Am. J. Hum. Genet. 49:735–745, 1991

Covariate-dependent Age-at-Onset Distributions for Huntington Disease

Michael Krawczak,* Barbara Bockel,* Lodewijk Sandkuijl,† Ulrike Thies,* Ian Fenton,† and Peter S. Harper†

*Institute of Human Genetics, Göttingen; and †Institute of Medical Genetics, Cardiff

Summary

A combined logistic regression and life-table analysis is presented on age-at-onset data for Huntington disease. Covariates included in the analysis were sex of the at-risk individual, parental age at onset, and sex of transmitting parent. Parental age at onset and parental sex were found to be significant covariates for age at onset in the offspring, and the appropriate logistic regression functions are calculated by maximum likelihood methods. These regression functions permit a more precise evaluation of carrier risks and likelihoods than hitherto was possible by simple computational means. We further introduce a novel method to account for sibship correlations in the significance assessment, using log-likelihood differences between different models.

Introduction

Huntington disease (HD) is an autosomal dominant, neurodegenerative disorder that is manifested in involuntary movements, dementia, and psychiatric anomalies. The prevalence of HD is approximately 1 in 20,000 among Caucasians, with considerably lower rates being observed in Asians and African blacks (Hayden 1981) and in the Finnish population (Paolo et al. 1987). The gene mutation causing HD has been located genetically to the distal region of the short arm of chromosome 4 (Gusella et al. 1983), but the structure and function of the putative HD gene still remain unknown. However, a number of closely linked DNA markers have been reported that improve the accuracy of risk assessment, and these may serve as starting points for the molecular genetic search for the HD gene (Gilliam et al. 1987; Smith et al. 1988; Wasmuth et al. 1988; MacDonald et al. 1989).

One of the most remarkable features of HD is its delayed onset, usually occurring around the fourth decade but in several cases ranging from early childhood to after the age of 70 years. This variation is of considerable relevance for both scientific purposes and genetic counseling, and several studies have therefore been performed in order to estimate the statistical distribution of age at onset (AO) for HD. Some of these studies, however, yielded strikingly different results, with mean AO ranging from 33.8 to 51.6 years (for review, see Hayden 1981). The reasons for these differences may be manifold, but in part they can be attributed to diagnostic and statistical problems. The definition of onset of HD is not unique at all, and, if the most conservative approach is adopted—i.e., the first appearance of any neurological or psychiatric signs—the accuracy of AO assessment is largely dependent on patients or families themselves. Further, some of the studies mentioned above considered only affected individuals, which would consequently result in a downward bias in the estimation of AO distributions. Carriers with late onset were systematically excluded from these studies, either because they had died or were lost to a study or because a study had simply been completed before manifestation actually took place. While the first problem is more a question of satisfying definitions and ascertainment criteria, the second problem—i.e., that of censored data—can be overcome by the choice of appropriate statistical methods.

Received March 4, 1991; revision received May 24, 1991.

Address for correspondence and reprints: Dr. Michael Krawczak, Institute of Human Genetics, Konstanty-Gutschow-Strasse 8, W-3000 Hannover 61, Germany.

0002-9297/91/4904-0006$02.00

In principle, two methods were proposed to obtain unbiased estimates of AO distributions in a retrospective manner. One is the use of remote cohorts, including only those affected individuals who were born before a certain early date. This approach, adopted in an early study by Wendt (1959), might control for the bias which is introduced by censoring due to the end of a study, but other sources (e.g., death) are not excluded. Thus, as noted also by Adams et al. (1988), studies based on remote cohorts do not seem to deal adequately with the censoring bias. Another, statistically more reliable approach is the use of life-table techniques. This method, which has been applied before by several other authors (Newcombe 1981; Adams et al. 1988; Cupples et al. 1989), deals with hazard rates within a certain time interval, instead of with empirical distribution functions. Life tables are known to yield unbiased results (if censoring is independent of AO) and should therefore be preferred for AO analyses.

Intrafamilial correlations of AO of HD have been observed in several studies showing that both the AO (Myers et al. 1982, 1985; Ridley et al. 1988) and the sex of the transmitting parent (Brackenridge 1973; Newcombe et al. 1981; Myers et al. 1982, 1985) are of considerable influence on the AO in the children. These findings suggest that onset appears earlier in children of male transmitting parents and that the AO is positively correlated between parents and children. The influence of such covariates on the AO has been accounted for by Chase et al. (1986) in a life-table design. A Cox proportional hazard model was used by these authors modeling annual hazards as a specific function of sex and race. Race, however, did not appear to be a statistically significant covariate.

In the present study, we adopt an approach similar to that of Chase et al. (1986), using sex of the at-risk individual, parental AO, and parental sex as covariates. Again, annual hazard rates are modeled as functions of the covariates, but with logistic regression applied instead of a Cox proportional hazard model. Logistic regression has several theoretical advantages over other models, and the application to genetic problems has been outlined in detail by Bonney (1986). Information on the offspring of at-risk individuals will also be included in our calculations, in order to adjust their prior risks. As a result of the analysis, we present regression functions on AO that will allow easy AO-based carrier risk assessment and likelihood calculation, accounting for intrafamilial correlations.

Material and Methods

Data Analysis

Life-table techniques involve the estimation of hazard rates, $h(t)$, which are defined as the probability of an individual at risk getting affected at time t, given that he or she has onset not earlier than t (Kalbfleisch and Prentice 1980). This can be written as $h(t) = \text{Prob}(\text{AO} = t \text{ given AO} > t - 1)$. From the $h(t)$, the distribution of AO can be calculated using the following formula:

$$\text{Prob}(\text{AO} = t) = h(t) \prod_{i=1}^{t-1} [1 - h(i)] .$$

$h(t)$ can be estimated from a set of data in different ways. As empirical distribution functions, they can be calculated directly as the number of individuals with onset at time t, divided by the total number of individuals who are at risk prior to time t. So, unaffected at-risk individuals censored at time t_0 will contribute to all $h(t)$ with $t < t_0$. To include covariates, data have to be split into classes depending on the distinct covariate levels. However, if either the number of covariates or the number of classes per covariate is large, some of the classes may not contain enough observations to obtain reliable results. Further, borderline cases are difficult to evaluate for continuous covariates.

A more convenient approach with respect to practical applications is to fit a specified function, $h(t, w)$, of time t and covariates w to the $h(t)$. For this approach, the data need not be split into smaller subsets, but the reliability of the results depends on the choice of an appropriate regression function. A method widely used in epidemiology for the analysis of binary outcomes (e.g., affected/unaffected) is linear logistic regression. For a set of covariates $w = w_1, \ldots, w_n$, this kind of calculus is based on the assumption that the logarithm of the risk ratio

$$R(w) = \frac{\text{Prob (affected given } w)}{\text{Prob (unaffected given } w)} \qquad (1)$$

is a linear combination of the covariates included in w, i.e.,

$$\log R(w) = T(w) = a_0 + a_1 w_1 + \ldots + a_n w_n . \qquad (2)$$

Rearranging equalities (1) and (2) yields

$$\text{Prob (affected given } w) = \frac{\exp T(w)}{1 + \exp T(w)}, \quad (3)$$

allowing maximum likelihood estimation of the parameters a_j from a set of affected and unaffected individuals for which the covariates w_j are known.

Life-table methods and logistic regression can be combined by assuming that the logarithm of the risk ratio

$$R(t,w) = \frac{h(t, w)}{1 - h(t, w)} \quad (4)$$

is again a linear combination, $\log R(t, w) = T(t, w)$, of t and the covariates w. Similar to equality (3) we get

$$h(t, w) = \frac{\exp T(t, w)}{1 + \exp T(t, w)}. \quad (5)$$

Now, let t_i denote either AO or age of censoring for individual i. From t_i, the covariates w_i, and the prior carrier risk p_i, the likelihood, L_i, of observing individual i is calculated as

$$L_i = h(t_i, w_i) \prod_{j=1}^{t_i - 1} [1 - h(j, w_i)] \quad (6)$$

if the individual is affected and

$$L_i = p_i \prod_{j=1}^{t_i - 1} [1 - h(j,w_i)] + 1 - p_i \quad (7)$$

if the individual is unaffected. The joint likelihood, L, of the whole data set is taken to be equal to the product of all individual likelihoods L_i. From this and formulas (6) and (7), the factors a_j of the regression function $T(t, w)$ can be calculated by maximum likelihood methods.

Data Material

The present study was based on the analysis of 1,230 at-risk individuals (274 affected) and their affected parents. These data were collected over a period of 20 years in West Germany and the United Kingdom. The vast majority of data come from south Wales, where complete ascertainment has been an aim for a period of 10 years. At-risk individuals are from 445 sibships ranging in size from one to 11. AO, as documented in the surveys included in the present study, was determined by the presence of the first neurological signs (e.g., ataxia, impairment of balance, uncoordinated behavior, chorea, etc.). A considerable amount of data were collected in a retrospective manner from relatives, through questionaires phrased in terms of the patient's functioning.

Likelihood calculations were performed using the following variables and covariates: t, individual's AO or age at censoring; w_1, parental AO; w_2, sex of transmitting parent; and w_3, individual's sex. Sex was coded as 0 for males and as 1 for females. A time scale of 1-year intervals was used for time t. Only individuals for whom all three covariates were known were included in the study. For 345 unaffected at-risk individuals, offspring data were used to modify the prior risk of being a carrier for HD. Any affected offspring changed this risk to 100%, i.e., new mutations were excluded, since they are known to be very rare (Wolff et al. 1989). From unaffected offspring, prior risks were modified using Bayes' formula applied to the published AO distribution given by Adams et al. (1988). This procedure resulted in a total mean prior risk of .494 (SD = 0.04) for unaffected at-risk individuals.

Significance Assessment

Covariates were tested for significant influence on AO in a stepwise manner. Different nested models were compared by a maximum likelihood ratio test, making use of the fact that twice the log-likelihood difference approximately follows a χ^2 distribution. The number of df equals the difference between the numbers of parameters entered into the models. This type of significance testing is, however, problematic if siblings are not independent under given covariates. Sibship correlations in AO have been reported ranging from .28 (Reed and Chandler 1958) to .64 (Bell 1934), which means that the significance of covariates will be systematically overestimated if the total sum of individual log likelihoods is used for χ^2 approximation. The log likelihood for an individual who has siblings in the data set would contribute too much to the total log likelihood if a considerable AO correlation within a sibship were present. In the most extreme case, where AOs within a sibship were strictly correlated (i.e., correlation coefficient $r = 1$), any individual log likelihood must be scored with $1/n$, n denoting sibship size, to obtain the correct log likelihood of the corre-

sponding sibship. Similar scores, derived for lower correlation coefficients, should be larger than $1/n$, approaching unity if r becomes zero. In fact, this scoring approach can be extended to any degree of positive correlation if AO for HD follows a normal distribution. A detailed description of this procedure is given in the Appendix. The scoring factors derived there are $1/[1 + (n - 1)r]$. These scoring factors, however, do not yield the correct log likelihoods, but they will result in lower limits for the log-likelihood differences between different models (see Appendix). For comparison of different models, different correlation coefficients have to be used. If one model (I) is nested within the other (II), then the appropriate r for comparison is the partial sibship correlation coefficient, r_{II}, for which all covariates included in model II are excluded from r. The approach outlined above and in the Appendix may also be useful for other types of analysis dealing with observations that are not independent per se.

Likelihood Maximization

In order to calculate the maximum likelihood estimates of the parameters involved in the logistic regression, we first tried nonlinear programming making use of the gradient of the likelihood function (e.g., the BFGS method). These algorithms, however, failed because of the complex structure of the function, so that we had to apply a direct-search method, the Nelder-Mead algorithm (Nelder and Mead 1964). To maximize a function with m variables, this algorithm starts with a simplex of $m + 1$ points within the parameter space. If L_{max} denotes the maximum value of the function attained on the simplex, the procedure attempts to find in each step a point which yields a value larger than L_{max}. This search proceeds along the line between that point of the simplex which yields the lowest value, a_{min}, and the point of gravity of the remaining points. If the search is successful, the new point replaces a_{min}; otherwise, the simplex is reduced. This search is repeated until all points of the simplex are within a given distance from each other.

Results

In order to validate the derivation of the scoring factors (see Appendix), AO data for the 274 affected individuals were first tested for normality. The Shapiro-Wilk statistic W (Shapiro and Wilk 1965) was calculated as $W = .9827$, indicating a good fit to a normal distribution with mean 36.4 years and SD 11.8.

Pearson correlation coefficients for AO, together with relevant partial correlation coefficients, were calculated from affected siblings and their affected parents by using the CORR procedure of the SAS software package (SAS Institute Inc., 1988). The results are given in table 1. In order to account for parental sex in the likelihood maximization, the sibship correlation coefficients, r and rw_1, were also calculated separately for male and female transmitting parents. Individual log likelihoods were scored using these correlation coefficients whenever models including parental sex as a covariate were compared (see table 2). Two correlation coefficients, r_1 and r_2, from samples of size n_1 and n_2, respectively, can be compared using

$$z = \frac{|\tanh^{-1}(r_1) - \tanh^{-1}(r_2)|}{[1/(n_1 - 3) + 1/(n_2 - 3)]^{0.5}}.$$

This z value follows a standard normal distribution under the hypothesis $r_1 = r_2$ (Pfanzagl 1974). For the sibship correlation coefficients listed in table 1 we obtain $z = 1.102$ ($p = .27$, two-sided) for r^m versus r^f and $z = 2.665$ ($P = .007$) for rw_1^m versus rw_1^f.

As outlined above, the model for logistic regression was extended stepwise by those covariates that yielded the maximum χ^2. Model descriptions, regression estimates, and log likelihoods are presented in table 2. As can be inferred from table 3, parental AO turns out to be a highly significant covariate (model Ib vs. IIIa; χ^2

Table 1

Correlation in *AO* between Siblings and Partial Correlation Coefficients Excluding Parental *AO*

Sex of Transmitting Parent and Abbreviation[a]	Correlation Coefficient
r	.43
rw_1	.32
Male:	
r^m	.46
rw_1^m	.39
Female:	
r^f	.36
rw_1^f	.12

[a] r = sibship correlation coefficient; rw_1 = partial correlation coefficient with parental *AO* excluded from r; superscripts m (male) and f (female) indicate sex of transmitting parent.

Table 2

Logistic Regression Analysis of *AO*

Model, Correlation Coefficient(s), and Factors Included	Regression Factor Estimate	− Log Likelihood[a]
Ia, 0:		
Constant	-6.79	1,298.98
t	9.37×10^{-2}	
Ib, rw_1:		
Constant	-6.71	722.60
t	9.43×10^{-2}	
Ic, r^m and r^f:		
Constant	-6.77	663.44
t	9.59×10^{-2}	
II, 0:		
Constant	-6.77	
t	9.39×10^{-2}	1,298.88
w_3	-6.51×10^{-2}	
IIIa, rw_1:		
Constant	-4.19	
t	1.15×10^{-1}	695.67
w_1	-7.13×10^{-2}	
IIIb, rw_1^m and rw_1^f:		
Constant	-3.87	
t	1.17×10^{-1}	765.98
w_1	-8.07×10^{-2}	
IV, r^m and r^f:		
Constant	-6.48	
t	9.52×10^{-2}	660.04
w_2	-5.36×10^{-1}	
V, rw_1^m and rw_1^f:		
Constant	-3.59	
t	1.19×10^{-1}	758.82
w_1	-7.97×10^{-2}	
w_2	-7.41×10^{-1}	

[a] Imprecise values (only used for χ^2 approximation).

Table 3

Comparison of Logistic Regression Models for *AO*

Models Compared[a]	χ^2 − 2 Log-Likelihood Difference	P, 1 df
Ia vs. II	.20	6.56×10^{-1}
Ib vs IIIa	53.86	$<10^{-5}$
Ic vs. IV	6.80	9.10×10^{-3}
IIIb vs. V	14.32	1.54×10^{-4}

[a] Covariates included in different models are individual's sex (model II), parental *AO* (models III and V), and sex of transmitting parent (models IV and V).

= 53.86, 1 df, $P < 10^{-5}$) for AO in the offspring. The same holds true for the sex of the transmitting parent (model Ic vs. IV; $\chi^2 = 6.80$, 1 df, $P = 9.1 \times 10^{-3}$) but not for the individual's sex (model Ia vs. II; $\chi^2 = 0.20$, 1 df, not significant). It should be noted that the influence of individual's sex on AO was tested without consideration of any sibship correlation. Thus, the (insignificant) χ^2 value of 0.20 still represents an overestimate. If parental AO, the most significant covariate, is included in the regression model, a further significant effect remains for parental sex (model IIIb vs. V; $\chi^2 = 14.32$, 1 df, $P = 1.54 \times 10^{-4}$).

The regression model finally adopted is model V, with regression factors as presented in table 2. Variances and covariances of the factor estimates can be calculated approximately from the inverse of the so-called information matrix, obtained from the second derivative of the log-likelihood surface at the maximum likelihood estimate (Silvey 1987). Approximations of the variance-covariance components for model V are given in table 4, but these figures only represent upper limits for the true variances and covariances; the log-likelihood surface resulting from the scoring procedure is flatter than the true one, and therefore the corresponding second derivatives are too small. From table 4 and the fact that maximum likelihood estimates are approximately normally distributed (Silvey 1987), we also get approximate 95% confidence intervals (mean ± 2 SD) for the regression factors involved in the final model (table 5).

Table 4

Variance-Covariance Approximation for Regression Model V

	Constant	t	w_1	w_2
Constant	1.66×10^{-1}	-4.74×10^{-4}	-3.08×10^{-3}	-2.38×10^{-2}
t		5.48×10^{-5}	-2.99×10^{-5}	-7.00×10^{-5}
w_1			9.25×10^{-5}	2.16×10^{-4}
w_2				3.43×10^{-2}

Table 5

Approximate 95% Confidence Limits for Regression Factors in Model V

Factor	Regression Estimate ± 2 SD
Constant	−3.59 ± .82
t	.119 ± .014
w_1	−.0797 ± .018
w_2	−.741 ± .370

Discussion

In the present paper, we have confirmed previous findings on the parental effects on AO for HD. Onset appears significantly earlier in the children of male transmitting parents, with an annual risk ratio Prob (getting affected)/Prob(remaining unaffected) that is exp(.741) = 2.10 times higher for fathers than for mothers. The corresponding AO distributions for differing sex of transmitting parent are presented in figure 1. The regression parameters used for the calculation of these AO distributions are those from model IV (table 2), which yield mean AO ± SD of 38.30 ± 12.20 years and 43.62 ± 12.67 years for children of male and female transmitting parents, respectively. These figures are in good agreement with those presented by other authors (Myers et al. 1982, 1985; Chase et al. 1986). In figure 1, the AO distribution derived through life-table methods by Adams et al. (1988) is also included. Up to the age of 50 years, this distribution is also close to the average distribution calculated from our data, but above this age point it is shifted to the right. There may be several explanations for this finding.

1. Linear logistic regression may generate regression functions that are not flexible enough to allow for the obvious linearization of the upper tail of the distribution function as observed by Adams et al. (1988). If the latter phenomenon is to be confirmed by others, then results from linear logistic regression would have to be used with care for the evaluation of carrier risks for old unaffected at-risk individuals. A possible solution to this methodological problem would be the use of regression functions of higher order. However, as this will increase the number of parameters that have to be estimated simultaneously and will make the structure of the log-likelihood surface more complex, a larger amount of data will be required in order to obtain reliable results.

2. Probands were excluded from the analysis by Adams et al. (1988), to control for ascertainment bias,

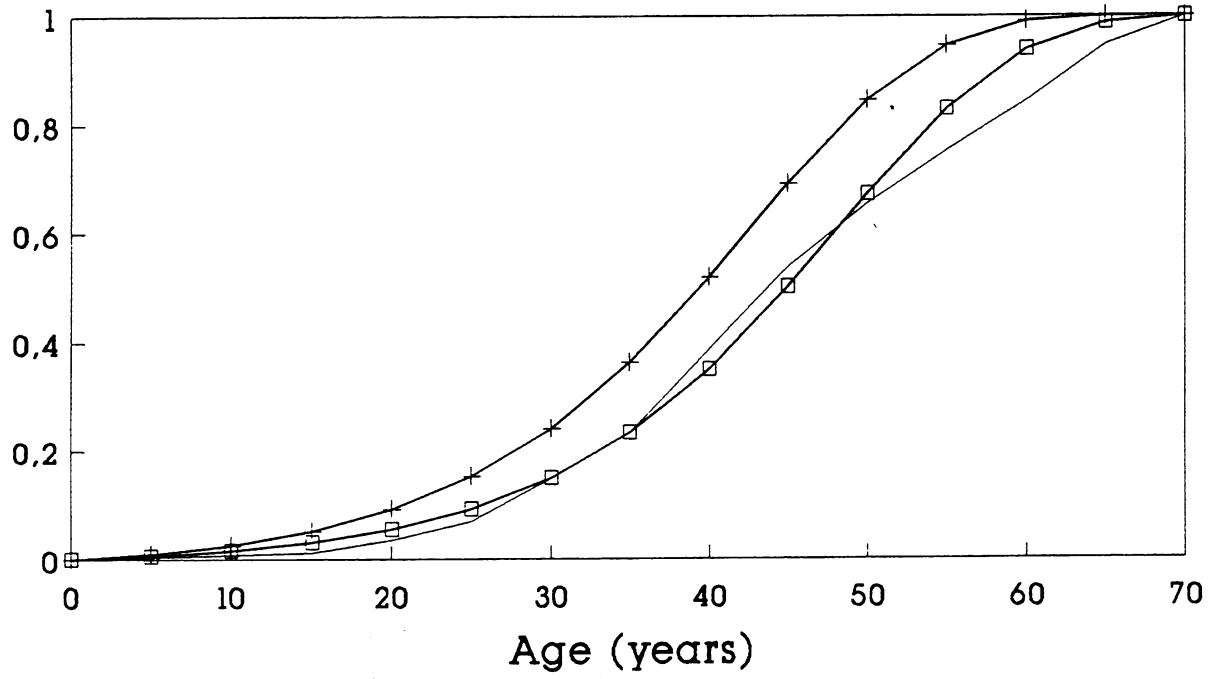


Figure 1 Cumulative AO distributions in offspring of male (—┼—) and female (—⊟—) transmitting parents, compared with cumulative AO distribution in offspring studied by Adams et al. (1988) (——).

which, by preferentially considering older family members, may cause an additional upward shift of the AO distribution. We do not, however, regard ascertainment to be a major source of bias for the study of AO here. HD has been well known and characterized for a long period of time, and therefore the chance of ascertainment of any HD family—and, hence, its structure of intrafamilial AO correlation—will hardly be influenced by the AO of the proband through which the family has come to clinical attention. Evidence for this suggestion also comes from a study by Cupples et al. (1989), in which only minor changes were found in the estimates of mean AO under different proband-exclusion strategies. Finally, as already mentioned above, the majority of our data come from a region where ascertainment can be assumed to be nearly complete.

3. Adams et al. (1988) did not consider offspring information for unaffected at-risk individuals in order to modify their prior carrier risks, as was done in our study. This means that some of the older potential gene carriers might have contributed to the estimates of $h(t)$, although their actual carrier risk was low. However, as both distributions are in good agreement for younger at-risk individuals, using the results of Adams et al. (1988) for prior risk modifications in our sample appears to be justified. Most (97%) unaffected offspring used for these modifications were younger than 55 years.

In figure 2, some selected AO distributions are presented for varying parental sex and AO. Positive correlation between parental and offspring AO, as reflected by these figures, has been noted by several authors (Myers et al. 1982, 1985; Ridley et al. 1988), and our results also indicate a highly significant influence of parental AO on offspring AO. Each year until parental onset had taken place reduces the annual risk ratio by a factor of exp $(-.0797) = .92$. Linear logistic regression is, however, unable to quantify anticipation effects controlling for parental sex. Additional parameters would again be required to allow for this approach, which in turn would reduce the reliability of the regression parameter estimates, for the reasons noted above.

Several genetic and nongenetic (environmental) factors could contribute to the consistently observed parent-offspring correlation in AO for HD. Although the sex of the transmitting parent appears to be a significant covariate, sex-linked genetic factors can be excluded from playing an important role here, as no differences in AO were observed between male and female HD patients. Ridley et al. (1988) discussed in detail the possible effects that different methylation patterns inherited through the male and female germ line might have on the expression of the HD gene ("genomic imprinting"). These authors claim that age-dependent demethylation of the mutant allele in somatic cells may cause onset of symptoms when a certain threshold is reached. Similarity in methylation status would result in similarities of the demethylation process—and, therefore, in AO similarities between patients and their affected offspring. A lower, or "more defective," degree of methylation in paternally derived germ cells would explain earlier onset in offspring of affected males.

A higher sibship correlation in AO for paternally inherited HD, as observed in our study, would further imply that the variability of the parental imprinting effect is smaller for male than for female transmitting carriers. This either could be due to a more variable demethylation process for maternally derived HD alleles or may reflect a higher influence of the "genetic background" in these cases. Further, if paternal HD genes are "remethylated" during embryonal oogenesis, in order to indicate their "maternal origin" (Ridley et al. 1988), then this may contribute a higher variability to the methylation status than would a similar but less "restorative" mechanism acting on maternally derived mutations during spermatogenesis.

One of the major advantage of our regression analysis of AO data is that individual carrier risks can quite easily be calculated even on a programmable pocket calculator, without the use of extensive risk tables. Given the corresponding covariates w, the annual hazard rates $h(t, w)$ result from formula (5) (see Methods) with the appropriate regression function selected from table 2. From the $h(t, w)$, the risk of an unaffected offspring of an affected carrier is calculated as

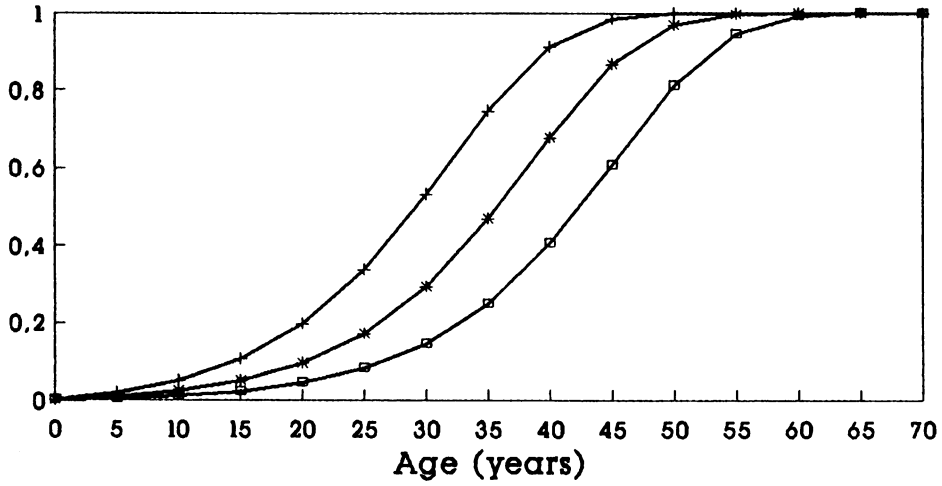
$$Z(t_0, w) = \frac{\text{Prob}(\text{AO} > t_0 \text{ given } w)}{\text{Prob}(\text{AO} > t_0 \text{ given } w) + 1},$$

where t_0 is the age of the at-risk individual and

$$\text{Prob}(\text{AO} > t_0 \text{ given } w) = \prod_{i=1}^{t_0} [1 - h(i, w)]$$

denotes the probability of onset after time t_0, given that the individual is a carrier. If in a line of inheritance the most recent affected individual dates back more than one generation, i.e., if the parent of the at-risk individual is not affected, then the probability Prob

a



b

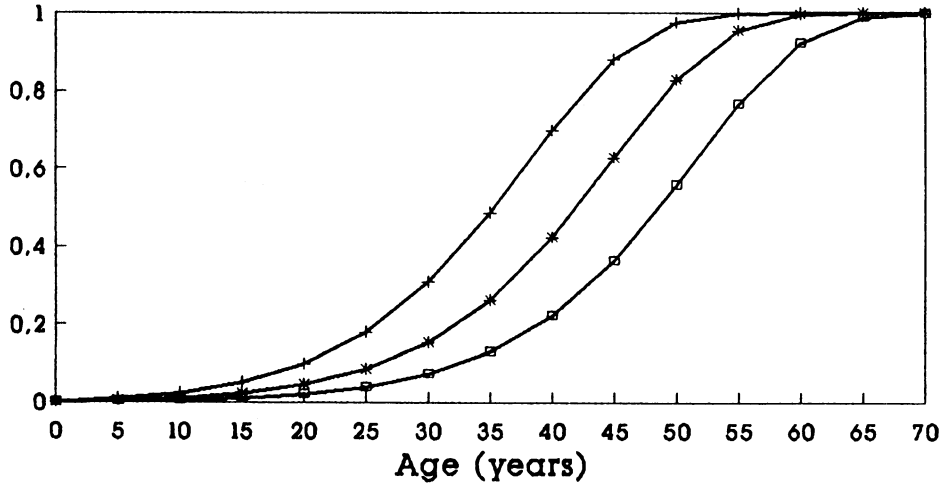


Figure 2 Cumulative AO distributions in offspring of male (*a*) and female (*b*) transmitting parents, controlling for parental AO of 30 years (—┼—), 40 years (—X—), and 50 years (—⊟—).

(AO > t_0 given w) has to be replaced by the average probability taken over all possible AOs in the parent, conditional on the covariates from the grandparent.

In any case, the results presented here will allow a more precise quantification of both carrier risk and likelihood in both genetic counseling and scientific studies. For the purposes of linkage analysis, logistic regression functions could either be incorporated directly into existing computer programs or be used to calculate the penetrances for predefined liability classes. Such applications need not be limited to the study of HD but can also be extended to other traits, for which AO and penetrance depend on familial or environmental covariates.

Acknowledgments

We thank A. Tyler (Cardiff), J. Schmidtke (Hannover), and D. N. Cooper (London) for helpful comments. W. Engel (Göttingen) is gratefully acknowledged for his support and encouragement. This work was supported by the Deutsche Forschungsgemeinschaft.

Appendix

Maximum Likelihood Comparison of Models under Positive Sibship Correlation of AO

Positive sibship correlation of AO would result in biased estimates of the significance of covariates if siblings are regarded as independent. To compare any two nested models, such correlations must be taken into account, and the main idea of our approach is to add only a fraction of an individual's log likelihood to the overall log likelihood, depending on the correlation coefficient and the sibship size. Let AO for HD follow a normal distribution with mean m and variance v. Then an individual's AO distribution is given by the density.

$$h_{m,v}(x) = \frac{1}{\sqrt{2\pi v}} \exp\left[-\frac{(x-m)^2}{2v}\right].$$

For a sibship of size n, the joint distribution of AO is given by the vector-valued density

$$H_{M,C}(X) = \frac{1}{\sqrt{2\pi \det(C)}} \exp\left[-\frac{1}{2}(X-M)^t C^{-1}(X-M)\right],$$

where X is an n vector $(x_1, \ldots, x_n)$, M is the mean vector $(m, \ldots, m)$, and C is the variance-covariance matrix

$$C = \begin{pmatrix} v & c & \ldots & c \\ c & v & \ldots & c \\ \vdots & \vdots & \vdots & \vdots \\ c & c & \ldots & v \end{pmatrix},$$

with v denoting variance and c denoting covariance of AO for any two siblings. The variable part of the function $H_{M,C}(X)$ can now be factorized as

$$f \cdot \sum_{i=1}^{n} \frac{(x_i-m)^2}{v} = (X-M)^t C^{-1}(X-M). \quad \text{(A1)}$$

The second factor on the left side of equation (A1) equals the variable part of $H_{M,C}(X)$ for a sibship correlation of zero (i.e., independence of AO between siblings). Now

$$C^{-1} = \frac{1}{D}\begin{pmatrix} v+(n-2)c & -c & \ldots & -c \\ -c & v+(n-2)c & \ldots & -c \\ \vdots & \vdots & \vdots & \vdots \\ -c & -c & \ldots & v+(n-2)c \end{pmatrix},$$

with $D = v^2 - (n-1)c^2 + (n-2) - c$. So equation (A1) becomes

$$\frac{f}{v}\sum_{i=1}^{n}(x_i-m)^2 = \frac{1}{D}\left\{[v+(n-2)c]\sum_{i=1}^{n}(x_i-m)^2 - c\sum_{i=1}^{n}\sum_{\substack{j=1 \\ j\neq i}}^{n}(x_i-m)(x_j-m)\right\},$$

and the factor f can be calculated as

$$f = f(x_1,\ldots,x_n)$$

$$= \frac{v[v+(n-2)c]}{D} - \frac{cv\sum_{i=1}^{n}\sum_{\substack{j=1 \\ j\neq i}}^{n}(x_i-m)(x_j-m)}{D\sum_{i=1}^{n}(x_i-m)^2}$$

$$= \frac{1+(n-2)r}{D^*} - \frac{r\sum_{i=1}^{n}\sum_{\substack{j=1 \\ j\neq i}}^{n}(x_i-m)(x_j-m)}{D^*\sum_{i=1}^{n}(x_i-m)^2},$$

with $D^* = 1 - (n-1)r^2 + (n-2)r = (1-r)[1+(n-1)r]$ and $r = c/v$, the sibship correlation coefficient of AO. Now, by replacing x_i and x_j with

the maximum x_{max} in the nominator and replacing x_i with the minimum x_{min} in the denominator, we get

$$\frac{1+(n-2)r}{D^*} - \frac{rn(n-1)(x_{max}-m)^2}{D^* n(x_{min}-m)^2} \leqslant f .$$

Thus, f is minimal if $x_{max} = x_{min} = x$.

$$\rightarrow \quad f_{min} = \frac{1}{D^*}[1+(n-2)r-(n-1)r]$$

$$= \frac{1-r}{D^*} = \frac{1}{1+(n-1)r} .$$

Let us now consider a sibship X of affected individuals and two joint distributions H_1 and H_2 with different means and variance-covariance matrices but with the same correlation coefficient r. IF H_1^* and H_2^* denote the corresponding density functions when we assume the independence of siblings (i.e., covariance $c = 0$), then

$$\begin{aligned} &|\log H_1(X) - \log H_2(X)| \\ &= |f(X)| \times |\log H_1^*(X) - \log H_2^*(X)| \\ &\geqslant |f_{min}| \times |\log H_1^*(X) - \log H_2^*(X)| . \end{aligned}$$

It can easily be shown that the same inequality also holds true for the corresponding log likelihoods if some unaffected at-risk individuals are present in the sibship, too. Thus, a lower limit for the log-likelihood difference between two models can be obtained by regarding siblings as independent but with their individual log likelihoods weighted with the corresponding factor f_{min}. If two nested models (I included in II) will be compared by maximum likelihood methods, then the appropriate correlation coefficient to use for log-likelihood comparison is the correlation coefficient belonging to model II. Since the information on the additional covariate is present for the evaluation of *both* models, the influence of the additional covariate also has to be excluded from the correlation coefficient in *both* models.

References

Adams P, Falek A, Arnold J (1988) Huntington disease in Georgia: age at onset. Am J Hum Genet 43:695–704

Bell J (1934) Huntington's chorea. In: A treasury of human inheritance, vol 4. Cambridge University Press, Cambridge, pp 1–77

Bonney GE (1986) Regressive logistic models for familial disease and other binary traits. Biometrics 42:611–625

Brackenridge CJ (1973) The relation of sex of affected parent to the age of onset of Huntington's disease. J Med Genet 10:333–336

Chase GA, Markson LE, Brookmeyer R, Folstein SE (1986) Covariate dependent genetic counseling in Huntington's disease. J Neurogenet 3:215–223

Cupples LA, Terrin NC, Myers RH, D'Agostini RB (1989) Using survival methods to estimate age-at-onset distributions for genetic diseases with an application to Huntington disease. Genet Epidemiol 6:361–371

Gilliam TC, Bucan M, MacDonald ME, Zimmer M, Haines JL, Cheng SV, Pohl TM, et al (1987) A DNA segment encoding two genes very tightly linked to Huntington's disease. Science 238:950–952

Gusella JF, Wexler NS, Conneally PM, Naylor SL, Anderson MA, Tanzi RE, Watkins RC, et al (1983) A polymorphic DNA marker linked to Huntington's disease. Nature 306:234–238

Hayden M (1981) Huntington's chorea. Springer, Berlin

Kalbfleisch JD, Prentice RL (1980) The statistical analysis of failure time data. Wiley, New York

MacDonald ME, Cheng SV, Zimmer M, Haines JL, Poustka A, Alitto B, Smith B, et al (1989) Clustering of multiallele DNA markers near the Huntington's disease gene. J Clin Invest 84:1013–1016

Myers RH, Cupples LA, Schoenfeld M, D'Agostino RB, Terrin NC, Goldmakher N, Wolf PA (1985) Maternal factors in onset of Huntington disease. Am J Hum Genet 37:511–523

Myers RH, Madden JJ, Teague JL, Falek A (1982) Factors related to onset age in Huntington's disease. Am J Hum Genet 34:481–488

Nelder JA, Mead R (1964) A simplex method for function minimization. Comput J 7:308–313

Newcombe RG (1981) A life table for onset of Huntington's chorea. Ann Hum Genet 45:375–385

Newcombe RG, Walker DA, Harper PS (1981) Factors influencing age at onset and duration of survival in Huntington's chorea. Ann Hum Genet 45:387–396

Paolo J, Somer H, Ikonen E, Karila L, Peltonen L (1987) Low prevalence of Huntington's disease in Finland. Lancet 2:805–806

Pfanzagl J (1974) Allgemeine Methodenlehre der Statistik, vol 2. de Gruyter, Berlin

Reed TE, Chandler JH (1958) Huntington's chorea in Michigan. I. Demography and genetics. Am J Hum Genet 10: 201–225

Ridley RM, Frith CD, Crow TJ, Conneally PM (1988) Anticipation in Huntington's disease is inherited through the male line but may originate in the female. J Med Genet 25:589–595

Shapiro SS, Wilk MB (1965) An analysis of variance test for normality (complete samples). Biometrika 52:591–611

Silvey SP (1987) Statistical inference. Chapman & Hall, Cambridge

Smith B, Skarecky D, Bengtsson U, Magenis RE, Carpenter

N, Wasmuth JJ (1988) Isolation of DNA markers in the direction of the Huntington disease gene from the G8 locus. Am J Hum Genet 42:335–344

Wasmuth JJ, Hewitt J, Smith B, Allard D, Haines JL, Skarecky D, Partlow E, et al (1988) A highly polymorphic locus very tightly linked to the Huntington's disease gene. Nature 332:734–736

Wendt VGG (1959) Das Erkankungsalter bei der Huntingtonschen Chorea. Acta Genet 9:18–31

Wolff G, Deutschl G, Wienker TF, Hummel K, Bender K, Lücking CH, Schumacher M, et al (1989) New mutation to Huntington's disease. J Med Genet 26:18–27