HAL
archives-ouvertes.fr

Semi-Parametric Estimation of Survival in Age-Dependent Genetic Disease. Application to the Transthyretin-related Hereditary Amyloidosis

Flora Alarcon, Gregory Nuel, Violaine Planté-Bordeneuve

▶ **To cite this version:**

Flora Alarcon, Gregory Nuel, Violaine Planté-Bordeneuve. Semi-Parametric Estimation of Survival in Age-Dependent Genetic Disease. Application to the Transthyretin-related Hereditary Amyloidosis. MAP5 2016-30. 2016. <hal-01343862v2>

HAL Id: hal-01343862
https://hal.archives-ouvertes.fr/hal-01343862v2

Submitted on 31 Oct 2016

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Semi-Parametric Estimation of Survival in Age-Dependent Genetic Disease

Application to the Transthyretin-related Hereditary Amyloidosis

Flora Alarcon [1], Gregory Nuel [2,3], Violaine Planté-Bordeneuve [4,5]

October 31, 2016

[1] Mathématiques appliquées Paris 5 (MAP5) CNRS : UMR8145 – Université Paris Descartes – Sorbonne Paris Cité, Paris, France

[2] Institute of Mathematics (INSMI), National Center for French Research (CNRS), Paris, France

[3] Laboratory of Probability (LPMA), Université Pierre et Marie Curie, Sorbonne Université, Paris, France

[4] Hôpital Universitaire Henri Mondor, Département de Neurologie Créteil, France

[5] Inserm, U955-E10, Créteil, France

Abstract

Mendelian diseases are determined by a single mutation in a given gene. However, in the case of diseases with late onset, the age at onset is variable; it can even be the case that the onset is not observed in a lifetime. Estimating the survival function of the mutation carriers and the effect of modifying factors such as the gender, mutation, origin, etc, is a task of importance to provide individual risk assessment, both for management of mutation carriers and for prevention. In this work, we present a semi-parametric method based on a proportional

to estimate the survival function of the mutation carriers using pedigrees ascertained through affected individuals (probands). Not all members of the pedigree need to be genotyped. The ascertainment bias is corrected by using only the phenotypic information from the relatives of the proband, and not of the proband himself. The method manage ungenotyped individuals through belief propagation in Bayesian networks and uses an EM algorithm to compute a Kaplan-Meier estimator of the survival function. The method is illustrated on simulated data and on samples of families with transthyretin-related hereditary amyloidosis, a rare autosomal dominant disease with highly variable age of onset.

keywords : Semi-Parametric Survival function, Kaplan-Meier estimator, familial data, Believe Propagation

1 Introduction

In monogenic disease with variable age of onset, a precise estimation of the survival function for mutation carrier individual is necessary as well as identification of potential factors that modulate this age. Indeed, this estimates allows to provide individual risk assessment to understand the underlying mechanisms of the disease and to establish prevention strategies. Even if the method can easily accommodate the incidence among non-carrier, we here consider the case where non-carrier individuals cannot be affected (i.e. only the carrier of the genetic mutation can develop the disease). But sometimes, the age at onset can be so late that a significant proportion of mutation carriers does not declare the disease in the lifetime. We called this phenomenon the incomplete penetrance. It should be noted that in this literature, the age-specific cumulative distribution function (CDF), named also penetrance function, is preferentially evoked. In this paper, we will use the classical survival function (which is simply the complementary of the CDF) to assess the probability of not being affected by the disease according to the age for mutation carrier individuals. Note that our survival function hence corresponds to the *cause-specific* survival (disease diagnosis) and not to the *overall* survival.

Because of the low carriage frequency and the hight cost of genetic test, random sampling is not a praticable approach to obtain a sample of sufficient size to draw reliable conclusions. Data are usually obtained from families ascertained through affected individuals. Indeed, as all affected individual necessary carry the mutation, the families ascertained in this way are informative to estimate the survival function. The drawback of this procedure is that the survival function can be significantly overestimated if the ascertainment process is not taking into account [1]. Therefore, an adjustment for the ascertainment bias is required.

The ascertainment correction problem is a very challenging problem. Vieland and Hodge explain in their articles [2, 3] that "*without knowledge of the true underlying pedigree structure (including who are the unobserved members of pedigree) it is not possible to write down a correct likelihood and the ascertainment correction problem becomes intractable*". However, different adjustments for ascertainment have already been suggested to provide valid risk estimate of a genetic disease [4, 5, 6].

In monogenic disease, as all affected individuals carry the mutation, an ascertainment through affected individuals is sufficient to have mutation carriers. When pedigree are ascertained through at least one affected individual, it is possible to correct ascertainment bias by modeling analytically the ascertainment correction [7, 8]. However, this prospective correction require additional parameters as π, the probability for an affected to be ascertained, which have to be estimated and make the strong assumption that all affected have the same chance to be ascertained.

Another more intuitive method, the PEL, have been proposed that corrects for ascertainment by simply removing the phenotypic information of the individual (called *proband*) who allowed his family to be selected [6] (i.e. proband's phenotype exclusion). A similar method had been proposed by Weinberg [9, 10] to correct for the ascertainment bias in the estimation of segregation ration (see also [11]). Moreover, it has been shown (in [6]) for various genetic models and selection schemes that PEL corrects better than the prospective method.

The PEL is a parametric method estimating age-dependent risks of monogenic diseases in mutation carriers using disease status and genotypic information of family members in pedigrees ascertained through affected individuals. In this parametric method, the age at onset is modeled by a Weibull distribution. Although this model is widely used in survival analysis because of its capacity to adjust to observed data, it can fail to fit properly the survival function in some cases. The advantage of a non-parametric estimation of the survival function has been shown in [12], as well as the ability of the proband's phenotype exclusion to correct for the ascertainment bias in this context. However, the method proposed in [12] assume that all genotypes are observed, which is a strong hypothesis that prevent any application on real data set.

In this article, we introduce a semi-parametric method based on a proportional hazard (Cox model) to estimate survival function from familial data. The presence of ungenotyped individuals in the families are managed through belief propagation in Bayesian networks which allows to estimate, for all unaffected individual, his probability to be a carrier. This probability is then taken into account in the survival function estimation through weights.

The method uses an EM algorithm to compute a Kaplan-Meier estimator of the survival function and correct for the ascertainment bias by excluding the proband's phenotype, like in the PEL.

Another advantage of this non parametric method is its ability to accommodate covariates (as gender, mutation, etc.) thanks to the Cox model. The EM algorithm can be summarized as follow : 1) Survival function are estimated with arbitrary weight standing for the individual probability to carry the mutation in the M-step. 2) Then the weight is assessed according to the estimated survival function in the E-step.

The PEL handles unobserved genotypes through the Elston-Stewart [13] algorithm that is computationally heavy and not able to manage for loops in pedigrees. For this reason, our new method will not be compared with the PEL in real data, but preferentially with a ≪Weibull estimation ≫that provide same results than PEL and that handles unobserved genotypes with belief propagation in Bayesian networks.

Section 2 presents first the estimation model and then it describes the believe propagation in this contexte as well as the E-M algorithm. Section 3 presents the results obtained on simulated data sets and Section 4 illustrates the method on transthyretin-related hereditary amyloidosis families from different origin (French, Portuguese, and Swedish). For French families, two different mutations are compared through a log-rank test. In Portuguese dataset, as only the Val30Met mutation is present, our non-parametric estimation is compared with a Weibull estimation. Finally, methodology and results are discussed in the Section 5.

2 Semi-parametric estimation of the survival function

2.1 The model

Let's consider n individuals indexed by $i = 1, \ldots, n$. For an individual i, we denote by (T_i, δ_i) the vector defined for $T_i \leq 0$ and $\delta_i \in \{0, 1\}$ as follows :

$$T_i = \begin{cases} \text{age at diagnosis} & \text{if } \delta_i = 1 \\ \text{age at last follow-up} & \text{if } \delta_i = 0 \end{cases}$$

We denote by $X_i \in \{00, 01, 10, 11\}$, the genotype of individual i where the first number represents the number of disease allele ($\in \{0, 1\}$) transmitted by the father and the second one represents the number of disease allele ($\in \{0, 1\}$) transmitted by the mother. So $X_i = 01$ means that the indi-

vidual i carry the mutation, that he is heterozygous and that his mutation have been transmitted to him by his mother. Note that this variable is often unobserved because individual are rarely genotyped, and that distinguishing betwee, "01" and "10" usually requires to take into account the whole pedigree structure.

Finally, we consider the vectors of dimension n of the sample : $\boldsymbol{T} = (T_1, \ldots, T_n)$, $\boldsymbol{\delta} = (\delta_1, \ldots, \delta_n)$, $\boldsymbol{X} = (X_1, \ldots, X_n)$, and we assume the following model:

$$\mathbb{P}(\boldsymbol{T}, \boldsymbol{\delta}, \boldsymbol{X}) = \underbrace{\prod_{i=1}^{n} \mathbb{P}(T_i, \delta_i | X_i)}_{\text{survival part}} \times \underbrace{\prod_{i=1}^{n} \mathbb{P}\left(X_i | X_{\text{father}_i}, X_{\text{mother}_i}\right)}_{\text{genetic part}}$$

where father_i and mother_i indicate the father and mother of individual i (empty information for the founders). Through the product over all the individuals of the pedigree, the model hence allows to take into account the full history of the disease and genetic dependance in the complete family. We will now detail both the survival and the genetic part of this model.

2.1.1 Survival part.

We assume that the disease of interest is distributed across non-carrier according to a proportional hazard model with known baseline and regression coefficient. For any rare mutation, these parameters are typically estimated from the general population or from a specific population of non-carrier. Note that for genetic disease such as a non-carrier cannot be affected (ex: transthyretin-related hereditary amyloidosis) we simply have:

$$\mathbb{P}(T_i, \delta_i | X_i = 00) = 1$$

For the sake of simplicity we will consider only this particular case from now on but nothing in the presented method forbid to consider the more general model where the disease can occur as well in non-carrier.

For the mutation carriers, we consider a dominant model with incomplete penetrance and proportional hazards (PH):

$$\log \mathbb{P}(T_i, \delta_i | X_i \neq 00) = -\Lambda(T_i) e^{\boldsymbol{Z}_i \boldsymbol{\beta}} + \delta_i \left(\lambda(T_i) + \boldsymbol{Z}_i \boldsymbol{\beta}\right)$$

where $\lambda(t)$ is the baseline hazard, $\Lambda(t) = \int_0^t \lambda(u) du$ is the baseline cumulative hazard, $\boldsymbol{Z}_i$ are the covariates of individual i, and $\boldsymbol{\beta}$ is a regression coefficient.

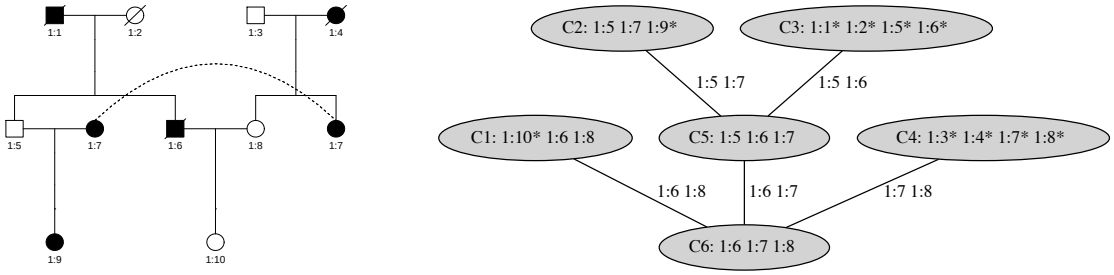
The term of ≪semi-parametric≫comes from the coexistence of a non-parametric part (i.e. $\Lambda(t)$) and a parametric part (i.e. $e^{\boldsymbol{Z}_i\boldsymbol{\beta}}$). In our method, λ is estimated with the Nelson–Aalen estimator that is a non-parametric piecewise constant baseline survival $S(t) = \exp(-\Lambda(t))$; the classical non-parametric choice in survival analysis (N.B.: in the particular case where there is no covariates in the model, this estimator is due to Kaplan-Meier and is therefore often improperly referred under this name even in the presence of covariates). However, other forms can be considered for λ (ex: Weibull, exponential, lognormal, etc.). In this case the model becomes entirely parametric and $\lambda(t)$ is the density of, for exemple, a Weibull distribution that is a classical parametric choice in the context of survival analysis. Other popular choices include the exponential or log-gamma distributions. The particular case of a Weibull distibution will be treated in the application section.

2.1.2 Genetic part.

We assume a classical genetic model : Hardy-Weinberg equilibrium is assumed in pedigree founders and the disease allele frequency q is assumed to be known for the founders. Moreover, Mendelian transmission of the alleles from parents to offspring is assumed. Since our n individuals might belong to completely independent families, it is clear that the genetic likelihood can be computed separately on this independent families, however, the notations are still valid, dramatically simpler by ignoring the family level.

Another important point is the fact that the true genotype X_i is at best partially observed. Indeed, a positive mutation search or an affected individual, only indicates that $X_i \neq 00$ is impossible. On the other hand, a negative mutation search indicates that $X_i = 00$ (assuming a 100% sensitivity of the mutation search procedure). More complex model allowing for genotyping errors or even pedigree errors (wrong filiation for example) can be incorporated like in [14]. In the present work, we decided to use the most basic (but reasonable) model.

In our method, unknown genotypes are taking into account thanks to Belief Propagation that provides, for each ungenotyped individual, his/her probability to carry the disease mutation. Belief propagation (BP) in pedigree is a very general method which can deal efficiently with very complex pedigree structure (ex: 2000 individuals with 50 loops). Unlike Elston-Stewart algorithm, BP does



(a) (b)

Figure 1: The "mating loop" example. Panel a: the pedigree (the dashed line indicates that individual "1:7" is represented twice in the pedigree but is in fact a single individual); Panel b: a junction tree corresponding to the problem (the '*' indicate the locations where individual evidences are injected into the junction tree).

not use loop breaking approaches to deal with loop pedigrees. Instead, BP use an auxiliary tree called the junction tree (JT) which basically is a clique decomposition of the moral graph corresponding to the pedigree problem. JT and BP are well known is the graph theory (ex: JT can be used to solve a graph coloring problem) and in the mathematical field of probabilistic graphical models (Bayesian network, hidden Markov model, decision trees, Markov networks, etc.).

In Figure 1a we represent a simple example pedigree with a mating loop. This is typically a pedigree which would require to perform loop breaking (for example on Individual 1:7) in order to be solved by Elston-Stewart. Here we build instead the JT of Figure 1b in which the *evidence* (see above) is injected prior to the BP. Then BP consists in computing and propagating recursively so-called *messages* (denoted M) from the leaves to the root. Here we use evidence of Individual 1:10 to compute $M_{1,6}$ from C_1 to C_6, then evidence of Individual $1:9$ for $M_{2,5}$, evidence of Individuals 1:1, 1:2, 1:5 and 1:6 for $M_{3,5}$, then $M_{2,5}$ and $M_{3,5}$ for $M_{5,6}$, then evidence of Individuals 1:3, 1:4, 1:7 and 1:8 for $M_{4,6}$, and finally $M_{1,6}$, $M_{5,6}$, and $M_{4,6}$ at the root. After this *inward* propagation, evidence can be recursively propagated back to the leaves (*outward* propagation) in order to obtain marginal posterior distribution of the variables.

Let us see what give BP on our example assuming that allele frequency is $q = 1\%$ and that

i	$x = 00$	$x = 10$	$x = 01$	$x = 11$
1:1	0.000	0.494	0.494	0.012
1:2	0.965	0.017	0.017	0.000
1:3	0.965	0.017	0.017	0.000
1:4	0.000	0.495	0.495	0.010
1:5	0.389	0.591	0.009	0.012
1:6	0.000	0.977	0.012	0.012
1:7	0.000	0.010	0.975	0.016
1:8	0.486	0.009	0.496	0.009
1:9	0.000	0.203	0.590	0.207
1:10	0.365	0.374	0.129	0.132

Table 1: Posterior distribution $\mathbb{P}(X_i = x|\text{ev})$ computed though BP for the "mating loop" example.

all affected are carrier (no other information is provided). The posterior marginal distribution for all individuals in the pedigree is given in Table 1. Without surprise, we observe that all affected individuals ($i = 1, 4, 6, 7, 9$) cannot have the non-carrier genotype 00. If we look to individual 1:4, she has genotypes 10 or 01 with equal probability 0.495 and hence, she can be an homozygous carrier with probability 0.01, the allele frequency, which is consistent. Now, individual 1:7 is also a carrier, but the fact that her mother is indeed a carrier makes much more likely that her genotype is 01, and this is clearly accounted by the BP.

2.2 The EM algorithm

Since $\boldsymbol{X}$ is only partially observed, we consider this variable as latent and use a classical Expectation-Maximization algorithm in order to maximize the model log-likelihood in parameter $\boldsymbol{\theta} = (\lambda, \boldsymbol{\beta})$ (allele frequency q is assumed to be known). For this purpose, we first need to incorporate the auxiliary Q function:

$$Q\left(\boldsymbol{\theta}|\boldsymbol{\theta}_{\text{old}}\right) \stackrel{\text{def}}{=} \int \mathbb{P}\left(\boldsymbol{X}|\text{ev};\boldsymbol{\theta}_{\text{old}}\right) \log \mathbb{P}(\boldsymbol{T}, \boldsymbol{\delta}, \boldsymbol{X}; \theta) d\boldsymbol{X} = \sum_{i=1}^{n} w_i \log \mathbb{P}(T_i, \delta_i | X_i \neq 00)$$

where ev denote the evidence (that means $\boldsymbol{T}$, $\boldsymbol{\delta}$, and any mutation search information), and where the weights w_i are defined as:

$$w_i \stackrel{\text{def}}{=} \mathbb{P}\left(X_i \neq 00 | \text{ev}; \boldsymbol{\theta}_{\text{old}}\right)$$

2.2.1 E-step

The auxiliary function Q is computed at this step. In our case, the marginal weights w_i are all we need to compute. Due to the complex dependency structure of the genotype in our pedigree, this is however a challenging task. In the particular case where the pedigree is a simple tree, the classical Elston-Stewart algorithm can be used [13]. When loops are present (consanguinity, mating loops, twins), Elston-Stewart must be combined with loop breaking approaches at the cost of an exponentional complexity with the number of loops [15]. Instead of Elston-Stewart we consider here the belief propagation algorithm (also called sum-product algorithm) in Bayesian network which can be see as a generalization[1] of Elston-Stewart to arbitrary pedigrees. See Section 2.1.2 for more details on belief propagation.

The only information we need to provide for this step is called the *evidence* ev and is defined as:

$$\text{ev}_i(x) = 1_{\{X_i = x \text{ compatible}\}} \times \begin{cases} 1 & \text{if } \delta_i = 1 \\ \mathbb{P}(T_i, \delta_i = 0 | X_i = x) & \text{if } \delta_i = 0 \end{cases}$$

(ex: $x \neq 00$ is incompatible with a negative mutation search on an affected individual). Note that the evidence 1 for affected individual can be used because the T_i is non-informative for the distribution of X_i in this particular case. This is also better for our non-parametric estimation which cannot provide hazard estimate without smoothing (ex: kernel smoothing) but only survival estimates.

2.2.2 M-step

The auxiliary function Q is maximized at this step. As seen above, our auxiliary function can be simply seen as the classical log-likelihood of a survival model where each individual observation receive the weight w_i. This is hence a very classical problem which can easily be handled via clas-

[1] even if belief propagation was developed independently by the probabilist graphical model community.

sical statistical optimization procedure using the programming software R [16] and the `survival` package [17, 18].

2.2.3 Practical implementation

Initialization is performed by affecting random weights w_i (ex: drawn from a uniform distribution on $[0, 1]$). EM iterations are stopped when we observe convergence on test survival estimates (ex: baseline survival at age $20, 40, 60, 80$). The 95% pointwise confidence intervals are simply provided by the standard (weighted) Kaplan-Meier estimation of the incidence.

3 Analysis of simulated datasets

3.1 Simulation of pedigrees

We simulated families with realistic size and structure from 35 French families with transthyretion-related hereditary amyloidosis (see the next section on analysis of real data). We duplicated families k times (here we fixed $k = 3$) in order to have a larger sample of 105 families. Ages, gender and the proband individual in each families is given by the real dataset. Genotypes are assigned respecting Mendelian transmission, with a disease allele frequency $q = 0.004$ in our simulated dataset. Moreover, the gender of the transmitting parent is not taking into account in this work (no distinction between $X = 01$ and $X = 10$). The age at event is simulated according to a piecewise constant hazard rate function, $\lambda(t)$, given as follows. An uniform censoring data between 15 and 80 years is added in order to censure 30% of individuals. After what, only families with at least one affected individual is retained in the dataset, representing the ascertainment. Finally, a sample of 2604 individuals is analyzed.

$$\lambda(t) = \begin{cases} 0 & \text{if } t \in [0, 20] \\ 0.02 & \text{if } t \in [20, 40] \\ 0.10 & \text{if } t \in [40, 60] \\ 0.05 & \text{if } t > 60 \end{cases}$$

The last step of the simulation is to set all genotypes to “unknown”, so that all simulated data are analysed without knowledge of the genotypes.

3.2 Method assessment on a simple simulated dataset

We first assessed the method on a simple simulated dataset regardless additional covariates in the model. Figure 2 shows that our semi-parametric method succeed in estimating the true survival curve (red curve) even if all genotypes are considered as missing. Furthermore, the sample size leads to smaller confidence intervals. As the true survival curve is very similar to the estimated one, we can also conclude that the ascertainment bias has been correctly corrected since families who have disease mutation but no affected individual have been not ascertained.

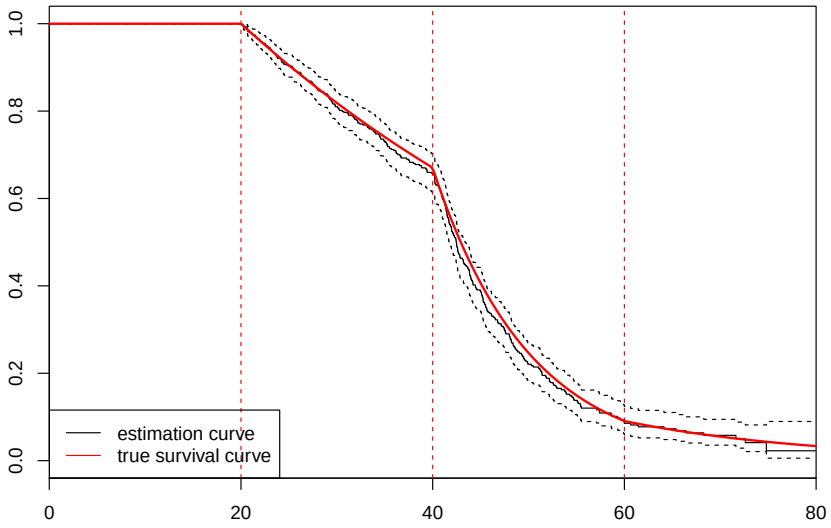


Figure 2: Estimation of the baseline survival function $S_0(t)$ with confidence intervals for a simulated dataset

3.3 Stratification or proportional hazard to take account for covariate

When covariates are available, we have the choice between stratify on these covariates or take account on the covariate in the proportional hazard model. Firstly, in our simulated dataset, we simulate the age at event separately according to different piecewise constant hazard rate depending on the gender of the individual. Indeed, the piecewise constant hazard rate function by gender is defined as follows :

$$\lambda_{\text{gender}=1}(t) = \begin{cases} 0 & \text{if } t \in [0, 20] \\ 0.02 & \text{if } t \in [20, 40] \\ 0.10 & \text{if } t \in [40, 60] \\ 0.05 & \text{if } t > 60 \end{cases}$$

$$\lambda_{\text{gender}=2}(t) = \begin{cases} 0 & \text{if } t \in [0, 30] \\ 0.01 & \text{if } t \in [30, 50] \\ 0.08 & \text{if } t \in [50, 70] \\ 0.02 & \text{if } t > 70 \end{cases}$$

Figure 3 shows estimations of survival curve (black lines) stratified for males (gender=2) and females (gender=1). 95% confidence intervals are provided through polygons. Here, a non-parametric version of our method is used, as the covariate "gender" is not considered in a cox model. We note that our method provides good estimated curves as noted previously.

Then, a proportional protector effect of the female gender (i.e. gender=1) are added in simulations through a cox model. Indeed, The β parametric parameter of the model have set to $\beta = -0.4$. Thus, the women's survival curve is higher than men's. This simulated dataset has been analyzed with our semi-parametric method with $Z =$ gender as covariate. Figure 4 shows estimations of the survival curve (with 95% confidence interval in dotted lines) for gender=1 and for gender=2 (black curves in thin lines). The β parameter was estimated by $\hat{\beta} = -0.56$. This bias in the estimation of β explains why the survival estimated for gender=1 is more biased than the survival estimated for gender=2.

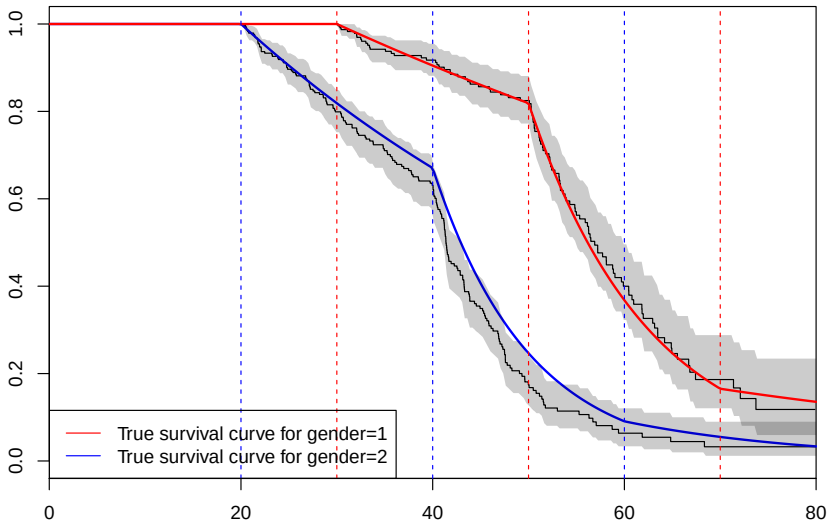


Figure 3: Estimation of Survival curve with stratified gender effect

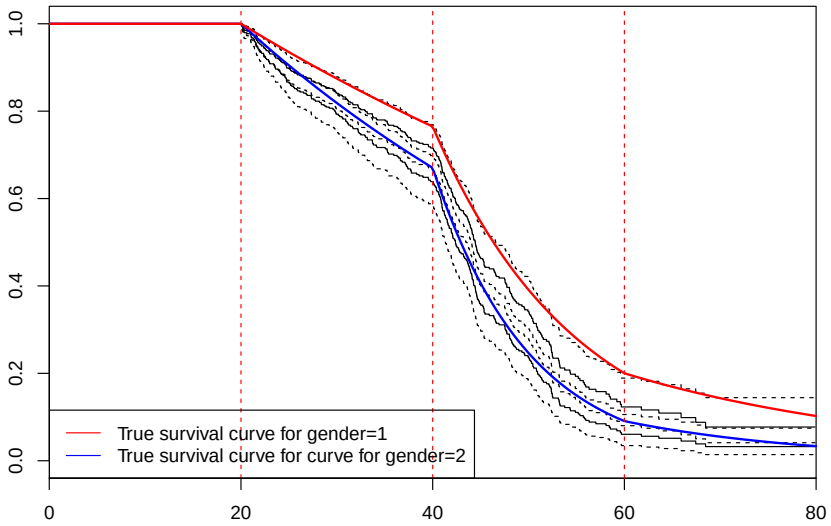


Figure 4: Estimation of Survival curve with a proportional gender effect

4 Analysis of real data

We illustrated the method on transthyretin-related hereditary amyloidosis, an autosomal dominant disease, caused by a mutation of the TTR gene, Val30Met (MET30) substitution being the most frequent mutation [19]. The age at onset ranges from early twenties to late seventies. Although distributed worldwide, the disease is often clustered in limited areas like in Portugal, Japan and Sweden with different genotypic and phenotypic variation. In France, we are dealing with two populations, i.e. of Portuguese and of French origins. While many pathogenic TTR variants have been detected among French population, only one variant, the MET30, was detected in the Portuguese population. In this setting, a better knowledge of the risk of being symptomatic for carriers (or penetrance that is the complementary of the survival function) is needed to ensure a better follow-up of carriers and to manage of patients at the very onset of symptoms. It may also give clues on our understanding of pheno-genotypic variability.

We analyse three data set constituted of 49 families of French descent, 33 families of Portuguese descent and 78 families of Swedish descent, ascertained through affected individuals (see Table 2). Data are analyzed excluding the proband to avoid ascertainment biais (as done in simulations) and the deleterious allele frequency was arbitrarily set to $q = 0.004$ and the *de novo* mutation was set to 0.

For the Portuguese and the Swedish dataset, we compared the semi-parametric estimation of survival curve obtained by our approach with that obtained when the hazard function is modeled by a Weibull distribution.

	French	Portuguese	Swedish
All	1238	1191	1361
Affected	87	178	151

Table 2: Description of French, Portuguese and Swedish families (total number of individuals)

4.1 A French dataset

Among the 30 different substitutions of the TTR observed in families of French descent, MET30 and Ser77Tyr (TYR77) are the most frequent accounting for about 50% of the kindreds. Age at first symptoms is significantly much older than in families of Portuguese descent but appear similar in both variants in the French families.

We analyse a French dataset affected by transthyretin-related hareditary amyloidosis. The sample set consist in 35 families with a mutation MET30 and 15 families with a mutation TYR77. Figure 5 shows the survival curves estimated stratified on the type of mutation. A log-rank test was performed with the R function *surdiff* in order to compare the two mutations. Thus, a significant difference is tested between survival curve for MET30 mutation (black curve) and TYR77 mutation (red curve) with a p-value estimated to 0.002. 95% confidence intervals are given through colored regions.

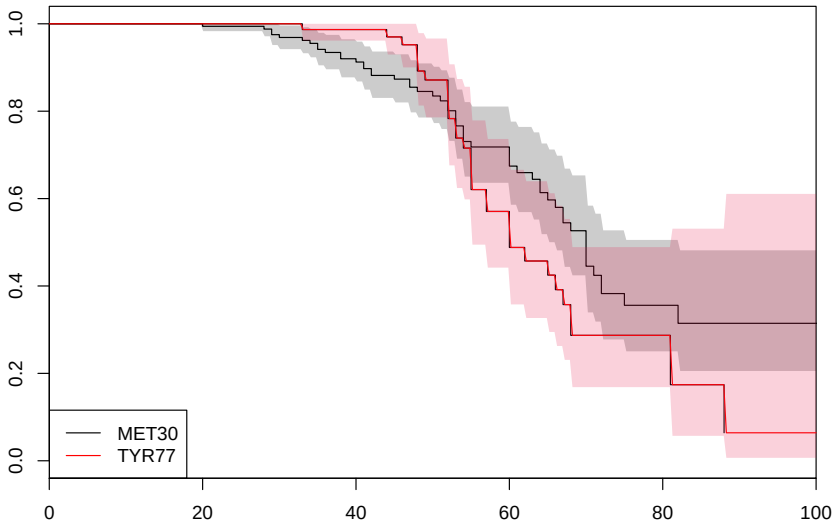


Figure 5: Estimation of Survival curve stratified on mutation in the French dataset

4.2 A Portuguese dataset

In this section, we analyse a data set constituted of 33 families of Portuguese descent, first described in [6]. Figure 6 shows the survival curves estimated with a proportional gender effect, given with 95% confidence intervals in dotted lines. The proportional parameter β is estimated to $\hat{\beta} = -0.327$ with a p-value $p = 0.034$, indicating that survival is higher for women than for men. We can note that the survival is lower in Portuguese data set than in French data set. This results have already been shown in [8].

Figure 7 shows the comparison between our semi-parametric method without any covariate and a Weibull parametric estimation assessed through a E-M algorithm with the R function *Survreg*. We observe that the Weibull estimation does not fit the non-parametric curve.

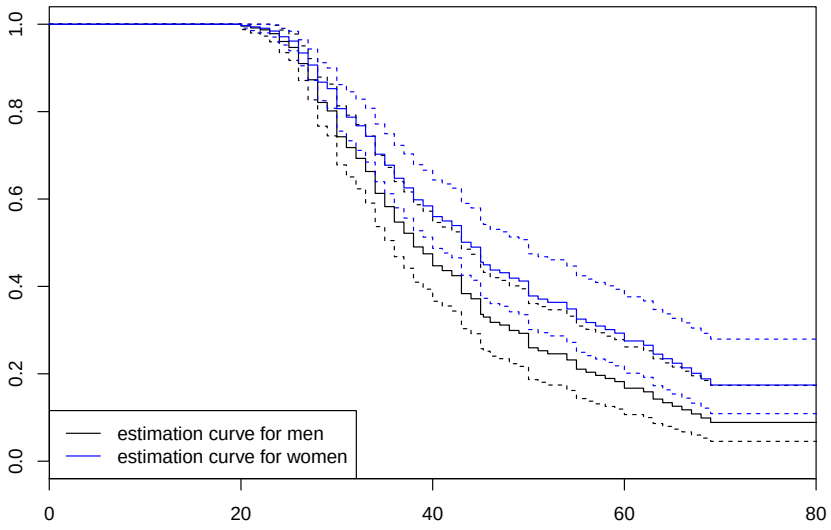


Figure 6: Estimation of the survival function with a proportional gender effect on Portuguese dataset

4.3 A Swedish dataset

In Swedish data, the proportionnal effect on gender was not significant with a p-value estimated to 0.43. Figure 8 shows estimation of the survival curve with our semi-parametric method without

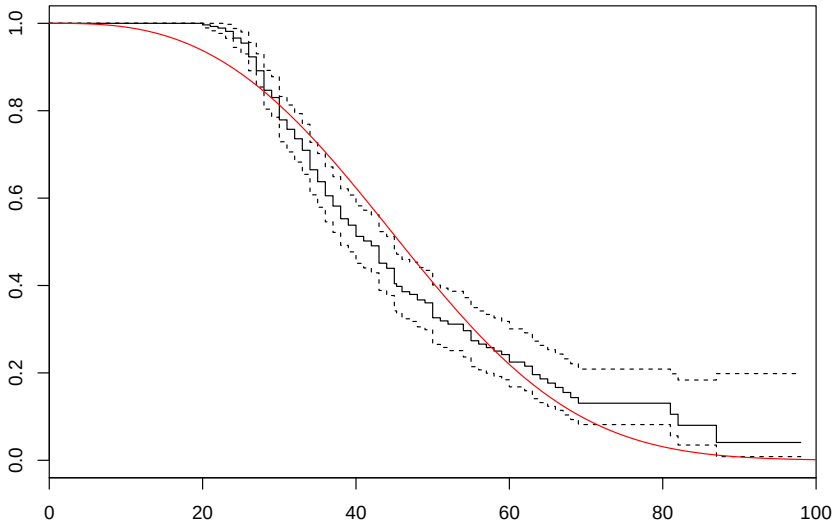


Figure 7: Comparison between a Weibull parametric method to estime the Survival curve in the Portuguese dataset

any covariate (i.e. with the Kaplan-Meier estimator) (black curve) and estimation obtained with a Weibull parametric estimation (red curve). In this case, the Weibull estimation fit almost perfectly the non-parametric curve, with the noticeable exception of the age 90 and more, where the Weibull distribution clearly underestimate the survival curve. Moreover, the survival estimated in the Swedish families is higher than in Portuguese and Val30Met French families. This results are consistent with those found in [20]

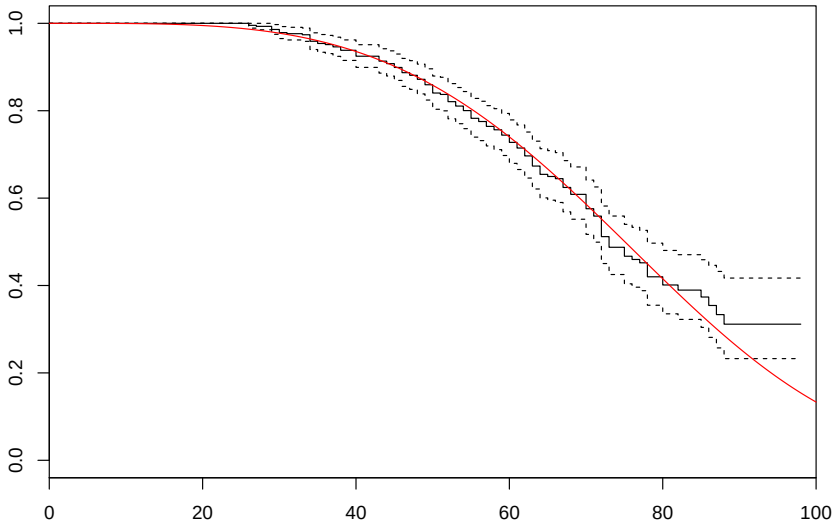


Figure 8: Comparison between a Weibull parametric method to estime the Survival curve in the Swedish dataset

5 Discussion

In this paper, we have proposed a semi-parametric method for estimating survival functions in age-dependent genetic disease using pedigrees with incomplete genotype information. We here considered the particular case where non-carrier cannot be affected (survival of 1.0) and where the genetic model is dominant. However, the method is straightforward to extend to more general models (recessive, relative hazard, etc.) as long as the incidence among non-carrier is known (ex: estimated from the general population).

In the suggested approach, latent genotypes are handled by believed propagation for pedigrees and a EM algorithm allows to estimate Survival curves with weights representing the probability to carry the mutation. The method can accommodate covariates in a proportional hazards model and account for potential stratification on covariates. The believed propagation method is implemented in C++ and EM algorithm is implemented in R.

As the pedigree are ascertained through an affected individual, the proband's phenotype exclusion method is used to avoid ascertainment biais. The problem of ascertainment in segregation

analysis arises when families are selected for study through ascertainment of affected individuals. An important part of the problem is how to handle the pedigree structure, and so to model correctly ascertainment in the likelihood. Statistically, the sampling scheme can be trough as a multistage sampling method (1- one or several probands are ascertained; 2- a sequential sampling scheme is applied). Vieland and al have shown [3] that "*modeling the ascertainment scheme is an intractable problem*". But she has used only sibships. This problem of ascertainment deserves more works and developments. For example, to generalize the Vieland's approaches to arbitrary pedigrees larger than sibships and to more general problems as penetrance function estimation for diseases with variable incidence with age.

Another possible source of bias, when some genotypes are missing among relatives, may be a misspecification of the de novo mutation rate or of the deleterious allele frequency that are commonly fixed to arbitrary low values. The robustness to an error on these two parameters has been already checked in [6].

In the Results Part, we have compared our non-parametric estimation to a Weibull parametric one and have seen that a Weibull parametric estimation fails to fit the survival curve estimated with our method. Additional parameters could be introduced into the Weibull model in order to improve its capacity of adjustment to the data but might involve overparametrization. Moreover, we have not been able to compare our non-parametric method to that introduced in[12] based on empirical likelihood because this last method does not handle unknown genotypes.

An interesting extension of this work would be to account for the possible correlation between member of the same family by including a frailty in the survival function. The familial frailty would typically represent an unknown shared exposure to some environmental factor or to some kind of polygenic effect. However, the estimation of such models is known to be challenging, especially in the context of non-parametric survival estimation [21, 22]. Further investigation will be conducted on this important subject in our forthcoming work.

As illustration, we have estimated Survival function in three samples of different origin : French, Portuguese and Swedich families. We have notices that Survival curves had different estimation according to the origin. Moreover, in comparing our semi-parametric estimation with a Weibull

parametric estimation in Portuguese families, we have observed that the Weibull model did not fit well the survival curve estimated with our method. In [6], Survival function was estimated with an extended Weibull model in which a parameter κ was introduced in order to take into account the possibility that some carriers will never develop the disease and the κ was estimated to 0.09 with a $p < 0.001$ showing that almost 10% of carrier will never develop the disease. We were not able to replicate this observation in the current analysis which clearly questions its relevance.

References

[1] J. Carayol, M. Khlat, J. Maccario, and C. Bonaïti-Pellié. Hereditary non-polyposis colorectal cancer: current risks of colorectal cancer largely overestimated. *Journal of Medical Genetics*, 39(5):335–339, 2002.

[2] Veronica J Vieland and Susan E Hodge. The problem of ascertainment for linkage analysis. *American journal of human genetics*, 58(5):1072, 1996.

[3] V.J. Vieland and S.E. Hodge. Inherent intractability of the ascertainment problem for pedigree data: a general likelihood framework. *American journal of human genetics*, 56(1):33, 1995.

[4] J. Carayol and C. Bonaïti-Pellié. Estimating penetrance from family data using a retrospective likelihood when ascertainment depends on genotype and age of onset. *Genetic Epidemiology*, 27(2):109–117, 2004.

[5] C. Le Bihan, C. Moutou, L. Brugieres, J. Feunteun, and C. Bonaïti-Pellié. ARCAD: a method for estimating age-dependent disease risk associated with mutation carrier status from family data. *Genet Epidemiol*, 12(1):13–25, 1995.

[6] Flora Alarcon, Catherine Bourgain, Marion Gauthier-Villars, Violaine Planté-Bordeneuve, D Stoppa-Lyonnet, and Catherine Bonaïti-Pellié. Pel: an unbiased method for estimating age-dependent genetic disease risk from pedigree data unselected for family history. *Genetic epidemiology*, 33(5):379–385, 2009.

[7] P. Kraft and D.C. Thomas. Bias and Efficiency in Family-Based Gene-Characterization Studies: Conditional, Prospective, Retrospective, and Joint Likelihoods. *The American Journal of Human Genetics*, 66(3):1119–1131, 2000.

[8] V. Plante-Bordeneuve, J. Carayol, A. Ferreira, D. Adams, F. Clerget-Darpoux, M. Misrahi, G. Said, and C. Bonaïti-Pellié. Genetic study of transthyretin amyloid neuropathies: carrier risks among French and Portuguese families. *J Med Genet*, 40(11):e120, 2003.

[9] Wilhelm Weinberg. Methoden und fehlerquellen der untersuchung auf mendelsche zahlen beim menschen. *Arch Rassenbiol*, 9:165–174, 1912.

[10] Wilhelm Weinberg. Mathematische grundlagen der probandenmethode. *Molecular and General Genetics MGG*, 48(1):179–228, 1928.

[11] RA Fisher. The effect of methods of ascertainment upon the estimation of frequencies. *Annals of Eugenics*, 6:13–25, 1934.

[12] F Alarcon, C Bonaıti-Pellié, and H Harari-Kermadec. A nonparametric method for penetrance function estimation. *Genetic epidemiology*, 33:38–44, 2009.

[13] R.C. Elston and J. Stewart. A general model for the genetic analysis of pedigree data. *Human Heredity*, 21(6):523–542, 1971.

[14] Alun Thomas. Gmcheck: Bayesian error checking for pedigreegenotypes and phenotypes. *Bioinformatics*, 21(14):3187–3188, 2005.

[15] K Lange and RC Elston. Extensions to pedigree analysis. *Human Heredity*, 25(2):95–105, 1975.

[16] R Core Team. Ra language and environment for statistical computing. vienna: R foundation for statistical computing, 2014.

[17] Terry Therneau. A package for survival analysis in s. r package version 2.37-4. *URL http://CRAN. R-project. org/package= survival. Box*, 980032:23298–0032, 2013.

[18] Terry M Therneau and Patricia M Grambsch. *Modeling survival data: extending the Cox model*. Springer Science & Business Media, 2000.

[19] Violaine Planté-Bordeneuve and Gerard Said. Familial amyloid polyneuropathy. *The Lancet Neurology*, 10(12):1086–1097, 2011.

[20] Urban Hellman, Flora Alarcon, Hans-Erik Lundgren, Ole B Suhr, Catherine Bonaïti-Pellié, and Violaine Planté-Bordeneuve. Heterogeneity of penetrance in familial amyloid polyneuropathy, attr val30met, in the swedish population. *Amyloid*, 15(3):181–186, 2008.

[21] Terry Therneau. Mixed effects cox models, 2015.

[22] Virginie Rondeau, Yassin Mazroui, and Juan R Gonzalez. frailtypack: an r package for the analysis of correlated survival data with frailty models using penalized likelihood estimation or parametrical estimation. *Journal of Statistical Software*, 47(4):1–28, 2012.