Clin Genet 2004: 65: 267–277
Printed in Denmark. All rights reserved

CLINICAL GENETICS
doi: 10.1111/j.1399-0004.2004.00241.x

Original Article

A new model for prediction of the age of onset and penetrance for Huntington's disease based on CAG length

Langbehn DR, Brinkman RR, Falush D, Paulsen JS, Hayden MR on behalf of an International Huntington's Disease Collaborative Group. A new model for prediction of the age of onset and penetrance for Huntington's disease based on CAG length.
Clin Genet 2004: 65: 267–277. © Blackwell Munksgaard, 2004

Huntington's disease (HD) is a neurodegenerative disorder caused by an unstable CAG repeat. For patients at risk, participating in predictive testing and learning of having CAG expansion, a major unanswered question shifts from "Will I get HD?" to "When will it manifest?" Using the largest cohort of HD patients analyzed to date (2913 individuals from 40 centers worldwide), we developed a parametric survival model based on CAG repeat length to predict the probability of neurological disease onset (based on motor neurological symptoms rather than psychiatric onset) at different ages for individual patients. We provide estimated probabilities of onset associated with CAG repeats between 36 and 56 for individuals of any age with narrow confidence intervals. For example, our model predicts a 91% chance that a 40-year-old individual with 42 repeats will have onset by the age of 65, with a 95% confidence interval from 90 to 93%. This model also defines the variability in HD onset that is not attributable to CAG length and provides information concerning CAG-related penetrance rates.

DR Langbehn[a,b], RR Brinkman[c], D Falush[d]*, JS Paulsen[a,e] and MR Hayden on behalf of an International Huntington's Disease Collaborative Group[c]†

[a]Department of Psychiatry, and [b]Department of Biostatistics, University of Iowa College of Medicine, Iowa City, IA, USA; [c]Department of Medical Genetics, Center for Molecular Medicine and Therapeutics, University of British Columbia, Vancouver, BC, Canada; [d]Mathematical Genetics Group, Department of Statistics, Oxford University, Oxford, Great Britain; [e]Department of Neurology, University of Iowa College of Medicine, Iowa City, IA, USA

Key words: genetics – Huntington's Disease – onset age – statistical modeling – trinucleotide repeat

Corresponding author: Michael R. Hayden, Department of Medical Genetics, Center for Molecular Medicine and Therapeutics, University of British Columbia, Vancouver, BC, Canada.
Tel.: +1 604 875 3535;
fax: +1 604 875 3819;
e-mail: mrh@cmmt.ubc.ca

Received 15 December 2003,
revised and accepted for publication
9 January 2004

Huntington's disease (HD, MIM 143100) is a progressive, neurodegenerative disorder that presents with motor disturbances, psychiatric symptoms, and cognitive decline. HD has been reported worldwide with an overall prevalence of about 10 per 100,000 in Caucasian populations (1, 2). The mutation associated with clinical manifestations of HD is a CAG trinucleotide repeat expansion in the HD gene (3). Persons affected with HD have a CAG repeat length (CAG) between 36 and 250, though some individuals with a CAG less than 42 will never show symptoms (4–8).

Numerous studies have described a significant inverse relationship between CAG and the age of onset (9–17), with CAG length accounting for up to 73% of the variation in the age of onset (4, 9–11). However, these studies have been of limited clinical use in predicting the mean age of onset for a particular CAG, because the precision of the predictions is relatively low, with 95% confidence limits up to 20% (4).

Using the largest worldwide cohort of both affected and at-risk individuals to date, we have

* Dr D Falush was affiliated with the Max-Planck Institut für Infektionsbiologie, Berlin, Germany during the conduct of this Research.

† For a complete list of participants, see *Appendix*.

now developed a parametric survival model that significantly reduces the average confidence limits of the predictions. Almost all of the previous reports (4–19) concerning age of onset and HD have been based only on onset ages for CAG-expanded individuals with manifest disease, leading to possible bias from ignoring age distributions among those who have not developed the illness (4). Our new model, incorporating information from those with onset and those still at risk, may be useful for predicting risk of onset for any person at risk of HD at any age. It also provides insight into lifetime penetrance at various CAG lengths.

Methods

Forty centers (*Appendix*) contributed anonymous data to this study, including centers in Europe (7), Asia (1), Africa (2), and North America (30). We obtained ethical approval for the study from the University of British Columbia, as did the centers supplying and analyzing the data from their local governing bodies. CAG length and age information was available for 3452 subjects, all with verified CAG expansions of at least 36. Our final model is based on a subset of 2913 of these subjects with CAG lengths between 41 and 56. Of these, 2298 had already experienced HD onset and 615 had not. For shorter repeats, bias due to underascertainment of asymptomatic individuals seemed likely to influence the results. We emphasize that all estimates presented later for this repeat range are extrapolations from the 41–56 range. For repeat lengths greater than 56, there was not enough data to assure stable estimates of the CAG-specific curves.

To minimize differences in CAG determination, only those individuals who had their CAG determined exclusive of the adjacent polymorphic CCGn stretch, using cloned standards for accurate sizing, were included (5, 19). Age at onset was defined as the first-time neurological signs representing a permanent change from the normal state was identified in a patient. The age used in the analysis of presymptomatic individuals was the oldest age when a physician last directly confirmed their clinical status.

Our model is based on the general theory of parametric survival analysis (20). This has several advantages: (1) Estimates are not based solely on those who have already experienced HD onset, but also on those without onset, as their age of observation. This reduces potential bias from considering only those coming to attention because of symptoms – possibly at an atypically early age; (2) Mathematical formulations of the relationships between CAG length and both the mean and variability in the age of onset are derived; (3) Relative to analyzing CAG lengths at one time, more efficient use of the data is possible; and (4) This allows for much narrower confidence intervals. Additional quantities of practical interest can be computed. One example is the conditional expected onset age, given that a person has reached his current age without developing HD. Initially, we did analyze the data one CAG length at a time to determine a family of probability distributions that consistently described age of onset across all lengths. This also allowed an approximate plot of the CAG relationships with the mean and variability of the age of onset and provided strong clues about potential ascertainment bias for short CAG lengths. This guided formulation of our overall model.

The first step in building our parametric model was finding a distribution family that gave a close fit to all of the observed (non-parametric) survival distributions from the individual CAG repeats. We examined 12 families of parametric survival distributions. Goodness-of-fit was checked by both visual inspection of quantile–quantile plots (20) and comparison of the likelihoods of each of the parametric models summed across CAG lengths. The plausibility of a Cox proportional hazards model (20) was tested similarly and rejected.

The second step consisted of finding classes of mathematical functions that adequately described the relationship between CAG and both the mean and dispersion of the age of onset. These functions and the parametric distribution selected in the first step were then combined to form our final parametric model.

We obtained maximum likelihood estimates (21) for the final model parameters via general purpose numerical optimization algorithms available in the programs MATLAB (22) and MATHEMATICA (23). We also derived the symbolic information matrix for the model parameters in MATHEMATICA and were then able to apply the delta method (25) to obtain approximate standard deviations and confidence intervals for quantities of practical interest, such as age-of-onset means.

Given that an individual is presymptomatic at age x, the conditional probability of onset between ages x and $(x+t)$ was calculated as $\{1–[S(x+t, CAG)/S(x, CAG)]\}$, where $S(x)$ is the probability of still being presymptomatic at age x (the survival function) (24,25). This follows from the definition of conditional probability. The form of the survival function will be given in the '*Results*' section.

Table 1. Distribution of affected and presymptomatic individuals at risk for Huntington's disease

Risk status		Numbers and ages for a CAG repeat of																																			
		36	37	38	39	40	41	42	43	44	45	46	47	48	49	50	51	52	53	54	55	56	57–60	61–65	66–75	76–250	Total
At risk	Number of individuals	16	14	21	53	93	90	121	100	103	62	45	37	23	12	7	5	5	1	2	1	1	3	2	1	0	818
	Average age at last examination	53	44	36	46	44	42	38	36	32	30	28	31	30	25	23	26	26	24	22	18	16	22	20	19		
	Minimum age at last examination	9	5	10	12	14	20	18	19	13	14	16	11	20	19	7	21	16	24	21	18	16	20	17	19		
	Maximum age at last examination	85	90	74	80	80	71	67	59	48	43	44	47	39	34	34	32	33	24	23	18	16	23	22	19		
	SD of age at last examination	21	25	15	17	13	11	10	10	7	6	6	8	6	5	9	5	8	NA	1	NA	NA	2	4	NA		
	% of presymptomatic individuals	2	2	3	6	11	11	15	12	13	8	6	5	3	1	1	1	1	0	0	0	0	0	0	0	0	
Affected	Number of individuals	3	10	13	53	134	242	314	349	299	277	205	147	125	95	65	39	41	38	20	20	22	46	35	22	20	2634
	Average age of onset	66	55	64	58	58	57	51	47	43	42	39	35	33	33	31	29	29	27	25	28	24	22	19	14	8	
	Minimum age of onset	59	40	35	35	31	25	25	17	20	18	20	19	17	13	12	18	13	16	17	18	16	13	12	14	2	
	Maximum age of onset	75	77	85	81	84	83	80	73	66	69	63	58	50	52	46	43	52	35	34	54	36	35	29	24	15	
	SD of the age of onset	8	14	17	11	11	11	9	8	8	7	7	7	6	7	7	6	7	5	6	8	4	4	5	4	4	
	% of affected individuals	<1	<1	<1	2	5	9	12	13	11	11	8	6	5	4	2	1	2	1	1	1	1	2	1	1	1	
Total	Number of individuals	19	24	34	106	227	332	435	449	402	339	250	184	148	107	72	44	46	39	22	21	23	49	37	23	20	3452
	% of total data	1	1	1	3	7	10	13	13	12	10	7	5	4	3	2	1	1	1	1	1	1	1	1	1	1	

Results

Subjects

Individuals were drawn from a cohort from 40 centers (*Appendix*). The distribution of affected and presymptomatic at-risk individuals with repeats greater than 35 is summarized in Table 1. A total of 3452 individuals (2634 affected and 818 presymptomatic) met this CAG criterion. There were no affected individuals with less than 36 repeats in any center.

The contributing centers differed considerably in their proportion of presymptomatic patients (*Appendix*). We divided the centers into three groups based on the proportion of unaffected subjects ascertained and found that the groups gave systematically different estimates for the mean and variance in age at onset for repeat lengths less than 41. However, they gave similar estimates for larger repeat lengths (analyses not presented). Consistent with a more detailed analysis of the potential effect of underascertaining unaffected individuals with a smaller CAG (not presented), these results imply that our data are generally representative for a CAG 41 or greater but may be increasingly biased by incomplete ascertainment for shorter repeat lengths. For the development of the parametric model, we therefore used only the 2913 individuals (84% of the sample) who had a CAG between 41 and 56. Model-based results for shorter CAG lengths are an extrapolation beyond this range.

Parametric model

The logistic distribution had the best average fit to the non-parametric survival curves across CAG lengths. Furthermore, the logistic fit was consistently good for each CAG length over 41, suggesting a common shape of the survival distributions. The curves failed the assumption of a proportional hazard relationship ($p < 0.0001$), a prerequisite for applying the familiar Cox model of survival analysis.

The mean and variance of the age of onset for each repeat length, as estimated one CAG length at a time under the logistic model, both suggested regular curvilinear relationships with CAG length (Fig. 1). We explored relatively simple mathematical functions that described these and found that exponential functions of the form $y = a + \exp[b–c\ (CAG)]$ provided excellent fits to both the mean and variance of the age of onset.

Using the functional relationships between CAG and mean and variance of onset described earlier and the subset of data with lengths

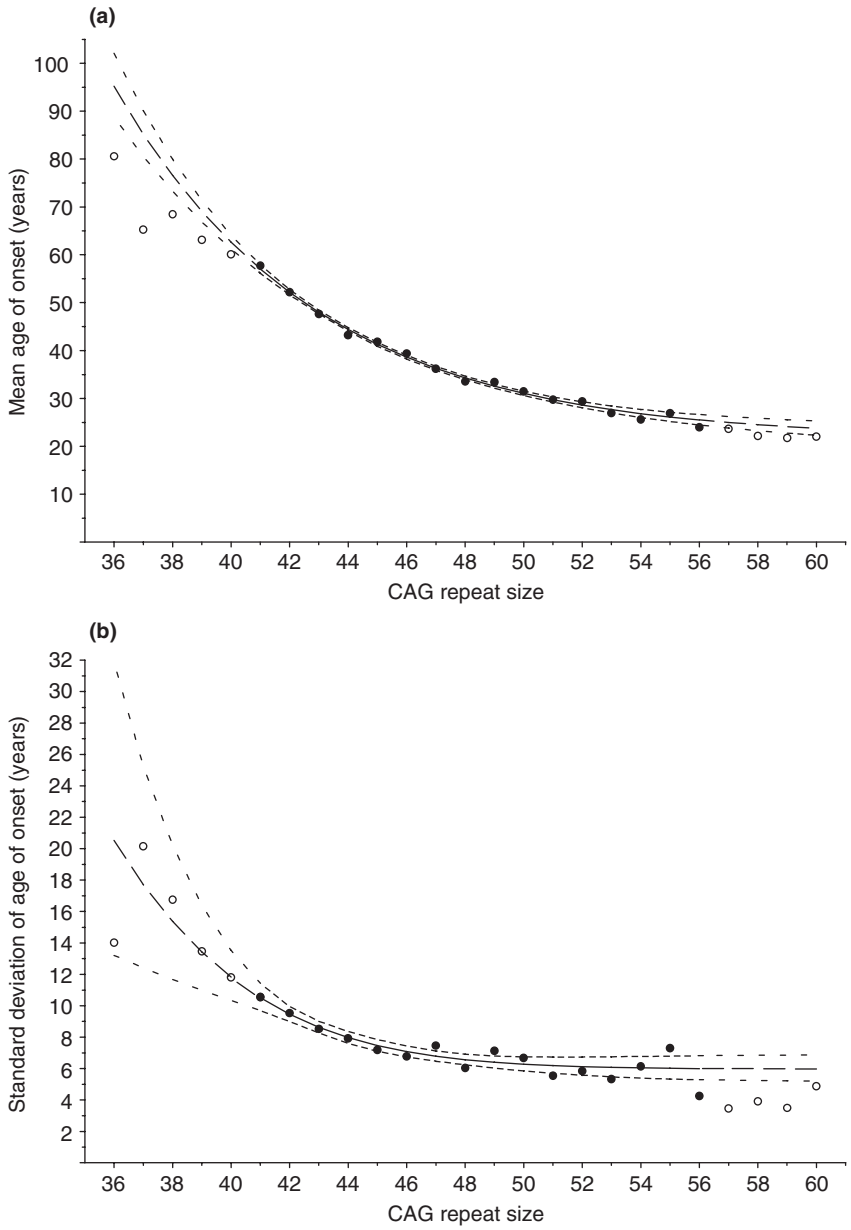

Fig. 1. Population estimates of the mean age of onset (a) and standard deviation of the age of onset (b) for CAG repeat lengths 36–60. The • symbols and solid line indicate the range of data that was used to fit the exponential curves. The O symbols and long dashed lines indicate CAG lengths for which the model's predictions were extrapolated. Small dashed lines indicate 95% confidence intervals, larger spaces between dashes indicate the region where the model's predictions were extrapolated.

between 41 and 56, our estimated parametric survival model for HD onset is

$$S(Age, CAG) = \left(1+\exp\left\{\frac{\pi}{\sqrt{3}}\,\frac{[-21.54-\exp(9.56-0.146CAG)+Age]}{\sqrt{35.55+\exp(17.72-0.327CAG)}}\right\}\right)^{-1},$$

where S(*Age*) is the probability of surviving without neurological symptoms until at least the given age.

The very close fits between the non-parametric survival curves and the above equation for CAGs of 41, 45, 49, and 53 are shown in Fig. 2. We found similar agreement for the entire range of repeats. The confidence intervals are much narrower for the parametric model, because the entire data set is used to estimate each parameter. This is evident when the widths of the parametric and non-parametric confidence intervals are compared in Fig. 2. The approximate p-values of each of the six terms estimated within the model were highly significant (minimum chi-square = 13.819, 1 d.f., p = 0.0002). Finally, the goodness-of-fit of our model was satisfactory when compared with a saturated model with a separate logistic distributions fit for each CAG (chi square = 35.165, 26 d.f., p = 0.108). When we re-estimated our model using only 80% of the data, there was no notable overfitting evident when the predictions were compared to the observed onsets in the other 20% "holdout" sample (results not presented).

Under the parametric model, the variance in the age of onset is larger for shorter repeats, as

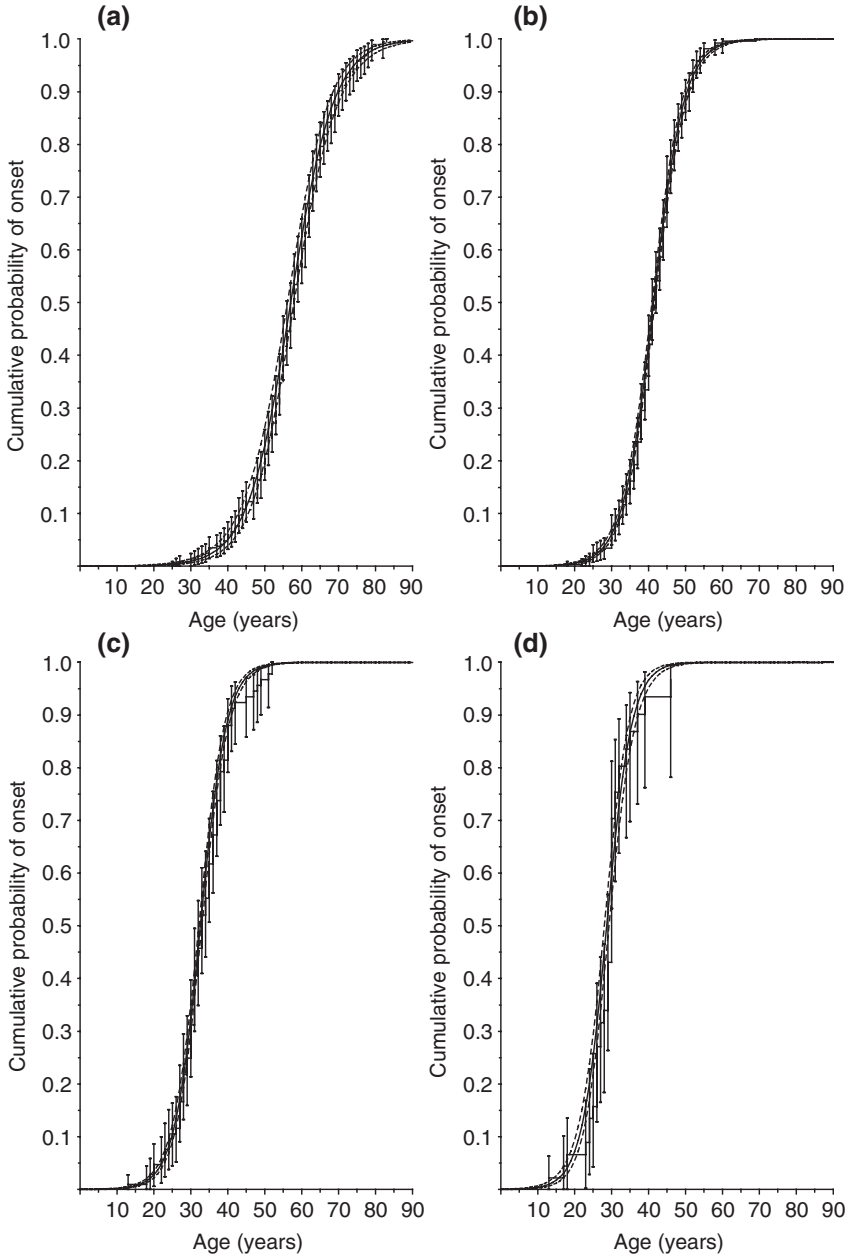

Fig. 2. Cumulative probability of onset for CAG repeats 41 (a), 45 (b), 49 (c), and 53 (d) CAG repeats. Staircase lines represent the non-parametric (Kaplan–Meier) analysis with bars representing 95% confidence intervals. Smooth curves represent prediction based on the parametric model with dashed lines representing 95% confidence intervals.

illustrated by the greater spread in the distribution of the age of onset (Figs 1b and 3). For high repeat lengths, most individuals are predicted to have onset within a relatively narrow age range. The transition in this variability is quite smooth as CAG varies (Figs 3 and 4).

Prediction of onset

Age-specific probability of onset, predicted at birth, is given by equation as given in the '*Methods*' section. An estimate that is of greater clinical relevance is the conditional probability, given a person's current age, of having onset by a particular point in the future. For example, a 40-year-old individual with 39 repeats could be told that we estimate a 52% chance of onset by the age of 70 (Table 2). However, should that same individual still be presymptomatic by the age of 65, they would have only a 26% chance of onset within the next 5 years (Tables 1 and 2). Similarly, a 40-year-old individual with 41 repeats could be told that he has a 9590% chance of onset by the age of 7570, while a 40-year-old individual with 44 repeats is almost certain to be affected by that age (Table 2).

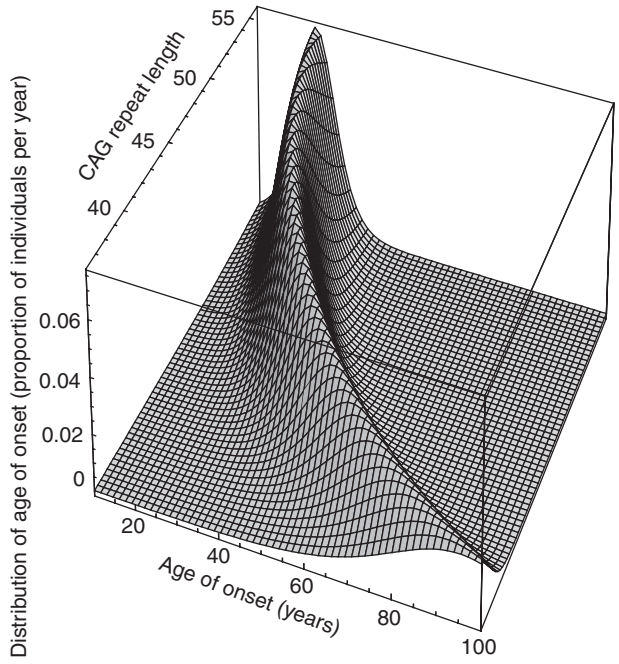

Fig. 3. Distribution of the age of onset for individuals with 36–56 CAG repeats based on the parametric model.

Conditional predictions, and mean and median ages of onset, for individuals aged less than 91 years old with a CAG between 36 and 56 are available in booklet form from the authors or at http://www.cmmt.ubc.ca/clinical/hayden.

Penetrance

A disease is partially penetrant if not all individuals manifest symptoms within a normal lifetime. The parametric model we have developed can be used to estimate the age and CAG-specific penetrance of HD [Indeed, this is simply 1–S(*Age*, *CAG*)]. Our model suggests there is a substantial probability that individuals with less than 40 repeats will not have onset within their lifespan. While this has been previously recognized (4), here we provide for the first time CAG-specific penetrance rates (Table 2). For example, we predict (with a caveat regarding extrapolation) only a 14% chance that a 40-year-old individual with 36 repeats will have onset before the age of 75 (Table 2).

Although ascertainment issues excluded these individuals from direct analysis, individual cases in the short CAG range support the plausibility of this low penetrance. For example, 12 of 20 individuals with CAG lengths 36–39 and over the age of 74 were still presymptomatic. Thirteen of the 410 individuals with a CAG less than 41 were older than 74 and still presymptomatic. The oldest presymptomatic individual was a 90-year-old individual with 37 repeats (Table 1). In contrast, the oldest presymptomatic individual with a CAG of 41 was 71 years old There were no presymptomatic individuals older than 70 with a

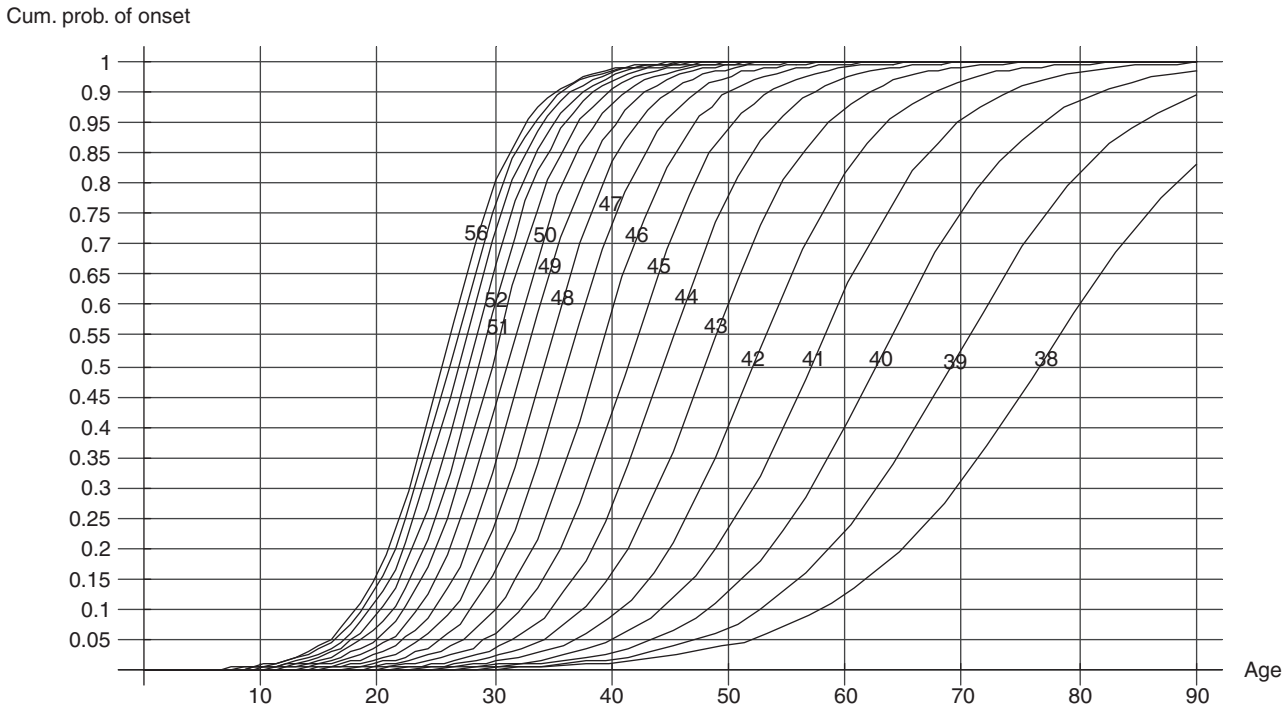

Fig. 4. Cumulative probability of failure curves for various CAG lengths based on the final model. Numbers indicate CAG repeat length. Cum. prob. onset = Cumulative probability of onset of Huntington's disease.

Table 2. Probability of onset for a 40-year-old at-risk individual at 5-year intervals for different CAGs

CAG repeat size	Conditional probability (95% CI) of onset for an individual 40* years by age						
	45 years	50 years	55 years	60 years	65 years	70 years	75 years
36	<0.01 (0–0.02)	0.01 (0–0.05)	0.02 (0–0.08)	0.04 (0.01–0.12)	0.06 (0.02–0.16)	0.09 (0.03–0.22)	0.14 (0.06–0.28)
37	0.01 (0–0.02)	0.02 (0.01–0.05)	0.03 (0.01–0.09)	0.06 (0.03–0.14)	0.10 (0.05–0.20)	0.17 (0.09–0.28)	0.25 (0.16–0.37)
38	0.01 (0–0.03)	0.03 (0.01–0.06)	0.06 (0.03–0.11)	0.11 (0.07–0.19)	0.19 (0.13–0.28)	0.31 (0.23–0.40)	0.45 (0.35–0.54)
39	0.02 (0.01–0.03)	0.05 (0.03–0.08)	0.11 (0.08–0.16)	0.21 (0.16–0.28)	0.35 (0.29–0.43)	0.52 (0.45–0.60)	0.68 (0.60–0.76)
40	0.03 (0.02–0.05)	0.10 (0.08–0.13)	0.21 (0.18–0.26)	0.38 (0.33–0.43)	0.58 (0.52–0.63)	0.75 (0.69–0.80)	0.87 (0.82–0.90)
41	0.06 (0.05–0.08)	0.19 (0.17–0.21)	0.38 (0.35–0.42)	0.60 (0.57–0.64)	0.79 (0.75–0.82)	0.90 (0.87–0.92)	0.95 (0.94–0.97)
42	0.12 (0.11–0.13)	0.34 (0.32–0.36)	0.59 (0.57–0.62)	0.80 (0.78–0.82)	0.91 (0.90–0.93)	0.96 (0.96–0.97)	0.99 (0.98–>0.99)
43	0.22 (0.21–0.24)	0.53 (0.51–0.55)	0.78 (0.76–0.80)	0.91 (0.90–0.92)	0.97 (0.96–0.97)	0.99 (0.99–0.99)	>0.99 (0.99–>0.99)
44	0.36 (0.34–0.38)	0.70 (0.67–0.72)	0.89 (0.87–0.90)	0.96 (0.95–0.97)	0.99 (0.98–0.99)	>0.99 (>0.99–>0.99)	>0.99 (>0.99–>0.99)
45	0.50 (0.47–0.52)	0.81 (0.79–0.83)	0.94 (0.93–0.95)	0.98 (0.98–0.99)	>0.99 (>0.99–>0.99)	>0.99 (>0.99–>0.99)	>0.99 (>0.99–>0.99)
46	0.60 (0.58–0.63)	0.87 (0.86–0.89)	0.96 (0.96–0.97)	>0.99 (>0.99–>0.99)	>0.99 (>0.99–>0.99)	>0.99 (>0.99–>0.99)	>0.99 (>0.99–>0.99)
47	0.67 (0.65–0.69)	0.91 (0.89–0.92)	0.98 (0.97–0.98)	>0.99 (>0.99–>0.99)	>0.99 (>0.99–>0.99)	>0.99 (>0.99–>0.99)	>0.99 (>0.99–>0.99)
48	0.71 (0.69–0.73)	0.92 (0.91–0.94)	0.98 (0.98–0.98)	>0.99 (>0.99–>0.99)	>0.99 (>0.99–>0.99)	>0.99 (>0.99–>0.99)	>0.99 (>0.99–>0.99)
49	0.74 (0.71–0.76)	0.93 (0.92–0.95)	0.98 (0.98–0.99)	>0.99 (>0.99–>0.99)	>0.99 (>0.99–>0.99)	>0.99 (>0.99–>0.99)	>0.99 (>0.99–>0.99)
50	0.75 (0.72–0.78)	0.94 (0.93–0.95)	0.99 (0.98–0.99)	>0.99 (>0.99–>0.99)	>0.99 (>0.99–>0.99)	>0.99 (>0.99–>0.99)	>0.99 (>0.99–>0.99)
51	0.76 (0.73–0.79)	0.94 (0.93–0.96)	0.99 (0.98–>0.99)	>0.99 (>0.99–>0.99)	>0.99 (>0.99–>0.99)	>0.99 (>0.99–>0.99)	>0.99 (>0.99–>0.99)
52	0.77 (0.73–0.80)	0.95 (0.93–0.96)	0.99 (0.98–>0.99)	>0.99 (>0.99–>0.99)	>0.99 (>0.99–>0.99)	>0.99 (>0.99–>0.99)	>0.99 (>0.99–>0.99)
53	0.77 (0.73–0.81)	0.95 (0.93–0.96)	0.99 (0.98–>0.99)	>0.99 (>0.99–>0.99)	>0.99 (>0.99–>0.99)	>0.99 (>0.99–>0.99)	>0.99 (>0.99–>0.99)
54	0.77 (0.73–0.81)	0.95 (0.93–0.96)	0.99 (0.98–>0.99)	>0.99 (>0.99–>0.99)	>0.99 (>0.99–>0.99)	>0.99 (>0.99–>0.99)	>0.99 (>0.99–>0.99)
55	0.78 (0.73–0.81)	0.95 (0.93–0.97)	0.99 (0.98–>0.99)	>0.99 (>0.99–>0.99)	>0.99 (>0.99–>0.99)	>0.99 (>0.99–>0.99)	>0.99 (>0.99–>0.99)
56	0.78 (0.73–0.82)	0.95 (0.93–0.97)	0.99 (0.98–>0.99)	>0.99 (>0.99–>0.99)	>0.99 (>0.99–>0.99)	>0.99 (>0.99–>0.99)	>0.99 (>0.99–>0.99)

*Predictions are available for other ages at http://www.cmmt.ubc.ca/hayden.

Table 3. CAG related penetrance rates of HD

Age (years)	Penetrance of HD (95% CI) for a CAG repeat of							
	36	37	38	39	40	41	42	43
70	0.10 (0.03–0.25)	0.17 (0.09–0.30)	0.32 (0.23–0.42)	0.53 (0.45–0.61)	0.76 (0.70–0.80)	0.90 (0.88–0.92)	0.97 (0.96–0.97)	0.99 (0.99–0.99)
75	0.14 (0.06–0.30)	0.26 (0.16–0.39)	0.45 (0.36–0.55)	0.69 (0.61–0.76)	0.87 (0.83–0.90)	0.96 (0.94–0.97)	0.99 (0.98–0.99)	1.00 (1.00–1.00)
80	0.21 (0.10–0.37)	0.37 (0.26–0.50)	0.60 (0.50–0.69)	0.81 (0.74–0.87)	0.93 (0.90–0.96)	0.98 (0.97–0.99)	1.00 (0.99–1.00)	1.00 (1.00–1.00)
85	0.29 (0.17–0.45)	0.49 (0.38–0.61)	0.73 (0.62–0.82)	0.90 (0.83–0.94)	0.97 (0.95–0.98)	0.99 (0.99–0.99)	1.00 (1.00–1.00)	1.00 (1.00–1.00)

CAG greater than 41, and the model predicts a 95% chance that a 40-year-old individual with 41 repeats will have onset by the age of 70 (Table 2).

Discussion

The predictive model we have developed is based on the largest collection of HD patients reported to date, from a worldwide collaboration of 40 HD centers. This study indicates that it is possible to derive a clinically useful model that expresses the relationship between having a certain repeat size and the probability that disease onset will occur by a certain age. By incorporating both affected and at-risk individuals, more powerful and less-biased statistical techniques have been applied. This allowed us to develop a parametric survival model that predicts the age-specific probability of onset with much narrower confidence limits.

Nonetheless, we must acknowledge certain limitations of this study. For our model to be useful, it should be representative of the HD population for which predictions are to be made. As with most earlier reported models, there is some risk of bias in our analysis resulting from the clinical based sampling of gene-tested subjects. Given that only 9–20% of at-risk individuals approached by testing centers take part in predictive testing, it is possible that people who choose to be tested are not representative of the HD population as a whole (26–28).

Beyond this, the issue of CAG-dependent ascertainment should be considered. Falush et al. (29) have presented a quantitative population model suggesting that clinical ascertainment of CAG expansions may be essentially complete for lengths of 44 or greater, over 80% for lengths 42 and 43 and 50–60% for length 41. Combining those findings with the consistent distribution shape that we observed for that range, we believe it is quite plausible that our data for repeat lengths 41–56 are minimally biased relative to the total population at risk.

In contrast, previous studies have suggested underascertainment is especially substantial for individuals with smaller repeats. Falush et al. (29) estimate that as few as one of 20 individuals with 38 repeats are ascertained as affected with neurological symptoms. This is consistent with our model that predicts that the mean age of onset for such individuals is 77 years, approaching the limit of a normal lifespan (Tables 1 and 2).

Also, given these considerations, a direct survival analysis of any group of known individuals with a CAG between 36 and 40 likely overestimates the probability of onset for that same range of the general population. We attempted to avoid this bias by estimating the parametric model using only individuals from the CAG range where such non-representative ascertainment seems much less likely. We have extrapolated from this model to provide less-pessimistic estimates for individuals with smaller repeats. While this approach seems likely to provide more accurate estimates for the 36–40 range, we must emphasize that it is a projection.

We have partially tested this projection by additional modeling of CAG lengths less than 40 under speculative assumptions: (1) that ascertainment probability is a function of onset age, with progressively lower ascertainment for later onset; and (2) the true onset distributions have the same symmetrical shape as observed at longer CAG lengths. All such models suggest the extrapolations to these CAG ranges presented in this article are, if anything, conservative (in the sense that the true distributions may be shifted to even later onset ages).

This is not the first report of reduced penetrance in individuals with less than 40 repeats (4, 30). However, previous studies were limited by their small sample sizes, which precluded accurate numerical estimation of the non-penetrance. Based on extrapolations from our model, many individuals with a CAG less than 41 will not show symptoms of HD within their lifetime, including up to 86% of those with 36 repeats (Table 2). This provides the first numeric estimate of penetrance of HD by age and repeat length.

All reported confidence limits were calculated assuming the observations were statistically independent. However, there may have been some residual dependencies, because many of the subjects came from common pedigrees. Our previous attempt to detect an effect of pedigree size on estimates did not, however, find any significant correlation, suggesting that the main source of pedigree-based dependence is almost certainly repeat length (4). Having accounted for this as the main predictor under study, we have accounted for much of the pedigree-based dependence. Additional calculations suggest that, at most, pedigree dependencies might possibly inflate the confidence limits that we list by 50% (Details available from DRL on request). We believe the correction would be much smaller, but even at this upper limit, we are left with relatively narrow confidence bands in Fig. 2 – much narrower than any previously reported. Interfamilial correlations also appear to be unaccounted for in all previous estimates of the CAG-onset relationship [with a qualified exception in Ranen et al. (15)].

The significant association between the variance of onset and CAG suggests that the contribution

of other modifiers (both genetic and environmental) is less obvious in individuals with higher repeats sizes (e.g. greater than 44) due to the overwhelming effect of polyglutamine length. Conversely, the larger onset variance for smaller repeats could be indicative of modifiers playing a greater role when the CAG size is less (Fig. 3). Differences in the CAG distribution of the cohorts used to investigate the influence of modifiers such as apolipoprotein E on the age of onset of HD could be responsible for varying findings of significance of the effect of different genetic modifiers on the age of onset. In the future, it might be helpful to compare the effects of potential modifiers on those individuals with lower (e.g. less than 42) *vs* higher repeat size.

CAG length is clearly not the only determinant of HD onset. After controlling of these lengths, parental age of onset has remained an additional strong predictor in previous studies and presumably reflects additional genetic and/or environmental influences (9–15). Paternal *vs* maternal transmission (13, 15), additional genetic polymorphisms (31–35), and an interaction between expanded CAG length and the length of the normal allele (36) have also been reported.

As it stands, this potential additional information remains hidden in the residual variability of our model. Our results are primarily valuable as a detailed description of CAG length influence rather than as the most accurate possible source of onset-age forecasting. We must emphasize, however, that these other prognostic factors have not been studied in as large or as representative of a sample as we have used. Because of the relatively large standard errors and potential bias attached to existing estimates of their effects, there is no guarantee that use of this information, when available, can currently provide a more accurate prognostic forecast that the conditional probabilities that we provide based only on age and CAG length.

The significant association between the variance of onset and CAG suggests that the contribution of other modifiers (both genetic and environmental) is less obvious in individuals with higher repeats sizes (e.g. greater than 44) due to the overwhelming effect of polyglutamine length. Conversely, the larger onset variance for smaller repeats could be indicative of modifiers playing a greater role when the CAG size is less (Fig. 3). Differences in the CAG distribution of the cohorts used to investigate the influence of modifiers such as apolipoprotein E on the age of onset of HD could be responsible for varying findings of significance of the effect of different genetic modifiers on the age of onset. In the future, it might be helpful to compare the effects of potential modifiers on those individuals with lower (e.g. less than 42) *vs* higher repeat size.

Several of the authors (DRL, MRH, and JSP) are now engaged in a prospective, longitudinal study of people with HD CAG expansions of known lengths. This study, PREDICT-HD, aims at identifying early markers of pathogenetic progression that may eventually be useful in the development and monitoring of prophylactic treatment. While definitive results will require prospective observation for actual disease conversion, we are finding conditional probabilities similar to those available on our website to be quite useful in the analysis of the baseline data.

Research applications similar to the one just described illustrate the utility of our model. With appropriate caveats that some bias toward early onset may remain, we believe that predictions from the model also have an important place in clinical practice. These should provide useful prognostic information to those who desire it. Indeed, for some, such information may be of the utmost importance as they plan their future. The lines of evidence suggesting that those with relatively short CAG length may have very late-life onset of disease (if indeed, they experience it at all) should be especially welcome news. Many of those people now live with the expectation that HD lies inevitably in their future.

Genetic tests are rightly celebrated as the first clinical fruits of the molecular biology revolution. We have used a large sample and specialized statistical methods to accurately model the distribution of a clinical outcome on the basis of a genetic test. Our results will allow the clinician to provide detailed prognostic information to patients wishing to know the significance of their CAG length. We believe that this may be a harbinger of the next clinical harvest of the revolution – similarly quantitative risk models for numerous other illnesses. It is our hope that such models will not only allow us to predict the future, but – via more specifically targeted clinical research and interventions – hasten the day when we are also at greater liberty to move from prediction of risk to prevention of disease.

Acknowledgements

The authors thank all doctors and patients who have contributed blood samples and clinical information to HD DNA banks and databases. We also thank Y. Zhao (UBC), D. Oakes (University of Rochester), and M. Jones (University of Iowa) for helpful comments regarding certain challenges in the statistical analysis. This work was supported by a MRC studentship (Canada) awarded to RRB and a CIHR grant to MRH. MRH

is a holder of a Canada Research Chair. DRL and JSP were partially supported in this research by NIH grant NS40068 (1 R01 NS40068-01A1).

References

1. Harper PS. Huntington's disease. In: Major problems in neurology (Warlow CP, van Gijn J, eds). Great Britain: The University Press, 1996.
2. Hayden MR. Huntington's chorea. New York: Springer-Verlag, 1981.
3. The Huntington's Disease Collaborative Research Group. A novel gene containing a trinucleotide repeat that is expanded and unstable on Huntington's disease chromosomes. Cell 1993: 72: 971–983.
4. Brinkman RR, Mezei MM, Theilmann J, Almqvist E, Hayden MR. The likelihood of being affected with Huntington disease by a particular age, for a specific CAG size. Am J Hum Genet 1997: 60: 1202–1210.
5. Bruland O, Almqvist EW, Goldberg YP, Boman H, Hayden MR, Knappskog PM. Accurate determination of the number of CAG repeats in the Huntington disease gene using a sequence-specific internal DNA standard. Clin Genet 1999: 55: 198–202.
6. Kremer B, Goldberg P, Andrew SE et al. A worldwide study of the Huntington's disease mutation. The sensitivity and specificity of measuring CAG repeats. N Engl J Med 1994: 330: 1401–1406.
7. Nance MA, Mathias-Hagen V, Breningstall G, Wick MJ, McGlennen RC. Analysis of a very large trinucleotide repeat in a patient with juvenile Huntington's disease. Neurology 1999: 52: 392–394.
8. Rubinsztein DC, Leggo J, Coles R et al. Phenotypic characterization of individuals with 30–40 CAG repeats in the Huntington disease (HD) gene reveals HD cases with 36 repeats and apparently normal elderly individuals with 36–39 repeats. Am J Hum Genet 1996: 59: 16–22.
9. Duyao M, Ambrose C, Myers R et al. Trinucleotide repeat length instability and age of onset in Huntington's disease. Nat Genet 1993: 4: 387–392.
10. Snell RG, MacMillan JC, Cheadle JP et al. Relationship between trinucleotide repeat expansion and phenotypic variation in Huntington's disease. Nat Genet 1993: 4: 393–397.
11. Andrew SE, Goldberg YP, Kremer B et al. The relationship between trinucleotide (CAG) repeat length and clinical features of Huntington's disease. Nat Genet 1993: 4: 398–403.
12. Stine OC, Pleasant N, Franz ML, Abbott MH, Folstein SE, Ross CA. Correlation between the onset age of Huntington's disease and length of the trinucleotide repeat in IT-15. Hum Mol Genet 1993: 2: 1547–1549.
13. Trottier Y, Biancalana V, Mandel J-L. Instability of CAG repeats in Huntington's disease: Relation to parental transmission and age of onset. J Med Genet 1994: 31: 377–382.
14. Lucotte G, Turpin JC, Riess O et al. Confidence intervals for predicted age of onset, given the size of (CAG)n repeat, in Huntington's disease. Hum Genet 1995: 95: 231–232.
15. Ranen NG, Stine OC, Abbott MH et al. Anticipation and instability of IT-15 (CAG)n repeats in parent–offspring pairs with Huntington disease. Am J Hum Genet 1995: 57: 593–602.
16. Brandt J, Bylsma FW, Gross R, Stine OC, Ranen N, Ross CA. Tri-nucleotide repeat length and clinical progression in Huntington's disease. Neurology 1996: 46: 527–531.
17. Vuillaume I, Vermersch P, Destee A, Petit H, Sablonniere B. Genetic polymorphisms to the CAG repeat influence clinical features at onset in Huntington diseases. J Neurol Neurosurg Psychiatry 1998: 64: 758–762.
18. Maat-Kievit A, Losekoot M, Zwinderman K et al. Predictability of age at onset in Huntington disease in the Dutch population. Medicine 2002: 81: 251–259.
19. Warner JP, Barron LH, Brock DJ. A new polymerase chain reaction (PCR) assay for the trinucleotide repeat that is unstable and expanded on Huntington's disease chromosomes. Mol Cell Probes 1993: 7: 235–239.
20. Cox DR, Oakes D. Analysis of survival data. London: Chapman & Hall, 1984: 80–110.
21. Hogg RV, Craig AT. Introduction to Mathematical Statistics, 5th edn. Englewood Cliffs, NJ: Prentice Hall, 1995: 261–262.
22. The Math Works, Inc. Optimization toolbox user's guide. Natick, MA: The Math Works, Inc, 2002.
23. Wolfram S. The Mathematica book, 5th edn. Champaign, IL: Wolfram Media, Inc., 2003.
24. Knight K. Mathematical Statistics. New York: Chapman & Hall/CRC, 2000:27.
25. Kalbfleisch JD, Prentice RL. The statistical analysis of failure time data, 2nd edn. New York: Wiley, 2002.
26. Meiser B, Dunn S. Psychological impact of genetic testing for Huntington's disease: an update of the literature. J Neurol Neurosurg Psychiatry 2000: 69: 574–578.
27. Tyler A, Ball D, Craufurd D. Presymptomatic testing for Huntington's disease in the United Kingdom. The United Kingdom Huntington's Disease Prediction Consortium. BMJ 1992: 304: 1593–1596.
28. Creighton S, Almqvist EW, Hayden MR. Predictive, prenatal and diagnostic genetic testing for Huntington's disease: The experience in Canada from 1987–2000. Clin Genet 2003: 63: 462–475.
29. Falush D, Almqvist EW, Brinkmann RR, Iwasa Y, Hayden MR. Measurement of mutational flow implies both a high new-mutation rate for Huntington disease and substantial underascertainment of late-onset cases. Am J Hum Genet 2000: 68: 373–385.
30. McNeil SM, Novelletto A, Srinidhi J et al. Reduced penetrance of the Huntington's disease mutation. Hum Mol Genet 1997: 6: 775–779.
31. Li JL, Hayden MR, Almqvist EW et al. A genome scan for modifiers of age at onset in Huntington disease: The HD MAPS Study. Am J Hum Genet 2003: 73: 682–687.
32. Rubinsztein DC, Leggo J, Chiano M et al. Genotypes at the GluR6 kainate receptor locus are associated with variation in the age of onset of Huntington disease. Proc Natl Acad Sci USA 1997: 94: 3872–3876.
33. MacDonald ME, Vonsattel JP, Shrinidhi J et al. Evidence for the GluR6 gene associated with younger onset age of Huntington's disease. Neurology 1999: 53: 1330–1332.
34. Kehoe P, Krawczak M, Harper PS, Owen MJ, Jones AL. Age of onset in Huntington disease: sex specific influence of apolipoprotein E genotype and normal CAG repeat length. J Med Genet 1999: 36: 108–111.
35. Panas M, Avramopoulos D, Karadima G, Petersen MB, Vassilopoulos D. Apolipoprotein E and presenilin-1 genotypes in Huntington's disease. J Neurol 1999: 246: 574–577.
36. Djousse′ L, Knowlton B, Hayden M et al. Interaction of normal and expanded CAG repeat sizes influences age at onset of Huntington disease. Am J Med Genet 2003: 119A: 279–282.

Appendix

Participating centers and number of patients (number of presymptomatic at risk, number of affected patients tested)

Austria

Aschauer Harald, Department of General Psychiatry, University Hospital for Psychiatry, Vienna (1, 8)

Belgium

Eric Legius, Center for Human Genetics, Leuven (2, 81)

Verellen Lannoy, Center de Genetique Humaine et Unite de Genetique Medicale, Bruxelles (24, 80)

Canada

Tillie Chiu, Children's Hospital of Eastern Ontario, Ottawa (14, 7)

Cathy Gillies, Janice Schween, Thunder Bay District Health Unit, Thunder Bay (5, 2)

Heather Hogg, Jill Beis, Christie Riddel, Medical Genetics, IWK Grace Health Center, Halifax (0, 36)

Odell Loubser, Ryan Brinkman, Elisabeth Almqvist, Susan Creighton, Michael Hayden Department of Medical Genetics, University of British Columbia, Vancouver (294, 668)

Wendy Meschino, Department of Genetics, North York General Hospital, North York (50, 31)

David Rosenblatt, Maria Galvez, Division of Medical Genetics, Department of Medicine, McGill University, Montreal (25, 26)

Anaar Sajoo, Sandra Farrell, The Credit Valley Hospital, Mississauga (3, 4)

Germany

Elke Holinski-Feder, Martin Daumer, Michael Scholz, Department of Medical Genetics, University of Munich, Munich (0, 52)

Italy

Paola Mandich, Emilio Di Maria, Department of Neurological Sciences and Vision, University of Genova, Genova and Andrea Novelletto, Department of Cell Biology, University of Calabria, Rende (40, 184)

Japan

Ichiro Kanazawa, Jun Goto, Department of Neurology, University of Tokyo Hospital, Tokyo (35, 182)

South Africa

Jacquie Greenberg, Alison September, Department of Human Genetics, University of Cape Town Medical School, Cape Town (3, 46)

Amanda Krause, Department of Human Genetics, South African Institute for Medical Research and University of the Witwatersrand, Johannesburg (3, 5)

Sweden

Gabrielle Ahlberg, Center of Molecular Medicine, Karolinska Hospital, Stockholm (7, 27)

Ingela Landberg, Ulf Kristoffersson, Department of Clinical Genetics University Hospital, Lund (2, 25)

United States of America

Bonnie Baty, Department of Pediatrics, University of Utah Health Sciences Center, Salt Lake City (4, 3)

Robin Bennett, Thomas Bird, Department of Medical Genetics, University of Washington, Seattle (32, 143)

Laurie Carr, Susan Perlma, Department of Neurology, University of California, Los Angeles (0, 42)

Kimberly Quaid, Indiana University/Purdue University, Indianapolis (22, 8)

Kathleen Delp, Spectrum Health Genetics, Grand Rapids (4, 3)

Mahala Earnhart, Brad Hiner, Movement Disorder Center, Marshfield Clinic, Marshfield (4, 7)

Carolyn Gray, Richard M. Dubinsky, Department of Neurology, University of Kansas Medical Center, Kansas City (6, 9)

Madaline Harrison, Department of Neurology, University of Virginia Health System, Charlottesville (7, 39)

Don Higgins, Departments of Neurology and Neuroscience, Ohio State University, Columbus (13, 49)

Danna Jennings, Yale, New Haven (18, 14)

John Johnson, Linda Beischel, Shodair Hospital, Helena (10, 15)

Karen Kovak, Oregon Health Sciences University, Portland (0, 2)

Katie Leonard, Baylor College of Medicine, Houston (17, 4)

Richard H. Myers, Nat Couropmitree, Beth Knowlton, Boston University School of Medicine, Boston (32, 226)

Martha Nance, Park Nicollet Clinic, St. Louis Park (12, 198)

Mark E. Nunes, United States Air Force Medical Genetics Center, Keesler Air Force Base, Mississippi (6, 8)

Jane Paulsen, Beth Turner, Departments of Psychiatry and Neurology University of Iowa, Iowa City (9, 2)

Guerry Peavy, Jody Corey-Bloom, Mark Jacobson, University of California, San Diego (9, 54)

Adam Rosenblatt, Christopher Ross, Johns Hopkins University School of Medicine, Baltimore (2, 158)

Kathleen Shannon, Rush Presb/St. Lukes Medical Center, Chicago (9, 29)

Elaine Spector, University of Colorado Health Sciences Center DNA Diagnostic Laboratory, Maureen Leehey, University of Colorado School of Medicine, Lauren Seeberger, Colorado Neurologic Institute, Denver (14, 23)

Carrie Stoltzfus, David R. Witt, Elaine Louie, Genetics Department, Kaiser Permanente, Northern California, San Jose (29, 38)

Andrea Zanko, Division of Medical Genetics, University of California, San Francisco (51, 96)