PEL: An unbiased method for estimating age dependent genetic disease risk from pedigree data unselected for family history

F. Alarcon [1,2], C. Bourgain[2,1], M. Gauthier-Villars[3], V. Planté-Bordeneuve[4], D. Stoppa-Lyonnet[3,5], C. Bonaïti-Pellié[2,1]

[1] Univ Paris-Sud , IFR69, UMR-S535, Villejuif, France
[2] INSERM, U535, Villejuif, France
[3] Genetic Oncology, Institut curie, Paris, France
[4] Department of Neurology, CHU Bicêtre, Le Kremlin Bicêtre, France
[5] Univ Paris-Descartes, Paris, France

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

ABSTRACT

Providing valid risk estimates of a genetic disease with variable age of onset is a major challenge for prevention strategies. When data are obtained from pedigrees ascertained through affected individuals, an adjustment for ascertainment bias is necessary. This paper focuses on ascertainment through at least one affected and presents an estimation method based on maximum likelihood, called the Proband's phenotype Exclusion Likelihood or PEL for estimating age-dependent penetrance using disease status and genotypic information of family members in pedigrees unselected for family history. We studied the properties of the PEL and compared with another method, the Prospective likelihood, in terms of bias and efficiency in risk estimate. For that purpose, family samples were simulated under various disease risk models and under various ascertainment patterns. We showed that, whatever the genetic model and the ascertainment scheme, the PEL provided unbiased estimates, whereas the Prospective likelihood exhibited some bias in a number of situations. As an illustration, we estimated the disease risk for transthyretin amyloid neuropathy from a French sample and a Portuguese sample and for *BRCA1/2* associated breast cancer from a sample ascertained on early-onset breast cancer cases.

Key Words: Ascertainment bias, risk estimation, penetrance function, maximum likelihood method.

INTRODUCTION

Some diseases with variable age of onset are due to the presence of predisposing gene mutations. Precise estimation of the age-specific cumulative risk (called penetrance function) for mutation carriers is very important for defining prevention strategies and understanding underlying mechanisms of the diseases.

Before the identification of genes responsible for diseases, penetrance had been estimated using the segregation of the disease in families. Because families were ascertained through affected individuals (probands), segregation analysis methods included a correction for ascertainment [Cannings and Thompson, 1977; Morton, 1959; Weinberg, 1912]. Such methods could also account for variable age of onset by using a survival analysis approach [Abel and Bonney, 1990] .

When the gene(s) responsible for a given disease has(ve) been identified, one might expect that the knowledge of genotypes of family members would allow a more precise estimation of penetrance, but that a correction for ascertainment would still be necessary [Carayol, et al. 2002]. When families have been ascertained regardless of family history, i.e. on the presence of at least one affected individual, the correction may be obtained by using a prospective likelihood that corrects by conditioning on the ascertainment event [Kraft and Thomas 2000; Le Bihan, et al. 1995; Plante-Bordeneuve, et al. 2003] and includes a survival analysis approach to account for variable age of onset.

Other methods have been proposed for estimating penetrance, in a different context. The kin-cohort design [Wacholder, et al. 1998] has been proposed to estimate penetrance for carriers of specific mutations of the *BRCA1* and *BRCA2* genes by comparing the proportions of affected relatives between mutation carriers and non carriers in a population in which the specific mutations are frequent. This design is more generally referred to as the genotype-proband design (GPD) by Gail et al. [1999] to emphasize that the proband only was genotyped and extended to the so-called GPDR where one or

two relatives are genotyped in each nuclear family. These population-based methods require very large samples of affected and unaffected individuals and do not need any correction for ascertainment. The Genotype Restricted likelihood (GRL) has been proposed for analysing pedigrees ascertained through familial criteria and for which a formal correction for ascertainment bias cannot easily be performed.

In this paper, we present a simple and intuitive approach, based on the Weinberg Proband Method in segregation analysis [Weinberg 1912], referred to as the Proband's phenotype Exclusion Likelihood (PEL), for estimating age-dependent penetrance using disease status and genotypic information of family members in pedigrees ascertained through affected individuals but unselected for family history. Using simulations, we studied the properties of the PEL and compared its properties to those of the Prospective likelihood which explicitly models the probability that a family is ascertained. The robustness of the PEL to underlying hypotheses was studied in various situations likely to be encountered in the analysis of family data.

Finally, both methods were applied to a monogenic disease and a complex disease with monogenic sub-entities: a sample of 27 French and 33 Portuguese families diagnosed with Transthyretin (*TTR*) amyloid neuropathy and a sample of 30 families with *BRCA1* or *BRCA2* associated breast and ovarian cancer ascertained on early-onset breast cancer cases.

METHODS

THE PROBAND'S PHENOTYPE EXCLUSION LIKELIHOOD (PEL)

The PEL is a Maximum Likelihood (ML) method that corrects for ascertainment bias when families, in which a deleterious mutation has been found, have been ascertained through at least one affected individual, i.e. regardless of family history. The penetrance function is estimated using the phenotypic

information, conditioned on genotype, from all family members, including those with unknown genotype. Correction for ascertainment is performed by removing the phenotypic information of the individual who allowed the family to be detected (the proband) and by duplicating families when there are several probands in a family.

This principle was introduced by Weinberg [1912] for estimating the segregation ratio in the offspring of two heterozygous parents under recessive inheritance. The argument for discarding the proband is a simple, intuitive one: each ascertained affected individual (proband) is regarded as providing the information that his(her) parents are capable of producing affected children; then the remaining members of the sibship provides an unbiased estimate of the ratio of affected to normal individuals. In case of single ascertainment (only one proband per family), the method has been shown by Fisher [1934] to be fully efficient for sibships, and Crow [1965] showed that, in case of multiple ascertainment, the method yielded a consistent estimate of the segregation frequency provided that sibships were replicated as many times as there were probands.

We consider the case of a dominant disease where almost all carrier individuals are heterozygotes for the deleterious allele. Let us denote $Phen$ the vector containing the n_f individuals' phenotypes and Gen_{obs} the vector of the observed genotypes for the family f :

$Phen = (Phen_1, ..., Phen_{n_f})$ with $Phen_i = 1$ if i is affected and $Phen_i = 0$ if i is unaffected, $Gen_i = 0$ if i is not a carrier and $Gen_i = 1$ otherwise.

To include non genotyped individuals in the likelihood, we denote Ω the number of all possible genotypic configurations (which is a function of the number of unknown genotypes) and Gen_ω, the vector of observed and unobserved genotypes corresponding to configuration ω.

The PEL uses the probability of the phenotypes of the family members other than the proband, denoted $Phen^{*}$, computed conditionally on all observed genotypes. For family *f*, the likelihood L_f is

$$L_f = \mathrm{P}(Phen^{*}/Gen_{obs})$$

Let P_ω be the probability of the genotypic configuration ω, i.e. the joint probability of genotypes of the family f in configuration ω:

$$P_\omega = \mathrm{P}(Gen_{1,\omega};\ldots;Gen_{nf,\omega}) = \prod_j \mathrm{P}(Gen_{j,\omega}) \prod_{\{l,m,n\}} \mathrm{P}(Gen_{l,\omega}/Gen_{m,\omega}, Gen_{n,\omega})$$

where the product on j is taken over all founders of the family and the product on $\{l, m, n\}$ is taken over all parent-offspring triplets.

P_ω is a function of both the frequency of the mutated allele (denoted fq) in the general population using Hardy-Weinberg proportions in the founders (parents' status unknown) and Mendelian transmission rates in the triplets, and the *de novo* mutation rate (denoted pn).

Practically, L_f is computed using the algorithm of Elston and Stewart [1971] and the conditioning on observed genotypes is obtained by restricting the summation in the likelihood to the set Ω of genotypes compatible with the observed ones. Under the assumption that phenotypes of relatives are independently distributed conditionally to their genotype:

$$L_f = \sum_{\omega=1}^{\Omega} P_\omega . \mathrm{P}(Phen^{*}/Gen_\omega)$$

where $\mathrm{P}(Phen^{*}/Gen_\omega)$ is the product over all family members i of the probability of phenotypes $\mathrm{P}(Phen_i/Gen_{i,\omega})$ given his/her genotype in the configuration ω, the proband's phenotype being set as unknown.

Finally, as the N families of the sample are assumed to be independent, the total likelihood may be written as:

$$L = \prod_{f=1}^{N} L_f$$

PENETRANCE FUNCTION AND CONTRIBUTION OF FAMILY MEMBERS TO THE LIKELIHOOD

Let us denote $F(t_i)$, the penetrance function for a carrier i at age t_i.

If the individual i is still unaffected at age t_i , his contribution to the likelihood at age t_i is:

$$\mathrm{P}(Phen_i / Gen_i) = 1 - F(t_i)$$

If i is affected at an age of onset included between t_i and $(t_i + 1)$, his contribution to the likelihood is:

$$\mathrm{P}(Phen_i / Gen_i) = F(t_i + 1) - F(t_i)$$

For the age-dependent penetrance function $F(t_i)$, we chose the Weibull model with parameters λ (scale parameter) and α (shape parameter) for the parametric function. The Weibull model is widely used in parametric risk estimation because of its flexibility to adjust to observed data.

We introduced two additional parameters into the model in order to improve its capacity of adjustment to the data. The possibility that some carriers will never develop the disease was accounted for by a parameter κ, the fraction of individuals that would never be affected. We also introduced a parameter δ, which sets an age before which the probability of being affected is equal to zero. In order to avoid an overparametrization of the model, δ was not estimated, but fixed on the basis of previous knowledge on the age distribution of the disease.

Finally, the penetrance function for carriers, using this extended Weibull model, can be written as follows:

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

The penetrance is assumed to be known for non-carriers and taken as the risk in the general population.

In our computations, (κ, λ, α) were estimated by Maximum Likelihood (ML) using the program GEMINI [Lalouel 1979].

PROPERTIES OF THE PEL

The properties of the PEL were assessed by simulating family samples under various disease risk models and ascertainment patterns with at least one affected member.

Genetic models

Two genetic disease models were considered: 1) monogenic diseases (MD) in which all affected individuals are carriers of a predisposing mutation and the presence of at least one affected family member is sufficient to detect the presence of a mutation in the family; 2) complex diseases with monogenic sub-entities (CDMS) in which only a minority of cases is due to rare mutations, such as breast or colorectal cancer, and the detection of genetic cases usually requires familial criteria to increase the probability that the affected individuals are mutation carriers. In such diseases, cases due to rare mutations usually occur at a substantially lower age than sporadic cases, and the inclusion of an age criterion is an alternative to familial criteria to identify families with carrier individuals [Bonadona, et al. 2005; Dunlop, et al. 1997].

Simulation of pedigree samples

The simulated pedigrees had a fixed size and structure: a couple of ancestors with 4 offspring, each with 2 offspring. Ages were simulated to fit French demographic data [Pennec 1996].

A genotype was randomly assigned to the pedigree founders and spouses with a frequency of the mutated allele fq. For the other family members, genotypes were randomly assigned using Mendel's laws. In the simulations, fq was set to 0.01 and pn, the frequency of *de novo* mutation was set to 0. Phenotypes were simulated with an age-dependent function, based on the Weibull model in which α was fixed at 3. For the sake of simplicity, we simulated phenotypes with a Weibull model in which κ and δ were set to null. For carriers, we considered two different values for the parameter λ. The first one corresponding to a cumulative risk of 0.5 by age 80 (called "low true penetrance") and the second one to a cumulative risk of 0.8 by age 80 (called "high true penetrance"). For non carriers, the cumulative risk by age 80 was set to null for monogenic diseases (MD), as and to 0.10 for complex diseases with monogenic sub-entities (CDMS).

Ascertainment process

To model a realistic ascertainment process, we defined time periods (denoted T) for ascertainment of probands. We assumed that only individuals affected during this period might be ascertained with a probability p_s. We considered two different periods of time: a period of 20 years ($T = 20$) in which essentially all affected individuals may be probands, and a period of 1 year ($T = 1$) in which the probability that more than one affected individual be a proband is negligible. The family was included in the sample if there was at least one proband.

Under the CDMS model, we introduced an age criterion for ascertainment to increase the probability of detecting mutation carriers. As 36 years is the criterion used in the breast and ovarian cancer families analysed in this paper, we used the same age criterion in our simulations. Then, in the CDMS model, probands are all carriers affected before 36 years during the period T and ascertained with a probability p_s.

Bias and efficiency

To study the behaviour of the PEL according to the ascertainment scheme and the model considered, we first evaluated, in each case, the average relative bias B estimated by the average on 1000 replicates of 100 pedigrees of the relative bias of risk estimate at age 70 years usually taken as the lifetime risk [Alarcon, et al. 2007; Easton, et al. 1995; Ford, et al. 1998; Gong and Whittemore 2003]:

$B = \frac{1}{1000}\sum_{i=1}^{1000}(\frac{\hat{R}_i - R_0}{R_0})$ where $\hat{R}_i$ is the penetrance estimated at 70 years for the replicate i and R_0 is the true one at the same age.

Various situations may affect the estimation of penetrance by the PEL. On the one hand, departure from the underlying hypotheses may induce a bias. For instance, one assumes that probands are unambiguously identified, which may not be the case and may prevent a correct replication of families. 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. Indeed, as the proportion of unknown genotypes is usually high, the existence of *de novo* mutations may be difficult to assess. Moreover, neither the value of the *de novo* mutation rate pn, nor the deleterious allele frequency fq are usually known in the population. Instead, these parameters are commonly fixed to arbitrary low values. On the other hand, to study the loss in efficiency due to unknown genotypes, and therefore, the interest of genotyping probands' relatives, we evaluated the asymptotic relative efficiencies (AREs) of penetrance estimates, that is, the inverse of the ratio of the variance estimated with various values of the proportion of unknown genotypes to the variance estimated when all genotypes are known. In each situation, the variance by age 70 was evaluated by simulating 1000 replicates of 100 families.

Finally, to analyse the interest of the two parameters κ and δ in the Weibull model for the age-dependant function, we simulated data using the extended Weibull model with various values of κ and δ, and analysed the data without taking one or both of these two parameters into account.

Comparison with the Prospective Likelihood

The Prospective likelihood [Kraft and Thomas, 2000; Plante-Bordeneuve, et al. 2003] is the probability of phenotypes given observed genotypes, conditioned on the ascertainment process that is explicitly modelled, as done in segregation analysis, by the probability that at least one individual is ascertained in the family, a function of the probability π that an individual is ascertained [Morton 1959]. This method implements the same age-dependent penetrance function as in the PEL.
We compared the properties of the PEL with those of the prospective likelihood in terms of bias and efficiency in various situations.

RESULTS

BEHAVIOUR OF THE PEL ACCORDING TO THE ASCERTAINMENT SCHEME

Table I shows the relative bias obtained with the PEL when all genotypes are known, under the two different models considered MD and CDMS with an age criterion of 36 years as described above, and the two time periods considered; we have checked that results were similar when various proportions of unknown genotypes were introduced. Under both models, the PEL provided unbiased or nearly so estimates in all the situations considered. Table II shows the relative efficiencies (AREs) under the MD model obtained when the proportion of unknown genotypes among relatives varies from 50% to 100%, compared to the case where all genotypes are known, in the two extreme situations of ascertainment i.e. high (p_s=1, T=20), and low (p_s=0.5, T=1) ascertainment probability. In all

situations, the loss in efficiency is quite small, whatever the true penetrance value. Under the CDMS model, the effect is quite similar, although slightly more important, with AREs varying from 56 to 69% according to the ascertainment probability and the true penetrance value, in the extreme situation where only the proband's genotype is known. Therefore, genotyping the relatives brings some additional information on the risk estimate, compared to the situation where only their phenotype is known, but the gain in efficiency is expected to be modest.

ROBUSTNESS TO A DEPARTURE FROM UNDERLYING HYPOTHESES

Misspecification of probands

Table III shows the relative bias in the penetrance estimate in the extreme case when only one proband (chosen at random in the simulation process) is identified in each family, so that families are not replicated, even though some of them should be. For the CDMS model, the results are relatively close to those obtained when replicating the families (Table I). For the MD model, when the time period for ascertainment of probands is 20 years (ascertainment of several affected individuals in the same family is more likely), the relative bias can be as high as -13%.

Misspecification of genetic model parameters

We have studied the sensitivity of penetrance estimates to misspecification of pn and fq in case of a MD model, with a high ascertainment probability (p_s=1, T=20) and for a low penetrance . We found that the method was not sensitive to a misspecification of the *de novo* mutation rate pn when it is fixed at different values (10^{-4} or 0) in the analysis whereas its true value is 10^{-5} and in case where 80% of genotypes are unknown. Regarding fq, when this parameter was set at a higher value (0.1) than its true value (0.01), the relative bias generated was almost negligible 3%) when all genotypes are

known while the relative bias was not negligible (-18%) when only the proband's genotypes are known. But this situation is not very realistic. On the other hand, when fq was set at a lower value (10^{-6}) in the analysis, the relative bias was almost negligible, 3% when all genotypes are known and 6% when only the proband's genotypes are known. So, we found that the method was quite robust to a misspecification of the frequency of the mutated allele fq.

Interest of the Extended Weibull model

Figure 1 shows the penetrance estimates when families are simulated with $\kappa = 0.10$, in the high penetrance case, and analyses are performed with Weibull models including a κ parameter or not. Ignoring κ in the model modifies the form of the penetrance curve and may induce an over estimation of the penetrance function. Note that, because κ has an effect on the shape of the curve, this result would have been missed, had we only measured the relative bias in penetrance estimated at 70 years. Moreover, we have simulated families with a δ equal to 25 years and a κ null, and we have verified a large misspecification (δ=10 or 0) only slightly affects the form of the penetrance curve.

Comparison with the prospective likelihood

Table IV presents the results obtained with the Prospective likelihood under the same ascertainment models as for the PEL. The Prospective likelihood provided almost unbiased estimates in case of MD, but exhibited a non negligible relative bias in CDMS, particularly when the true penetrance is low.
In terms of efficiency, we evaluated AREs only under the MD model where the Prospective likelihood provided unbiased estimated. The prospective method provided similar efficiency as the PEL in a wide range of situations, varying from 0.5 to 1.9 according to the ascertainment probability, with the highest relative efficiency when ascertainment probability was high (the reference here is the

estimation with the PEL). This method appeared as robust as the PEL to a high proportion of unknown genotypes, with AREs varying from 59 to 100% according to the ascertainment model and the proportion of unknown genotypes (results not shown).

APPLICATION

The PEL was applied to data illustrating the MD and CDMS models considered in this paper: 1) families ascertained from patients affected by a genetic disease with variable age of onset, the transthyretin amyloid neuropathy, 2) a sample of breast and ovarian cancer families with *BRCA1* or *BRCA2* mutations, ascertained through early-onset breast cancer cases.

Transthyretin (TTR) amyloid neuropathy

Transthyretin (*TTR*) amyloid neuropathy is an autosomal dominant condition characterised by deposition of amyloid substance made up of mutated *TTR*. This severe condition, firstly described in Portugal, involves mainly the peripheral nervous system and the heart. Although distributed worldwide, the disease is often clustered in limited areas like in Portugal, Japan and Sweden with different genotypic and phenotypic variation including the age of first symptoms. In France, we are dealing with 2 populations i.e. of Portuguese and of French origins. Virtually all the families are referred to the department of Neurology of Bicêtre Hospital which is the national center of reference for this rare disease. Many pathogenic transthyretin variants have been detected among the French population, but only one variant, the Val30Met, was detected in the Portuguese population. In a previous paper, we had analyzed a sample of 79 families (46 French and 33 Portuguese) investigated in the neurology department of Bicêtre Hospital and found a strong difference in penetrance function between these two populations [Plante-Bordeneuve, et al. 2003]. In the present paper, we restricted the

analysis to Val30Met carriers (20 French and 33 Portuguese kindreds). TTR genotype was available for 108 and 139 relatives in French and Portuguese families of whom respectively 47 and 50 were carriers. The proportions of unknown genotypes among relatives were respectively 72% and 68%. The deleterious allele frequency was arbitrarily set to 0.001, and the *de novo* mutation was set to 0 in both analyses.

Figure 2 shows the penetrance functions estimated with the PEL in the French and Portuguese families. Confidence intervals (obtained by bootstraps) are given for ages 50, 60 and 70. The plateau in Portuguese estimation shows the existence of a proportion of carriers which will never be affected (i.e. $\kappa = 0.09$, significantly different from zero with a $p_{value} < 0.001$) which illustrates the importance of implementing κ in the Weibull model. In both samples, the penetrance curve estimated by the Prospective likelihood was very close the one obtained with the PEL, which illustrates that the choice of the method is not crucial in such a situation.

Breast cancer due to BRCA1 or BRCA2 mutations

Families had been selected through 317 women suffering from invasive breast cancer, diagnosed before 36 years between January 1990 and January 1998, and followed up at the Institut Curie. Genetic counselling was proposed to all women and 153 of them came to the appointment to the Institut Curie cancer clinic. A *BRCA1* and *BRCA2* and *TP53* genetic screening was systematically proposed whatever the family history, and 145 of them underwent genetic testing [Chompret, et al. 2001]. The entire coding sequence of both genes was analyzed by a combination of DGGE, DHPLC, and PTT [Stoppa-Lyonnet, et al. 1997; Wagner, et al. 2000]. Sixteen and 14 patients with respectively a germline *BRCA1* or *BRCA2* mutation were identified. In these families, genetic testing was proposed to relatives as recommended by the French guidelines [Eisinger, et al. 1998]. Among the 30 families,

genotype was available for 33 relatives of whom 17 were found to be carriers. In 16 families, the proband was the only one individual tested in the family as no relative asked for genetic testing, and the overall proportion of unknown genotypes among relatives was 91%.

The frequency of the mutated allele, fq, was set to 0.001 and the *de novo* mutation was set to 0.

The cumulative risk for non-carriers was taken as the risk in the general population.

Figure 3 shows the estimations of the penetrance function using the PEL and the Prospective likelihood. The sample being relatively small, families with BRCA1 (16 families) and BRCA2 (14 families) mutations were pooled. Confidence intervals were estimated for ages 50, 60 and 70 years. The two methods provide different penetrance curves, with smaller risks using the PEL, which illustrates that the two methods are not equivalent under such a model. However, the difference is not statistically significant, probably because of the small sample size.

DISCUSSION

As underlined by Vieland and Hodge [1995], the problem of correction for ascertainment is, in most situations, literally intractable. These authors recommended that future efforts focus on the development of robust approximate approaches to the problem. The aim of our study was to propose a simple method to estimate the penetrance function from pedigrees ascertained on one affected case, and where a variable number of relatives have been genotyped, and to determine whether this method fulfilled such requirements.

Using simulations under various genetic models and various ascertainment schemes, the PEL, based on a principle of a classical method proposed in segregation analysis [Weinberg, 1912], turned out to generally provide very satisfactory results. This method has the advantage of being very simple and of leading to unbiased penetrance estimations, provided that the families with several probands are

counted as many times as there are probands in the case of multiple ascertainment. When probands are not specified and families cannot be replicated, we showed that the risks might be underestimated but with a small relative bias. We have shown a relatively low sensitivity to a misspecification of the parameters in the genetic model, and found that the extended Weibull model allowed in general a better, although modest, fit to the data. We also showed that efficiency of the PEL was moderately affected when a substantial proportion of genotypes was unknown under both models, and performed well even when a limited number of relatives have been genotyped.

The Prospective likelihood, which also provides penetrance estimates using genotypes of relatives in pedigrees, usually appears to perform almost equally well as the PEL under the MD model, but not under the CDMS model where this method leads to biased estimates, particularly when the penetrance is low. A possible explanation of the bias observed when using the Prospective likelihood is that it assumes that all individuals have an equal probability of being included in the sample, and in particular that ascertainment events are independent within families. This assumption is usually not fulfilled because, for instance, affected individuals in the older generations may have a null probability of being ascertained, and two siblings are not usually independently ascertained. Moreover, adding an age criterion for the ascertainment implies that all affected individuals do not have the same probability of being ascertained and this is probably one of the reasons why the Prospective likelihood is biased in the CDMS model for which the ascertainment criteria are quite stringent.

The other methods that have been mentioned in the introduction for estimating penetrance apply to a completely different context. The kin-cohort design [Wacholder, et al. 1998], and more generally the genotype-proband design (GPD) and the GPDR [Gail, et al. 1999] cannot be applied to the same samples as the PEL . Indeed, the GPD and GPDR are population-based designs which require very large samples of affected and unaffected individuals and do not need any correction for ascertainment. Moreover these methods have been designed for common mutations in common diseases, and can be

applied neither to the MD model nor to the general situations of mutations predisposing to common cancers in which each mutation is very rare and carriers are difficult to identify at the population level. The Genotype Restricted likelihood (GRL) has been developed for estimating penetrance from pedigrees ascertained on familial criteria, and corrects for ascertainment by conditioning the likelihood on all phenotypes of pedigree members [Carayol and Bonaiti-Pellie 2004]. This method has the advantage of being unbiased, whatever the selection criteria, but, as the retrospective likelihood [Kraft and Thomas, 2000], has the drawback of lacking efficiency, particularly when there are numerous unknown genotypes in the family [Alarcon, et al. 2007]. Therefore, the GRL should be restricted to samples of families ascertained on multiple affected individuals in which ascertainment correction cannot easily be performed. When families are ascertained on specific familial criteria, a Prospective likelihood correcting for these criteria may be used [Le Bihan, et al. 1995], but the effects of a departure from the underlying hypotheses have not been studied.

As an illustration, we estimated the penetrance function for transthyretin amyloid neuropathy from a French sample and a Portuguese sample and for breast cancer from a sample ascertained on early-onset breast cancer cases. For transthyretin amyloid neuropathy in the Portuguese sample and in the French sample, penetrance curves estimated with the two methods were quite similar, as expected according to the simulation study. Interestingly, the importance of introducing a kappa parameter in the Weibull model is well illustrated in the Portuguese sample.

For breast cancer families, the inclusion of an age criterion implies a very low ascertainment probability, in which case the PEL is expected to provide a lower relative bias than the Prospective likelihood. Thus the results obtained by the PEL are more likely to be the correct estimates. We must however keep in mind that not all families fulfilling the inclusion criterion were investigated and that the patients who underwent genetic testing might have been motivated by a stronger family history. Such a potential bias is difficult to correct but should be considered in the interpretation of results.

Note that the risks obtained by the PEL are close to those obtained by Bonadona et al [2005] from a population-based series of early-onset breast cancer cases (age at diagnosis less than 45), and slightly smaller than those obtained by Antoniou et al [2003] from 22 studies unselected for family history, 10 of which were ascertained on early-onset breast cancer cases (limit 36 to 50 years), although risks estimated in these studies are not strictly comparable to our estimations because we pooled *BRCA1* and *BRCA2* mutations.

Confidence intervals were obtained by bootstraps. The use of bootstraps implicitely assumes that families are independent, which is not the case when families are replicated, and in particular for transthyretin amyloid neuropathy families. However, we have checked, by simulations, that estimated variance were similar whether families are replicated or not in case of multiple probands.

ACKNOWLEDGEMENTS

We thank Muriel Belotti for the management of families with *BRCA1 and BRCA2* mutations.

REFERENCES

Abel L, Bonney G. 1990. A time-dependent logistic hazard function for modeling variable age of onset in analysis of familial diseases. Genet Epidemiol 7:391-407.

Alarcon F, Lasset C, Carayol J, Bonadona V, Perdry H, Desseigne F, Wang Q, Bonaiti-Pellie C. 2007. Estimating cancer risk in HNPCC by the GRL method. Eur J Hum Genet 15(8):831-6.

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 young women: results from a prospective population-based study in France. Genes Chromosomes Cancer 43(4):404-13.

Cannings C, Thompson E. 1977. Ascertainment in the sequential sampling of pedigrees. Clin Genet 12:208-12.

Carayol J, Bonaiti-Pellie C. 2004. Estimating penetrance from family data using a retrospective likelihood when ascertainment depends on genotype and age of onset. Genet Epidemiol 27(2):109-17.

Carayol J, Khlat M, Maccario J, Bonaiti-Pellie C. 2002. Hereditary non-polyposis colorectal cancer: current risks of colorectal cancer largely overestimated. J Med Genet 39(5):335-9.

Chompret A, Abel A, Stoppa-Lyonnet D, Brugieres L, Pages S, Feunteun J, Bonaiti-Pellie C. 2001. Sensitivity and predictive value of criteria for p53 germline mutation screening. J Med Genet 38(1):43-7.

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.

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.

Easton DF, Ford D, Bishop DT. 1995. Breast and ovarian cancer incidence in BRCA1-mutation carriers. Breast Cancer Linkage Consortium. Am J Hum Genet 56(1):265-71.

Eisinger F, Alby N, Bremond A, Dauplat J, Espie M, Janiaud P, Kuttenn F, Lebrun J, Lefranc J, Pierret J and others. 1998. Recommendations for medical management of hereditary breast and ovarian cancer: the French National Ad Hoc Committee. Annals of Oncology 9:939-950.

Elston RC, Stewart J. 1971. A general model for the genetic analysis of pedigree data. Hum Hered 21(6):523-42.

Ford D, Easton DF, Stratton M, Narod S, Goldgar D, Devilee P, Bishop DT, Weber B, Lenoir G, Chang-Claude J and others. 1998. Genetic heterogeneity and penetrance analysis of the BRCA1 and BRCA2 genes in breast cancer families. The Breast Cancer Linkage Consortium. Am J Hum Genet 62(3):676-89.

Gail MH, Pee D, Benichou J, Carroll R. 1999. Designing Studies to Estimate the Penetrance of an Identified Autosomal Dominant Mutation: Cohort, Case-Control, and Genotyped-Proband Designs. Genet Epidemiol 16:15-39.

Gong G, Whittemore AS. 2003. Optimal designs for estimating penetrance of rare mutations of a disease-susceptibility gene. Genet Epidemiol 24(3):173-80.

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

Lalouel J. 1979. GEMINI : a computer program for optimization of general non linear function. Technical report no 14. Salt Lake City : University of Utah, Department of Medical Biophysics and Computing.

Le Bihan C, Moutou C, Brugieres L, Feunteun J, Bonaiti-Pellie 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.

Morton NE. 1959. Genetic Tests Under Incomplete Ascertainment. Am J Hum Genet 11(1):1-16.

Pennec S. 1996. La place des familles à quatre générations en France. Population 1:31-60.

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: carrier risks among French and Portuguese families. J Med Genet 40(11):e120.

Stoppa-Lyonnet D, Laurent-Puig P, Essioux L, Pages S, Ithier G, Ligot L, Fourquet A, Salmon RJ, Clough KB, Pouillart P and others. 1997. BRCA1 sequence variations in 160 individuals referred to a breast/ovarian family cancer clinic. Institut Curie Breast Cancer Group. Am J Hum Genet 60(5):1021-30.

Vieland VJ, Hodge SE. 1995. Inherent intractability of the ascertainment problem for pedigree data: a general likelihood framework. Am J Hum Genet 56(1):33-43.

Wacholder S, Hartge P, Struewing JP, Pee D, McAdams M, Brody L, Tucker M. 1998. The kin-cohort study for estimating penetrance. Am J Epidemiol 148(7):623-30.

Wagner T, Stoppa-Lyonnet D, Fleischmann E, Muhr D, Pages S, Sandberg T, Caux V, Moeslinger R, Laugbauer G, Borg A and others. 2000. Denaturing high performance liquid chromatography (DHPLC) detects BRCA1 and BRCA2 mutations with high sensitivity. Genomics 62:369:376.

Weinberg. 1912. Method und Fehlerquellen der Untersuchung auf Mendleschen Zahlen Beim Menschen. Arch. Rass.u.Ges.Biol. 9:165:174.

Figure legends:

Figure 1: Sensitivity of the penetrance function estimate to the presence of a parameter κ in the Weibull model

Figure 2: Penetrance of the TTR mutation involved in the Transthyretin amyloid neuropathy estimated using the PEL and the Prospective likelihood in the French and in the Portuguese family sample.

Figure 3: Penetrance of BRCA1/2 carriers in the French breast cancer families.

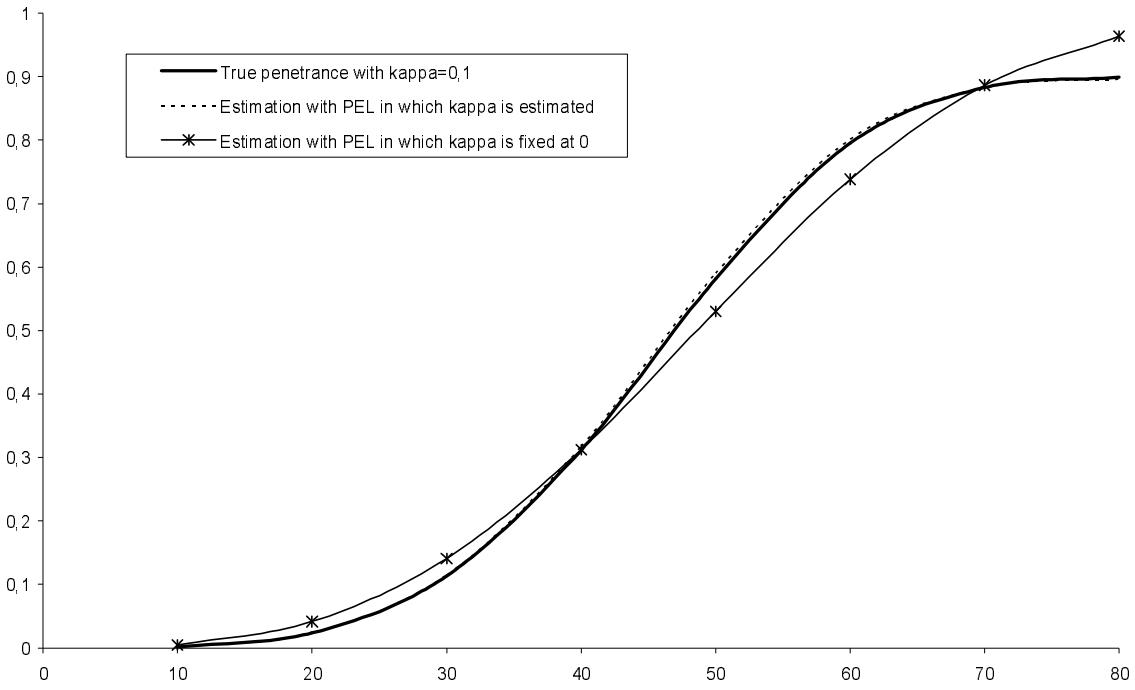
True penetrance with kappa=0,1
Estimation with PEL in which kappa is estimated
Estimation with PEL in which kappa is fixed at 0
1
0,9
0,8
0,7
0,6
0,5
0,4
0,3
0,2
0,1
0
0
10
20
30
40
50
60
70
80

Figure 1

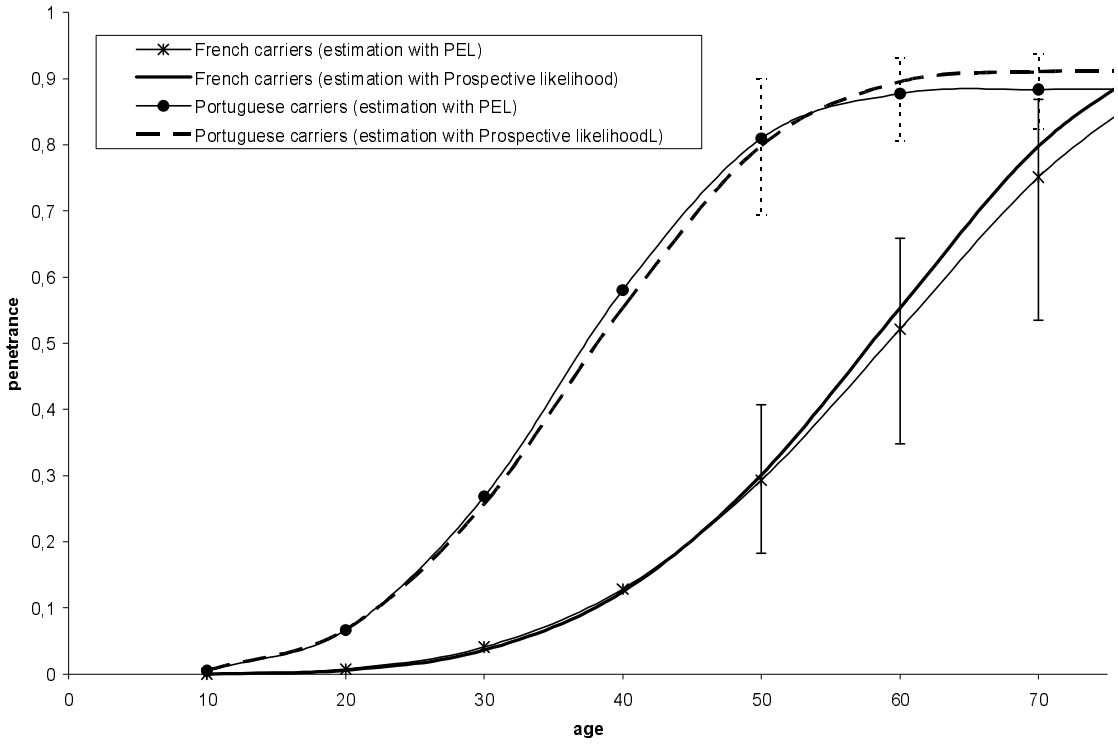
French carriers (estimation with PEL)
French carriers (estimation with Prospective likelihood)
Portuguese carriers (estimation with PEL)
Portuguese carriers (estimation with Prospective likelihoodL)
penetrance
age
0
0,1
0,2
0,3
0,4
0,5
0,6
0,7
0,8
0,9
1
0
10
20
30
40
50
60
70

Figure 2

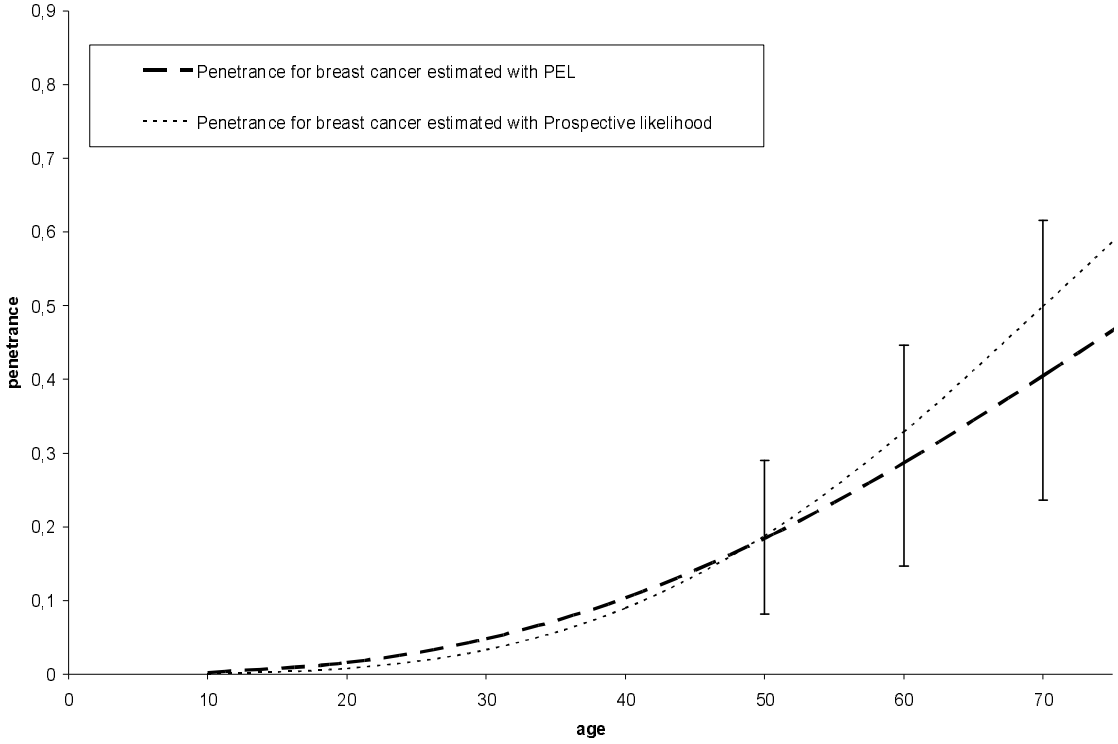
Penetrance for breast cancer estimated with PEL
Penetrance for breast cancer estimated with Prospective likelihood
penetrance
age
0,9
0,8
0,7
0,6
0,5
0,4
0,3
0,2
0,1
0
0
10
20
30
40
50
60
70

Figure 3

Table I : Quantification of the relative bias in penetrance estimate at 70 years using the PEL

Genetic model (*)	p_s (*)	Time period T (in years)	Relative bias (%)	
			High true penetrance	Low true penetrance
MD	1	20	2	2
		1	2	3
	0.5	20	2	3
		1	2	3
CDMS	1	20	2	3
		1	2	4
	0.5	20	2	3
		1	3	3

(*) MD: Monogenic disease; CDMS: complex disease with monogenic sub-entities; p_s : probability of being ascertained for individuals affected during the period T;

Table II : Relative efficiency of the PEL according to the proportion of unknown genotypes under the MD model

Ascertainment probability	Proportion of unknown genotypes among relatives (%)	Asymptotic relative efficiency (reference : all relatives tested)	
		High true penetrance	Low true penetrance
high	50	0.97	0.92
	75	0.92	0.88
	100*	0.82	0.74
low	50	0.98	0.95
	75	0.97	0.91
	100*	0.87	0.84

* Situation in which only the proband's genotype is known

Table III: Quantification of the relative bias in penetrance estimate at 70 years using the PEL without any replication of families in case of multiple probands

Models (*)	p_s (*)	Time period T (in years)	Bias without replication (%)	
			High true penetrance	Low true penetrance
MD	1	20	- 4	- 13
		1	1	1
	0.5	20	- 1	-4
		1	2	2
CDMS	1	20	0	1
		1	2	3
	0.5	20	1	2
		1	2	3

(*) MD: Monogenic disease; CDMS: complex disease with monogenic sub-entities; p_s **:** probability of being ascertained for individuals affected during the time period T.

Table IV: Quantification of the relative bias in penetrance estimate at 70 years using the prospective likelihood

Genetic model (*)	p_s (*)	Time period T (in years)	Relative bias (%)	
			High true penetrance	Low true penetrance
MD	1	20	0	-3
		1	3	2
	0.5	20	1	0
		1	1	-2
CDMS	1	20	5	14
		1	9	21
	0.5	20	5	14
		1	10	21

(*) MD: Monogenic disease; CDMS: complex disease with monogenic sub-entities; p_s : probability of being ascertained for individuals affected during the time period T.