A nonparametric method for penetrance function estimation.

Flora Alarcon, Catherine Bonaïti-Pellié, Hugo Harari-Kermadec

A non parametric method for penetrance function estimation

F. Alarcon[1,2], C. Bonaïti-Pellié[2,1], H. Harari-Kermadec[3,4]

May 16, 2008

[1] *Univ. Paris-Sud, IFR69, UMR-S535, F-94817 Villejuif*

[2] *INSERM U535, BP 1000, F-94817 Villejuif*

[3] *CREST, Statistics Laboratory*

[4] *Université Paris-Dauphine, Ceremade*

Corresponding author

ALARCON Flora
INSERM U535
BP 1000
F-94817 VILLEJUIF
France
Tel. : +33 1 45 59 53 89
E-mail: alarcon@vjf.inserm.fr

Abstract

In diseases caused by a deleterious gene mutation, knowledge of age-specific cumulative risks is necessary for medical management of mutation carriers. When pedigrees are ascertained through at least one affected individual, ascertainment bias can be corrected by using a parametric method such as the Proband's phenotype Exclusion Likelihood, or PEL, that uses a survival analysis approach based on the Weibull model. This paper proposes a non parametric method for penetrance function estimation that corrects for ascertainment on at least one affected: the Index Discarding EuclideAn Likelihood or IDEAL. IDEAL is compared with PEL, using family samples simulated from a Weibull distribution and under alternative models. We show that, under Weibull assumption and asymptotic conditions, IDEAL and PEL both provide unbiased risk estimates. However, when the true risk function deviates from a Weibull distribution, we show that the PEL might provide biased estimates while IDEAL remains unbiased.

Key Words: Risk estimation, ascertainment bias, non parametric method

Introduction

In monogenic diseases with variable age of onset a precise estimation of the cumulative risk of being affected by a given age (called the penetrance function) for mutation carriers is important both to understand the underlying mechanisms of the diseases and for prevention strategies. The only data available to estimate the penetrance function are families selected through affected individuals. If the ascertainment process is not taken into account in the estimation, the penetrance function is likely to be biased. Different adjustments for ascertainment have been proposed to provide valid risk estimates of a genetic disease [Carayol and Bonaiti-Pellie 2004; Le Bihan, et al. 1995].

Selection schemes usually depend on the disease characteristics. In this paper, we focus on samples of family selected through at least one affected individual (i.e. unselected for family history). For genetic diseases in which all affected individuals are carriers of the predisposing mutations, this selection is sufficient to provide informative data on the penetrance function. But in common diseases in which only a minority of cases is due to the rare mutation (referred to as monogenic sub-entities), an age criterion has to be introduced to increase the probability that the cases sampled are mutation carriers [Bonadona, et al. 2005; Dunlop, et al. 1997].

When families are selected through at least one affected, two methods taking into account the ascertainment bias have been proposed to estimate the penetrance function: the Proband's phenotype Exclusion Likelihood (or PEL) [alarcon, et al. 2008] and the Prospective likelihood [Kraft and Thomas 2000; Le Bihan, et al. 1995; Plante-Bordeneuve, et al. 2003]. Both are maximum likelihood methods implementing survival analysis. The Prospective likelihood corrects for the ascertainment with an analytical expression of the ascertainment probability

while PEL is a more intuitive method that corrects for the ascertainment by simply removing the individual (the proband) who allowed his family to be selected. It has been shown for various genetic models and selection schemes that PEL is practically unbiased while the Prospective method is biased in several situations, see [alarcon, et al. 2008]. However, the penetrance function implemented in the method is modeled with a Weibull distribution. Although this model is widely used in survival analysis because of its capacity to adjust to observed data, the assumption of a Weibull distribution for the penetrance can be a tricky limitation in some applications. A strategy to relax the constraints of the Weibull model is to extend it by adding new parameters. The model is then more general and can fit more situations. But, the complexity of the estimation procedure increases dramatically with the number of parameters ("the curse of dimensionality") and may turn to be intractable.

In this paper, we propose a non parametric method for penetrance function estimation, correcting for the ascertainment bias: the Index Discarding EuclideAn Likelihood (IDEAL). The method is applicable for all selection criteria and disease models with at least one affected. Instead of building a likelihood based on the Weibull model, we use a non parametric likelihood that does not assume any parametric family for the distribution. We use the Euclidean Likelihood, a fast version of Owen's Empirical Likelihood [Owen 2001]. To the best of our knowledge, this paper is the first attempt to estimate a penetrance function by means of this approach.

The paper is organized as follows: we first introduce the two estimation methods, PEL and IDEAL. Then, simulations corresponding to real situations are presented under various risk models and various selection patterns. Both methods are applied on data simulated from a Weibull distribution as well as from other distributions (Uniform and Cauchy).

Methods

This section introduces the two estimations methods. First, PEL is briefly presented and then IDEAL is precisely defined. Finally, the simulation processes are explained and the different selection schemes are described.

The Proband's phenotype Exclusion Likelihood

PEL [alarcon, et al. 2008] is an estimation method based on Maximum Likelihood (ML) using a survival analysis approach and correcting for ascertainment bias when families are selected through at least one affected individual. It estimates the penetrance function by using the phenotypic information from family members, genotyped or not, conditionally on observed genotypes. For an individual i of family f, the phenotype is denoted $P_{i,f}$ and the genotype $G_{i,f}$. $P_{i,f} = 1$ (respectively $G_{i,f} = 1$) if i is affected (resp. carrier) and $P_{i,f} = 0$ (resp. $G_{i,f} = 0$) if i is not affected (resp. not carrier). The penetrance function $F(t)$ of a carrier i at age t is modeled using an extended Weibull function [Plante-Bordeneuve, et al. 2003] and is therefore given by:

$$F(t) = (1 - \kappa)[1 - \exp(-\lambda(t - \delta)^{\alpha})],$$

where κ, λ and α are the parameters of the model estimated by ML using the maximization procedure implemented in the program GEMINI [Lalouel 1979]. To avoid an over-parametrization, the parameter δ is not estimated but fixed on the basis of previous knowledge on the age distribution of the disease. The parameters κ and δ extend the classical Weibull model given by the simpler form $F(t) = 1 - \exp(-\lambda t^{\alpha})$ (κ is the fraction of individuals that would never be affected and δ is the age before which the probability of being affected is equal to zero).

The principle of PEL, based on the Weinberg Proband Method in segregation analysis [Weinberg 1912], is to correct for ascertainment by ignoring the proband's phenotype and by duplicating families that contain several probands. Briefly, PEL can be written as follows:

$$PEL(\kappa, \lambda, \alpha) = \prod_f PEL_f = \prod_f \mathbb{P}(P_f^* | G_{f,obs})$$

where $\mathbb{P}(X)$ denotes the probability of X under the Weibull model, P_f^* is the phenotypic vector of the family f in which the phenotype of the proband is set as unknown and $G_{f,obs}$ is the vector of the observed genotypes for the family f. When there is more than one proband in the family, the family is duplicated as many times as there are probands and the phenotype of each proband, referred to as the index, is set as unknown alternately.

IDEAL

In this paper, we propose to consider a non parametric approach based on Empirical likelihood to estimate the penetrance function, the Index Discarding EuclideAn Likelihood (IDEAL). Like PEL, IDEAL corrects for the ascertainment by using the discarding method described by Weinberg [Crow 1965; Weinber 1912]. In addition, IDEAL provides confidence bands for the penetrance function F, i.e. two functions that bind the penetrance a each age t with a given probability.

In this subsection, we first present Empirical Likelihood and its Euclidean version for a cumulative function. Then, we show how to apply this method to the estimation of the penetrance function and we present the modifications introduced to correct for ascertainment. Finally, we describe the construction of confidence bands for the penetrance function.

Empirical likelihood

We present here the Empirical Likelihood method for the estimation of a cumulative distribution function. A more complete exposition of this method and its numerous applications can be found in Owen's book [Owen 2001]. This method has been designed to avoid the choice of a model for the distribution. It can be applied as soon as the true value θ_0 of the parameter of interest is defined as the solution of an estimating equation: for some random variable X and some function m, $\mathbb{E}[m(X, \theta_0)] = 0$, where $\mathbb{E}[.]$ stands for the expectation. This means that, according to the observations $X_1, \ldots, X_n$, in order to estimate θ_0, one looks for a value of θ such as the sample of the $m(X_i, \theta)$'s has zero-mean:

$$\frac{1}{n}\sum_{i=1}^{n} m(X_i, \theta) = 0.$$

The problem of the estimation of the cumulative distribution function F can be formulated in this context as follows: for any $t > 0$,

$$\mathbb{E}[m(X, \theta_0)] = \mathbb{E}[\mathbb{1}_{A \leq t} - F(t)] = 0,$$

with correspondence $\theta_0 = F(t)$, $X = (A, G, P, Pb)$, $m(x, y) = \mathbb{1}_{x \leq t} - y$ and where $\mathbb{1}_{x \leq t}$ is the indicator function of the event $x \leq t$. In the following, θ_0 is $F(t)$ and θ stands for a potential value of θ_0. X resumes the information of an individual and contains the age at onset of the disease A, the genotype G, the phenotype P and the fact that the individual is either a proband or not Pb ($Pb = 1$ if the individual is a proband and 0 else).

Empirical Likelihood is built by means of the multinomial distributions on the sample

$X_1, \ldots, X_n$:

$$\mathbb{Q}(x) = \begin{cases} q_i & \text{if } \exists i,\ x = X_i, \\ 0 & \text{otherwise} \end{cases}$$

with $0 < q_i < 1$ and $\sum q_i = 1$.

This lead to Empirical Likelihood (Owen, 2001):

$$\begin{aligned} EL(\theta, t) &= \sup_{\mathbb{Q}} \left\{ \prod_{i=1}^{n} \mathbb{Q}(X_i) \middle| \mathbb{E}_{\mathbb{Q}}[\mathbb{1}_{A \leq t} - \theta] = 0 \right\} \\ &= \sup_{(q_1, \ldots, q_n)} \left\{ \prod_{i=1}^{n} q_i \middle| \sum_{i=1}^{n} q_i (\mathbb{1}_{A_i \leq t} - \theta) = 0,\ \sum_{i=1}^{n} q_i = 1 \right\}. \end{aligned}$$

The estimator is given by $\hat{\theta} = \text{argmax}_\theta \{EL(\theta, t)\}$ and is an asymptotically normal estimator of θ_0 whatever the distribution of the data. This is the main property of this non parametric method: it is not necessary to suppose that the distribution belongs to a given parametric family (not even the multinomial family).

It is interesting to note that the Kullback discrepancy K appears in the expression of the log-likelihood ratio corresponding to EL:

$$\begin{aligned} -2 \log \left(\frac{EL(\theta, t)}{EL(\hat{\theta}, t)} \right) &= -2 \log \left(\frac{EL(\theta, t)}{\sup_\theta \{EL(\theta, t)\}} \right) \\ &= -2 \log \left(\frac{\sup_{\mathbb{Q}} \{ \prod_{i=1}^{n} \mathbb{Q}(X_i) | \, \mathbb{E}_{\mathbb{Q}}[\mathbb{1}_{A_i \leq t} - \theta] = 0 \}}{\sup_{\theta,\, \mathbb{Q}} \{ \prod_{i=1}^{n} \mathbb{Q}(X_i) | \, \mathbb{E}_{\mathbb{Q}}[\mathbb{1}_{A_i \leq t} - \theta] = 0 \}} \right) \\ &= 2n \inf_{\mathbb{Q}} \{ K(\mathbb{Q}, \mathbb{P}_n) | \, \mathbb{E}_{\mathbb{Q}}[\mathbb{1}_{A \leq t} - \theta] = 0 \}. \end{aligned}$$

where $K(\mathbb{Q}, \mathbb{P}_n) = -\int \log \left(\frac{d\mathbb{Q}}{d\mathbb{P}_n} \right) d\mathbb{P}_n$, and $\mathbb{P}_n$ is the multinomial maximizing the likelihood:

$$\mathbb{P}_n(x) = \begin{cases} \dfrac{1}{n} & \text{if } \exists i,\ x = X_i, \\ 0 & \text{otherwise.} \end{cases}$$

Thus, the Empirical Likelihood method consists in minimizing the Kullback discrepancy between $\mathbb{Q}$ and $\mathbb{P}_n$. Nevertheless, other choices of discrepancy can be used: the Hellinger distance, the Relative Entropy and the Euclidean distance are the more common, but the method can be generalized way beyond (see [Bertail, et al. 2004]). We propose here to use the Euclidean distance (denoted χ^2) instead of the Kullback discrepancy in the expression of the log-likelihood ratio because it leads to a closed form for the likelihood that strongly reduces the computational time.

Euclidean Likelihood

The statistic corresponding to the Euclidean distance, that we refer to as the EAL, is then:

$$\begin{aligned} EAL(\theta, t) &= 2n \inf_{\mathbb{Q}} \left\{ \chi^2(\mathbb{Q}, \mathbb{P}_n) \middle| \, \mathbb{E}_{\mathbb{Q}}[\mathbb{1}_{A \leq t} - \theta] = 0 \right\} \\ &= 2n \inf_{\mathbb{Q}} \left\{ \int \left(\frac{d\mathbb{Q}}{d\mathbb{P}_n} - 1 \right)^2 d\mathbb{P}_n \middle| \, \mathbb{E}_{\mathbb{Q}}[\mathbb{1}_{A \leq t} - \theta] = 0 \right\}. \end{aligned}$$

As for $EL(\theta, t)$, maximizing $EAL(\theta, t)$ in θ gives an asymptotically normal estimator of θ_0.

Estimation of penetrance

In the context of penetrance estimation, specific modifications of the Euclidian likelihood in the reference probability measure $\mathbb{P}_n$ are necessary. Hereafter, $\mathbb{W}$ stands for the modified versions of the reference measure.

In order to estimate the value of the penetrance at an age t, one should consider only the population of carriers aged of t or more years. Therefore, at any fixed t, the individual i is considered only if $G_i = 1$ (i.e. if i is affected) and if the current age Y_i of i is bigger than t.

The technical effect of this remark is that the reference measure varies as a function of t:

$$\mathbb{W}(x) = \begin{cases} \dfrac{1}{n} & \text{if } \exists i,\ x = X_i,\ G_i = 1 \text{ and } Y_i \geq t, \\ 0 & \text{otherwise.} \end{cases}$$

For unaffected individuals i, the age at onset A_i does not exists. From a mathematical point of view, the available information is that A_i is bigger than the current age Y_i. Therefore, it is technically convenient to set $A_i = +\infty$. For large values of t, the proportion of such A_i corresponds to the κ of the extended Weibull model [alarcon, et al. 2008], i.e. the proportion of individuals that will never be affected.

Ascertainment correction

Because of the ascertainment scheme, the sample is currently biased with an excess of affected individuals and $\theta_0 = F(t)$ does not verify the estimating equation: $\mathbb{E}_{\mathbb{P}_0}[\mathbb{1}_{A \leq t} - \theta_0] \neq 0$, where $\mathbb{P}_0$ is the distribution generating the observed (biased) data. We propose a method related to Weinberg's that consists in correcting for the ascertainment bias by underweighting the probands: if a family contains k potential probands, they should be weighted by $1 - \frac{1}{k}$, see Appendix . Under this modified distribution the estimating equation $\mathbb{E}[\mathbb{1}_{A \leq t} - \theta_0] = 0$ can be used and leads to a non biased estimate of θ_0. Therefore, we apply this modification to our reference measure $\mathbb{W}$:

$$\mathbb{W}(x) = \begin{cases} \dfrac{1}{n} & \text{if } \exists i,\ x = X_i,\ Pb_i = 0,\ G_i = 1 \text{ and } Y_i \geq t, \\ \dfrac{1}{n}\left(1 - \dfrac{1}{k}\right) & \text{if } \exists i,\ x = X_i,\ Pb_i = 1,\ G_i = 1 \text{ and } Y_i \geq t, \\ 0 & \text{otherwise.} \end{cases}$$

Confidence bands

A very strong property of Empirical Likelihood and related method is that it provides confidence bands for the cumulative distribution function (CDF), see [Owen 2001], chapter 7. This means that for any given level, for example 95%, we can give two CDF G et H such as with probability 95% and for all $t > 0$,

$$G(t) \leq F(t) \leq H(t).$$

This is stronger than a sequence of confidence intervals (CIs) given t by t: with probability 95% the function F remains between G and H for all t. The sequence of CIs is local (given t by t), whereas the confidence band is global (valid for all t):

$$\text{sequence of CIs} \quad \forall t > 0,\ \mathbb{P}\left(G(t) \leq F(t) \leq H(t)\right) = 95\%,$$

$$\text{confidence band} \quad \mathbb{P}\left(\forall t > 0,\ G(t) \leq F(t) \leq H(t)\right) = 95\%.$$

An additional enjoyable property is that the confidence band is not asymptotic: it is actually reached at the current value of the sample size n. G and H are defined as follows:

$$G(t) = \min\left\{\theta | \text{IDEAL}(\theta, t) \leq c_n\right\}$$

$$H(t) = \max\left\{\theta | \text{IDEAL}(\theta, t) \leq c_n\right\}$$

where critical values of c_n are tabulated, see [Owen 2001], page 159. For a confidence level of 95% and $n \leq 1000$, c_n writes:

$$n \leq 100 \quad c_n = 3.0123 + 0.4835 \log(n) - 0.00957 \log(n)^2 - 0.001488 \log(n)^3,$$

$$100 < n \leq 1000 \quad c_n = 3.0806 + 0.4894 \log(n) - 0.02086 \log(n)^2.$$

Simulation

IDEAL was compared to PEL by simulating family samples under various situations. We chose a simulation process in which the family size and structure are fixed (see Figure 1). To ensure asymptotic conditions, sample size was fixed to 5 000 families and therefore to 90 000 individuals.

A genotype was randomly assigned to the pedigree founders with a frequency of 10% for the mutated allele. This value was chosen in order to limit the computational time. For the other family members, genotypes are randomly assigned using Mendel's laws. The frequency of *de novo* mutation was set to 0 and we restricted to the case where all genotypes are known. We used French demographic data to simulate the ages of the individuals. For non carriers, we considered either a risk of 0% for monogenic diseases (MD) or a cumulative risk of 10% at 80 years for complex diseases with monogenic sub-entities (CDMS) where only a fraction of cases are due to a mutation.

First, to compare IDEAL with PEL under a Weibull model, phenotypes were simulated with an age-dependent function, based on the Weibull model corresponding to a cumulative risk of 50% by age 80 for carriers. Secondly, to enlighten the difference between the parametric method (PEL) and our non parametric method (IDEAL), phenotypes were simulated with an age-dependent function not based on the Weibull model but respectively on an Uniform distribution and on a Cauchy distribution. These distributions have been chosen for their substantial difference with the Weibull model.

As in [alarcon, et al. 2008], in order to model a realistic selection process, we defined a period length T to select the probands : only individuals affected during the last T years could

be selected, and this with a probability p_s. We considered two different period lengths : a period of 20 years ($T = 20$) and a period of 1 year ($T = 1$). The probability p_s was introduced to simulate the fact that, in real situations, some individuals affected during the study are not detected and therefore do not become probands. A family was included in the sample as soon as one of its member was a proband. Under the CDMS model, we introduced an age criterion for selection (35 years) to increase the probability of detecting families with mutation carriers [Claus, et al. 1990].

Results

We studied the behavior of IDEAL in two extreme situations. First, we considered the case of a low ascertainment probability for affected by simulating a CDMS model with an age criterion of 35 years for the selection as described above, a probability $p_s = 0.5$ and a period $T = 1$ in the selection process. Then, we considered the case of a high ascertainment probability for affected by simulating a MD model with a probability $p_s = 1$ and a period $T = 20$ in the selection process. To ensure asymptotic conditions, sample size were fixed to 5 000 families after selection and we considered the case where all genotypes are known. It has already been shown in [alarcon, et al. 2008] that PEL is practically unbiased in these two situations when the penetrance function belongs to the Weibull family. Then, we compared IDEAL with PEL when the Weibull assumption fails. Finally, we study the robustness of the two methods to sample size by analyzing a Weibull distributed sample of 200 families after selection.

Behavior of IDEAL under a Weibull model

Figure 2 shows estimation by IDEAL in the case of a low ascertainment probability. IDEAL is unbiased, the true penetrance (curve with stars) and the estimated penetrance by IDEAL (dotted curve) are superposed. The plain curves represent the confidence band.

The case of high ascertainment probability is shown in Figure 3. As in Figure 2, IDEAL is unbiased and the estimated penetrance is indistinguishable from the true penetrance. The PEL estimator being also superposed with the true penetrance, it is not represented on Figures 2 and 3.

For both figures, the confidence bands are quite thin and contain the true penetrance at all t, as expected.

Comparison of IDEAL and PEL under Uniform and Cauchy models

Results are only presented in the case of a CDMS model, with an age criterion of 35 for the selection, with $p_s = 0.5$ and $T = 20$ for the selection. All other cases we considered in preliminary investigations lead to similar results.

Figure 4 shows that PEL does not fit the curve while IDEAL is perfectly unbiased (the estimate with IDEAL is indistinguishable from the true penetrance). Figure 5 shows estimations with the same previous parameters for the selection, when penetrance is simulated under a Cauchy distribution (with parameter 5). We can see again that IDEAL is perfectly unbiased while PEL has a non negligible bias.

Sensitivity of the two methods to the sample size

We compare in this paragraph the two methods when the penetrance function belongs to the Weibull family, for a realistic sample size of 200 families. We only report here the case of a MD model with probability $p_s = 1$ and with period $T = 20$ in the selection process, but the CDMS model leads to the same results. Figure 6 shows that PEL is less biased than IDEAL when asymptotic conditions failed. IDEAL remains unbiased for low ages but the method is biased for high ages. Moreover, PEL's confidence intervals are smaller than IDEAL's confidence bands.

Discussion

In this paper, we have proposed an estimation method (IDEAL) that adapts to the penetrance function model and that corrects for the ascertainment bias when families are ascertained through at least one affected. We have compared this method with a parametric method (PEL) also designed to take into account the ascertainment bias. First, we have shown that IDEAL corrects for the ascertainment bias and that it leads to unbiased estimates. Then, we have shown through simulations on large samples, that IDEAL performs as well as PEL when the true penetrance is Weibull distributed and significantly better when this assumption fails. This adaptability of IDEAL allows to estimate penetrance functions in new contexts without risking a bias due to a model misspecification.

In the simulation part, we have only reported results for the theoretical situations where all genotypes are known. Unknown genotypes can easily be taken into account in IDEAL method by means of weighting. For example, we can estimate a probability p_i of mutation for each

individual from the known genotypes of the family and use $\frac{p_i}{n}$ as a reference instead of $\frac{1}{n}$.

An other important situation considered in this paper is the behavior for a small number of families. In this case, the parametric framework (i.e. the extended Weibull model) is useful to complete the lack of information by forcing the shape of the distribution and PEL provides better results than IDEAL particularly for high ages. But this means that the assumption that the penetrance belong to a given model is then overriding, when the model holds. The performances of the estimators are then even more dependent of the validity of the model.

A additional feature of IDEAL is that it gives confidence bands directly on the penetrance function $F(t)$, instead of a confidence interval for a parameter derivated from a model. Simulation results show that the width of this confidence band increases with t. This can be explained by the fact that only individuals of age larger than t are considered to estimate the penetrance at t. The population considered is therefore decreasing with t. Thus, the confidence band informs on the precision of the estimator of the penetrance in function of t.

As a last point, it can be remarked that in the literature, the penetrance function is usually modeled using a parametric survival analysis approach. But in practice, the real distribution is not known and, in the best of our knowledge, the used models have never been validated. Thus, it would be interesting to use IDEAL as a validation method for parametric models; First estimate the penetrance curve both with IDEAL and with a parametric method and then confront the two estimators. If the difference is too important, the parametric model can be questioned. Moreover, when considering a new disease, information on the penetrance structure is unlikely available. The most natural approach is then to consider directly a non parametric method like IDEAL. The resulting estimations can be used to motivate the use of a parametric family like the Weibull.

Acknowledgments

We would particularly like to acknowledge the comments and corrections of Catherine Bourgain.

We also acknowledge the reviewers for their comments.

References

[1] Alarcon F, Bourgain C, M. Gautier-Villars, Planté-Bordeneuve V, Stoppa-Lyonnet D, Bonaïti-Pellié C. 2008.PEL: An unbiased method for estimating genetic disease risk from pedigree data unselected for family history. *(submited)*

[2] Bertail P, Harari-Kermadec H, Ravaille D. 2004. φ-Divergence empirique et vraisemblance empirique généralisée. To appear in "Annales d'Economie et de Statistique".

[3] Bonadona V, Sinilnikova OM, Chopin S, Antoniou AC, Mignotte H, Mathevet P, Bremond A, Martin A, Bobin JY, Romestaing P and others. 2005. Contribution of BRCA1 and BRCA2 germ-line mutations to the incidence of breast cancer in youg women : results from a prospective population-based study in France. Genes Chromosomes Cancer 43(4):404-13.

[4] Carayol J and Bonaïti-Pellié C. 2004. Estimating penetrance from family data using a retrospective Likelihood when ascertainment depends on genotype and age of onset. Gen Epidemiol 27(2):109-17.

[5] Claus EB, Risch NJ, Thompson WD. 1990. Age of onset as an indicator of familial risk of breast cancer. Am J Epidemiol 131:961-72.

[6] Crow J. 1965. Problems of ascertainment in the analysis of family data. In:(Neel J.V, Shaw M.W, Schull W.J). Genetics and the epidemiology of chronic disease. Public Health "'source"' Publication Washington D.C.

[7] Dunlop MG, Farrington SM, Carothers AD, Wyllie AH, Sharp L, Burn J, Liu B, Kinzler KW, Vogelstein B. 1997. Cancer risk associated with germline DNA mismatch repair gene mutations. Hum Mol Genet 6(1):105-10.

[8] Kraft P and Thomas D. 2000. Bias and efficiency in family-based gene-characterization studies : conditional, prospective, retrospective, and joint likelihoods. Am J Hum Genet 66(3):1119-31.

[9] Lalouel JM. 1979. GEMINI: a computer program for optimization of general non linear functionc. Technical report no 14. Salt Lake City : university of Utah, Department of Medical Biophysics and Computing.

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

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

[12] Owen AB. 2001. Empirical Likelihood. Chapman and Hall/CRC, Boca Raton.

[13] Weinberg. 1912. Methode und Fehlerquellen der Untersuchung auf Mendleschen Zahlen beim Menschen. Arch. Rass.u.Ges.Biol. 9:165-174.

Appendix: Index discarding

Weinberg proposes in [Weinberg 1912] a method to correct for the ascertainment based on discarding the probands: for a family with k probands, the family is replicated k times and each time a different proband is discarded. Crow has shown the validity of this method in [Crow 1965]. This procedure as been designed to estimate the segregation ratio and can straightforwardly transpose in our context of penetrance estimation. The important result is that θ, the penetrance at time t, is given by the ratio of the statistical mean of the number of affected individuals in a ascertained and replicated family by the statistical mean of the number of carriers individuals in a ascertained and replicated family. In the following P_i and G_i are respectively the phenotype and the genotype of the individual i.

$$\theta = \frac{\mathbb{E}\left[\sum_{i=1}^{r} k\mathbb{1}_{P_i=1}\mathbb{1}_{Pb_i=0} + \sum_{i=1}^{r}(k-1)\mathbb{1}_{P_i=1}\mathbb{1}_{Pb_i=1}\right]}{\mathbb{E}\left[\sum_{i=1}^{r} k\mathbb{1}_{G_i=1}\mathbb{1}_{Pb_i=0} + \sum_{i=1}^{r}(k-1)\mathbb{1}_{G_i=1}\mathbb{1}_{Pb_i=1}\right]}$$

where r is the length of the family, k the number of probands and $Pb_i = 1$ if i is a proband. This can be rewritten:

$$\mathbb{E}\left[\sum_{i=1}^{r}\left(\mathbb{1}_{P_i=1} - \theta\mathbb{1}_{G_i=1}\right)\left(k\mathbb{1}_{Pb_i=0} + (k-1)\mathbb{1}_{Pb_i=1}\right)\right] = 0.$$

Dividing by k, we get:

$$\mathbb{E}\left[\sum_{i=1}^{r}\left(\mathbb{1}_{P_i=1} - \theta\mathbb{1}_{G_i=1}\right)\left(\mathbb{1}_{Pb_i=0} + \left(1-\frac{1}{k}\right)\mathbb{1}_{Pb_i=1}\right)\right] = 0.$$

Therefore, if we set $\mathbb{W}_0$ as follows:

$$\mathbb{W}_0(x) = \begin{cases} 1 & \text{if } i \text{ is not a proband and }, x = X_i, \\ 1 - \dfrac{1}{k} & \text{if } i \text{ is a proband,} \\ 0 & \text{otherwise.} \end{cases}$$

then θ is given as the solution of

$$\mathbb{E}_{\mathbb{W}_0}\left[\sum_{i=1}^{r} (\mathbb{1}_{P_i=1} - \theta \mathbb{1}_{G_i=1})\right] = 0.$$

$\mathbb{W}_0$ has mass $r-1$ and must be normalized to be a probability measure. Our reference measure $\mathbb{W}$,

$$\mathbb{W}(x) = \begin{cases} \dfrac{1}{n} & \text{if } \exists i,\ x = X_i,\ Pb_i = 0,\ G_i = 1 \text{ and } Y_i \geq t \\ \dfrac{1}{n}\left(1 - \dfrac{1}{k}\right) & \text{if } \exists i,\ x = X_i,\ Pb_i = 1,\ G_i = 1 \text{ and } Y_i \geq t, \\ 0 & \text{otherwise.} \end{cases}$$

converges (once normalized) to the normalized version of $\mathbb{W}_0$. Therefore, the estimate $\hat{\theta}$ given as the solution of $\mathbb{E}_{\mathbb{W}}\left[\sum_{i=1}^{r} (\mathbb{1}_{P_i=1} - \theta \mathbb{1}_{G_i=1})\right] = 0$ converges to the parameter of interest θ. This motivates the ascertainment correction used in IDEAL.

Now we show that the estimating equation can be rewritten as in our statement. First, under $\mathbb{W}$, G_i is constant and equals to 1 and can therefore be omitted. Secondly, for a fixed t, the fact that the individual i as contracted the disease, i.e. $P_i = 1$, is equivalent that it occurred before t, i.e. $A_i \leq t$. Therefore:

$$\mathbb{E}_{\mathbb{W}}\left[\sum_{i=1}^{r} (\mathbb{1}_{P_i=1} - \theta \mathbb{1}_{G_i=1})\right] = \mathbb{E}_{\mathbb{W}}\left[\sum_{i=1}^{r} (\mathbb{1}_{A_i \leq t} - \theta)\right].$$

Figure 1: Family structure.

Figure 2: Simulation in case of a low ascertainment probability

Figure 3: Simulation in case of a high ascertainment probability

Figure 4: Comparison of PEL and IDEAL under a Uniform distribution, in case of a CDMS model with an age criterion of 35 for the selection, $p_s = 0.5$ and $T = 20$ for the selection.

Figure 5: Comparison of PEL and IDEAL under a Cauchy distribution with parameter 5, in case of a CDMS model with an age criterion of 35 for the selection, $p_s = 0.5$ and $T = 20$ for the selection.

Figure 6: Simulation in case of a MD model