HAL
open science

Estimating penetrance from multiple case families with predisposing mutations: Extension of the "genotype-restricted likelihood" (GRL) method

Bernard Bonaiti, Valérie Bonadona, Hervé Perdry, Nadine Andrieu, Catherine Bonaïti-Pellié

To cite this version:

Bernard Bonaiti, Valérie Bonadona, Hervé Perdry, Nadine Andrieu, Catherine Bonaïti-Pellié. Estimating penetrance from multiple case families with predisposing mutations: Extension of the "genotype-restricted likelihood" (GRL) method. European Journal of Human Genetics, 2010, 19 (2), pp.173 - 179. 10.1038/ejhg.2010.158 . hal-00583530

HAL Id: hal-00583530

https://hal.science/hal-00583530v1

Submitted on 6 Apr 2011

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.

Estimating penetrance from multiple case families with predisposing mutations: Extension of the "genotype-restricted likelihood" (GRL) method

Bernard Bonaïti[1,2], Valérie Bonadona[3,4], Hervé Perdry[2,5], Nadine Andrieu[6,7,8] and Catherine Bonaïti-Pellié[2,5]

[1] INRA-GABI, Jouy-en-Josas, France

[2] INSERM, U669, Villejuif, France

[3] Université Lyon 1, UMR CNRS 5558, Lyon, France

[4] Centre Léon Bérard, Lyon, France

[5] Univ Paris-Sud, Villejuif, France

[6] INSERM, U900, Paris, France

[7] Institut Curie, Paris, France

[8] Ecole des Mines de Paris, ParisTech, Fontainebleau, France

Running title : extension of the GRL method

Corresponding author

Catherine Bonaïti-Pellié

INSERM U669

Bâtiment 15/16

Hôpital Paul Brousse

94807 Villejuif Cedex

catherine.bonaiti@inserm.fr

Abstract

Some diseases are due to germline mutations in predisposing genes, such as cancer family syndromes. Precise estimation of the age-specific cumulative risk (penetrance) for mutation carriers is essential for defining prevention strategies. The genotype-restricted likelihood (GRL) method is aimed at estimating penetrance from multiple case families with such a mutation. In this paper, we proposed an extension of the GRL to account for multiple trait disease and to allow for a parent-of-origin effect. Using simulations of pedigrees, we studied the properties of this method and the effect of departures from underlying hypotheses, misspecification of disease incidence in the general population or misspecification of the index case, and penetrance heterogeneity. In contrast with the previous version of the GRL, accounting for multiple trait disease allowed unbiased estimation of penetrance. We also showed that accounting for a parent-of-origin effect allowed a powerful test for detecting this effect. We found that the GRL method was robust to misspecification of disease incidence in the population, but that misspecification of the index case induced a bias in some situations for which we proposed efficient corrections. When ignoring heterogeneity, the penetrance estimate was biased toward that of the highest risk individuals. A homogeneity test performed by stratifying the families according to the number of affected members was shown to have low power and seems useless for detecting such heterogeneity. These extensions are essential to better estimate the risk of diseases and to provide valid recommendations for the management of patients.

Key-words: penetrance, bias, pleiotropy, parent-of-origin, families

INTRODUCTION

Some diseases with variable age of onset are due to the presence of predisposing gene mutations, such as mismatch repair (*MMR*) genes in hereditary nonpolyposis colorectal cancer (HNPCC) or *BRCA1* and *BRCA2* in breast-ovarian cancer syndrome. These genes may be responsible for hereditary forms of these diseases. Precise estimation of the age-specific cumulative risk (penetrance function) for mutation carriers is essential for defining prevention strategies.

Families in which such mutations have been identified can contribute to estimate these risks, as long as adjustment is made for these families generally ascertained because of several affected members [1]. Such families are usually referred by physicians to genetic counsellors who propose genetic testing when specific criteria are fulfilled. For hereditary cancers, for example, most criteria used for recommending genetic testing are based on familial aggregation of specific cancers [2,3]. When a mutation is identified in an affected member (defined as index case), genetic testing is proposed to close relatives who, if carriers, will be offered intensive surveillance, which will improve the prognosis, or prophylactic surgery when possible.

An ascertainment-adjusted method, based on maximum likelihood, has been proposed for estimating the age-specific cumulative risk (penetrance) of a given disease associated with a deleterious mutation from families in which such a mutation has been identified [4]. This method, called the "genotype-restricted likelihood" (GRL) method, provides unbiased penetrance estimates whatever criteria used for ascertainment of families and without having to model the ascertainment process. The GRL method corrects for the bias due to the selection on carrier genotype of the index case since only families with an identified carrier individual are informative for penetrance estimation. This method is especially appropriate for hereditary predispositions to common diseases when numerous and complex familial criteria involving several affected relatives are used to recommend genetic testing. It has been shown to be independent of selection criteria, in particular on the number of affected individuals and on the age at diagnosis of affected family members [4].

Beside the numerous advantages mentioned above, the GRL method relies on assumptions which may not be fulfilled in some situations. In order to evaluate its robustness to a departure from these hypotheses, we studied the sensitivity of the GRL in various situations such as disease frequency of

the general population over- or under-estimated or a genetic heterogeneity not taken into account. Additionally, the previous GRL version allows estimating the penetrance of only one trait once at a time and this might bias penetrance estimates in family syndromes where several different traits may occur (pleiotropy), like different tumor localizations. Therefore, we proposed in this paper to extend the method to account for pleiotropy. We also extended the method to allow for a parent-of-origin effect, i.e. penetrance functions differing according to the gender of the parent who transmitted the deleterious mutation. This effect has been described in some diseases [5-8] but has not been yet accounted for in penetrance estimation. These extensions are essential to better estimate the risk of diseases and to provide valid recommendations for the management of patients.

METHODS

The GRL method

The GRL [4,9] uses a retrospective likelihood conditioned on the phenotypes of all family members and on the genotype of the index case. It estimates penetrance parameters for mutation carriers by maximizing the probability of observed genotypes (G) of family members who have been tested for the mutation found in the index case, conditional on observed phenotypes (P) and on index case being a carrier (I). Due to the fact that the index case is always tested, the conditional probability may be written as:

$$\Pr(G/P,I) = \Pr(G,P) / \Pr(P,I)$$

Following the cure model of De Angelis [10], we considered that a proportion κ of individuals will never be affected and a Weibull model for penetrance for the others. The cumulative risk by age t is:

$$Pr(\ T < t\ ) = F\ (\ t; \kappa_l, \alpha_l, \lambda_l\ ) = (\ 1 - \kappa_l\ ) \left\{ 1 - exp\ \left[- \left(\lambda_l t \right)^{\alpha_l} \right] \right\}$$

where l is the genotype for the mutation (l = 1, 2, 3 for AA, Aa and aa genotypes respectively, A being the mutated allele), α_l and λ_l are the Weibull shape and scale parameters, κ_l the probability of never being affected given l, and T is the age at disease occurrence.

Let $y_{ij}=(t_{ij}, \delta_{ij}, g_{ij})$ the set of observations on the jth member of the ith family, t_{ij} the age at onset of the disease or the age at censoring (earliest date among dates of prophylactic surgery, death or last

news), δ_{ij} the indicator of occurrence of the disease before the age at censoring and g_{ij} the observed genotype which is coded as 0 when unknown. Let $F(t;\ \theta_l)$ the cumulative risk by age t for the lth genotype and $h(t;\ \theta_l)$ the corresponding hazard functions with $\theta_l=(\alpha_l,\ \lambda_l,\ \kappa_l)$. The probability of the set of observations (y_{ij}) given the latent genotype (u_{ij}) is:

$$\Pr\left(y_{ij}/u_{ij}=l\right)=\left[1-F\left(t_{ij},\theta_l\right)\right]\times\left[h\left(t_{ij},\theta_l\right)\right]^{\delta_{ij}}\times\varphi\left(g_{ij},l\right)$$

where $\varphi(g_{ij},l)$ is the probability of observed genotype knowing the latent genotype : $\varphi(g_{ij},l) = 1$ if g_{ij} =0 or l and 0 otherwise.

Finally it is possible to calculate the probability Pr(G,P) using $\Pr\left(y_{ij}/u_{ij}=l\right)$ with the Elston-Stewart algorithm [11,12]. The denominator, Pr(P,I), is computed in the same way with g_{ij} being known only for the index individual.

The maximization of the conditional likelihood Pr(G/P,I) was performed by the L-BFGS-B algorithm[13]. When analysing a real family sample, confidence intervals of penetrance estimates are obtained by bootstraps.

The method assumes that the mutated allele frequency in the general population is known as well as the penetrance function in non carriers which is set equal to the incidence in the general population.

Extensions of the GRL

Multiple trait phenotype

The contribution to the likelihood of each individual was modified to simultaneously take into account the phenotype of a variable number of possible traits, under the hypothesis that, given genotype, the occurrence of a disease does not modify the risk of developing subsequently other diseases. If t_{ijk} is the age at onset or the age at censoring of the kth disease and δ_{ijk} the indicator of the occurrence of the kth disease before censoring time on the jth member of the ith family, $y_{ij}=(t_{ij1}\ \ t_{ijk}..,\ \delta_{ij1}..\delta_{ijk}..,\ g_{ij})$ with probability :

$$Pr(y_{ij}/u_{ij}=l)=\prod_k\left\{\left[1-F(t_{ijk},\theta_{kl})\right]\times\left[h(t_{ijk},\theta_{kl})\right]^{\partial_{ijk}}\right\}\times\varphi(g_{ij},l)$$

Parent-of-origin effect

To take into account a parent-of-origin effect, we modified the likelihood by splitting the heterozygote genotype in two different genotypes according to the paternal or maternal origin of the mutated allele. Four genotypes were considered: *AA*, *Aa*, *aA* and *aa*, where *Aa* and *aA* are the ordered heterozygous genotypes in which the first allele is transmitted by the father and the second one by the mother. The matrix of genotype probabilities from one generation to another was modified accordingly and three penetrance functions instead of two were considered for gene carriers.

Simulations of pedigrees

To study the statistical properties of the method, and in particular its robustness to departures from underlying hypotheses, as well as the interest of the extensions implemented in the method, we simulated four generation families with a fixed number of relatives: starting from the index case, we simulated a couple of his(her) parents and two couples of grand-parents. The grand-parents and parents of the index case (generation 1 and 2) had respectively two and three children. Each of them had two children as well as all their offspring until the fourth generation (figure 1). Genotypes of family members were randomly generated according to the mutated allele frequency in ancestors and spouses and to Mendel's laws for offspring except the index case who had at least one mutated allele. To simulate realistic situations, we chose to compute parameters to fit the data of a large national French survey of 537 families with Lynch syndrome [14]. Age at last news (or age at death) for each individual was assumed to be normally distributed with 63, 60, 50, 35 years for means and 17, 16, 13 and 11 years for standard deviations respectively for generations 1 to 4. For each family member (including the index case), age at disease occurrence was randomly generated given penetrance function according to genotype. The individual was set as affected or not by comparing the age at disease occurrence and the age at last news. Finally, a family was selected if the index case was affected and if the family fulfilled selection criteria on the minimal number of affected (MNA) individuals.

We considered a dominant genetic model for disease transmission with equal penetrance functions for *AA* and *Aa* genotypes which has been often observed in cancer predisposition so far. The *A* allele

was set to 0.001, so that most of the mutation carriers have the *Aa* genotype. The sets of penetrance functions used for the simulations are shown in table 1. Three sets of penetrance fonctions (high, medium and low risks) were used in the case when only one trait may be observed. To simulate realistic situations for the several different traits exercise, parameters for the simulation of families were chosen to fit the risks published for colorectal, endometrial, ovarian, and urologic tract cancers in Lynch syndrome [9,15-18].

As only some family members are usually tested in families, we had also to simulate when genotypes are known or not, with probabilities fitting to realistic situations in cancer genetics. That is, we fixed some individuals as known or unknown for their genotype and generated the availability on the other family members' genotypes using a probability π, subsequently referred as the genotyping rate, of being tested as follows:

- The index case and his(her) parents are always tested
- Grand-parents and spouses are never tested, as well as the sib of the non carrier parent and his (her) offspring
- The sib of the carrier parent is tested with probability π and, if non carrier, his(her) offspring is not tested
- The offspring of family members found to be carriers is tested with probability π.

Properties of the GRL

Estimates and standard errors of penetrance at different ages between 20 and 80 years were obtained from their average on 200 replicates in different situations according to penetrance among carriers and non carriers, genotyping rate π (0.1, 0.3, 0.5, 0.8, 1.0), number of families in the sample (100, 200, 500), and MNA (2, 3 or 4) individuals for a family to be ascertained. Maximum likelihood estimation should be asymptotically unbiased but some bias may exist with small sample size. It was measured by the average difference between estimations and the true value.

Robustness to departures from underlying hypotheses was investigated by invalidating these hypotheses and by evaluating the bias induced on penetrance estimates when analysing the simulated data. We considered two sources of error when using the GRL method: 1) misspecification of the

disease incidence in the general population; 2) misspecification of the index case (cf. infra). In addition, the GRL assumes that the probability of phenotypes is the same for all mutation carriers, i.e. no heterogeneity in penetrance. However, heterogeneity of penetrance may exist because of modifier factors and one may expect that families with higher penetrance are likely to have more affected individuals than families with lower penetrance. We considered the possibility of penetrance heterogeneity and studied estimates obtained when ignoring this heterogeneity. We also studied the possibility of detecting heterogeneity using the number of affected individuals in the family as a surrogate. Homogeneity tests were performed using likelihood ratio tests that compare the maximum likelihood value L_1 obtained assuming the same penetrance in the different subgroups to the maximum likelihood value L_2 obtained when allowing for a difference among subgroups. The statistic used, $-2 \text{Ln} [L_1/L_2]$, follows a Chi square distribution with a number of degrees of freedom equal to the difference between the number of parameters estimated in L_1 and in L_2 respectively.

The interest of the extension to a possible parent-of-origin effect was evaluated by simulating such an effect in two scenarios i) strong effect with large difference in penetrance, ii) small effect with moderate difference in penetrance according to the parental origin of the mutation. In these two situations, we estimated the power to detect the effect using a likelihood ratio test with a 0.05 type I error. The interest of the extension to a multiple trait phenotype was evaluated by comparing the results of the joint analysis to those obtained by performing separate analyses for each trait, where individuals are considered as unaffected when affected by another trait than the one under study.

RESULTS

Bias and efficiency of the GRL

The bias on penetrance estimates due to limited sample size is shown in table 2 for various situations. The bias is most often positive (i.e. penetrance is overestimated) but very small. Only in extreme situations when the sample size is small (i.e. 100 families), few individuals are tested in the family (e.g. π=0.1) and penetrance is low, the bias may be non negligible compared to the penetrance. This bias increases with age due to the decrease in genotyped individuals among older family

members. There is no variation in bias according to the MNA individuals required for the selection of pedigrees (data not shown).

The efficiency, measured by the standard error of the estimates, also depends on the age, the sample size, the proportion of known genotypes and the true risk values. The results are given in table 2 and show small standard errors for most of estimates. For example, a sample of 200 families with a low genotyping rate (π=0.3) would provide a standard error of 0.07 by age 70 years for a disease with medium risk. The standard errors do not vary with the MNA individuals required for the selection of pedigrees (data not shown).

Robustness to departures from underlying hypotheses

Misspecification of the disease incidence in the general population

The incidence of common diseases in the general population is known when regional or national registries provide accurate rates. However, these registries are usually available for recent time periods only: e.g. the first cancer registry was established in France in 1978. As cancer incidence has varied with time, the incidence specified in the analysis may not be valid for the past generations. In many other diseases for which there are no registries, the incidence may be estimated with some error. We estimated the bias induced by a large error, i.e. by dividing or multiplying the risks by 2 in the analysis, for various values of population incidence. The results are given in table 3. The bias induced by an error on population incidence is rather small and less than 0.05 in a majority of situations. It barely varies with π and the MNA in the families (data not shown).

Misspecification of the index case

The index case is defined as the first individual who was tested in the family. Index cases are often incident cases who asked for genetic counselling because one or several relative(s) are also affected. However, when there are only prevalent cases in a family, the geneticist has to choose, among the affected members, the person who will first be tested and will be considered as the index case. In practice, the geneticist chooses the one with the highest prior probability of being a carrier, in general the youngest affected one. Such a choice is expected to induce a bias on penetrance estimated as

explained on a specific example given in the appendix. To investigate this potential bias, we defined as the index case the youngest affected one in the analyses. The results are shown in table 4 only for MNA=3 or 4. Indeed, no effect was observed for MNA=2 since, in the simulated data, the index case was often the youngest affected one. In general the bias is small, but may be substantial for high and medium penetrance values, with an underestimation of the penetrance for young ages and an overestimation for old ones.

To possibly correct for this bias, we proposed in table 4 two strategies of analysis that free to know the "true" index: 1) indicate as the index case in the analysis an individual chosen at random among affected family members, 2) condition the likelihood on the event that at least one affected individual is a mutation carrier as proposed by Quehenberger et al. [16], with two options for affected individuals participating to the conditioning i) all affected individuals, tested or not; ii) only tested affected individuals. The three methods most often provide a decrease in bias compared with choosing the youngest one; globally, the smallest bias is obtained when choosing the index at random. Regarding the second strategy, a smaller bias was obtained when using the second option, i.e. only tested affected individuals participate to the conditioning.

Penetrance heterogeneity

We considered the case where the population of carriers would be a mixture (50% each) of low risk (penetrance by age 40, 70 and end of life: .05, .20 and .22 respectively) and high risk families (penetrance by age 40, 70 and end of life: .20, .80 and .90 respectively). We analyzed the data under the assumption of homogeneity, and also by stratifying on the number of affected individuals (less than 3 vs. at least three, and less than 4 vs. at least 4) in the family and performed a homogeneity test at the 5% level. The power was estimated as the proportion of 400 replicates with a significant homogeneity test (table 5). As expected, the penetrance estimated when ignoring heterogeneity was toward the penetrance value of high risk families, due to selection on multiple case families. When we stratified the families according to the number of affected members, we obtained different risk estimates in the two strata, but the power for detecting heterogeneity was very low whatever the number of affected members used for stratification.

Extensions of the GRL

Multiple trait phenotype

We considered four different traits, with different values of penetrance between men and women for some traits in order to fit the risks of cancers described in Lynch syndrome as indicated in table 1. The results are given in table 6. The penetrance estimates using single trait analyses are generally underestimated when compared with values in simulated data and the bias is stronger when the MNA individuals in the pedigrees is high, except for colorectal cancer in men. For instance, for endometrial cancer where the true penetrance by age 70 years was simulated at 0.28 and the penetrance estimated separately, i.e. by considering women affected by other cancers as unaffected, was respectively 0.27, 0.22 and 0.18 when families were ascertained on at least two, three, and four affected individuals, respectively. In contrast, the multiple trait analysis provided unbiased estimates.

Parent-of-origin effect

Table 7 shows the power to detect a parent-of-origin effect in samples of various sizes and variable genotyping rate π, for two sets of differences in penetrance according to the gender of the parent having transmitted the mutated allele: a large difference between penetrances by age 70 of respectively 0.30 and 0.60, and a small one of respectively 0.40 and 0.50. As for penetrance heterogeneity, the power was estimated as the proportion of replicates with a significant homogeneity test. The power to detect the parent-of-origin effect was found to be very high, even for samples of moderate size, when the difference was large, but decreased dramatically when the difference between penetrances decreased.

DISCUSSION

Since 1994, several genes have been identified the mutations of which are responsible for hereditary forms of cancer, in particular breast/ovarian cancer and colorectal cancer; guidelines for genetic testing and clinical management have been published and the detection of mutations is now

routinely organized [19,20]. In France, more than 2500 mutations in *BRCA1* and *BRCA2* were found in index cases and more than 7000 relatives had genetic testing between 2003 and 2007. For *MMR* genes, a mutation was found in more than 1000 index cases and nearly 3000 relatives were tested (http://www.e-cancer.fr/v1/fichiers/public/synthese_evolution_activite_2003_2007.pdf). Such familial data provide unique information on carrier risk, as long as adjustment for selection criteria of these families is adequately performed. These criteria are complex and have evolved with time. Moreover, these criteria are proposed as guidelines for genetic counsellors and therefore are not purely applied. Therefore, a formal correction for these criteria is completely unfeasible and the GRL method appears particularly appropriate for estimating disease risks in carriers.

In this paper, we investigated the statistical properties of the GRL method and proposed some extensions.

As any maximum likelihood estimator, the GRL estimator is unbiased only asymptotically. Indeed, we observed a small bias in penetrance estimate for samples of moderate size, particularly at old ages. This bias should be kept in mind when analyzing small samples of families. Regarding precision of the estimates, we showed that the standard errors in current situations were rather small, although retrospective likelihood methods are known to be poorly efficient [21,22].

We studied the robustness of the GRL to departures from underlying hypotheses. We found that the method was very robust to a misspecification of disease incidence in the general population, even for important errors. Such an error could be due to a cohort effect as observed in some cancers [23]. Our results indicated that the method is expected to perform well when using an average disease incidence at an intermediate period between the periods of oldest and youngest generation of the families.

Regarding the specification of the index case, there is no ambiguity if the index is an incident case who asked for genetic counselling and the GRL should be used as it. If there is any doubt on the index identification in some families, when all affected family members are prevalent cases, our recommendation is to use, for these families only, either the random procedure or the conditioning on the genotype of all individuals affected by a disease under study [16], provided that they had genetic testing proving their carrier status. In general, the random procedure provides the least biased estimates but Quehenberger's method[16] is a good alternative.

Regarding penetrance heterogeneity, we found that the estimated penetrance was close to that of highest risk individuals because of selection on multiple case families. A homogeneity test performed by stratifying the families according to the number of affected members was shown to have low power and appears useless for detecting such heterogeneity. However, one must keep in mind that nearly all families referred to genetic counselling so far are multiple case families and that the average penetrance estimated from a sample of such families may be the most appropriate one to predict disease occurrence and to recommend or not genetic testing in such context.

Another source of bias could be that we did not take into account a possible family-specific random effect due to low penetrance genes with multiplicative effects, such as those found by Antoniou et al. [24] in breast cancer, nor modifier genes such as those found to influence cancer risk in *BRCA2* carriers [25,26]. Such an effect could affect major gene penetrance estimate [22,27,28]. The GRL will have to be extended to take into account such effects when better identified.

The GRL method assumes that the frequency of deleterious mutations is low in the general population. A departure from this hypothesis would invalidate our approximation for risks in non carriers that are set to be equal to the risks in the general population. The sample size needed to estimate the risk in non carriers by the GRL is prohibitive and this approximation is necessary. Therefore, we do not recommend using the GRL for mutations with frequency that would exceed 0.01 for a dominant predisposition, or 0.10 for a recessive one, in the general population.

Finally, we could evaluate the efficiency of the extensions of the GRL that we proposed, i.e. possible multiple trait phenotype and parent-of-origin effect. Regarding the former, we found that analyzing separately each trait induced a bias which was corrected when using the multiple trait option. The bias is due to the fact that a single trait analysis uses an inaccurate ascertainment correction when conditioning on only one trait whereas several traits led to the ascertainment of families. Indeed, our results showed that the risks of CRC in men were unbiased when performing a single trait analysis, which is most probably due to the fact that CRC is almost the only tumour observed in carrier men. Therefore, this extension, in addition to sparing time by performing only one analysis, avoids bias, particularly for low risk traits. This is most probably the reason why we found a lower risk of endometrial cancer than other studies when we analyzed a sample of 36 French families

using our first version of the GRL [9]. Among studies that used a retrospective likelihood for estimating cancer risk in Lynch syndrome [9,16,17,29], note that only Quehenberger et al.[16] used a multiple trait approach. This may partly explain the variability of estimates among studies.

The test for a parent-of-origin effect was shown to be very efficient with samples of moderate size when the difference in penetrance was large. It would be interesting to apply the method for testing such an effect in hereditary forms of diseases, such as breast and ovarian cancers associated to mutations of *BRCA1* or *BRCA2* or Lynch syndrome associated with mutations of the MMR genes, as this hypothesis has not been tested yet.

The two extensions of the GRL have also been implemented in the proband's phenotype exclusion likelihood (PEL) that estimates penetrance in case of single ascertainment [30]. Both methods are two different options of the GENERISK software which can be obtained from the authors (bernard.bonaiti@inserm.fr, BONADONA@lyon.fnclcc.fr, nadine.andrieu@curie.net).

CONFLICT OF INTEREST

The authors declare no conflict of interest

REFERENCES

1. Carayol J, Khlat M, Maccario J, Bonaiti-Pellie C: Hereditary non-polyposis colorectal cancer: current risks of colorectal cancer largely overestimated. *J Med Genet* 2002; **39:** 335-339.

2. Eisinger F, Bressac B, Castaigne D *et al*: Identification et prise en charge des prédispositions héréditaires aux cancers du sein et de l'ovaire. *Bull Cancer* 2004; **91:** 219-237.

3. Lindor NM, Petersen GM, Hadley DW *et al*: Recommendations for the care of individuals with an inherited predisposition to Lynch syndrome: a systematic review. *JAMA* 2006; **296:** 1507-1517.

4. Carayol J, Bonaiti-Pellie C: Estimating penetrance from family data using a retrospective likelihood when ascertainment depends on genotype and age of onset. *Genet Epidemiol* 2004; **27:** 109-117.

5. Shete S, Yu R: Genetic imprinting analysis for alcoholism genes using variance components approach. *BMC Genet* 2005; **6 Suppl 1:** S161.

6. Green J, O'Driscoll M, Barnes A *et al*: Impact of gender and parent of origin on the phenotypic expression of hereditary nonpolyposis colorectal cancer in a large Newfoundland kindred with a common MSH2 mutation. *Dis Colon Rectum* 2002; **45:** 1223-1232.

7. Gorlova OY, Lei L, Zhu D *et al*: Imprinting detection by extending a regression-based QTL analysis method. *Hum Genet* 2007; **122:** 159-174.

8. Klutz M, Brockmann D, Lohmann DR: A parent-of-origin effect in two families with retinoblastoma is associated with a distinct splice mutation in the RB1 gene. *Am J Hum Genet* 2002; **71:** 174-179.

9. Alarcon F, Lasset C, Carayol J *et al*: Estimating cancer risk in HNPCC by the GRL method. *Eur J Hum Genet* 2007; **15:** 831-836.

10. De Angelis R, Capocaccia R, Hakulinen T, Soderman B, Verdecchia A: Mixture models for cancer survival analysis: application to population-based data with covariates. *Stat Med* 1999; **18:** 441-454.

11. Elston RC, Stewart J: A general model for the genetic analysis of pedigree data. *Hum Hered* 1971; **21:** 523-542.

12. Thompson E: *Pedigree Analysis in Human Pedigrees*, Baltimore: Johns Hopkins University Press, 1986.

13. Zhu C, Byrd RH, Nocedal J: L-BFGS-B: Algortithm 778: L-BFGS-B, FORTRAN routines for large scale bound constrained optimization. *ACM Transactions on Mathematical Software* 1997; **23:** 550-560.

14. Bonadona V, Bonaïti B, Olschwang S *et al*: Cancer risks associated with germline mutations in *MLH1*, *MSH2* and *MSH6* genes in Lynch syndrome: results from the large nationwide French study ERISCAM. *(submitted)*.

15. Dunlop MG, Farrington SM, Carothers AD *et al*: Cancer risk associated with germline DNA mismatch repair gene mutations. *Hum Mol Genet* 1997; **6:** 105-110.

16. Quehenberger F, Vasen HF, van Houwelingen HC: Risk of colorectal and endometrial cancer for carriers of mutations of the hMLH1 and hMSH2 gene: correction for ascertainment. *J Med Genet* 2005; **42:** 491-496.

17. Jenkins MA, Baglietto L, Dowty JG *et al*: Cancer risks for mismatch repair gene mutation carriers: a population-based early onset case-family study. *Clin Gastroenterol Hepatol* 2006; **4:** 489-498.

18. Watson P, Vasen HF, Mecklin JP *et al*: The risk of extra-colonic, extra-endometrial cancer in the Lynch syndrome. *Int J Cancer* 2008; **123:** 444-449.

19. Schwartz GF, Hughes KS, Lynch HT *et al*: Proceedings of the international consensus conference on breast cancer risk, genetics, & risk management, April, 2007. *Breast J* 2009; **15:** 4-16.

20. Lynch HT, Lynch JF, Lynch PM, Attard T: Hereditary colorectal cancer syndromes: molecular genetics, genetic counseling, diagnosis and management. *Fam Cancer* 2008; **7:** 27-39.

21. Kraft P, Thomas DC: Bias and efficiency in family-based gene-characterization studies: conditional, prospective, retrospective, and joint likelihoods. *Am J Hum Genet* 2000; **66:** 1119-1131.

22. Choi YH, Kopciuk KA, Briollais L: Estimating disease risk associated with mutated genes in family-based designs. *Hum Hered* 2008; **66:** 238-251.

23. Belot A, Grosclaude P, Bossard N *et al*: Cancer incidence and mortality in France over the period 1980-2005. *Rev Epidemiol Sante Publique* 2008; **56:** 159-175.

24. Antoniou AC, Pharoah PD, McMullan G *et al*: A comprehensive model for familial breast cancer incorporating BRCA1, BRCA2 and other genes. *Br J Cancer* 2002; **86:** 76-83.

25. Antoniou AC, Spurdle AB, Sinilnikova OM *et al*: Common breast cancer-predisposition alleles are associated with breast cancer risk in BRCA1 and BRCA2 mutation carriers. *Am J Hum Genet* 2008; **82:** 937-948.

26. Antoniou AC, Sinilnikova OM, Simard J *et al*: RAD51 135G-->C modifies breast cancer risk among BRCA2 mutation carriers: results from a combined analysis of 19 studies. *Am J Hum Genet* 2007; **81:** 1186-1200.

27. Carroll RJ, Gail MH, Benichou J, Pee D: Score tests for familial correlation in genotyped-proband designs. *Genet Epidemiol* 2000; **18:** 293-306.

28. Li H, Thompson E: Semiparametric estimation of major gene and family-specific random effects for age of onset. *Biometrics* 1997; **53:** 282-293.

29. Stoffel E, Mukherjee B, Raymond VM *et al*: Calculation of risk of colorectal and endometrial cancer among patients with Lynch syndrome. *Gastroenterology* 2009; **137:** 1621-1627.

30. Alarcon F, Bourgain C, Gauthier-Villars M, Plante-Bordeneuve V, Stoppa-Lyonnet D, Bonaiti-Pellie C: PEL: an unbiased method for estimating age-dependent genetic

disease risk from pedigree data unselected for family history. *Genet Epidemiol* 2009; **33:** 379-385.

1 Table 1. Penetrance functions used for simulations

2

phenotype	Disease	gender	Penetrance in mutation carriers by age			Penetrance in non carriers by age		
			40	50	70	40	50	70
Single trait	Low risk	all	0.01	0.02	0.05	0.0001	0.0006	0.005
	Medium risk	all	0.10	0.19	0.40	0.001	0.006	0.05
	High risk	All	0.20	0.40	0.80	0.02	0.04	0.10
Multiple trait	CRC	M	0.06	0.14	0.40	0.0009	0.013	0.032
		F	0.04	0.10	0.30	0.0008	0.009	0.020
	END	F	0.05	0.11	0.28	0.0002	0.004	0.011
	OVA	F	0.02	0.04	0.07	0.001	0.006	0.010
	URE	M	0.005	0.009	0.02	0.0001	0.0005	0.0009
		F	0.005	0.009	0.02	0.0001	0.0003	0.0004

3 M: male, F: female; CRC: colorectal cancer, END: endometrial cancer, OVA: ovarian cancer, URE:

4 cancer of the urologic tract

5

Table 2. Maximum likelihood bias (and standard errors) on penetrance estimate by ages 50 and 70 years, according to sample size and genotyping rate (π), for low, medium and high penetrance values (selection of families on at least two affected individuals, 200 replicates)

		Bias on penetrance estimate (standard error)					
		By age 50			By age 70		
Penetrance	π	Number of families			Number of families		
		100	200	500	100	200	500
Low	0.1	0.05	0.02	0.01	0.06	0.03	0.01
		(0.08)	(0.06)	(0.03)	(0.08)	(0.06)	(0.03)
	0.3	0.02	0.02	0.01	0.03	0.02	0.01
		(0.05)	(0.03)	(0.02)	(0.05)	(0.03)	(0.02)
	0.8	0.02	0.01	0.00	0.02	0.01	0.01
		(0.03)	(0.02)	(0.01)	(0.04)	(0.02)	(0.01)
Medium	0.1	0.02	0.02	0.01	0.03	0.03	0.01
		(0.11)	(0.08)	(0.05)	(0.12)	(0.09)	(0.05)
	0.3	0.01	0.01	0.00	0.02	0.02	0.01
		(0.08)	(0.05)	(0.03)	(0.10)	(0.07)	(0.04)
	0.8	-0.01	0.00	0.00	0.01	0.01	0.00
		(0.05)	(0.04)	(0.02)	(0.08)	(0.06)	(0.04)
High	0.1	0.02	0.01	0.01	0.00	0.00	0.00
		(0.09)	(0.07)	(0.04)	(0.09)	(0.07)	(0.04)
	0.3	0.01	0.00	0.01	0.00	0.00	0.00
		(0.07)	(0.05)	(0.03)	(0.08)	(0.06)	(0.03)
	0.8	0.01	0.00	0.00	-0.01	0.00	0.00
		(0.04)	(0.03)	(0.02)	(0.06)	(0.04)	(0.02)

Table 3. Bias induced by a misspecification of population incidence on penetrance estimate (500 families, π =0.5, selection of families on at least two affected individuals, 200 replicates)

		Bias on penetrance estimate by age (years)				
True penetrance value	Disease incidence Multiplied by *	30	40	50	60	70
Low	0.5	0.00	0.00	0.00	-0.01	-0.02
	1	0.01	0.01	0.01	0.01	0.01
	2	0.01	0.01	0.02	0.04	0.05
Medium	0.5	0.04	0.03	-0.01	-0.05	-0.10
	1	0.00	0.00	0.00	0.00	0.00
	2	-0.01	-0.01	0.02	0.06	0.11
High	0.5	-0.01	-0.03	-0.04	-0.03	-0.03
	1	0.00	0.00	0.00	0.00	0.00
	2	0.01	0.03	0.04	0.05	0.04

* When this factor is equal to 1, the true population incidence value is specified and the bias is due to limited sample size (see table 2)

1 Table 4. Bias induced by a misspecification of the index case (500 families, π =0.5, 200 replicates)

2

True penetrance value	Choice of index	Bias on penetrance estimate			
		MNA = 3		MNA = 4	
		Age 50	Age 70	Age 50	Age 70
Low	Youngest	0.00	0.02	0.00	0.02
	Random	0.01	0.01	0.01	0.01
	At least 1 A	0.02	0.06	0.02	0.06
	At least 1 AT	0.01	0.03	0.02	0.03
Medium	Youngest	- 0.04	0.04	- 0.03	0.05
	Random	- 0.02	0.01	- 0.01	0.01
	At least 1 A	-0.02	0.06	- 0.01	0.09
	At least 1 AT	- 0.01	0.06	0.00	0.07
High	Youngest	- 0.06	0.00	- 0.06	0.01
	Random	- 0.01	- 0.02	- 0.01	- 0.01
	At least 1 A	- 0.03	- 0.06	- 0.02	- 0.04
	At least 1 AT	0.01	0.01	0.01	0.01

3 MNA: minimal number of affected individuals required for selection of families; Youngest: analysis
4 performed with the youngest affected specified as the index. Random: index chosen at random among
5 affected individuals known as carriers. At least 1 A: no index specified but likelihood conditioned on
6 at least one carriers among affected individual, tested or not; At least 1 AT: likelihood conditioned on
7 at least one carrier among affected and tested individuals.

Table 5. Power to detect penetrance heterogeneity in case of mixture (50% each) of high and low risk families using a stratification on the number of families (500 families, π =0.5, MNA=2, 400 replicates)

Estimated penetrance						Power for detecting heterogeneity
By age	Assuming homogeneity	After stratification of families according to the number of affected individuals				
		< 3	≥ 3	< 4	≥ 4	
40	0.17	0.12	0.14	0.18	0.19	0.08
70	0.66	0.51	0.58	0.71	0.75	0.10

1 Table 6. Comparison between single trait and multiple trait estimation for 4 cancer localizations by

2 age 70 years according to MNA (500 families, π =0.5, 200 replicates)

3

Tumor	True penetrance value	Single trait estimation			Multiple trait estimation		
	MNA	2	3	4	2	3	4
CRC (men)	0.40	0.40	0.40	0.40	0.39	0.40	0.41
CRC (women)	0.30	0.28	0.24	0.21	0.30	0.31	0.31
END	0.28	0.27	0.22	0.18	0.26	0.27	0.28
OVA	0.07	0.07	0.06	0.05	0.08	0.08	0.08
URE	0.02	0.02	0.015	0.014	0.04	0.04	0.04

4 CRC: colorectal cancer, END: endometrial cancer, OVA: ovarian cancer, URE: cancer of the urologic

5 tract

6

7

Table 7. Power for detecting a parent-of-origin effect according to number of families and genotyping rate (MNA=2, 200 replicates)

Penetrance values according to origin of mutation	Number of families	Genotyping rate (π)	Power
P40=.25 & .40 P70=.30 & .60	100	0.2	0.61
		0.5	0.64
	200	0.2	0.97
		0.5	0.98
	500	0.2	1.00
		0.5	1.00
P40=.30 & .35 P70=.40 & .50	100	0.2	0.01
		0.5	0.01
	200	0.2	0.05
		0.5	0.04
	500	0.2	0.17
		0.5	0.15

Legend for figure

Figure 1. Family structure used for the simulations (this is only an example as the sex of each individual is randomly generated)

Appendix: example of misspecification of the index case

In 2005, a woman was affected by colorectal cancer at age 52 years and died two years later from ovarian cancer (case 1). Nobody suspected Lynch syndrome at that time although one of her uncle, currently aged 70 years, was affected at age 45 years by colorectal cancer (case 2). In 2009, the brother of the latter, aged 71 years, also developed colorectal cancer (case 3). The sister of case 1 suspected a hereditary syndrome and asked for genetic counselling. As case 1 was not available any more and case 3 could be a sporadic case because of late-onset diagnosis, the geneticist proposed genetic testing of case 2. Therefore, case 2 was designed as the index case although case 3 was obviously the incident case that would have been the 'natural' index.

Why should this choice induce a bias in the analysis? As the GRL is conditioned not only on the phenotypes of family members, but also on the genotype of the index case, the choice of case 2 as the index case 'cancels' his contribution to the likelihood and replaces it by the contribution of case 3. As case 3 was affected at an older age than case 2, this tends to overestimate age of onset and to decrease the penetrance estimate.
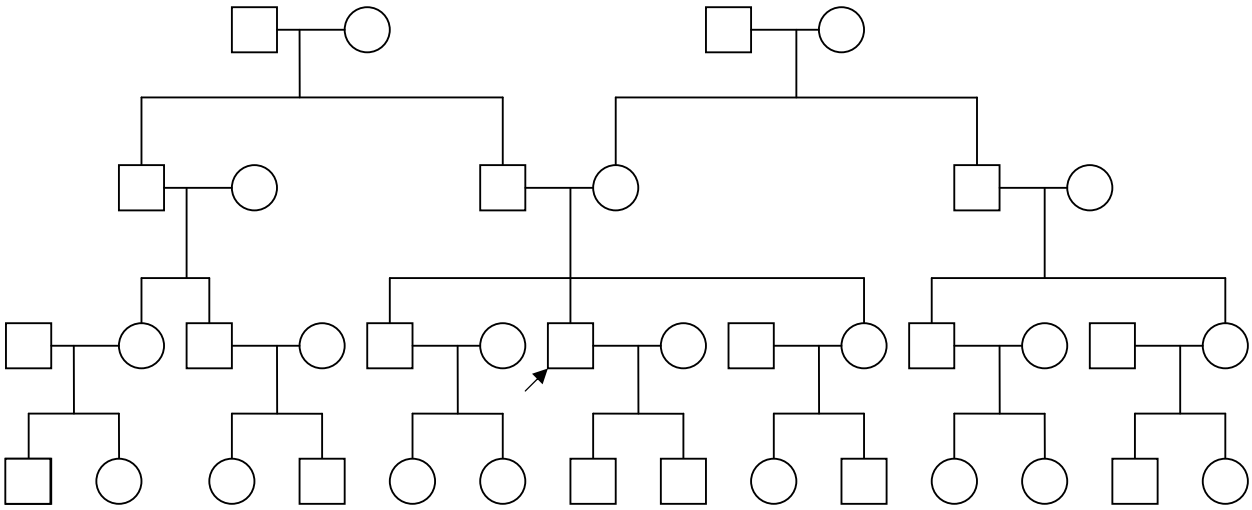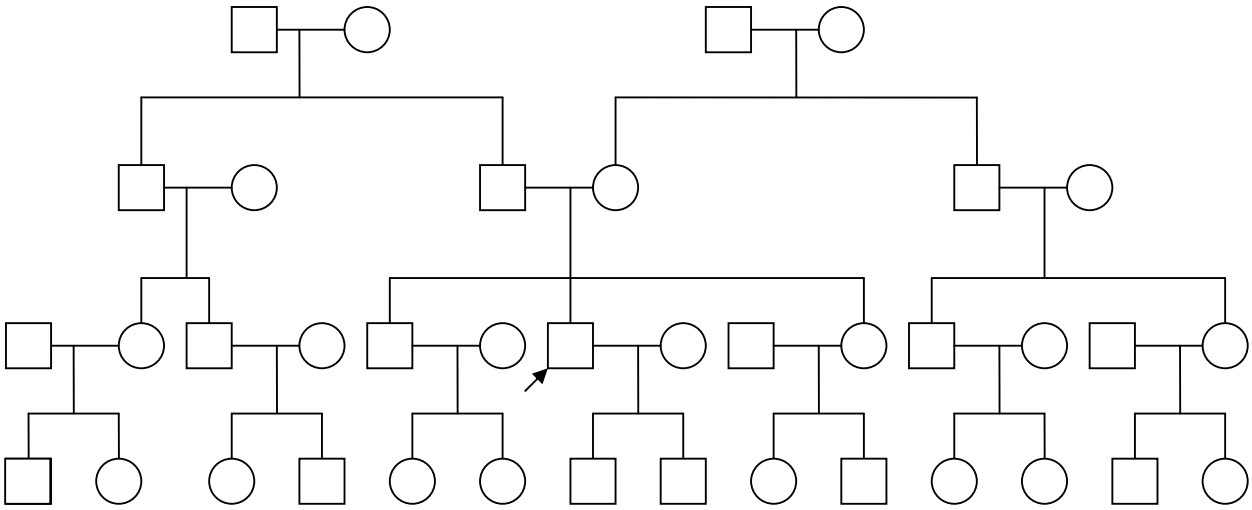
Figure 1