HHS Public Access

Author manuscript

Handb Clin Neurol. Author manuscript; available in PMC 2022 March 19.

Published in final edited form as:

Handb Clin Neurol. 2017 ; 144: 47–61. doi:10.1016/B978-0-12-801893-4.00004-3.

Natural history of Huntington's disease: evolution of modeling onset

Tanya P. Garcia [Assistant Professor],
Department of Epidemiology and Biostatistics, Texas A&M Health Science Center, TAMU 1266, College Station, TX

Karen Marder [Professor],
Department of Neurology, Psychiatry (at the Sergievsky Center and Taub Institute), Columbia University Medical Center

Yuanjia Wang[1] **[Associate Professor]**
Department of Biostatistics, Mailman School of Public Health, Columbia University, New York, NY 10032

Abstract

Huntington's disease (HD) is a unique disease caused by a CAG trinucleotide expansion in the Huntingtin gene and with the power to predict age-at-onset from subject-specific features like motor and neuroimaging measures. In clinical trials, properly modeling onset age is important because it improves power calculations and directs clinicians to recruit subjects with certain features. We discuss the history of modeling onset, from simple linear and logistic regression to advanced survival models. We highlight their advantages and disadvantages, emphasizing the methodological challenges when genetic mutation status is unavailable. We also discuss the potential bias and higher variability incurred from the uncertainty associated with subjective definitions for onset. Methods to adjust for the uncertainty in survival models are still in its infancy, but would be beneficial for HD and neurodegenerative diseases with long prodromal periods like Alzheimer's and Parkinson's disease.

Keywords

Age at onset; Diagnosis definitions; Kin-cohort study; Measurement error; Penetrance; Prediction

1 Introduction

Huntington's disease (HD) is a unique, genetic disease caused by a CAG repeat expansion in the Huntingtin gene and with the power to predict age at onset as it relates to subject-specific features such as neuroimaging measures and measures from the Unified Huntington's Disease Rating Scale (UHDRS). Such a relationship is statistically powerful because it allows one to estimate the effects of phenotypic features on age-at-onset, and the likeliness of onset, which in the case of HD has been defined by motor signs occurring by a given

[1]Corresponding Author (tpgarcia@sph.tamhsc.edu;).

age. In clinical trials, properly modeling age-at-onset in relation to subject-specific features is important because it improves power calculations and improves the trial's efficiency by directing the clinicians' efforts to recruit subjects with specific features. In this chapter, we discuss the history of models used to predict age-at-onset and estimate its distribution function (i.e., the likeliness of onset occurring by a given age). We especially focus on the utility and limitations of the different models proposed.

How one defines age of onset has evolved throughout the years. Early criteria included when a first abnormality (Andrew et al., 1993) or combination of abnormalities (Vuillaume et al., 1998) appeared; these included chorea, dystonia, inability to perform complex hand movements, as well as psychiatric and cognitive impairments. However, the greater emphasis on extrapyramidal signs due to their specificity led to the current, semi-quantitative motor assessments and diagnosis (Huntington Study Group, 1996). The motor assessments form part of the UHDRS, and their results are summarized into two summary scores. The first is a total motor score, ranging from 0 to 124, with higher numbers indicating greater impairment. The second is a diagnostic confidence level (DCL) from 0 to 4 that indicates the clinician's confidence that the subject's extrapyramidal signs are unequivocally associated with HD; a score of 4 indicates ≥99% confidence, and is the point of HD diagnosis.

For nearly two decades, clinicians have used the UHDRS exams and DCL score to diagnose HD. Recently, however, an interest to re-evaluate the criteria has emerged given that the current standard may not comprehensively account for all relevant information needed to make a diagnosis(Biglan et al., 2013; Reilmann et al., 2014). Specifically, the DCL score solely focuses on the motor aspect of HD, whereas it is well known that HD develops insidiously (Ross et al., 2014) with, for example, cognitive impairments emerging years before a motor-diagnosis (Stout et al., 2011). Hence, limiting the criteria to only one aspect may yield delays in diagnosis, and may affect results in future intervention studies where onset is a primary endpoint. To remedy this limitation, newer criteria (Table 1) have been proposed that emphasize a collective analysis of multiple aspects of HD. In particular, Biglan et al. (2013) proposed a multidimensional diagnosis which essentially asks clinicians to base diagnosis on motor, cognitive, behavioral, and functional components. Reilmann et al. (2014) proposed a natural history-based diagnosis which builds on the multidimensional-diagnosis by including a subject's medical history and stratifying criteria according to individuals who are "genetically confirmed" (i.e., subject has ≥ 36 CAG repeats) or not. Both definitions have the potential to lead to earlier diagnosis than the current standard; Biglan et al. (2013) showed in a recent study that subjects assessed by their proposed multidimensional definition (Table 1) tended to receive an earlier diagnosis than when assessed with the current standard.

1.1 Potential for uncertainty in age-at-onset

Though both definitions show promise in yielding earlier diagnoses—an aspect favored in intervention studies—they also have the potential for incorrectly determining age-at-onset. While efforts are made to ensure that a subject is viewed by the same rater at each visit, this is not always the case. For the multidimensional (Biglan et al., 2013) or natural

history-based (Reilmann et al., 2014) diagnoses, having different raters at each visit and different raters evaluating the motor, cognitive, behavioral, functional components may lead to uncertainty. For example, Biglan et al. (2013) showed that if the same rater assessed subjects based on the multidimensional and standard motor-based diagnoses, then a similar age-at-diagnosis was more likely to occur (89.8%) than when different raters were used. Different ages-at-onset between multiple raters is common in any rating scale, and is a prime motivation for more objective, non-rater dependent measures (Ross et al., 2014).

Differences in age-at-onset can also stem from the non-quantitative criteria in the multidimensional and natural history-based diagnoses definitions. For example, Reilmann et al. (2014) suggest that a clinician should identify a subject as having manifest HD if “clear changes” occur in the functional scales of UHDRS. However, no quantitative threshold is provided that would indicate a ‘clear’ functional change. Hence, in their current forms, both new diagnoses definitions largely rely on subjective measures which can ultimately lead to poor interrater reliability—an issue that is problematic if the criteria is used in intervention studies where accurate diagnoses are needed.

The potential for incorrectly determining age-at-onset can also induce bias in statistical models used to predict onset or estimate its distribution. The uncertain onset ages are referred to in the statistical literature as mismeasured onset ages. That is, a mismeasured onset age is equal to the true, unobservable onset age plus some error; the error is the so-called measurement error. A range of statistical methods have been developed to handle different models with measurement error; see Carroll et al. (2006) for an excellent review of different methods. However, to the best of our knowledge, none of the statistical measurement error techniques has been utilized in the HD literature. In Section 4, we highlight the impact of measurement error in the models developed over the past two decades. We also discuss potential methodologies needed to handle mismeasured ages of diagnoses—techniques which could also be useful in other neurodegenerative diseases such as Alzheimer’s or Parkinson’s.

2 Regression models for age-at-onset

Regression models to predict age-at-onset have been extensively developed in the HD literature. We discuss the history of these models and highlight their advantages and disadvantages, with potential remedies in the latter case. In addition, we discuss the impact of mismeasured on set ages on these models. A summary of our findings are listed in Table 2.

2.1 Correlation analysis and linear regression

A first approach to assessing the relationship between age-at-onset and subject-specific features is through the Pearson (1895) correlation coefficient: a measure of the strength of the linear relationship between two variables. After the discovery of the Huntingtin gene in 1993, several studies used the Pearson correlation coefficient to show that age-at-onset is significantly and inversely correlated with CAG repeat length (Andrew et al., 1993; Duyao et al., 1993; Stine et al., 1993; Snell et al., 1993; Trottier et al., 1994; Lucotte et al., 1995; Huntington Study Group, 1996; Brandt et al., 1996; Rubinsztein et al., 1997; Vuillaume

et al., 1998; Foroud et al., 1999). This relationship has proved useful in HD studies, but correlation analysis alone is limiting. It is strongly influenced by outliers such as juvenile onset where onset occurs in individuals ≤ 20 years (Quarrell et al., 2012) which is much younger than the average adult onset of 40 years (Myers, 2004). Similarly, outliers are prevalent in the number of CAG repeats where some studies have reported individuals (e.g., juveniles) with 250 repeats (Nance et al., 1999). Such a high CAG repeat number far exceeds the threshold used to genetically confirm individuals as carriers of the HD gene (i.e., 36 CAG repeats). Moreover, correlation analysis does not reveal a potential nonlinear relationship between two variables. For example, computing the correlation coefficient between age-at-onset and CAG repeats does not reveal that the relationship between the two is actually a curved S-shape as shown by Langbehn et al. (2004).

An improvement over correlation analysis is linear regression which expresses the linear relationship between age-at-onset and subject-specific features via:

$$T_{AAO} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \epsilon \quad (1)$$

Here, T_{AAo} represents the age-at-onset (response variable), $x_1, \cdots, x_p$ are p subject-specific features (covariates) and ϵ is the residual model error assumed to be normally distributed with mean zero and variance σ^2. The parameters $\beta_0, \cdots, \beta_p$ and σ^2 are estimated using standard regression techniques (i.e., ordinary least squares). Linear regression permits quantifying the effect each feature has on age-at-onset, and assessing how much of the variability of age-at-onset is explained by the features (i.e., coefficient of determination). Linear regression techniques have led to the discovery that CAG repeat-length accounts for 50–70% of the variance in age-at-onset (Andrew et al., 1993; Duyao et al., 1993; Stine et al., 1993; Snell et al., 1993; Trottier et al., 1994; Lucotte et al., 1995), with the remaining variance attributed to other features. These include, for example, genetic markers such as the GluR 6 kainate receptor which accounts for 1 3 % of the variability of age-at-on set after accounting for C A G repeats (Rubinsztein et al., 1997) and a Δ2642 mutation which significantly decreases age-at-onset (Vuillaume et al., 1998); high passive activity scores indicating that a subject prefers less physical or intellectual activities which significantly decrease age-at-onset (Trembath et al., 2010); neuroimaging markers which have varying associations with age-at-on set (Stout et al., 2011; Ciarmiello et al., 2012) and higher caffeine intake before disease onset which significantly decreases age-at-on set (Simonin et al., 2013). The aforementioned results are all from linear regression analyses; more advanced analyses have led to other features associated with age-at-onset (Paulsen et al, 2014).

While linear regression techniques have aided to identify and estimate the effects of influential features on age-at-onset, a limiting factor is that it requires the onset ages to be fully observed. Subjects with missing or unknown onset are ignored in the analysis. This is problematic since missing observations can limit the study sample sizes to less than 100 subjects, as in Trottier et al. (1994), Brandt et al. (1996), Vuillaume et al. (1998), and Ciarmiello et al. (2012). Missing onset ages can occur, for example, with subjects having 36 to 39 CAG repeats; such subjects exhibit variable penetrance (Walker, 2007) and late onset, meaning that their onset ages may occur after the study period ends. One potential

remedy for missing observations is to replace or impute the unobserved onset ages with data-driven, biologically meaningful values (Rubin, 1996; Little and Rubin, 2002). For example, one could impute the missing ages-at-onset with the average of the observed onset ages, or with a predicted age from a linear regression model based on data from subjects whose onset ages are fully observed. Doing so, however, may lead to incorrect imputations since, as previously mentioned, subjects whose onset ages are observed are not necessarily representative of subjects for whom onset ages are missing.

2.2 Logistic regression with binary phenoconversion status outcome

A limiting factor of correlation analysis and linear regression is that both rely on all ages-at-onset being observed; that is, individuals who have not phenoconverted during the study period are ignored. However, information from individuals who have not phenoconverted is also beneficial in predicting age-at-onset since it is important to understand what features these individuals have that might delay age at onset. To use information from all study subjects, whether phenoconverted or not, one modeling approach is logistic regression. Here, a binary response indicating diagnostic status (i.e., whether or not a subject has phenoconverted) is related to subject-specific features (covariates) through log odds (Vittinghoff et al., 2012, chap. 5). More exactly, letting $Y_{AAO} = 1$ if a subject has phenoconverted, and 0 if not, a logistic model takes the form:

$$\log\left\{\frac{pr\left(Y_{AAO}=1 \mid X_1,\cdots,X_p\right)}{1-pr\left(Y_{AAO}=1 \mid X_1,\cdots,X_p\right)}\right\}=\beta_0+\beta_1X_1+\cdots+\beta_pX_p\,. \tag{2}$$

This equation implies that the log odds of onset occurring (left-hand-side) is linearly related to p subject-specific features $x_1, \cdots, x_p$ (right-hand-side).

The logistic model in equation (2) has been used in the HD literature, not only to assess what features significantly impact the log odds of onset, but other outcomes as well. Aylward et al. (2012) used logistic regression to show that putamen volumes are significantly associated with the log odds of diagnostic status, and the log odds of UHDRS motor score groups (individuals with a motor score ≥10 or <10). In another study, Beglinger et al. (2010) used logistic regression to show that the log odds of functional decline is significantly predicted by motor, cognitive, and depressive symptoms as measured on the UHDRS.

Logistic regression is beneficial in that it extracts information from subjects who have and have not experienced the outcome of interest (e.g., motor-onset). One limitation, however, is that it ignores any time differences between subjects experiencing the outcome. More specifically, if the outcome of interest is phenoconversion, then logistic regression treats a subject who phenoconverts after 6 months the same as a subject who phenoconverts after 18 months. That is, both will have $Y_{AAO} = 1$ in equation (2) although the 12-month time difference between the two subjects may be clinically relevant. For example, subjects with fewer CAG repeats will have delayed onset and capturing the relationship between CAG repeats and this delayed onset may provide insights for establishing guidelines in clinical studies and understanding a diagnosis in genetic counseling sessions. A remedy to this loss

of time information is defining multiple binary variables indicating if onset occurs in certain time intervals: $Y_{AAO\ in\ [t_1,t_2],} = 1$ if age-at-onset occurs in the time interval $[t_1,t_2]$, where the time intervals are non-overlapping. Specifying the time intervals requires care, however, since an interval where no subjects experience onset would lead to a numerically unstable situation of logistic regression with rare events (King and Zeng, 2011). Time intervals with rare events are more likely to occur in studies with varying lengths of follow-up; see Section 3 for an alternative method that better handles high follow-up variability.

3 Survival models for age-at-onset

Up to now, we have discussed two important regression models used to model age-at-onset: linear regression where all onset ages must be observed, and logistic regression where not all onset ages are observed and the focus is on whether or not each subject has phenoconverted. A strong limitation of logistic regression is it ignores any time differences between subjects experiencing phenoconversion—information which is useful to understanding how subject-specific features affect when onset occurs. An approach better suited to capturing the time differences is a survival model: a model specifically designed to represent time-to-event data in relation to subject-specific features without losing information due to varying lengths of follow-up.

3.1 Brief overview of survival analysis

Before discussing the evolution of survival-type models in the HD literature, we first provide a brief overview of the key features in survival analysis: the ability to handle varying lengths of follow-up, the ability to estimate so-called survival and hazard functions, and the ability to handle random censoring as defined next.

Random censoring occurs when the onset age for an individual is unobserved either because the subjects leaves the study, is lost to follow-up or does not phenoconvert during the study period due to reasons related to their observed characteristics (covariates), but not their underlying likeliness of HD onset in the future. It is also referred as non-informative censoring in the survival analysis literature (Kleinbaum and Klein, 2011). Unlike in linear regression where censored subjects are generally removed from the analysis, a survival model adjusts for censoring using a technique called likelihood-based methods (Vittinghoff et al., 2012, chap. 6). Likelihood methods involve a likelihood function: a product formed by the likeliness that each subject's onset age is censored or observed at different time points. The likelihood function can incorporate the effects of subject-specific features which can then be unbiasedly estimated using maximum likelihood (Aldrich, 1997), Expectation-Maximization (Dempster et al., 1977), or other techniques. Though likelihood-based methods are designed to handle censoring, too much censoring is also not helpful. A high percentage of censoring indicates that a significant part of the target population is not being included. One solution is to have more representative sampling, but this may be difficult in HD studies since there is a varied prevalence of nonsymptomatic CAG expansion in the 36 to 40 range, for example, and including subjects with such a range in clinical samples is rare. Large sample from multiple sites as are being assembled in the ENROLL-HD study may allow such specific recruitment and help to alleviate this issue.

The survival and hazard functions in survival analysis help to describe the distribution of event times. The survival function is the probability of not experiencing onset (i.e., "surviving") up to a pre-specified time t, and the hazard function is the rate at which a subject will experience onset given he has not yet done so. Modeling the survival and hazard function is generally achieved using one of three approaches: parametric, semiparametric and nonparametric methods. Parametric methods involve positing a particular model for the survival and/or hazard function; the posited model is fully specified except for a small, finite number of parameters that need to be estimated from the data. A parametric model may be reasonable, for example, when the researcher has enough biological understanding of the underlying disease mechanism. If the posited model is indeed correct, then parametric methods are numerically simple, and yield unbiased and efficient estimates. Correctness of the posited model is important, however, since inadequacy results in bias and incorrect inference conclusions.

At the other end of the spectrum, nonparametric methods make no a priori assumptions about the survival/hazard function form; instead, the forms are completely estimated from the data. A popular nonparametric method is the Kaplan and Meier (1958) product limit estimator which estimates the survival function free of any restrictive assumptions on the underlying time-to-event process. Lastly, semiparametric methods are a compromise to parametric and nonparametric methods. Semiparametric methods involve a priori assumptions for some model components and complete flexibility for others. A well-known semiparametric model is the Cox proportional hazards model which links the linear effect of subject-specific features to a ratio of hazard functions through a log-transformation. Unfortunately, it is not always valid in HD studies as shown by Langbehn et al. (2004). Still, semiparametric and nonparametric methods are not susceptible to misspecification of the survival/hazard functions and are especially useful when the researcher does not have sufficient biological knowledge about the underlying disease mechanism, which is particularly common in HD studies.

In the subsequent sections, we discuss different forms of parametric, semiparametric and nonparametric models proposed in the HD literature, and highlight their advantages and disadvantages in better understanding age-at-onset and its distribution. The subsections are divided according to whether the analysis is performed under known or unknown genetic information, the latter of which occurs in kin-cohort studies (see Section 3.3).

3.2 Models with genetic mutation status known

A variety of parametric and semi- or nonparametric survival models have been developed for HD studies where genetic information of study individuals is available. Given the strong associations between CAG repeat-length and age-at-onset, the available genetic information helps to more accurately model the relationship between the two. A summary of our findings is in Table 3.

3.2.1 Parametric methods—Parametric survival models in the HD literature include the work of Gutierrez and MacDonald (2002), Langbehn et al. (2004), and Zhang et al. (2011), all of which can predict age-at-onset and its distribution based on subject-specific

features (i.e., CAG repeat length, age). The Gutierrez and Macdonald (2002) model was constructed from the data of Brinkman et al. (1997), which included 1,049 subjects (CAG repeats 29–121) whose information was collected retrospectively. Roughly 69.4% of the subjects had reported onset and 30.6% did not (i.e., onset ages were censored). Similarly, the Langbehn et al. (2004) model was constructed based on 2,913 subjects (CAG repeats 41 to 56) whose information of onset was also collected retrospectively. Roughly 78.9% of the subjects reported onset and 21.1% did not. The model of Zhang et al. (2011) was constructed based on 730 prodromal individuals from PREDICT-HD (Paulsen et al., 2008) followed prospectively for up to 7 years (CAG repeats ≥36). Roughly 18.8% of the subjects experienced onset (i.e., phenoconverted), and 81.2% did not. The larger censoring rates in the Zhang et al. (2011) study certainly increase the difficulty of identifying adequate model fits, but do not invalidate the analysis. Plausible reasons for the higher censoring in Zhang et al. (2011) compared to that in Gutierrez and MacDonald (2002) and Langbehn et al. (2004) are (i) the differences in the distribution of CAG repeat lengths: Gutierrez and MacDonald (2002) and Langbehn et al. (2004) involved many subjects with a high number of CAG repeats who would more likely experience onset; and (ii) different inclusion criteria for retrospective and prospective studies where the former may have been more relaxed to include more onset cases.

All three studies followed similar strategies to formulate their models, but used different criteria for best-fit. First, the age-at-onset data was used to find an appropriate distribution function; i.e., the probability age-at-onset T_{AAO} occurs by age t given subject specific features $\boldsymbol{x}$, represented mathematically as $\text{pr}(T_{\text{AAO}} < t \mid \boldsymbol{x})$ Second, a parametric model was identified to best relate age-at-onset (or its summary measures) and the subject-specific features.

Following this two-part strategy, Gutierrez and MacDonald (2002) fit different Kaplan-Meier curves for the ages-at-on set stratified by CAG repeat length. To avoid issues of under ascertainment, Gutierrez and MacDonald (2002) focused on the subset of subjects with 40–50 CAG repeats. After studying the general shapes formed from the Kaplan-Meier curves, the authors found that a gamma distribution with a linear function for the effect of CAG repeats best fit the age-at-onset distribution. A gamma distribution is a flexible model of two parameters commonly used to represent time-to-event data (Kleinbaum and Klein, 2011). Best-fit was assessed graphically by comparing different parametric models to the nonparametric Kaplan-Meier curves.

In a similar spirit, Langbehn et al. (2004) and Zhang et al. (2011) also used Kaplan-Meier curves stratified by CAG repeat length to learn the general shape of the age-at-onset distribution. Langbehn et al. (2004) then examined 12 different parametric families of functions to see which function fit the average shape of the Kaplan Meier curves best. Unlike the gamma distribution of Gutierrez and Macdonald (2002), Langbehn et al. (2004) found that the best fit was given by the following logistic distribution:

$$\text{pr}(T_{\text{AAO}} < t \mid \boldsymbol{x}) = \frac{1}{1 + exp\{-\alpha_1(\boldsymbol{x}) - \alpha_2(\boldsymbol{x})u(t)\}}. \quad (3)$$

Here, $\boldsymbol{x}$ is a vector of subject-specific features and the functions, $\alpha_1(\boldsymbol{x})$, $\alpha_2(\boldsymbol{x})$ $u(t)$, are as defined below. The logistic distribution in equation (3) is much simpler than the gamma distribution of Gutierrez and MacDonald (2002) because the former does not involve complex numerical integration.

The forms of $\alpha_1(\boldsymbol{x}), \alpha_2(\boldsymbol{x}), u(t)$ in Langbehn et al. (2004) were determined by assessing the relationship between CAG repeat-length and the mean and dispersion of the ages-at-onset. They ultimately found that a so-called exponential-linear model adequately described the relationship. This form was later validated using prospective data from PREDICT-HD (Langbehn et al., 2010) involving 610 subjects with at least 1 year of follow-up, where 13.3% experienced motor-onset and 86.7% did not. In spite of the high censoring rates, Langbehn et al. (2010) demonstrated that their proposed model captured the general experience of the prospective data fairly well: the 95% confidence band from the predicted onset rates covered the observed onset rates from the prospective data, but the predicted onset rates and observed onset rates did not overlap in general.

In comparison, the forms of $\alpha_1(\boldsymbol{x}), \alpha_2(\boldsymbol{x}), u(t)$ in Zhang et al. (2011) were determined by relating age-at-onset to both CAG repeat-length and age at study-entry. Thirty-five different forms were assessed and ultimately an exponential-linear model different from that of Langbehn et al. (2004) was determined appropriate. The model was shown to perform well in terms of predicting individuals who would be diagnosed in the next two years using a longitudinal receiver operating characteristic analysis (Heagerty and Zheng, 2005).

To the best of our knowledge, only a formal comparison between the parametric survival models of Gutierrez and MacDonald (2002) and Langbehn et al. (2004) has been carried out (Langbehn et al., 2010). It has been shown that the model of Langbehn et al. (2004) provides earlier estimates of onset probabilities compared to those of Gutierrez and MacDonald (2002). The differences are most likely due to the Gutierrez and MacDonald (2002) study using subjects with 40–50 CAG repeats, whereas Langbehn et al. (2004) used subjects with 41–56 CAG repeats, and purposely excluded subjects with 40 CAG repeats to avoid under ascertainment bias (Langbehn et al., 2010).

All three parametric models provide substantial benefits to the HD literature. They (i) provide means to use subject-specific features $\boldsymbol{x}$ to predict the likeliness of onset occurring by a given age; (ii) predict the probability of onset occurring in s-year intervals (e.g., 5-year intervals); (iii) estimate penetrance rates at different CAG repeat-lengths; (iv) model the explicit effect of subject-specific features on the age-at-onset distribution; and (v) provide clinically meaningful divisions of subjects so that they can be compared cross-sectionally.

In regards to benefit (v), a clinically meaningful division from the Langbehn et al. (2004) model is the division of subjects into so-called Far-Mid-Near categories for the estimated time to diagnosis. A subject is said to be in the Far category if his estimated time to diagnosis is ≥15 years; he is in the Mid category if his estimated time to diagnosis is between 9 and 15 years; and h e i s in the Near category if his estimated time to diagnosis is <9 years. The estimated time to diagnosis is based on the Lanbehn et al. (2004) model using equation (3) with the specific $\alpha_1(\boldsymbol{x}), \alpha_2(\boldsymbol{x}), u(t)$. In comparison, estimates from the Zhang et

al. (2011) model lead to a different set of clinically meaningful divisions. The Zhang et al. (2011) division is defined by the so-called CAG-Age-Product (CAP) score which divides subjects into Low-Med-High categories that emphasize the amount of disease burden. A subject is in the Low category if he has less than 50% chance of being diagnosed in the next 5 years; a subject is in the Med category if he has 50% chance of being diagnosed in the next 5 years; and a subject is in the High category if he has more than 50% chance of being diagnosed in the next 5 years. Zhang et al. (2011) demonstrated that classifying individuals in Far-Mid-Near compared to Low-Med-High generally moves individuals from being in less severe categories to more severe ones. The authors point out that such a migration is useful, especially in the PREDICT-HD cohort which has aged since the Far-Mid-Near categorization and thus are more susceptible to being diagnosed.

While the explicit parametric survival models have clear advantages, they also have limitations. One limitation is that CAG repeat length and age-at-entry are the only subject-specific features in the model. This is relatively understandable given that two of the studies (Gutierrez and Macdonald, 2002; Langbehn et al., 2004) are retrospective and so collecting additional information on subjects may be infeasible. In addition, finding explicit mathematical formulations that adequately describe the effects of CAG repeat-length and age-at-entry on age-at-onset is challenging. Searching for an explicit formulation is challenging with one or two features, and the difficulty only grows exponentially as the number of features increase. Still, advancements in the field have shown numerous other subject-specific features having significant impact on age-at-onset. These include the effect of parental age of onset which presumably reflects additional genetic and/or environmental influences (Duyao et al., 1993; Snell et al., 1993; Andrew et al., 1993; Stine et al., 1993; Trottier et al., 1994; Lucotte et al., 1995; Ranen et al., 1995); paternal vs maternal transmission (Trottier et al., 1994; Ranen et al., 1995); additional genetic polymorphisms (Li et al., 2003; Rubinsztein et al., 1997; MacDonald et al., 1999; Kehoe et al., 1999; Panas et al., 1999), and neuroimaging measures (Tabrizi et al., 2013), among others. A model that explains the cumulative effect of these features on the distribution of age-at-onset would certainly be advantageous in designing clinical studies and providing guidance during genetic counseling sessions. However, assessing a cumulative effect may be challenging in the parametric setting, and finding a correct explicit formulation with so many features may be near impossible. One possible way to avoid incorrect explicit formulations is to use more flexible techniques such as semiparametric or nonparametric models as described next.

3.2.2. Nonparametric methods—An implicit assumption of the parametric models in Section 3.2.1 is that they accurately relate the effect of subject-specific features on the age-at-onset distribution. When this assumption is in doubt, however, the models can be misleading. To counteract potential model misspecification, more flexible nonparametric techniques can be used.

A popular nonparametric estimator is the Kaplan-Meier which has been used to flexibly estimate the age-at-onset distribution stratified by the number of CAG repeats (Brinkman et al., 1997; Maat-Kievit et al., 2002; Langbehn et al., 2004; Gutierrez and Macdonald, 2002, 2004). In all studies, log rank tests agree that penetrance significantly differs by the number of CAG repeats, with complete penetrance observed for CAG repeats ≥ 42, and

reduced penetrance for 36–41 CAG repeats (Brinkman et al., 1997). A limitation of the Kaplan-Meier estimator, as argued by Langbehn et al. (2010), is that it does not permit an explicit relationship between the effect of CAG repeats and the age-at-onset distribution as done so by the parametric models of Section 3.2.1. Still, these parametric models used fairly strong assumptions, especially in the forms of $\alpha_1(\boldsymbol{x}), \alpha_2(\boldsymbol{x})$ in equation (3). To reduce the chance that $\alpha_1(\boldsymbol{x}), \alpha_2(\boldsymbol{x})$ are misspecified, a more flexible way to model them is through kernel functions or smoothing spline techniques (De Boor, 2001) as explored by Ma and Wang (2014b).

Kernel functions and smoothing splines are nonparametric techniques that estimate $\alpha_1(\boldsymbol{x}), \alpha_2(\boldsymbol{x})$ flexibly and without making any a priori assumptions of their particular forms. For example, we do not assume $\alpha_1(\boldsymbol{x}), \alpha_2(\boldsymbol{x})$ follow an exponential-linear formula as in Langbehn et al. (2004) or Zhang et al. (2011). Instead, the forms for $\alpha_1(\boldsymbol{x}), \alpha_2(\boldsymbol{x})$ are completely decided by the data using kernel and smoothing functions implemented in standard software (e.g., SAS, R, STATA); the flexible estimation helps to avoid issues of model misspecification. The flexibility, however, does mean that the final estimates for $\alpha_1(\boldsymbol{x}), \alpha_2(\boldsymbol{x})$ are usually very complex and cannot be written down explicitly. For this reason, it is nearly impossible to have an explicit and clinically meaningful division of the subjects based on a model with a nonparametric estimate for, $\alpha_1(\boldsymbol{x}), \alpha_2(\boldsymbol{x})$ In summary, nonparametric methods provide wider flexibility but at the incurring cost of computational challenges.

In an analysis of the Cooperative Huntington's Observational Research Trial (COHORT), Ma and Wang (2014b) compared their estimates of the age-at-onset distribution with $\alpha_1(\boldsymbol{x}), \alpha_2(\boldsymbol{x})$ estimated using nonparametric kernel functions to the estimates obtained from the Langbehn et al. (2004) model with parametric forms for $\alpha_1(\boldsymbol{x}), \alpha_2(\boldsymbol{x})$ The nonparametric estimates differ from those of the Langbehn et al. (2004) model in that the parametric model forces the age-at-onset distribution to be an increasing function of CAG repeats. That is, the model enforces that the likeliness of onset occurring increases with longer CAG repeats. Surprisingly, however, this increasing trend is not supported by the COHORT data at ages 15, 25, and 35 years (see Figure 3 in Ma and Wang (2014b)). Instead, the nonparametric model of Ma and Wang (2014b) shows evidence that for younger ages, the age-at-onset distribution remains constant across CAG repeats. The results suggest that further studies are needed and that the clinical impression of higher CAG repeats increasing the risk of disease may need to be re-evaluated especially for younger individuals. The observed discrepancies for younger individuals may be due in part to the low sample size of subjects who are under 35 in the COHORT study. Also, the phenotype for very young individuals differs from older subjects with the phenotype resembling more Parkinsonism than chorea.

3.3 Models with gene mutation status unknown

Up to now, we have discussed analytical techniques for studies where the gene mutation status of individuals is known. Due to high costs or reluctance to undergo genetic testing, genetic mutation status is not always available. A recent body of literature has explored techniques for analyzing such data as collected in kin-cohort studies. The overall aim of

these methods is to find ways to extract all meaningful information from the kin-cohort studies so as to not dismiss subjects with missing genetic mutation status.

Kin-cohort studies involve a sample of (usually diseased) subjects referred to as probands that are genotyped. Disease history and age-at-onset in the probands' first-degree relatives is obtained through validated interviews (Marder et al., 2003). The relatives' genotype information, however, is not collected because of practical considerations. Instead, the probability that the relative has the genetic mutation or not is computed under the assumption of Mendelian transmission using information about the relative's relationship to the proband and the proband's mutation status (Khoury et al., 1993, sec. 8.4). The relatives' ages-at-onset and probability of having the gene mutation or not are then used to estimate the age-at-onset distribution via survival models. Information from the probands can also be utilized, but one may prefer to not do so if there are concerns of ascertainment bias: when the mutation carrier probands are not a representative sample of the population of HD mutation carriers. Ascertainment bias is difficult to adjust (Begg, 2002), and easily avoided by excluding the probands from the analysis.

The model for the age-at-onset distribution in kin cohort studies takes the form of

$$\begin{aligned}\mathrm{pr}(T_{\mathrm{AAO}} < t \mid \boldsymbol{x}) = \pi \mathrm{pr}(T_{\mathrm{AAO}} < t \mid \boldsymbol{x}, \mathrm{CAG} \geq 36) + (1 - \pi \\ )\mathrm{pr}(T_{\mathrm{AAO}} < t \mid \boldsymbol{x}, \mathrm{CAG} < 36),\end{aligned} \tag{4}$$

which is essentially a mixture of two distributions: the distribution for the mutation carrier group,pr $\mathrm{pr}(T_{\mathrm{AAO}} < t \mid \boldsymbol{x}, \mathrm{CAG} \geq 36)$ and the distribution for the mutation non-carrier group, $\mathrm{pr}(T_{\mathrm{AAO}} < t \mid \boldsymbol{x}, \mathrm{CAG} < 36)$. Each distribution is multiplied by $\boldsymbol{\pi}$ or 1-$\boldsymbol{\pi}$ which denotes, respectively, the probability an individual has the gene mutation, or not. This mixture of distributions is needed because the exact genotype status is unknown, and only probabilities (or proportions) of the status are known via Mendelian assumptions. The type of model above is referred to in the statistics literature as a genetic mixture model.

We now discuss different parametric and nonparametric survival models for kin cohort studies; see Table 4 for a summary of the models. Compared to the models in Section 3.2, the exact form of the models here are much more involved and beyond the scope of this chapter. Instead, we provide the appropriate references for readers interested in the mathematical details.

3.3.1 Extensions of Langbehn et al (2004) model—Chen et al. (2012) extended the parametric model of Langbehn et al. (2004) to a kin cohort study by incorporating information from the probands and relatives, along with the mixture proportion $\boldsymbol{\pi}$. The authors modeled $\mathrm{pr}(T_{\mathrm{AAO}} < t \mid \boldsymbol{x}, \mathrm{CAG} \geq 36)$ in equation (4) using the Langbehn et al. (2004) model. For relatives of probands where covariates $\boldsymbol{x}$ are not observed, the authors assume that a relative inherits the same CAG repeat length as the proband if the proband has the gene mutation. Such an assumption may be an oversimplification of the gamete transmission process, but a model that better represents the changing CAG repeat-length can be later incorporated into the mixture proportion $\boldsymbol{\pi}$. A future area of research is finding models to adequately represent the gamete transmission process (the fact that transmission from fathers

may result in an increased CAG repeat length compared to transmission from mothers) and seeing its effects on these genetic mixture models.

Under the simplified assumption for gamete transmission and using a so-called Expectation-Maximization algorithm (Dempster et al., 1977) to estimate model parameters, Chen et al. (2012) ultimately found that their model which combines history of HD observed in first-degree relatives and probands leads to lower estimates of penetrance than when using probands alone. One potential reason contributing to this difference is that the proband data may consist of a biased clinical sample of premanifet or HD-affected subjects. Probands with more severe disease or with earlier onset may be more likely to participate. Premanifest probands might be undersampled and therefore may not b e a fair representative sample of the entire HD population, especially underrepresenting subjects at risk. 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, not their own HD disease status.

Ma and Wang (2014b) also modeled $\mathrm{pr}(T_{\mathrm{AAO}} < t \mid \boldsymbol{x}, \mathrm{CAG} \geq 36)$ in equation (4) using the Langbehn et al. (2004) model except with $\alpha_1(\boldsymbol{x}), \alpha_2(\boldsymbol{x})$ in equation (3) estimated nonparametrically. Ma and Wang (2014b) made the same gamete transmission assumptions as in Chen et al. (2012), and used techniques such as local kernel and backfitting to estimate the model parameters for the genetic mixture model. Analogous to the results of Chen et al. (2012), Ma and Wang (2014b) had lower estimates of penetrance than observed by the Langbehn et al. (2004) model; see Figure 3 in Ma and Wang (2014b). There are several possible reasons for these differences. The model outcome, age-at-onset, might be considered to be slightly different in the two models. The event in Langbehn et al. (2004) 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, which was the event being considered in Ma and Wang (2014b). Possible systematic variability between the clinicians in the two studies may also account for the differences in the estimates.

3.3.2 Nonparametric Methods—Beyond extensions of the Langbehn et al. (2004) model, a wide range of methods have been developed to estimate the mixture distributions $\mathrm{pr}(T_{\mathrm{AAO}} < t \mid \boldsymbol{x}, \mathrm{CAG} \geq 36)$ and $\mathrm{pr}(T_{\mathrm{AAO}} < t \mid \boldsymbol{x}, \mathrm{CAG} < 36)$ in equation (4) nonparametrically. The methods vary in computational difficulty and correctness, but in general provide the following key advantages: the resulting estimated cumulative risk curve can (i) serve as time-dependent positive and negative predictive values of the HD mutation test (Heagerty and Zheng, 2005); (ii) provide a numerical summary of cumulative risk by a certain age associated with a positive mutation test; (iii) predict the risk of onset for a subject based on his genetic test results and demographic information; (iv) predict conditional probabilities of developing HD in the next s-years (i.e., 5-years) given the subject's current age current age of a subject. The advantages are similar to those as stated for parametric models in Section 3.2.1, except that the nonparametric estimation for $\mathrm{pr}(T_{\mathrm{AAO}} < t \mid \boldsymbol{x}, \mathrm{CAG} \geq 36)$ and $\mathrm{pr}(T_{\mathrm{AAO}} < t \mid \boldsymbol{x}, \mathrm{CAG} < 36)$ completely avoid model misspecification issues.

We now discuss the range of nonparametric estimators for the genetic mixture model, emphasizing their utility and limitations. Wacholder et al. (1998), Chatterjee and Wacholder (2001) and Fine et al. (2004) proposed different nonparametric maximum likelihood estimators (NPMLEs), all of which make no a priori assumptions about the mixture distributions. The so-called type I NPMLE (Wacholder et al., 1998) models the mixture distributions using a linearly transformed combination of NPMLEs. The so-called type II NPMLE (Chatterjee and Wacholder, 2001) is based on an Expectation-Maximization algorithm. The NPMLE of Fine et al. (2004) is constructed assuming independence between censoring times and the event of interest (i.e., age-at-onset). Given the independence assumption, the NPMLE of Fine et al. (2004) is referred to as IND NPMLE.

Although these methods are well-established in the statistics literature, they unfortunately are inadequate for kin cohort studies. Using COHORT data, Wang et al. (2012) and Ma and Wang (2014a) demonstrated their limitations. First, the type I NPMLE estimates the cumulative risk of onset as 40% over all ages—an estimate that completely disagrees with clinical findings (Langbehn et al., 2004). Second, Wang et al. (2012) demonstrated that the type II NPMLE gives biased and unreliable estimates of cumulative risk of onset. Lastly, the IND NPMLE estimates the risk of onset at 65 years as being greater than 1, thus indicating that the IND NPMLE violates the constraints that estimated risk, being a probability, must be between zero and one.

Motivated by these inadequacies, Wang et al. (2012) and Ma and Wang (2014a) developed a series of nonparametric estimators that are asymptotically unbiased (i.e., consistent), efficient (i.e., small variability), and provide clinically meaningful estimates of the cumulative risk of onset. The estimators stem from deriving all potential unbiased estimators (i.e., functions) that appropriately represent the distribution for kin-cohort studies as modeled in equation (4); then among all these estimators, we identify which estimator has the smallest variability (Tsiatis, 2006). A consistent estimator with smallest variability is desired so as to achieve reliability and high power in detecting significant differences. Doing the rigorous procedure mentioned, Wang et al. (2012) identified three novel classes of nonparametric estimators: an inverse probability weighting (IPW) estimator which gives more weight to subjects who are under-represented because of censoring; an augmented IPW estimator which has substantially less variability than the IPW estimator; and an imputation-based estimator which has similar variability to the augmented IPW but is substantially more complicated to compute.

A limitation of the Wang et al. (2012) estimators is the computational difficulties. The imputation-based estimator is complex and time-consuming, and the two IPW-based estimators can be numerically unstable since the inverse weighting function can lead to division by zero in some instances. As a remedy to these computational challenges, Ma and Wang (2014a) developed a fourth novel class of nonparametric estimators: a so-called weighted least squares estimator that is computationally effortless and has substantially less variability than any of the NPMLEs or the estimators of Wang et al. (2012). Using the kin-cohort data from the COHORT study, Ma and Wang (2014a) demonstrated the computational simplicity of their estimator over the aforementioned ones in estimating the cumulative risk curve.

Lastly, a final limitation of all nonparametric estimators mentioned, including the weighted least squares, is that none is guaranteed to satisfy the mathematical properties of a distribution function: be positive and non-decreasing over time. Wang et al. (2012) and Ma and Wang (2014a) remedied this flaw by a post-estimate adjustment to ensure that mathematical assumptions were met. Rather than a post-estimate adjustment, a more direct approach is that of Qin et al. (2014) which involved combining an Expectation-Maximization algorithm and isotone regression (Ayer et al., 1955) to guarantee the assumptions were met. The nonparametric estimator of Qin et al. (2014) is also consistent, efficient and has high power in detecting differences between age-at-onset distributions for different subpopulations (i.e., genetic mutation carriers and non-carriers). Incorporating the isotone regression into the estimator of Ma and Wang (2014a) may be worth pursuing to make further improvements.

4. Discussion

We discussed the range of models for age-at-onset developed over the past 30 years. Given the potential move to more subjective measures of onset—the multidimensional (Biglan et al., 2013) and natural-history based (Reilmann et al., 2014) ones—a natural question that arises is how would models for predicting onset change if the onset ages are mismeasured? See Table 5 for a brief overview of how mismeasured onset ages affect the modeling approaches.

Pearson correlations (Section 2.1) would be biased downward (Fan, 2003), meaning that the observed correlation will be weaker. Such weaker correlation may be present in the studies of Aylward (2014) or Tabrizi et al. (2011, 2013) where correlations are computed between structural magnetic resonance imaging (MRI) and estimated ages-at-onset, the latter of which could presumably be mismeasured.

For linear regression (Section 2.1), the answer depends on whether or not the measurement error is correlated with any subject-specific covariates in the model. If the mismeasured onset ages and covariates are correlated, then the estimated covariate effects would be incorrect (Abrevaya and Hausman, 2004). This could occur, for example, for subjects who do not regularly visit the clinic because they are more symptomatic (covariates) and are in denial. The irregular visits could lead to a mismeasured onset age.

In contrast, if the mismeasured onset ages and covariates are uncorrelated, then the mismeasurements have minimal impact. The estimated covariate effects would be unbiased but more variable (Carroll et al., 2006, chap. 15.1), meaning that the effect of covariates is correct but the power to identify significant features would be reduced. The unbiasedness and reduced power is because the measurement error in the age-at-onset (response) can be absorbed into the residual model error in equation (1) so long as the measurement error is independent of covariates (Abrevaya and Hausman, 2004).

In comparison, logistic regression (Section 2.2) is highly sensitive to error, known in the literature as misdiagnosis or misclassification. Misclassification occurs when a clinician erroneously concludes that a subject has manifest HD (i.e., a DCL score of 4) when s/he

does not, or the converse. Ignoring misclassification in logistic regression has a profound effect: estimates for the slope parameters $\beta_1,\cdots,\beta_p$ in equation (2) are severely biased and have high variability (Magder and Hughes, 1997; Carroll et al., 2006). A range of methods have been developed to address these concerns (Paulmgren and Ekholm, 1987; Copas, 1988; Neuhaus, 1999, 2002; Ramalho, 2002; Prescott and Garthwaite, 2002; Paulino et al., 2003), with the majority revolving around estimates for sensitivity and specificity. Sensitivity is the proportion of individuals who have manifest HD and are correctly identified as having phenoconverted, and specificity is the proportion of individuals who do not have manifest HD onset and are correctly identified as not having phenoconverted.

When sensitivity (denoted by π_1) and specificity (denoted by π_0) are completely unknown, yet believed to be independent of subject-specific covariates (i.e., π_1,π_0 remain constant regardless of the covariate values), then one may use so-called maximum likelihood or Bayesian approaches (Neuhaus, 2002) to unbiasedly estimate the parameters in equation (2). Unfortunately, these methods require extremely large sample sizes to unbiasedly estimate the sensitivity, π_1, and specificity, π_0 (Neuhaus, 2002), which may be impractical in many HD studies.

In the rare case that sensitivity and specificity are known, then unbiased estimates of the parameters can be easily obtained using standard logistic regression techniques (i.e., the method of re-weighted least squares). In this case, while the estimates remain unbiased, their variability is increased as a result of the misclassification.

Otherwise, when sensitivity and specificity can be estimated via validation data (i.e., the diagnostic status Y_{AAO} is correctly observed on a subset of the study data) or replicate data (i.e., subjects are assessed by multiple raters so that we observe multiple diagnostic status Y_{AAO} for each subject), then one may use this additional information to estimate π_1,π_0. Appropriate use of the validation or replicate data allows one to unbiasedly estimate the model parameters (Prescott and Garthwaite, 2002), but care must be taken since incorrect use can result in biased estimates and inference conclusions (Carroll et al., 2006, chap. 15.3.2.3), or low power (Carroll et al., 2006, chap. 15.3.2.5).

Lastly, while there is some methodological developments in measurement error for survival response (Snappin, 1998; Skinner and Humphreys, 1999; Gelfand and Wang, 2000; Richardson and Hughes, 2000; Balasubramanian and Lagakos, 2001, 2003; Meier et al., 2003; Margaret, 2009; Adeniji et al., 2013), the methods have been developed under restrictive modeling assumptions and/or under the assumption of no censoring. Much work is still needed to properly handle censoring and the more complex case of genetic mixture models, which is especially relevant given that genetic information is not always available (Ma and Wang 2014a, Qin et al. 2014).

Acknowledgments

This research is supported in part by the Huntington's Disease Society of America Human Biology Project Fellowship, National Institute of Neurological Disease and Stroke (NS073671, NS082062, 5R01NS40068), National Center for Advancing Translational Sciences (2UL1RR024156-06), CHDI Foundation, and Chamber's Family Fund.

References

Abrevaya J and Hausman JA (2004). Response error in a transformation model with an application to earnings-equation estimation. The Econometrics Journal 7:366–388.

Adeniji AK, Belle SH, and Wahed AS (2013). Journal of Applied Statistics 41:60–72.

Aldrich J (1997). R. A. Fisher and the making of maximum likelihood 19121922. Statistical Science 12:162–176.

Andrew SE, Goldberg YP, Kremer B, et al. (1993). The relationship between trinucleotide (CAG) repeat length and clinical features of Huntington's disease. Nature Genetics 4:398–403. [PubMed: 8401589]

Ayer M, Brunk HD, Ewing GM, et al. (1955). An empirical distribution function for sampling with incomplete information. The Annals of Mathematical Statistics 26:641–647.

Aylward EH (2014). Magenetic resonance imaging striatal volumes: A biomarker for clinical trials in Huntington's disease. Movement Disorders 29:1429–1433. [PubMed: 25164586]

Aylward EH, Liu D, Nopoulos PC, et al. (2012). Striatal volume contributes to the prediction of onset of Huntington disease in incident cases. Biological Psychiatry 71:822–828. [PubMed: 21907324]

Balasubramanian R and Lagakos SW (2001). Estimation of the timing of perinatal transmission of HIV. Biometrics 57:1048–1058. [PubMed: 11764243]

Balasubramanian R and Lagakos SW (2003). Estimation of a failure time distribution based on imperfect diagnostic tests. Biometrika 90:71–182.

Begg CB (2002). On the use of familial aggregation in population-based case probands for calculating penetrance. Journal of the National Cancer Institute 94:1221–1226. [PubMed: 12189225]

Beglinger LJ, O'Rourke JJ, Wang C, et al. (2010). Earliest functional declines in Huntington disease. Psychiatry Research 178:414–418. [PubMed: 20471695]

Biglan K, Zhang Y, Long JD, et al. (2013). Refining the diagnosis of Huntington disease: the PREDICT-HD study. Frontiers in Aging Neuroscience 5: Article 12.

Brandt J, Bylsma FW, Gross R, et al. (1996). Trinucleotide repeat length and clinical progression in Huntington's disease. Neurology 46:527–531. [PubMed: 8614526]

Brinkman RR, Mezei MM, Theilmann J, et al. (1997). The likelihood of being affected with Huntington disease by a particular age, for a specific CAG size. American Journal of Human Genetics 60:1202–1210. [PubMed: 9150168]

Carroll RJ, Ruppert D, Stefanski LA, and Crainiceanu C (2006). Measurement error in nonlinear models: a modern perspective. CRC Press, London, 2nd edn.

Chatterjee N and Wacholder S (2001). A marginal likelihood approach for estimating penetrance from kin-cohort designs. Biometrics 57:245–252. [PubMed: 11252606]

Chen T, Wang Y, Ma Y, et al. (2012). Predicting disease onset from mutation status using proband and relative data with applications to Huntington's disease. Journal of Probability and Statistics 2012:1–19.

Ciarmiello A, Giovacchini G, Orobello S, et al. (2012). ^{18}F-FDG PET uptake in the preHuntington disease caudate affects the time-to-onset independently of CAG expansion size. European Journal of Nuclear Medicine and Molecular Imaging 39:1030–1036. [PubMed: 22526956]

Copas J (1988). Binary regression models for contaminated data (with discussion). Journal of the Royal Statistical Society B 50:1314–1328.

De Boor C (2001). A Practical Guide to Splines (Revised Edition). Springer, New York.

Dempster AP, Laird NM, and Rubin DB (1977). Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society. Series B (Methodological) 39:1–38.

Duyao M, Ambrose C, Myers R, et al. (1993). Trinucleotide repeat length instability and age of onset in Huntington's disease. Nature Genetics 4:387–392. [PubMed: 8401587]

Fan X (2003). Two approaches for correcting correlation attenuation caused by measurement error: implications for research practice. Educational and Psychological Measurement 63:915–930.

Fine JP, Zou F, and Yandell BS (2004). Nonparametric estimation of the effects of quantitative trait loci. Biostatistics 5:501–513. [PubMed: 15475415]

Foroud T, Gray J, Ivashina J, and Conneally PM (1999). Differences in duration of Hunting-ton's disease based on age at onset. Journal of Neurology, Neurosurgery, and Psychiatry 66:52–56.

Gelfand AE and Wang F (2000). Modelling the cumulative risk for a false-positive under repeated screening events. Statistics in Medicine 19:1865–1879. [PubMed: 10867676]

Gutierrez C and Macdonald A (2002). Huntington's disease and insurance I: a model of Huntington's disease. Genetics and Insurance Research Centre (GIRC), Edinburgh.

Gutierrez C and Macdonald A (2004). Huntington's disease, critical illness insurance and life insurance. Scandinavian Actuarial Journal 4:279–313.

Heagerty PJ and Zheng Y (2005). Survival model predictive accuracy and ROC curves. Biometrics 61:92–105. [PubMed: 15737082]

Huntington Study Group (1996). Unified Huntington's disease rating scale: reliability and consistency. Movement Disorder 11:136–142.

Huntington Study Group (1996). Unified Huntington's disease rating scale: reliability and consistency. Movement Disorders 11:136–142. [PubMed: 8684382]

Kaplan EL and Meier P (1958). Nonparametric estimation from incomplete observations. Journal of the American Statistical Association 53:457–481.

Kehoe P, Krawczak M, Harper PS, et al. (1999). Age of onset in Huntington disease: sex specific influence of apolipoprotein E genotype and normal CAG repeat length. Journal of Medical Genetics 36:108–111. [PubMed: 10051007]

Khoury MJ, Beaty TH, and Cohen BH (1993). Fundamentals of genetic epidemiology. Oxford University Press, New York.

King G and Zeng L (2001). Logistic Regression in Rare Events Data. Political Analysis 9:137–163.

Kleinbaum DG and Klein M (2011). Survival analysis: a self-learning text. Springer, New York, 3rd edn.

Langbehn DR, Brinkman RR, Falush D, et al. (2004). A new model for prediction of the age of onset and penetrance for Huntington's disease based on CAG length. Clinical Genetics 65:267–277. [PubMed: 15025718]

Langbehn DR, Hayden MR, Paulsen JS, et al. (2010). 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 153B:397–408. [PubMed: 19548255]

Lee JH, Lee JM, Ramos EM, et al. (2012). TAA repeat variation in the GRIK2 gene does not influence age at onset in Huntington's disease. Biochemical and Biophysical Research Communications 424:404–408. [PubMed: 22771793]

Li JL, Hayden MR, Almqvist EW, et al. (2003). A genome scan for modifiers of age at onset in Huntington disease: The HD MAPS Study. American Journal of Human Genetics 73:682–687. [PubMed: 12900792]

Little RJA and Rubin DB (2002). Statistical analysis with missing data. Wiley, New York, 2nd edn.

Lucotte G, Turpin JC, Riess O, et al. (1995). Confidence intervals for predicted age of onset, given the size of (CAG)n repeat, in Huntington's disease. Human Genetics 95:231–232. [PubMed: 7860073]

Ma Y and Wang Y (2014a). Estimating disease onset distribution functions in mutation carriers with censored mixture data. Journal of the Royal Statistical Society C 63:1–23.

Ma Y and Wang Y (2014b). Nonparametric modeling and analysis of association between Huntington's disease onset and CAG repeats. Statistics in Medicine 33:1369–1382. [PubMed: 24027120]

Maat-Kievit A, Losekoot M, Zwinderman K, et al. (2002). Predictability of age at onset in Huntington disease in the Dutch population. Medicine (Baltimore) 81:251–259. [PubMed: 12169880]

MacDonald ME, Vonsattel JP, Shrinidhi J, et al. (1999). Evidence for the GluR6 gene associated with younger onset of Huntington's disease. Neurology 53:1330–1332. [PubMed: 10522893]

Magder L and Hughes JP (1997). Logistic regression when the outcome is measured with uncertainty. American Journal of Epidemiology 146:195–203. [PubMed: 9230782]

Marder K, Levy G, Louis ED, et al. (2003). Accuracy of family history data on Parkinson's disease. Neurology 61:18–23. [PubMed: 12847150]

Marder K, Zhao H, Myers RH, et al. (2000). Rate of functional decline in Huntington's disease. Neurology 54:452–458. [PubMed: 10668713]

Margaret A (2009). Incorporating validation subsets into discrete proportional hazards models for mismeasured outcomes. Statistics in Medicine 28:1999–2011. [PubMed: 19455509]

Meier A, Richardson B, and Hughes J (2003). Biometrics 59:947–954. [PubMed: 14969473]

Myers RH (2004). Huntington's disease genetics. NeuroRx: the Journal of the American Society for Experimental NeuroTherapeutics 1:255–262. [PubMed: 15717026]

Nance MA, Mathias-Hagen V, Breningstall G, et al. (1999). Analysis of a very large trinucleotide repeat in a patient with juvenile Huntington's disease. Neurology 52:392–394. [PubMed: 9932964]

Neuhaus JM (1999). Bias and efficiency loss due to misclassified responses in binary regression. Biometrika 86:843–855.

Neuhaus JM (2002). Analysis of clustered and longitudinal binary data subject to response misclassification. Biometrics 58.

Panas M, Avramopoulos D, Karadima G, et al. (1999). Apolipoprotein E and presenilin-1 genotypes in Huntington's disease. Journal of Neurology 246:574–577. [PubMed: 10463359]

Paulino CD, Soares P, and Neuhaus J (2003). Binomial regression with misclassification. Biometrics 59:670–675. [PubMed: 14601768]

Paulmgren J and Ekholm A (1987). Exponential family non-linear models for categorical data with errors of observation. Applied Stochastic Models and Data Analysis 3:111–124.

Paulsen JS, Langbehn DR, Stout JC, et al. (2008). Detection of Huntington's disease decades before diagnosis: The Predict HD study. Journal of Neurology, Neurosurgery and Psychiatry 79:874–880.

Paulsen JS, Long J, Ross C et al. (2014). Prediction of manifest Huntington's disease with clinical and imaging measures: a prospective observational study. The Lancet Neurology 13: 1193–1201. [PubMed: 25453459]

Pearson K (1895). Notes on regression and inheritance in the case of two parents. Proceedings of the Royal Statistical Society of London 58:240–242.

Prescott GJ and Garthwaite PH (2002). A simple Bayesian analysis of misclassified binary data with a validation substudy. Biometrics 58:454–458. [PubMed: 12071421]

Qin J, Garcia TP, Ma Y, et al. (2014). Combining isotonic regression and EM algorithm to predict genetic risk under monotonicity constraint and unknown genotypes. Annals of Applied Statistics 8: 1182–1208.

Quarrell O, Donovan KLO, Bandmann O, and Strong M (2012). The Prevalence of Juvenile Huntington's disease: A Review of the Literature and Meta-Analysis. PLOS Currents Huntington Disease. Edition 1, doi: 10.1371/4f8606b742ef3.

Ramalho EA (2002). Regression models for choice-based samples with misclassification in the response variable. Journal of Econometrics 106:171–201.

Ranen NG, Stine OC, Abbott MH, et al. (1995). Anticipation and instability of IT-IF (CAG) N repeats in parent-offspring pairs with Huntington's disease. The American Journal of Human Genetics 57:593–602. [PubMed: 7668287]

Reilmann R, Leavitt BR, and Ross C (2014). Diagnostic criteria for Huntington's disease based on natural history. Movement Disorders 29:1335–1341. [PubMed: 25164527]

Richardson BA and Hughes JP (2000). Product limit estimation for infectious disease data when the diagnostic test for the outcome is measured with uncertainty. Biostatistics 1:341–354. [PubMed: 12933514]

Ross CA, Aylward EH, Wild EJ, et al. (2014). Huntington disease: natural history, biomarkers and prospects for therapeutics. Nature Reviews Neurology 10:204–2016. [PubMed: 24614516]

Rubin D (1996). Multiple imputation after 18+ years. Journal of American Statistical Association 91:473–489.

Rubinsztein DC, Leggo J, Chiano M, et al. (1997). Genotypes at the GluR6 kainate receptor locus are associated with variation in the age of onset of Huntington disease. Proceedings of the National Academy of Sciences of the United States of America 94:3872–3876. [PubMed: 9108071]

Simonin C, Duru C, Salleron J, et al. (2013). Association between caffeine intake and age at onset in Huntington's disease. Neurobiology of Disease 58:179–182. [PubMed: 23732677]

Skinner CJ and Humphreys K (1999). Weibull Regression for Lifetimes Measured with Error. Lifetime Data Analysis 5:23–37. [PubMed: 10214000]

Snappin SM (1998). Survival analysis with uncertain endpoints. Biometrics 54:209–218. [PubMed: 9574966]

Snell RG, MacMillan JC, Cheadle JP, et al. (1993). Relationship between trinucleotide repeat expansion and phenotypic variation in Huntington's disease. Nature Genetics 4:393–397. [PubMed: 8401588]

Stine OC, Pleasant N, Franz ML, et al. (1993). Correlation between the onset age of Huntington's disease and length of the trinucleotide repeat in IT-15. Human Molecular Genetics 2:1547–1549. [PubMed: 8268907]

Stout JC, Paulsen JS, Queller S, et al. (2011). Neurocognitive signs in prodromal Huntington disease. Neuropsychology 25:1–14. [PubMed: 20919768]

Tabrizi SJ, Scahill RI, Durr A, et al. (2011). Biological and clinical changes in premanifest and early stage Huntington's disease in the TRACK-HD study: The 12-month longitudinal analysis. The Lancet Neurology 10:31–42. [PubMed: 21130037]

Tabrizi SJ, Scahill RI, Owen G, et al. (2013). Predictors of phenotypic progression and disease onset in premanifest and early-stage Huntington's disease in the TRACK-HD study: analysis of 36-month observational data. The Lancet Neurology 12:637–649. [PubMed: 23664844]

Trembath MK, Horton Z, Tippett L, et al. (2010). A retrospective study of the impact of lifestyle on age at onset of Huntington disease. Movement Disorders 25:1444–1450. [PubMed: 20629137]

Trottier Y, Biancalana V, and Mandel JL (1994). Instability of CAG repeats in Huntington's disease: relation to parental transmission and age of onset. Journal of Medical Genetics 31:377–82. [PubMed: 8064815]

Tsiatis AA (2006). Semiparametric theory and missing data. Springer, New York.

Vittinghoff E, Glidden DV, Shiboski SC, and McCulloch CE (2012). Regression Methods in Biostatistics. Springer, New York, 2nd edn.

Vuillaume I, Vermersch P, Destée A, et al. (1998). Genetic polymorphisms to the CAG repeat influence clinical features at onset in Huntington diseases. Journal of Neurology, Neurosurgery, and Psychiatry 64:758–762.

Wacholder S, Hartge P, Struewing J, et al. (1998). The kin-cohort study for estimating penetrance. American Journal of Epidemiology 148:623–630. [PubMed: 9778168]

Walker F (2007). Huntington's disease. Lancet 369:218–228. [PubMed: 17240289]

Wang Y, Garcia TP, and Ma Y (2012). Nonparametric estimation for uncensored mixture data with application to the Cooperative Huntington's Observational Research Trial. Journal of the American Statistical Association 107:1324–1338. [PubMed: 24489419]

Zhang Y, Long JD, Mills JA, et al. (2011). Indexing disease progression at study entry with individuals at-risk for Huntington disease. American Journal of Medical Genetics Part B: Neuropsychiatric Genetics 156B:751–763.

Table 1

DEFINITIONS OF HD DIAGNOSIS.

Pre-UHDRS	Age of first clearly defined abnormality (Andrew et al., 1993) or combination of abnormalities (Vuillaume et al., 1998) such as involuntary movement, psychiatric or cognitive abnormality or inability to perform complex movements.
Motor-based diagnosis (Huntington Study Group, 1996)	Age when a clinician believes with ≥ 99% confidence that the subject's extrapyramidal signs are unequivocally associated with HD (i.e., DCL=4).
Multidimensional diagnosis (Biglan et al., 2013)	Age when a clinician believes with ≥ 99% confidence that based on the entire UHDRS (motor, cognitive, behavioral, and functional components), the subject has manifest HD.
Natural history-based diagnosis (Reilmann et al., 2014)	Age when a clinician believes with ≥99% confidence that based on the entire UHDRS (motor, cognitive, behavioral and functional components) and all available history, the subject has manifest HD. The definition further involves different criteria for "genetically confirmed" subjects (i.e., ≥36 CAG repeats) and those who are not. That is, a subject has manifest HD if • The subject, whether genetically confirmed or not, exhibits significant cognitive symptoms, shows evidence of progression from previous UHDRS exams, and has DCL ≥2. • Else, the subject exhibits significant changes in the Total Functional Capacity (TFC) and Functional Assessment (Marder et al., 2000; Beglinger et al. (2010); Tabrizi et al., 2013) that are only attributable to HD. In addition, the subject is genetically confirmed and has DCL ≥ 3; or the subject is genetically unconfirmed and has DCL 4.

Table 2

ADVANTAGES AND DISADVANTAGES OF DIFFERENT REGRESSION METHODS FOR MODELING AGE OF MOTOR-ONSET.

Method	Advantages	Disadvantages
Correlation Analysis	• Simple and available in standard software. • Assesses strength of linear relationship between two variables (e.g., age-at-onset and subject-specific features).	• Missing observations are dropped from the calculation. • Results can be perturbed by outliers. • Does not reveal any nonlinear relationship between two variables.
Linear Regression	• Simple and available in standard software. • Quantifies the linear effect of subject-specific features (covariates) on age-at-onset (response). • Assesses how much of the variability of age-atonset is explained by the features (i.e., coefficient of determination).	• Missing observations are dropped from the analysis. • Results can be perturbed by outliers.
Logistic Regression	• Simple and available in standard software. • Assesses the linear effect that subject-specific features have on the log odds of onset occurring. • Extracts information from individuals who have experienced onset and from those who have not by incorporating a binary response variable in the model.	• Ignores time differences between subjects experiencing onset; that is, subjects who experience onset at different times are treated similarly. • Loss of time information can be remedied with multiple binary response variables but can result in numerically unstable estimation.

Table 3

ADVANTAGES AND DISADVANTAGES OF DIFFERENT SURVIVAL TYPE METHODS FOR MODELING AGE OF MOTOR-ONSET WHEN GENETIC MUTATION STATUS IS KNOWN.

Method	Advantages	Disadvantages
Parametric methods (Gutierrez and MacDonald, 2002; Langbehn et al., 2004; Zhang et al., 2011)	• Simple implementation. • Predicts likeliness of onset occurring by a given age as an explicit function of subject-specific features. • Predicts probability of onset in *s*-year intervals (e.g., 5-year intervals). • Estimates penetrance rates at different CAG repeat-lengths. • Provides clinically meaningful divisions of subjects to make crosssectional comparisons.	• Current parametric methods only incorporate CAG repeat-length and age-at-study entry, although many other subject-specific features have been identified to influence age-at-onset. • Searching for explicit relationships between multiple subject-specific features and age-at-onset is challenging. • Prone to model misspecification.
Semiparametric Cox model	• Available in standard software. • Relates linear effect of subject-specific features to a ratio of hazard functions through a log-transformation. • Baseline hazard function left completely unspecified to allow model flexibility.	• Proportional hazards assumption not always satisfied in HD studies (Langbehn et al., 2004).
Nonparametric Kaplan-Meier	• No assumptions about underlying age-at-onset distribution. • Estimates age-at-onset distribution stratified by different subject-specific features (e.g., CAG repeat-length).	• No explicit relationship between the effect of CAG repeat-length and the age-at-onset distribution.
Nonparametric Kernel and Smoothing Splines (Ma and Wang, 2014b)	• Effect of subject-specific features estimated using smooth functions that do not necessarily adhere to any particular explicit formula.	• Computationally demanding. • No explicit final model available and hence no clinically meaningful divisions of subjects.

Table 4

ADVANTAGES AND DISADVANTAGES OF DIFFERENT SURVIVAL TYPE METHODS FOR MODELING AGE OF MOTOR-ONSET WHEN GENETIC MUTATION STATUS IS UNKNOWN.

Method	Advantages	Disadvantages
1 Parametric model (Chen et al. 2012) **2 Nonparametric model** (Ma and Wang, 2014b)	• Extends Langbehn et al. (2004) model to genetic mixture models for kin-cohort studies. • Ma and Wang (2014b) model assumes nonparametric forms of subject-specific effects, whereas Chen et al (2012) uses same explicit parametric form as Langbehn et al. (2004). • Consistent estimator.	• Overly simplified assumption of the gamete transmission process: assumes a family member inherits the same CAG repeat-length as his proband if the proband has the gene mutation. • Ma and Wang (2014b) model is computationally demanding.
Type II Nonparametric Maximum Likelihood Estimator (NPMLE) (Chatterjee and Wacholder, 2001)	• Directly maximizes the nonparametric likelihood using an Expectation-Maximization algorithm.	• Biased and unreliable estimates of cumulative risk of onset. • Computationally demanding.
1 Type I NPMLE (Wacholder et al., 1998) **2 Independent NPMLE** (Fine et al., 2004) **3 Inverse Probability Weighting** (IPW) estimator (Wang et al, 2012) **4 Imputation Estimator** (Wang et al, 2012) **5 Weighted least squares estimator** (Ma and Wang, 2014a) **6 Isotone regression** (Qin et al., 2014)	• Consistent estimator. • Resulting estimated cumulative risk curve can - Serve as time-dependent positive and negative predictive values of the HD gene mutation test. - Provide a numerical summary of cumulative risk associated with a positive mutation test. - Predict the risk of onset for a subject based on his genetic test result and demographic information. - Predict conditional probabilities of developing HD in next *s*-years. • Augmented IPW estimator and Imputation estimator have least variability. • Weighted least squares estimator is easiest to compute. • Isotone regression estimator is guaranteed to satisfy the mathematical properties of a distribution function.	• Type I NPMLE has high variability and its estimates of cumulative risk of onset disagrees with clinical findings. • Independent NPMLE can violate mathematical constraints on probability risk. • IPW has high variability. • IPW, Augmented IPW are susceptible to division by zero. • IPW, Augmented IPW, Imputation and Weighted least squares estimators are not guaranteed to satisfy the mathematical properties of a distribution function and may violate the constraints of a probability where values must be between zero and one. • Imputation and isotone regression estimator are computationally demanding.

Table 5

EFFECT OF MISMEASURED ONSET AGES IN DIFFERENT METHODS FOR MODELING AGE OF MOTOR-ONSET.

Method	Effect of mismeasured onset age
Correlation Analysis	• Pearson correlation is biased downward.
Linear Regression	• Biased estimated effects of subject-specific features (covariates) if the measurement error in age-at-onset is correlated with the features.
Logistic Regression	• Highly sensitive to response error (i.e., misclassification). Ignoring misclassification leads to estimated model parameters being severely biased and highly variable.
Survival Models	• Evidence of bias but exact effects need to be explored, especially for genetic mixture models where genetic information is not always available.