* Mutant huntingtin fragmentation in immune cells tracks Huntington's disease progression. *J Clin Invest* 2012; 122: 3731-6.
18. Tabrizi SJ, Scahill RI, Durr A, *et al.* Biological and clinical changes in premanifest and early stage Huntington's disease in the TRACK-HD study: the 12-month longitudinal analysis. *Lancet Neurol* 2011; 10: 31-42.

FIGURE LEGENDS

Figure 1 Spaghetti plot of disease progression (top panels) and the best-fit line (bottom panels) of individual HD patients whose disease progression can be described in a linear (n=41, left panels) or quadratic (n=17, right panels) fashion.

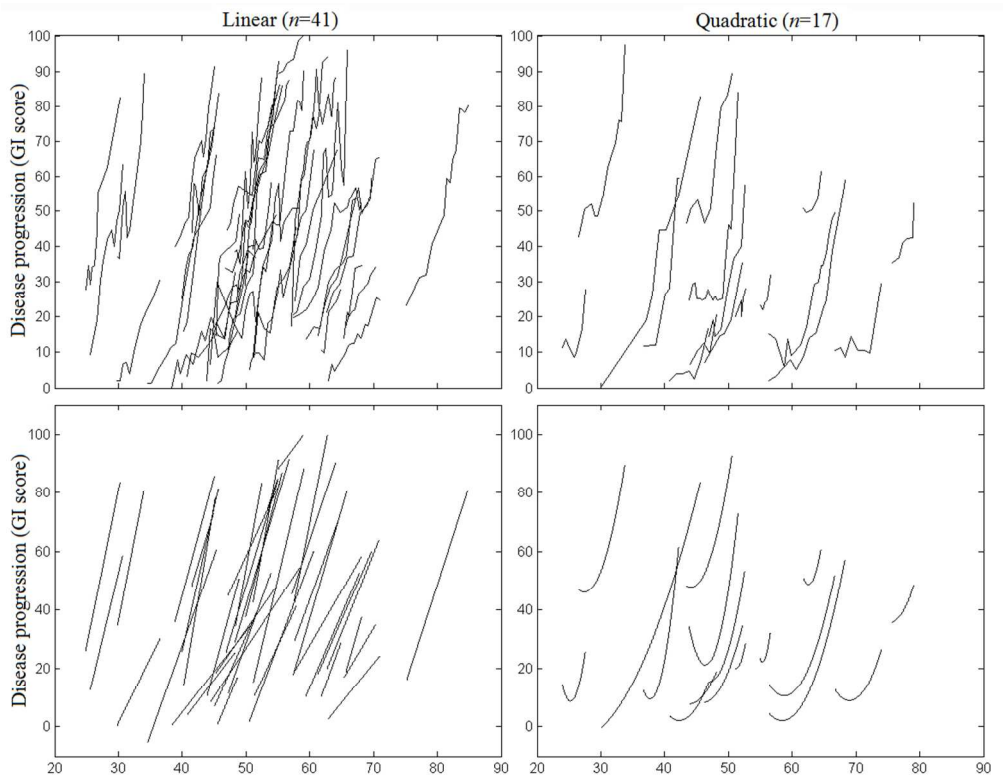
Figure 2 Comparison of the mean age of disease onset for CAG repeat lengths 40–65, between that estimated by the Langbehn *et al.*,[5] or from HD patients in our cohort exhibiting a linear disease progression (n=41).

Figure 3 The effect of factor time of prediction from the last visit used for modelling (four categories) and factor number of data sets from prior clinical visits used for modelling (six categories) on prediction error after removing the validation index. Vertical bars indicate mean prediction error calculated as percentage change between the predicted and actual general index score during the latest clinical visit. Whiskers indicate standard error. Post-hoc differences are indicated for factor time of prediction from the last visit which yielded a significant main effect. Numbers in the bottom of the vertical bars indicate number of data sets (n) within each category. Data was subjected to square root transformation prior to analysis. * p<0.05, ** p<0.01, and *** p<0.001, respectively.

Figure 4 The effect of factor time of prediction from the last visit used for modelling (four categories) and factor number of data sets from prior clinical visits used for modelling (six categories) across 40 random trials during which participants were randomly reassigned into either 'training' or 'testing' groups. Vertical bars indicate mean prediction error across the 40 trials calculated as percentage change between the predicted and actual GI score during the latest clinical visit. Whiskers indicate mean standard error across the 40 trials. Mean Post-hoc differences are indicated for factor time of prediction from the last visit which yielded a mean significant main effect. Numbers in the bottom of the vertical bars indicate number of data sets (n) within each category. * p<0.05, ** p<0.01, and *** p<0.001, respectively.

Figure 5 The association between the actual and predicted GI in the holdout validation. (A) The overall mean and standard error of actual and predicted general index. (B) The correlation between predicted and general index (Spearman Rho). Calculated (continuous) and ideal (dashed) regression lines are indicated. (C) The difference between the actual (A) and predicted (P) general index overall and for each participant and each category of time elapsed since the last clinical assessment (four categories). (D) The ratio between the predicted versus actual GI for all and for each category of time elapsed since the last clinical assessment.

Figure 6 Optimum values for B_1 (rate of disease progression) versus CAG repeats for patients exhibiting linear disease progression (n=41) for motor and functional indices. Values of B_1 for each index are normalised to the mean B_1 value.



Spaghetti plot of disease progression (top panels) and the best-fit line (bottom panels) of individual HD patients whose disease progression can be described in a linear (n=41, left panels) or quadratic (n=17, right panels) fashion.
98x77mm (300 x 300 DPI)

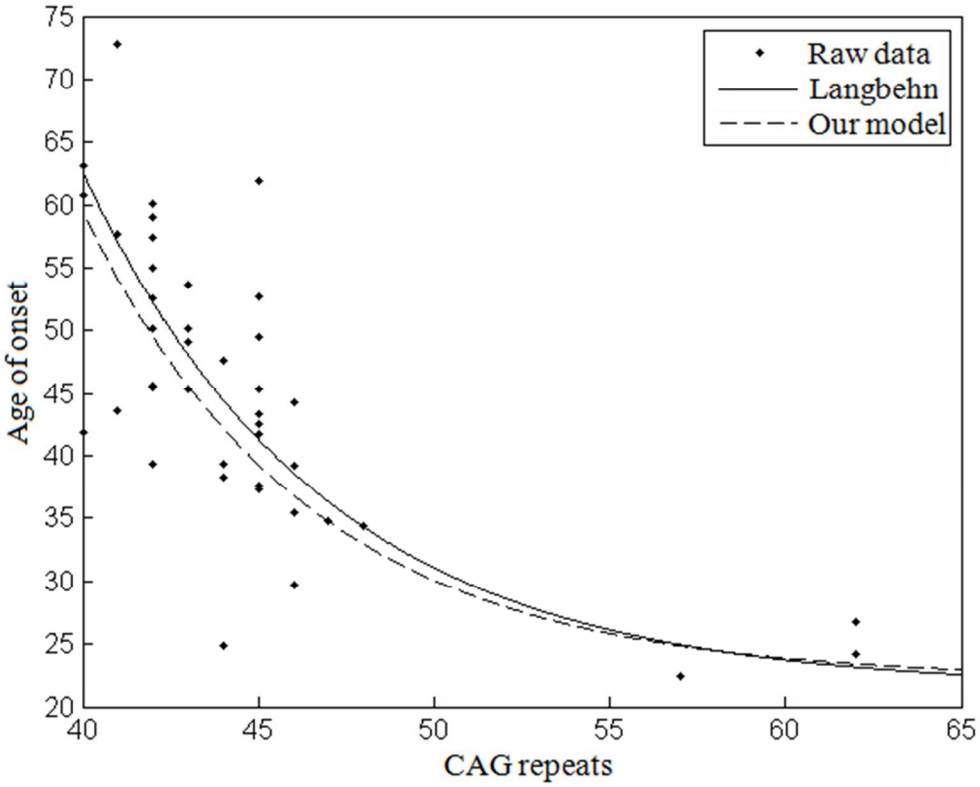
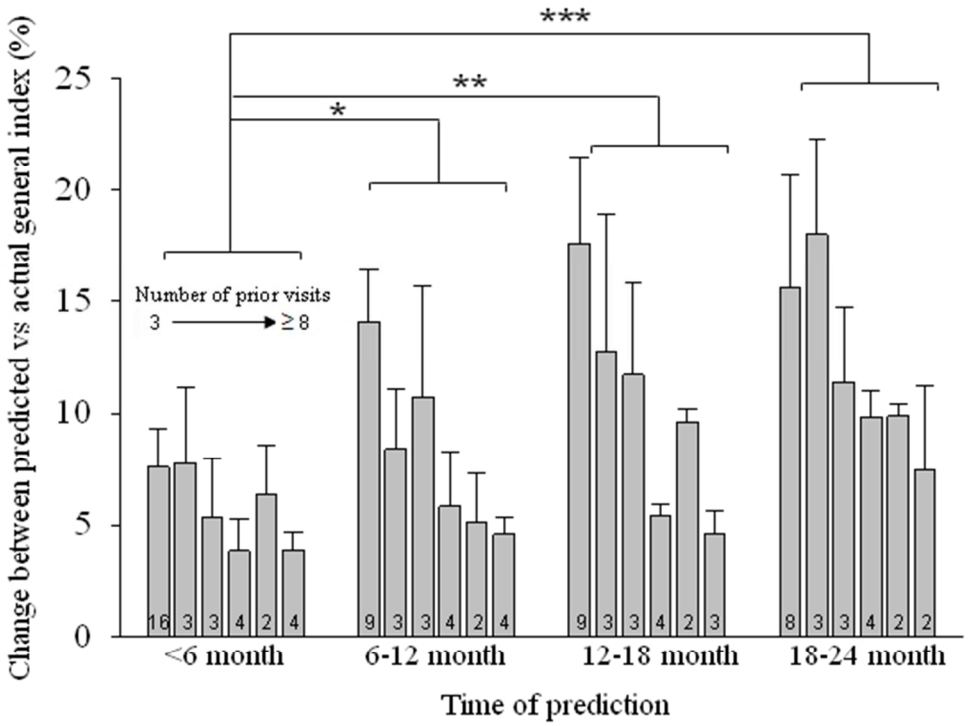
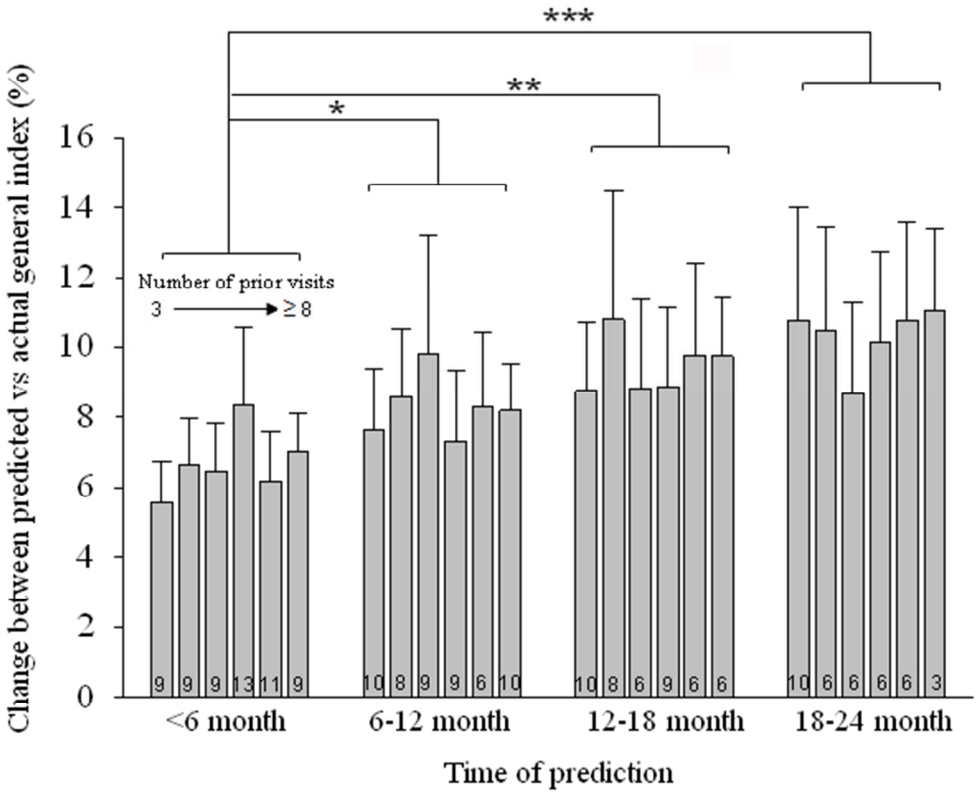


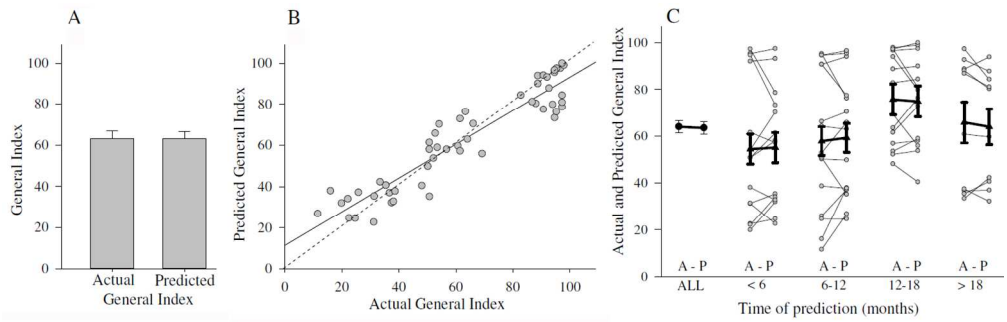
Figure 2 Comparison of the mean age of disease onset for CAG repeat lengths 40–65, between that estimated by the Langbehn et al.5, or from HD patients in our cohort exhibiting a linear disease progression (n=41).
50x40mm (300 x 300 DPI)



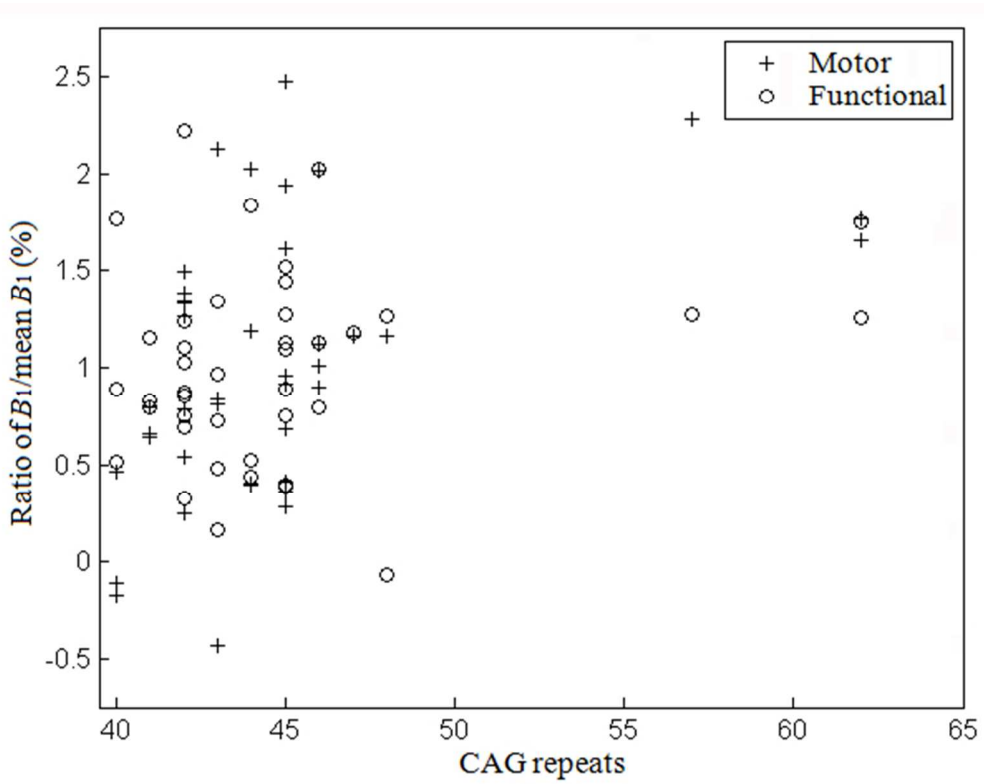
The effect of factor time of prediction from the last visit used for modelling (four categories) and factor number of data sets from prior clinical visits used for modelling (six categories) on prediction error after removing the validation index. Vertical bars indicate mean prediction error calculated as percentage change between the predicted and actual general index score during the latest clinical visit. Whiskers indicate standard error. Post-hoc differences are indicated for factor time of prediction from the last visit which yielded a significant main effect. Numbers in the bottom of the vertical bars indicate number of data sets (n) within each category. Data was subjected to square root transformation prior to analysis. * $p<0.05$, ** $p<0.01$, and *** $p<0.001$, respectively.
51x40mm (300 x 300 DPI)



The effect of factor time of prediction from the last visit used for modelling (four categories) and factor number of data sets from prior clinical visits used for modelling (six categories) across 40 random trials during which participants were randomly reassigned into either 'training' or 'testing' groups. Vertical bars indicate mean prediction error across the 40 trials calculated as percentage change between the predicted and actual GI score during the latest clinical visit. Whiskers indicate mean standard error across the 40 trials. Mean Post-hoc differences are indicated for factor time of prediction from the last visit which yielded a mean significant main effect. Numbers in the bottom of the vertical bars indicate number of data sets (n) within each category. * $p<0.05$, ** $p<0.01$, and *** $p<0.001$, respectively.
51x42mm (300 x 300 DPI)



The association between the actual and predicted GI in the holdout validation. (A) The overall mean and standard error of actual and predicted general index. (B) The correlation between predicted and general index (Spearman Rho). Calculated (continuous) and ideal (dashed) regression lines are indicated. (C) The difference between the actual (A) and predicted (P) general index overall and for each participant and each category of time elapsed since the last clinical assessment (four categories). (D) The ratio between the predicted versus actual GI for all and for each category of time elapsed since the last clinical assessment.
131x41mm (300 x 300 DPI)



Optimum values for B1 (rate of disease progression) versus CAG repeats for patients exhibiting linear disease progression (n=41) for motor and functional indices. Values of B1 for each index are normalised to the mean B1 value.
51x40mm (300 x 300 DPI)

Supplementary material:

1. Introduction

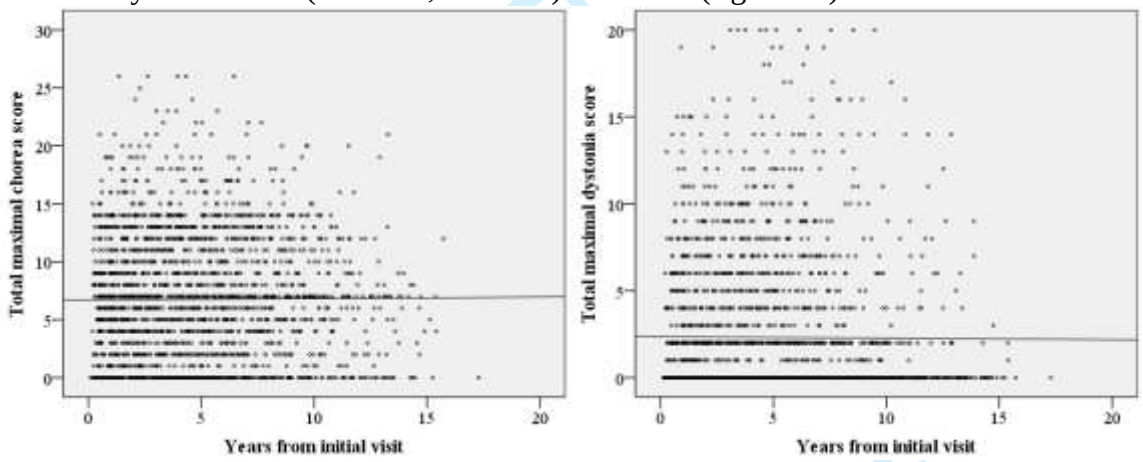
This supplemental material contains the details of the mathematical tools used for analysis of the data. It also explains the choice of model classes and validation criteria. It is divided into two parts. The first part discusses the data available and the filtering and normalisation criteria, while the second part focuses on the mathematical modelling of that data. Model validations can be found in the results section of the main manuscript.

2. Data management

For this work we used data from regular patient visits collected by Prof. Barker from 1995 until 2013 for 343 patients. For every visit of a patient, it contains all the scores that lead to the total Unified Huntington's Disease Rating Scale (UHDRS), plus patient ID, age, gender, education years, left or right hand sided, CAG minor and major.

2.1 Filtering the data - Mapping to a feature vector

To increase the robustness and reliability for training the model, some of these parameters, such as chorea and dystonia, were removed from modelling since to our experience they can show great variations in a relatively short time as well as having high inter-rater variability. Furthermore, in our data set there was no significant correlation between either the UHDRS total maximal chorea score (r=0.008, P=0.713) or the total maximal dystonia score (r=-0.008, P=0.718) over time (figure S1).



Some of the sample sets were incomplete. Hence, we could either interpolate the missing data (by using the rest of the data to estimate the missing ones) or just not use these data. The latter was chosen, as the interpolation method could introduce undesirable noise.
From the resulting data, each patient visit was reduced to a feature vector containing the information required for our analysis. The feature vector contained the following information:

Patient's ID.
Patient's age at the time the sample was taken.
The length of their CAG repeats.
The percentage of symptoms in the UHDRS motor scores at that time.
The percentage of symptoms in the UHDRS functional scores at that time.
The General Index (GI), which was the average of motor and functional indices.

This feature vector allowed the reduction of the dataset vector, while keeping all the information required for classification and prediction.

2.2 Filtering out not suitable patients

The next step was to choose an appropriate set of patients for training the model. The following three criteria were introduced to filter out patients:

The first criterion was to use patients that were beyond the prodromal stage of the disease. Quantitatively, a patient must have at least 15% of all the features at his most recent assessment to be considered for modelling. The criterion was introduced after noticing that at the prodromal stage of the disease there was not a clear trend in the data. This is probably caused by factors independent of the disease that dominate the test results, such as ageing and noise (e.g. quantisation of data scores).

Secondly, in order to maximise the validity of the model, patients were required to have a relatively large number of assessments. To make the mathematical modelling problem well-posed, it is often required that the number of samples is much larger than that of parameters in the model. However, increasing the required number of assessments to more than five resulted in having too few patients for modelling.

Difference in GI between two consecutive samples (%)	Change in validity score
≥ 5	+5
0 - 5	+1
-3 - 0	-5
$\leq$ -3	-10

Table S1 Dependence of validity scores on the change of GI between consecutive samples.

Thirdly, a validity score was created based on the assumption that a patient was not expected to improve with time and was penalising the validity of patients that showed improvements. Due to a somewhat random behaviour, patients with low validity were harder to model and, hence, were not used for modelling. The validity score was calculated as shown in table S1. For each visit, a patient's validity score was updated according to the difference between his current and previous GI scores. Patients with a validity score of less than -8 were filtered out. This boundary value was selected by observing the progression of the disease in several patients and comparing their validity score. Patients with scores below that value could not be used for modelling.

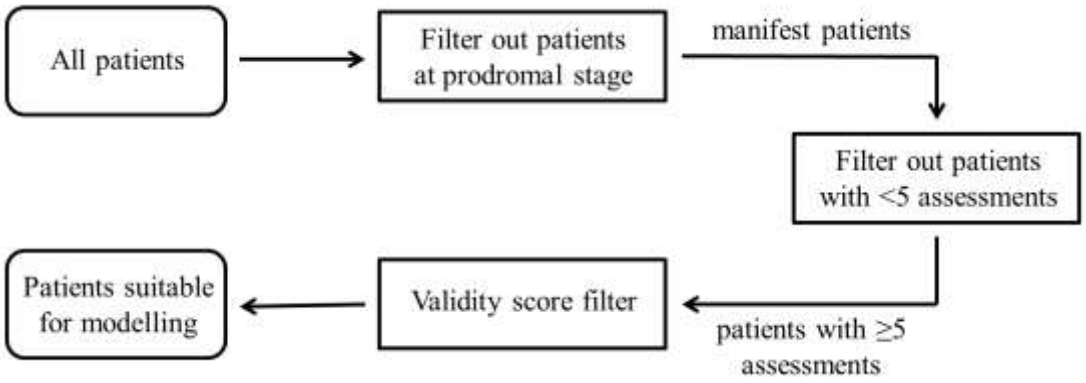


Figure S2 Schematic of the filtering procedure to obtain suitable for modelling patients.

At the end of the filtering stage there were 68 patients from the initial 343 that could be used for modelling. Note that the major reason for excluding patients was because of them not having enough assessments. The data filtering procedure is summarised in figure S2.

3. Modelling, classification and prediction

This section is dedicated to the modelling aspects of the project. It is divided into two subsections that provide the necessary details to repeat our analysis. The first part describes how the data was used for the creation of a model. The second part explains different methods to validate the models.

3.1 Modelling procedure

3.1.1 Fitness function

To understand the quality of a model, we need a way to quantify its performance. Adopted from standard system identification textbooks[1], in this paper we use the following measure of how well a curve fits the data

$$\text{fitness} = 1 - \frac{\sum_{k=1}^{N}(\hat{y}_k - y_k)^2}{\sum_{k=1}^{N}(y_k - \bar{y})^2} \qquad (3.1)$$

where $\hat{y}_k$ is the value of the index as predicted from the model, y_k is the real value and $\bar{y}$ is the mean value of all the data we have in that sample. Note that we obtain 100% fitness for a perfect fit. In a sense, it quantifies how well a certain model fits the data compared with a constant model that always equals the mean of these data. In this paper, we considered that models with a fitness higher than 70% described reasonably well the data.

3.1.2 Fitting the optimum function for each patient

To understand the classes required for modelling, we started by modelling individual patients, instead of searching for one of a few models that explained all the patients. Hence, we isolated each patient's data and tried to find a model that would best describe the patient's GI growth. The candidate models were linear, quadratic and exponential and implemented as follows:

1. The linear model assumed the General index (GI) was proportional to age. This was implemented using the following equation

$$GI = B_0 + B_1 * Age \quad (3.2)$$

where, for each patient, the values for B_0 and B_1 minimum mean squared error and maximum fitness. Rewriting the above equation as $Ax = B$, with appropriate choices for A, B and the unknowns x, the problem reduces to a standard least squares minimisation problem which has a solution given by

$$x = (A^T A)^{-1} (A^T B) \quad (3.3)$$

In our case, x is a vector containing the optimum values for B_0 and B_1, A is an N x 2 matrix, where N is the amount of samples obtained from a specific patient. The left column of A contains only the number 1 while the right column contains the patient's age at the time when the sample was taken. Matrix B is of size N x 1 and contains the GI index's value for each sample.

2. The quadratic model assumed that the GI index varied with the square of the age and is given by the following equation

$$GI = B_0 + B_1 * Age + B_2 * Age^2 \quad (3.4)$$

As in the linear model's case, we estimated the minimum mean squared error values for B_0, B_1, and B_2 using equation 3.3. In this case, x contained these three values.

3. Finally, the exponential model assumed that the GI index depended on the exponential of the patient's age and is given by

$$GI = \exp[B_0 + B_1 * Age] \quad (3.5)$$

Again, the optimum coefficients were obtained from equation 3.3.

Patient ID	Fitness			Optimum parameters						
	Linear model	Quadratic model	Exponential model	Linear		Quadratic			Exponential	
				B_1	B_0	B_2	B_1	B_0	B_1	B_0
4	0.948	0.960	0.897	8.65	-394	-0.62	74	-2102	0.155	0.0168
15	0.956	0.975	0.970	6.76	-492	0.35	-50	1790	0.139	0.0007
16	0.973	0.983	0.981	5.41	-319	0.24	-27	858	0.134	0.0053
51	0.941	0.976	0.972	7.25	-314	0.63	-57	1323	0.130	0.0691
58	0.798	0.798	0.755	7.45	-410	-0.05	14	-599	0.167	0.0015
62	0.897	0.914	0.841	5.99	-360	-1.11	147	4819	0.344	0.0000
79	0.869	0.946	0.780	7.61	-481	-2.90	395	-13451	0.297	0.0000

Table S2 Table with fitness and optimum coefficients for the three candidate models for seven patients.

Hence, for each patient we obtained the fitness for each of the three models together with the optimum coefficients that produced it. Table S2 shows a sample of seven out of the 68 patients.

3.1.3 Classification of patients to model classes

After obtaining optimum values for the parameters of the three candidate models, the fitness of each model was calculated for every patient, using equation 3.1. We then observed the following remarks:

- The fitness of quadratic models was always better than linear ones. This was expected, since the extra parameter made the quadratic model more flexible.
- The fitness of the exponential model was always worse than the quadratic. Hence, we did not consider this model any further.

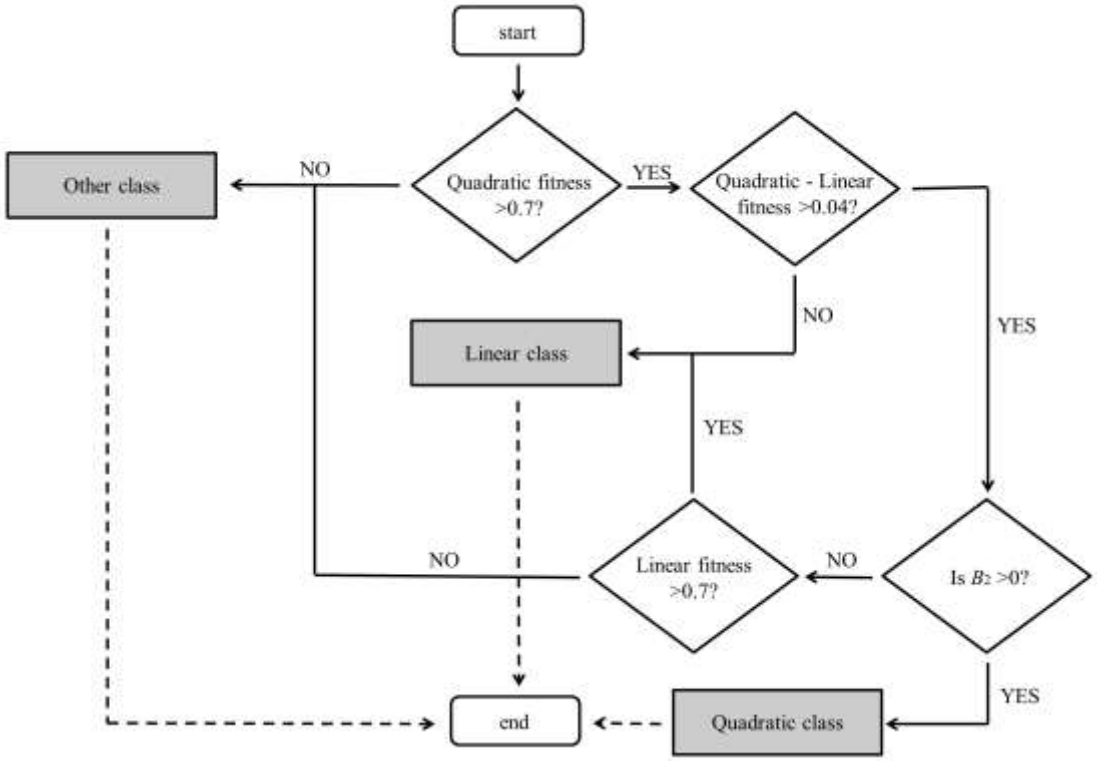


Figure S3 Flow chart of the classification procedure.

The following criteria were used to choose which model (linear, quadratic or others) was more appropriate for each patient. The first step was to check if either model could capture the data of a particular patient. If not, the patient was classified as "Others". It was decided that the acceptable fitness for this was 70%. Hence, a patient with fitness less than 70% on either model was classified as "Others". For patients with fitness higher than 70% on at least one model, the second step was to choose the model that best

describes the patient. Since quadratic fitness was always better than linear, we decided that quadratic class was chosen only if its fitness was at least 4% greater than the linear model's fitness. This is to avoid the likely overfitting, i.e. if the difference was lower than 4%, the extra fitness did not justify the extra model complexity. Hence, in this case, the linear model was selected. The third and last step was to address an issue with quadratic models. In some patients with a high fitness in a quadratic class but a negative B_2 (recall that the quadratic class was given by $GI=B_0+B_1*Age+B_2*Age^2$). This predicted that the disease would progress at a slower rate for later stages and eventually the patient would improve. Hence, it would most likely result in bad predictions. Therefore, patients were excluded from the Quadratic class when they had a negative B_2. The classification algorithm is demonstrated schematically in figure S3.

3.1.4 Aggregate optimal coefficients

Class	Linear		Quadratic		
Parameter	B_1	B_0	B_2	B_1	B_0
Minimum	2.67	-588	0.21	-732	132
Maximum	12.95	-72	6.62	-10	2030

Table S3 Range of coefficients for each patient's optimum equation.

Linear model: for the linear models, the previous part showed that its coefficients exhibited high variability (see table S3 for the range of these values). Hence, an aggregate single model resulted in a poor description of the whole linear class so we decided to find two subclasses for the linear model class. In the linear models, the parameter B_0 is very dependent on each patient and is related to the age of onset of the disease. Hence, this parameter needs to be calculated for each patient. The most important parameter is, in fact, B_1 since this captures the rate at which the disease progresses. Hence, this is the parameter that aggregated in two subgroups. For a given value of B_1, we calculated the optimum value of B_0 using a similar expression to 3.3, where A is a N-sized vector of ones (N is the number of samples for each patient) and B is an N-sized vector consisting of the GI index's value minus the value of B_1 times the age when the sample was taken. The value of x gave the optimum value for B_0 for that patient.
We used a greedy algorithm that computed all possible combinations to simultaneously evaluate two linear models with different values for B_1 between the range 2.67 to 12.95 (table S3). This resulted in the globally optimum solution for two linear subclasses. The solution was one class with B_1 = 6.86, fitting 29 patients well with at least 70% fitness, and a second class with B_1 = 3.30, fitting well 11 patients. Hence, the linear class was divided into two subclasses: Linear A with B_1 = 6.86 and Linear B with B_1 = 3.30. Patients were classified into either Linear A or Linear B, by choosing the corresponding highest fitness (always required being above 70%). The dividing point between both Linear classes was 4.92, which was chosen to maximise the margin between the two classes. It was calculated as the average of the maximum B_1 coefficient between the patients of the Linear B class and the minimum B_1 coefficient between the patients in Linear A. We also tried dividing the Linear group into more than two classes. However, although the overall fitness increased, the data was over fitted and, in some circumstances, resulted in poor predictions.

Quadratic model: as above, the most important parameter for the quadratic model is B_2, which ranges between 0.21 to 6.62 (table S3). The other two parameters, B_0 and B_1, were very dependent on the individual patient. For each value of B_2, to obtain the optimum coefficients B_0 and B_1, equation 3.3 was again used with x having the values of B_0 and B_1, vector A containing a column of ones and a column of ages and vector B containing the GI index values minus B_2 times the age squared. We then optimised over B_2, with the best value of 0.64, and giving at least 70% fitness for 16 out of 17 quadratic patients.

3.2 Prediction stage

With models in place, the next step was to check how good the models could predict future data. To make a prediction, the first N samples of each patient were used to classify the patient and to create a model that would describe the progression of the disease. This model was used to make predictions for up to two years (if the amount of samples allowed it). The metric used to evaluate the results was the standard error between our predicted values and the real ones. This section explains how to classify a patient given a number of visits and the criteria that make a prediction trustworthy.

3.2.1 Classification

Before making any predictions, we first need to classify each patient in one of the groups Linear A, Linear B, Quadratic or Others using only the first N samples. The classification method was similar to the one in the initial classification (when all the data was available). To differentiate patients between Linear A and Linear B classes, patients were in Linear A if their optimum coefficient for B_1 was greater than 4.92 and Linear B otherwise.

An equation for each patient was then created as follows. Depending on the model class of the patient determined from the classification stage, the most significant coefficient was fixed to one of the three values: $B_1 = 6.86$ for Linear A, $B_1 = 3.30$ for Linear B and $B_2 = 0.64$ for Quadratic group (with B_1 as defined in equation 3.2 and B_2 as in equation 3.4). The remaining coefficients were found using the method explained in section 3.1.2.

3.2.2 Developing and evaluating the predictions

The method was evaluated by calculating the proportion of correctly classified patients when only their first N data points were used. We made predictions on all patients except those where:

- Only N points were available, since in this case no future data was available to evaluate the prediction,
- And the patient was classified in the group Others, since in this case the prediction could not be trusted.

Predictions were made for each patient for up to four future time steps. Each time step represented a period of six months (hence, the fourth step prediction was for 18 - 24 months).

4. Additional Findings from our Analysis

4.1 Rate of change for Linear class with CAG length.

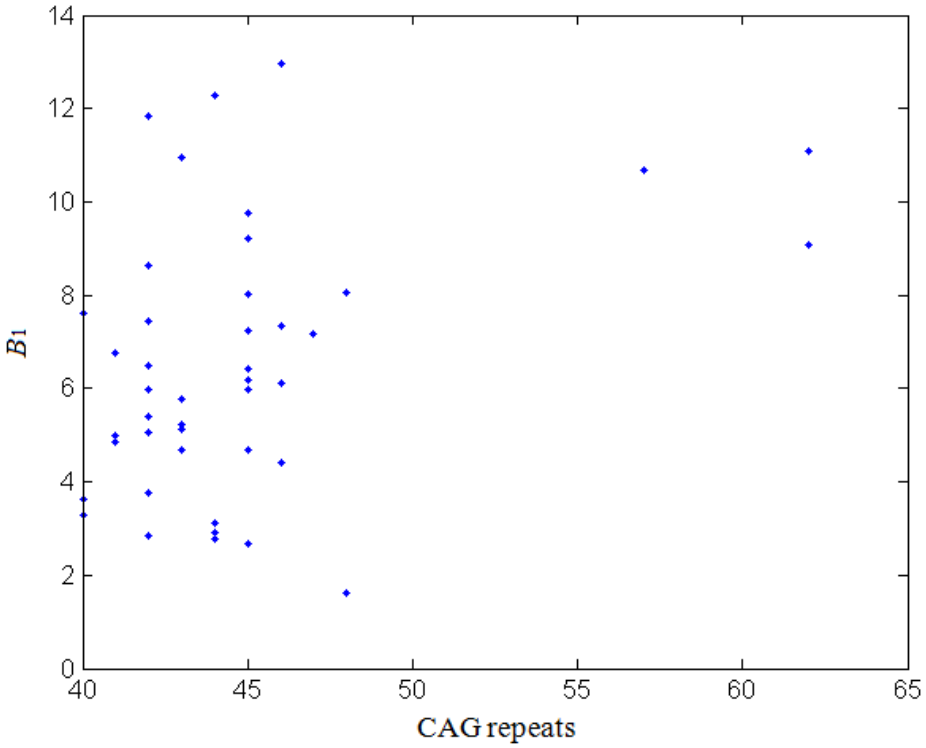


Figure S4 Optimum values for B_1 versus CAG repeats for the 41 patients of the Linear class.

After obtaining the optimum coefficients for each patient (as explained in section 3.1.2) and classifying our patients (section 3.1.3), we attempted to create an equation that would predict B_1 value's for the linear model using only the length of patient's CAG repeats. The values for optimum B_1 for each patient classified in the Linear class (either Linear A or Linear B) with their corresponding lengths of CAG repeats are shown on figure S4. From this figure, given the variability of B_1 for given CAG repeats, it can be deduced that a prediction of B_1 based on CAG repeats alone is not possible.

CAG repeats	Linear (n=41)	Linear A (n=29)	Linear B (n=11)
40-44	6.05	7.44	3.38
45-49	6.74	7.87	3.93
>49	10.28	10.28	No data

Table S4 Mean values of B_1 for certain ranges of CAG repeats for classes Linear, Linear A and Linear B.

The correlation of B_1 values with CAG repeats was found to be 0.42 using Spearman's test. The P value to reject the hypothesis that B_1 and CAG repeats are correlated was less than 1%. Hence, from this analysis, it can be concluded that CAG repeats is one factor

that affects the value of B_1 but it is not the only one. As table S5 suggests, a higher value of B_1 is expected when a larger CAG repeats' length is encountered.

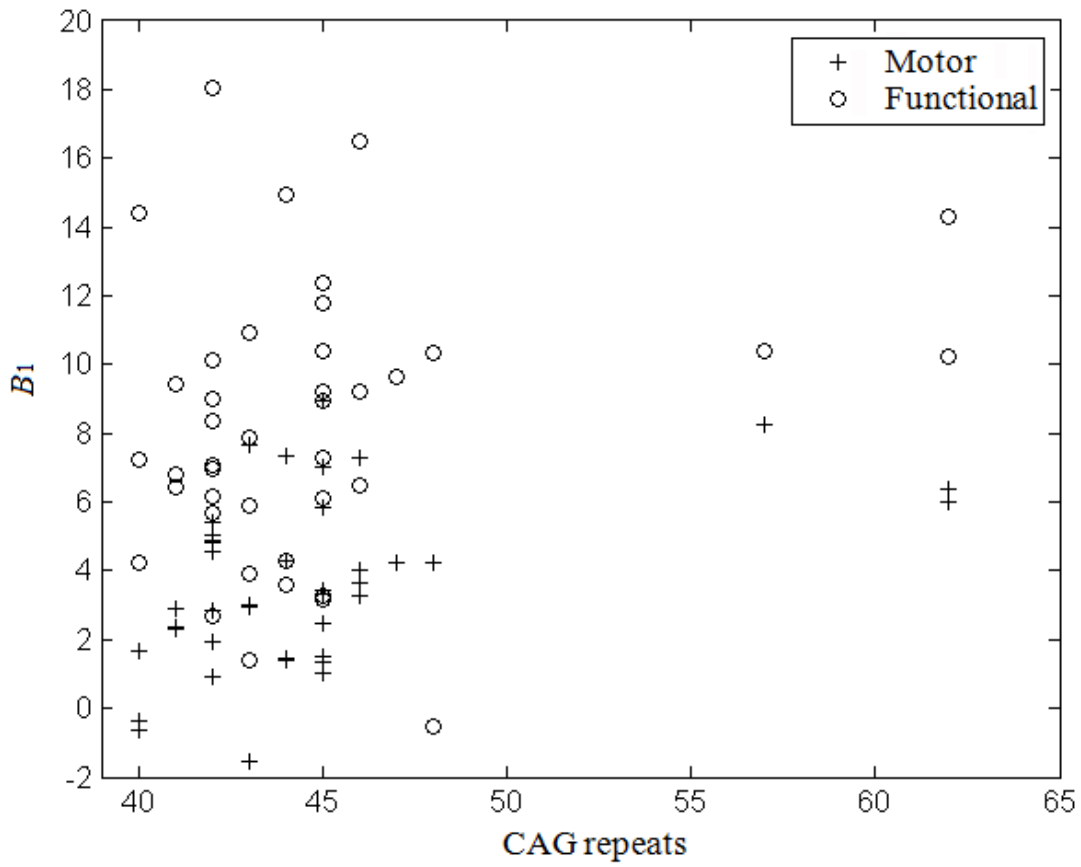


Figure S5 Optimum values for B_1 versus CAG repeats for the 41 patients in the Linear class for Motor and Functional indices.

4.2 Analysis of motor and functional symptoms

Thus far, only the overall GI was considered, obtained as the mean of motor and functional indices. Next we investigated how the rate of change of progression of the disease differed between the two indices. For this, we used only patients in the Linear group and calculated the best fit line for each index and patient. The method to obtain the coefficients of this line was described in section 3.1.2. The values of B_0 and B_1 that minimised the mean square error between the data and equation GI = B_0+B_1*Age were then calculated for each index. The optimum coefficient was again B_1, which described the rate of increase of each index (motor or functional). The results are summarised in figure S5, which shows that, in general, the optimum B_1 for functional indices were larger than the corresponding ones for motor index. The mean value of B_1 for motor index was 3.62 and for functional index 8.16. We then investigated the following:

- Whether the expected value of B_1 increases with an increase in the CAG repeats.
- Whether the relative increase of B_1 between low and high CAG repeats for the

To ease the comparison, the values of B_1 for the two indices were normalised by their averages. This is shown on figure 6 of the main manuscript, where both indices have means of 1.0. To get the trends for the two indices, we obtained the average value of B_1 for each index in different ranges of CAG repeats. The results are shown in table 2 of the main manuscript. This table reveals that the expected value of B_1 for each index increases with the length of CAG repeats. Columns four and five, obtained by dividing columns two and three by the average of all values for each index, show that the expected value of the motor index for B_1 becomes almost twice the average when a longer CAG repeat is encountered. If this value is compared with the corresponding one for low CAG lengths, the ratio becomes almost 250% (1.96 over 0.79). For the functional index, that ratio is close to 150%. From these results, we conclude that an increase in CAG repeats has a much stronger effect in motor progression than in functional deterioration.

4.3 No difference of medications between patient subgroups

In order to compare whether there was any difference in medications between subgroups of patients (Table 1), the number of times each patient was taking medications was divided by his/her number of visits. Hence for each patient, a value between 0 and 1 for each medication was obtained. This value was the proportion of visits that the patient received that particular medication. The metric used for each medication was the average of this value for all patients. We could not observe any difference between medications taken among different subgroups of patients (figure S6). Furthermore we did not find any evidence how medications taken affected the rate of disease progression.

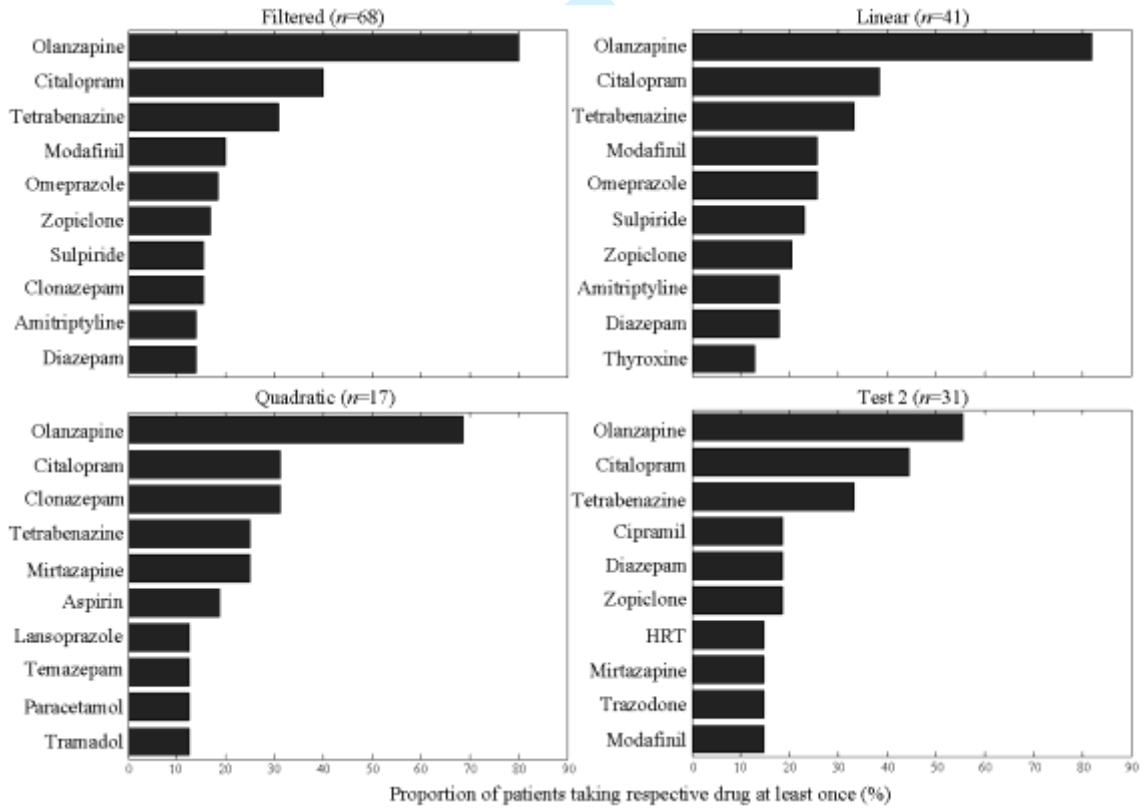


Figure S6 Top 10 medications being taken by patients among different subgroups.

REFERENCE

1. Ljung L. System identification (2nd edition): theory for the user. *Upper Saddle River, NJ, USA*; Prentice Hall PTR; 1999.