Oncogenetic pedigrees: relation between design and ability to predict mutation

Fabrice Kwiatkowski, Laurent Serlet, Andrzej Stos

Oncogenetic pedigrees: relation between design and ability to predict mutation

FABRICE KWIATKOWSKI[1,2], LAURENT SERLET[1,3], ANDRZEJ STOS[1,4]

[1] Université Clermont Auvergne, CNRS, Laboratoire de Mathématiques Blaise Pascal (UMR 6620), F-63000 CLERMONT-FERRAND, FRANCE.

[2] ORCID 0000-0002-4041-3999

[3] ORCID 0000-0002-3867-0312

[4] ORCID 0000-0002-4986-901X

Abstract. We consider the risk of a disease caused by the presence within the genome of one deleterious mutation or two interacting mutations. For an individual, the probability of being mutated/doubly mutated can be estimated knowing the phenotype of the other members in the family pedigree. We study the performance of this process as a function of the size, the shape of the family tree and the parameters of the model. We carry out simulations using the parameters pertaining to breast/ovarian cancer in BRCA-mutated families.

Keywords: oncogenetics, risk modelisation, family pedigree, simulation

Code availability: custom Python code freely downloadable

MSC Classification 62P10, 92D10, 92C50, 92-10, 92-08

Acknowledgements:

Computations have been performed on the supercomputer facilities of the Mésocentre Clermont Auvergne.

1 Introduction :

Diseases caused by genetic mutations are numerous. One of the most representative cases is likely breast/ovarian cancer favored by BRCA mutations. But other mutations are suspected for many cancers as well as in other types of diseases like cardiovascular ones. Neurodegenerative diseases may soon join the list of pathologies associated to mutations. Even if sequencing the genome - or, as far as we are concerned, part of it - has become faster and much less expensive, it cannot yet be systematically performed for various reasons and its analysis requires to have fully identified the mutations involved. Therefore we still need reliable computer tools first to evaluate a familial predisposition considering its medical data, i.e. the pedigree, second to limit the genetic inquiry to a reduced number of susceptibility genes and third, to estimate the individual risk to inherit a deleterious genetic mutation.

Scoring systems (BRACAPRO [Berry, 2002], Manchester [Evans, 2009], Eisinger [Eisinger, 2004]…) have been built in the context of HBOC (Hereditary Breast/Ovarian Cancer) and various refinements have been proposed [Bonaïti, 2011]. These scoring methods only require a few additions, so their implementation is easy. However, the amount of family information they use is limited and their performance is not optimal. Nowadays, family history is becoming better and better reported, going back several generations and stored in databases. Relatively complex calculations are therefore easy to perform. One can in particular calculate the probability of mutation knowing all the family information which is the best predictor of the mutation risk. This method has long been described by Elston and Stewart [Elston, 1971]. The prognosis can of course be confirmed or reversed after sequencing, in the case of mutations that are clearly identified. And in all cases, this risk calculation is a valuable tool in the personalization of screening and medical monitoring.

In this context of an ever-increasing availability of data and of calculation possibilities, the aim of this article is to optimize the design of this method and to examine its performance using simulations.

A few words about the underlying model first. Since the prognosis method we are discussing is based only on probability calculations, it does not require a detailed knowledge of the nature of the mutation concerned (or the mutations concerned). We will assume, as in most of the known cases, that the deleterious effect of the mutation consists of a strong increase in penetrance i.e. the probability of declaring the disease at some point in one's life and an earlier age of onset. These effects are evidently seen in the case of cancer, but we aim in this work to establish a general mathematical framework that is not restricted to a specific disease. In what follows, the disease caused by the mutations will be called and noted K.

Whether the antecedents are collected by the oncogeneticist especially for this prognosis purpose or whether they are pre-existing in a database, it is necessary to question the desirable period of this family exploration, that is the size and shape of the family tree to take into account, provided of course that such a tree can be completed in relation to the individual concerned. Several issues naturally arise:

- Is the prognosis always better when a larger number of parents are included?
- Is it desirable to go back the generations as much as possible?
- Does taking siblings into account in the ancestral line improve the result? Are cousins useful?
- Is the prognosis easier in families with several cases of illness?
- Do descendants provide information as valuable as ascendants to evaluate the mutation risk?
- What sensitivity and specificity can be typically achieved by this kind of method and what is their variability according to the parameters of the model?

We will answer these questions, with answers that are sometimes surprising, in Section 3. Before that, we present the modelling of the data and the principle of the computational algorithm in the following Section 2. We end this paper in Section 4 with a discussion about the limitations and possible extensions of our work.

2 Model and data description

The basis for constructing a random model is that individuals from the same family have genotypes stochastically linked to each other by Mendel's laws (50% chance of inheriting a mutation from one of its parents). But these genotypes are hidden and all statistically usable data (phenotypes) are random expressions of these genotypes. Mathematically, this is a hidden Markov process indexed by a tree. Two difficulties then arise:

- The combinatorial explosion in the number of possible genotypes makes unrealistic - except for small families - any approach based on the exhaustive examination of all possible genotypes, i.e. a brute force method.
- Because the considered mutations are relatively rare in frequency in the general population, the effect to be analyzed remains narrow and it must be distinguished from the random noise which here takes the form of sporadic onset of the disease, generally at an older age.

In the following numerical calculations, the probability of the disease occurrence (as a function of the age), depending on whether a mutation is involved or not, will have the same distribution as what is observed for breast/ovarian cancers with or without a BRCA mutation:

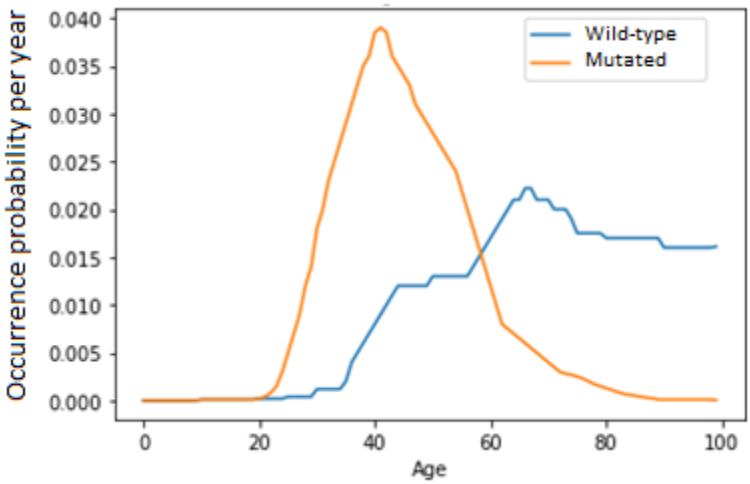


Figure 1: probability laws used to model the onset of cancer according to age.

In addition, we only consider the case of heterozygous mutations, knowing that homozygous mutations are generally lethal during embryogenesis. We therefore consider that the genome of each individual is a random quantity having two possible states {wild-type, mutated} with the following associated probabilities:

Table 1: Probability for an individual to inherit a deleterious mutation from his parents depending on whether or not they are carriers

↓ mother / father →	wild-type	mutated
wild-type	0	0.50
mutated	0.50	0.75

We assume that all the parameters of the model are known:

- the mutation frequency in the general population, p_m
- the penetrance of K for non-mutated subjects, p_0
- the penetrance of K for the mutated subjects, p_1

However, these parameters can be varied in order to consider penetrances linked to less deleterious mutations or even interactions between mutations and polymorphisms that are very poorly penetrating alone. The methods for evaluating these different contexts will be discussed in Section 4.

Here, we assume that the data available per individual is the year of birth, whether or not the disease occurred, and if so, the age at reporting. Such an approach, although it appears succinct, has the advantage of being easily generalizable. To quantify the performance of the algorithms used, we are going to perform data simulations according to the following model: for each family pedigree, we will first generate the (hidden) genotypes: for any individual, his genotype is obtained randomly but according to that of his parents if they are present; if they do not exist in the family pedigree, we will use the frequency p_m. We then randomly

assign birth years to the family members so that ages are consistent with the generations. Finally, the phenotypes are generated by taking into account the genotypes, the value of the parameters p_0 and p_1 and the laws described at Figure 1.

2.1 Principle of the mutation risk computation

The calculation of the risk of mutation considering the disease cases in the family by the conditional probability calculation, as reported below, has been described long ago [Elston, 1971]. However, its implementation for each specific case is complex. For this reason, practitioners rather use scoring methods which only require making a few additions or subtractions but without being able to claim the same performance [Bonaïti, 2014].

Let us denote by F the set of phenotypes of all family members - known family data - and denote by G the compilation of the (unknown) genotypes of individuals, in particular G(i) will denote the genotype of the i^{th} individual. For the single mutation model which we are working with until stated otherwise, we will note $G(i) = 1$ when individual i is mutated and $G(i) = 0$ otherwise.

An elementary probability calculation from Bayes' theorem shows that:

$$\mathbb{P}(G(i) = 1 \mid F) = \frac{1}{1 + \frac{\text{numerator}}{\text{denominator}}} \quad (1)$$

where

$$\text{numerator} = \sum_{g\,:\,g(i)=0} \mathbb{P}(F \mid G = g)\ \mathbb{P}(G = g)$$

and

$$\text{denominator} = \sum_{g\,:\,g(i)=1} \mathbb{P}(F \mid G = g)\ \mathbb{P}(G = g)$$

In each of these two expressions, the sum is extended to all possible genotypes satisfying the mentioned constraint on the genotype $g(i)$ of individual i. Explicit expressions for the conditional probability of the phenotype F knowing that the genotype is g and for the probability that $G = g$ are given in Section 2.4.3.

Note that the calculation of the numerator and the denominator require either scanning all possible genotypes or proceeding by a Monte-Carlo method. The first method, called "by brute-force", is possible for small families and requires the calculation of the likelihood of each genotype g. Note, moreover, that a fraction of genotypes are impossible, therefore of zero probability; the summation is *de facto* reduced to eligible genotypes. The Monte-Carlo method consists in simulating a series of genotypes, either by direct process or by a Metropolis algorithm [Metropolis, 1953]. With this method, formula (1) is rewritten:

$$(2) \qquad \frac{\text{numerator}}{\text{denominator}} = \lim_{N \text{ large}} \frac{\sum_{j=1}^{N} \mathbf{1}_{\{G_j(i)=0\}} \, \mathbb{P}(F|G_j)}{\sum_{j=1}^{N} \mathbf{1}_{\{G_j(i)=1\}} \, \mathbb{P}(F|G_j)}$$

where G_1, G_2, ... forms a series of random draws among the set of possible genomes, according to Mendel's laws and the contextual hypotheses mentioned above. The 1_A function refers to the binary indicator function which is equal to 1 if A is carried out and 0 otherwise. Numerically, it is necessary to ensure that there is a sufficient number N of simulations so that both the numerator and especially the denominator are estimated correctly.

The (conditional) probability of mutation obtained by formula (1) is a quantification of the risk of mutation of an individual but also indirectly of the family risk. One may prefer a "mutated/non-mutated" binary predictor, which will be obtained by thresholding: for a threshold level s, the individual i is prognosticated as mutated if $P(\, G(i) = 1 \mid F) > s$. From a certain point of view, one may regret this conversion in binary prognosis of a quantitative risk which is *a priori* more informative. However, this conversion then makes it possible to analyze the predictor with the usual performance indicators, i.e. in terms of sensitivity (percentage of well predicted among the mutated) and specificity (percentage of well predicted among the non-mutated). We will present the traditional Receiver Operating Characteristics (ROC) curve and in the rest of this chapter, the different predictors will be compared using the relative position of their ROC curves and the corresponding areas under curve (AUC). We can also consider evaluating the performance on the whole family tree, which amounts to averaging the predictors on all the individuals of the pedigree. The results are generally better, but less pertinent so we will leave them aside to focus on the individual case.

2.2 Standardization of family trees

One of our main purposes is to address the influence of the size and shape of the family tree considered. In a real situation, the family tree is constrained in its form by the number of children born of the different couples and by the ability to obtain reliable information about the phenotype of each individual. This last point applies particularly to older generations, because they have had time to present the disease. As a result, the size and especially the shape can vary a lot.

To facilitate comparisons, we are going to use a generic form of regular pedigree, with two parameters which are the size (number of children per couple) and the height (number of generations), and to answer specific questions later, we will define other typical shapes. We call regular pedigree with $n_g \geq 2$ generations and $n_e \geq 1$ children per couple and we denote by $\text{Reg}(n_g, n_e)$ a pedigree whose skeleton is the ancestral line of an individual numbered 0 which goes back to n_g generations including the ancestors of two genders and to this skeleton, we add n_e children for each couple. Here is a diagram of a Reg(4, 3) type pedigree:

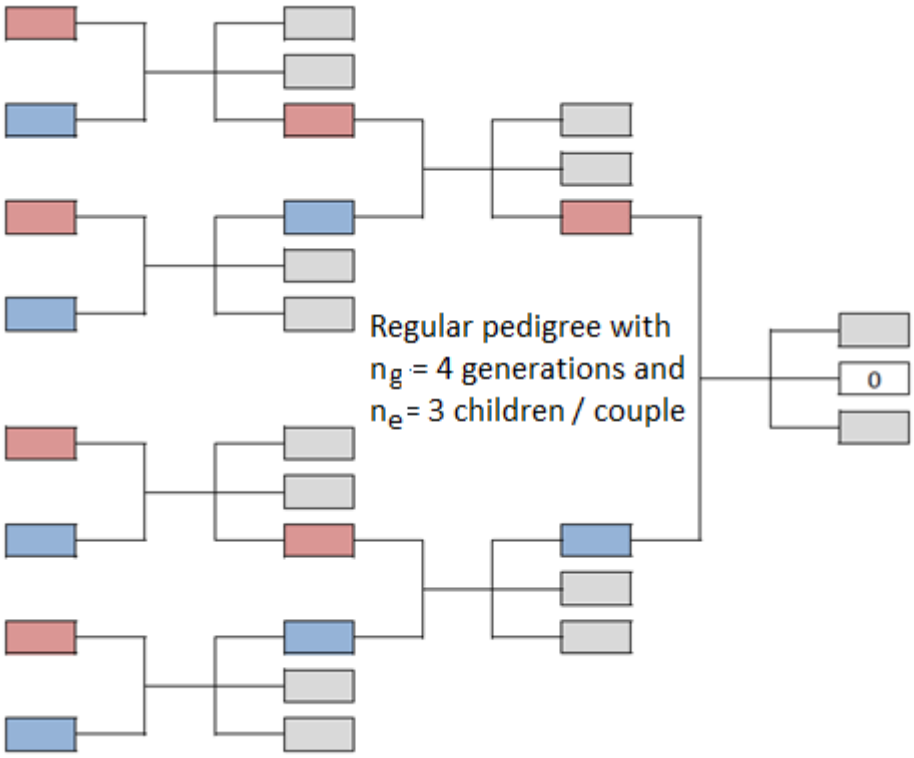


Figure 2: Diagram of a regular pedigree of 4 generations and 3 children per couple. Women are colored in pink, men in blue and regardless of gender in grey. Individual quoted 0 is the one considered for the calculation of the mutation risk.

For comparisons, we can fix n_g and vary $n_e \geq 1$, or fix $n_e \geq 1$ and vary $n_g \geq 2$. An intuitive idea -- maybe naïve -- is that the more individuals are taken into account (richness of phenotype), the more the performance of the predictor increases, probably up to a certain limit. A related but more subtle question is to assess the importance of fixed-size tree shapes: is it better to have more generations but fewer individuals per generation than the opposite? In particular, we will compare the performance of the predictors for two pedigrees both comprising 15 individuals. The first is Reg(4, 1) which we will call “Grt Grd Par” because it involves all the great-grandparents. Here it is:

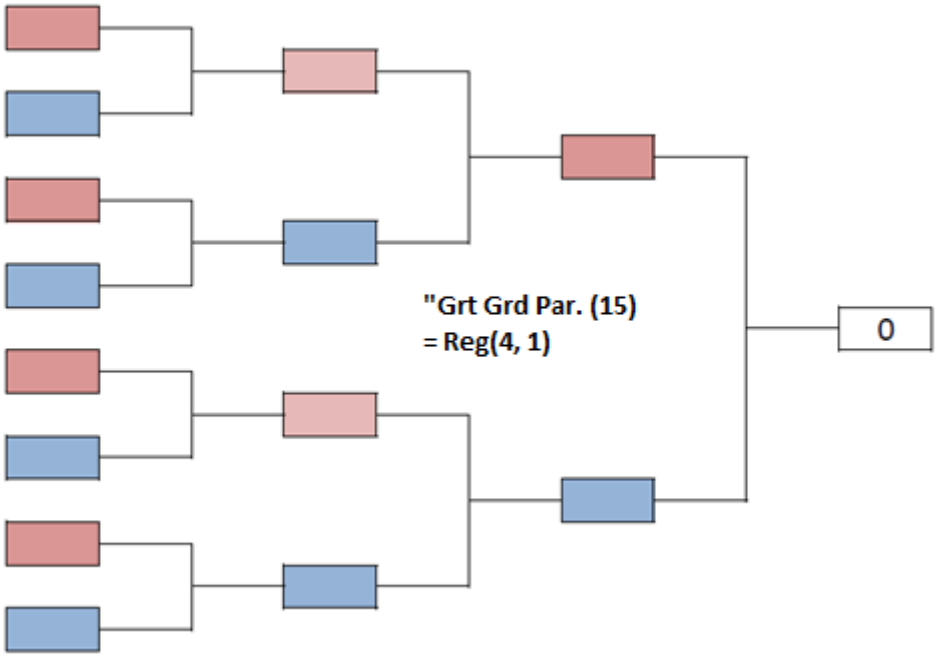


Figure 3: Regular pedigree with 4 generations and 1 child per couple

The second that we call “Wide 3 gen.” also includes 15 members but only 3 generations because of the addition of uncles / aunts and cousins.

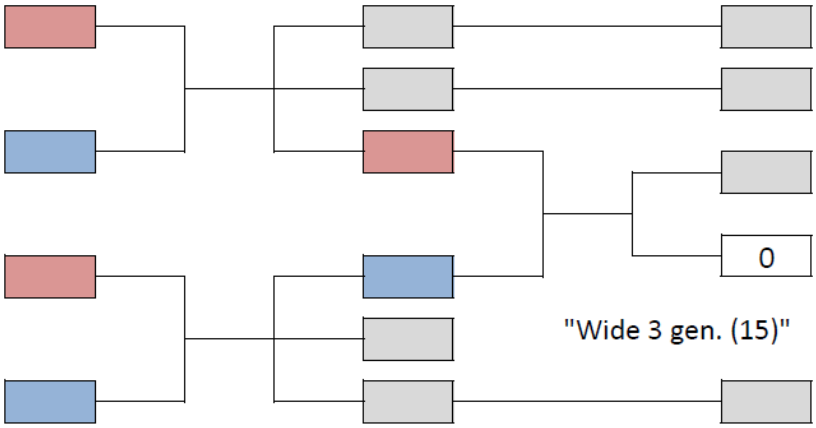
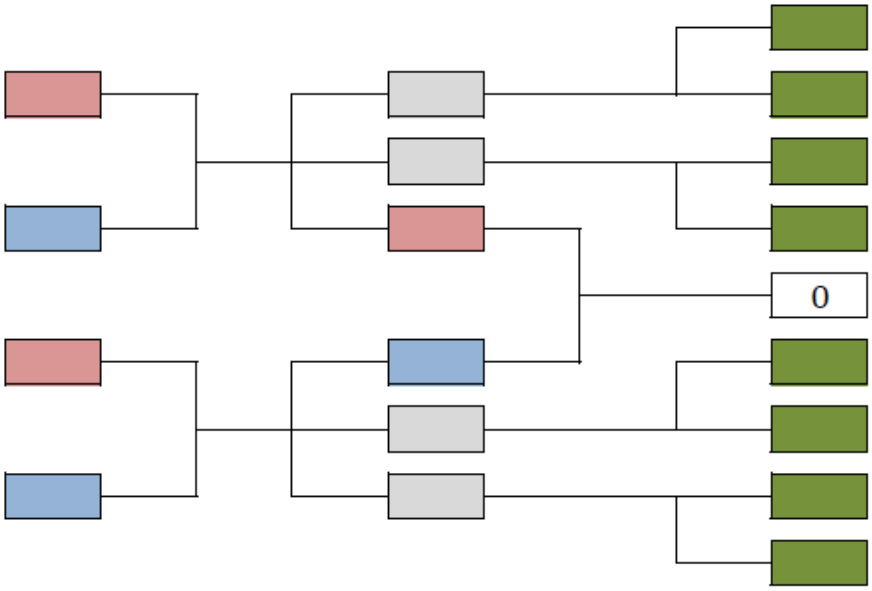


Figure 4: Design of the pedigree « wide 3 gen. » with 15 members.

In the case of the regular tree Reg(n_g, n_e) with $n_g \geq 3$ and $n_e \geq 2$, individual 0 has uncles (or aunts) but his cousins do not appear. In general, it might be interesting to isolate the influence of individuals a little further from the direct ascending line of individual 0. We will compare the family “Uncles”, a family with 11 individuals, to the family “Cousins” with 19 individuals. This latter strictly contains the family “Uncles” but 8 cousins are added, drawn in green in the diagram below:



"Uncles" (11) + 8 (in green) = "cousins" (19)

Figure 5 : design of the pedigree "Cousins" (19) after the addition of 8 cousins (in green) to the pedigree "Uncles" (11)

Studying the influence of the shape, we also study the case of an asymmetric family pedigree in the sense that only one of the parents of individual 0 is entered. For example, we compare the predictors obtained for the following two families of almost identical sizes:

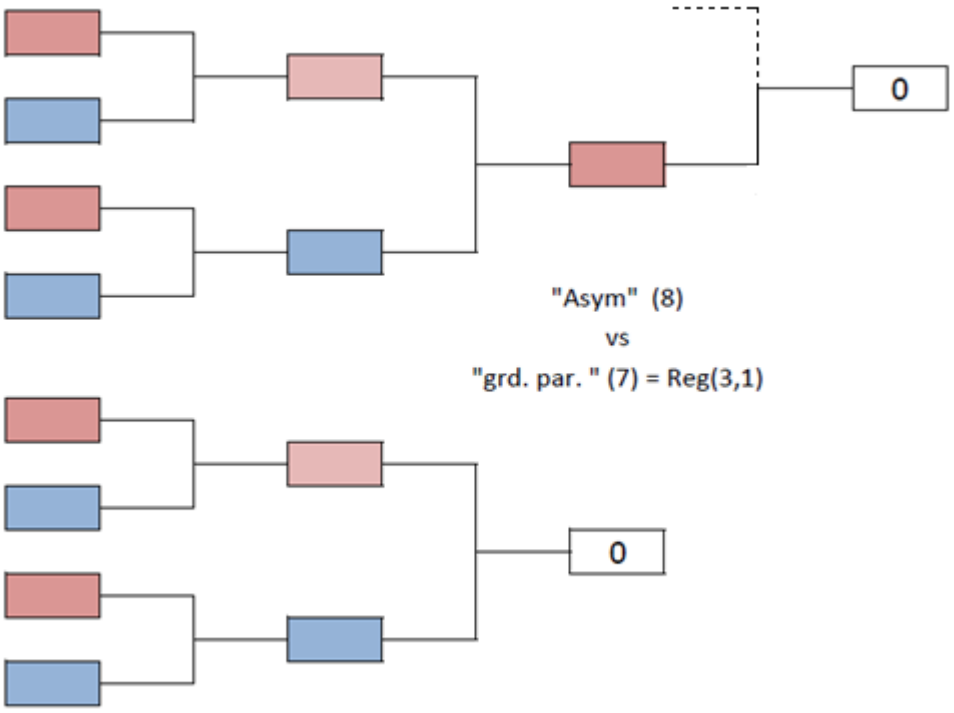


Figure 6: Asymmetry created between two pedigrees Reg(3, 1) by moving individual 0 down of one generation and omitting his father.

In the previous diagrams, individual 0, for which the mutation risk is evaluated, is positioned on the right and his risk is calculated according to individuals of the same or previous generations represented at his left.

One might question to what extent descendants can also provide valid information about an individual's genotype. Of course, this can be informative only if an individual is old enough to have children and these children should not be too young themselves. This may limit its practical value, but this problem remains an interesting question. In particular, for a fixed pedigree size, do ascendants give more information than descendants?

Recall that to obtain our numerical results we simulate families with inevitably a choice on the law of the birth year by generation. Here, the years of birth are chosen so that the (potential) age of the individuals is uniform in [30, 50], [50, 70] and [70, 90] respectively for the first 3 generations. We use the word “potential” because the individual may in fact have died earlier, either from the disease K studied or from another cause. It is essential to take age into account because the likelihood of having contracted the disease K before a given date depends on how long the individual has lived. Consequently, "young" individuals, for example of the first generation, *a priori* provide little information, which suggests that prediction with ascendants is more efficient than that with descendants. To carry out this comparison, we will compare the predictors associated with pedigrees "Grd Par." = Reg(3, 1) including 7 individuals and “Grt Grd Par." = Reg(4, 1) containing 15 individuals respectively to those calculated with the following pedigrees:

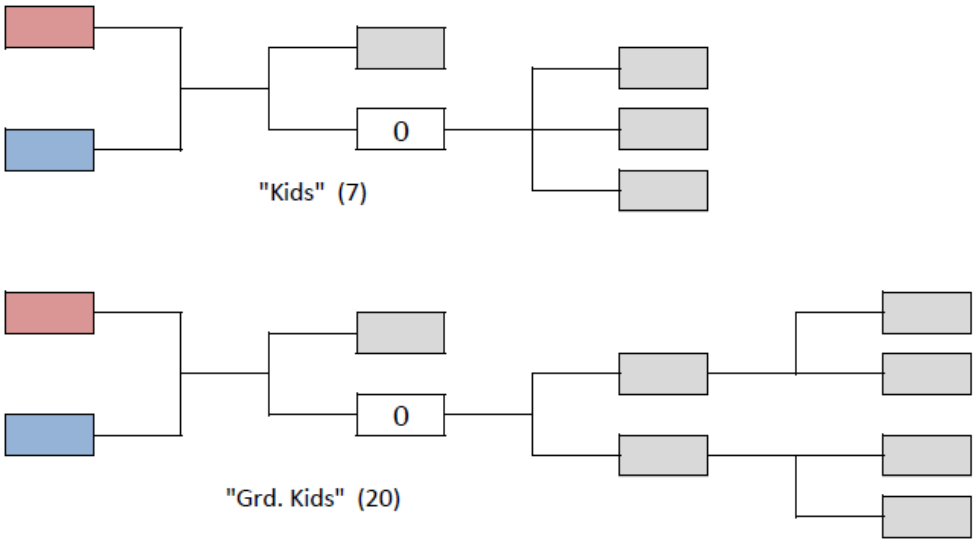


Figure 7: Two kinds of pedigree with one or two descending generations

2.3 The importance of conditioning

Nowadays, the individual mutation diagnosis is primarily offered to families considered at risk and a family is considered as such if one or even several cases of

sickness K have occurred. This induces a strong bias on data in present databases such as the ones collected by anti-cancer centers. In particular, estimating some parameters in these databases will lead to values radically different from the ones for the general population. In the future, larger and less biased databases will probably emerge and risk calculation will be computed systematically as a prevention approach. Consequently, in the following numerical experiments we will consider both the case of unconstrained family and the case of families conditioned to have (at least) a certain number of cases of illness. The performances and ranking of the predictors can change significantly between the two cases.

2.4 Notation and formulas used in software

To ensure reproducibility of our work, our demonstration software (in Python) can be downloaded at the address: https://drive.uca.fr/f/a07b89c2df52460686bf/?dl=1

With minor adaptations this software can execute most of the experiments that follow.

To complete the mathematical description of the risk computation algorithm and make the software clear, we give in the present subsection the necessary notation and the explicit formulas for the laws (likelihood) of the random objects involved.

2.4.1 Coding the family tree structure

Mathematically a family tree is a set of n (related) individuals, numbered from 1 to n, for whom data have been collected. The genealogical relations between these individuals are modeled —both in the formulas below and in the companion software— by two vectors $(f(i), 1 \leq i \leq n)$ and $(m(i), 1 \leq i \leq n)$ where, for $1 \leq i \leq n$, the value $f(i)$ [resp. $m(i)$] indicates the index of the father [resp. mother] of the individual of index i, when these parents belong to the family tree i.e. the set of individuals for whom data are available. Otherwise, we set $f(i) = -1$ if the father of the individual i is out of the tree and we write $i \in U_f$. Similarly, $i \in U_m$ if $m(i) = -1$, meaning the mother of i is out of the tree.

2.4.2 Genotype: law and generation

For a given family tree, the genotype $G = (G(i), 1 \leq i \leq n)$ is the collection of the genotypes of all individuals in the tree. For $1 \leq i \leq n$, we have $G(i) = 1$ *or* 0 according to the individual being mutated or wild-type, respectively. The only parameter useful in the generation and the expression of the law of the genotype is p_m, the probability for an individual drawn at random in the population to be mutated. This is used for the genotype of individuals whose parents are out of the tree. For the other individuals, the probability of being mutated is expressed conditionally to the state of their parents, as prescribed by Mendel laws. To this end, we introduce the 3×2 matrix

$$M = \begin{pmatrix} 1 & 0 \\ 0.5 & 0.5 \\ 0.25 & 0.75 \end{pmatrix} = (M[i,j],\ 0 \leq i \leq 2,\ 0 \leq j \leq 1)$$

which gives the probability for an individual to be wild-type (column 0) or mutated (column 1) according to the number of mutated parents: 0 for line 0, 1 for line 1, 2 for line 2. Recall Table 1 for justifications. Taking into account the previous remarks, the law of the family genotype G is given by:

$$\begin{aligned}
&\mathbb{P}(G = (g(i),\ 1 \leq i \leq n)) \\
&= \prod_{i \in U_f \cap U_m} (p_m\ g(i) + (1 - p_m)\ (1 - g(i))) \\
&\times \prod_{i \notin U_f \cup U_m} M[g(f(i)) + g(m(i)),\ g(i)] \\
&\times \prod_{i \in U_m \setminus U_f} (p_m\ M[g(f(i)) + 1,\ g(i)] + (1 - p_m)\ M[g(f(i)),\ g(i)]) \\
&\times \prod_{i \in U_f \setminus U_m} (p_m\ M[g(m(i)) + 1,\ g(i)] + (1 - p_m)\ M[g(m(i)),\ g(i)])
\end{aligned}$$

Note that this probability can be zero since the matrix M contains a null term corresponding to the probability for an individual to be mutated when neither of his parents are.

2.4.3 Phenotype: law and generation

The family phenotype

$$(F = (F(i), \quad 1 \leq i\ \leq n)$$

is the collection of the phenotypes of the n individuals under consideration and, in the present work, these phenotypes are reduced to the occurrence or not of a cancer and, in the positive case, the age of onset (which in reality is more often the age of diagnosis). The phenotype of one individual is thus the age of diagnosis between 0 and 99 or the value 200 meaning by convention that no cancer has occurred.

We use the two parameters introduced earlier: p_1[resp. p_0] is the probability that a mutated [resp. wild-type] individual develops a cancer throughout his life. Given that an individual does develop a cancer, the age of diagnosis is given by one of two probability laws $(d_g(t), 0 \leq t \leq 99)$ depending on whether the individual is mutated $g = 1$ or not $g = 0$. The curves we use in practice for our numerical studies are drawn in Figure 1. The ages $(a(i), 1 \leq i \leq n)$ of the family members are randomly generated throughout the tree in a coherent manner as specified before. For old generations, age has to be understood as potential age and is set to 99.

With the previous notations, the conditional law of the phenotype F knowing the genotype $G = g$ can be expressed as

$$\mathbb{P}\left(F = (k(i),\ 1 \le i \le n) \mid G = (g(i),\ 1 \le i \le n)\right)$$
$$= \prod_{i=1}^{n} \left[\mathbf{1}_{\{k(i)=200\}} \left(1 - p_{g(i)} \sum_{j=0}^{a(i)} d_{g(i)}(j) \right) + \mathbf{1}_{\{k(i)\le a(i)\}}\ p_{g(i)}\ d_{g(i)}(k(i)) \right]$$

3 Results

We now present results of numerical experimentations which address the questions stated in the introduction. For most of them a choice of parameter values had to be made. Unless stated otherwise, we chose $p_m = 0.03$, $p_0 = 0.02$, $p_1 = 0.7$. Although this appears arbitrary, we have carried out enough experimentations to be convinced that the phenomena we emphasize hold for other "typical" values of the parameters. Moreover, the available software enables the reader to check by himself for any set of values.

3.1 Reliability of the inference: confidence intervals

In what follows we will compare different predictors on the basis of their ROC curves and in particular the area under these curves (AUC). These ROC curves are obtained after hours of intensive computation but are still submitted to the randomness of the samples drawn to construct them. In other words, the AUC is affected by a random noise which has two distinct origins: first the sampling of families on which the predictors are tested, second the computation of the predictor itself when using simulation to evaluate the conditional probability. Note that this second source of randomness disappears when using a brute force method.

Consequently, it would be safe to obtain precise confidence intervals for all the AUC that we will provide but that seems too ambitious regarding the computation time needed. In particular, the noise on the AUC does not seem to be close to Gaussian with the current sample sizes and real confidence intervals are much wider than Gaussian ones. Consequently we have chosen to evaluate confidence intervals in the specific case of the Reg(3,2) tree which allows rather quick computation both by brute force or simulation method. The size of the confidence intervals in this case will allows us to calibrate the software in terms of numbers of simulations, taking also into account the limits in computing time and it will guide us on the conclusions that can be drawn from the other ROC curves.

Below are the graphs produced by the demonstration software in its original form. It gives the performance of the mutation predictor for a regular family tree with 3 generations and 2 children per couple. It shows the ROC curve i.e. resulting from the sensitivity and specificity of the predictor to assign individual 0 to its correct group (i.e. mutated or not). Since this family contains 10 individuals, there are 2^{10}

genotypes (including the inadmissible ones) so a brute force method for the computation of the conditional probability is possible. Also, as we have already mentioned, conditioning the family by a certain number of occurrences of K makes sense and we will tell more on this subject in the next subsection. For the moment we consider both the unconditioned case and the case of at least one occurrence of K. Hence the figure below shows four graphs corresponding to the different cases with brute force method represented by continuous lines and simulation method represented by crosses.

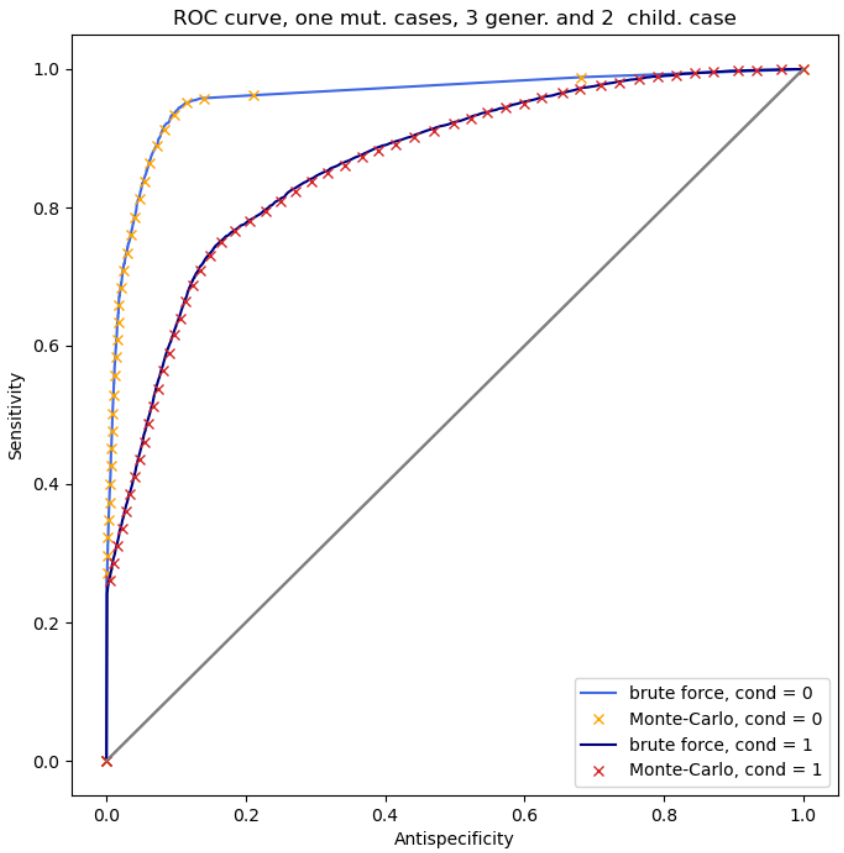


Figure 8: ROC curves for the mutation predictor; context: Reg(3,2) family tree conditioned by at least one case of K or unconditioned.

First, we see that the brute force and the simulation methods give almost exactly the same results. This is a reassuring fact about the accuracy of the simulation method recalling that, for most of the cases we will study, it will be the only method available. We performed 50 000 simulations of the family tree and we imposed, for the simulation method based on Formula (2), that both evaluations of the denominator and the numerator in (2) get a least 5000 contributions. These are the numbers we will use by default for the other ROC curves in this article. For large families, it is close to the limit imposed by the computation time.

In the unconditioned case represented above, the empirical confidence interval for

the AUC based on 100 attempts is [0.949, 0.959] for a confidence level of 90%. The conditioned case shows a confidence interval of similar length. This means that ranking predictors can safely be made when AUC show differences typically larger than 0.02. This will often be the case in the following experiments, but not always. For tighter differences on AUCs, we may simply conclude that the corresponding predictors have quite similar performances which is certainly sufficient in practice but if a ranking is nevertheless desired, we will raise the number of simulations or multiply the number of parallel computations.

3.2 Influence of conditioning

The performance of a predictor evolves when we condition the total number of ill people in the family. For instance, for a family Reg(4, 2) with 22 individuals, that allows very good performances (as we will see later), we "force" the minimum number of patients successively to 0, 1, 2 or 3 and we obtain the following ROC curves:

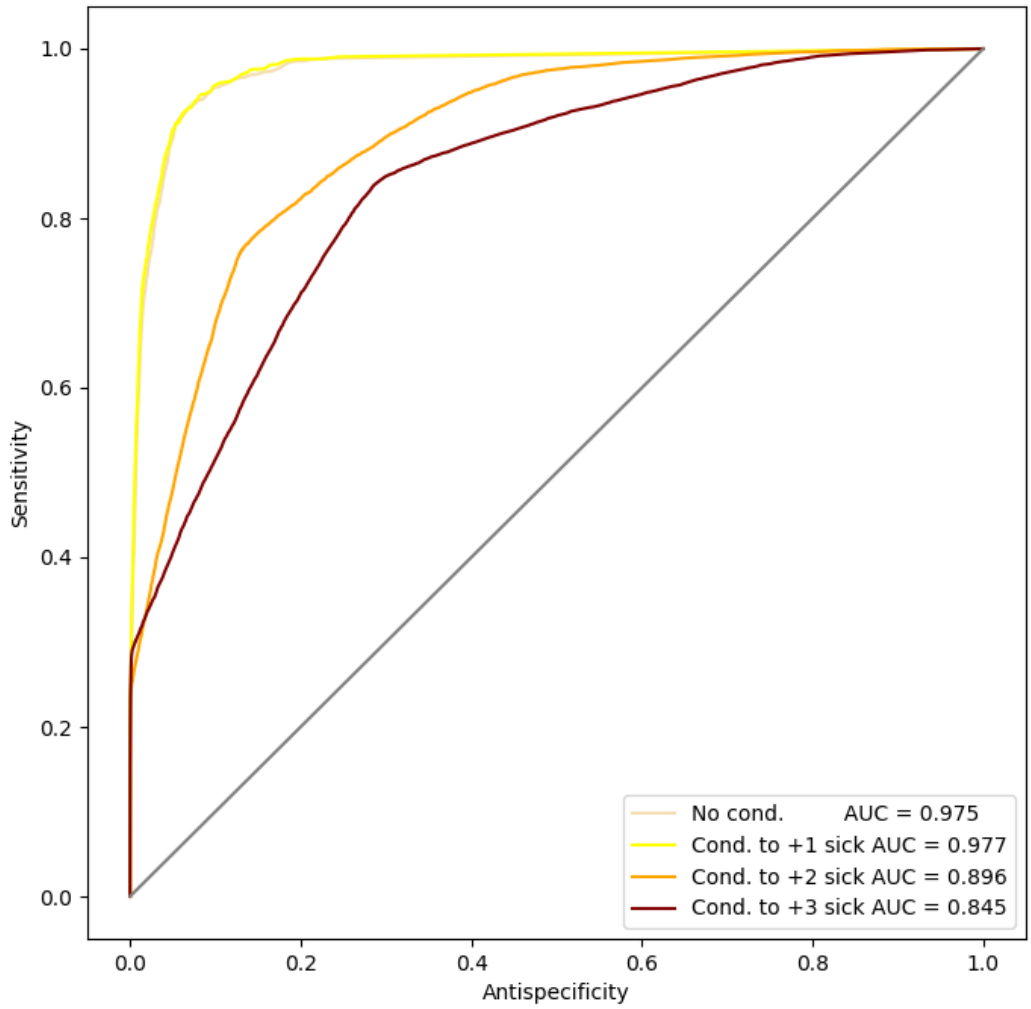


Figure 9: Performance of predictor depending on the conditioning of a minimal number of disease case per family (context: Reg(4, 2) pedigree, one deleterious mutation)

The first two curves overlap and hardly stand out. We notice that beyond 1 case of disease, conditioning decreases the effectiveness of the predictor, which in the end

is quite logical because it becomes more and more difficult for the model to distinguish the mutated families from the others as they all systematically present cases of illness.

3.3 Influence of the generation number

We consider regular pedigrees Reg(n_g, 1) for a number of n_g generations varying from 2 to 6 and a fixed number of children n_e per couple equal to 1. The ROC curves for the associated predictors are as follows:

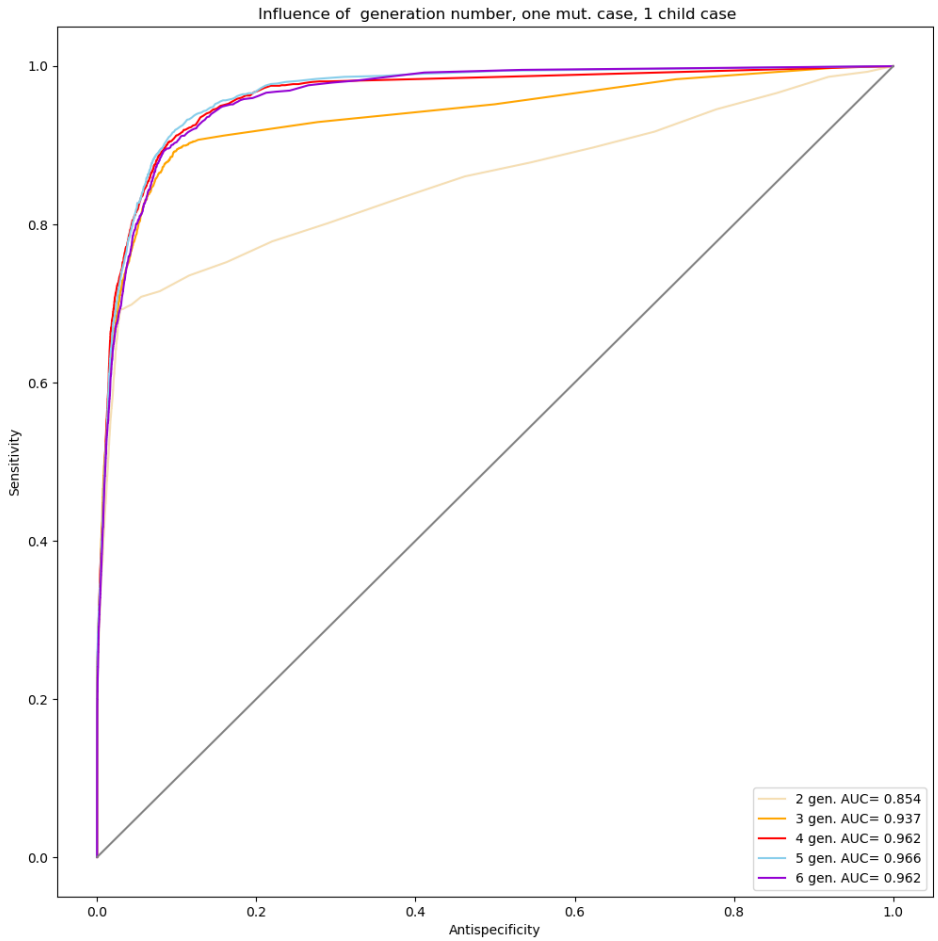


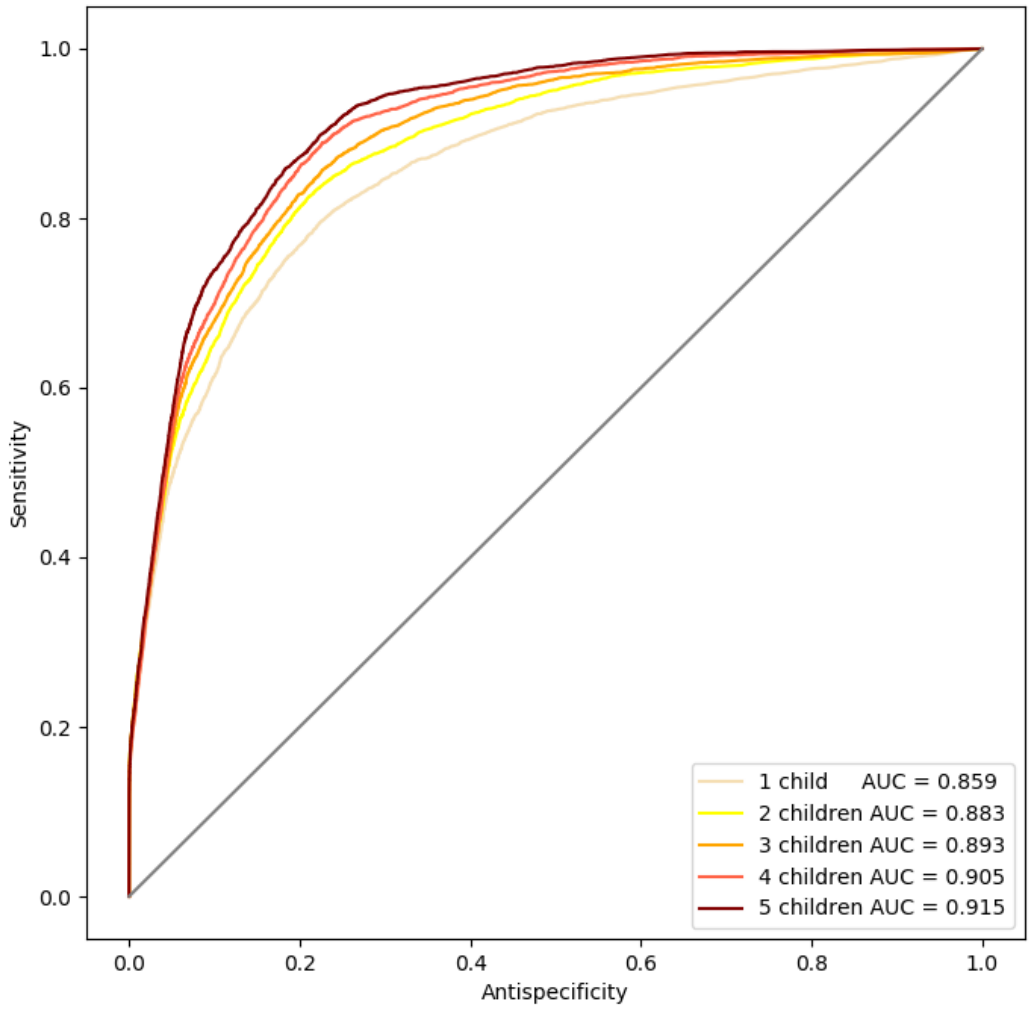
Figure 10: Influence of the generation number on the model performance (context: one mutation, conditioning = 0 and 1 children per couple).

Obviously, the data from only 2 generations, i.e. the individual and his parents are not sufficient to obtain a successful prognosis. Rising to 3 generations allows a significant gain and the addition of the 4th generation still provides a small improvement. Beyond, any addition of information from previous generations is useless or even harmful: performance seems to regress slightly with the 6th

generation. This is confirmed by performing 20 independent attempts: the AUC for 6 generations is less than the AUC for 5 generation (yet infinitesimally) for all attempts. This phenomenon also holds when the families are conditioned to at least one occurrence of the illness and essentially the same thing happens when we set the number of children per couple at 2. Our interpretation is that too distant individuals bring more randomness than information.

3.4 Number of children

Investigating the performance according to the number of children in pedigrees is more theoretical than the previous issue. Indeed, in the previous case, the doctor / statistician wishing to make a prognosis can choose the number of generations that he provides. For children, their existence, a fact, cannot be the object of a choice, even if the practitioner can choose to include only one child: the one who is in the ascending line of the individual studied (ex. the index case[1]). The ROC curves obtained for the families Reg(4, n_e) where n_e varies between 1 and 5 are as follows:



[1] A case in a family is called "index" if he is the first family member who comes to consultation. Most often, he has reported a cancer that corresponds to the familial syndrome, although it is not mandatory.

Figure 11: Influence of children number per couple on model performance; context: one mutation and 4 generations.

We observe that the larger the pedigree, the better the prediction with a significant gain added between one child and two children. Beyond two children the gain is lower. This means that adding information from individuals directly connected to the lineage is beneficial. However, since the AUC increases slightly less as the number of children increases, there may be a limit at which the performance of the predictors peaks.

3.5 Height against width

We have just seen that the predictors performances increase with the number of generations up to 4 (height) and also with the number of children per couple (width). Hence the question: for the same size (same number of individuals), is it better to have a large pedigree or a tall one? If we refer to the following ROC curves relating to pedigrees "Grt Grd Par." and "Wide 3 gen." of 15 individuals each, it would seem that a large pedigree does slightly better than a tall one.

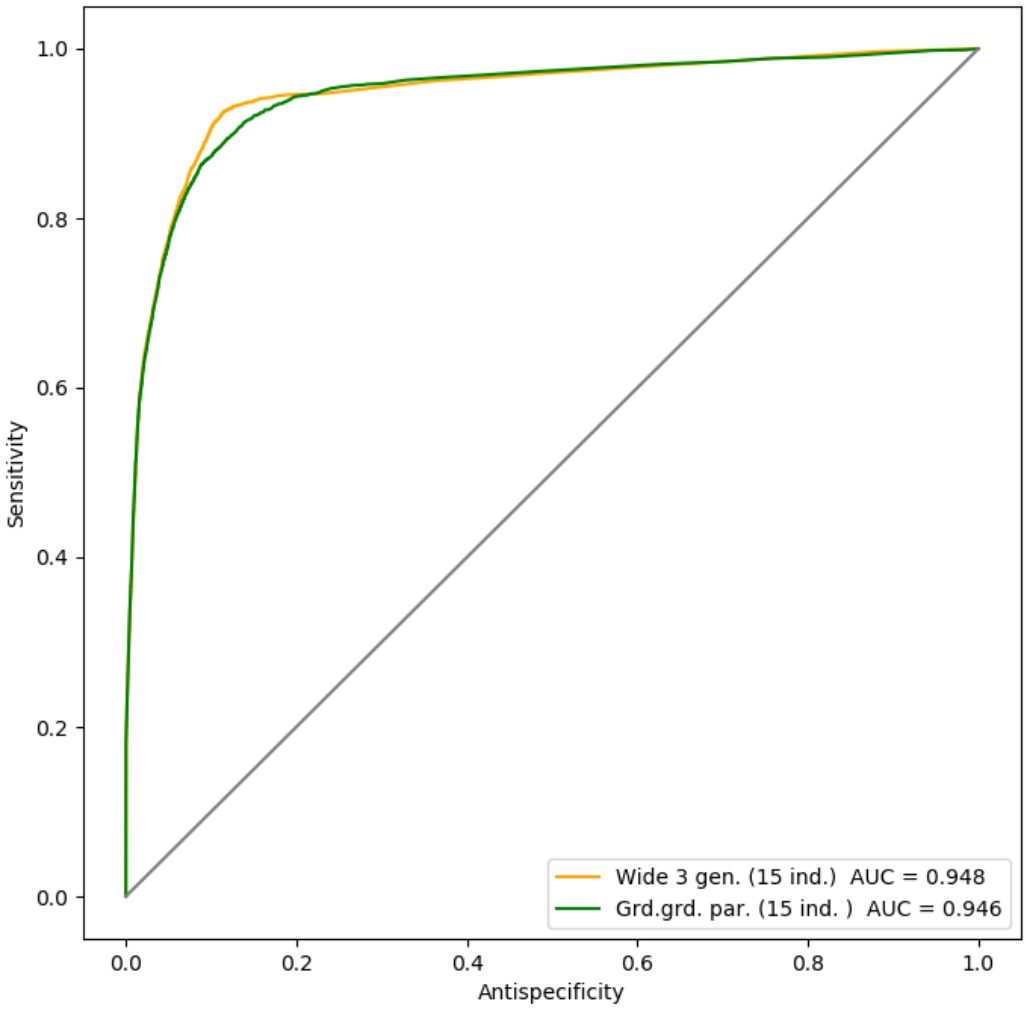


Figure 12: Performance comparison of two pedigrees of 15 members, one large and the other one high

However, conditioning the presence of at least one disease case per pedigree is sufficient to produce a markedly opposite result:

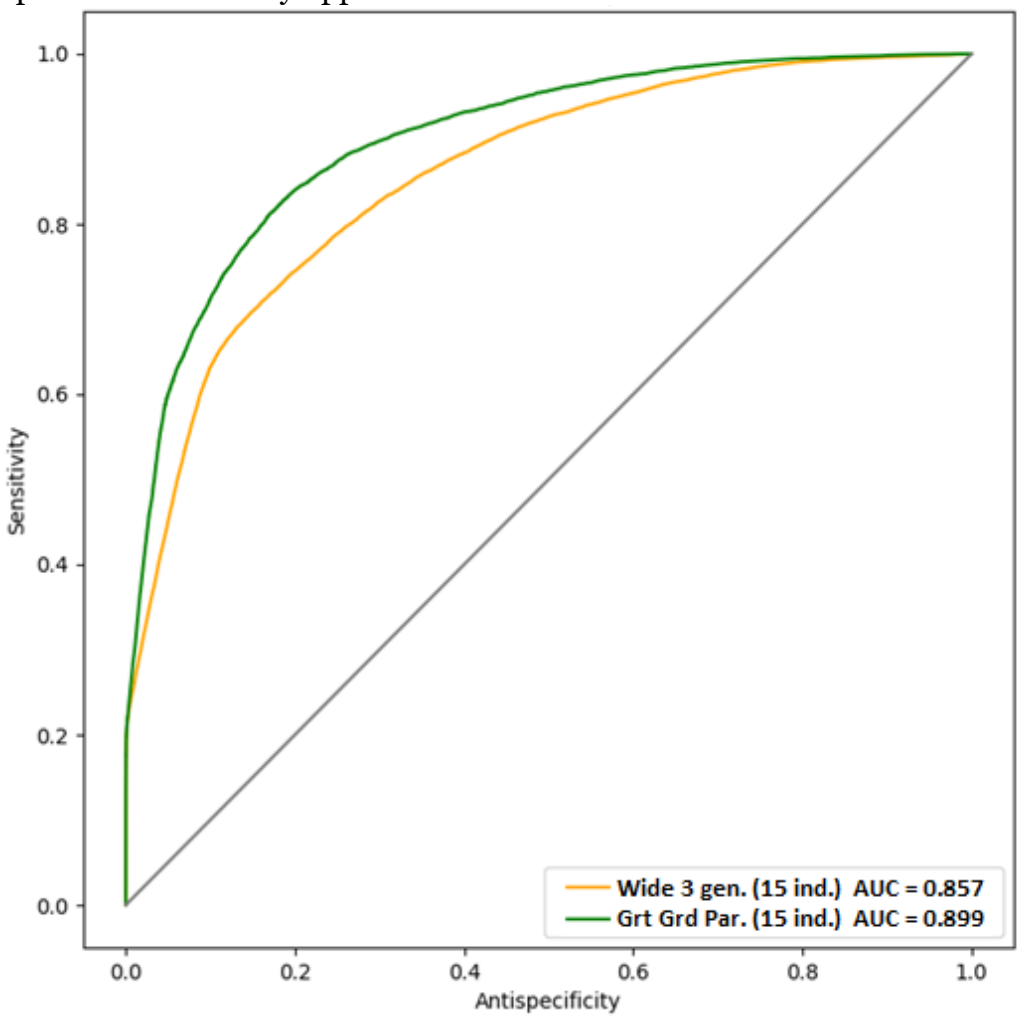


Figure 13: Performance comparison of two pedigrees of 15 members, one large and the other one high but constraining the presence of at least one disease case per pedigree

Given the presence of cancer cases, either sporadic or familial, in pedigrees usually observed in oncogenetics, we can estimate that the second case is more relevant. Therefore, with an identical number of members, the height of a pedigree seems more informative than its width, even if, as seen previously, beyond 4 generations compiling information does not bring anymore benefit.

3.6 Utility of cousins

We have previously shown the value of the information corresponding to individuals directly connected to the ancestral line of the individual studied. It makes sense now to study the case of somewhat more distant relatives, starting with cousins. We thus compare the predictors associated with the family "uncles" and "cousins" among themselves but also with the pedigree Reg(3, 1) without uncle or cousin:

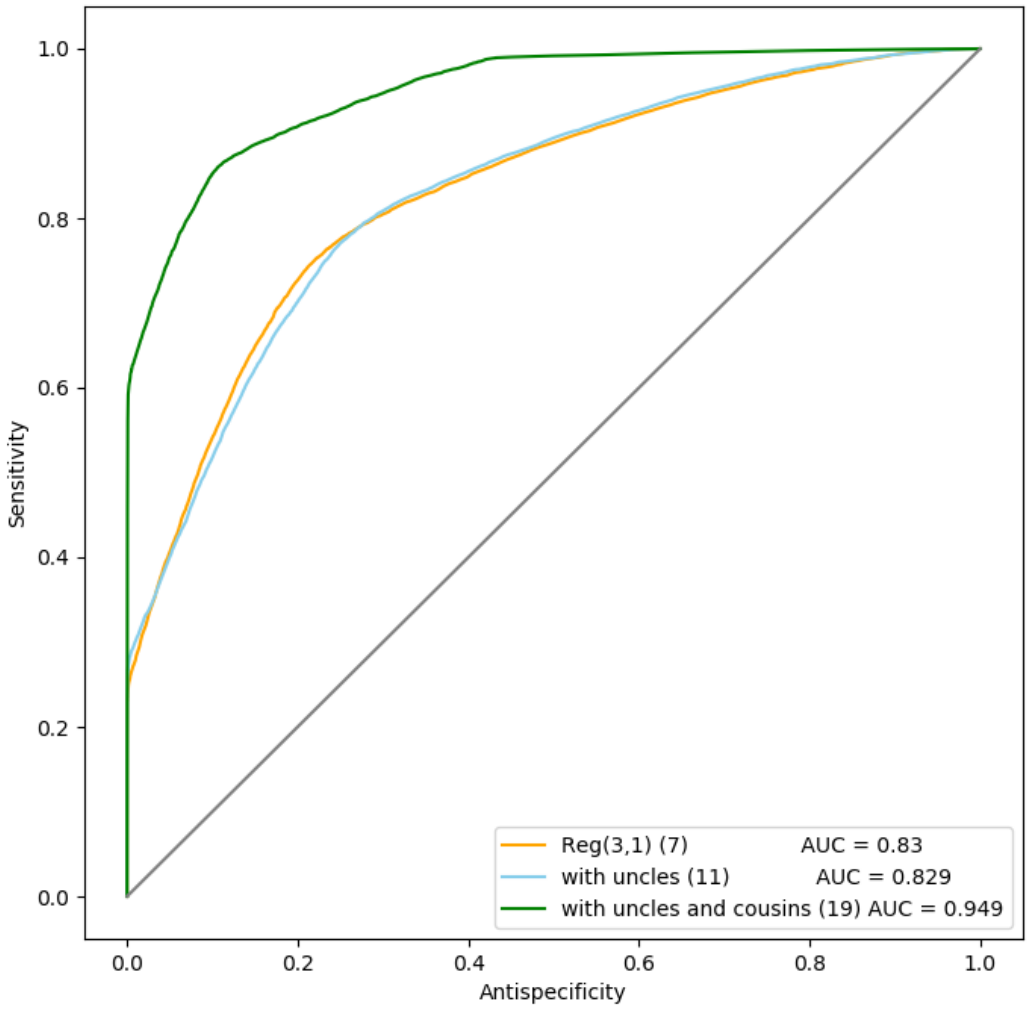


Figure 14: Influence on predictors of the presence of uncles and cousins in pedigrees (context: 1 mutation, 3 generations, 1 child per couple, except when adding uncles and cousins, and condit. of 1 disease case)

This figure is interesting because it shows that indirect links can also be very informative – this is echoing the conclusion as to the number of children per couple. If the addition of uncles does not improve ROC AUC, the addition of cousins when you have uncles, brings clearly contributing elements. This could be explained by the fact that if a cancer is found in cousins, usually rather young, these cancers are very informative about a possible familial mutation.

3.7 Asymmetric pedigrees

We study the extreme case where all phenotypic information on either the father or the mother and all their ancestry is missing. Is it then illusory to hope for a mutation prognosis? We compare the associated predictors of the following three pedigrees:

- The “Asym” family made up of the father and the paternal grand and great-grandparents
- The “Grt grd par." with grand and great-grandparents on both sides.

- The “Grd par." with grandparents but no great-grandparents on both sides.

See Figures 3 and 6 for a visualization of these families. The following ROC curves are obtained:

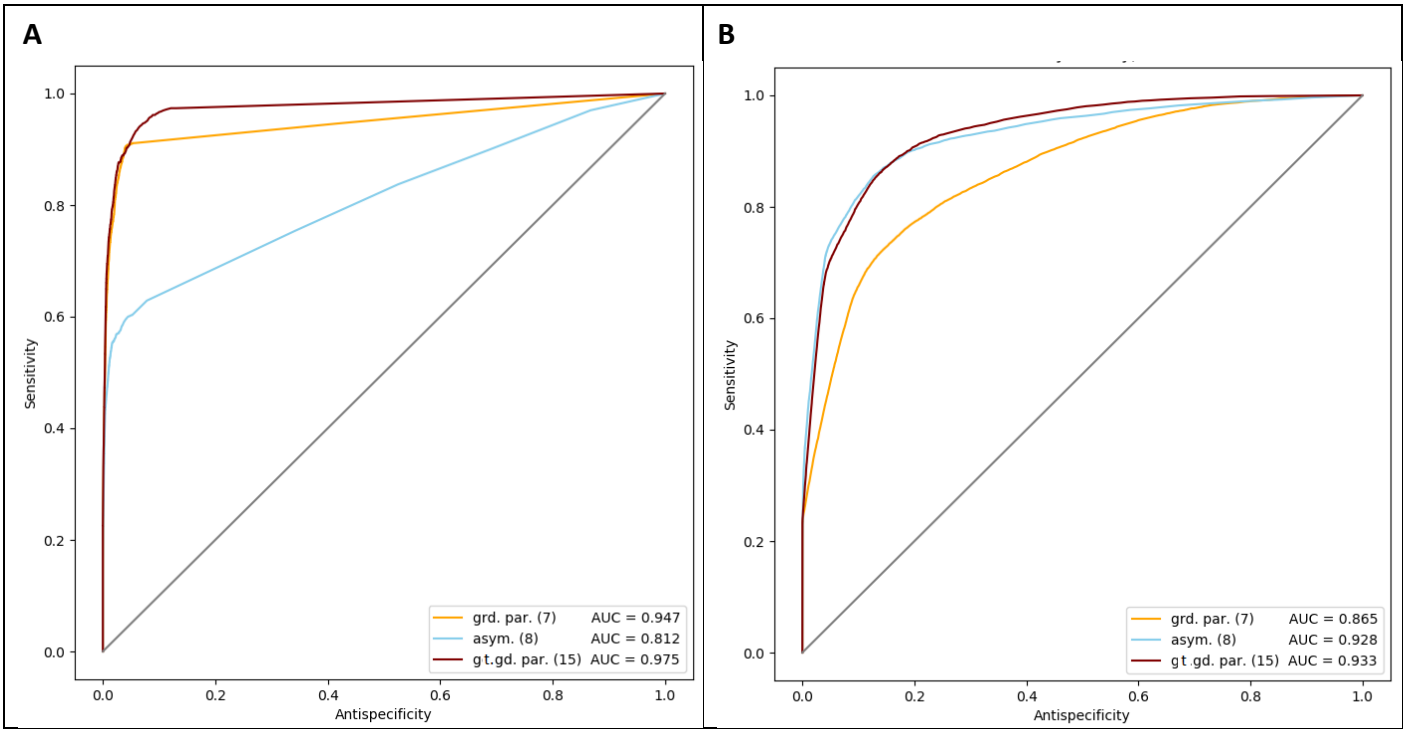


Figure 15: Influence of asymmetry on predictors performance
(context: 1 mutation and one child per couple; A - no condit. and B - condit. of 1 case)

Depending on the conditioning, the loss of symmetry (i.e. a loss of one familial branch) may induce or not a significant loss of information. When we do not force the presence of a case of disease (0-conditionning), the loss of a branch can remove the pivotal genetic information. This is the case of figure 15-A: the blue ROC curve is much less sensitive than the orange one although the number of members is similar. On the contrary, when the information has to appear in the pedigree, thus in the remaining branch, the other one becomes less necessary, and the pedigree works as well as a pedigree containing its two branches (figure 15-B). For the oncogeneticist, knowing only one branch of a pedigree may be sufficient as long as this branch includes at least one cancer case.

3.8 Ascendants against descendants

We previously suggested that the prediction using the descendants should be worse than the one including the ancestors since recent generations are less exposed to the disease risk. We tested this hypothesis and here is the result (A) in the standard context then (B) when we condition to at least one disease case per pedigree:

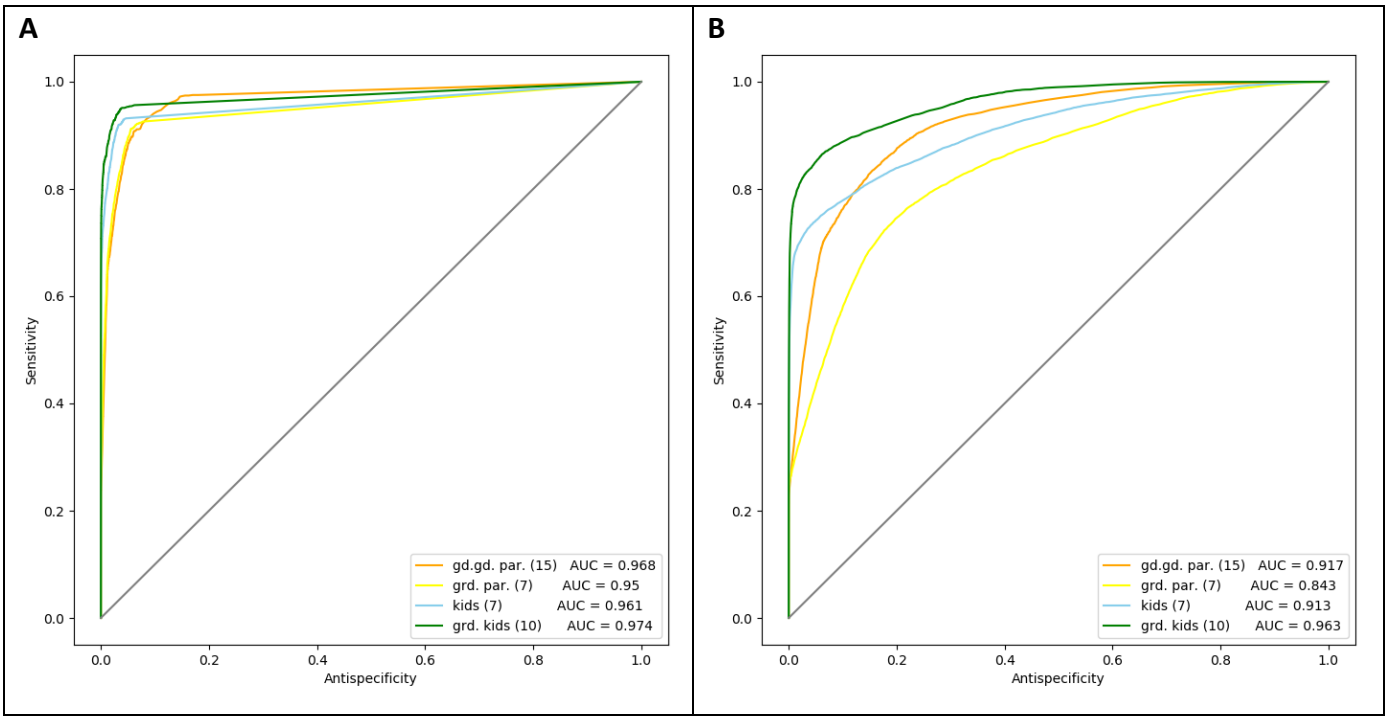


Figure 16: Comparison of predictors whether pedigrees contain rather ascending or descending generations and whether we condition (B) or not (A) the occurrence of at least one disease by pedigree.

These results contradict our initial guess. The prediction using children or grandchildren is better at identical number of members per pedigree. Perhaps this is happening because a pedigree containing children and/or young adults imposes a high age requirement on individual 0. Also, conditioning one cancer at least per family implies that they occur more likely at younger ages and therefore must be caused by a deleterious mutation.

3.9 Influence of parameters p_m, p_0 and p_1

For Reg(4, 2)-type pedigree containing 22 individuals, the five ROC curves below correspond to increasing mutation frequencies but respectively weaker and weaker areas under the curve (AUC). Thus, we see that the predictor is all the better, both in sensitivity and in specificity, as the prevalence of the mutation is low.

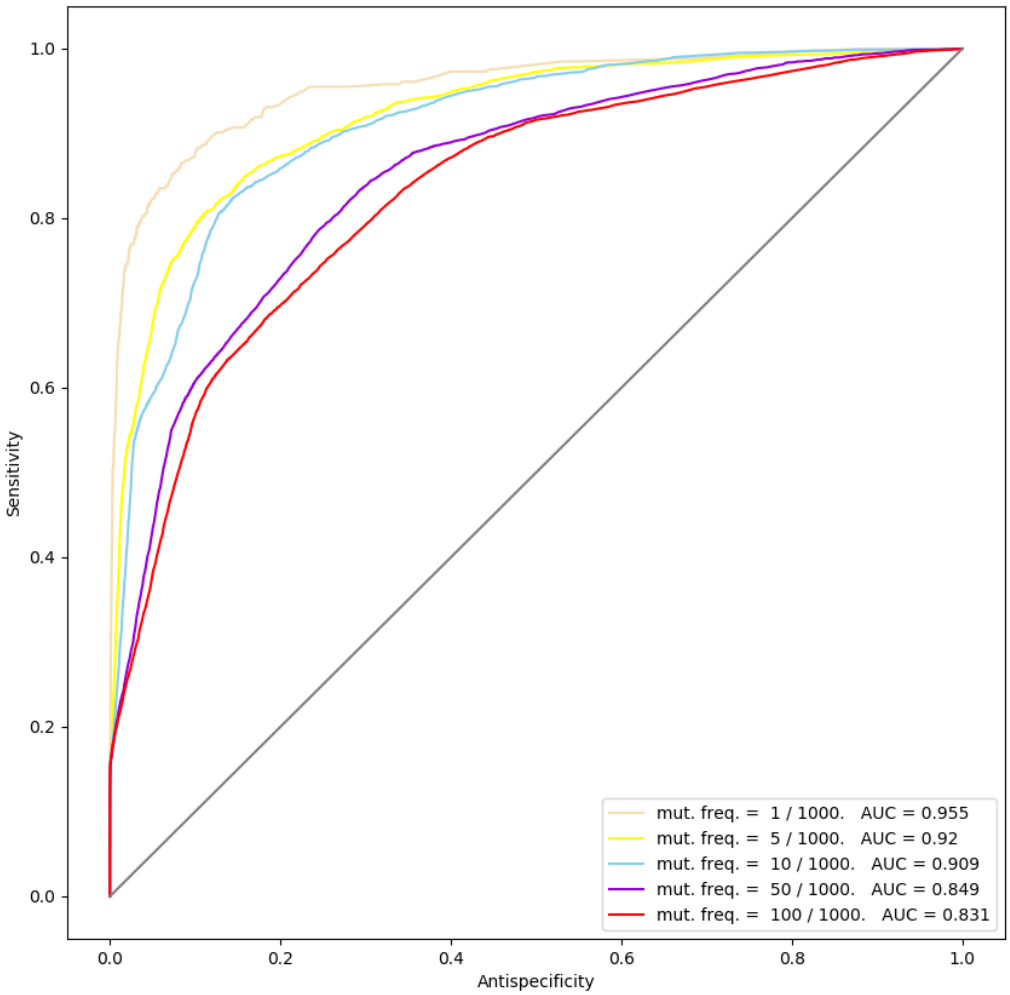


Figure 17: Predictors performance according to the mutation prevalence (context: 4 generation pedigree and 2 children per couple, thus 22 individuals, condit. of 1 case of disease))

Another parameter of the model is the penetrance of disease K in mutation carriers. We tested a penetrance p_1 equal to 20%, 40% and 70% (this latter similar to that of BRCA mutations). For its part, the p_0 risk of occurrence of the disease in non-mutated individuals (wild-type) during its lifetime can vary greatly and we have tested values ranging from 0.1% to 10%:

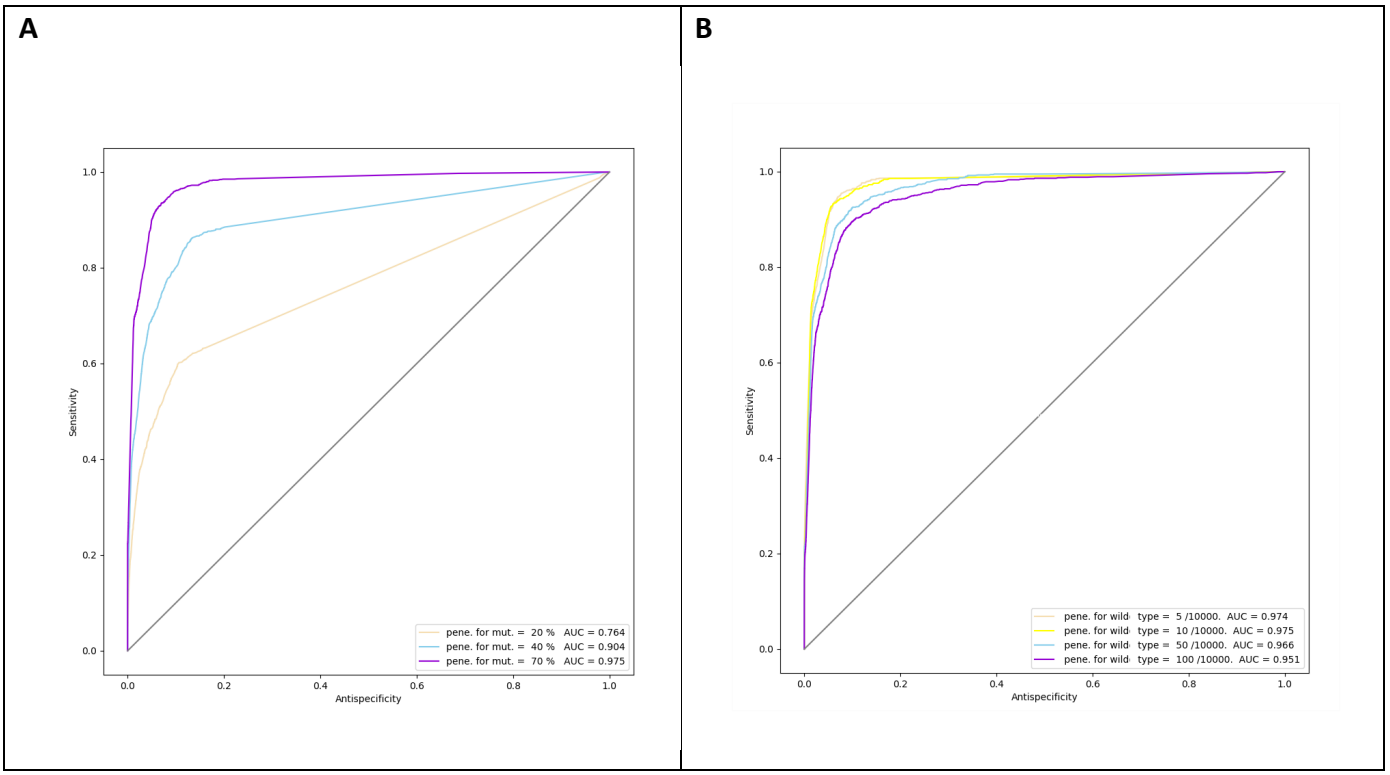


Figure 18: Influence on the prediction of the deleterious mutation penetrance (A) or its incidence in non-mutated individuals (B).

It appears that a high penetrance in case of mutation improves the quality of the predictor while conversely, a higher incidence of the disease in the general population decreases this quality. Clearly, the greater the gap between penetrance and incidence of disease at any age, the better the predictor becomes.

Overall, these effects are in line with what we would expect: the more the disease K is confined to mutated individuals, the easier the prediction. However, these parameters do not all have the same influence. The incidence of the disease in wild-type individuals is not very influential, while penetrance in mutation carriers has a strong impact on the performance of the prediction. The frequency of mutation is also very influential and its direction of variation seems less intuitive: the more frequent the mutation, the more difficult its detection.

3.10 Conclusion on the cost / effectiveness trade-off

Previous experiments with family pedigrees of growing size show that there is an optimal size for the prediction. This notion of optimum is reinforced by the increase in the computational cost with the pedigree size, which might even shift the practical optimum towards smaller sizes. A difference should also be made regarding the cost of collecting phenotypes, depending on whether the data is pre-existing in a database or collected specifically for mutation diagnosis. Typically, we

will seek to get information regarding parents, grandparents and great-grandparents. This ancestral line can be completed, albeit with a relative gain in performance. One should probably avoid going too far, as illustrated by the three ROC curves below:

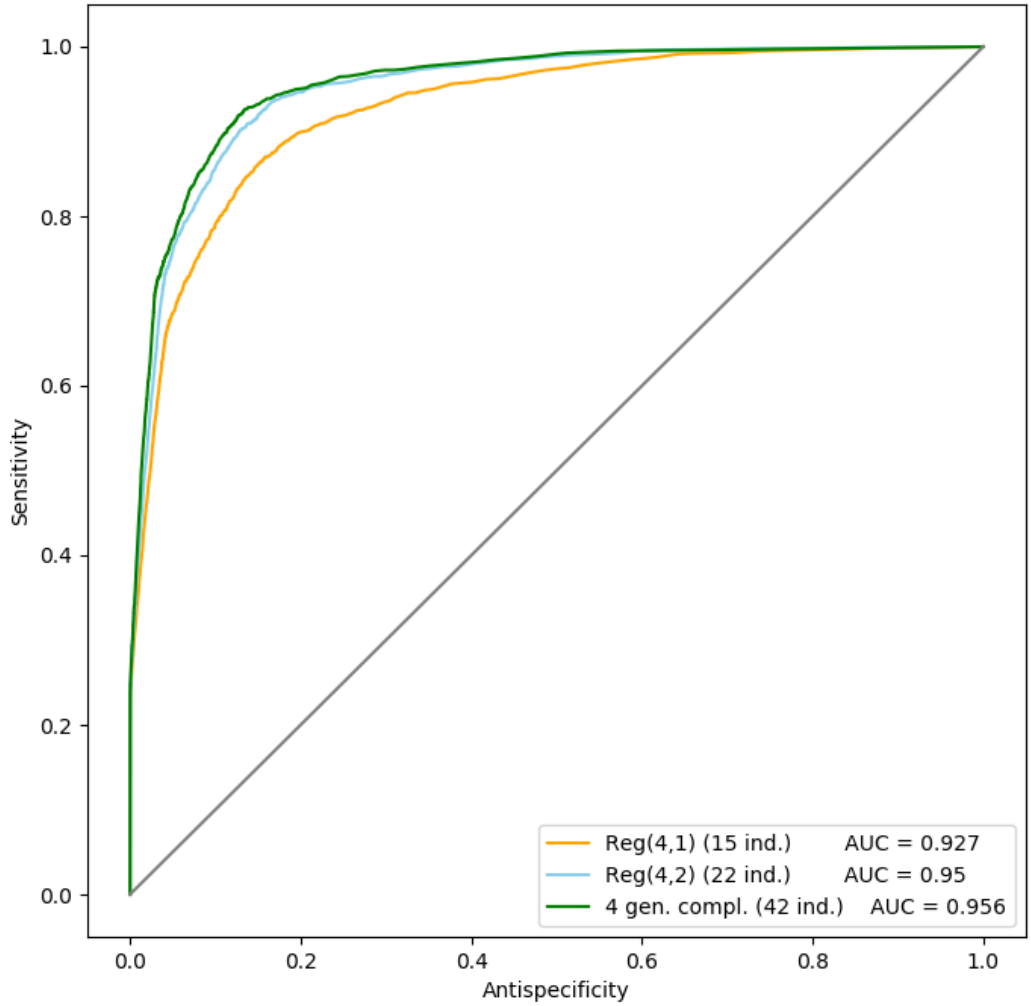


Figure 19: Predictor performance when pedigree size increases from 15 members to 22, then to 42.

We can observe in this figure that by adding the last 20 members no gain is obtained, which suggests that a very exhaustive search for collateral relatives may cost more than it helps. The three pedigrees concerned are represented nested in the following diagram:

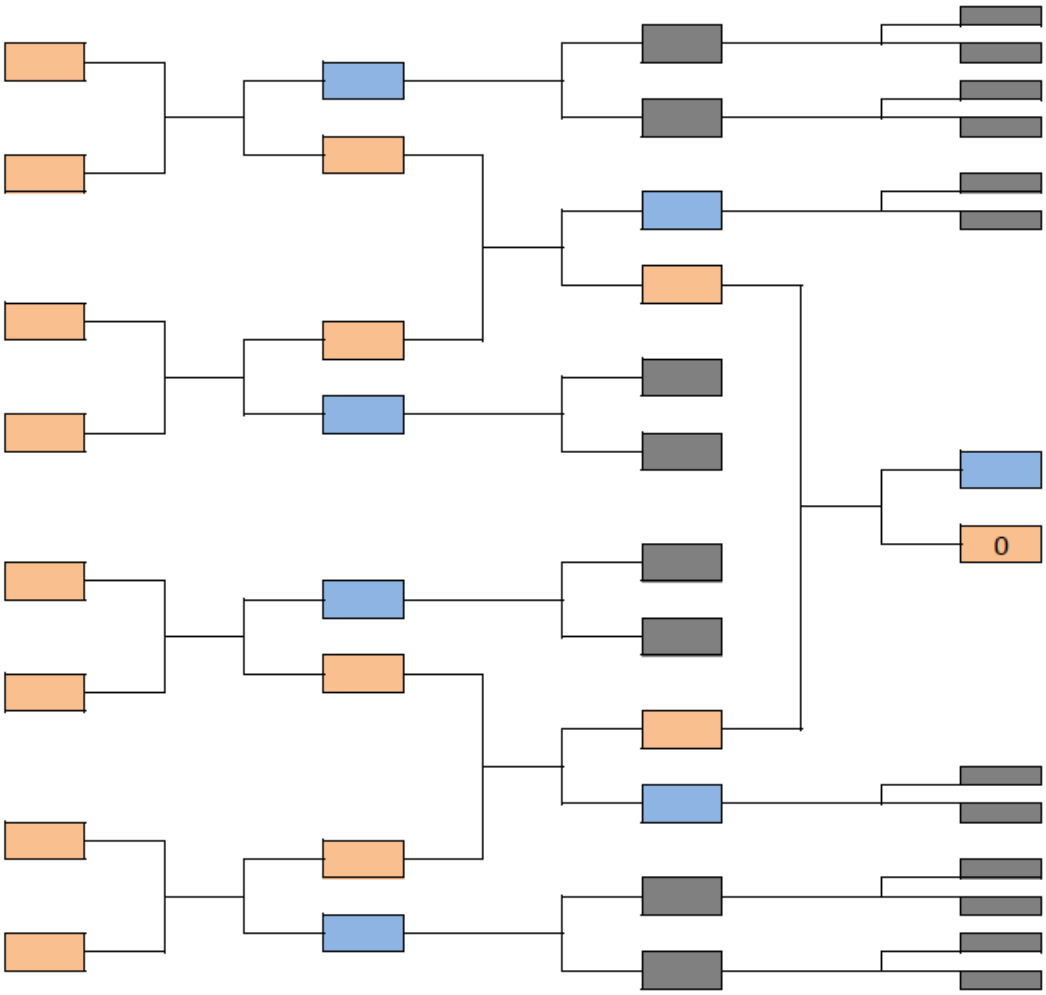


Figure 20: The 3 embedded pedigrees corresponding to the 3 previous ROC curves: 15 members in pink for the basal pedigree, + 7 members in blue for the second (N = 22) and finally + 20 members in grey for the last one (N = 42)

4 Generalization and extensions

4.1 Model with two mutations

In this model, two cross-effect mutations are implemented. We could probably call them “polymorphisms” because separately they have minimal penetrance and no impact on the age of onset, but when they are present together in the genome, the risk of cancer increases considerably as well as its precocity, with a penetrance similar to previous single mutation models. The phenotype is therefore regulated in a similar way, but it is not the presence of a mutation that generates the increased risk but the simultaneous presence of both.

The difference between the single and double models lies in the transmission of

mutations which, of course, obeys Mendel's laws, but results in different characteristics. Thus, in the double model we consider that the genome of an individual takes its values from a set of 4 elements:

{ wild-type, mutated 1, mutated 2, mutated 1 and 2 } = { (0, 0), (1, 0), (0, 1), (1, 1) }

For each mutation 1 or 2, the probability of transmission from parents to their children is still governed by Table 1. In the single mutation model, a child of two parents carrying the same heterozygous mutation has the probability 3/4 to be a carrier. In the double mutation model, a child whose two parents are doubly mutated (always heterozygous), now has only a probability $(3/4)^2 \approx 0.56$ of being in turn doubly mutated. Note that we continue to consider only heterozygous mutations by ignoring homozygous cases, previously thought to be lethal. But in the case of polymorphisms that are not frankly deleterious, the possibilities of homozygosity could legitimately also be considered. This would complicate the models[2] so much without likely providing any additional information that we have ruled them out.

The parameters of this double-mutation model are therefore slightly more numerous, namely:

- The frequency of mutation 1 in the general population, f_{mut1}
- The frequency of mutation 2 in the general population, f_{mut2}
- The penetrance of K for non-mutated subjects, p_0
- The penetrance of K for carriers of mutation 1 only, p_1
- The penetrance of K for the carriers of mutation 2 only, p_2
- The penetrance of K for doubly mutated subjects, $p_{1,2}$

In the cases studied below, the values of p_0, p_1 and p_2 are close and low while the value of $p_{1,2}$ is high. Besides, the law of the diagnosis age was assumed identical, for carriers of a double mutation, to the case of carriers of a single mutation in the previous sections. The problem therefore comes down to the prediction of the doubly mutated state which is the only one to constitute a significant risk of disease. The double mutation probability formula is an obvious adaptation of the single mutation model (1).

(3) $$P\left(G(i) = (1,1) \mid F\right) = \frac{1}{1 + \frac{numerator}{denominator}}$$

where

$$numerato\,r = \sum_{g:g(i)\neq(1,1)} P(F|G = g)\, P(G = g)$$

and

$$denominator = \sum_{g:g(i)=(1,1)} P(F|G = g)\, P(G = g)$$

[2] In particular, it would be necessary to rule on the lethal nature or not of a double homozygosity, on a difference in penetrance depending on whether one of the two mutations is homozygous or not, etc.

In this two-mutation model, the number of genotypes to be scanned in the brute-force method is 4^N where N denotes the size of the pedigree. So, except for extremely small trees, the simulation method should be preferred and can be written as an obvious variant of (2).

The performances obtained in this model with two mutations are overall a little lower than in the one-mutation model. But the phenomena described above persist. In particular, performance reaches a practical optimum around fifteen individuals as shown by the little gain in efficiency when going to 22 members, after what a loss is observed when going to 42 members:

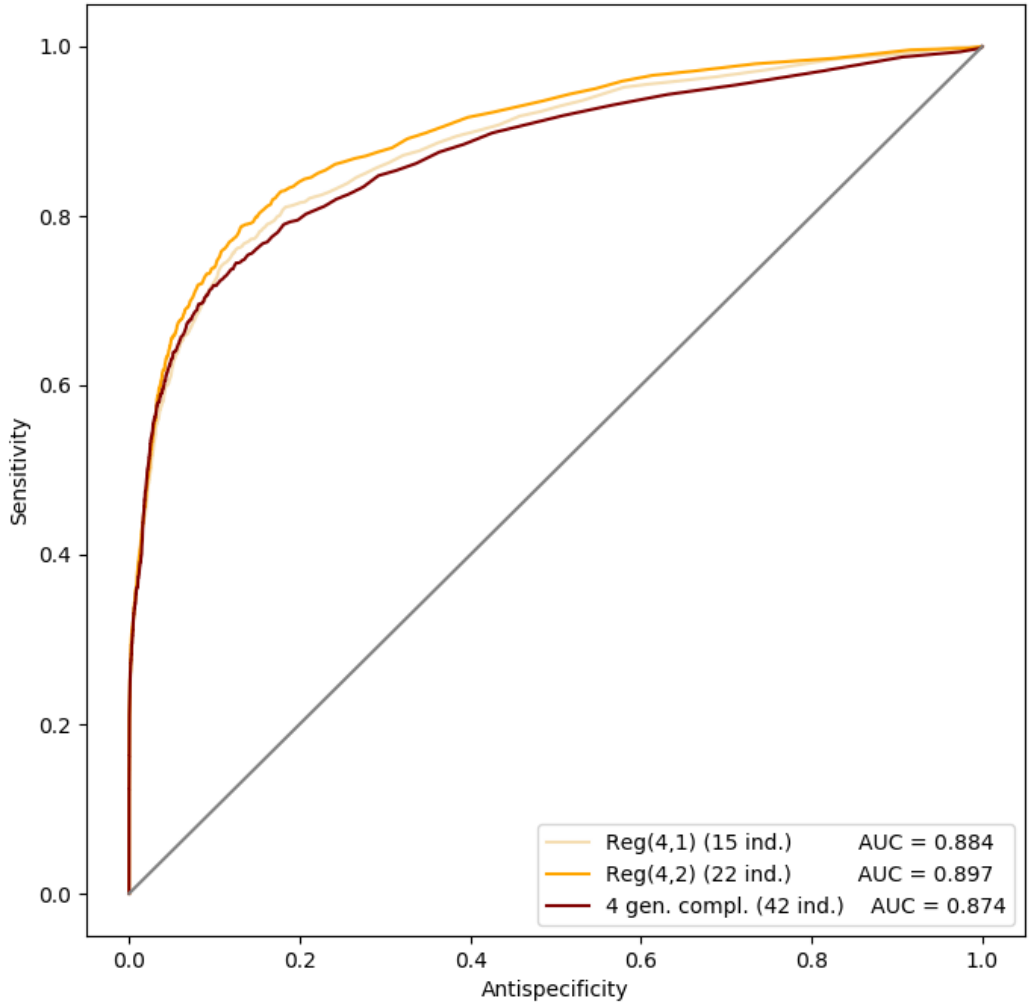


Figure 21: Comparison of the predictor performance according to the number of members included in the pedigrees; context: 2 mutations interacting and conditioning of at least one disease case per family

4.2 Other predictors

Calculating the probability of the genotype conditional on the phenotype is the most reliable estimate of an individual's risk of mutation. However, other predictors naturally come to mind. Given the popularity of likelihood maximization methods, one predictor might be the genome value that maximizes the likelihood of a phenotype. For small families, this maximization calculation can be done by brute-force, but for more realistic sizes, it is necessary to move towards a simulated annealing algorithm. Let us consider the case of the single mutation model and, as always, denote by F [resp. G] the variable which compiles the phenotypes [resp. genotypes] of individuals in the family. We still rate 0 the individual for whom the prognosis is to be made. Let us denote by $\hat{G}_0$ and $\hat{G}_1$ respectively the genotypes which maximize the likelihood of the phenotype with the individual 0 non-mutated or mutated respectively, i.e.:

$$L(F|G = \hat{G}_i) = \max\{ L(F|G = g); g(0) = i\} \quad for\ i \in \{0,1\}$$

A natural predictor of mutation is:

$$p_1 = 1_{\{ L(F|\hat{G}_1) \geq L(F|\hat{G}_0) \}}$$

That is, the mutational state of individual 0 would be the one found in the genome that makes the phenotype most likely. Unfortunately, the performance of this predictor is catastrophic. It can be modified to obtain a usable variant by introducing:

$$p_s = 1_{\{ L(F|\hat{G}_1) \geq s\, L(F|\hat{G}_0) \}}$$

where s is a threshold to adjust. By varying s, we thus obtain a family of predictors based on the likelihood ratio. In the case of small family pedigrees, these predictors yield results very close to the thresholding prediction of the conditional probability that we have used so far. Indeed, on the graph below, the ROC curves obtained by the brute-force method for each of the two types of predictor merge. The same is true for the curve corresponding to the predictor based on thresholding the conditional probability, but this time calculated using simulations.

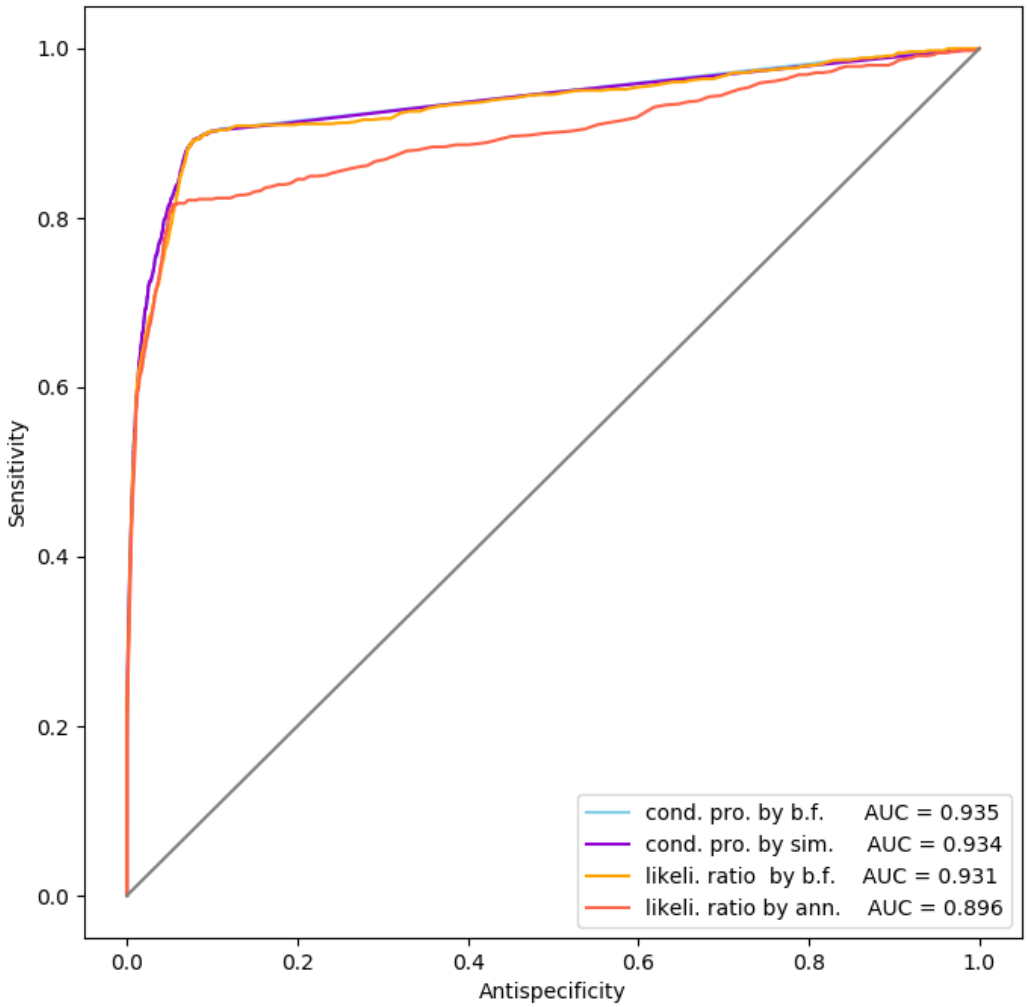


Figure 22 : Comparison of performance of two predictors (thresholding the conditional probability or likelihood ratio) calculated either by brute-force, or by simulated annealing algorithm ("*ann.*" for *simulated annealing*); context: families Reg(3, 1) of 7 individuals, one deleterious mutation and no conditioning of disease case

In the case of maximizing the likelihood of the genome by simulated annealing, the result is not as good and we must certainly conclude that this simulated annealing should be improved either in the choice of the kernel enabling to explore the space of the genomes, or in the setting of the temperature scheme (number and duration of temperature steps).

If we condition at least one case of disease, we obtain, as reported above, a lower performance but with an identical hierarchy:

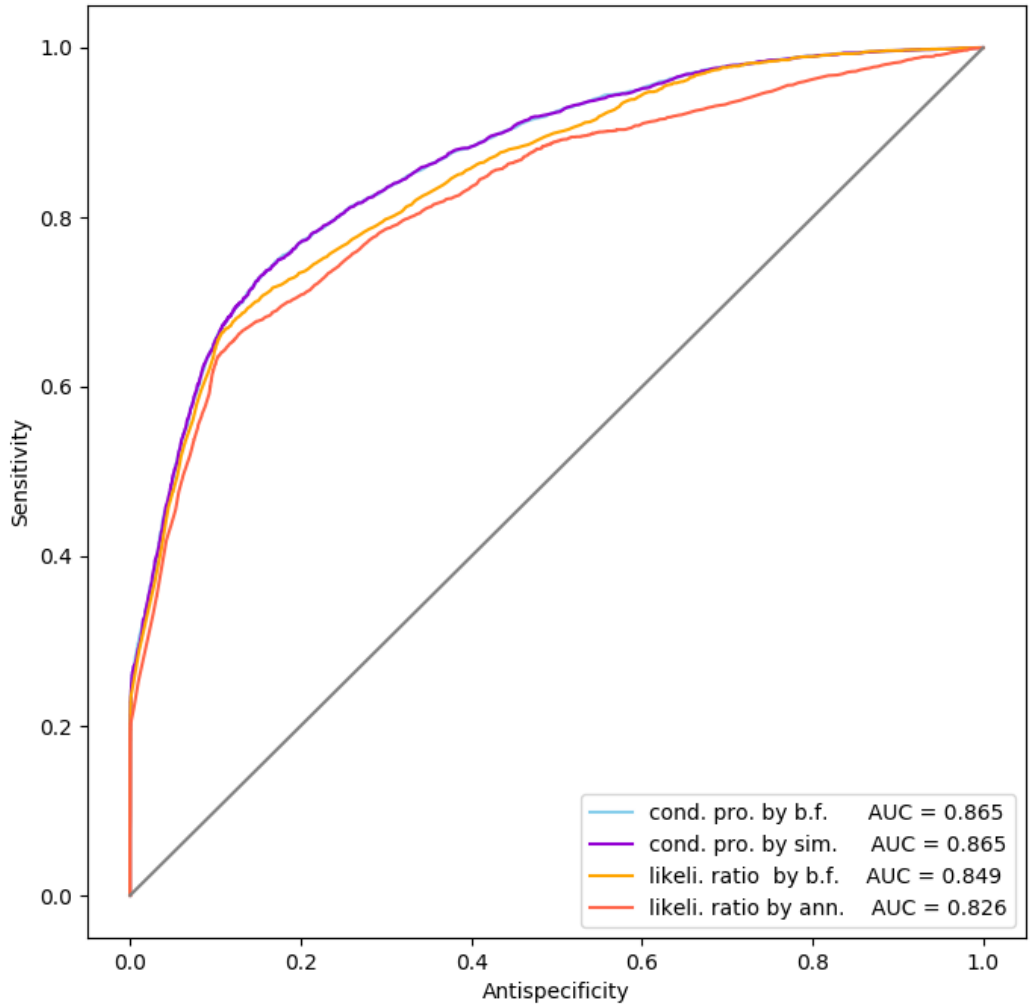


Figure 23 : Comparison of the performance of both predictors (thresholding the conditional probability or the likelihood ratio) obtained either by brute-force, or simulated annealing (context: Reg(3, 1) families of 7 individuals thus and one deleterious mutation and conditioning of at least one disease case per pedigree)

4.3 Additions considered

We have limited the statement of our results to cases that we suppose the most interesting. The computing time is also a limiting factor because to get regular and - hopefully - reliable ROC curves as above, the computing time is counted in hours or even days. However, various extensions are possible:

- To carry out for the double mutation model all the experiments done for the single mutation model
- To evaluate a double mutation model taking into account homozygous mutations
- To improve the annealing method based on the likelihood-ratio predictor so that it equals the brute-force method (when the latter is feasible).

- To examine the sensitivity of predictors to parameter estimation errors or even to model change, i.e. simulate families according to the one-mutation model and analyze them with the model for two mutations, and vice versa.

4.4 Limitations

General approaches like the one used in the article face several limitations. The first one concerns the type of disease that was implemented in our models: this disease was uniform and occurred according to two laws of age, one for mutation carriers and another for non-mutated individuals. When we consider the cancer predispositions encountered in the oncogenetic consultation, for example those caused by BRCA mutations, several cancer locations are favored by the mutations but with an incidence varying a lot by location: breast and ovaries are the organs that are the most often affected, but case of cancer striking other organs can also occur. Besides, breast cancers do not constitute a single and simple category since there are different histological types and a classification according to hormone receptors. All these subclasses influence both the age at disease detection and the prognosis. Therefore, the oncogeneticist needs not only to consider in a pedigree the number of cancers, but also their type in detail.

5 Conclusion

This study, although based on rather simple algorithms, yields practical conclusions about the daily work of the oncogeneticist. In particular, it suggests that there is an optimal size for familial pedigrees: bigger is not always better. It also quantifies the prediction quality for other shapes and sizes of pedigrees. Our computations have focused on a single-mutation model for computational feasibility, but we have also considered a double-mutation model with interaction, with similar results. This second type of model fits present oncogenetics issues since nowadays, many of the hereditary cancer risks seem to originate from the interaction between several non-pathogenic variants.

REFERENCES

Berry DA, Iversen ES Jr, Gudbjartsson DF, Hiller EH, Garber JE, Peshkin BN, Lerman C, Watson P, Lynch HT, Hilsenbeck SG, Rubinstein WS, Hughes KS, Parmigiani G. J. BRCAPRO (2002) Validation, sensitivity of genetic testing of BRCA1/BRCA2, and prevalence of other breast cancer susceptibility genes. Clin Oncol. 20(11) 2701-2712. doi: 10.1200/JCO.2002.05.121.

Bonaiti B, Alarcon F, Bonadona V, Pennec S, Andrieu N, Stoppa-Lyonnet D, Perdry H, Bonaïti-Pellié C & Groupe Génétique et Cancer (2011) A new scoring system for the diagnosis of BRCA1/2 associated breast-ovarian cancer predisposition. Bulletin du cancer, 98(7) 779-795.

Eisinger F, Bressac B, Castaigne D et al. (2004) Identification et prise en charge des prédispositions héréditaires aux cancers du sein et de l'ovaire (mise à jour 2004). Bull Cancer 91, 219-237

Elston RC, Stewart J (1971) A general model for the genetic analysis of pedigree data. Human Heredity 21:523-542.

Evans DG, Lalloo F, Cramer A et al. (2009) Addition of pathology and biomarker information significantly improves the performance of the Manchester scoring system for BRCA1 and BRCA2 testing. J Med Genet, 46:811-817

Hanley JA, McNeil BJ (1983) A method of comparing the areas under Receiver Operating Curves derived from the same cases. Radiology 148(3): 839-43

Metropolis N, Rosenbluth AW, Rosenbluth MN, Teller AH (1953) Equation of state calculations by fast computing machines. J. Chem. Phys. 21(6) : 1087