Article

SMMR
STATISTICAL METHODS IN MEDICAL RESEARCH

Statistical Methods in Medical Research
2019, Vol. 28(10–11) 2924–2936

Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/0962280218791338
journals.sagepub.com/home/smm

SAGE

Estimating the penetrance of pathogenic gene variants in families with missing pedigree information

Marianne A Jonker,[1,2] Johannes A Rijken,[3] Frederik J Hes,[4] Hein Putter[5] and Erik F Hensen[3,6]

Abstract

Accurate assessment of the age-dependent disease risk conferred by germline variants in disease susceptibility genes is often hampered by the way the data are collected. Cohort-based data sets frequently contain an overrepresentation of patients (i.e. carriers of the gene variant of interest affected with the associated disease), and an underrepresentation of disease-free carriers. In order to overcome this problem, penetrance estimates can be based on family-based study designs, through the evaluation of index patients and their family members. This approach facilitates the identification of asymptomatic germline variant carriers. By adjusting for the way these family data are ascertained, an estimate for the penetrance of the pathogenic gene variant can be obtained. However, the family structure is often incomplete or missing. This complicates the estimation of the penetrance, because full adjustment of the likelihood is not possible. We present a conditional likelihood for the estimation of the penetrance of pathogenic gene variants, based on a cohort of multiple families comprising index patients, disease-free and affected non-index carriers, but with missing information on pedigree structure. The proposed estimator corrects for the ascertainment in a robust way and is shown to be more accurate than the frequently used Kaplan–Meier estimator of the penetrance function.

Keywords

Age-at-onset, conditional maximum likelihood method, missing data, SDHB

1 Introduction

Accurate estimates of the age-dependent disease risk (the penetrance) for carriers of germline gene variants that affect function (hereafter variants) in specific disease-susceptibility genes are important in counseling carriers and their families and in optimising cascade screening and follow-up protocols (surveillance). Especially in the case of rare genetic variants and diseases, obtaining data through a population-based cohort study design is usually not a realistic option, since a large number of individuals is needed in order to identify enough carriers of the gene variants of interest for an accurate estimation of its penetrance. Therefore, carriers of specific gene variants are often identified through a family-based study design, involving genetic cascade screening of the family of index patients (i.e. the first identified individual within a family carrying a pathogenic gene variant and expressing the associated disease phenotype). Different family-based study designs have been suggested for penetrance estimations.[1] These designs allow for unbiased

[1]Department of Epidemiology and Biostatistics, VU University Medical Center, Amsterdam, the Netherlands
[2]Department for Health Evidence, Biostatistics Section, Radboud University Medical Center, Nijmegen, the Netherlands
[3]Department of Otorhinolaryngology – Head and Neck Surgery, VU University Medical Center, Amsterdam, the Netherlands
[4]Department of Clinical Genetics, Leiden University Medical Center, Leiden, the Netherlands
[5]Department of Medical Statistics and Bioinformatics, Leiden University Medical Center, Leiden, the Netherlands
[6]Department of Otorhinolaryngology – Head and Neck Surgery, Leiden University Medical Center, Leiden, the Netherlands

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

estimates of the penetrance via the maximum likelihood method, provided that adequate adjustment of the likelihood is applied to the manner in which data are ascertained.[2,3] For this purpose, prospective and retrospective likelihoods are often used. The prospective likelihood can only be applied if the ascertainment rules are clear, have been followed exactly, and are easy to model. Otherwise the retrospective likelihood should be used. This likelihood is applicable if the ascertainment is based on the phenotypes only.[4,5] A frequently employed method for the estimation of the penetrance uses the Kaplan–Meier estimator based on data of the relatives of index patients only (the index patients themselves are left out of the analysis to correct for the ascertainment bias to prevent overestimation of the penetrance). However, this method still yields biased estimates as it does not actually correct for the way the data are ascertained. The estimation procedure ignores the fact that ascertainment is based on the phenotype of the whole pedigree and not on one particular individual in the pedigree. Moreover, leaving out the index patients from the analysis means discarding valuable information, especially in rare and low-penetrant disease.

Here we consider the situation where essential pedigree information is missing. This is frequently the case, due to a plethora of possible reasons: the patient may be unaware of the family history, the hereditary nature of the disease may not have been recognised during treatment of a single patient, or the data were initially not collected for research or cascade screening purposes. Usually, pedigree data are obtained via an index patient in a pedigree. Patients who express the disease phenotype and carry the genetic variant of interest are asked to inform their family members about their potential risk. Some of these relatives will consent to genetic counseling and DNA testing. Detected carriers of the germline variant will be followed and regularly screened for the disease. When aiming to estimate the penetrance of the disease, the follow-up data of all known carriers (the index patients and their relatives with the variant) are collected from the medical records. However, the relation between the carriers and the index patient is often lacking. This missing information hampers the correction for the way the data were ascertained. A frequently used method for penetrance calculations in this setting is the Kaplan–Meier estimator, because of its simplicity and because up to now, there was no adequate alternative. However, the Kaplan–Meier estimator is prone to ascertainment bias and discards useful information. Here we describe a novel method for the estimation of the penetrance function that is designed especially for the situation described above. The performance of this newly proposed estimator is evaluated by means of simulation studies and compared with the traditionally used Kaplan–Meier estimator.

Information on the family members at risk that do not consent to clinical surveillance or DNA testing is usually entirely missing. These data may not be missing at random, as relatives of index patients who experience similar symptoms are presumably more inclined to undergo surveillance or DNA testing than family members who do not. This problem is especially poignant in low-penetrant disease, as a relatively large number of family members who carry the gene variant of interest will be asymptomatic. As it is impossible to account for this missing data in the estimation method if there is no indication of the number of family members that are missing from the dataset, the effect of the missing data is quantified with a simulation study.

With the growing availability and accessibility of high-throughput genetic screening techniques such as whole exome sequencing and whole genome sequencing, our knowledge of the genetic determinants of hereditary disease is growing rapidly. The advent of these techniques has facilitated the identification of causative genes in rare diseases, and of gene variants that confer a relatively low disease risk. Accurate estimation of disease risk is challenging especially in these rare and low-penetrant gene variants. The need for accurate methods for estimating disease risk that address ascertainment correction will increase with the rising number of individuals known to harbour a hereditary predisposition to disease.

2 Methods

2.1 Data, model assumptions and likelihood

Here it is assumed that the disease is not lethal. At the end of Appendix 1, it is explained in what way the likelihood will change if this assumption does not hold and individuals may die from the disease.

Once an individual is diagnosed with the disease of interest and with a function-affecting gene variant that predisposes for the disease of interest, relatives are often offered genetic testing and clinical surveillance of the disease in case of identified carriership. When aiming for estimating the disease penetrance the follow-up data of the carriers is collected from the medical records.

For gene variant carriers, the following notation is used: the age at diagnosis of the disease is denoted by T with cumulative distribution F and density f, and the age at the end of the study (collection of the data), or the age at death (whichever occurs first) by C, with distribution G and density g. It is assumed that T and C are independent.

For every subject it is known whether he or she is an index patient or a relative of an index patient, but information on the family (structure) is not available. For patients ($T \leq C$), T is observed. The age at censoring, C, is also observed. For the unaffected carriers ($T > C$), C is always observed, but T is unobserved. The indicator function $\Delta = 1_{\{T \leq C\}}$ equals 1 if $T \leq C$ and 0 if $T > C$. Note that for index patients $\Delta = 1$. In order to distinguish between the different individuals, an underscore is used; for individual $i, i = 1, \ldots, n$: T_i, C_i, Δ_i. It is assumed that the phenotypes (disease status and age-at-onset) of individuals are independent given their carrier status. In total, data of n mutation-carriers from r pedigrees (r index patients), with $n > r$, is available. The number of individuals in family j in the data-set (i.e. index j and its relatives in the data-set) is denoted as n_j, $j = 1, \ldots, r$. These numbers cannot be retrieved from the data due to missing information on the families.

If the estimation method is not corrected for the way the data were ascertained, this would result in overestimation or underestimation of the disease penetrance. Correction of the likelihood can be done by conditioning on the ascertainment event (see Appendix 1 for the details) and the corrected likelihood is equal to

$$\frac{\prod_{i=1}^{n} f(T_i)^{\Delta_i}(1 - F(C_i))^{1-\Delta_i} g(C_i)}{\prod_{j=1}^{r}\{1 - [\int [1 - F(s)] dG(s)]^{n_j}\}} \tag{1}$$

In the likelihood, no notation is used (and necessary) to distinguish between families, because it is not known from the data which individuals are related.

2.2 Estimation

The conditional likelihood in equation (1) contains the unknown distributions F and G with densities f and g, and the sizes of the families in the data-set $n_j, j = 1, \ldots, r$. In the setting discussed in this paper, no family information is present to retrieve these numbers n_j from the data. Therefore, the average family size in the data set is taken instead: $\bar{n} = \sum_{j=1}^{r} n_j / r$. This average can be calculated as the number of index patients is assumed to be known. However, if more information is available for determining the values of n_j, this information can also be used. Inserting these averages into equation (1) yields

$$\frac{\prod_{i=1}^{n} f(T_i)^{\Delta_i}(1 - F(C_i))^{1-\Delta_i} g(C_i)}{\prod_{j=1}^{r}\{1 - [\int [1 - F(s)]\, dG(s)]^{\bar{n}}\}} \tag{2}$$

This likelihood is used for estimating F and G. As the average $\bar{n}$ might deviate from the true values $n_1, \ldots, n_r$, this may affect the unbiasedness of the estimators. This is studied in section 3.

An estimator for F is obtained by maximising this likelihood. To make maximisation computationally feasible within a foreseeable computing time, it is assumed that $F = F_\theta$ belongs to a class of parametric distribution functions (for instance the class of Weibull distributions with unknown shape and the scale parameter).

Two methods for estimating G are considered. In the first method, it is assumed that G belongs to a class of parametric distributions. This class could be chosen based on the form of the Kaplan–Meier curve or the empirical distribution of the observed censoring times C of relatives with or without the germline variant (if data for the latter is available). Next, the estimates of the unknown parameters that determine G and θ (the unknown parameter of F_θ) can be found by simultaneous maximisation of the likelihood in equation (2). A confidence interval for θ and a pointwise confidence interval for F_θ can be constructed using the Delta method.[6]

A second estimate for G can be constructed as follows. For relatives who turn out to be non-carriers after genetic testing, there is no follow-up. However, if the age at the time of testing is available, the censoring times can be determined. Then, G can be estimated by the empirical distribution function of these censoring times, denoted as $\hat{G}_{NC}$ (NC stands for non-carrier). After inserting $\hat{G}_{NC}$ into the likelihood, it can be maximised with respect to θ to obtain an estimate for F_θ. The construction of a pointwise confidence interval for F_θ can be done based on the likelihood ratio statistic (see Appendices 2 and 3). If the censoring times cannot be determined for non-carriers, G could be estimated based on the censoring times of all carriers except the index patients, denoted as $\hat{G}_C$ (C stands for carrier). Simulation studies show that this latter estimate gives slightly biased estimates for F_θ (see section 3.1).

If people may die from the disease, then C may be right censored by death due to the disease, and the likelihood that is derived at the end of Appendix 1 should be used in stead. However, the two-step estimation method as described in this section will yield exactly the same estimate for F_θ (see Appendix 1 for more information).

R-code for computing the estimates and the confidence interval is available upon request.

3 Results

3.1 Simulation study

In this section, the results of simulation studies are evaluated to study the effect of replacing $n_j, j = 1, \ldots, r$ by the average $\sum_k n_k/r$ in the likelihood and the effect of the way of estimating G (as described in section 2.2) on the accuracy of the estimate for the penetrance F_θ. Furthermore, the performance of the Kaplan–Meier estimator for F_θ based on simulated data of the carriers excluding the index patient is studied.

The simulation of data is performed in five steps.

Step 1. The number of tested carriers of the genetic variant in, say, pedigree j in the data, n_j, is simulated as described below.

Step 2. For all n_j individuals, the age at diagnosis of the disease, T, is simulated from a Weibull distribution with shape and scale parameters equal to 2.5 and 90 (motivated by the real data example below).

Step 3. The age at time of censoring, C, is simulated from a uniform distribution at the interval [20, 80] (this choice for G is based on the estimate in the application, but other distributions have been considered as well).

Step 4. If at least one of the individuals has an age-at-onset T with $T \leq C$, the pedigree is included in the data-set.

This simulation schedule is repeated 10,000 times. The total number of individuals or pedigrees that are selected depends on the choice of the distribution for the values n_j and the chosen distribution F_θ.

Step 5. Independently of the simulated data so far, 1000 censoring times C are simulated from the same distribution as chosen in step 4. These are the simulated censoring times for the non-carriers in the ascertained pedigrees in step 4. Its empirical distribution function will be used for estimating G (see the settings A and B below).

In total, three simulation studies are performed. In the first simulation, n_j can take the values 1, 2, or 3 with probabilities 0.5, 0.25, and 0.25. In the second simulation, these values are 1, 2, 3, 4, and 5 and they are sampled with equal probabilities (probability 1/5) and in the third study the values are 1, 2, 3, 4, 5, and 10 which are also sampled with equal probability (probability 1/6). This leads to three data-sets of approximately 3500, 5000 and almost 6000 families. Next, the model parameters are estimated as described in the previous section. Furthermore, the Kaplan–Meier estimate based on the data of all carriers except the index patients is determined. The whole procedure, simulation and estimation, is repeated 250 times.

Estimates for F_θ are obtained by maximising likelihood (1) (i.e. n_j is known) or likelihood (2) (i.e. n_j is replaced by $\bar{n}$), after inserting estimate $\hat{G}_{NC}$ or $\hat{G}_C$ into the likelihood. These four settings are labeled as:

A: The likelihood in equation (1) with G estimated by $\hat{G}_{NC}$ is maximised.
B: The likelihood in equation (2) with G estimated by $\hat{G}_{NC}$ is maximised.
C: The likelihood in equation (1) with G estimated by $\hat{G}_C$ is maximised.
D: The likelihood in equation (2) with G estimated by $\hat{G}_C$ is maximised.

Figure 1 shows boxplots of the estimates of $F_\theta(50)$ (left column) and $F_\theta(70)$ (right column) in the different settings and studies. The true values $F_\theta(50)$ and $F_\theta(70)$ are displayed by the horizontal dashed curves (and are the same in all simulation studies). The three rows in Figure 1 correspond with the three sets from which is simulated. The boxplots in the first row present the results from the first simulation (the numbers n_j are simulated from 1, 2, and 3), those in the second row for the second study (n_j simulated from 1, 2, 3, 4, and 5), and the boxplots in the third row for the third study (n_j is simulated from 1, 2, 3, 4, 5, and 10). Boxplots based on the estimates of the Kaplan–Meier curve are missing, because these are far from the other estimates, but the results are given in Table 1.

In the figure it can be seen that the penetrance function F_θ is slightly overestimated if the true sizes n_j in the likelihood are replaced by the average $\bar{n}$, but the bias is small. Furthermore, if G is estimated by $\hat{G}_{NC}$ the estimator of F_θ seems to be unbiased, whereas the estimators show a systematic bias downwards if G is estimated by $\hat{G}_C$. The biases in the third simulation (third row) seem to be larger than in the first simulation (first row). This is probably due to the degree of skewness of the distribution of the sizes n_j. By lack of information in the data on the distribution of n_j, more robust alternative estimates for n_j are not available.

Table 1 shows the medians of the estimates for every setting as well as for the Kaplan–Meier estimate, denoted as F_{naive}. It can be seen that this estimator strongly underestimates the penetrance function. The amount of bias depends on the values of n_j; it decreases with increasing size n_j as can be seen in Table 1. By increasing the pedigree

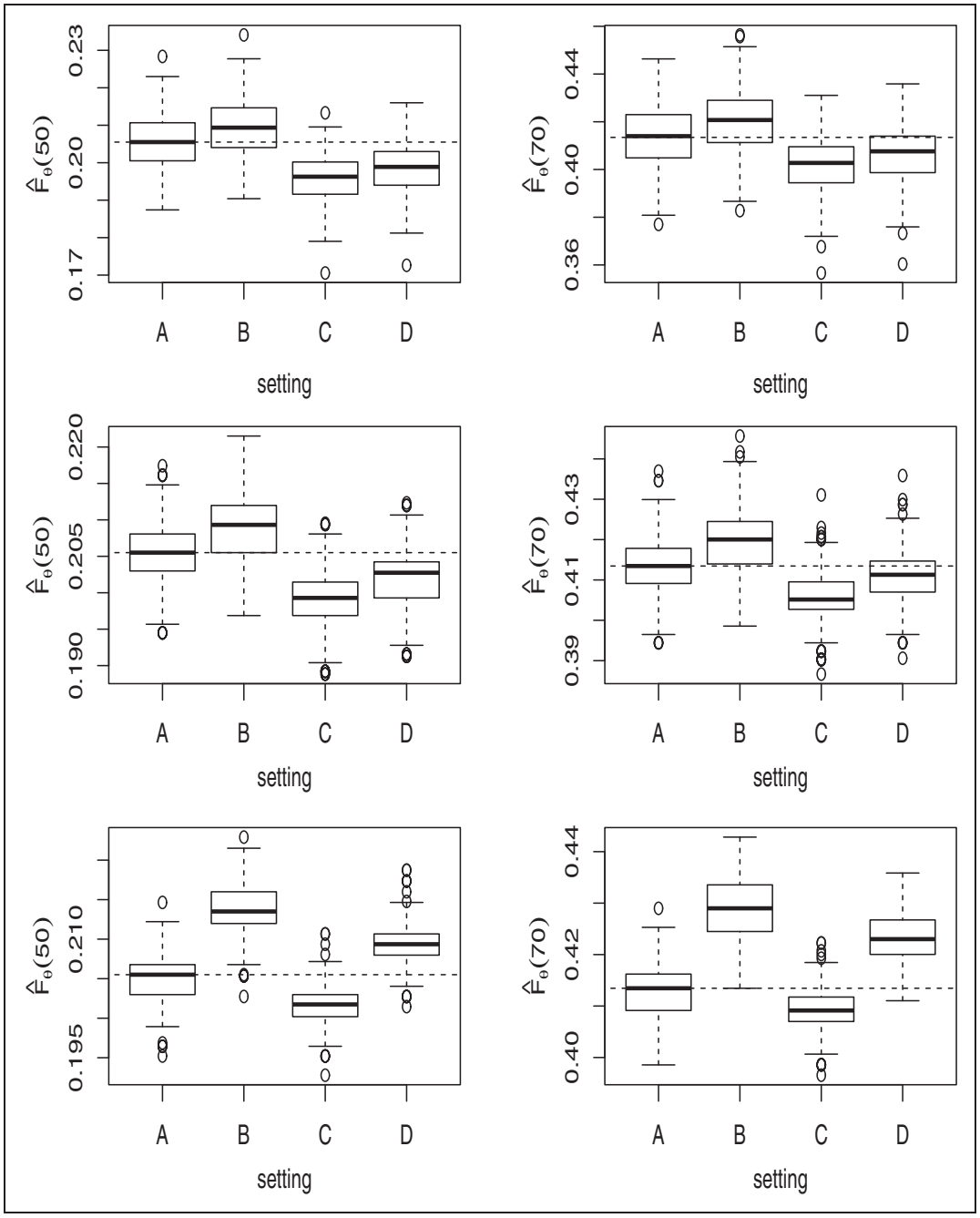

Figure 1. Boxplots for the results of the simulation studies with F_θ equal to the Weibull distribution with the shape and scale parameters equal to 2.5 and 90, at ages 50 (left) and 70 (right).

sizes even further to $n_j = 50$ for all families, the bias of the naive estimator decreases further: $\hat{F}_{naive}(50) = 0.193$ and $\hat{F}_{naive}(70) = 0.392$ while the true values equal 0.205 and 0.413, respectively.

In the application section, the maximum likelihood estimate of F_θ is plotted together with the Kaplan–Meier estimates based on data of relatives with and without index patients (left plot in Figure 3). It can be seen that the Kaplan–Meier estimator that includes the index patients overestimates and the Kaplan–Meier estimator that excludes the index patients underestimates the penetrance. The figure shows the need for correction for the way the data were ascertained.

Several more simulations have been performed (see Table 2 in Appendix 4 for the results). From the results in Tables 1 and 2, it is concluded that the proposed estimation method performs well, whereas the naive

Table 1. Median of the estimates of F at the ages 50 and 70 by the estimates in the settings A, B, C and D and in the three different studies, and the naive estimator $\hat{F}_{naive}$ (the Kaplan–Meier estimator as described before).

	Study 1		Study 2		Study 3	
age	50	70	50	70	50	70
F	0.205	0.413	0.205	0.413	0.205	0.413
A	0.206	0.414	0.206	0.413	0.206	0.413
B	0.209	0.421	0.209	0.420	0.214	0.429
C	0.196	0.403	0.199	0.405	0.202	0.409
D	0.199	0.408	0.203	0.411	0.209	0.423
$\hat{F}_{naive}$	0.124	0.260	0.130	0.272	0.142	0.294

Note: The first row yields the true values.

Kaplan–Meier estimator based on data of all carriers except the index patients dramatically underestimates the true penetrance curve. If the distribution of n_j is extremely skewed to the right, the penetrance seems to be systematically overestimated, but performs still better than the naive Kaplan–Meier estimator.

3.2 Application

The development of the method described in this paper was inspired by a study that was aimed at estimating the risk of developing a paraganglioma or pheochromocytoma (rare, usually benign neuroendocrine neoplasms) in carriers of germline SDHB gene variant.[7] The penetrance calculations were based on the age at detection of a paraganglioma/pheochromocytoma in affected carriers and the present age in unaffected living carriers or the age of death in unaffected deceased carriers, of 61 index patients and 133 relatives, 20 of which were affected with the disease at the end of the study and 113 of which were not. The study population was identified through the clinical genetics centers in The Netherlands. All individuals had consented to DNA testing, and all included individuals were identified as carriers of a SDHB gene variant predisposing to paraganglioma/pheochromocytoma. No data from familial non-carriers were available, nor from family members that did not consent to DNA testing. In the data analysis, it is assumed that all individuals without a family history of the associated disease at the time of the DNA screening are index patients; they are the first diagnosed patients in the family. Furthermore, it is assumed that those individuals with a positive family history at the time of DNA screening are relatives of an index patient.

Three different families of distributions are considered for the penetrance function F_θ: a Weibull distribution, a gamma distribution both with unknown shape and scale parameters, and a chi-squared distribution with an unknown degree of freedom.

The average pedigree size in the data-set $\bar{n}$ is equal to $(133 + 61)/61 = 3.2$. Since there is no data available of non-carriers, the distribution G is estimated by the empirical distribution based on the censoring times of the relatives only (the plot on the right in Figure 2). The straight line in the same plot is the distribution function for the uniform distribution at the interval [20,80].

The three models (with the different families of distributions for F_θ) are fitted to the data and their fits are compared based on their AIC. The models in which F_θ is assumed to be a Gamma or a Weibull distribution fit equally well and the estimates for F_θ overlap, whereas the fit of the model with the chi-squared distribution for F_θ performed badly. The maximum likelihood estimate for the penetrance based on likelihood (2) and under the assumption of a Weibull distribution is shown in Figure 2. The shape and scale parameters equal 2.47 and 90. (The maximum likelihood estimate of the shape and scale parameter under the assumption of a gamma distribution equal 3.57 and 24.6. The two maximum likelihood estimates of F_θ overlap on the interval of interest: up to 75 years.)

As demonstrated in Figure 3 (left), the Kaplan–Meier estimator based on the data including the index patients without correction for the way the data are ascertained, overestimates the penetrance function, whereas it underestimates the penetrance if data of non-index carriers is used only.

The maximum likelihood estimator is accurate if no data from relatives is missing and all carriers within a pre-specified degree have been tested. However, from the available data the number of carriers that have not been DNA-tested cannot be deduced, and since pedigree structures are not present, it is impossible to include obligate carriers in the data set manually or correct for family members with uncertain carrier status. The bias introduced by not including these individuals depends on the number of missing individuals and their genotypic and

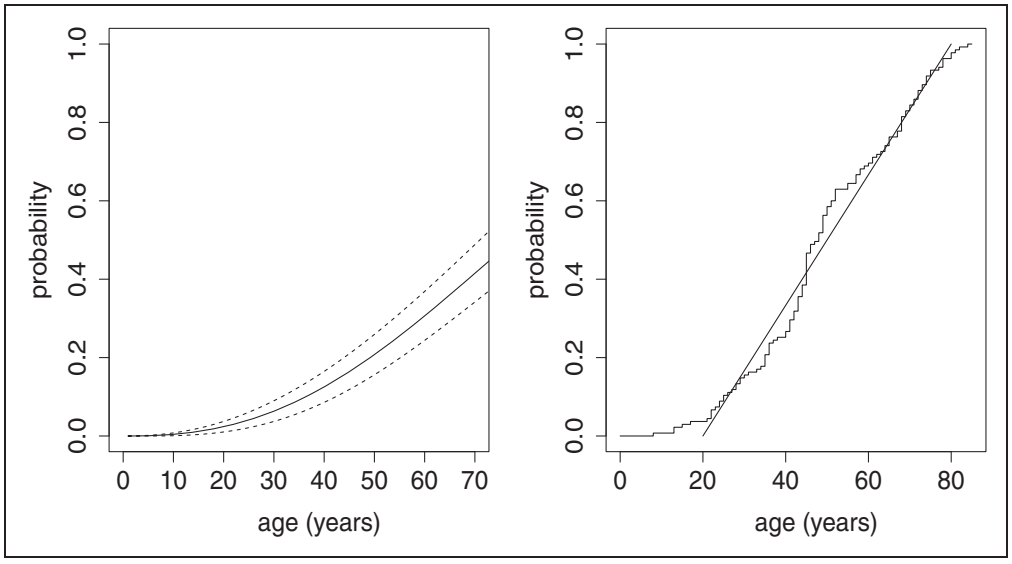

Figure 2. Left: Maximum likelihood estimate for F_θ (black) and confidence interval (dashed). Right: Empirical distribution for G based on the censoring times of the relatives. Straight line: distribution function for the uniform [20,80] distribution.

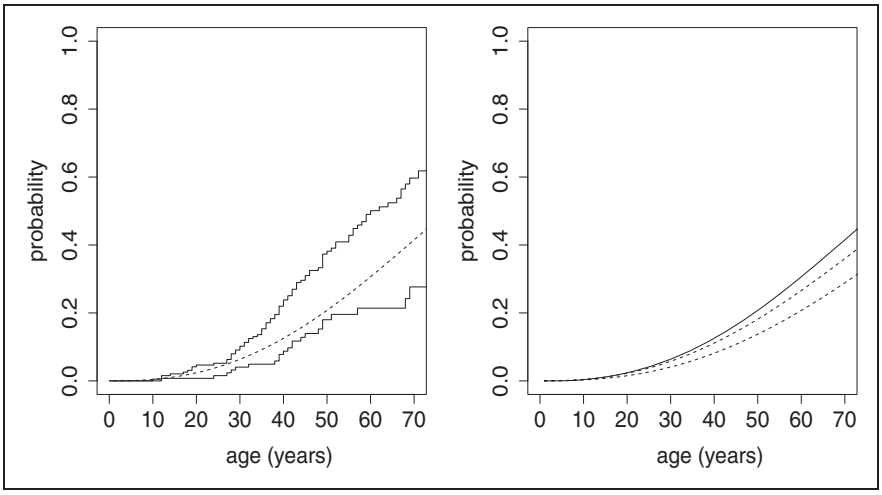

Figure 3. Left:Kaplan–Meier estimates with (upper step-function) and without (lower step-function) index patient and the maximum likelihood estimate (continuous dashed line). Right: Maximum likelihood estimate (continuous line). The dashed lines indicate the range of possible bias due to missing of 20% individuals. Only a bias towards overestimation in the dataset is evaluated.

phenotypic status. It is likely that the missing family members are not entirely missing at random. It is likely that family members who experience symptoms are more inclined to consent to DNA testing than family members who are asymptomatic, which may lead to an over-representation of affected family members and thus an overestimation of the penetrance. To quantify a possible bias, we performed a simulation study for a scenario of maximal bias. We assumed that 20% of the relatives (33 individuals) are not tested and are therefore missing in the data set, and that all missing individuals are disease-free carriers (one could also assume that all missing individuals are affected carriers, but as affected family members are less likely to be missing from the data set, this possibility is not further evaluated here). From all unaffected carriers in the dataset, 33 are selected at random (with replacement) and added to the data. Next, the maximum likelihood estimate for the penetrance F_θ is computed as before. This procedure is repeated 250 times, leading to 250 estimates of F_θ. All 250 simulations resulted in a penetrance function between the two dashed lines in Figure 3 (at the right). The continuous line represents the maximum likelihood estimate based on our true cohort. As demonstrated, a lack of 20% of the

pedigree data leads to an overestimation of the penetrance if all missing individuals are disease-free carriers of the genetic alteration of interest. This bias increases with age but remains limited even at advanced age (approximately 0.1 at the age of 70).

4 Discussion

In this paper, we present a method for the estimation of the age-related penetrance of pathogenic gene variants in a cohort of multiple families with missing family data. A conditional likelihood is proposed, corrected for the way the data were ascertained. Since essential information is missing, full correction is not possible. Simulation studies show that the proposed estimator is only slightly biased in the simulation settings considered. Moreover, it is able to correct for the way the data are ascertained in a more accurate and robust way than conventional methods such as the Kaplan–Meier estimator based on non-index carriers.

The proposed estimator was based on a maximum likelihood method. These estimators are known to be almost always asymptotically unbiased and have a minimum variance.[8] As in any statistical analysis, some assumptions have been made: the first assumption is that the ages at onset of gene variant carriers are statistically independent. Environmental or genetic factors other than the gene variant of interest may also affect the age at onset of the disease and may be shared among family members. As it is unclear which individuals in the data-set belong to the same family and which do not, it is impossible to address these possible modifiers of disease risk and to correct for it.

For some diseases the age-at-onset distribution may depend on genetic characteristics, like the variant at the disease-susceptibility gene or the number of repeat units (for instance for the disease FSHD). To account for these genetic characteristics, the penetrance function F could be modeled by a (parametric) Cox-regression model that includes these genetic characteristics as covariates.

The estimation of the penetrance of pathogenic gene variants is ideally based on a large number of affected pedigrees, collected in a study with clear ascertainment rules, with a high uptake of DNA tests by family members and detailed descriptions of the phenotype of all variant carriers. However, many factors such as the severity and actionability of the disease, the availability of clinical or genetic information, familial relationships and dynamics, and the way the data are ascertained, may cause relevant data to be missing even after considerable effort to retrieve the information. This is especially true for rare and low-penetrant hereditary disease, as the awareness of the segregation of the hereditary trait within the family, the clinical characteristics and the perceived disease risk are generally lower in families with a limited number of affected family members, reducing the inclination to undergo DNA testing and clinical surveillance. As a consequence, penetrance estimations for rare and low-penetrant hereditary disease are usually based on cohorts comprising a limited number of affected patients and their nuclear families. To quantify the possible bias induced by the missing data as just described, simulation studies can be performed as was demonstrated in the section on the SDHB data.

With the growing insight into the genetic determinants of disease and the growing number of known carriers of a hereditary predisposition to disease, the need for accurate estimation of the disease risk of pathogenic gene variants will increase. Especially in rare and low-penetrant disease, current methods are prone to ascertaiment bias. The method described in this manuscript is aimed at providing more robust penetrance estimates when pedigree information is incomplete, a common situation in daily practice, and is shown to outperform the commonly used Kaplan–Meier estimator.

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. Gong G and Whittemore AS. Optimal design for estimating penetrance of rare mutations of a disease-susceptibility gene. *Genet Epidemiol* 2003; **24**: 173–180.

2. Carayol J and Bonaiti-Pellie C. Estimating penetrance from family data using a retrspective likelihood when ascertainment depends on genotype and age of onset. *Genet Epidemiol* 2004; **27**: 109–117.
3. Choi YH. *On combining family data from different study designs for estimating disease risk associated with mutated genes*. Epidemiology Insights, Croatia: INTECH Open Access Publisher, 2012.
4. Hsu L, Ping Zhao L and Aragaki C. A note on a conditional approach for family-based association studies of candidates genes. *Hum Hered* 2000; **50**: 194–200.
5. Kraft P and Thomas DC. Bias and efficiency in family-based gene-characterization studies: conditional, prospective, retrospective, and joint likelihoods. *Am J Hum Genet* 2000; **66**: 1119–1131.
6. van der Vaart AW. *Asymptotic statistics*. Cambridge, UK: Cambridge University Press, 1998.
7. Rijken JA, Niemeijer ND, Jonker MA, et al. The penetrance of paraganglioma and pheochromocytoma in sdhb germline mutation carriers. *Clinical Genetics* 2018; **93**: 60–66.
8. Bijma F, Jonker MA and van der Vaart AW. *An introduction to mathematical statistics*. Amsterdam: Amsterdam University Press, 2017.
9. Jonker MA and van der Vaart AW. On the correction of the asymptotic distribution of the likelihood ratio statistic if nuisance parameters are estimated based on an external source. *Int J of Biostat* 2014; **10**: 123–142.

Appendix 1. Derivation of the likelihood in equation (1)

In the setting of independent observations T and C, the likelihood for a single person is given by

$$f(T)^{\Delta}(1-F(C))^{1-\Delta}g(C)$$

like in the standard right censoring case (the indices are left out for simplification of the notation). If the age at onset of the disease T is observed, then $\Delta=1$ and the likelihood is equal to $f(T)\,g(C)$ (age at time of censoring is always observed). Similarly, if the age at onset is not observed, then $\Delta=0$ and the likelihood is given by $(1-F(C))\,g(C)$; the product of the conditional probability that $T>C$ given C, and the density g at C.

Remind that n_j is defined as the number of individuals in the data-set that are related to index j. Under the assumption that the ages at onset of the disease and the ages at time of censoring between carriers (within a family) are independent, the likelihood for the data of the individuals related to index j is given by the product

$$\prod_{i=1}^{n_j} f(T_{ij})^{\Delta_{ij}}(1-F(C_{ij}))^{1-\Delta_{ij}}g(C_{ij})$$

where the double index (ij) refers to individual i among those related to index j.

If pedigrees are ascertained via affected individuals and their sibships (like the setting we are interested in this paper), the likelihood must be adjusted for this to obtain unbiased estimates. Let A_j be the event that individuals in pedigree j are ascertained. Then, the conditional likelihood for the individuals in pedigree j, given the pedigree was ascertained equals

$$L_j=\frac{\prod_{i=1}^{n_j} f(T_{ij})^{\Delta_{ij}}(1-F(C_{ij}))^{1-\Delta_{ij}}g(C_{ij})}{\mathrm{P}(A_j)}$$

The denominator in the likelihood, $\mathrm{P}(A_j)$, corrects for the way the data were ascertained. This probability is a function of the unknown distributions F and G. Often, people get DNA testing offered, once a relative (the index-patient) expresses the disease phenotype and carries the genetic variant of interest. Those who turn out to carry the variant will be followed and regularly screened for the disease. When aiming for estimating the penetrance of the disease, the follow-up data (until that moment) of all known carriers (the index patients and their relatives with the variant) are collected. So, disease data are collected of families with at least one affected individual (the index and possibly relatives). Then the ascertainment event A_j is defined as the event that at least one individual is affected with the disease and $\mathrm{P}(A_j)$ equals the probability that this occurred and is equal to 1 minus the probability that none of them was affected at the time of collecting the data: $\mathrm{P}(A_j)=1-\mathrm{P}(T_{1j}>C_{1j},\ldots,T_{n_jj}>C_{n_jj})$. This probability can be rewritten as

$$\begin{aligned}\mathrm{P}(A_j)&=1-\mathrm{P}(T_{1j}>C_{1j},\ldots,T_{n_jj}>C_{n_jj})\\&=1-\mathrm{P}(T_{1j}>C_{1j})^{n_j}=1-\left(\int 1-F(s)\mathrm{d}G(s)\right)^{n_j}\end{aligned}$$

by independence and because $P(T_{ij} > C_{ij}) = \int (1 - F) dG$. This yields the likelihood

$$L_j = \frac{\prod_{i=1}^{n_j} f(T_{ij})^{\Delta_{ij}} (1 - F(C_{ij}))^{1-\Delta_{ij}} g(C_{ij})}{1 - [\int [1 - F(s)] \, dG(s)]^{n_j}}$$

for family j.

Then, the full likelihood for all pedigrees and all individuals in the data-set equals

$$\prod_{j=1}^{r} L_j = \frac{\prod_{j=1}^{r} \prod_{i=1}^{n_j} f(T_{ij})^{\Delta_{ij}} (1 - F(C_{ij}))^{1-\Delta_{ij}} g(C_{ij})}{\prod_{j=1}^{r} \{1 - [\int [1 - F(s)] dG(s)]^{n_j}\}}$$

The numerator can be rewritten so that no distinction in the notation is made between families. To simplify notation all individuals in the data-set have a single unique index $i, i = 1, \ldots, n$. This yields the likelihood

$$L = \frac{\prod_{i=1}^{n} f(T_i)^{\Delta_i} (1 - F(C_i))^{1-\Delta_i} g(C_i)}{\prod_{j=1}^{r} \{1 - [\int [1 - F(s)] dG(s)]^{n_j}\}}$$

as given in equation (1). Note that for computing this likelihood it is not necessary to know the composition of the families. This is essential because this information is not available from the data.

In the foregoing it was assumed that T and C are stochastically independent; the disease is not lethal. In the following, the likelihood is derived in case the disease is lethal.

Define Z_i as the age at time of death due to the disease for individual i, with distribution G_Z and density g_Z. Moreover, define Σ_i as the indicator function that equals 1 if $C_i < Z_i$ and 0 if $C_i \geq Z_i$. Then, a likelihood can be derived as described above. This likelihood is very similar to the likelihood in the previous display, with the only exception that the term $g(C_i)$ in the numerator of the conditional likelihood should (only) be replaced if $\Delta = 1$; it should be replaced by $g(C_i)(1 - G_Z(C_i))/(1 - G_Z(T_i))$ if $\Delta_i = \Sigma_i = 1$ and by $(1 - G(Z_i)) g_Z(Z_i)/(1 - G_Z(T_i))$ if $\Delta_i = 1$ and $\Sigma_i = 0$. Although the likelihood has got a slightly different form, it will not change the estimate for the penetrance F found by maximising the likelihood after inserting an estimate for G, since G_Z is not present in the denominator and is present in the numerator only as a multiplication factor (it does not have to be estimated to obtain an estimate for F).

Appendix 2. Construction of the confidence interval

When constructing a confidence interval, ignoring the fact that G is unknown and estimated may yield an interval with a coverage lower than it should be. An alternative approach for constructing a confidence interval is described in this appendix. Suppose the distribution function F_θ is a parametric distribution with unknown parameter $\theta \in \Theta$. The distribution G, the nuisance parameter, is estimated by the empirical distribution function based on the censoring times of m non-carriers, so that the estimator, denoted as $\hat{G}_{NC}$, is independent of the data used for estimating θ. After inserting the estimator $\hat{G}_{NC}$ into the likelihood, an estimate for θ is found by maximising this partially estimated likelihood: $\hat{\theta} = \text{argmax}_\theta \log L(\theta; \hat{G}_{NC})$. In the paper[9] the asymptotic distribution of the likelihood ratio statistic for testing the parameter of interest θ is derived under the null hypothesis that θ equals a value, say ξ, after the nuisance parameter in the model, G, is estimated with data from an external independent data-set of size m. The theorem given in Jonker and van der Vaart[9] states that for $Y_1, Y_2, \ldots, Y_r$ independent and identically distributed, the corresponding likelihood ratio statistic

$$\Lambda_{r,m} = 2 \log \frac{\max_{\theta \in \Theta} L(Y; F_\theta, \hat{G}_{NC})}{L(Y; F_\xi, \hat{G}_{NC})}$$

is asymptotically distributed as $\sum_{i=1}^{d} (1 + \lambda d_i) Z_i^2$ with $\lambda = \lim_{r,m} r/m$ and $d_1, \ldots, d_d$ the eigenvalues of the covariance matrix $i_{\theta_0, G_0}^{-1/2} j_{\theta_0, G_0} i_{\theta_0, G_0}^{-1/2}$ (with θ_0 and G_0 the true parameter values) and $Z_1^2, \ldots, Z_d^2$ independent random variables with a chi-squared one distribution. The matrix i_{θ_0, G_0} is defined as the Fisher information matrix for θ in the model where $G = G_0$ is known and j_{θ_0, G_0} is the covariance matrix of the asymptotic distribution of $\sqrt{m} P_{\theta_0, \hat{G}_{NC}} \dot{l}_{\theta_0, G_0}$.

From the theorem it can be seen that if m is much bigger than r, λ will be close to zero and the asymptotic distribution of $\Lambda_{r,m}$ will be close to a chi-square distribution with the number of degrees of freedom equal to the dimension of θ. Then, the distribution of the test-statistic is hardly affected by the fact that the nuisance parameter is unknown and estimated. However, this also holds if r/m is not close to zero, but the eigenvalues $d_1, \ldots, d_d$ are. It is not only the sample sizes of the two samples that determines the deviation from the chi-squared distribution, but also the estimation precision of the estimators (the asymptotic variances) is important. Since the proposed estimator for G, the empirical distribution based on the m censoring times, converges at a high rate, it is expected that the eigenvalues will be small (it is known that $\sqrt{m}||\hat{G}_{NC} - G_0||_\infty$, the Kolmogorov–Smirnov statistic, converges in distribution (to the supremum of a Gaussian process), where $||.||_\infty$ is the notation for the supremum norm[6]).

The likelihood ratio test-statistic can be used to derive a confidence interval for θ: the confidence interval for θ equals all values ξ for which the null hypothesis $H_0 : \theta = \xi$ is not rejected.[8] After constructing an interval for θ, a pointwise interval for $F_\theta(t)$ can be constructed.

In our application, the number of independent observations r equals the number of independent pedigrees (or index patients). Since the pedigrees do not necessarily have the same size, the data are not identically distributed. However, under the assumption of equal pedigree sizes the theory holds. The Fisher information matrix can be estimated as $-\log \ddot{L}(\hat{\theta})$, with $\ddot{L}$ the second derivative. Once a parametric family for F_θ has been chosen, this derivative can be found by straightforward calculations. The calculations that lead to an estimator for j_{θ_0,G_0} are tedious and are given in Appendix 3. The theorem just described holds under some assumptions which are summed up in Jonker and van der Vaart.[9] These assumptions have been checked by straightforward calculations (calculations not shown), and they are satisfied. In the application in this paper, the eigenvalues turned out to be small (0.02305 and 0.003286), small enough to ignore and to set λ equal to zero.

It is appealing to estimate G based on the censoring times of all individuals, relatives with or without mutations, especially if m is equal to zero (i.e. no data of non-mutation carriers are available) or small compared to the number of non-index mutation carriers. However, by doing this, the independence assumption (of two independent data-sets) is violated and the asymptotic distribution of the likelihood ratio statistic may change. Moreover, the empirical distribution for G based on these data is not asymptotically unbiased, but will show a small bias, because of ascertainment. However, if data of non-mutation carriers are not available, this is a good alternative way for finding a confidence interval.

Appendix 3. Computation of the matrix $j_{\theta_0,G}$

In this appendix, the asymptotic covariance matrix of $\sqrt{m}P_{\theta_0,\hat{G}_{NC}}\dot{l}_{\theta_0,G}$, the matrix $j_{\theta_0,G}$, is computed. Since $P_{\theta_0,G}\dot{l}_{\theta_0,G} = 0$, it holds that

$$\begin{aligned}\sqrt{m}P_{\theta_0,\hat{G}_{NC}}\dot{l}_{\theta_0,G} &= \sqrt{m}(P_{\theta_0,\hat{G}_{NC}} - P_{\theta_0,G})\dot{l}_{\theta_0,G} \\ &= \sqrt{m}(P_{\theta_0,\hat{G}_{NC}} - P_{\theta_0,G})\left\{(p-1)\left(\Delta\frac{\dot{f}_\theta(T)}{f_\theta(T)} - (1-\Delta)\frac{\dot{f}_\theta(C)}{1-F_\theta(C)}\right)\right. \\ &\quad \left. + \left(\frac{\dot{f}_\theta(T_i)}{f_\theta(T_i)} - \frac{d}{d\theta}\log\int F_\theta(s)\mathrm{d}G(s)\right)\right\}\end{aligned}$$

We compute the terms one by one.

$$\begin{aligned}\sqrt{m}(P_{\theta_0,\hat{G}_{NC}} - P_{\theta_0,G})\Delta\frac{\dot{f}_\theta(T)}{f_\theta(T)} &= \sqrt{m}\int\int_{t<c}\frac{\dot{f}_\theta(t)}{f_\theta(t)}f_\theta(t)\mathrm{d}t\mathrm{d}(\hat{G}_{NC} - G) \\ &= \frac{d}{d\theta}\sqrt{m}\int\int_{t<c} f_\theta(t)\mathrm{d}t\mathrm{d}(\hat{G}_{NC} - G) = \frac{d}{d\theta}\sqrt{m}\int F_\theta(c)\mathrm{d}(\hat{G}_{NC} - G) \\ &= \sqrt{m}\int \dot{F}_\theta(c)\mathrm{d}(\hat{G}_{NC} - G) = \sqrt{m}\mathbb{G}_m\dot{F}_\theta\end{aligned}$$

and

$$\sqrt{m}(P_{\theta_0,\hat{G}_{NC}} - P_{\theta_0,G})(1-\Delta)\frac{\dot{f}_\theta(C)}{1-F_\theta(C)} = \sqrt{m}\int\int_{t>c}\frac{\dot{f}_\theta(c)}{1-F_\theta(c)}f_\theta(t)\mathrm{d}t\mathrm{d}(\hat{G}_{NC}-G)$$
$$= \sqrt{m}\int_c \frac{\dot{f}_\theta(c)}{1-F_\theta(c)}(1-F_\theta(c))\mathrm{d}(\hat{G}_{NC}-G)$$
$$= \sqrt{m}\int_c \dot{f}_\theta(c)\mathrm{d}(\hat{G}_{NC}-G) = \sqrt{m}\mathbb{G}_m\dot{f}_\theta$$

and

$$\sqrt{m}(P_{\theta_0,\hat{G}_{NC}} - P_{\theta_0,G})\frac{\dot{f}_\theta(T)}{f_\theta(T)} = \sqrt{m}\int_c\int_t \frac{\dot{f}_\theta(t)}{f_\theta(t)}f_\theta(t)\mathrm{d}t\mathrm{d}(\hat{G}_{NC}-G)$$
$$= \sqrt{m}\int_c\int_t \dot{f}_\theta(t)\mathrm{d}t\mathrm{d}(\hat{G}_{NC}-G) = 0$$

and

$$\sqrt{m}(P_{\theta_0,\hat{G}_{NC}} - P_{\theta_0,G})\frac{d}{d\theta}\log\int F_\theta(s)\mathrm{d}G(s) = 0$$

Taking the results together yields that

$$\sqrt{m}P_{\theta_0,\hat{G}_{NC}}\dot{l}_{\theta_0,G} = (p-1)\sqrt{m}\mathbb{G}_m(\dot{F}_\theta + \dot{f}_\theta)$$

is asymptotically normal with mean zero and a covariance matrix j_θ. If $\theta = (\eta, \gamma)$ is two-dimensional, j_θ is estimated as the sample covariance matrix

$$\hat{j}_\theta^{(1,1)} = (p-1)^2\left\{\frac{1}{m}\sum_{i=1}^m(\hat{\dot{F}}_\eta(C_i)+\hat{\dot{f}}_\eta(C_i))^2 - \left(\frac{1}{m}\sum_{i=1}^m\hat{\dot{F}}_\eta(C_i)+\hat{\dot{f}}_\eta(C_i)\right)^2\right\}$$
$$\hat{j}_\theta^{(2,2)} = (p-1)^2\left\{\frac{1}{m}\sum_{i=1}^m(\hat{\dot{F}}_\gamma(C_i)+\hat{\dot{f}}_\gamma(C_i))^2 - \left(\frac{1}{m}\sum_{i=1}^m\hat{\dot{F}}_\gamma(C_i)+\hat{\dot{f}}_\gamma(C_i)\right)^2\right\}$$
$$\hat{j}_\theta^{(1,2)} = (p-1)^2\left\{\frac{1}{m}\sum_{i=1}^m(\hat{\dot{F}}_\gamma(C_i)+\hat{\dot{f}}_\gamma(C_i))(\hat{\dot{F}}_\eta(C_i)+\hat{\dot{f}}_\eta(C_i))\right.$$
$$\left. -\frac{1}{m}\sum_{i=1}^m(\hat{\dot{F}}_\gamma(C_i)+\hat{\dot{F}}_\gamma(C_i))\frac{1}{m}\sum_{i=1}^m(\hat{\dot{F}}_\eta(C_i)+\hat{\dot{f}}_\eta(C_i))\right\}$$

where $\hat{\dot{F}}_\eta$ and $\hat{\dot{f}}_\eta$ equal the derivative functions of F_θ and f_θ with respect to η and the unknown parameters are replaced by its estimates. The calculations for γ can be done analogues..

Appendix 4. Additional simulation studies

Additional to the simulations in Section 3.1, more simulations have been performed to study the performance of the estimators (proposed in Section 2.2) in different settings. In this appendix, the results of these simulations are described. The way of simulating the data is exactly the same as described in Section 3.1, but the distributions and the values of the parameters vary (Study 1, 2, 3 described below).

- Study 1. Lower penetrance function
  The penetrance function F_θ equals a Weibull distribution with shape and scale parameters equal to 1.0 and 300. The values n_j are drawn from numbers 5 to 8 all with probability equal to 0.25. The distribution for the ages at the time of censoring G equals the uniform distribution at the interval [20,80] as before.

Table 2. Median of the estimates of F at the ages 50 and 70 by the estimates in the settings A, B, C and D and in the three different studies, and the naive estimator $\hat{F}_{naive}$ (the Kaplan–Meier estimator as described before).

	Study 1		Study 2		Study 3	
age	50	70	50	70	50	70
F	0.154	0.208	0.205	0.413	0.0527	0.152
A	0.154	0.208	0.206	0.413	0.0527	0.152
B	0.154	0.209	0.226	0.448	0.0527	0.153
C	0.153	0.207	0.204	0.413	0.0512	0.150
D	0.154	0.208	0.224	0.448	0.0515	0.150
$\hat{F}_{naive}$	0.0936	0.128	0.169	0.347	0.0281	0.0850

Note: The first row yields the true values.

- Study 2. Skew distribution for n_j
  In this simulation study, the penetrance function F_θ equals the Weibull distribution with shape and scale parameters equal to 2.5 and 90. The distribution G is chosen as before. The values n_j are simulated from numbers 1 to 5 and 25, all with equal probability. Since pedigrees with size $n_j = 25$ have a higher probability to be ascertained (to have at least one individual with the disease: $T \leq C$), these pedigrees will be over represented in the data set.
- Study 3. Gamma distribution
  In the third simulation study, the penetrance function is taken equal to a gamma distribution with shape and scale parameters equal to 5.0 and 25. The values n_j are drawn from numbers 1 to 4 all with equal probability. The distribution G is as before.

In all different studies, the penetrance function F_θ is estimated by the five estimators as described in Section 3.1. For every simulation studies, data are simulated and parameters are estimated 150 times, yielding 150 estimates for every estimator for F_θ. For every estimator, the median of the estimates for $F_\theta(50)$ and $F_\theta(70)$ are calculated. The results are summarised in Table 2.