Research Article

Predicting Disease Onset from Mutation Status Using Proband and Relative Data with Applications to Huntington's Disease

Tianle Chen,[1] Yuanjia Wang,[1] Yanyuan Ma,[2] Karen Marder,[3] and Douglas R. Langbehn[4]

[1] *Department of Biostatistics, Mailman School of Public Health, Columbia University, 722 West 168th Street, New York, NY 10032, USA*
[2] *Department of Statistics, Texas A&M University, College Station, TX 77843, USA*
[3] *Departments of Neurology and Psychiatry and Sergievsky Center and the Taub Institute, Columbia University Medical Center, New York, NY 10032, USA*
[4] *Department of Psychiatry and Biostatistics (Secondary), University of Iowa, Iowa City, IA 52242, USA*

Correspondence should be addressed to Yuanjia Wang, yw2016@columbia.edu

Huntington's disease (HD) is a progressive neurodegenerative disorder caused by an expansion of CAG repeats in the IT15 gene. The age-at-onset (AAO) of HD is inversely related to the CAG repeat length and the minimum length thought to cause HD is 36. Accurate estimation of the AAO distribution based on CAG repeat length is important for genetic counseling and the design of clinical trials. In the Cooperative Huntington's Observational Research Trial (COHORT) study, the CAG repeat length is known for the proband participants. However, whether a family member shares the huntingtin gene status (CAG expanded or not) with the proband is unknown. In this work, we use the expectation-maximization (EM) algorithm to handle the missing huntingtin gene information in first-degree family members in COHORT, assuming that a family member has the same CAG length as the proband if the family member carries a huntingtin gene mutation. We perform simulation studies to examine performance of the proposed method and apply the methods to analyze COHORT proband and family combined data. Our analyses reveal that the estimated cumulative risk of HD symptom onset obtained from the combined data is slightly lower than the risk estimated from the proband data alone.

1. Introduction

Huntington's disease (HD) is a severe, autosomal dominantly inherited neurodegenerative disorder that affects motor, cognitive, and psychiatric function and is uniformly fatal. HD is caused by the expansion of CAG trinucleotide repeats at the huntingtin gene (IT15)

[1, 2]. Affected individuals typically begin to show motor signs around 30–50 years of age and typically die 15–20 years after the disease onset [3]. Despite identification of the causative gene, there is currently no treatment that modifies disease progression.

One large genetic epidemiological study of HD, the Cooperative Huntington's Observational Research Trial (COHORT), including 42 Huntington study group research centers in North America and Australia, was initiated in 2005 and concluded in 2011 [4–6]. Participants in COHORT (probands) underwent a clinical evaluation and DNA from whole blood was genotyped for the length of the CAG-repeat huntingtin mutation. Since 2005, COHORT probands from sites with IRB approval have participated in family history interviews and have provided information on HD affection status in their family members. While CAG repeat length is ascertained in probands, the high cost of conducting in-person interviews of family members prevents the collection of all family members' blood samples. However, family members' age-at-onset (AAO) of HD and vital status are obtained through systematic interviews of the probands or the family members themselves. Although a relative's HD genotype is unavailable, the corresponding distribution of the HD gene can be estimated based on the relative's relationship with the proband, the proband's mutation status, and assumptions regarding within-family similarity of CAG length [7, 8].

In a genetic counseling setting, subjects with CAG repeats of 36 or greater are defined as carrying the HD mutation (carrier; [9]), and CAG less than 36 is defined as screened negative, or noncarrier [9]. It is known that there is an inverse association between the CAG repeat length and AAO of HD, that is, the longer the repeat length, the earlier the motor onset [10]. Modeling such a relationship as well as the conditional distribution of HD onset given CAG repeat length accurately and precisely is important for genetic counseling and the design of clinical trials for HD. The AAO of HD onset is subject to right censoring by constraints of the observation periods. Carriers who have not been diagnosed with HD are right-censored for AAO. Several formulae were proposed in the literature to estimate the survival function of age at HD diagnosis given CAG repeat length (e.g., [9–11]). Langbehn et al. [10] have shown that the standard semiparametric survival models, such as the Cox proportional hazards model, do not fit the HD data and proposed a new logistic-exponential parametric model. Specifically, the conditional distribution of HD onset given the CAG repeat length is modeled as a logistic function, with a location and a scale parameter both depending on CAG through nonlinear relationships. Using a large clinical data set, they observed that separate exponential relationships with CAG length gave excellent empirical goodness of fit to both the mean AAO and its variance. Other parametric models, such as Gamma distribution, have also been proposed in the literature [12, 13]. Langbehn et al. [14] examine several AAO models in the literature and show the superior performance of Langbehn et al. [10] in terms of predicting the two-year probability of new HD diagnosis with independent prospective data.

None of the aforementioned existing methods can be directly used to analyze COHORT family data because family members are not always genotyped and their HD mutation status is unknown. The inclusion of family data contributes additional information; however, the unobserved HD mutation sharing status in family members (CAG-elongated or not) complicates the analysis. To see this, note that the affected parent carrying huntingtin mutation has a 50% chance of transmitting the mutation to an offspring. An added complexity is that the likelihood of the offspring having a higher CAG repeat than the parent is higher if the parent is the father. Since the offspring is not genotyped, whether he or she carries expanded CAG repeats is unknown. In this work, we treat the unknown huntingtin gene sharing status in first-degree family members (CAG-elongated or not) as missing data and

use the EM algorithm to carry out the maximum likelihood estimation of the proband and family data jointly. Conditionally on the transmission status in family members, we use the logistic-exponential model in Langbehn et al. [14] to model the AAO as a function of CAG repeat length. We perform simulation studies to examine finite sample performances of the proposed methods. Finally, we apply these methods to analyze the COHORT proband and family combined data. Our results show a slightly lower estimated cumulative risk of HD symptom onset using the combined data compared to using proband data alone.

2. Methods

We start by introducing some notations. For the ith subject, let T_i denote the age-at-onset of HD, let δ_i be the event indicator, let C_i denote the censoring time, and let $X_i = \min(T_i, C_i)$. Let A_i denote the CAG repeat length. Langbehn et al. [10] model distribution of T_i given A_i by a logistic function. The cumulative distribution function (CDF) given A_i is

$$F(t \mid A_i) = \Pr(T_i \leq t \mid A_i) = \frac{1}{1 + e^{-[t-\mu(A_i)]/s(A_i)}}, \tag{2.1}$$

and the density function is

$$f(t \mid A_i) = \frac{e^{-[t-\mu(A_i)]/s(A_i)}}{s(A_i)\{1 + e^{-[t-\mu(A_i)]/s(A_i)}\}^2}. \tag{2.2}$$

Here $\mu(A_i)$ is a location parameter depending on the covariate A_i and $s(A_i)$ is a scale parameter depending on A_i. Let $S(t \mid A_i) = 1 - F(t \mid A_i)$ denote the survival function of HD onset. The location and scale parameters have the following relationship with the mean and variance of T_i given A_i:

$$E(T_i A_i) = \mu(A_i), \qquad \text{var}(T_i A_i) = \pi^2 3 s^2(A_i). \tag{2.3}$$

Various parametric functions for the location and scale parameters were compared in Langbehn et al. [10, 14], and the exponential function provides the best fit. Therefore, we use the same model where

$$\mu(A_i) = \mu_1 + \exp(\mu_2 - \mu_3 A_i),$$
$$\text{var}(A_i) = \sigma_1 + \exp(\sigma_2 - \sigma_3 A_i). \tag{2.4}$$

Substitute these into $F(t \mid A_i)$ and $f(t \mid A_i)$ to obtain a parametric model for the distribution of AAO of HD with six parameters, $\beta = (\mu_1, \mu_2, \mu_3, \sigma_1, \sigma_2, \sigma_3)^T$. Langbehn et al. [10] fitted estimates of $\beta = (21.54, 9.56, 0.146, 35.55, 17.72, 0.327)^T$.

2.1. Proband-Only Analysis

First, consider probands' data where all A_i's are observed. Since a subject's AAO of HD is subject to the right censoring, the likelihood function is

$$L(\beta) = \prod_{i=1}^{n} f^{\delta_i}(X_i \mid A_i; \beta) S^{1-\delta_i}(X_i \mid A_i; \beta), \tag{2.5}$$

and the log-likelihood is

$$l(\beta) = \sum_{i=1}^{n} \left\{ -\delta_i \log[s(A_i)] - \frac{X_i - \mu(A_i)}{s(A_i)} - (1 + \delta_i) \log\left[1 + e^{-(X_i - \mu(A_i))/s(A_i)}\right] \right\}. \tag{2.6}$$

The maximum likelihood estimate (MLE) of the parameters, $\widehat{\beta}$, can be obtained via a general-purpose optimization algorithm such as Newton-Raphson or Nelder-Mead implemented in the R program version 2.13.1. The variance-covariance matrix of $\widehat{\beta}$ is estimated by the inverse of the estimated Hessian matrix

$$\widehat{\mathrm{cov}}\left(\widehat{\beta}\right) = \left[H\left(\widehat{\beta}\right)\right]^{-1}. \tag{2.7}$$

The standard error of the estimated survival function, $\widehat{S}(t \mid A_i)$, is then estimated by the Delta method, that is,

$$\widehat{\mathrm{var}}\left[\widehat{S}(t \mid A_i)\right] = G^T\left(\widehat{\beta}\right) \widehat{\mathrm{var}}\left(\widehat{\beta}\right) G\left(\widehat{\beta}\right), \tag{2.8}$$

where the gradient vector

$$G\left(\widehat{\beta}\right) = \left.\frac{\partial S(t \mid A_i)}{\partial \beta}\right|_{\beta=\widehat{\beta}}. \tag{2.9}$$

Since the parameters are estimated by maximum likelihood, it is straightforward to carry out likelihood ratio tests (LRTs) to compare the model fit from the COHORT data with the one obtained by applying parameters from other studies such as Langbehn et al. [10] to the COHORT data. Here, twice the difference in the log-likelihood follows an asymptotic chi-square distribution with 6 degrees of freedom.

2.2. Incorporating Family Members

Next, we consider incorporating family members' AAO data. We do not directly observe whether a family member shares the huntingtin mutation with the proband, but we do have data regarding family members' age-at-onset of the first symptoms, as well as the family members' current ages. When we incorporate the additional family data, the likelihood for the survival takes a mixture form. Let p_i denote the probability of the ith subject sharing

a deleterious allele with a proband and therefore becoming a carrier. Such probabilities are calculated based on Mendelian transmission and a family member's relationship to the proband [8]. For example, offspring and siblings of a carrier proband have a probability of 50% of receiving the huntingtin allele that contains the CAG expansion (Homozygotes for HD are extremely rare since prevalence of HD in general population is rare). We assume that, conditioning on a family member receiving the expanded huntingtin allele, the CAG repeat length is the same as observed in the proband, although this is a simplification [7]. For subjects who receive a wild-type allele (CAG < 36), their probability of developing HD is zero, thus $f(t \mid A_i < 36) = 0$, and $S(tA_i < 36) = 1$, for all t. For the family members, the likelihood is

$$L(\beta) = \prod_{i=1}^{n} \left[p_i f^{\delta_i}(X_i \mid A_i; \beta) S^{1-\delta_i}(X_i \mid A_i; \beta) + (1 - p_i)(1 - \delta_i) \right], \tag{2.10}$$

where the above second term follows from the assumption that noncarriers do not develop HD. Note that for all carrier probands we observe $p_i = 1$, thus the likelihood reduces to (2.5).

The above likelihood can be maximized by a combination of EM and Newton-Raphson algorithms. Let G_i denote the unobserved carrier status indicator for the ith family member (i.e., $G_i = 1$ indicates a family member receives a mutation and $G_i = 0$ indicates otherwise). Then the complete data log-likelihood is

$$\sum_{i=1}^{n} I(G_i = 1) \{ \delta_i \log[f(X_i \mid A_i; \beta)] + (1 - \delta_i) \log[S(X_i \mid A_i; \beta)] \}. \tag{2.11}$$

At the $(k+1)$th iteration of the E-step, we compute the conditional expectation of the complete data log-likelihood, given the observed data. Essentially, we compute

$$\begin{aligned} w_i^{(k+1)} &= E\left[I(G_i = 1) \mid X_i, \delta_i, \beta^{(k)} \right] \\ &= \frac{p_i f^{\delta_i}(X_i \mid A_i; \beta^{(k)}) S^{1-\delta_i}(X_i \mid A_i; \beta^{(k)})}{p_i f^{\delta_i}(X_i \mid A_i; \beta^{(k)}) S^{1-\delta_i}(X_i \mid A_i; \beta^{(k)}) + (1 - p_i)(1 - \delta_i)}. \end{aligned} \tag{2.12}$$

In the M-step, we update $\beta^{(k+1)}$ by maximizing the weighted log-likelihood

$$\sum_{i=1}^{n} w_i^{(k+1)} \{ \delta_i \log[f(X_i \mid A_i; \beta)] + (1 - \delta_i) \log[S(X_i \mid A_i; \beta)] \} \tag{2.13}$$

using the Newton-Raphson algorithm developed for the proband data.

Since for the combined analysis, the parameters are estimated by maximizing the likelihood through an EM algorithm, the standard asymptotic theory applies and the standard errors of parameters can be estimated by inverting the expected or observed information matrix based on the log-likelihood of the observed data. When there is missing data and an EM algorithm is used to obtain the MLE, the information matrix based on the observed data likelihood can be difficult to compute analytically or computationally. In such situations, Louis [15] proposed to compute the observed information matrix in terms of the conditional

moments of the first and second derivatives of the complete data log likelihood which can be obtained easily under the EM algorithm framework. In some cases, these moments are easier to compute than the corresponding derivatives of the incomplete, observed data log-likelihood.

However, in our application, the derivatives of the observed data log likelihood are easy to compute. Thus, we computed the gradient and Hessian matrix of the observed data log-likelihood directly and estimated the standard errors of $\hat{\beta}$ by the inverse of the Hessian matrix and estimated the standard errors of $\hat{F}(t)$ by the Delta method similar to the proband-only analysis. Simulation studies in the next section show satisfactory performance of this direct and relatively simpler approach.

3. Simulation Studies

We conducted two simulation studies closely related to the observed COHORT data to illustrate the performance of the Newton-Raphson optimization and the EM algorithm [16]. In all our optimization procedures, we centered both A_i and X_i. Since the direct optimization and EM algorithm need reasonable initial values, we fitted two nonlinear least square (NLS) to the observed sample mean and variance of the AAO on subjects with $\delta_i = 1$. To be specific, we fit

$$m_1(a_i) = \mu_1 + \exp(\mu_2 - \mu_3 a_i), \qquad s_1^2(a_i) = \sigma_1 + \exp(\sigma_2 - \sigma_3 a_i), \tag{3.1}$$

where $m_1(a_i)$ and $s_1^2(a_i)$ are the sample mean and variance for all subjects with $A_i = a_i$, respectively. The six NLS estimators were used as the initial values for further optimization. We denoted the estimated β from the centered data as $\hat{\beta}_c$. For each simulation, the uncentered $\hat{\beta}$ were then calculated based on $\hat{\beta}_c$ and the sample mean of A_i and X_i.

We restricted simulations to CAG repeat lengths between 41 and 56 to guard against sensitivity to the extremely high or low CAG repeats to be consistent with Langbehn et al. [10]. For the analysis of proband data, we generated a sample of 2000 subjects, each with a CAG length ranging from 41 to 56 that follows a multinomial distribution in which the probability $\mathrm{pr}(A_i = a)$ equals to the observed proportion of $A_i = a$ in the COHORT proband data set. The failure times T_i were simulated from the distribution (2.1), where the parameters β were fixed at the values fitted from the COHORT proband data (see next section for their values). The censoring times, C_i, were generated from a rescaled Beta distribution with a scale and shape parameter of four. The parameters for the Beta distribution were chosen so that the proportion of censored subjects is the same in the simulated data and the observed COHORT proband data.

For the analysis of the combined proband and family data, we generated a sample of 4000 subjects. We assume the same proportion of the probands and relatives as observed in the combined COHORT data. For the family members, the probabilities p_i were generated by resampling the observed p_i's in the COHORT data. With a given p_i for each subject, we simulated his or her huntingtin carrier status from a Bernoulli distribution with success probability p_i. For family members simulated to receive an expanded CAG repeat (carriers), their CAG repeats A_i were set to be the same as the probands and their failure times were simulated from (2.5) with β fixed at estimates from the COHORT combined data. For
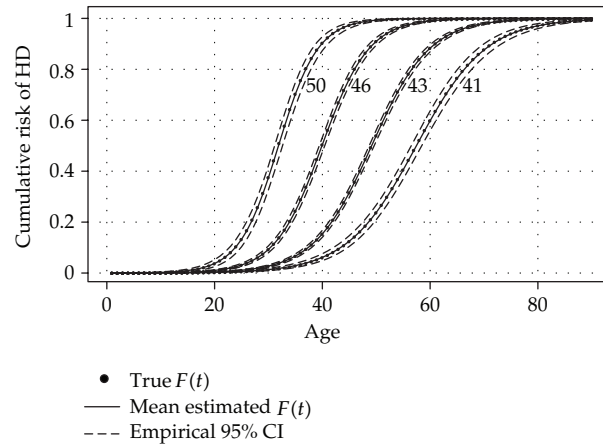
Figure 1: Estimated CDF of HD onset for $A_i = 41, 43, 46$, and 50 with simulated proband data.
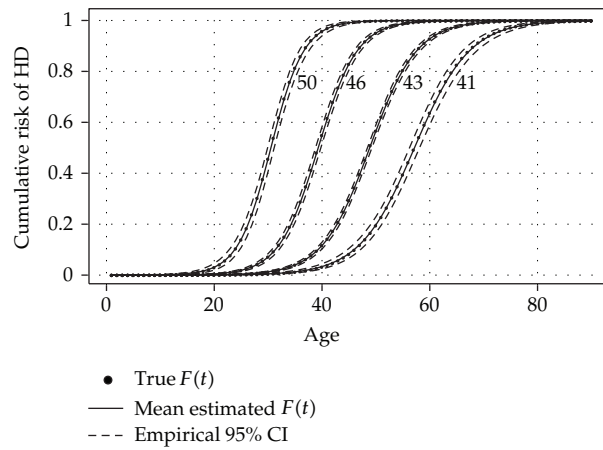

Figure 2: Estimated CDF of HD onset for $A_i = 41, 43, 46$, and 50 with simulated combined proband and relative data.

noncarrier family members, their failure times were set to be infinity and their $X_i = C_i$. We used the same censoring distribution for generating C_i as in the first simulation study.

We provide simulation results of the proband only and combined analyses in Tables 1 and 2. We present mean $\widehat{F}(t \mid A_i)$, empirical standard deviation of $\widehat{F}(t \mid A_i)$, and the mean estimated standard error of $\widehat{F}(t \mid A_i)$ at various ages in. We see from these tables that mean $\widehat{F}(t \mid A_i)$ is very close to true $F(t \mid A_i)$ in both studies. The mean estimated standard errors of $\widehat{F}(t \mid A_i)$ are close to the empirical standard deviations, indicating that the estimation of variability is appropriate. Figures 1 and 2 present three curves of $\widehat{F}(t \mid A_i)$ at $A_i = 41, 46, 50$ and their 95% empirical confidence intervals for the proband data and combined data, respectively. We see that $\widehat{F}(t \mid A_i)$ coincide with the circles representing true $F(t \mid A_i)$ at various ages.

Table 1: Simulation 1 (proband data). Estimated CDF and standard errors from the direct optimization of proband-only analysis, $n = 2000$, 1000 replications.

Age	CAG = 41				CAG = 46				CAG = 50			
	$F(t\mid A_i)$	Mean $\hat{F}(t\mid A_i)$	Empi: sd	Mean $\widehat{sd}$	$F(t\mid A_i)$	Mean $\hat{F}(t\mid A_i)$	Empi: sd	Mean $\widehat{sd}$	$F(t\mid A_i)$	Mean $\hat{F}(t\mid A_i)$	Empi: sd	Mean $\widehat{sd}$
10	0.0001	0.0001	0.0000	0.0000	0.0003	0.0003	0.0001	0.0001	0.0011	0.0012	0.0005	0.0004
20	0.0006	0.0007	0.0002	0.0002	0.0049	0.0048	0.0008	0.0009	0.0301	0.0309	0.0066	0.0060
30	0.0046	0.0049	0.0011	0.0012	0.0717	0.0709	0.0066	0.0068	0.4560	0.4578	0.0253	0.0248
40	0.0322	0.0335	0.0051	0.0054	0.5492	0.5487	0.0171	0.0162	0.9577	0.9572	0.0084	0.0077
50	0.1944	0.1972	0.0162	0.0160	0.9505	0.9509	0.0052	0.0056	0.9984	0.9983	0.0007	0.0006
60	0.6368	0.6358	0.0227	0.0219	0.9967	0.9967	0.0006	0.0007	0.9999	0.9999	0.0000	0.0000
70	0.9272	0.9252	0.0102	0.0108	0.9998	0.9998	0.0001	0.0001	1.0000	1.0000	0.0000	0.0000
80	0.9893	0.9887	0.0025	0.0026	1.0000	1.0000	0.0000	0.0000	1.0000	1.0000	0.0000	0.0000
90	0.9985	0.9984	0.0005	0.0005	1.0000	1.0000	0.0000	0.0000	1.0000	1.0000	0.0000	0.0000

Table 2: Simulation 2 (combined proband and relative data). Estimated CDF and standard errors from the EM algorithm with combined proband and family analysis, $n = 4000$, 1000 replications.

Age	CAG = 41				CAG = 46				CAG = 50			
	$F(t\mid A_i)$	Mean $\widehat{F}(t\mid A_i)$	Empi: sd	Mean $\widehat{sd}$	$F(t\mid A_i)$	Mean $\widehat{F}(t\mid A_i)$	Empi: sd	Mean $\widehat{sd}$	$F(t\mid A_i)$	Mean $\widehat{F}(t\mid A_i)$	Empi: sd	Mean $\widehat{sd}$
10	0.0006	0.0006	0.0002	0.0002	0.0010	0.0010	0.0002	0.0002	0.0025	0.0026	0.0008	0.0008
20	0.0028	0.0029	0.0007	0.0007	0.0102	0.0102	0.0014	0.0014	0.0373	0.0374	0.0069	0.0068
30	0.0134	0.0137	0.0023	0.0023	0.0928	0.0928	0.0069	0.0070	0.3754	0.3751	0.0241	0.0238
40	0.0609	0.0616	0.0069	0.0069	0.5041	0.5042	0.0148	0.0143	0.9031	0.9030	0.0139	0.0132
50	0.2373	0.2378	0.0149	0.0146	0.9099	0.9100	0.0076	0.0074	0.9931	0.9930	0.0020	0.0019
60	0.5987	0.5979	0.0200	0.0188	0.9901	0.9901	0.0015	0.0014	0.9996	0.9995	0.0002	0.0002
70	0.8773	0.8761	0.0133	0.0125	0.9990	0.9990	0.0002	0.0002	1.0000	1.0000	0.0000	0.0000
80	0.9717	0.9710	0.0050	0.0047	0.9999	0.9999	0.0000	0.0000	1.0000	1.0000	0.0000	0.0000
90	0.9940	0.9937	0.0015	0.0014	1.0000	1.0000	0.0000	0.0000	1.0000	1.0000	0.0000	0.0000

4. COHORT Data Analysis Results

COHORT is a multicenter observational study of individuals in the HD community. COHORT recruitment is open to subjects who have HD symptoms and signs (manifest HD), subjects who have an expanded CAG repeat but have not yet developed symptoms of HD (presymptomatic), subjects who have an HD affected parent but have not been tested and do not have symptoms (at risk), subjects who have an affected grandparent (secondary risk), and control subjects who are not at risk for HD. Information available on participating probands include genetic status (whether or not they carry HD mutation, and the number of CAG repeats), clinical diagnosis of HD, and the timing of symptom onset and timing of diagnosis. In our analyses, only probands with expanded CAG (CAG ≥ 36) and their family members were included. Details of the cohort are cited in a publication in press [6].

We first describe the proband and family data in the COHORT study. Information on CAG repeat length and age was available for 1357 probands with CAG repeats varying from 36 to 100 (Table 3). There were 3409 first-degree relatives available from 675 probands. We do not have information on whether some of the probands are from the same family. We show the descriptive statistics for the relatives stratified by relationship type in Table 4. Each proband potentially has three versions of age-at-the-first-symptom (rater's report, subject's self-report, and a family member's report). We gave the rater reported AAO of symptom the highest priority. If the rater reported version is not available, we then used subject report. If neither rater nor subject's self-report is available, we then used the family member's report. Twenty-one subjects whose self-reported and rater-reported AAO of symptom differed by greater than 15 years were removed. Our proband data set has 1151 subjects with CAG length between 41 and 56 and was used for the proband-only analysis. Similar to Langbehn et al. [10], we restricted the analysis to CAG repeat lengths between 41 and 56 to guard against sensitivity to the extremely high or low CAG repeats and against bias due to likely under ascertainment (relative to the population) of subjects with CAG length between 36 and 40.

Information on CAG repeat length, age at time of evaluation and the probability of being a carrier (receiving huntingtin mutation from the proband) was available for 2851 family members of 1151 probands. In the proband data set, both individuals with manifest HD and presymptomatic carriers (24%) are included. Their age-at-diagnosis and age-at-first-motor sign were recorded. Among 1151 probands, 876 (76%) subjects had experienced HD onset and the average AAO of the HD diagnosis was 44 years of age (standard deviation: 10.7). There were 54% females and 94% Caucasians. Our combined proband and family data set has 4002 subjects. In this combined data set, 51% were females and 35% subjects had experienced HD onset. Among the 4002 subjects, 467 are singletons (probands with no family member included). The other 3535 subjects belong to 623 pedigrees with an average size of 5.674 (sd = 2.609) members. In the combined data, there are two different probabilities of being a carrier: $p_i = 1$ (1199 subjects with known CAG expansions or known HD onset) or $p_i = 0.5$ (2803 subjects). Among the 2851 family members, 966 are parents of the probands, 1095 are siblings of the probands, and 790 are children of the probands.

When using the age-at-diagnosis in our proband data as T_i, the estimated cumulative risk of HD is

$$\widehat{F}(t \mid A_i) = \left(1 + \exp\left\{ -\frac{\pi}{\sqrt{3}} \frac{[t - 16.284 - \exp(8.325 - 0.111A_i)]}{\sqrt{22.379 + \exp(15.657 - 0.284A_i)}} \right\} \right)^{-1}. \tag{4.1}$$

Table 3: Descriptive statistics of the COHORT proband data.

		Numbers and ages for a CAG repeat length																						Total
		36	37	38	39	40	41	42	43	44	45	46	47	48	49	50	51	52	53	54	55	56	57+	
At risk	Number	2	5	15	21	43	55	68	57	31	28	18	9	5	2	1	0	2	0	0	0	0	0	362
	Ave age	61	64	48	55	50	45	42	39	37	31	34	34	27	23	30		35						42
	Min age	60	61	26	37	25	21	21	18	19	19	20	21	20	21	30		23						18
	Max age	62	69	66	70	88	67	71	62	51	44	51	53	40	25	30		47						88
	sd	1	3	11	9	14	11	11	10	9	7	9	9	9	3	.		17						13
	%	0.6	1.4	4.1	5.8	11.9	15.2	18.8	15.7	8.6	7.7	5.0	2.5	1.4	0.6	0.3	0.0	0.6	0.0	0.0	0.0	0.0	0.0	
Affected	Number	2	1	6	7	67	128	148	144	143	93	83	47	34	21	18	9	7	10	6	3	3	15	995
	Ave age	54	68	55	53	60	55	51	48	44	41	38	36	33	31	31	30	28	23	26	26	23	20	45
	Min age	49	68	46	25	37	28	17	19	21	16	25	21	20	19	22	23	22	11	18	25	17	12	11
	Max age	59	68	67	77	82	76	76	67	67	58	53	48	46	44	39	35	35	29	31	29	28	27	82
	sd	7	.	7	19	10	9	9	8	8	8	6	6	6	6	5	4	5	6	5	2	6	4	12
	%	0.2	0.1	0.6	0.7	6.7	12.9	14.9	14.5	14.4	9.3	8.3	4.7	3.4	2.1	1.8	0.9	0.7	1.0	0.6	0.3	0.3	1.5	
Total	Number	4	6	21	28	110	183	216	201	174	121	101	56	39	23	19	9	9	10	6	3	3	15	1357

Table 4: Descriptive statistics of the first-degree relatives of COHORT proband subjects stratified by relationship.

		Relationship			
		Parents	Siblings	Children	Total
Not affected	Number	739	1110	931	2780
	Ave age	70	50	26	42
	Min age	27	0	0	18
	Max age	111	93	62	88
	sd	13	15	14	13
	%	26.6	39.9	33.5	
Affected	Number	379	237	13	629
	Ave age	45	42	36	45
	Min age	18	7	23	11
	Max age	82	70	44	82
	sd	11	11	7	12
	%	60.3	37.7	2.1	
Total	Number	1118	1347	944	3409

Table 5: Mean and standard deviation of the AAO estimated from the model (2.1) for four analyses.

	Langbehn data		COHORT data					
			Probands diagnosis*		Probands symptom**		Combined symptom†	
CAG	Mean	SD	Mean	SD	Mean	SD	Mean	SD
41	57.06	10.50	59.84	8.78	57.74	9.13	59.33	11.68
43	48.06	8.62	51.17	7.31	49.32	7.90	50.63	9.60
46	38.66	7.08	41.29	5.97	39.66	6.57	41.20	7.59
48	34.32	6.57	36.31	5.47	34.75	5.95	36.69	6.79
50	31.08	6.28	32.32	5.16	30.80	5.50	33.21	6.28

* : using proband age-at-diagnosis data;
**: using proband age-at-first-symptom data;
†: using proband and relative combined age-at-first-symptom data.

The estimated parameters for the CDF from the proband-only analysis are slightly different from the ones obtained from Langbehn et al. [10]. Our estimated mean and standard deviation of the AAO of HD is about 1 to 3 years later than the ones obtained in Langbehn et al. [10], and the standard deviation (SD) is slightly smaller (Table 5). In addition, the estimated CDF is smaller for most A_i values using COHORT data. We ran a joint likelihood ratio test on the goodness-of-fit of parameters obtained in Langbehn et al. [10] and the P value was less than 0.001 (test statistic = 66.0). When analyzing the age-at-first-symptom in our proband data, the estimated cumulative risk of HD is

$$\widehat{F}(t \mid A_i) = \left(1 + \exp\left\{-\frac{\pi}{\sqrt{3}} \frac{[t - 14.266 - \exp(7.987 - 0.104 A_i)]}{\sqrt{28.933 + \exp(17.130 - 0.312 A_i)}}\right\}\right)^{-1}. \tag{4.2}$$

We present $\widehat{F}(t \mid A_i)$ curves for age-at-diagnosis and age-at-symptom at various CAG lengths and their 95% confidence intervals for the proband data in Figure 3. It can be seen that with a given A_i, the estimated probability of having the first symptoms of HD is higher than
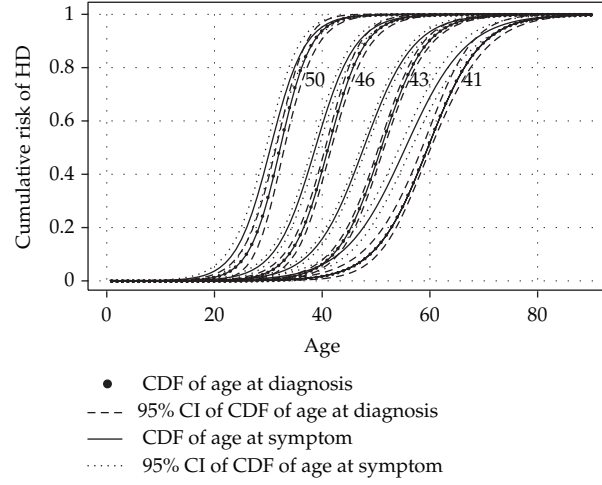
Figure 3: Estimated CDFs of age-at-diagnosis and age-at-first-symptom of HD for $A_i = 41, 43, 46$, and 50 with COHORT proband data.
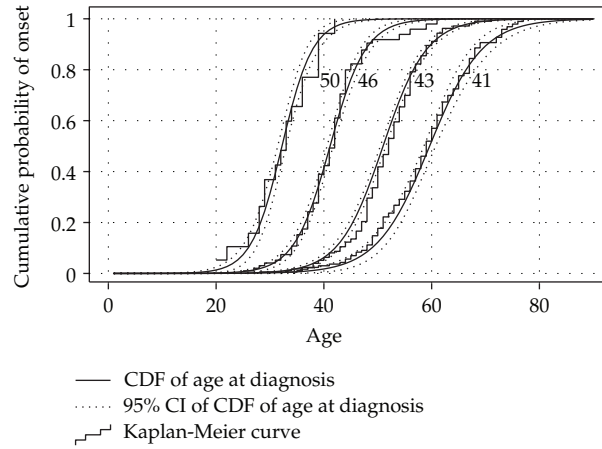

Figure 4: Kaplan-Meier curve and estimated CDF of age-at-diagnosis of HD for $A_i = 41, 43, 46$, and 50 with COHORT proband data.

the probability of a diagnosis of HD at the same age. This is consistent with the intuition that symptoms of HD will be observed before a diagnosis. The mean AAO of first symptom is estimated to be about 2 years earlier than AAO of diagnosis (Table 5) and the standard deviation of the former is slightly larger, indicating that reported age-at-first-symptom is more variable. It is unclear to what extent this difference represents true physical variability in illness development versus possibly lower reliability in the retrospective reporting of symptom onset [17].

As a sensitivity analysis, we compared the estimated CDF based on the parametric model with a nonparametric Kaplan-Meier estimator for subjects with a given A_i. Figure 4 presents this comparison using probands' age-at-diagnosis data. We show in the figure that the parametric model fit is consistent with the Kaplan-Meier fit. However, as expected, the confidence interval for the parametric model estimate at a given age is narrower than
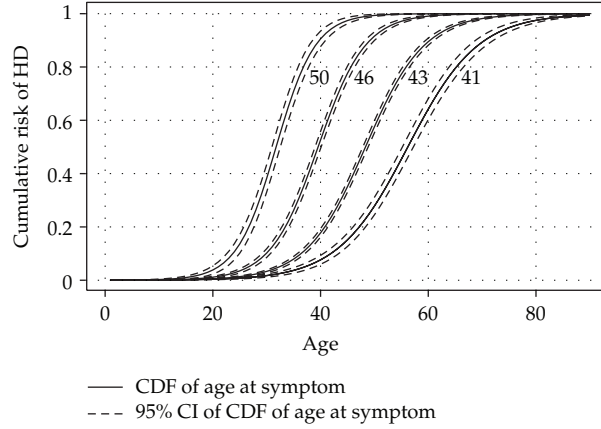
Figure 5: Estimated CDF of age-at-first-symptom of HD for $A_i = 41, 43, 46,$ and 50 with COHORT combined proband and relative data.

the Kaplan-Meier estimate (results not shown). The figure comparing age-at-symptom models is similar and therefore omitted.

We reanalyzed only the AAO of the first symptom using the combined proband and family data, since the age-at-diagnosis was not available for family members who are not seen in person. The estimated cumulative risk of HD at age t is

$$\widehat{F}(t \mid A_i) = \left(1 + \exp\left\{-\frac{\pi}{\sqrt{3}} \frac{[t - 18.832 - \exp(8.461 - 0.118 A_i)]}{\sqrt{32.365 + \exp(14.823 - 0.248 A_i)}}\right\}\right)^{-1}. \tag{4.3}$$

The corresponding $\widehat{F}(t \mid A_i)$ curves at various CAG lengths and their 95% confidence intervals are shown in Figure 5. In Table 5, we compare the estimated mean and SD of the AAO from the proband and combined data. We can see that the estimated mean AAOs for several CAGs are similar regardless of whether family members are included. The SD estimated from the model is larger for the combined data. This is a reflection of the observed data in that there is a wider range of AAO in the combined data than in the proband data. For example, the SD for CAG = 41 of the former is 11 years, whereas it is 10 years in the probands, and the SD for CAG = 42 is 10 in the combined and 8 in the probands.

One of the utilities of the estimated curves is to estimate the conditional probability of having an HD onset (or staying HD free) in the next five or ten years, given a subject has not had an onset by a given age. Similar to Langbehn et al. [10], in Table 6, we present such conditional probabilities in five-year intervals for a subject without HD at age 40 and with given CAG repeats. For example, a 40-year presymptomatic subject with a CAG of 42 has a probability of 34% (CI: 32%, 36%) for developing HD in the next 10 years (by age 50), while for a subject with a CAG of 50 this probability increases to 0.93 (CI: 0.91, 0.95).

5. Discussion

We propose methods to predict disease risk from a known mutation (or to estimate the penetrance function). For most complex diseases, predicting the AAO of a disease

Table 6: Conditional survival probabilities estimated from the COHORT combined data.

CAG	45 years	50 years	55 years	60 years	65 years	70 years
36	0.01 (0.00, 0.02)	0.02 (0.00, 0.04)	0.04 (0.00, 0.08)	0.07 (0.01, 0.13)	0.11 (0.05, 0.20)	0.17 (0.07, 0.28)
37	0.01 (0.00, 0.02)	0.03 (0.01, 0.06)	0.06 (0.02, 0.11)	0.11 (0.05, 0.18)	0.18 (0.11, 0.27)	0.28 (0.17, 0.39)
38	0.02 (0.01, 0.03)	0.05 (0.02, 0.08)	0.10 (0.06, 0.15)	0.18 (0.12, 0.25)	0.29 (0.21, 0.38)	0.43 (0.33, 0.53)
39	0.03 (0.02, 0.04)	0.08 (0.05, 0.11)	0.17 (0.12, 0.21)	0.29 (0.23, 0.35)	0.44 (0.36, 0.52)	0.60 (0.52, 0.69)
40	0.05 (0.04, 0.06)	0.14 (0.11, 0.16)	0.27 (0.23, 0.31)	0.44 (0.39, 0.50)	0.62 (0.55, 0.68)	0.77 (0.72, 0.82)
41	0.08 (0.07, 0.09)	0.22 (0.19, 0.24)	0.41 (0.37, 0.44)	0.61 (0.57, 0.65)	0.78 (0.74, 0.81)	0.88 (0.86, 0.91)
42	0.13 (0.12, 0.14)	0.34 (0.32, 0.36)	0.57 (0.54, 0.60)	0.77 (0.74, 0.79)	0.89 (0.87, 0.90)	0.95 (0.94, 0.96)
43	0.21 (0.20, 0.22)	0.48 (0.46, 0.51)	0.72 (0.70, 0.75)	0.87 (0.86, 0.89)	0.95 (0.93, 0.95)	0.98 (0.97, 0.98)
44	0.31 (0.29, 0.33)	0.63 (0.60, 0.65)	0.83 (0.81, 0.85)	0.93 (0.92, 0.95)	0.97 (0.96, 0.98)	0.99 (0.99, 0.99)
45	0.43 (0.40, 0.45)	0.74 (0.72, 0.77)	0.90 (0.88, 0.92)	0.96 (0.96, 0.97)	0.99 (0.98, 0.99)	>0.99 (0.99, >0.99)
46	0.53 (0.50, 0.56)	0.82 (0.80, 0.85)	0.94 (0.93, 0.95)	0.98 (0.97, 0.99)	0.99 (>0.99, >0.99)	>0.99 (>0.99, >0.99)
47	0.61 (0.57, 0.64)	0.87 (0.85, 0.89)	0.96 (0.95, 0.97)	0.99 (0.98, 0.99)	>0.99 (>0.99, >0.99)	>0.99 (>0.99, >0.99)
48	0.66 (0.63, 0.70)	0.90 (0.88, 0.92)	0.97 (0.96, 0.98)	0.99 (0.99, >0.99)	>0.99 (>0.99, >0.99)	>0.99 (>0.99, >0.99)
49	0.70 (0.66, 0.74)	0.92 (0.90, 0.94)	0.98 (0.97, 0.99)	0.99 (0.99, >0.99)	>0.99 (>0.99, >0.99)	>0.99 (>0.99, >0.99)
50	0.73 (0.68, 0.77)	0.93 (0.91, 0.95)	0.98 (0.97, 0.99)	>0.99 (0.99, >0.99)	>0.99 (>0.99, >0.99)	>0.99 (>0.99, >0.99)
51	0.74 (0.69, 0.80)	0.94 (0.91, 0.96)	0.98 (0.98, 0.99)	>0.99 (0.99, >0.99)	>0.99 (>0.99, >0.99)	>0.99 (>0.99, >0.99)
52	0.76 (0.70, 0.82)	0.94 (0.91, 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)
53	0.77 (0.70, 0.83)	0.95 (0.92, 0.98)	0.99 (0.98, >0.99)	>0.99 (0.99, >0.99)	>0.99 (>0.99, >0.99)	>0.99 (>0.99, >0.99)
54	0.77 (0.70, 0.85)	0.95 (0.92, 0.98)	0.99 (0.98, >0.99)	>0.99 (0.99, >0.99)	>0.99 (>0.99, >0.99)	>0.99 (>0.99, >0.99)
55	0.78 (0.70, 0.86)	0.95 (0.92, 0.99)	0.99 (0.98, >0.99)	>0.99 (0.99, >0.99)	>0.99 (>0.99, >0.99)	>0.99 (>0.99, >0.99)
56	0.78 (0.70, 0.87)	0.95 (0.92, 0.99)	0.99 (0.98, >0.99)	>0.99 (0.99, >0.99)	>0.99 (>0.99, >0.99)	>0.99 (>0.99, >0.99)

from genetic markers such as single-nucleotide polymorphisms (SNPs) continue to be a challenging issue [18]. Even with diseases like HD where the gene is identified, the predictive model can be complicated: a special feature of HD is that the mutation severity is quantifiable and varies significantly among the affected population. This contrasts with the typical categorical approach needed, for example, in genome-wide association studies. The proposed methods are also applicable to other expanded trinucleotide repeat diseases similar to HD.

One of the contributions of this work is to use the family data as well as the proband data to maximize available information in building a model. Our results reveal that the estimated risk obtained from the combined proband and family data is slightly lower than the risk estimated from the proband data alone. It is possible that the proband data consists of a biased clinical sample of gene positive or HD-affected subjects (e.g., subjects with more severe disease or with earlier onset may be more likely to participate; presymptomatic subjects might be undersampled) and is therefore not a fair representative sample of the entire HD population, especially underrepresenting subjects at risk. The plausibility of such underascertainment is so strong for CAG lengths of 40 or less [7] that we did exclude observations within that range from analysis. The family data may be a better representative of the population since the family members are included in the analysis only through the inclusion of the probands. Although proband may participate the study because they had HD or they had more severe symptoms of HD, the relatives were not included based on their CAG repeat lengths or affection status. Of course, some of the family members will not share an expanded CAG repeat huntingtin with the probands and therefore are noncarriers who will never develop HD.

Note that our estimated cumulative risk of onset of a positive HD diagnosis in the proband data is also slightly lower than Langbehn et al. [10] which also examined age-at-HD diagnosis. We estimated later mean AAO for each CAG repeat length shorter than 54 than did Langbehn et al. [10]. For example, the mean AAO of HD diagnosis for probands with a CAG of 42 in the former data was 3 years later and, for a CAG of 43, it was 4 years later (Table 3). On average, for all subjects with a CAG between 41 and 50, the mean AAO in Langbehn data was 2 years earlier than in the COHORT data. More detailed comparisons are presented in Table 5. There are several possible reasons for these differences. The model end point, AAO, should probably be considered to be slightly different in the two models. The outcome in Langbehn et al. [10] was earliest age at which a clinician documented an irreversible objective sign of the illness. This may occur earlier than the point at which an actual diagnosis of manifest HD is given. (Many clinicians wait until there are several such signs.) This may also occur, however, at a point that is later than the proband's or family's first report of subjective symptoms or their first perception of disease signs. In the CAG range of 41–49, the Langbehn et al. means are very close to the symptom onset means in the current data. For longer CAG lengths, the Langbehn et al. estimates more closely resemble the current models for disease diagnosis. Possible systematic variability between the clinicians in the two studies may also account for the differences in the estimates.

Other potential differences between the data sources include potential research-center-specific heterogeneity in diagnostic and rating conventions and slight variations in the methods used to determine CAG repeat length. In the Langbehn study, these were measured by a variety of laboratories while in the COHORT they were all measured in the same laboratory.

We do note that the differences between the fitted models here and those in Langbehn et al. are substantially smaller than differences among other formulae in the literature [14]. AAO probabilities, conditioned on current age, are especially similar. In HD research and

genetic counseling, these conditional probabilities are perhaps the most commonly used statistic deriving from these formulae. Finally, the logistic-exponential form of the parametric model proposed in Langbehn et al. [10] does indeed fit the empirical AAO distributions quite well in the COHORT data. This validates use of this relatively complicated survival model for HD AAO research and may encourage considerations of quantitative biological mechanisms that would generate exponential relationships between CAG and both AAO and its variance.

There has often been ambiguity in the modeling literature concerning the exact meaning of HD "onset." The first onset of observable signs or reportable symptoms of HD generally occurs before the actual diagnosis of clinically manifest HD is given. Much of the earlier modeling literature, reviewed in Langbehn et al. [14], does not clearly address this distinction, although the resultant formulas have often been used for subsequent prediction of HD diagnosis [14]. The event modeled in Langbehn et al. [10] was "the first time that neurological signs representing a permanent change from the normal state was identified in a patient." This might be considered to the concept of "subject's first noted symptom" rather than age of diagnosis. Nonetheless, this model has been used frequently as a predictor of future diagnosis in HD [14]. In the current study, we do distinguish between first symptom onset and diagnosis.

Here, we assumed Mendelian transmission of huntingtin without interference so that the CAG length does not change from parents to offspring. There are several possible violations of these assumptions. CAG lengths do, in reality, vary somewhat among family members, and those inheriting the gene from their father have, on average, a slightly longer CAG repeat length than their father. The probability of this occurring is much lower if inheritance is from the mother [19]. An explanation is that there are many more biological opportunities for the CAG length to change in the father's process of sperm formation than in the mother's process of egg formation. These processes and their dynamics have been studied extensively in vitro [7, 20], but we know of no well-verified in vivo dynamic population genetics models. Assuming the CAG length does not change from father to offspring may lead to a slightly lower estimated risk for affected fathers of probands.

Consistent with Langbehn et al. [10] and other studies [20, 21], we estimated reduced penetrance for lower CAG repeat lengths (≤ 40). We point out that the parameter estimates from the current model do not include subjects with CAG less than 41; therefore, the risk estimates for these subjects are extrapolations. However, it is conceivable that as long as the inverse relationship between AAO and CAG still holds for the lower CAGs, the life time disease risk for these subjects will be less than 100%, since the life time risk for a CAG of 41 is about 100%.

In the literature, no proportional odds model has been fitted to model the age-at-onset of HD. Proportional odds model, or along a similar line, transformation model, belongs to the semiparametric model framework and is beyond the scope of this paper. We are currently investigating semiparametric models other than the Cox proportional hazards model.

Finally, we stress that our current model does not include other observed covariates, such as additional genetic polymorphisms. In addition, we assumed conditional independence of family members' age-at-onset (AAO) of HD given their CAG repeats. This assumption implies that we do not account for residual correlation among family members' AAO caused by factors other than the CAG repeats, such as life style factors. When there exists such residual correlation, point estimates from our current approach are still consistent hence still valid, although the standard error estimates are no longer correct. A practical limitation of using family members' AAO data is that they may be less reliable than the data directly collected from the probands. This limitation applies to all other diseases, especially those

with late onset. This limitation can be more pronounce when there is incomplete penetrance and variability of phenotype. Future work would consider incorporating such measurement error in the analysis. Lastly, the proposed methods do not include possible unobserved effects that may be site or clinician-specific and perhaps related to the interpretation of the point of "onset." Future research will focus on incorporating observed covariates and adding family-specific random effects to account for residual familial aggregation.

Acknowledgments

References

[1] C. A. Ross, "When more is less: pathogenesis of glutamine repeat neurodegenerative diseases," *Neuron*, vol. 15, no. 3, pp. 493–496, 1995.

[2] C. A. Ross and S. J. Tabrizi, "Huntington's disease: from molecular pathogenesis to clinical treatment," *The Lancet Neurology*, vol. 10, pp. 83–98, 2010.

[3] T. Foroud, J. Gray, J. Ivashina, and P. M. Conneally, "Differences in duration of Huntington's disease based on age at onset," *Journal of Neurology Neurosurgery and Psychiatry*, vol. 66, no. 1, pp. 52–56, 1999.

[4] K. Kieburtz and Huntington Study Group, "The unified Huntington's disease rating scale: reliability and consistency," *Movement Disorder*, vol. 11, pp. 136–142, 1996.

[5] E. R. Dorsey, C. A. Beck, M. Adams et al., "TREND-HD communicating clinical trial results to research participants," *Archives of Neurology*, vol. 65, no. 12, pp. 1590–1595, 2008.

[6] E. R. Dorsey and Huntington Study Group COHORT Investigators, "Characterization of a large group of individuals with Huntington disease and their relatives enrolled in the COHORT study," *PLoS ONE*, vol. 7, no. 2, Article ID e29522, 2012.

[7] D. Falush, E. W. Almqvist, R. R. Brinkmann, Y. Iwasa, and M. R. Hayden, "Measurement of mutational flow implies both a high new-mutation rate for huntington disease and substantial under ascertainment of late-onset cases," *The American Journal of Human Genetics*, vol. 68, pp. 373–385, 2000.

[8] Y. Wang, L. N. Clark, E. D. Louis et al., "Risk of Parkinson disease in carriers of Parkin mutations: estimation using the kin-cohort method," *Archives of Neurology*, vol. 65, no. 4, pp. 467–474, 2008.

[9] D. C. Rubinsztein, J. Leggo, R. Coles 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," *American Journal of Human Genetics*, vol. 59, no. 1, pp. 16–22, 1996.

[10] D. R. Langbehn, R. R. Brinkman, D. Falush, J. S. Paulsen, and M. R. Hayden, "A new model for prediction of the age of onset and penetrance for Huntington's disease based on CAG length," *Clinical Genetics*, vol. 65, no. 4, pp. 267–277, 2004.

[11] O. C. Stine, N. Pleasant, M. L. Franz, M. H. Abbott, S. E. Folstein, and C. A. Ross, "Correlation between the onset age of Huntington's disease and length of the trinucleotide repeat in IT-15," *Human Molecular Genetics*, vol. 2, no. 10, pp. 1547–1549, 1993.

[12] C. Gutierrez and A. MacDonald, *Huntington Disease and Insurance. I: A Model of Huntington Disease*, Genetics and Insurance Research Centre (GIRC), Edinburgh, UK, 2002.

[13] C. Gutierrez and A. MacDonald, "Huntington disease, critical illness insurance and life insurance," *Scandinavian Actuarial Journal*, vol. 4, pp. 279–313, 2004.

[14] D. R. Langbehn, M. R. Hayden, and J. S. Paulsen, "CAG-repeat length and the age of onset in Huntington disease (HD): a review and validation study of statistical approaches," *American Journal of Medical Genetics*, vol. 153, no. 2, pp. 397–408, 2010.

[15] T. Louis, "Finding the observed information matrix when using the EM algorithm," *Journal of the Royal Statistical Society, Series B*, vol. 44, pp. 226–233, 1982.

[16] N. M. Laird and J. H. Ware, "Random-effects models for longitudinal data," *Biometrics*, vol. 38, no. 4, pp. 963–974, 1982.

[17] K. Marder, G. Levy, E. D. Louis et al., "Accuracy of family history data on Parkinson's disease," *Neurology*, vol. 61, no. 1, pp. 18–23, 2003.

[18] J. Kang, J. Cho, and H. Zhao, "Practical issues in building risk-predicting models for complex diseases," *Journal of Biopharmaceutical Statistics*, vol. 20, no. 2, pp. 415–440, 2010.

[19] B. Kremer, E. Almqvist, J. Theilmann et al., "Sex-dependent mechanisms for expansions and contractions of the CAG repeat on affected Huntington disease chromosomes," *American Journal of Human Genetics*, vol. 57, no. 2, pp. 343–350, 1995.

[20] C. T. McMurray, "Mechanisms of trinucleotide repeat instability during human development," *Nature Reviews Genetics*, vol. 11, no. 11, pp. 786–799, 2010.

[21] R. R. Brinkman, M. M. Mezei, J. Theilmann, E. Almqvist, and M. R. Hayden, "The likelihood of being affected with Huntington disease by a particular age, for a specific CAG size," *The American Journal of Human Genetics*, vol. 60, no. 5, pp. 1202–1210, 1997.