Computing Individual Risks based on Family History in Genetic Disease in the Presence of Competing Risks

Grégory Nuel, Alexandra Lefebvre, Olivier Bouaziz

Research Article: Computing Individual Risks based on Family History in Genetic Disease in the Presence of Competing Risks

G. Nuel*[1,2], A. Lefebvre[3,4] and O. Bouaziz[5]

[1]LPMA, UMR CNRS 7599, Paris, France
[2]UPMC, Sorbonne universités, Paris, France
[3]UPSud, Paris-Saclay, Orsay, France
[4]Institut Curie, Paris, France
[5]MAP5, UMR CNRS 8145, Paris, France

September 13, 2017

Abstract

When considering a genetic disease with variable age at onset (ex: diabetes, familial amyloid neuropathy, cancers, etc.), computing the individual risk of the disease based on family history (FH) is of critical interest both for clinicians and patients. Such a risk is very challenging to compute because: 1) the genotype X of the individual of interest is in general unknown; 2) the posterior distribution $\mathbb{P}(X|\mathrm{FH}, T > t)$ changes with t (T is the age at disease onset for the targeted individual); 3) the competing risk of death is not negligible.

In this work, we present a modeling of this problem using a Bayesian network mixed with (right-censored) survival outcomes where hazard rates only depend on the genotype of each individual. We explain how belief propagation can be used to obtain posterior distribution of genotypes given the FH, and how to obtain a time-dependent posterior hazard rate for any individual in the pedigree. Finally, we use this posterior hazard rate to compute individual risk, with or without the competing risk of death.

Our method is illustrated using the Claus-Easton model for breast cancer (BC). This model assumes an autosomal dominant genetic risk factor such

*corresponding author, gregory.nuel@math.cnrs.fr

as non-carriers (genotype 00) have a BC hazard rate $\lambda_0(t)$ while carriers (genotypes 01, 10 and 11) have a (much greater) hazard rate $\lambda_1(t)$. Both hazard rates are assumed to be piecewise constant with known values (cuts at 20, 30, ..., 80 years). The competing risk of death is derived from the national French registry.

Keywords: piecewise constant hazard, Bayesian network, belief propagation, Hardy-Weinberg, Mendelian transmission.

1 Introduction

Complex diseases with variable age at onset typically have many interacting factors such as the age, lifestyle, environmental factors, treatments, genetic inherited components. The genetic component is generally composed of one or several genes including major genes for which a deleterious mutation rises significantly the risk of the disease and/or minor genes which participation in the disease is moderate by itself.

The mode of inheritance can be monogenic if a mutation in a single gene is transmitted or polygenic if mutations in several genes are transmitted. As an example of a major gene in a complex disease, the BRCA1 gene is well known to be strongly correlated with ovarian and breast cancer since the 90s (Hall et al., 1990; Claus et al., 1994). Carriers of a deleterious mutation in BRCA1 gene have a much higher risk to be affected with relative risks ranging from 20 to 80 but deleterious mutations in BRCA1 gene only explain 5 to 10 % of the disease (Mehrgou and Akouchekian, 2016) as many other implicated known or unknown genes exist along with sporadic cases (cases with no inherited component).

In other rare genetic diseases such as the Transthyretin-related Hereditary Amyloidosis (THA), no sporadic cases are found and therefore the incidence is equal to zero among non-carriers and all affected individuals are necessarily carriers of a deleterious mutation (Plante-Bordeneuve et al., 2003; Alarcon et al., 2009).

The family history (FH) of such diseases is often the first tool for clinicians to detect a family of carriers of a deleterious mutation as any unusual accumulation of cases in relatives leads to suspect a deleterious allele in the family. With the appropriate model and computation, the FH can be used to better target the most appropriate individuals for a genetic testing and / or to identify high-risk individuals who require special attention (monitoring and/or treatments).

The first challenge to compute such a model comes from the fact that genotypes are mostly (if not totally) unobserved and that posterior carrier probability computations must sum over a large number of familial founders' genotypes configurations. Once such computations are carried out, deriving posterior individual

disease risk is also a challenging task since the posterior carrier distribution changes over time and must be accounted for. Finally, for diseases with possibly late age at onset (*e.g.* cancer), the competing risk of death is not negligible and must be accounted for.

A competing risk situation occurs when an event (called a competing event) precludes the occurrence of the event of interest. This is typically the case for late-onset diseases as the risk of death is not negligible for advanced age. Ignoring the risk of death would amount to assume that death cannot happen and would therefore lead to overestimate the cumulative incidence (the probability of having the disease before any time point). Famous examples of such situations include dementia where the patients are of a particularly advanced age and have a high risk of dying as in Jacqmin-Gadda et al. (2014) or Wanneveich et al. (2016), or studies on geriatric patients (see for instance Berry et al., 2010).

Classical familial risk models such as Claus-Easton (Claus et al., 1991; Easton et al., 1993), BOADICEA (Antoniou et al., 2004), or the BayesMendel models (BRCAPRO, MMRpro, PancPRO and MelaPRO, see Chen et al., 2006) do not take into account the competing event of death. As a result, it is likely that individual predictions will tend to be overestimated from these models (De Pauw, 2012). The main result of the present work is that we show how to derive individual risk predictions from the family history while taking into account the competing risk of death, which is a new contribution to the best of our knowledge.

Another interesting point is that, unlike most similar publications, we here provide all the necessary details to integrate the likelihood over the unobserved genotypes and to compute posterior genotype distributions using Bayesian network and sum-product algorithms. One should not that these models and algorithms clearly are often used in the context of genetics (see Lauritzen, 1996; O'Connell and Weeks, 1998; Fishelson and Geiger, 2002; Lauritzen and Sheehan, 2003; Palin et al., 2011, for a few examples), but rarely fully detailed (see Chen et al., 2006, for example).

It should also be noted that the genetics community usually prefers to rely on simple *peeling* algorithms rather than Bayesian network for pedigree computations but the two concepts are in fact totally equivalent, and the sum-product algorithm presented in this paper can indeed be seen as a simple Bayesian network based reformulation of the most general peeling-based algorithm developed so far (Totir et al., 2009).

The paper is organized as follows: firstly, in Section 2.1 we introduce a formal generic Bayesian network model adaptable to any genetic disease with variable age at onset. Secondly, in Section 2.2, we provide in this context all the necessary details to carry belief propagation on this model, and express the marginal posterior carrier distribution using Bayesian network's potentials. Thirdly, in Section 2.3, we

give closed-form formulas for the posterior individual disease risk, and introduce a simple numerical algorithm allowing to take into account the competing risk of death. Finally, in Section 3, all the methods are illustrated with the Claus-Easton model for breast cancer using the disease model and the parameters of Claus et al. (1991); Easton et al. (1993). In particular, individual predictions derived by taking into account the competing risk of death or ignoring it are compared, which emphasizes the importance of properly taking into account competing risk of death in such models.

2 Materials and Methods

In this section, we first introduce our model (Section 2.1) as a Bayesian network. We next explain how to perform belief propagation in order to obtain posterior carrier distributions (Section 2.2). Finally, we provide all the details needed to derive disease risks predictions from these posterior distributions, including taking into account the competitive risk of death (Section 2.3).

2.1 The Bayesian Network

We consider a total of n (related) individuals. With $\mathcal{I} = \{1, \ldots, n\}$, we denote by $\mathcal{F} \subset \mathcal{I}$ the subset of the founders (i.e. individuals without ancestors in the pedigree) and we denote by $\mathcal{I} \setminus \mathcal{F}$ the set of non-founders (i.e. with ancestors in the pedigree). Let $\boldsymbol{X} = (X_1, \ldots, X_n) \in \{00, 01, 10, 11\}^n$ be the genotypic distribution[1] of the whole family, where X_i denotes the genotype of Individual i. Let $\boldsymbol{T} = (T_1, \ldots, T_n) \in \mathbb{R}^n$ be the time vector representing the age at diagnosis of all individuals. The joint distribution of $(\boldsymbol{X}, \boldsymbol{T})$ is given by:

$$\mathbb{P}(\boldsymbol{X}, \boldsymbol{T}) = \underbrace{\prod_{i \in \mathcal{F}} \mathbb{P}(X_i) \prod_{i \in \mathcal{I} \setminus \mathcal{F}} \mathbb{P}\left(X_i | X_{\mathrm{pat}_i}, X_{\mathrm{mat}_i}\right)}_{\text{genetic part}} \times \underbrace{\prod_{i \in \mathcal{I}} \mathbb{P}(T_i | X_i)}_{\text{survival part}} \tag{1}$$

which corresponds to the definition of a Bayesian Network (BN). See Koller and Friedman (2009) for more details. The genetic part of Eq. (1) only relies on the "classical" Mendelian assumption that the distribution of a non-founder genotype only depends on the parental genotypes. The survival part makes the strong assumption that all T_i are conditionally independent given X_i. This assumption is clearly not true when considering any other familial effect on the disease (*e.g.* polygenic effect, environmental exposure, etc.) which is often taken into account

[1]For the sake of simplicity, we consider here a simple bi-allelic gene but multi-allelic genes can obviously be easily considered.

using a familial random effect (often called *frailty* in the survival context). Such familial random effect is for example assumed to account for a polygenic effect in the BOADICEA model (Antoniou et al., 2002, 2004). Note that for the sake of simplicity, the symbol "$\mathbb{P}$" corresponds through the whole paper either to a probability measure or to a density.

The extension of the present model to frailty models such as BOADICEA is clearly possible and, in many ways, quite straightforward. However, for the sake of simplicity, we focus here on a simpler model and will briefly discuss the extension in the conclusion section. However, even with the strong assumption that T_i only depends on X_i, since (the basically unobserved) $\boldsymbol{X}$ has a strong correlation structure within the pedigree, so does $\boldsymbol{T}$.

We can see on Fig. 1 an example of a moderate size (hypothetical) family with a severe history of breast and ovarian cancer. This family has a total of $n = 12$ individuals with $\mathcal{F} = \{1, 2, 3, 4\}$ and $\mathcal{I} \setminus \mathcal{F} = \{5, 6, 7, 8, 9, 10, 11, 12\}$. There is no inbreeding (mating between individuals with a common ancestor) in this family but a mating loop (two families joined more than once by mating) due to the two brothers of the first nuclear family having children with two sisters of the second nuclear family. Such looped pedigree can be tricky to represent and this explains why Individual 7 appears twice (with an identity link) in Fig. 1.

One should note that loops in pedigree are not the same as cycles in the Bayesian networks framework in the sense that the underlying conditional dependence structure of the model remains a proper directed acyclic graph even in the presence of pedigree with loops.

Genetic Part. For the genetic part, we assume that founders' genotypes are distributed according to the Hardy-Weinberg distribution with disease allele frequency f. It means that for any founder $i \in \mathcal{F}$ we have $\mathbb{P}(X_i = 00) = (1 - f)^2$, $\mathbb{P}(X_i = 01) = \mathbb{P}(X_i = 10) = f(1 - f)$, and $\mathbb{P}(X_i = 11) = f^2$. This assumption is extremely frequent in family genetics and usually reasonable since it corresponds to the stationary distribution we observe in a population under mild assumptions. However, one should note that other distributions can easily be considered if necessary (*e.g.* genotype 11 forbidden because it is lethal). For the non-founder we simply assume a Mendelian transmission of the alleles, but unbalanced transmission patterns can also be considered.

The genetic part of the model can also be easily extended to account for various constraints. For example, the presence of monozygous twins, say individuals i and j, only requires one to add an identity variable between the two genotypes: $I_{i,j} \in \{0, 1\}$ such as $\mathbb{P}(I_{i,j}|X_i, X_j) = \mathbf{1}\{X_i = X_j\}$. Genetic tests (including error or not) can also be incorporated as additional variables G_i such as $\mathbb{P}(G_i|X_i)$ corresponding to the test specificity and sensibility. Finally, assuming lethal genotypes (*e.g.*

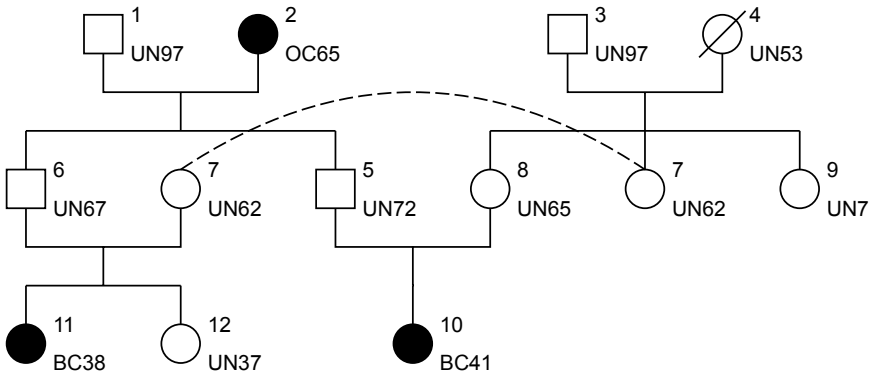

Figure 1: A hypothetical family with a severe FH of cancer. Squares correspond to males, circles to females, and affected individual are filled in black. Individual id on the top-right of the nodes, personal history of cancer (UN=UNaffected; BC=Breast Cancer; OC=Ovarian Cancer) on the bottom-right. The dashed line represents an identity link used to represent the mating loop (due to the mating between individuals 5/8 and 6/7) between brothers 5 and 6, and sisters 7 and 8.

genotype 11) is done straightforwardly by setting to 0 the probability of carrying such genotype. This is equivalent to working conditionally on $\{X_i \neq 11 \text{ for all } i\}$ which obviously alter all genotype distributions, including Hardy-Weinberg for founders.

Survival Part. We place ourselves in the classical survival framework, denoting by $\lambda(t)$ the (time dependent) hazard function, by $S(t)$ the survival function defined as $S(t) = \exp(-\Lambda(t))$ where $\Lambda(t) = \int_0^t \lambda(u)du$ is the cumulative hazard.

We assume an autosomal dominant model where non-carriers have a disease incidence $\lambda_0(t)$ and carriers have a disease incidence $\lambda_1(t)$. This simple assumption results in the following expression of the survival part of the model:

$$\begin{cases} \mathbb{P}(T_i > t | X_i = 00) = S_0(t) \quad \text{and} \quad \mathbb{P}(T_i = t | X_i = 00) = S_0(t)\lambda_0(t) \\ \mathbb{P}(T_i > t | X_i \neq 00) = S_1(t) \quad \text{and} \quad \mathbb{P}(T_i = t | X_i \neq 00) = S_1(t)\lambda_1(t) \end{cases} . \tag{2}$$

As explained above, the symbol "$\mathbb{P}$" corresponds to a (conditional) probability measure for the event $\{T_i > t\}$ and to a density for the punctual event $\{T_i = t\}$.

For example, in the context of the THA, non-carriers cannot be affected ($\lambda_0(t) \equiv 0$) and only carriers have an age-dependent incidence. In the context of breast cancer, $\lambda_0(t)$ might be the incidence for non BRCA carriers and $\lambda_1(t)$ the incidence for BRCA carriers (BRCA1 or BRCA2).

Of course the simple model suggested in Eq. (2) can easily be extended to account for other genetic models (*e.g.* recessive, additive, gonosomal (*i.e.* non-

autosomal), with parent-of-origin effect, etc.) as well as for any known covariates (*e.g.* BMI, smoking, other diseases, etc.) using a classical proportional hazard model.

Hazard rates $\lambda_0(t)$ and $\lambda_1(t)$ are typically described by the literature as piecewise constant hazards (PCHs), but our model allows for any parametric or non-parametric shape as long as hazard rates are provided (*e.g.* hazard rates of Weibull distributions, Gaussian survival, etc.).

2.2 Carrier Risk

For all individual i let us denote by PH_i his/her personal history of the disease. In the case where Individual i was diagnosed with the disease at age t_i we have $\mathrm{PH}_i = \{T_i = t_i\}$. If Individual i was unaffected at age t_i (age at the last follow-up), the variable T_i is right-censored and we have $\mathrm{PH}_i = \{T_i > t_i\}$. From now on, we denote by FH the family history of the disease. This includes the personal history of all individuals and all possible additional constraints or informations (*e.g.* monozygous twins, genetic tests, lethal alleles, etc.). Formally, we can define $\mathrm{FH} = \cup_i(\mathrm{PH}_i \cup \{X_i \in \mathcal{X}_i\})$ where $\mathcal{X}_i \subset \{00, 01, 10, 11\}$ is the subset of allowed values for X_i (*e.g.* $\mathcal{X}_i = \{00, 01, 10\}$ if we know that the genotype 11 is lethal, $\mathcal{X}_i = \{00\}$ if we know that a particular individual is a non-carrier, etc.). Even with genetic testing, it is essential to understand that $\boldsymbol{X}$ is, at best, partially observed. Indeed, even with a (hypothetical and unrealistic) 100% specificity/sensitivity test, a positive heterozygous carrier status cannot distinguish between genotypes 01 and 10. Moreover, genetic tests are in general only available for a few individuals in the whole pedigree. Accounting for the unobserved genotypes is therefore of utmost importance.

Following the classical BN notations, we write the so-called *evidence* $\mathbb{P}(\mathrm{FH})$ as the simple following sum-product of *potentials*:

$$\mathbb{P}(\mathrm{FH}) = \sum_{X_1} \dots \sum_{X_n} \prod_{i=1}^{n} K_i\left(X_i | X_{\mathrm{pa}_i}\right) \tag{3}$$

where the potentials are defined by:

$$K_i\left(X_i | X_{\mathrm{pa}_i}\right) = \mathbb{P}(\mathrm{PH}_i | X_i) \times \begin{cases} \mathbb{P}\left(X_i | X_{\mathrm{pat}_i}, X_{\mathrm{mat}_i}\right) & \text{if } i \in \mathcal{I} \setminus \mathcal{F} \\ \mathbb{P}\left(X_i\right) & \text{if } i \in \mathcal{F} \end{cases} \tag{4}$$

where $\mathbb{P}(\mathrm{PH}_i|X_i)$ is either $\mathbb{P}(T_i = t_i|X_i)$ or $\mathbb{P}(T_i > t_i|X_i)$ and can be obtained through Eq. (2). Note that $\mathrm{pa}_i \subset \mathcal{I}$ denote the parental set of Individual i (empty for founders), and that $X_{\mathcal{J}} = (X_j)_{j \in \mathcal{J}}$ for any $\mathcal{J} \subset \mathcal{I}$. As explained above, any additional information or constraint might and should be added directly into the potentials.

Since $\boldsymbol{X}$ has 4^n possible configurations in the worst case, it is clearly impossible to simply enumerate these configurations even for moderate size pedigrees (e.g., for $n = 10$ or $n = 20$). We therefore need a more efficient algorithm to compute Eq. (3). An efficient solution is provided by the Elston-Stewart algorithm (Elston et al., 1992) in the particular (and frequent) case where the pedigree has no loop. The basic idea is to eliminate variables from the sum-product (*peeling* in the Elston-Stewart literature) from the last generations up to the oldest common ancestor. The resulting complexity $\mathcal{O}(n \times 4^3)$ clearly allows one to deal with arbitrary pedigree size as long as there is no loop.

Unfortunately, loops (inbreeding or mating) are not totally uncommon in pedigrees and therefore have to be accounted for. A simple extension of the Elston-Stewart algorithm consists in using loop breakers: working conditionally to a few number of key genotypes that can be considered as duplicated individuals with known genotypes in a pedigree with no loop. For example, in Fig. 1, Individual 7 is a possible loop breaker. By performing a classical Elston-Stewart algorithm for each genotypic configuration of the loop breakers, $\mathbb{P}(\text{FH})$ can be computed with complexity $\mathcal{O}(n \times 4^{\ell+3})$ where ℓ is the number of loop breakers.

In the context of Bayesian networks, computing $\mathbb{P}(\text{FH})$ (and, in fact, the whole $\mathbb{P}(\boldsymbol{X}, \text{FH})$ distribution) is typically done through *belief propagation* (BP)[2] with a $\mathcal{O}(n \times 4^k)$ complexity where k is the *tree-width* of the graphical model (see Koller and Friedman, 2009, for more details). For a pedigree with no loop, $k = 3$ and the BP complexity is strictly the same than Elston-Stewart, but for more complex pedigrees, k usually increases much slower than $\ell + 3$ and, as a result, BP is often dramatically faster than Elston-Stewart with loop breakers.

In order to achieve this, BP basically eliminates variables from the sum-product of Eq. (3) in a suitable order. In that sense, it is very similar to the notion of *cutset* long used to compute likelihoods in complex pedigrees (see Lange et al., 2013, for a recent reference on the MENDEL package). But BP has the noticeable advantage to allow obtaining the full posterior distribution $\mathbb{P}(\boldsymbol{X}|\text{FH})$ for the same algorithmic complexity while likelihood-based approaches need to repeat many cutset eliminations to achieve the same results. As a consequence, it should not be surprising to see that, in parallel with the classical genetic literature (Elston et al., 1992; Kruglyak et al., 1996; Lange et al., 2013) many authors have been using BP and BN to deal with genetic models (Lauritzen, 1996; O'Connell and Weeks, 1998; Fishelson and Geiger, 2002; Lauritzen and Sheehan, 2003; Palin et al., 2011).

Let us finally point out that the genetics community has put considerable efforts in developing Elston-Stewart algorithms for any Bayesian network counterpart, claiming that *peeling-based* algorithms are more natural for geneticists than junction-tree based ones. Note however that the most general version of these

[2]Also called *sum-product* algorithm.

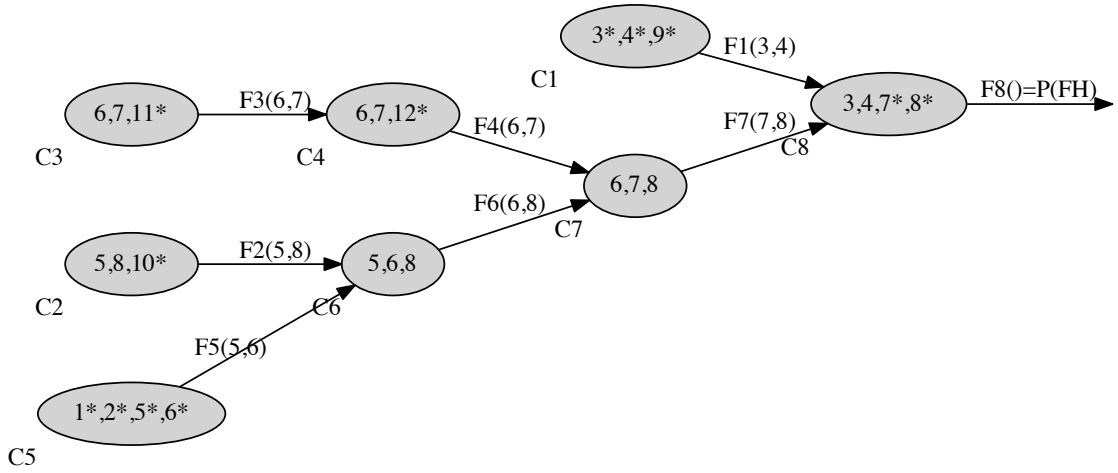

Figure 2: Junction tree of our hypothetical family with the following elimination order: $X_9, X_{10}, X_{11}, X_{12}, X_{1,2}, X_5, X_6, X_{3,4,7,8}$.

peeling algorithms (Totir et al., 2009) is in fact *exactly* equivalent to the classical junction-tree based forward/backward algorithm presented below.

For completeness, we will now briefly recall all the minimal necessary results to implement BP in the context of our model. We nevertheless encourage the interested reader to refer to more classical references like Lauritzen and Sheehan (2003) or Koller and Friedman (2009) for more details.

Variable Elimination and Junction Tree. As an example, we consider the pedigree of Fig. 1 and want to compute $\mathbb{P}(\mathrm{FH})$ by successive variable elimination. We use the following elimination order: X_9, X_{10}, X_{11}, X_{12}, $X_{1,2}$, X_5, X_6, and $X_{3,4,7,8}$. Here follow the quantities obtained in the process:

$$F_1(X_{3,4}) = \sum_{X_9} K_3(X_3)K_4(X_4)K_9(X_{3,4,9});\ F_2(X_{5,8}) = \sum_{X_{10}} K_{10}(X_{5,8,10});$$

$$F_3(X_{6,7}) = \sum_{X_{11}} K_{11}(X_{6,7,11});\ F_4(X_{6,7}) = \sum_{X_{12}} F_3(X_{6,7})K_{12}(X_{6,7,12});$$

$$F_5(X_{5,6}) = \sum_{X_{1,2}} K_1(X_1)K_2(X_2)K_5(X_{1,2,5})K_6(X_{1,2,6});\ F_6(X_{6,8}) = \sum_{X_5} F_2(X_{5,8})F_5(X_{5,6});$$

$$F_7(X_{7,8}) = \sum_{X_6} F_4(X_{6,7})F_6(X_{6,8});\ \mathbb{P}(\mathrm{FH}) = \sum_{X_{3,4,7,8}} F_1(X_{3,4})F_7(X_{7,8})K_7(X_{3,4,7})K_8(X_{3,4,8}).$$

We therefore can obtain $\mathbb{P}(\mathrm{FH})$ by considering only $6 \times 4^3 + 2 \times 4^4 = 896$ configurations over the $4^{12} \simeq 16.8 \times 10^6$ total number of $\boldsymbol{X}$ configurations. Note that a memory bounded version of the variable elimination exists, see Darwiche (2001) for more details.

Fig. 2 is a graphical representation of this particular sequence of elimination and is also a *junction tree* defined as a set of K *cliques* $C_1, \ldots, C_K$ with $C_j \subset \{X_1, \ldots, X_n\}$ with the following properties:

i) tree: each clique j is connected to a a subsequent *clique* $\mathrm{to}_j \in \{j+1, \ldots, K\}$ ($\mathrm{to}_K = \text{root}$ by convention). We also define $\mathrm{from}_k = \{j, \mathrm{to}_j = k\}$ ($\mathrm{from}_1 = \emptyset$) and $S_j = C_j \cap C_{\mathrm{to}_j}$ (with the convention that $S_K = \emptyset$).

ii) covering: for all $i \in \{1, \ldots, n\}$ it exists a j such as $\{X_i, X_{\mathrm{pa}_i}\} \subset C_j$. We then define $\mathrm{of}_i = \min\{j, (X_i, X_{\mathrm{pa}_i}) \subset C_j\}$ and $C_j^* = \{X_i \in C_j, \mathrm{of}_i = j\}$.

iii) running intersection: for all $i \in \{1, \ldots, n\}$ the subgraph formed by $\{C_j, X_i \in C_j\}$ (and the from/to relationships) is a tree.

In the graph theory, junction trees are used as an auxiliary structure for many applications (*e.g.* graph coloring). The proof that any elimination sequence gives a junction tree can be found in Koller and Friedman (2009). The *tree-width* of an elimination sequence / junction tree is defined as the size of its largest clique. Finding the elimination sequence with the smallest tree-width is NP-hard in general, but many heuristics are available (Koller and Friedman, 2009). The elimination order of Fig. 2 has been obtained using the well-known minimum fill-in heuristic.

Belief Propagation. We assume that a suitable elimination order / junction tree has been obtained. For all $j \in \{1, \ldots, K\}$ we hence define the potential of clique C_j as $\Phi_j(C_j) = \prod_{X_i \in C_j^*} K_i\left(X_i | X_{\mathrm{pa}_i}\right)$ and we have the following result:

Theorem 1. *(posterior distribution) For all $i \in \{1, \ldots, n\}$, let $k = \mathrm{of}_i$ and we have:*

$$\mathbb{P}(X_i, \mathrm{FH}) = \sum_{C_k \setminus \{X_i\}} \left\{ \prod_{j \in \mathrm{from}_k} F_j(S_j) \times \Phi_k(C_k) \times B_k(S_k) \right\}$$

where the forward *quantities are defined for $k = 1, \ldots, K$ by:*

$$F_k(S_k) = \sum_{C_k \setminus S_k} \left\{ \prod_{j \in \mathrm{from}_k} F_j(S_j) \times \Phi_k(C_k) \right\}$$

and the backward *quantities are defined by $B_K(S_K = \emptyset) = 1$ (convention) and for $k = K, \ldots, 2$, for all $i \in \mathrm{from}_k$:*

$$B_i(S_i) = \sum_{C_k \setminus S_i} \left\{ \prod_{j \in \mathrm{from}_k, j \neq i} F_j(S_j) \times \Phi_k(C_k) \times B_k(S_k) \right\}.$$

Proof. See Appendix A. □

Using Theorem 1, it is therefore possible to obtain $\mathbb{P}(\text{FH})$ and *all* $\mathbb{P}(X_i|\text{FH})$ by just recursively computing once all forward and backward quantities.

2.3 Disease Risk

While the previous section covered the computation of the posterior probability $\mathbb{P}(X_i|\text{FH})$ for all individuals in the pedigree, we now focus in this section on computing individual posterior disease risks, with or without the competing risk of death.

Risk without competing events. We consider an individual i with a posterior carrier probability π at age τ, that is $\pi = \mathbb{P}(X_i \neq 00|\text{FH}, T_i > \tau)$. Conditionally to the family history, we denote the survival and hazard functions respectively by S and λ such that, for $t \geq \tau$, $S(t) = \mathbb{P}(T_i > t|\text{FH}, T_i > \tau)$ and $S(t) = \exp(-\int_\tau^t \lambda(u)du)$. We have the following result.

Theorem 2. *For any $t \geq \tau$, we have:*

$$S(t) = \pi \frac{S_1(t)}{S_1(\tau)} + (1-\pi)\frac{S_0(t)}{S_0(\tau)}$$

$$\mathbb{P}(X_i \neq 00|\text{FH}, T_i > t) = \frac{1}{S(t)}\pi\frac{S_1(t)}{S_1(\tau)}$$

$$\lambda(t) = \frac{1}{S(t)}\left[\pi\frac{S_1(t)}{S_1(\tau)}\lambda_1(t) + (1-\pi)\frac{S_0(t)}{S_0(\tau)}\lambda_0(t)\right] \tag{5}$$

Proof. See Appendix B. □

Risk with death as a competing event. As explained in the introduction, death precludes the occurence of the disease. This needs to be taken into account by defining the hazard rate of the disease conditionally to the fact that both disease and death have not occurred yet. From a statistical point of view, such a situation can be seen as a competing risk situation or as an illness-death model; see Andersen et al. (1993) or Andersen and Keiding (2012) for a presentation of such models. We define T^* as the minimum between age at disease onset and age at death and we keep the notation T to denote the age at disease onset. Given an individual i with a family history FH, its hazard rate for the disease is defined as

$$\lambda_\alpha(t) = \lim_{\Delta t \to 0} \frac{\mathbb{P}(t \leq T_i < t + \Delta t | T_i^* \geq t, \text{FH})}{\Delta t}$$

We denote by λ_β and S_β the hazard and survival functions of T_i^* (conditionally to the family history) and we assume that λ_α and λ_β are piecewise constants with common cuts $\tau = c_0 < c_1 < \ldots < c_N$ (that is $\lambda_\alpha(t) = \alpha_j$ and $\lambda_\beta(t) = \beta_j$ for $t \in]c_{j-1}, c_j]$).

Lemma 3. *For $j = 1, \ldots, N$, $t \in]c_{j-1}, c_j]$, we have*

$$\mathbb{P}(T_i \leq t | T_i > c_{j-1}, \mathrm{FH}) = \int_{c_{j-1}}^{t} \lambda_\alpha(u) S_\beta(u) du = \frac{\alpha_j}{\beta_j} \left[S_\beta(c_{j-1}) - S_\beta(t) \right]$$

Proof. See Appendix B. □

Practical computations. We assume that one individual has a carrier probability π at age τ (his age without the disease in the FH). We denote by λ_{death} his/her hazard of death. Then the posterior disease risk with the competing risk of death can be computed through the following steps:

1) choose a fine enough discretization $\tau = c_0 < c_1 < \ldots < c_N = t_{\max}$ (ex: all $c_j - c_{j-1} = 0.1$ year);

2) compute $\alpha_j = \lambda_\alpha(c_j)$ using Eq. (5);

3) compute $\beta_j = \alpha_j + \lambda_{\mathrm{death}}(c_j)$;

4) then the marginal posterior probability of being diagnosed with the disease before age c_k, in the presence of death as a competing risk, is given for $k = 1, \ldots, N$ by:

$$\mathbb{P}(T_i \leqslant c_k | \mathrm{FH}) = \sum_{j=1}^{k} \frac{\alpha_j}{\beta_j} \left[S_\beta(c_{j-1}) - S_\beta(c_j) \right].$$

3 Results and Discussion

3.1 The Claus-Easton Model

In order to illustrate our method, we will use the model of illness and the parameters of the Claus-Easton model developed from the Cancer and Steroid Hormone Study in the 90s (Claus et al., 1991; Easton et al., 1993).

The Claus-Easton model is a classical genetic model composed of a genotypic part and a phenotypic part with only the family history (FH) as covariate. It assumes an autosomal dominant mode of inheritance, and a piecewise constant hazard rate by steps of 10 years. The penetrance ($F(t) = 1 - S(t)$) and the

	$20-30$	$30-40$	$40-50$	$50-60$	$60-70$	$70-80$	>80
non carriers	2.00	26.04	112.94	139.94	235.17	232.16	232.03
carriers	168.35	1391.49	3153.21	3222.22	3281.25	3289.86	3286.43
relative risk	84.17	53.44	27.92	23.03	13.95	14.17	14.16

Table 1: Annual incidence (for 100,000) of breast cancer (BC) for carriers/non-carriers and relative risks by age (in years) in the Claus-Easton model.

$20-30$	$30-40$	$40-50$	$50-60$	$60-70$	$70-80$	$80-85$
23.85375	46.86641	130.5396	308.9539	599.914	1493.6	3845.406
$85-90$	$90-95$	$95-99$	$99-100$	$100-101$	$101-102$	$102-103$
8114.203	16400.99	27912.22	35644	38696.22	43033.07	45647.85

Table 2: Annual female death incidence (for 100,000) by age (in years) in the metropolitan French population between 2012 and 2014 (INED, 2017).

density ($f(t) = \lambda(t)S(t)$) are given in Table 2 from Easton et al. (1993) for both carriers and non-carriers at ages $25, 35, \ldots, 85$. The hazard rates can therefore be derived from these data using the formula $\lambda(t) = f(t)/(1 - F(t))$. The results of these computations are given in Table 1. The frequency of the mutated allele has been estimated at $f = 0.0033$ (Claus et al., 1991). The death incidences needed in the competing risk section are given in Table 2.

Figure 3 presents the incidence and survival for BC (carriers and non-carriers) as well as death. We can notice that the breast cancer incidences in carriers are always much higher than in non-carriers at any age and the relative risk between carriers and non-carriers is especially large (RR > 50) before age 40 (see Table 1) but then decreases with aging. We notice that the death incidence stays above the BC incidence for non-carriers at all ages and exceeds even the BC incidence for carriers from age 80. This shows the importance of taking it into consideration especially over a certain age.

3.2 Carrier Risk

In this section we will use the belief propagation in Bayesian networks to obtain the posterior distribution of individual genotypes given the FH. We get the posterior probabilities of each genotype (non-carrier, heterozygous carrier with a paternal mutated allele, heterozygous carrier with a maternal mutated allele and homozygous carrier).

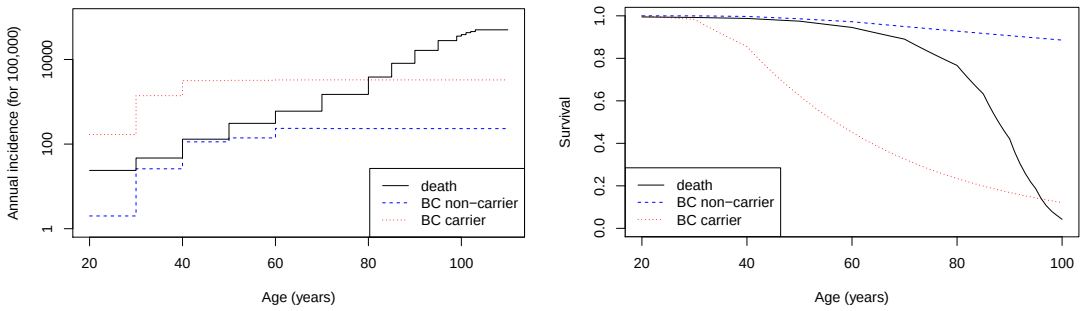

Figure 3: Left-panel: annual (female) death incidence and annual non-carrier/carrier breast cancer incidence. Right-panel: death survival and percentage of non-carrier/carrier individuals without diagnosed breast cancer.

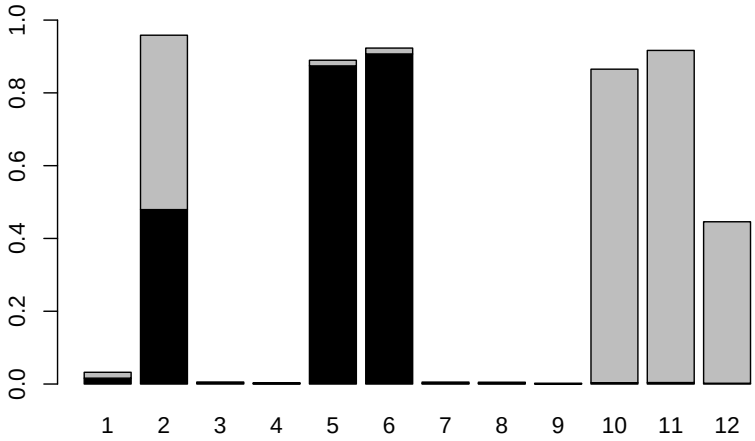

Figure 4: Posterior probabilities for the carrier genotypes of each individual (Individual 1 to 12) in our hypothetical family (Figure 1). The posterior probability of being a paternal carrier $\mathbb{P}(X = 10|\text{FH})$ (resp. maternal carrier $\mathbb{P}(X = 01|\text{FH})$) is colored in black (resp. in grey). The deleterious allele being very rare in the general population ($f = 0.33\%$), the probability of the monozygous carrier genotype is almost zero for each individual and it is therefore not represented here.

Figure 4 represents the marginal posterior probability $\mathbb{P}(X_i = x|\text{FH})$ for all individuals i and for $x = 10$ (paternal carrier) and $x = 01$ (maternal carrier). Note that the posterior probability of the monozygous carrier genotype ($x = 11$) being almost zero for each individual, it is not shown here. The posterior probability of the non-carrier genotype can be easily deduced.

We can notice that the probabilities of being a non-carrier for 1, 3, 4, 7, 8 and 9 are all by far the highest despite the severe phenotype of relatives (granddaughter, niece or daughter). This result is consistent with the personal history of Individual 2 (ovarian cancer at age 51) which points her out as the most likely origin of the mutation in the family. Let us note that since we have no additional information on the ancestors of Individual 2, it is impossible to determine whether her mutation was transmitted by her father or her mother. As a consequence, the posterior carrier probability is equally shared between the paternal and maternal carrier genotypes.

Considering the severe personal history of cancer of Individuals 10 and 11, the most likely situation would be that they both received the mutation of their grandmother through their respective fathers (Individuals 6 and 5 respectively). The posterior probabilities are clearly consistent with this scenario: Individuals 5 and 6 have a probability of $\simeq 90\%$ to be maternal carriers, and Individuals 10 and 11 have similar probabilities to be paternal carriers. Note that Individual 12, being unaffected at age 37 (which is not very informative) basically have 50% chance to have received the mutation from her father.

Figure 5 shows some examples of the variations of the posterior marginal distribution of the genotypes in a same family structure according to different FH. We first notice that with no information (FH1) the posterior probabilities are exactly those of the general population: $\mathbb{P}(X_i \neq 00|\text{FH1}) = 1 - (1 - f)^2 \simeq 0.0066$.

Note that Individual 2 has a severe personal history of cancer (ovarian cancer at age 51) in all other examples. As a consequence, Individual 1, as a male with no personal history of cancer, is mostly totally uninformative therefore not included in the forthcoming analyses.

Individual 4 having no children, she is independent from the rest of the family conditionally to her phenotype and her parent's genotype. With no information about her phenotype in any FH, her probability of being a carrier is therefore almost half her mother's one in each FH (because her father is almost uninformative). If we compare the posterior distribution of the genotype of Individual 3 in FH2, FH3 and FH4, we can notice that the ovarian cancer of her mother which increased her mother's probability of being a carrier raises her probability of being a carrier (FH2). A protective information about her phenotype such as no cancer until age 61 lowers her posterior probability of being a carrier (FH3). On the contrary, the cancer at young age of her daughter which increases her daughter's

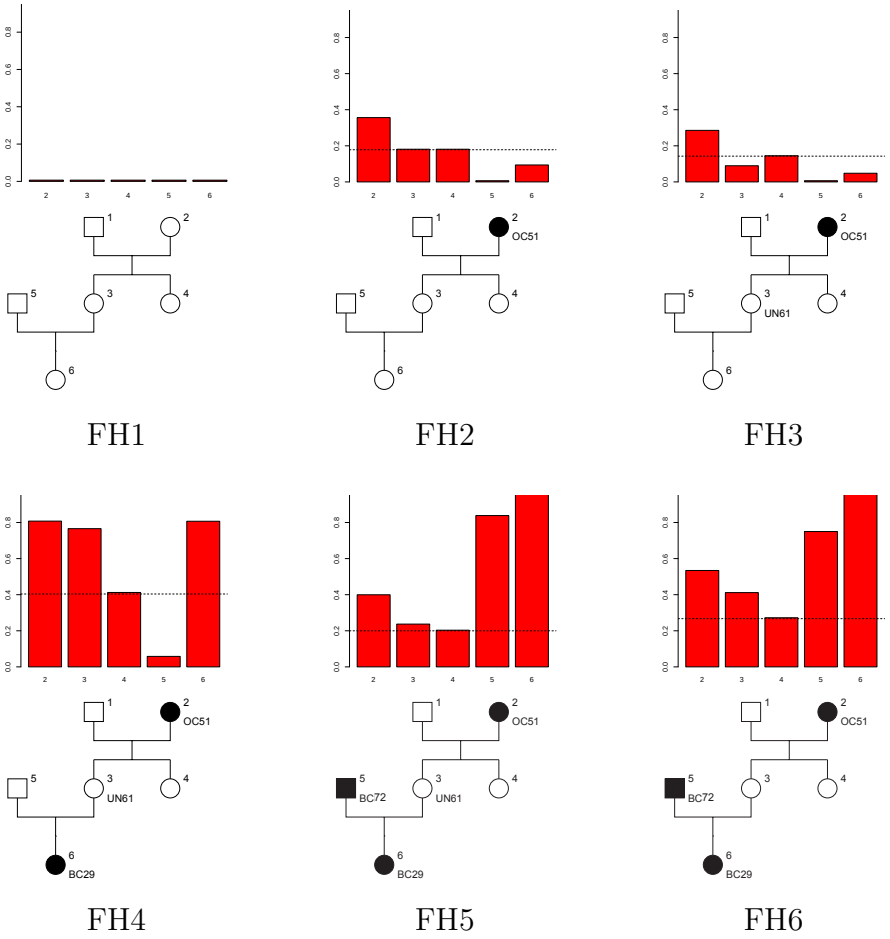

Figure 5: Posterior marginal carrier distribution for a total of 6 FH with increasing degree of severity on the same pedigree structure with 6 individuals. Dashed line represent half the marginal carrier probability of Individual 2.

probability of being a carrier raises her own probability of being a carrier (FH4-6).

We also notice the causal relationships in a whole branch of the family with the transmission between Individuals 2, 3 and 6 of the deleterious allele being highly probable which raises the probability of being a carrier for Individual 3 even in the presence of a protective phenotype (unaffected at age 61) in FH4.

We finally observe the influence of the spouse's genotype when having children (FH5). The higher risk of being a carrier for Individual 5 (because of his cancer at age 72) strongly decreases the carrier probability of his spouse (in comparison with FH4) since the paternal origin of the disease mutation naturally becomes the most likely event. On the other side, the increase of risk for Individual 3 when suppressing her protective phenotype (FH6) also has a consequence on the marginal posterior distribution of her spouse in lowering his probability of being a carrier as his participation in the risk for their daughter is lowered.

To summarize, one's probability of being a carrier mainly depends on: 1) one's probability of having at least one carrier parent, which is correlated to the history of cancer of one's ancestors; 2) one's probability of having transmitted the mutation to one's offspring which is correlated to the history of cancer of one's descendant relatives and one's spouse probability of being a carrier.

Remark: As introduced in the "Disease Risk" section, we know that posterior carrier probabilities should decrease with time for unaffected individuals. For example, if we assume that Individual 4 is unaffected at age 40 in FH6, her probability of being a carrier is 24%. If she stays unaffected up to age 60 (resp. age 80), her probability of being a carrier decreases to 15% (resp. 8.5%).

Table 3 gives a practical illustration of the dependence and conditional independence in a trio grandparent - parent - child. We compare the posterior joint distribution and the product of the posterior marginal distributions of genotypes X_2 and X_6 in FH4 with various information on X_3. We can see that these two quantities are not equal when X_3 is not observed while they are exactly the same when X_3 is fixed. This example demonstrates how X_2 and X_6 are not conditionally independent given FH but they are, conditionally to FH and X_3. Note that when $X_3 = 11$, the mutation is necessarily found in both parents (Individual 1 and 2) as well as in her daughter (Individual 6).

3.3 Cancer Risk

As in Section 2.3 we now consider a female individual i who is unaffected at age τ (*i.e.* $\{T_i > \tau\} \subset \text{FH}$) and denote by $\pi = \mathbb{P}(X_i \neq 00|\text{FH})$ its posterior carrier probability. The purpose of this section is to compute the posterior risk of cancer for this individual (with or without the competing risk of death). As previously explained, these risks only depend on π and τ.

X_2/X_6	NC/NC	NC/C	C/NC	C/C
FH4				
marginal	0.0371306	0.1551811	0.1559446	0.6517438
joint	0.1443102	0.0480015	0.0487650	0.7589233
FH4 and $X_3 = 10$				
marginal	0.0092840	0.7741949	0.0025657	0.2139554
joint	0.0092840	0.7741949	0.0025657	0.2139554
FH4 and $X_3 = 01$				
marginal	0.0000000	0.0000000	0.0118497	0.9881503
joint	0.0000000	0.0000000	0.0118497	0.9881503
FH4 and $X_3 = 11$				
marginal	0.0000000	0.0000000	0.0000000	1.0000000
joint	0.0000000	0.0000000	0.0000000	1.0000000

Table 3: product of the posterior marginal probabilities $\mathbb{P}(X_2|\text{FH})\mathbb{P}(X_6|\text{FH})$ and joint posterior probability $\mathbb{P}(X_2, X_6|\text{FH})$ in the context of known and unknown X_3. NC: non-carrier; C: carrier.

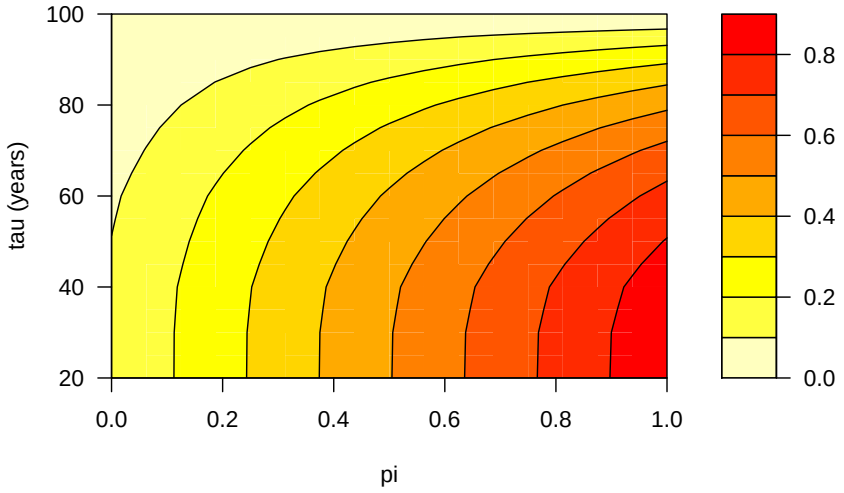

Figure 6: Individual risk of breast cancer without the competing risk of death and for various π and τ

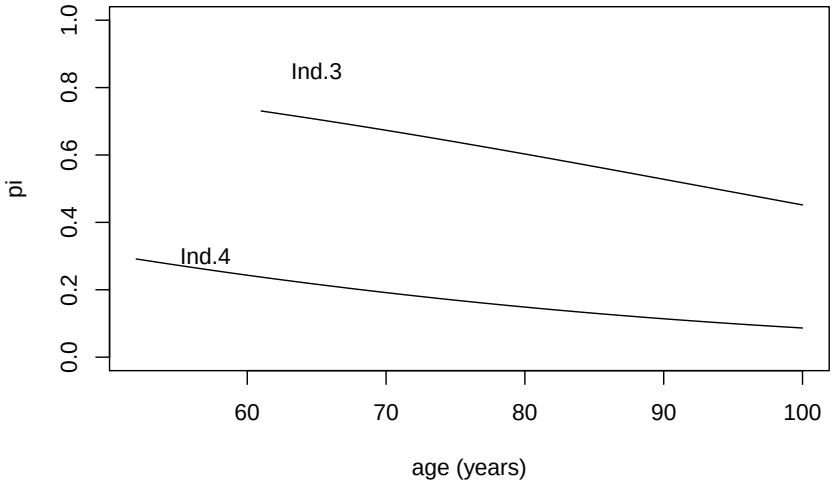

Figure 7: Posterior probabilities of being a carrier according to the time for Individuals 3 and 4 in FH4 assuming Individual 4 is 52 at the time of the censoring.

Figure 6 represents the individual risk of breast cancer up to age 100[3] without the competing risk of death and variant π and τ. We can see that the individual risk of BC rises as π increases and τ decreases. This result is quite intuitive as the younger a patient is, the longer she will be at risk until age 100; the greater her probability of carrying a deleterious allele, the greater her risk to develop a cancer.

As introduced in the previous section the probability of being a carrier for an unaffected individual decreases with time if she stays unaffected. Assuming Individual 4 was 52 in FH4, Figure 7 shows the evolution of the probability of being a carrier for Individual 3 and Individual 4 in FH4. As they stay unaffected we can clearly see the decrease of this probability which has to be taken into account in the computation of the individual risk of breast cancer over time (see Section 2.3).

As explained in Section 2.3, computing risk with the competing risk of death requires a numerical discretization of age by a fixed step Δt. In order to calibrate Δt we used $\Delta t = 0.01$ as a reference, and observed that $\Delta t = 0.1$ is a reasonable balance between accuracy and computational efficiency (data not shown).

Figure 8 represents the individual risk of breast cancer for Individual 7 ($\pi =$

[3]Note that we obtain qualitatively similar results with a lower age limit (*e.g.* age 80), but quantitative results are more illustrative with age 100.

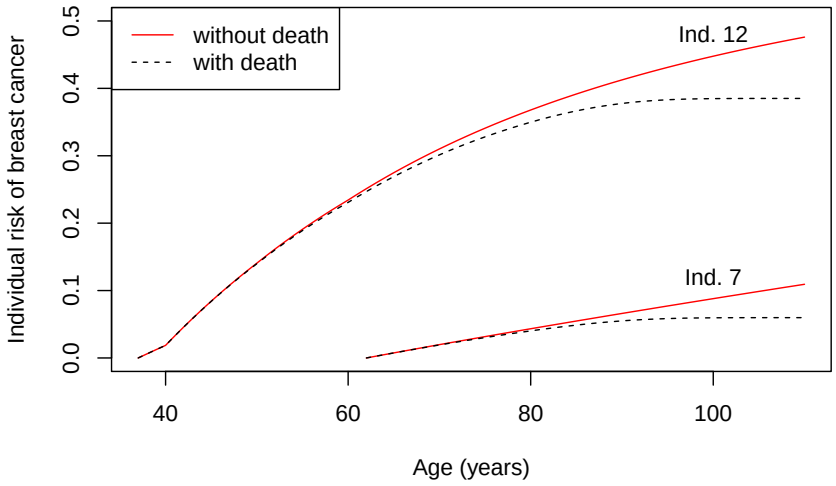

Figure 8: Individual risk of breast cancer with and without the competing risk of death for individual 7 and 12 of our hypothetical family from τ to 100 years with and without the competing risk of death.

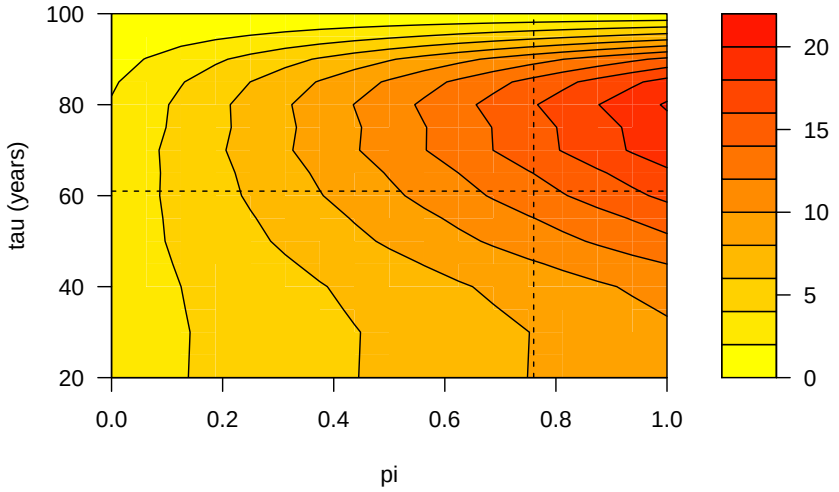

Figure 9: Difference (in percentage) between the individual risk of breast cancer up to 100 years without and with the competing risk of death for various π and τ. Specific values $\pi = 76.59\%$ and $\tau = 61$ are given by the dashed lines.

0.553% and $\tau = 62$ years) and Individual 12 ($\pi = 44.6\%$ and $\tau = 37$ years) in our hypothetical family from τ to 100 years with and without taking into account the competing risk of death. We can see that the difference between the two curves for each individual is increasing with the age. The age from which the difference becomes significant varies with the couple (π, τ). We also observe that the individual risk of breast cancer eventually reaches a plateau which corresponds to the point where the incidence of breast cancer becomes negligible compared to the incidence of death in the elderly.

Quantitatively, the importance of taking into account the competing risk of death is pointed out in the Figure 9 which represents the difference between the individual risk of breast cancer up to the age of 100 years for variant couples (π, τ). For example for Individual 3 in FH4 ($\pi = 76.59\%$, $\tau = 61$, see Figure 5), the error while calculating her individual risk of breast cancer up to the age of 100 years reaches almost 14 %. If it is clear that the competing risk of death can have a limited effect on the global risk of cancer for certain couples (π, τ) its effect is never totally negligible, and since we provide a rigorous way to take it into account we strongly advocate its use in all circumstances.

4 Conclusions

We presented here a general model for genetic disease with variable age at onset. This model, a Bayesian network, combines classical genetic modeling with survival analysis. In order to deal with the (mostly) unobserved genotypes, we first explained in detail how belief propagation can be used to perform likelihood and posterior probability computations. Secondly, we focused on the challenging problem of computing posterior individual disease risks, with or without taking into account the competing risk of death. Finally, we illustrated theses results with the Claus-Easton model for breast and ovarian cancer. The R source codes are available upon request for the interested readers.

For the sake of simplicity, we only considered a bi-allelic locus with standard distribution (autosomal, Hardy-Weinberg, Mendelian allele transmission) but extensions (*e.g.* multi-loci, unbalanced allele transmission, lethal genotypes, etc.) are straightforward. For the survival model, we presented a simple dominant effect without covariates, but again, extensions to any proportional hazard model (*e.g.* recessive, additive, with covariates, etc.) are easy to implement. Incorporating random effects (at the individual and/or familial level) in the model (like in the BOADICEA model, see Antoniou et al., 2002, 2004) is clearly also possible, but slightly more challenging.

Computation of posterior carrier distributions remains almost unchanged except for the random effect support which must be discretized (five values are claimed to be sufficient in the BOADICEA literature) and for the belief propagation which must be performed once for each of the possible value of the random effect. For posterior risks, calculations get slightly more complex since the posterior individual hazard must now be integrated over the (changing over time) posterior joint distribution of the individual genotype and of the random effect. Basically, all computations are slightly more intensive with random effects, but most results of Section 2.3 remain very similar.

One of the important limitations of the present work is the fact that we assume that all model parameters are known. However, it should be noted that likelihood and conditional likelihood might be easy to compute through the belief propagation which means that we basically provide all the necessary means to estimate the model parameters from actual data. In that context, it is nevertheless critical to deal efficiently with ascertainment issues: the fact that the family ending up in the database are usually precisely the one with the most severe disease family history. But standard methods like the PEL (Alarcon et al., 2009), which basically are conditional likelihood computations, are known to deal relatively well with the problem.

In order to take into account the competing risk of death, we used death from all causes, which was obtained from registry data (INED, 2017). However, only

death without cancer precludes the onset of cancer and we are not interested into death from all causes. Since registry data usually do not report the causes of death it is a difficult task to estimate the risk of death without cancer. This has been studied for instance in Wanneveich et al. (2016) through a illness-death model, using registry data and differential equations to model the specific causes of death. Nevertheless, it is very likely that the gain in terms of predictions would be minor as mortality from all causes is likely to be close to mortality without cancer.

Further work includes all the extensions described above (*e.g.* more complex genetic model, genetic tests, familial random effects, etc.) as well as the development of a clinical web application for the Claus-Easton model in close collaboration with the cancer genetics department of the *Institut Curie*. From the methodological point of view, we plan to focus on the computation of more complex posterior distribution like the number of carriers in any subgroup of individuals and/or the familial posterior risk (time before any family member at risk is diagnosed).

Acknowledgments

We would first like to thank both anonymous reviewers for their constructive comments and remarks. This work received the support of both the *Institut de Recherche en Santé Publique* (IRESP) and the *Ligue National Contre le Cancer* (LNCC). Alexandra Lefebvre's internship was funded by the *Institut Curie*. We finally warmly thank Antoine de Pauw (*Institut Curie*) for his continuous friendly support and for suggesting the hypothetical family presented here.

Conflict of Interest

The author(s) declare(s) that there is no conflict of interest regarding the publication of this paper.

A Proofs for the Carrier Risk Section

For all $k \in \{1, \dots, K\}$ we recursively define: $u_k = \{k\} \cup_{j \in \text{from}_k} u_j$, $U_k = \cup_{j \in u_k} C_j$, and $V_k = \cup_{j \notin u_k} C_j$. Then we can compute the so-called *forward* and *backward* quantities over any separator $S_j = C_j \cap C_{\text{to}_j}$:

$$F_j(S_j) = \sum_{U_j \setminus S_j} \prod_{X_i \in U_j^*} K_i\left(X_i | X_{\text{pa}_i}\right) \quad \text{and} \quad B_j(S_j) = \sum_{V_j \setminus S_j} \prod_{X_i \in V_j^*} K_i\left(X_i | X_{\text{pa}_i}\right)$$

where $U_j^* = \{X_i \in U_j, \exists k \in u_j, \text{of}_i = k\}$ and $V_j^* = \{X_i \in V_j, \exists k \notin u_j, \text{of}_i = k\}$.

The key is then to prove that, for all $j \in \{1, \ldots, K\}$ we have:

$$\mathbb{P}(S_j, \mathrm{FH}) = F_j(S_j) B_j(S_j) \tag{6}$$

$$\mathbb{P}(C_k, \mathrm{FH}) = \Phi_k(C_k) \times \prod_{j \in \mathrm{from}_k} F_j(S_j) \times B_k(S_k). \tag{7}$$

For proving Eq. (6), we start by noticing that the JT (Junction Tree) properties (Koller and Friedman, 2009) give: $\{X_1, \ldots, X_n\} \setminus S_j = (U_j \setminus S_j) \uplus (V_j \setminus S_j)$ and $\{X_1, \ldots, X_n\} = U_j^* \uplus V_j^*$ (both being disjoint unions). We therefore have:

$$\begin{aligned}
\mathbb{P}(S_j, \mathrm{FH}) &= \sum_{U_j \setminus S_j} \sum_{V_j \setminus S_j} \prod_{X_i \in U_j^*} K_i\left(X_i | X_{\mathrm{pa}_i}\right) \prod_{X_i \in V_j^*} K_i\left(X_i | X_{\mathrm{pa}_i}\right) \\
&= \underbrace{\left(\sum_{U_j \setminus S_j} \prod_{X_i \in U_j^*} K_i\left(X_i | X_{\mathrm{pa}_i}\right) \right)}_{F_j(S_j)} \times \underbrace{\left(\sum_{V_j \setminus S_j} \prod_{X_i \in V_j^*} K_i\left(X_i | X_{\mathrm{pa}_i}\right) \right)}_{B_j(S_j)}
\end{aligned}$$

the factorization between the first and second equation being possible thanks to the fact that $\left(\cup_{X_i \in U_j^*} \{X_i, X_{\mathrm{pa}_i}\}\right) \cap \left(\cup_{X_i \in V_j^*} \{X_i, X_{\mathrm{pa}_i}\}\right) = S_j$ (JT properties again).

The proof is basically the same for Eq. (7) using $\{X_1, \ldots, X_n\} \setminus C_k = \uplus_{j \in \mathrm{from}_k} (U_j \setminus S_j) \uplus (V_k \setminus S_k)$ we get:

$$\begin{aligned}
\mathbb{P}(C_k, \mathrm{FH}) &= \sum_{\{X_1, \ldots, X_n\} \setminus C_k} \prod_{X_i \in \{X_1, \ldots, X_n\}} K_i\left(X_i | X_{\mathrm{pa}_i}\right) \\
&= \Phi_k(C_k) \prod_{j \in \mathrm{from}_k} \sum_{U_j \setminus S_j} \sum_{V_k \setminus S_k} \prod_{X_i \in U_j^*} K_i\left(X_i | X_{\mathrm{pa}_i}\right) \prod_{X_i \in V_k^*} K_i\left(X_i | X_{\mathrm{pa}_i}\right) \\
&= \Phi_k(C_k) \prod_{j \in \mathrm{from}_k} \underbrace{\sum_{U_j \setminus S_j} \prod_{X_i \in U_j^*} K_i\left(X_i | X_{\mathrm{pa}_i}\right)}_{F_j(S_j)} \underbrace{\sum_{V_k \setminus S_k} \prod_{X_i \in V_k^*} K_i\left(X_i | X_{\mathrm{pa}_i}\right)}_{B_k(S_k)}.
\end{aligned}$$

The factorisation being possible as $\uplus_{j \in \mathrm{from}_k} (U_j \setminus S_j) \cap (V_k \setminus S_k) = \emptyset$ (running intersection) and $\forall j, \forall k, U_j^* \subseteq U_j$ and $V_k^* \subseteq V_k$.

Finally, the recursive expression of the forward and backward quantities can be easily derived from equations (6) and (7):

$$\begin{aligned}
\mathbb{P}(S_k, \mathrm{FH}) &= \sum_{C_k \setminus S_k} \mathbb{P}(C_k, \mathrm{FH}) \\
F_k(S_k) \cancel{B_k(S_k)} &= \sum_{C_k \setminus S_k} \prod_{j \in \mathrm{from}_k} F_j(S_j) \times \Phi_k(C_k) \times \cancel{B_k(S_k)}
\end{aligned}$$

which gives the forward recursion by simplifying the $B_k(S_k)$ term.

B Proofs for the Disease Risk Section

Proof of Theorem 2. For clarity, we recall that $S_0(t) = \mathbb{P}(T_i > t|X_i = 00)$, $S_1(t) = \mathbb{P}(T_i > t|X_i \neq 00)$, $\pi = \mathbb{P}(X_i \neq 00|\text{FH}, T_i > \tau)$ and $S(t) = \mathbb{P}(T_i > t|\text{FH}, T_i > \tau)$, for $i = 1, \ldots, n$, and that $\{T_i > \tau\} \subset \text{FH}$. Since the T_i are independent conditionally to the X_i, the distribution of T_i conditionally on X_i obviously does not depend on FH (for values of X_i which are not forbidden by FH). This is why FH can be omitted almost everywhere in the following proof as soon as π has been computed.

We have $S(t) = \sum_{X_i} \mathbb{P}(T_i > t, X_i|T_i > \tau, \text{FH})$, where the notation $\sum_{X_i}$ represents the summation over the different possible values of X_i, that is $X_i = 00$ or $X_i \neq 00$. Using Bayes' rule,

$$\begin{aligned}\mathbb{P}(T_i > t, X_i \neq 00|T_i > \tau, \text{FH}) &= \mathbb{P}(T_i > t|X_i \neq 00, T_i > \tau, \text{FH}) \times \mathbb{P}(X_i \neq 00|T_i > \tau, \text{FH}) \\ &= \frac{\mathbb{P}(T_i > t, X_i \neq 00, \text{FH})}{\mathbb{P}(T_i > \tau, X_i \neq 00, \text{FH})} \times \pi \\ &= \frac{\mathbb{P}(T_i > t|X_i \neq 00, \text{FH})}{\mathbb{P}(T_i > \tau|X_i \neq 00, \text{FH})} \times \pi = \frac{S_1(t)}{S_1(\tau)}\pi,\end{aligned}$$

where we used the fact that $\mathbb{P}(T_i > t|X_i \neq 00, \text{FH}) = \mathbb{P}(T_i > t|X_i \neq 00)$. We similarly prove that $\mathbb{P}(T_i > t, X_i = 00|T_i > \tau, \text{FH}) = (1 - \pi)S_0(t)/S_0(\tau)$.

The next result is proved using Bayes' rule:

$$\begin{aligned}\mathbb{P}(X_i \neq 00|\text{FH}, T_i > t) &= \frac{\mathbb{P}(X_i \neq 00, \text{FH}, T_i > t)}{\mathbb{P}(\text{FH}, T_i > t)} \\ &= \frac{\mathbb{P}(T_i > t|X_i \neq 00, T_i > \tau)}{\mathbb{P}(T_i > t|\text{FH}, T_i > \tau)}\mathbb{P}(X_i \neq 00|\text{FH}, T_i > \tau),\end{aligned}$$

where we also used the fact that $\mathbb{P}(T_i > t|X_i \neq 00, \text{FH}, T_i > \tau) = \mathbb{P}(T_i > t|X_i \neq 00, T_i > \tau)$.

We then directly have $\mathbb{P}(T_i > t|X_i \neq 00, T_i > \tau) = S_1(t)/S_1(\tau)$ from Bayes' rule, $\mathbb{P}(X_i \neq 00|\text{FH}, T_i > \tau) = \pi$ and $\mathbb{P}(T_i > t|\text{FH}, T_i > \tau) = S(t)$ which concludes the proof.

Finally, in order to prove Equation (5), we recall that

$$\begin{aligned}\lambda(t) &= \lim_{\Delta t \to 0} \frac{\mathbb{P}(t \leq T_i < t + \Delta t|T_i \geq t, \text{FH})}{\Delta t} \\ \lambda_0(t) &= \lim_{\Delta t \to 0} \frac{\mathbb{P}(t \leq T_i < t + \Delta t|T_i \geq t, X_i = 00)}{\Delta t} \\ \lambda_1(t) &= \lim_{\Delta t \to 0} \frac{\mathbb{P}(t \leq T_i < t + \Delta t|T_i \geq t, X_i \neq 00)}{\Delta t}\end{aligned}$$

Then,

$$\begin{aligned}\mathbb{P}(t \leq T_i < t+\Delta t | T_i \geq t, \mathrm{FH}) &= \sum_{X_i} \mathbb{P}(t \leq T_i < t+\Delta t, X_i | T_i \geq t, \mathrm{FH}) \\ &= \sum_{X_i} \mathbb{P}(t \leq T_i < t+\Delta t, X_i, \mathrm{FH}) / \mathbb{P}(T_i \geq t, \mathrm{FH}) \\ &= \sum_{X_i} \mathbb{P}(t \leq T_i < t+\Delta t | X_i) \mathbb{P}(X_i | T_i \geq t, \mathrm{FH}),\end{aligned}$$

using Bayes' rule and the fact that $\mathbb{P}(t \leq T_i < t+\Delta t | X_i, \mathrm{FH}) = \mathbb{P}(t \leq T_i < t+\Delta t | X_i)$ and $\mathbb{P}(X_i, \mathrm{FH} | T_i \geq t, \mathrm{FH}) = \mathbb{P}(X_i | T_i \geq t, \mathrm{FH})$. Dividing by Δt and taking the limit as Δt tends to 0 gives

$$\lambda(t) = \lambda_1(t) \times \mathbb{P}(X_i \neq 00 | T_i \geq t, \mathrm{FH}) + \lambda_0(t) \times \mathbb{P}(X_i = 00 | T_i \geq t, \mathrm{FH})$$

We showed previously that $\mathbb{P}(X_i \neq 00 | T_i \geq t, \mathrm{FH}) = \pi S_1(t)/(S(t)S_1(\tau))$ and $\mathbb{P}(X_i = 00 | T_i \geq t, \mathrm{FH}) = (1-\pi)S_0(t)/(S(t)S_0(\tau))$ which concludes the proof. □

Proof of Lemma 3. The first part of the equality is a standard result in the competing risk setting: we have, from Bayes' rule,

$$\lambda_\alpha(u) = \lim_{\Delta t \to 0} \frac{\mathbb{P}(t \leq T_i < t+\Delta t | \mathrm{FH})}{\Delta t \, \mathbb{P}(T_i^* \geq t | \mathrm{FH})}$$

and consequently $\lambda_\alpha(u) S_\beta(u)$ is equal to the density of T conditionally to FH. Then, since $\lambda_\alpha(u) = \alpha_j$ for $u \in]c_{j-1}, c_j]$ we have

$$\begin{aligned}\mathbb{P}(T_i \leq t | T_i > c_{j-1}, \mathrm{FH}) &= \int_{c_{j-1}}^{t} \lambda_\alpha(u) S_\beta(u) du = \alpha_j \int_{c_{j-1}}^{t} S_\beta(u) du \\ &= \alpha_j \int_{c_{j-1}}^{t} \exp\left(- \int_0^u \lambda_\beta(v) dv \right) du\end{aligned}$$

Now, for $u \in]c_{j-1}, t]$, $t \leq c_j$,

$$\int_0^u \lambda_\beta(v) dv = \int_0^{c_{j-1}} \lambda_\beta(v) dv + \beta_j (u - c_{j-1})$$

and

$$\begin{aligned}\int_{c_{j-1}}^{t} \exp\left(- \int_0^u \lambda_\beta(v) dv \right) du &= \exp\left(- \int_0^{c_{j-1}} \lambda_\beta(v) dv \right) \int_{c_{j-1}}^{t} \exp(-\beta_j (u - c_{j-1})) du \\ &= S_\beta(c_{j-1}) \int_{c_{j-1}}^{t} \exp(-\beta_j (u - c_{j-1})) du\end{aligned}$$

The integral on the right side of the equation is straightforward to compute. This gives,

$$S_\beta(c_{j-1}) \int_{c_{j-1}}^{t} \exp(-\beta_j(u - c_{j-1}))du = \frac{1}{\beta_j}\Big(S_\beta(c_{j-1}) - S_\beta(c_{j-1}) \exp(-\beta_j(t - c_{j-1}))\Big)$$

Finally, we conclude by noticing that

$$\begin{aligned} S_\beta(t) &= \exp\left(-\int_0^{c_{j-1}} \lambda_\beta(u)du - \int_{c_{j-1}}^{t} \lambda_\beta(u)du\right) \\ &= S_\beta(c_{j-1}) \exp\left(-\beta_j(t - c_{j-1})\right) \end{aligned}$$

□

References

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

P. K. Andersen, Ø. Borgan, R. D. Gill, and N. Keiding. *Statistical models based on counting processes*. Springer Series in Statistics. Springer-Verlag, New York, 1993. ISBN 0-387-97872-0.

Per Kragh Andersen and Niels Keiding. Interpretability and importance of functionals in competing risks and multistate models. *Statistics in medicine*, 31 (11-12):1074–1088, 2012.

AC Antoniou, PDP Pharoah, G. McMullan, NE Day, MR Stratton, J. Peto, BJ Ponder, and DF Easton. A comprehensive model for familial breast cancer incorporating BRCA1, BRCA2 and other genes. *British journal of cancer*, 86(1):76–83, 2002. ISSN 0007-0920.

AC Antoniou, PPD Pharoah, P. Smith, and DF Easton. The BOADICEA model of genetic susceptibility to breast and ovarian cancer. *British journal of cancer*, 91(8):1580–1590, 2004. ISSN 0007-0920.

Sarah D Berry, Long Ngo, Elizabeth J Samelson, and Douglas P Kiel. Competing risk of death: an important consideration in studies of older adults. *Journal of the American Geriatrics Society*, 58(4):783–787, 2010.

Sining Chen, Wenyi Wang, Shing Lee, Khedoudja Nafa, Johanna Lee, Kathy Romans, Patrice Watson, Stephen B Gruber, David Euhus, Kenneth W Kinzler, et al. Prediction of germline mutations and cancer risk in the lynch syndrome. *Jama*, 296(12):1479–1487, 2006.

Elisabeth B Claus, N Risch, and W Douglas Thompson. Genetic analysis of breast cancer in the cancer and steroid hormone study. *American journal of human genetics*, 48(2):232, 1991.

Elizabeth B Claus, Neil Risch, and W Douglas Thompson. Autosomal dominant inheritance of early-onset breast cancer. implications for risk prediction. *Cancer*, 73(3):643–651, 1994.

Adnan Darwiche. Recursive conditioning. *Artificial Intelligence*, 126(1-2):5–41, 2001.

Antoine De Pauw. *Estimation des risques de cancer du sein et de l'ovaire des femmes sans mutation des gènes BRCA1et BRCA2: apport des modèles de calcul de risque*. PhD thesis, Paris 7, 2012.

DF Easton, DT Bishop, D Ford, and GP Crockford. Genetic linkage analysis in familial breast and ovarian cancer: results from 214 families. the breast cancer linkage consortium. *American journal of human genetics*, 52(4):678, 1993.

Robert C Elston, Varghese T George, and Forrestt Severtson. The elston-stewart algorithm for continuous genotypes and environmental factors. *Human heredity*, 42(1):16–27, 1992.

Maáyan Fishelson and Dan Geiger. Exact genetic linkage computations for general pedigrees. *Bioinformatics*, 18(suppl 1):S189–S198, 2002.

J.M. Hall, M.K. Lee, B. Newman, J.E. Morrow, L.A. Anderson, B. Huey, and M.C. King. Linkage of early-onset familial breast cancer to chromosome 17q21. *Science*, 250(4988):1684, 1990. ISSN 0036-8075.

INED. Death incidence in France, period 2012-2014. `https://www.ined.fr/en/everything_about_population/data/france/deaths-causes-mortality/mortality-tables/`, 2017.

Hélène Jacqmin-Gadda, Paul Blanche, Emilie Chary, Lucie Loubère, Hélène Amieva, and Jean-François Dartigues. Prognostic score for predicting risk of dementia over 10 years while accounting for competing risk of death. *American journal of epidemiology*, 180(8):790–798, 2014.

Daphne Koller and Nir Friedman. *Probabilistic graphical models: principles and techniques*. MIT press, 2009.

Leonid Kruglyak, Mark J Daly, Mary Pat Reeve-Daly, and Eric S Lander. Parametric and nonparametric linkage analysis: a unified multipoint approach. *American journal of human genetics*, 58(6):1347, 1996.

Kenneth Lange, Jeanette C Papp, Janet S Sinsheimer, Ram Sripracha, Hua Zhou, and Eric M Sobel. Mendel: the swiss army knife of genetic analysis programs. *Bioinformatics*, 29(12):1568–1570, 2013.

Steffen L Lauritzen. *Graphical models*, volume 17. Clarendon Press, 1996.

Steffen L Lauritzen and Nuala A Sheehan. Graphical models for genetic analyses. *Statistical Science*, pages 489–514, 2003.

Amir Mehrgou and Mansoureh Akouchekian. The importance of brca1 and brca2 genes mutations in breast cancer development. *Medical Journal of the Islamic Republic of Iran*, 30:369, 2016.

Jeffrey R O'Connell and Daniel E Weeks. Pedcheck: a program for identification of genotype incompatibilities in linkage analysis. *The American Journal of Human Genetics*, 63(1):259–266, 1998.

Kimmo Palin, Harry Campbell, Alan F Wright, James F Wilson, and Richard Durbin. Identity-by-descent-based phasing and imputation in founder populations using graphical models. *Genetic epidemiology*, 35(8):853–860, 2011.

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

Liviu R Totir, Rohan L Fernando, and Joseph Abraham. An efficient algorithm to compute marginal posterior genotype probabilities for every member of a pedigree with loops. *Genetics Selection Evolution*, 41(1):52, 2009.

Mathilde Wanneveich, Hélène Jacqmin-Gadda, Jean-François Dartigues, and Pierre Joly. Impact of intervention targeting risk factors on chronic disease burden. *Statistical methods in medical research*, page 0962280216631360, 2016.