Article

SMMR
STATISTICAL METHODS IN MEDICAL RESEARCH

Estimating the age at onset distribution of the asymptomatic stage of a genetic disease based on pedigree data

Statistical Methods in Medical Research
2020, Vol. 29(8) 2344–2359

DOI: 10.1177/0962280219893400
journals.sagepub.com/home/smm
SAGE

Marianne A Jonker[1], Priya Vart[1] and Mar Rodriguez Girondo[2]

Abstract

Information on the age at onset distribution of the asymptomatic stage of a disease can be of paramount importance in early detection and timely management of that disease. However, accurately estimating this distribution is challenging, because the asymptomatic stage is difficult to recognize for the patient and is often detected as an incidental finding or in case of recommended screening; the age at onset is often interval-censored. In this paper, we propose a method for the estimation of the age at onset distribution of the asymptomatic stage of a genetic disease based on ascertained pedigree data that take into account the way the data are ascertained to overcome selection bias. Simulation studies show that the estimates seem to be asymptotically unbiased. Our work is motivated by the analysis of data on facioscapulohumeral muscular dystrophy, a genetic muscle disorder. In our application, carriers of the genetic causal variant are identified through genetic screening of the relatives of symptomatic carriers and their disease status is determined by a medical examination. The estimates reveal an early age at onset of the asymptomatic stage of facioscapulohumeral muscular dystrophy.

Keywords

Age at onset, ascertainment, current status data, family data, outcome-dependent sampling

1. Introduction

A reliable and accurate estimate of the age at onset distribution of a disease is of great importance for optimizing follow-up protocols of high-risk patients, aiming at early detection of the disease and timely start of treatment. For most diseases treatment is more effective if it is started in an early stage than at a later moment when the disease has progressed.

Diseases usually progress in stages. The early stage is typically asymptomatic. In this phase of the disease the patient is not aware of having the disease, because no symptoms with which the disease is usually associated are experienced. Diagnosis is most often in the symptomatic stage when symptoms appear. Although the disease is not apparent during the asymptomatic stage, some pathological changes may be detectable with a medical test. For common diseases like breast and colon cancer, population screening programs have been set up in many countries to identify the disease process during this asymptomatic phase so that intervention can be started at an early stage in the disease process. However, population screening is only offered for a limited scope of diseases, because of cost-effectiveness and possible profit for the patient. Small-scale screening is performed in high-risk sub-populations, typically characterized for certain rare genetic variants.[1,2]

[1]Department for Health Evidence, Section Biostatistics, Radboud University Medical Center, Nijmegen, the Netherlands
[2]Department of Biomedical Data Sciences, Leiden University Medical Center, Leiden, the Netherlands

Corresponding author:
Marianne A Jonker, Department for Health Evidence, Section Biostatistics, Radboud University Medical Center, Geert Grooteplein-Noord 21, Nijmegen 6525 EZ, the Netherlands.
Email: marianne.jonker@radboudumc.nl

Since genetic traits aggregate within families, these high-risk sub-populations are often identified via family members carrying the genetic variant and being affected by the disease. Specifically, families who satisfy certain ascertainment criteria (i.e. "at least one affected carrier") are selected and all family members (up to a certain degree) are invited for a genetic test. Individuals who carry the genetic variant of interest may be then (regularly) screened for the presence of the disease in the asymptomatic stage. The screening scheme and especially the age at which screening starts should depend on the risk the carrier will become asymptomatic. To determine this risk, an estimate of the age at onset distribution for the asymptomatic stage is needed.

Estimation of this age at onset distribution for the asymptomatic stage is challenging, because it should rely on data of selected families (based on the presence of the disease). Moreover, the exact age at onset is never observed for any of the individuals in the data-set. Instead, at every moment of screening it is observed whether an individual has the disease in asymptomatic stage or not; age at onset is interval-censored by the ages at time of screening. Sometimes, if the disease is not lethal or no treatment is available, individuals are screened only once. Then, only the age at time of screening is observed, as well as whether or not the individual is asymptomatic before examination took place. This type of censoring is referred to as interval-censored type I or current status.[3]

With interval-censored data, the exact age at onset is never observed for any individual; therefore, this kind of censoring differs greatly from right censoring. If the data are obtained from a population of independent, non-selected individuals, for instance with population screening, the distribution of the age at onset of the asymptomatic disease can be estimated with the non-parametric maximum likelihood estimator (NPMLE) (see, e.g. Zhang and Sun,[3] Groeneboom and Wellner,[4] Jewell and van der Laan,[5] and Witte et al.[6]). In case of selected families, one often tries to correct for ascertainment bias by leaving out the index patients and assuming independence between the relatives. Under this independence assumption, the NPMLE could be used, but the estimator is still biased because of the suboptimal ascertainment correction (leaving out the index): it does not take into account the fact that ascertainment is based on the phenotype of all individuals in the pedigree and not on one particular person. The bias is especially large in data-sets with small pedigrees.[7] Moreover, leaving out the data of the index patients from the analysis, means that valuable information is discarded, what is especially unfortunate in rare diseases for which data-sets are often small. A more sophisticated analysis has to be performed to properly correct for ascertainment and to use the data optimally. This could be done by considering an adjusted likelihood which conditions on the ascertainment event, for instance by maximizing the retrospective likelihood, the prospective likelihood, or a joint conditional likelihood (e.g. Carayol and Bonaiti-Pellie[8] and Kraft and Thomas[9]).

In this paper, we propose a maximum likelihood-based method, adjusted for the ascertainment, to estimate the distribution of the age at onset of the asymptomatic stage of a disease. The likelihood function has a complex form as it takes into account all the available information: the interval-censored age at onset of the asymptomatic stage, the (right-censored) age at onset of the symptomatic stage, the ascertainment criteria, and also family characteristics that may affect the age at onset distributions. Simulation studies are performed to study the performance of the proposed estimator.

For many rare disease susceptible genetic variants, the number of detected pedigrees with this variant is small, like for the disorder that motivated this research. This complicates estimation of the age at onset distributions and hence the proposed statistical model cannot be too large (too many unknown parameters) to overcome overfitting of the data. A balance has to be found between the complexity of the model and the information in the data. The proposed model can also be applied if the sample size is small.

This work is motivated by a study on facioscapulohumeral muscular dystrophy (FSHD), a genetic muscle disorder. In order to learn more about the progression and causes of the disease, a cross-sectional observational study of families with at least two symptomatic family members was performed. All family members were offered a genetic test and a physical examination to determine if they had already entered the asymptomatic stage of the disease. No extra data were collected after the moment of examination. The age at onset for the asymptomatic stage is hence type I interval-censored. The date of examination was the same for all the participants in the study (cross-sectional).

The rest of the paper is organized as follows. In the second section, we introduce notation and establish a general framework for the problem. Ascertained-corrected conditional likelihood estimation is proposed in the third section. A simulation study is presented in the fourth section, while in the fifth section the methods are applied to a data-set of familial FSHD in the Netherlands. Main conclusions and a final discussion follow in the final section.

2. Data, notation, and assumptions

In case of genetic diseases, carriers of the disease susceptible variant are often identified via already detected carriers; relatives of affected carriers are invited for genetic testing. For data analysis and estimation of the age at onset distribution of the disease, usually only data of the proven carriers from ascertained families are included in the data-set (and the non-carriers are left out). Below, notation for carriers of the causal variant is introduced.

Let U and T be the ages at onset of the asymptomatic and symptomatic stage, with $U \leq T$ almost surely. The age at examination (genetic test and physical examination) is denoted by C which we assume to be independent of U and T. This is a reasonable assumption in our setting: due to the cross-sectional nature of the study all pedigrees fulfilling the ascertainment criteria are retrospectively selected and subsequently screened at the same chronological time (see also Section 5, the application). Further, define the indicator function $\Delta = I(T \leq C)$ as 1 if $T \leq C$ and 0 otherwise. The indicator function $\Sigma = I(U \leq C)$ is defined similarly.

The three possible configurations of observations are $C < U < T$, $U \leq C < T$, and $U \leq T \leq C$. In the first situation, $C < U < T$, the individual does not have any symptoms of the disease at the time of examination; U and T might occur after C, but it could also be that they never occur before the death of the individual. In the second situation, $U \leq C < T$, the disease is diagnosed during the examination, but the individual does not present any symptoms yet; the individual is in the asymptomatic stage. In the third situation, $U \leq T \leq C$, the individual has the disease and also notices symptoms at the time of examination, the individual is symptomatic.

We assume that all individuals with a positive genetic test for the variant of interest are physically examined and included in the data-set. For a single carrier, the triple $(\Sigma, T\Delta, C)$ is observed. The age at time of onset of symptoms, T, is only observed if $T \leq C$, so if $\Delta = 1$. This means that the age of appearance of symptoms is known for symptomatic patients only. Further, U is never observed, but the presence or absence of symptoms at examination allows to observe Σ, i.e. we can determine if the asymptomatic stage is entered before or after examination. The age at time of examination, C, is always observed (if an individual is symptomatic at age C, (s)he will not be examined again, but it is known at what time this should have taken place if the individual was not symptomatic). The observed data for each of the three situations are represented in Figure 1.

We assume that U and T follow (parametric) distributions, $U|x \sim H_{\eta,x}$ and $T|x \sim F_{\theta,x}$, where x is a vector of covariates and η and θ are unknown parameters. The covariates in $H_{\eta,x}$ and $F_{\theta,x}$ may differ, but for ease of notation we use x for both. The distribution for C, the age at time of examination, is denoted by G and is left unspecified for the moment. An overview of the notation is given in Table 1. The number of pedigrees in the data-set is denoted as r, the number of carriers in these pedigrees as $n_j, j = 1, \ldots, r$, and the total number of carriers in the data-set as $n = \sum_{j=1}^{r} n_j$.

In the paper it is assumed that, conditional on the covariates and the genetic status of the individuals, the ages at onset of the asymptomatic and the symptomatic stages are independent between individuals.

For tested non-carriers, the ages at time of genetic testing are known (from the records). These data can be used for estimating the distribution G. This is under the assumption that the distributions of the age at testing are equal for carriers and non-carriers. This is a reasonable assumption, because carriers and non-carriers do not know their carrier status when they decide to be genetically tested or not.

3. Ascertained-corrected conditional likelihood estimation

Our main goal is to obtain an accurate estimator of the distribution of the age at onset of the asymptomatic stage for carriers of the variant, $H_{\eta,x}$. We propose a two-stage approach. The main idea is to first estimate G and θ by maximizing a likelihood function, L^{symp}, for the symptomatic data $(T\Delta, C)$ only. These estimates are inserted in the likelihood function, L^{full}, for all data $(\Sigma, T\Delta, C)$ and this function is maximized with respect to η.

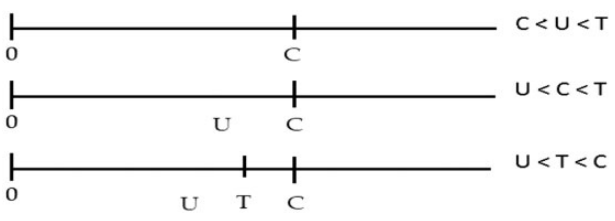

Figure 1. Presentation of the observations in the three configurations. There are no observations after C and the exact moment U took place is not indicated, because this is unknown.

Table 1. Definition of variables and distributions.

Variable	Meaning	Distribution
T	Age at onset of symptomatic disease for a carrier	$F_{\theta,x}, f_{\theta,x}$
U	Age at onset of asymptomatic disease for a carrier	$H_{\eta,x}, h_{\eta,x}$
C	Age at time of examination/screening for a carrier	G, g
Δ	Indicator function $\Delta = I(T \leq C)$	
Σ	Indicator function $\Sigma = I(U \leq C)$	

In the derivation of the likelihood functions, we assume that, conditional on the covariates, the observations of all carriers are independent, notwithstanding the pedigrees they belong to. In the notation, no discrimination with respect to the pedigrees is necessary and all carriers are identified with a single and unique index i: the variables $(U_i, T_i, C_i, \Sigma_i, \Delta_i)$ for individual i.

In Appendix 1, a derivation is given of the ascertainment-corrected prospective likelihood functions L^{symp} and L^{full}. The likelihood function based on the symptomatic data $(T_i\Delta_i, C_i)$, $i = 1, \ldots, n$ only, is given by

$$L^{symp} = \frac{\prod_{i=1}^{n} f_{\theta,x_i}(T_i)^{\Delta_i}(1 - F_{\theta,x_i}(C_i))^{1-\Delta_i} g(C_i)}{\prod_{j=1}^{r} \mathrm{P}(A_j)} \tag{1}$$

where A_j is the ascertainment event for pedigree j and $\mathrm{P}(A_j)$ the probability this event occurred. The ascertainment-corrected prospective likelihood function, L^{full}, for the full data $(\Sigma_i, T_i\Delta_i, C_i)$, $i = 1, \ldots, n$ is given by

$$L^{full} = \frac{\prod_{i=1}^{n} g(C_i)(1 - H_{\eta,x_i}(C_i))^{(1-\Sigma_i)(1-\Delta_i)}(H_{\eta,x_i}(C_i) - F_{\theta,x_i}(C_i))^{\Sigma_i(1-\Delta_i)} f_{\theta,x_i}(T_i)^{\Sigma_i\Delta_i}}{\prod_{j=1}^{r} \mathrm{P}(A_j)} \tag{2}$$

Since the pedigrees are not randomly sampled from the population, the likelihood functions are conditional on the ascertainment events $A_j, j = 1, \ldots, r$. The exact expression of the probability $\mathrm{P}(A_j)$ in the denominator of the likelihood functions depends on the ascertainment rules. These rules may vary over studies, depending on the severity and prevalence of the disease, but are usually based on observations on the symptomatic stage only. As a consequence, the product $\prod_{j=1}^{r} \mathrm{P}(A_j)$ is a function of the distributions G and $F_{\theta,x}$ only. As an example, suppose that a pedigree is ascertained if at least one individual among the carriers is symptomatic at the time of examination. The probability this ascertainment event occurs for family j, $\mathrm{P}(A_j)$, equals 1 minus the probability that none of the carriers was symptomatic. That means that

$$\mathrm{P}(A_j) = 1 - \left(\int 1 - F_{\theta,x} \mathrm{d}G \right)^{n_j} \tag{3}$$

In Section 5, we give an explicit expression of $\mathrm{P}(A_j)$ for our real data application based on the ascertainment of at least two symptomatic carriers at the time of examination.

3.1. Estimation of parameters

From a theoretical perspective the full likelihood, L^{full}, could be maximized with respect to all unknown parameters to obtain their maximum likelihood estimates. However, in practice we noticed that optimization algorithms often do not converge to the global maximum, probably because the parameter space is too big and the algorithm stops at a local and not the global maximum. Therefore, we propose to perform the estimation of the unknown parameters in the model in two steps. Simulation studies in Section 4 show good performance of this two-step procedure. The two steps are given by:

1. Distribution G and parameter θ are estimated based on the likelihood function L^{symp} in equation (1). More details are given below.

2. After inserting the estimates $\hat{G}$ and $F_{\hat{\theta},x}$ into the likelihood function L^{full} in equation (2), it is maximized with respect to η. If the ascertainment is based on the symptomatic stage only (what is usually the case), this boils down to maximizing

$$\prod_{i=1}^{n} (1 - H_{\eta,x_i}(C_i))^{(1-\Sigma_i)(1-\Delta_i)} (H_{\eta,x_i}(C_i) - F_{\hat{\theta},x_i}(C_i))^{\Sigma_i(1-\Delta_i)}$$

with respect to η.

In the first step of the estimation algorithm, G and θ are estimated without using data on the asymptomatic status (Σ). In principle, this does not affect the unbiasedness of the estimators, but because not all information is used standard errors might slightly increase. If G is assumed to follow a parametric distribution, G and θ can be estimated by maximizing the likelihood function L^{symp} in equation (1). Alternatively, G could be estimated by the empirical distribution of the ages at time of examination of the relatives with a negative genetic test (non-carriers). This estimator is asymptotically unbiased. If no data of non-carriers are available, the ages at examination of the carriers could be used instead. This yields an estimator that is asymptotically slightly biased. An extensive discussion on this topic, including simulation studies, is given in Jonker et al.[7]

Our two-step estimation procedure implies that two maximizations are required, but the dimension of the parameters space in each maximization problem is reduced. This lowers the computational complexity of the problem. Of course, the likelihood L^{full} in equation (2) could also be maximized with respect to all parameters simultaneously, but, as mentioned before, the two-step procedure shows more stable results (this has been checked with simulations).

3.2. Variance estimation

Variance estimates of $\hat{\eta}$ and $\hat{\theta}$ are derived with a parametric bootstrap. A bootstrap resample is obtained as follows:

1. Randomly select a pedigree from the original data-set. Say pedigree j is selected.
2. For each of the n_j family members i of the selected pedigree j, draw U_i^* and T_i^* from $H_{\hat{\eta},x_i}$ and $F_{\hat{\theta},x_i}$, respectively as described in Section 4.1.
3. To guarantee that the resulting bootstrap resample is similar to the original data-set in terms of family size and structure, the previous step is repeated until the simulated phenotypes of the carriers in the pedigree satisfy the ascertainment condition.

These three steps are repeated until the bootstrap resample contains the same number of pedigrees as the original data-set. For each bootstrap resample b, we apply the proposed estimation procedure described in Section 3.1. to obtain bootstrap estimates of the parameters of interest $(\hat{\eta}^b, \hat{\theta}^b)$. This procedure is repeated a large number of times, B, and the variances of $\hat{\eta}$ and $\hat{\theta}$ are computed empirically from the B bootstrap resamples. Confidence intervals for θ and η can be constructed based on these variance estimates. Further, pointwise confidence intervals for $H_\eta(t)$ and $F_\theta(t)$ can be constructed by their 2.5 and 97.5% quantiles of the bootstrap estimates of $H_{\hat{\eta}^b}(t)$ and $F_{\hat{\theta}^b}(t)$.

By constructing the confidence intervals for the unknown parameters in this way, the inaccuracy of the estimators for G and θ is reflected in the width of the confidence interval for η.

4. Simulation study

We conduct a simulation study to illustrate the performance of the proposed estimation procedure. The setup and the results are described below.

4.1. Simulation setup

We assume that the variables for age at onset of the asymptomatic and symptomatic stages, U and T, follow gamma distributions. Further, inspired by our real data example, we assume that the variable for the age at time of examination, C, is independent of (U, T). We consider three basic scenarios which resemble relevant situations in practice. In scenario 1, we assume a situation in which the asymptomatic stage is short; the onset of the asymptomatic and symptomatic phases of the disease is close to each other. Namely, we assume that the expected

age at onset of the asymptomatic stage, U, has mean $EU = 60$ with standard deviation $sd(U) = 8.5$, and that the expected age at onset of the symptomatic stage, T, is $ET = 66$ with $sd(T) = 8.9$. Scenario 2 corresponds to a situation with an average longer time between the onset of the asymptomatic and symptomatic phases. Specifically, we assume that the expected age at onset of the asymptomatic stage, U, has mean $EU = 48$ and $sd(U) = 7.6$, and that the expected age at onset of the symptomatic stage, T, is $ET = 72$ with $sd(T) = 9.2$. Scenario 3, which closely resembles our real data application, is characterized by an early expected age at onset of the asymptomatic phase ($EU = 20$), large latency period before the onset of the symptomatic phase ($ET = 60$), and large variation in the age-of-onset of both the asymptomatic and symptomatic phases ($sd(U) = 20$ and $sd(T) = 34.6$) (see Figure 2 for the curves). In the steps below, the corresponding shape and scale parameters are given.

We simulate $M = 1000$ Monte Carlo trials, each consisting of m families (m is taken equal to 100, 500, or 1000). Additionally, in scenario 3 we also considered $m = 25$ to mimic our real data setting with regard to the reduced number of observed families. $m = 25$ was not considered in scenarios 1 and 2 since it led to extremely low numbers of observed families so that inference was meaningless.

For each family j, $j = 1, \ldots, m$, we follow the steps given below:

1. Simulate family size n_j. In order to check the impact of the family size on the performance of our method, two situations are considered: populations composed of "small" families (n_j is sampled from $\{1, 2, 3, 4\}$ with equal probability) and populations composed of "large" families (n_j is sampled from $\{3, 6, 9, 12\}$ with equal probability).
2. For each family member $i, i = 1, \ldots, n_j$, simulate U_i from a gamma distribution with shape and scale parameters k_1 ($k_1 = 50$ in scenario 1, $k_1 = 40$ in scenario 2, $k_1 = 1$ in scenario 3) and θ_1 ($\theta_1 = 1.2$ in scenarios 1 and 2, $\theta_1 = 20$ in scenario 3).
3. For each family member i, $i = 1, \ldots, n_j$, simulate the time between entering the asymptomatic and the symptomatic stage, $\tilde{U}_i$, from a gamma distribution with shape and scale parameters $\tilde{k}_1$ ($\tilde{k}_1 = 5$ in scenario 1, $\tilde{k}_1 = 20$ in scenario 2, $\tilde{k}_1 = 2$ in scenario 3) and scale parameter θ_1 (equal to the scale parameter used to generate U_i).
4. For each family member i, $i = 1, \ldots, n_j$, define $T_i = U_i + \tilde{U}_i$ as the age at onset of the symptomatic stage of the disease. Since U_i and $\tilde{U}_i$ are independent, T_i follows a gamma distribution with shape parameter $k_1 + \tilde{k}_1$ and scale parameter $\theta_1 = 1.2$ in scenarios 1 and 2 and 20 in scenario 3.
5. For each family member i, $i = 1, \ldots, n_j$, simulate the age at time of examination C_i from a uniform distribution at the interval [20, 70].
6. If the ascertainment event A_j occurred, the n_j carriers of family j are ascertained. In our simulation setting, families are selected if at least one symptomatic family member is identified (at least one family member with $T_i \leq C_i$).
7. Independently of the simulations in the previous steps, a sample of 1000 observations from the uniform distribution at [20, 70] is simulated (the same distribution from which was simulated in step 5). The simulated values represent the ages at time of genetic testing of the individuals who got a negative test result. These data are used to estimate G (as explained in Section 3.1).

In the simulation study, families are ascertained if at least one family member is symptomatic at the time of examination, like in the example given earlier. As a result of the ascertainment, the effective sample that is used for estimation is smaller than m families; the effective sample size is denoted by r. To compute confidence intervals of the estimators, we set the number of bootstrap resamples equal to $B = 500$.

In practice, often some carriers in an ascertained pedigree do not participate in the study. It is likely that those carriers do not show symptoms of the disease. To check the impact of this form of informative missing data, we perform a second simulation study in which every non-symptomatic family member ($\Delta = 0$) is excluded from the sample with a probability of either 0.20 or 0.50.

All computations are performed using the statistical software R (R Core Team, 2018). The function `nlminb` `R` function is employed to maximize the log-likelihood functions (1) and (2).

4.2. Simulation results

The main results of the simulation study are shown in Figure 2 and in Table 2. In each graphic in Figure 2, the solid black and gray lines represent the true distributions of $H_{(k_1,\theta_1)}$ and $F_{(k_1+\tilde{k}_1,\theta_1)}$, respectively. The dashed lines represent the 2.5 and 97.5% estimated pointwise percentile curves across the M = 1000 Monte Carlo trials.

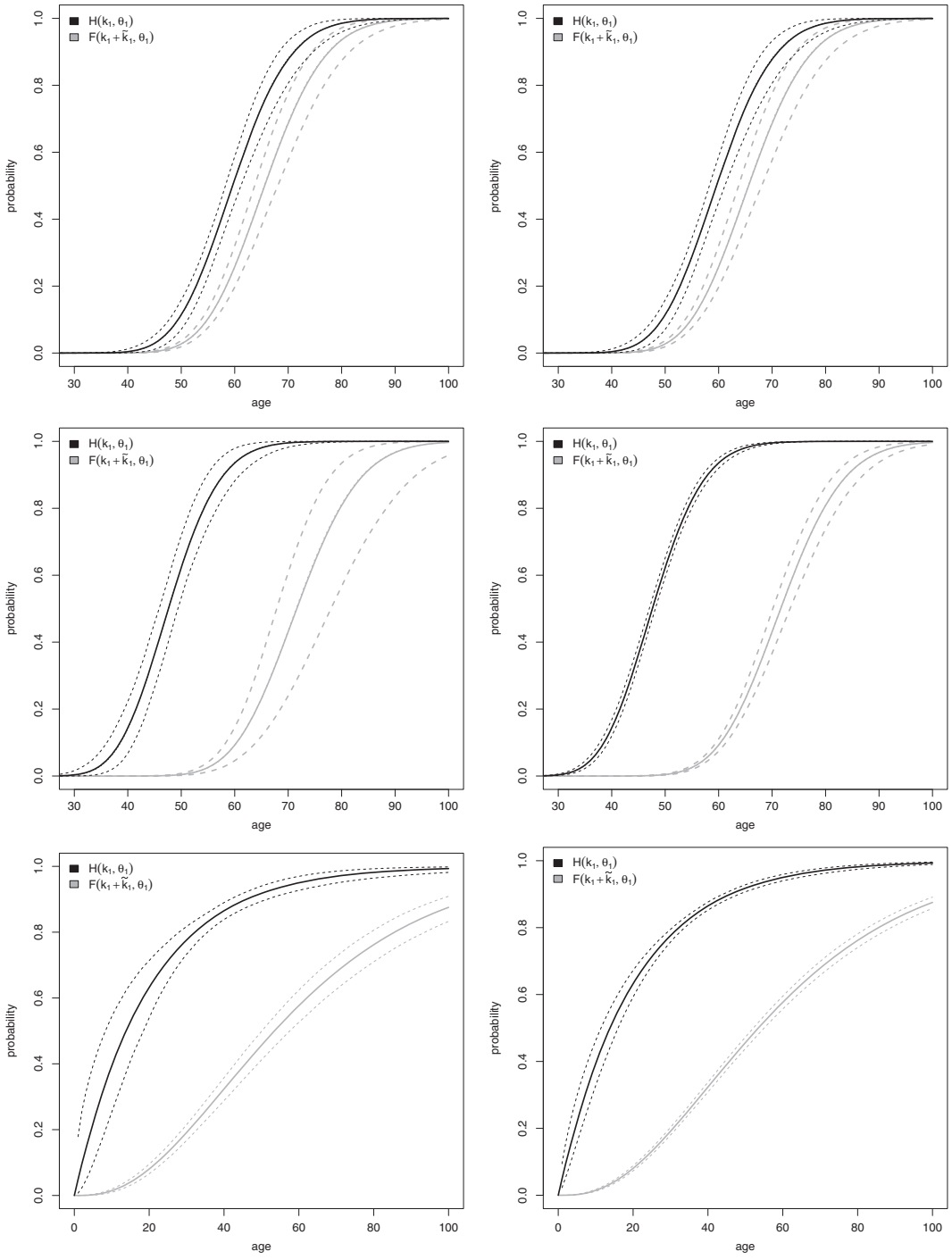

Figure 2. Results of the simulation study with $m = 500$ families. Top: scenario 1 ($k_1 = 50$, $\tilde{k}_1 = 5$, $\theta_1 = 1.2$). Middle: scenario 2 ($k_1 = 40$, $\tilde{k}_1 = 20$, $\theta_1 = 1.2$). Bottom: scenario 3 ($k_1 = 1$, $\tilde{k}_1 = 2$, $\theta_1 = 20$). Left panels: small families ($n_j \in \{1, 2, 3, 4\}$). Right panels: large families ($n_j \in \{3, 6, 9, 12\}$). Solid black and gray lines are the true distribution functions $H_{(k_1,\theta_1)}$ and $F_{(k_1+\tilde{k}_1,\theta_1)}$, respectively. Dashed lines form a band based on estimated pointwise 2.5 and 97.5% percentiles based on $M = 1000$ Monte Carlo trials.

The left graphics in Figure 2 refer to the "small families" situation ($n_j \in \{1, 2, 3, 4\}$), whereas the right graphics refer to the "large families" situation ($n_j \in \{3, 6, 9, 12\}$). Top panels refer to scenario 1 with $k_1 = 50$, $\tilde{k}_1 = 5$ and $\theta_1 = 1.2$, middle panels refer to scenario 2 with $k_1 = 40$, $\tilde{k}_1 = 20$ and $\theta_1 = 1.2$, and the bottom panels refer to scenario 3 with $k_1 = 1$, $\tilde{k}_1 = 2$ and $\theta_1 = 20$.

Table 2. reBias, standard deviation (SD), and coverage probabilities of the 95% confidence intervals (Cov) for the location and scale parameters of the age at onset gamma distributions of U (k_1, θ_1) and T ($k_1 + \tilde{k}_1, \theta_2$) along 1000 trials for several sample sizes ($m \in \{100, 500, 1000\}$ families) and effective sample sizes, defined as the mean number of ascertained families along the 1000 trials ($\bar{r}$).

		Small families				Large families			
Parameter	m	$\bar{r}$	reBias	SD	Cov	$\bar{r}$	reBias	SD	Cov
Scenario 1									
	100	26	0.736	–	–	58	0.054	10.594	0.921
	500	132	0.076	16.088	0.802	289	0.014	4.299	0.917
$k_1 = 50$	1000	264	0.033	9.791	0.792	579	0.002	3.076	0.896
	100	26	0.656	–	–	58	−0.013	0.244	0.900
	500	132	0.005	0.337	0.937	289	−0.006	0.105	0.926
$\theta_1 = 1.2$	1000	264	0.002	0.225	0.956	579	0.002	0.076	0.919
	100	26	0.093	–	–	58	0.025	8.623	0.962
	500	132	0.020	8.000	0.969	289	0.008	3.771	0.958
$k_1 + \tilde{k}_1 = 55$	1000	264	0.009	5.656	0.951	579	0.003	2.485	0.960
	100	26	0.076	–	–	58	0.001	0.201	0.940
	500	132	0.003	0.189	0.932	289	−0.002	0.088	0.948
$\theta_2 = 1.2$	1000	264	0.003	0.136	0.942	579	−0.001	0.059	0.960
Scenario 2									
	100	13	1.384	–	–	34	0.070	10.464	0.969
	500	65	0.186	20.123	0.941	167	0.013	4.288	0.925
$k_1 = 40$	1000	130	0.067	10.792	0.922	333	0.008	2.836	0.946
	100	13	−0.050	–	–	34	−0.013	0.268	0.939
	500	65	−0.036	0.402	0.904	167	−0.002	0.125	0.939
$\theta_1 = 1.2$	1000	130	<0.001	0.279	0.945	333	−0.003	0.084	0.961
	100	13	0.389	–	–	34	0.041	15.561	0.975
	500	65	0.046	15.168	0.971	167	0.007	6.391	0.958
$k_1 + \tilde{k}_1 = 60$	1000	130	0.014	10.036	0.961	333	0.005	4.483	0.957
	100	13	0.324	–	–	34	0.031	0.341	0.937
	500	65	0.036	0.400	0.928	167	0.006	0.144	0.944
$\theta_2 = 1.2$	1000	130	0.023	0.246	0.946	333	0.001	0.101	0.953
Scenario 3									
	100	66	0.309	–	–	92	0.027	0.332	0.930
	500	329	0.046	0.320	0.891	462	0.009	0.139	0.935
$k_1 = 1$	1000	658	0.015	0.219	0.897	924	0.003	0.097	0.937
	100	66	0.552	–	–	92	0.061	6.241	0.946
	500	329	0.038	5.822	0.958	462	0.009	2.378	0.954
$\theta_1 = 20$	1000	658	0.024	3.888	0.966	924	0.005	1.648	0.951
	100	66	0.021	–	–	92	0.004	0.224	0.953
	500	329	0.006	0.186	0.942	462	0.001	0.101	0.944
$k_1 + \tilde{k}_1 = 3$	1000	658	0.002	0.017	0.943	924	<0.001	0.068	0.954
	100	66	0.015	–	–	92	0.005	4.051	0.954
	500	329	<0.001	1.786	0.942	462	0.002	0.904	0.949
$\theta_2 = 20$	1000	658	0.002	1.310	0.937	924	<0.001	0.624	0.954

For small families and $m = 100$, the SDs are not reliable and left out from the table. reBias: relative bias.

From these graphics, the proposed estimators seem to be unbiased. The median estimated curves based on the median parameter estimate along the $M = 1000$ Monte Carlo trials cannot be distinguished from the theoretical distributions, and the bands formed by the 2.5 and 97.5% Monte Carlo percentiles nicely cover the theoretical curves in all the studied scenarios. In each Monte Carlo trial, data of $m = 1000$ families are simulated. Since not all of these m families satisfied the ascertainment criteria, the effective number of families included for analysis in each Monte Carlo trial, denoted as r, is considerably lower (see Table 2 for details). This might be the reason for the wide percentile band in the right upper graphic in Figure 2.

Table 2 complements Figure 2 and provides further results of the simulation study. For each of the studied scenarios, we provide results on mean estimated relative bias (reBias) (defined as the difference between the

simulated mean and true parameter value divided by the true value), empirical standard deviation, and coverage probabilities across the 1000 Monte Carlo trials of the scale and shape parameters of the distribution of $H_{(k_1,\theta_1)}$ and $F_{(k_1+\tilde{k}_1,\theta_1)}$. The results regarding bias reinforce some of the findings observed in Figure 2. Within the same scenario, the populations composed of large families provide better results than the populations composed of small families, as is expected. Results also improve by increasing the sample size, showing lower reBias and less variability (SD). With regard to the coverage probabilities, we observe that the coverage probabilities are close to 0.95 in all the studied scenarios. This indicates the good performance of our bootstrap approach to estimate the standard errors.

Special mention deserves scenario 3, large families and $m = 25$ (corresponding to a mean number of ascertained families along the 1000 trials of $\bar{r} = 17$), since it resembles our real data setting and hence gives valuable information about the expected performance of our method in our motivating data-set. In this case, we observe reasonable values of reBias (0.202 for k_1, 0.667 for θ_1, 0.025 for $k_1 + \tilde{k}_1$, and 0.005 for θ_2) and coverage probabilities close to 0.95. In terms of standard deviation (0.753 for k_1, 237 for θ_1, 0.467 for $k_1 + \tilde{k}_1$, and 4.153 for θ_2), the results are also reasonable with exception of k_1, with a very large value of standard deviation due to extreme results in some of the 1000 trials. This should be interpreted as symptoms of instability of our method with very small samples even if the overall performance is reasonably good in this setting.

In the second simulation study, we evaluate the level of introduced bias in our estimates due to informative missing. For scenario 1 ($k_1 = 50, \tilde{k}_1 = 5, \theta_1 = 1.2$), and scenario 2 ($k_1 = 40, \tilde{k}_1 = 20, \theta_1 = 1.2$), and for missing probabilities 0.20 and 0.50 of non-symptomatic family members, this bias is visualized in Figure 4 in Appendix 3 (scenario 1 in the graphics on the left and scenario 2 on the right). In scenario 1, the distributions of the asymptomatic and symptomatic stages are both overestimated if non-symptomatic family members do not participate in the study. The bias is smaller for $H_{(k_1,\theta_1)}$ than for $F_{(k_1+\tilde{k}_1,\theta_1)}$, which is interesting since the estimation $H_{(k_1,\theta_1)}$ is our main objective. In scenario 2 this phenomenon is even more pronounced for $F_{(k_1+\tilde{k}_1,\theta_1)}$ and the bias in the estimation of $H_{(k_1,\theta_1)}$ is negligible. Also, in general, the reBias is larger if the proportion of missing family members increases. Assuming a missing probability of 0.20 in scenario 1, the reBias for $F_{(k_1+\tilde{k}_1,\theta_1)}$ is around 13% at 50 and 60 years old, and reduces to 7% at age 70. The reBias for $H_{(k_1,\theta_1)}$ is lower at all ages, around 4% at ages 50 and 60 years, and reduces to 2% at age 70. Assuming a missing probability of 0.50, the reBias for $F_{(k_1+\tilde{k}_1,\theta_1)}$ increases to 40% at the ages 50 and 60 years, and to 20% at age 70. The reBias for $H_{(k_1,\theta_1)}$ also increases, reaching 10% at ages 50 and 60 years and to 5% at age 70. Similar results are found in scenario 2 in the estimation of $F_{(k_1+\tilde{k}_1,\theta_1)}$ while the bias in the estimation of $H_{(k_1,\theta_1)}$ remains negligible when increasing the proportion of missing individuals. The small bias in the estimation of $H_{(k_1,\theta_1)}$ is likely due to the fact that the missing data mechanism relies on the indicator $\Delta = I(T > C)$ and hence it is only informative about U given its association with T. As a result, less bias for estimating $H_{(k_1,\theta_1)}$ than for $F_{(k_1+\tilde{k}_1,\theta_1)}$ is expected. The association between T and U is larger in scenario 1 than in scenario 2 which explains the lower observed bias in scenario 2. In summary, even if missing members is a potential problem and it introduces systematic bias, we expect its impact will be limited.

5. Motivating example

This work was motivated by a study on FSHD, a genetic muscle disorder. The severity of this disease is associated with a specific form of genetic lesion, the loss of repetitions of the D4Z4 unit.[10] Individuals without a loss of units (they have at least 10 units) are considered to be healthy and are not susceptible to develop the muscle disorder. It is expected that the age at onset of the asymptomatic and the symptomatic stage is also associated with the number of repetitions of this unit.

The data come from a cross-sectional study in which at a fixed and non-informative calendar time, all affected pedigrees in the Netherlands with at least one affected member among the first and second degree of the index patient (the first diagnosed patient in a family) were invited to participate in the study (see Wohlgemuth et al.[11] for details). So, the ascertained families have at least two affected individuals with both a loss of repetitions of the D4Z4 unit: the index patient and a relative. Data of 10 pedigrees consisting of in total 155 individuals are available. Of these 155 individuals, 69 present loss of repetitions of the D4Z4 unit at some degree (the so-called carriers) and 86 have no genetic alteration (the non-carriers). All individuals within a pedigree with a loss of repetitions have an equal number of repetitions. An overview of the carrier-data is given in Table 3.

We consider two parametric models: the Weibull distribution for both $H_{\eta,x}$ and $F_{\theta,x}$ and the gamma distribution for both (see Appendix 2 for details) and covariate x equals an indicator function that indicates whether an individual has less than 7 or at least 7 (i.e. 7, 8, or 9) repetitions of the D4Z4 unit. (We also included the number of

Table 3. Overview of data: For every pedigree, the number of units, carriers, and symptomatic and asymptomatic carriers are given.

Pedigree no.	1	2	3	4	5	6	7	8	9	10
No. of D4Z4 units among carriers	4	5	5	6	6	6	7	7	9	9
No. of carriers in the pedigree	5	9	8	5	7	13	5	3	8	6
No. of symptomatic carriers	5	9	7	5	2	3	3	2	2	2
No. of asymptomatic carriers	0	0	1	0	5	6	1	1	2	1

A carrier is defined as an individual with a loss of repetitions of the D4Z4 unit.

repetitions as a continuous variable, but since the linearity assumption does not seem to hold, it is not considered further.)

For n_j the number of individuals in pedigree j in the data-set, the probability the ascertainment event occurs (at least two symptomatic patients at examination), equals 1 minus the probability that none or only one of the n_j individuals is symptomatic

$$\mathrm{P}(A_j) = 1 - \left(\int 1 - F_{\theta,x}\mathrm{d}G\right)^{n_j} - n_j\left(\int 1 - F_{\theta,x}\mathrm{d}G\right)^{n_j-1}\int F_{\theta,x}\mathrm{d}G \tag{4}$$

where the term $\left(\int 1 - F_{\theta,x}\mathrm{d}G\right)^{n_j}$ equals the probability that none of the family members has symptoms at the time of examination and $n_j\left(\int 1 - F_{\theta,x}\mathrm{d}G\right)^{n_j-1}\int F_{\theta,x}\mathrm{d}G$ equals the probability that exactly one individual has symptoms at the time of examination. This probability is inserted in the denominators of the likelihood functions in equations (1) and (2).

We estimate G by the empirical distribution function of the ages at time of examination C of the individuals with and without a loss of repetitions together. Since the number of observations is low, we chose to combine the data when estimating G. Next, we follow the estimation procedure as described in Section 3.

Based on the value of the likelihood (or AIC), the gamma model fits slightly better than the Weibull model. The actual estimates of $H_{\eta,x}$ and $F_{\theta,x}$ for both models are very similar. Only the estimates of $F_{\theta,x}$ for the group with 7–9 units and after the age of 45 seem to diverge; the estimate based on the gamma distribution increases to almost 0.6 at the age of 60, whereas the estimate based on the Weibull distribution reaches 0.4 at this age. This is probably due to the fact that the data-set is relatively small and most individuals in the data-set have not reached this age at time of examination. The form of the estimated parametric distributions (i.e. parameters) is therefore mainly determined by the events before the age of 45 and the curve is extrapolated after the age of 45. As a consequence one should be careful with drawing conclusions at higher ages. The estimates based on the gamma models are given in Figure 3. The estimation procedure was repeated for G estimated by the empirical distribution of data of individuals with no genetic alteration; the results are similar. The pointwise 95% confidence intervals constructed with the parametric bootstrap method are wide, especially for $H_{\eta,x}$ at younger age. This is possibly because no data of individuals below the age of 20 are available.

The estimates of the age at onset distribution of the asymptomatic and the symptomatic stage of FSHD show that these functions depend on the covariate repeat size and both increase until late adulthood. These estimates can be used in counseling and help in understanding progression of the disease over time. However, the number of individuals on which the estimates are based is low and should be interpreted with care.

R-code for maximizing the log-likelihood functions is provided in Appendix 4.

6. Discussion

In this paper, we have proposed a maximum likelihood-based method for estimating the age at onset distribution for the asymptomatic stage of a genetic disease using clinically ascertained pedigree data. Simulation studies showed that as long as the sample is not too small, our estimation method yields accurate results. Estimates of this distribution are of great importance for setting up follow-up programs in high-risk families, for instance families with a genetic variant associated with susceptibility of a disease.

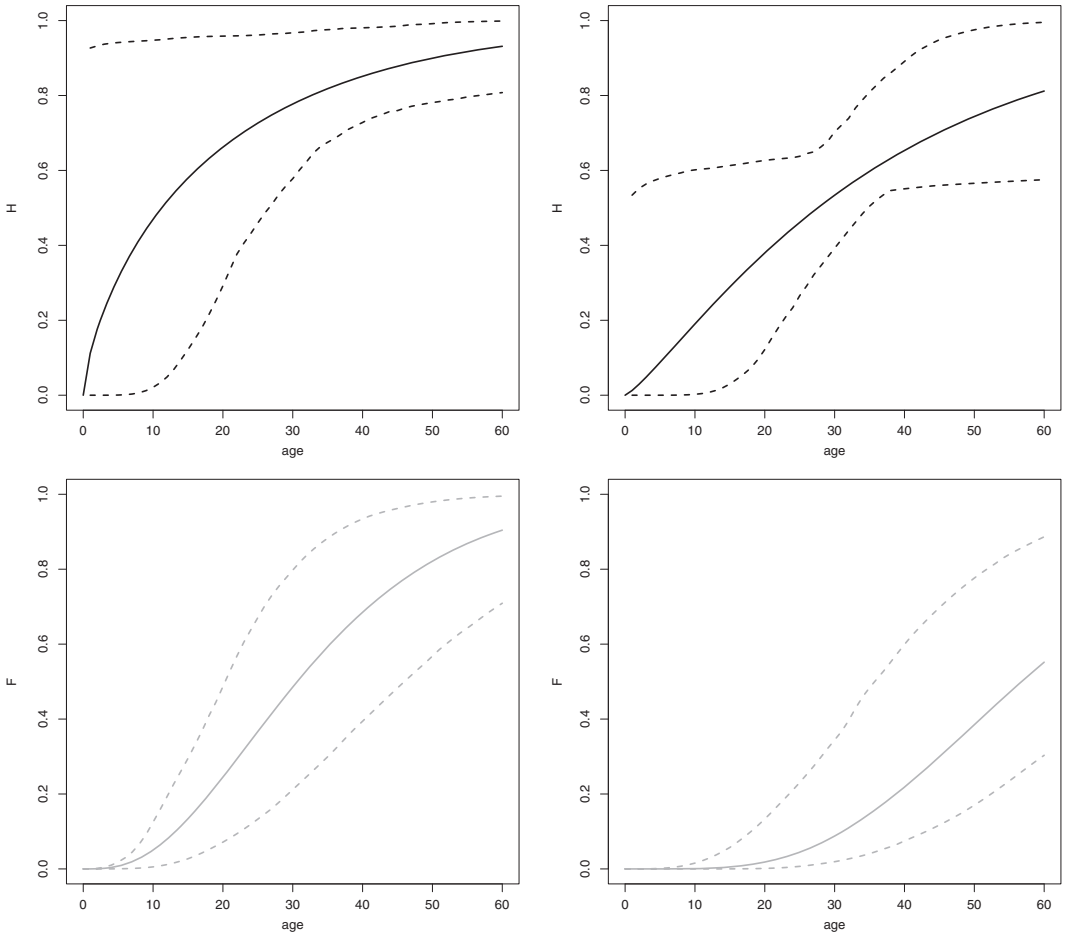

Figure 3. Solid lines: Estimates of $H_{\eta,x}$ (top panels) and $F_{\theta,x}$ (bottom panels) for 4, 5, or 6 repetitions of the D4Z4 unit (left) and for 7, 8, or 9 repetitions (right). Dashed lines: corresponding pointwise 95% confidence intervals.

The estimates of the age at onset distribution of the asymptomatic and symptomatic stage of FSHD, found in the application, can be used to learn more about the progression of the disease. Since there is no treatment available and the disease is not life threatening, one could argue whether screening is necessary, but eventually this decision will be made by the family and their medical doctor.

In this paper, we considered the situation of a cross-sectional study which took place at a randomly chosen moment (set by the researcher), without any follow-up afterwards. However, regular screening of patients with a high disposition of a slowly developing disease is quite common.[1,2] To fit these kind of screening data, the expressions of the likelihood functions need to be adjusted, but the underlying principle of estimating the age at onset distributions for the asymptomatic and the symptomatic stage in two steps remains valid.

Measured family characteristics can be included in the model via covariates. To account for unmeasured family characteristics, a frailty term (random family effect) could be added to the model (see, e.g. Gong et al.,[12] Hsu et al.,[13] Hsu and Gorfine,[14] and Gorfine et al.[15]). In a shared frailty model, every family has its own frailty that describes the susceptibility of the family members to develop the disease compared to the population of interest. This model could be further generalized to correlated frailty models in which every individual in the family has its own frailty term, but these terms are correlated within families (and are independent between families). This gives the model more flexibility. However, for fitting these models sufficient data must be available; the number of pedigrees and the number of individuals in the pedigrees must be sufficiently large. This is certainly not the case in our application, but including frailties in the model is an interesting topic for further research.

Since we are considering the distinct stages (healthy, asymptomatic, symptomatic), our data could be modeled with a multi-state model.[16] However, as far as we know, multi-state modeling with data that are ascertained based on the outcome is still an open problem.

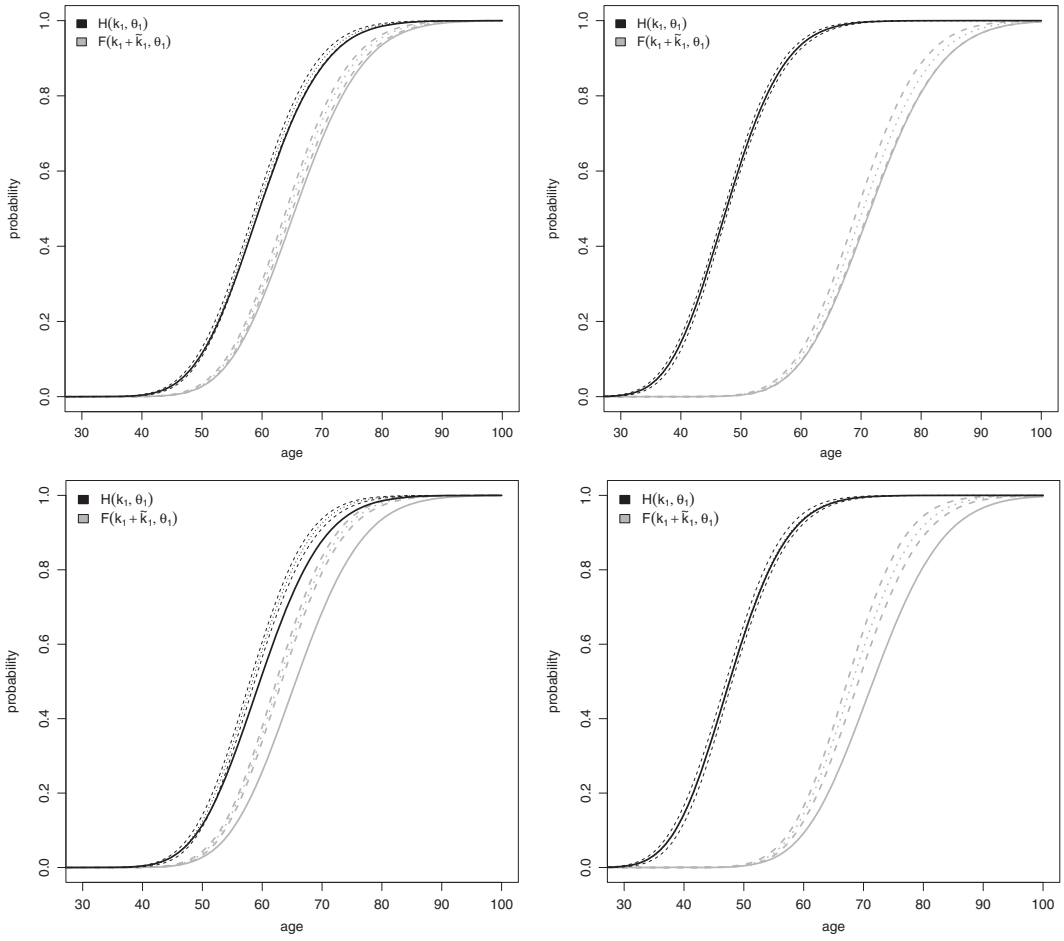

Figure 4. Results of the simulation study based on populations of families with $n_j \in \{5, 10, 15, 20\}$. Left: scenario 1 ($k_1 = 50$, $\tilde{k}_1 = 5, \theta_1 = 1.2$). Right: scenario 2 ($k_1 = 40$, $\tilde{k}_1 = 20, \theta_1 = 1.2$). Top: exclusion probability of 0.20. Bottom: exclusion probability of 0.50. Solid black and gray lines are the true distribution functions $H_{(k_1,\theta_1)}$ and $F_{(k_1+\tilde{k}_1,\theta_1)}$, respectively. Dashed lines form a band based on estimated pointwise 2.5 and 97.5% percentiles based on $M = 1000$ Monte Carlo trials.

In this paper, we have assumed that the medical tests at screening moments are fully sensitive and specific. In population screening programs like breast and colon cancer, the screening tests are often imperfect; the sensitivity and specificity of the test are below 100%.[6] To preclude unnecessary treatment, a screening test is followed by a confirmative test in case of a positive screening test. The confirmative test is assumed to be 100% sensitive and specific. In family studies, only high-risk individuals (carriers) are screened. For this purpose usually the most accurate medical test is applied and an assumption of full sensitivity and specificity is reasonable. Otherwise, the expression of the likelihood must be adapted, but the methodology will be the same.

To conclude, reliable estimates of the age at onset distribution of the asymptomatic stage are important for setting up personal follow-up screening in high-risk families. The methodology described in this paper is therefore of great relevance.

Acknowledgements

We like to thank M. Wohlgemuth and N. Voermans for collecting and making available the FSHD-data that were analyzed in Section 5. Further, we like to thank the reviewers for their helpful comments.

Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

References

1. Halk AB, Potjer TP, Kukutsch NA, et al. Surveillance for familial melanoma: recommendations from a national centre of expertise. *Br J Dermatol* 2019; **181**(3): 594–596.
2. Kennelly RP, Gryfe R and Winter DC. Familial colorectal cancer: patient assessment, surveillance and surgical management. *Eur J Surg Oncol* 2017; **43**: 294–302.
3. Zhang Z and Sun J. Interval censoring. *Stat Method Med Res* 2010; **19**: 53–70.
4. Groeneboom P and Wellner JA. *Information bounds and nonparametric maximum likelihood estimation*. Berlin: Birkhauser Verlag, 1992.
5. Jewell NP and van der Laan M. Current status data: review, recent developments and open problems. In: Balakrishnan N and Rao CR (eds) *Handbook of statistics*. Elsevier, Vol. 23. 2003, pp.625–642.
6. Witte B, Berkhof J and Jonker MA. An EM algorithm for nonparametric estimation of the cumulative incidence function from repeated imperfect test results. *Stat Med* 2017; **36**(21): 3412–3421.
7. Jonker MA, Rijken J, Hes FJ, et al. Estimating the penetrance of pathogenic gene variants in families with missing pedigree information. *Stat Method Med Res* 2019; **28**(10–11): 2924–2936.
8. Carayol J and Bonaiti-Pellie C. Estimating penetrance from family data using a retrospective likelihood when ascertainment depends on genotype and age of onset. *Genet Epidemiol* 2004; **27**: 109–117.
9. Kraft P and Thomas DC. Bias and efficiency in family-based gene-characterization studies: conditional, prospective, retrospective, and joint likelihoods. *AJHG* 2000; **66**: 1119–1131.
10. Mull K, Lemmers RJLF, Horlings CGC, et al. Phenotype-genotype relations in facioscapulohumeral muscular dystrophy type 1. *Clin Genet* 2018; **94**: 521–527.
11. Wohlgemuth M, Lemmers RJ, Jonker MA, et al. A family-based study into penetrance in facioscapulohumeral muscular dystrophy type 1. *Neurology* 2018; **91**(5): e444–e454.
12. Gong G, Hannon N and Whittemore AS. Estimating gene penetrance from family data. *Genet Epidemiol* 2010; **34**: 373–381.
13. Hsu L, Chen L, Gorfine M, et al. Semiparametric estimation of marginal hazard function from case-control family studies. *Biometrics* 2004; **60**: 936–944.
14. Hsu L and Gorfine M. Multivariate survival analysis for case-control family data. *Biostatistics* 2006; **7**: 387–398.
15. Gorfine M, Hsu L and Parmigiani G. Frailty models for familial risk with application to breast cancer. *J Am Stat Assoc* 2013; **108**: 1205–1215.
16. Putter H, Fiocco M and Geskus RB. Tutorial in biostatistics: competing risks and multi-state models. *Stat Med* 2006; **26**: 2389–2430.

Appendix 1

A.1. Derivation of the likelihood functions

In this appendix we derive the likelihood functions L^{symp} and L^{full} given in equations (1) and (2). An observation for individual i in pedigree j is denoted with a subscript ij, with $j = 1, \ldots, r$ and $i = 1, \ldots, n_j$ with n_j the number of individuals in pedigree j in the data-set. That means that for individual i from pedigree j, the variables are denoted as $(U_{ij}, T_{ij}, \Delta_{ij}, \Sigma_{ij}, C_{ij})$.

A.1.1. Derivation of likelihood L^{symp}

Suppose no information on the asymptomatic stage of the disease at the time of examination is present, Σ_{ij} is not observed. Then the observations reduce to $(T_{ij}\Delta_{ij}, C_{ij})$. Every individual in the data-set is either non-symptomatic $(T_{ij} > C_{ij})$ or symptomatic $(T_{ij} \leq C_{ij})$ at the time of examination. The corresponding unconditional likelihood function for pedigree j equals

$$\prod_{i=1}^{n_j} g(C_{ij}) f_{\theta,x_{ij}}(T_{ij})^{\Delta_{ij}} (1 - F_{\theta,x_{ij}}(C_{ij}))^{1-\Delta_{ij}}$$

the likelihood in case of right-censoring in which the censoring time C_{ij} is always observed.

For A_j be the ascertainment event for pedigree j and $\mathrm{P}(A_j)$ its corresponding probability, the conditional likelihood for pedigree j, given it was ascertained (the event A_j), is given by

$$L_j^{symp} = \frac{\prod_{i=1}^{n_j} f_{\theta,x_{ij}}(T_{ij})^{\Delta_{ij}}(1 - F_{\theta,x_{ij}}(C_{ij}))^{1-\Delta_{ij}} g(C_{ij})}{\mathrm{P}(A_j)}$$

The exact form of $\mathrm{P}(A_j)$ in terms of the distribution functions depends on the ascertainment criteria. Under the assumption of independence between pedigrees, the conditional likelihood equals

$$L^{symp} = \prod_{j=1}^{r} L_j^{symp} = \frac{\prod_{j=1}^{r}\prod_{i=1}^{n_j} f_{\theta,x_j}(T_{ij})^{\Delta_{ij}}(1 - F_{\theta,x_j}(C_{ij}))^{1-\Delta_{ij}} g(C_{ij})}{\prod_{j=1}^{r}\mathrm{P}(A_j)}$$

Under the assumption that the observations of all relatives are independent, conditionally the carrier status and the covariates in the model, no distinction between the pedigrees has to be made. As a consequence, for all individuals in the data-set the double index (ij) can be replaced by a single unique index i. This simplifies the likelihood function to the likelihood in equation (1).

A.1.2. Derivation of the full likelihood L^{full}

For individual i in pedigree j, the observation is given by the triple $(T_{ij}\Delta_{ij}, \Sigma_{ij}, C_{ij})$. For every individual in the data-set, there are three possible situations (recall Figure 1):

1. Healthy: $C_{ij} < U_{ij} < T_{ij}$. The observation is $(C_{ij}, \Sigma_{ij} = \Delta_{ij} = 0)$ and the corresponding likelihood function equals $g(C_{ij})(1 - H_{\eta,x_{ij}}(C_{ij}))$
2. Asymptomatic: $U_{ij} \leq C_{ij} < T_{ij}$. The observation is $(C_{ij}, \Sigma_{ij} = 1, \Delta_{ij} = 0)$ and the corresponding likelihood equals $g(C_{ij})(H_{\eta,x_{ij}}(C_{ij}) - F_{\theta,x_{ij}}(C_{ij}))$
3. Symptomatic: $U_{ij} \leq T_{ij} \leq C_{ij}$. The observation is $(C_{ij}, T_{ij}, \Sigma_{ij} = \Delta_{ij} = 1)$, with corresponding likelihood $g(C_{ij})f_{\theta,x_{ij}}(T_{ij})$

Combining the three gives the conditional likelihood function for pedigree j

$$L_j^{full} = \frac{\prod_{i=1}^{n_j} g(C_{ij})(1 - H_{\eta,x_{ij}}(C_{ij}))^{(1-\Sigma_{ij})(1-\Delta_{ij})}(H_{\eta,x_{ij}}(C_{ij}) - F_{\theta,x_{ij}}(C_{ij}))^{\Sigma_{ij}(1-\Delta_{ij})} f_{\theta,x_{ij}}(T_{ij})^{\Sigma_{ij}\Delta_{ij}}}{\mathrm{P}(A_j)}$$

Under the assumption of stochastic independence between the observations of the individuals in different pedigrees, the conditional likelihood function equals $L^{full} = \prod_{j=1}^{r} L_j^{full}$. Furthermore, conditionally the covariates in the double index can be replaced by a single unique index for every individual in the data-set. This yields the likelihood function (2).

Appendix 2

A.2. Parametric models

For analyzing the FSHD data (see Section 5), the Weibull and the gamma distributions are considered. A link function between the parameters of these distributions and a covariate x is proposed here.

For the Weibull distribution, the shape parameter is denoted as k and the scale parameter is chosen to be of the form $\nu_x = \mu\exp(-(\beta/k)x)$. The corresponding hazard function is given by

$$\lambda_{\theta,x}(t) = \frac{f_{\theta,x}(t)}{1 - F_{\theta,x}(t)} = \frac{kt^{k-1}}{\nu^k} = \frac{kt^{k-1}}{\mu^k\exp(-\beta x)} = \frac{kt^{k-1}}{\mu^k}\exp(\beta x) = \lambda_0(t)\exp(\beta x)$$

with parameter $\theta = (k, \mu, \beta)$ and $\lambda_0(t) = kt^{k-1}/\mu^k$, a proportional hazards regression model with a specific parametric baseline hazard. The density and the distribution functions for the age at onset distribution for the asymptomatic stage of the disease are defined similarly.

The gamma distribution with shape (k) and inverse scale (ν) parameters has expectation and variance equal to k/ν and k/ν^2. We assume that $k/\nu = \exp(\beta_0 + \beta_1 x)$, with x the covariate.

Appendix 3

A.3. Additional simulation results

In this appendix additional simulation results are presented in case a part of the healthy individuals is missing. The estimated curves are plotted in Figure 4. The simulation setting is described in the caption of the figure.

Appendix 4

A.4. R-code for maximizing the likelihood

In this appendix we present the R-code that was used for numerical maximization of the log-likelihood functions L^{symp} and L^{full} in the application section.

```
# MloglikFgamma calculates −L^{symp}, given starting values and data.
MloglikFgamma <- function(par,obs) {
    scale <- par[1]
    shape <- 1/scale * exp(par[2]+par[3]*obs$unit)
    x <- sum(obs$Delta*log(dgamma(obs$T,shape=shape,scale=scale))+
      (1−obs$Delta)*log(1−pgamma(obs$C,shape=shape,scale=scale)))
    z <- 0
    for (i in 1:length(obs$pedsize))
    { shapeped <- 1/scale * exp(par[2]+par[3]*obs$pedunit[i])
      intt <- sum(1−pgamma(obs$Crel,shape=shapeped,scale=scale))/length(obs$Crel)
      intt2 <- obs$pedsize[i] *
            (sum(1−pgamma(obs$Crel,shape=shapeped,scale=scale))/length(obs$Crel))
                  ^(obs$pedsize[i]−1)
            * sum(pgamma(obs$Crel,shape=shapeped,scale=scale))/length(obs$Crel)
    z <- z + log(1−intt^obs$pedsize[i]−intt2)
  }
  −(x − z)
}

# Mloglikgamma calculates -L^{full} given starting values and observed data
Mloglikgamma <- function(par,obs){
  scaleu <- par[1]
  shapeu <- 1/scaleu * exp(par[2]+par[3]*obs$unit)
  Fscale0 <- obs$Fscale0
  Fshape0 <- 1/Fscale0 * exp(obs$Fbeta0+obs$Fbeta1*obs$unit)

  # Delta=0, Sigma=0
  t1 <- log(1−pgamma(obs$C[obs$Sigma==0],shape=shapeu,scale=scaleu))
  L1 <- sum(t1)
  # Delta=0, Sigma=1
  t2 <- pgamma(obs$C[obs$Sigma*(1−obs$Delta)==1],shape=shapeu,scale=scaleu)
      −pgamma(obs$C[obs$Sigma*(1−obs$Delta)==1],shape=Fshape0,scale=Fscale0)
  t2 <- as.numeric(t2>0)*log(t2) + as.numeric(t2<=0)*−1000
  L2 <- sum(t2)
  −(L1+L2)
}
```

```
# EXECUTION #
obsdata <- list(T=T,C=C,Crel=CCall,Sigma=Sigma,Delta=Delta,unit=unitdi,
  pedsize=pedsize,pedunit=pedunit)

p <- c(5.5, −4.5, 1.5) #Starting values: scale, beta0, beta1.
res1 <- nlminb(start=p, objective=MloglikFgamma, obs=obsdata,control = list(trace=TRUE))

obsdata$Fscale0 <- res1$par[1]
obsdata$Fbeta0 <- res1$par[2]
obsdata$Fbeta1 <- res1$par[3]
p <- c(22, −6, 1.3) #Starting values.
res2 <- nlminb(start=p, objective=Mloglikgamma, obs=obsdata,control = list(trace=TRUE))
```

Remark: the vectors “pedunit” and “pedsize” are defined as the number of units and the pedigree sizes. Their lengths equal the number of pedigrees in the data-set. The two vectors should be ordered in the same way.