Estimation des risques de maladies dues à des mutations génétique à partir de données familiales

Flora Alarcon

Université Paris XI **Ecole Doctorale 420**

Faculté de Médecine Paris-Sud

Estimation des risques de maladies dues à des mutations génétiques à partir de données familiales

THÈSE

présentée et soutenue publiquement le 7 Juillet 2009

pour l'obtention du

Doctorat de l'université Paris XI

(spécialité Génétique Statistique)

par

Flora ALARCON

Composition du jury

Président :	Florence Demenais
Rapporteurs :	Maria Martinez
	Jean-Louis Golmard
Examinateurs :	Nadine Andrieu
	Dominique Stoppa-Lyonnet
Directrice de thèse :	Catherine Bonaïti-Pellié

Remerciements

"Un seul mot, usé, mais qui brille comme une vieille pièce de monnaie : merci !" *(Pablo Neruda)*

Je tiens à remercier tout particulièrement Catherine Bonaïti, ma directrice de thèse, pour sa disponibilité, ses conseils. Merci pour tout ce que vous m'avez appris.
Je remercie également Catherine Bourgain pour avoir relu et corrigé mes articles et pour m'avoir remonté le moral bien des fois.
Je remercie Françoise Clerget de m'avoir accueillie dans l'unité U535.

Je remercie sincèrement les membres de mon jury de thèse. Merci à Jean-Louis Golmard et Maria Martinez qui ont bien voulu rapporter cette thèse. Merci à Nadine Andrieu, Dominique Stoppa-Lyonnet et Florence Demenais d'avoir accepté de faire partie de mon jury.
Je souhaite remercier très chaleureusement Alfred Spira et Emmanuel Barillot de m'avoir fait confiance et de m'avoir donner l'envie de poursuivre dans la recherche.

Je tiens à remercier mes compagnons de route, ceux qui sont passés et ceux qui sont restés.
Je commence bien sûr par Céline avec qui j'ai partagé mon bureau durant ces trois années. Merci d'avoir pris le temps d'écouter mes problèmes de programme, d'écriture d'article. Merci pour toutes ces discussions enrichissantes. Et je te renvoie le compliment, Céline, "Ta présence m'aura terriblement manqué lors de la rédaction de ma thèse" et je bénis gmail !
Merci au "quatuor non cantinien" de m'avoir accompagnée dans cette dernière ligne droite. Merci à Marie-Claude, Hervé, Emeline et Hélène pour tous ces déjeuners plaisants.
Comment remercier Marie-Claude ? Car tu es l'un de nos deux GI préférés mais tu as aussi relu mes articles et presque toutes mes présentations. Enfin, tu as relu cette thèse. De tout ça, je te remercie !
Un merci particulier aussi à Hervé. Merci de m'avoir aidée dans mes programmes, de m'avoir écoutée

me plaindre, de m'avoir réconfortée dans les moments difficiles.
Merci à Michel, l'autre GI et gentil tout court (et merci à Monique pour tes bons gâteaux !).
Merci aussi à Audrey avec qui j'ai encore bien du plaisir a discuter autour d'un déjeuner.
Un merci très particulier à Salma avec qui j'ai partagé des moments inoubliables (et c'est le moins que l'on puisse dire !).
Merci à Rémi, Pascal, Mathieu, Emmanuelle, Jacqueline, Déwi, clair et Vincent (merci de m'avoir fait un makefile qui m'aura servit tout au long de ma thèse)) !
Merci à Pierre (mon voisin du 1-3), Patricia, Anne-Louise, Bernard, Marie-Hélène, Philippe.
Un grand merci à Christine notre secrétaire qui a pris le temps de relire cette thèse et qui sait, en plus, faire de si beaux petits chaussons.
Je tiens également à remercier Sophie Hassid, Jérome Carayol, Hamida d'avoir pris le temps de me répondre. Merci pour votre gentillesse !

L'un des grand moment de cette thèse fut ma collaboration avec Hugo. En plus d'être un ami cher, tu es le compagnon de travail idéal. Tu m'as beaucoup appris, bien plus encore que la EL...

J'aimerai aussi remercier mes amis. Merci à Pierre qui m'a initiée à la génétique. J'espère que nous aurons l'occasion de travailler ensemble.
Merci aussi à Thieums, tu vas nous manquer !
Merci à Anne, Chacha, Virginie, Brubru, Guigui pour votre présence si précieuse.

Je remercie ma "petite famille" pour leur soutien. Merci de vous être toujours tellement intéressé à tout ce que je faisais ! Merci à mon père, à mon parrain Philippe, à Tonio et à Jo ! Merci à Max !
Merci de tout mon coeur à ma mère qui fait toujours tout pour me rendre la vie plus simple. Merci de m'aider, de m'aimer comme tu le fais !

Enfin, merci merci merci...à mes tendres, mes doux, mes merveilleux amours !
Colombine, tu fais de ma vie un enchantement !
Fred, c'est bien toi ma plus belle rencontre...

Flora

A Colombine, la parenthèse enchantée de cette thèse.

Table des matières

Table des figures

Liste des tableaux

Introduction

L'épidémiologie génétique s'intéresse à la mise en évidence des facteurs de risque et, plus particulièrement ceux d'origine génétique, dans les maladies humaines. Les premiers facteurs génétiques ont été découverts pour des maladies dites monogéniques ou mendéliennes, qui sont des maladies dues à une mutation délétère rare dans un seul gène. Pour ces maladies, une agrégation familiale a été démontrée grâce à l'analyse de ségrégation, et la ségrégation du trait a été expliquée par un mode génétique simple, récessif ou dominant.

Aujourd'hui, les études en épidémiologie génétique portent également sur des maladies dites multifactorielles, qui résultent de la combinaison complexe de facteurs multiples aussi bien génétiques qu'environnementaux. Parmi ces maladies complexes, certaines possèdent des sous-entités mendéliennes, c'est-à-dire des formes rares, dues à une ou plusieurs mutation(s) dans un seul gène. Ces sous-entités, que l'on peut aussi appeler monogéniques, sont caractérisées par un risque beaucoup plus élevé et/ou un âge d'apparition plus jeune que celui de la maladie en population générale.
Pour ces sous-entités aussi bien que pour les maladies mendéliennes, l'estimation du risque morbide associée à une mutation génétique (que l'on appelle fonction de pénétrance) en fonction de l'âge, est un défi majeur en santé publique car il permet non seulement une prise en charge adaptée mais également de mieux comprendre la maladie.

L'estimation des risques se fait presque toujours à partir de données familiales. Ces familles ne sont pas recensées aléatoirement dans la population mais à partir d'individus atteints dont les apparentés sont susceptibles de porter le ou les génotypes à risque. L'estimation de la péné-

trance se fait donc à partir de l'information phénotypique et éventuellement génotypique de ces familles, recensées sous certains critères.
La problématique est différente selon que les familles ont été recueillies selon des recommandations pour la recherche de mutations dans le cadre de la prise en charge ou, selon un protocole élaboré dans le cadre d'un projet de recherche. Dans un cadre de protocole de recommandation pour la prise en charge, les critères de recensement sont souvent variables dans le temps et plus ou moins bien appliqués tandis que lorsque les familles sont recensées dans un cadre de recherche, les critères sont mieux définis.
Dans la plupart des études, la méthode d'estimation ne tient pas compte du biais que représente la sélection des familles sur certains critères et ceci conduit à des estimations biaisées.
De plus, il est rare que l'ensemble des individus d'une famille soit génotypé et, même si les individus génotypés apportent des informations sur les individus non génotypés du fait de leur lien de parenté, des difficultés peuvent survenir lorsqu'un grand nombre de génotypes est inconnus.

L'objectif de ce travail de thèse a été d'étudier puis de développer des méthodes permettant l'estimation de la fonction de pénétrance en fonction de l'âge, à partir de données familiales, en tenant compte du mode de recensement des familles.
Nous nous sommes intéressés, dans un premier temps, à l'étude d'une méthode asymptotiquement sans biais, fournissant des estimations de la pénétrance au moyen de familles recensées sur des critères pouvant être complexes ou mal définis. Nous avons étudié l'efficacité relative de cette méthode dans le cas où plusieurs individus n'étaient pas génotypés dans la famille. Nous nous sommes intéressés également au comportement de la méthode selon les apparentés inclus ou non dans la généalogie.
Ensuite, nous avons entrepris de développer une méthode, basée sur une approche paramétrique, permettant d'estimer la fonction de pénétrance en prenant en compte le fait que les familles ont été recensées sur l'existence d'au moins un individu atteint. Puis, nous avons étudié cette méthode et l'avons comparé à une méthode déjà existante.
Enfin, nous avons développé une seconde méthode, permettant également d'estimer la pénétrance dans le cas de familles recensées sur l'existence d'au moins un atteint dans la famille

mais à l'aide d'une approche non-paramétrique.

Ce manuscrit est construit en cinq chapitres. Après une brève description des méthodes d'analyse de ségrégation qui ont permis les premières estimations de risque, le chapitre 1 présentera un état de l'art sur l'estimation des risques de maladies dues à des mutations génétiques rares. Le second chapitre présentera la méthodologie commune utilisée dans les chapitres 3 et 4 pour l'estimation de la fonction de pénétrance à partir de données familiales. Dans le troisième chapitre, nous présenterons l'étude d'efficacité relative que nous avons menée sur une méthode d'estimation de la pénétrance à partir de familles recensées sur des critères complexes. Dans un quatrième chapitre, nous exposerons une méthode d'estimation paramétrique que nous avons développée afin d'estimer la pénétrance en tenant compte du biais de recensement à partir de familles recensées sur des critères indépendants de l'histoire familiale. Nous illustrerons notre méthode à l'aide de plusieurs jeux de données, l'un portant sur des familles françaises atteintes d'une maladie mendélienne : la neuropathie amyloïde héréditaire ; l'autre, portant sur des familles françaises atteintes d'une maladie complexe à sous-entité monogénique : le cancer du sein. Nous présenterons également une application sur des familles suédoises atteintes de neuropathie amyloide héréditaire. Dans le chapitre 5, nous présenterons une méthode que nous avons développée. Il s'agit d'une méthode non-paramétrique basée sur la vraisemblance empirique, qui permet également l'estimation de la fonction de pénétrance.
Enfin, nous conclurons ce travail en essayant de dégager les points importants de cette recherche et les perspectives à envisager.

Chapitre 1

État de l'Art

L'estimation des risques en génétique, à partir de données familiales, permet d'évaluer le risque de maladie associée à une ou plusieurs mutations génétiques. Il s'agit d'une problématique difficile à traiter mais dont les résultats peuvent s'avérer essentiels pour définir des stratégies de prise en charge pour les individus ayant des risques élevés de développer la maladie mais aussi pour permettre une meilleure compréhension des mécanismes sous-jacents de la maladie.

Dans ce cas, le problème majeur est la prise en compte du mode de recensement des familles.

L'analyse de ségrégation est le premier modèle de prise en compte du recensement. De plus, l'**analyse de ségrégation** a permis l'identification de gènes "majeurs" reconnus responsables de certaines maladies ou de certaines formes de maladie . On parle alors de maladies à prédisposition monogénique (c'est-à-dire dues à la mutation dans un seul gène). Pour ces maladies, l'estimation précise du risque cumulé d'être atteint associé à la mutation prédisposante en fonction de l'âge (appelé **fonction de pénétrance**) est possible, à condition de tenir compte du mode de **recensement des familles**.

L'analyse de ségrégation permet cette estimation mais n'utilise que l'information phénotypique des individus de la famille. D'autres méthodes ont été développées par la suite pour utiliser les génotypes des individus qui apportent une information supplémentaire importante pour l'estimation du risque.

Dans ce chapitre, nous commencerons par présenter l'analyse de ségrégation dans ses grandes lignes. Ensuite, nous décrirons les différents critères de recensement des familles ainsi que le

type de modèle génétique associé. Enfin, nous présenterons les différentes méthodes permettant d'estimer la fonction de pénétrance en prenant en compte le recensement des familles mais aussi les **âges variables de début de la maladie** ainsi que l'information génotypique des individus.

1.1 L'analyse de ségrégation

L'analyse de ségrégation vise à détecter l'existence et à spécifier la nature d'un facteur génétique susceptible d'expliquer les distributions familiales observées d'un phénotype donné. Elle utilise comme unité d'échantillonnage un groupe de sujets apparentés, la famille. Il s'agit de la première étape nécessaire pour déterminer, à partir de données familiales, le mode de transmission d'un phénotype, avec comme but principal, l'identification d'un gène ayant un effet fort que nous appellerons gène "majeur".
Classiquement, la **méthode d'analyse de ségrégation dite "simple"** consistait à tester si la transmission du caractère dans les familles était en accord avec les lois de l'hérédité mendélienne monogénique.
Or, dans de nombreux cas, un tel mécanisme simple ne permet pas d'expliquer les concentrations familiales observées. Dans cette situation, on peut chercher s'il existe un "gène majeur", c'est-à-dire un gène qui explique une partie importante de la variabilité du caractère, parmi l'ensemble des facteurs qui déterminent ce caractère. Des **méthodes d'analyse de ségrégation dites "complexes"** ont été développées dans ce but.

Généralement, pour que les observations apportent suffisamment d'information sur la ségrégation du trait, les familles sont recensées à partir de la population d'étude, non pas aléatoirement, mais en fonction de certaines considérations portant sur le phénotype étudié. Ces familles ne sont pas détectées "en bloc" mais par l'intermédiaire d'un (ou plusieurs) de ses membres ayant un phénotype particulier. La première étape consiste à recueillir un échantillon de malades par les méthodes classiques, en s'adressant aux secteurs susceptibles de voir de tels malade : services hospitaliers, instituts spécialisés, etc. La deuxième étape consiste à contacter la famille de chaque malade et à établir un diagnostic pour les membres de la famille que l'on

peut étendre plus ou moins loin. On a alors un échantillon de familles.

Le schéma de recensement des individus par lesquels la famille est sélectionnée (on les appelle les **proposants**) doit donc être pris en compte dans les méthodes d'analyse de ségrégation.

1.1.1 Les méthodes dites "simples"

Les méthodes d'analyse de ségrégation "simples" testent si la transmission du caractère dans les familles est en accord avec les lois de l'hérédité mendélienne monogénique. Lorsque les caractères étudiés sont des maladies, les familles sont recensées par l'intermédiaire de malades et l'étude de la répartition des individus atteints dans ces familles doit tenir compte du mode de recensement. Différentes méthodes ont été développées pour prendre en compte le biais de recensement. Nous en citons trois.

- La "méthode du proposant" de Weinberg [47, 14]. Il s'agit d'une méthode intuitive basée sur le constat qu'un proposant (c'est-à-dire un individu atteint et recensé) fournit l'information que ses parents sont capables d'engendrer un enfant malade. Le reste des individus de la fratrie, sans le proposant, fournit alors une estimation sans biais du rapport entre individus malades et individus sains. Dans le cas de plusieurs proposants par familles, l'auteur propose de dupliquer les familles autant de fois qu'il y a de proposants. Crow a montré que "la méthode du proposant" de Weinberg fournissait une estimation consistante du ratio de ségrégation [14]. La preuve est donnée en Annexe A.

- Morton a développé un modèle de recensement [33] qui s'applique à des familles nucléaires et qui a été très utilisé à cause de sa simplicité dans la procédure de correction. Il introduit la probabilité π qu'un cas, choisi aléatoirement dans la population, soit sélectionné comme étant un proposant. Dans la situation où $\pi = 1$ (que l'on appelle sélection tronquée), toutes les familles avec au moins un cas seront incluses dans l'analyse. Dans la situation contraire où $\pi \to 0$ (que l'on appelle sélection unique), la probabilité qu'une

famille soit recensée est proportionnelle au nombre d'atteints dans la famille. Dans ce cas, les familles avec plusieurs atteints seront sur-représentées.
Dans les situations intermédiaires où $0 < \pi < 1$ (que l'on appelle sélection multiple), les familles possédant plusieurs proposants sont très probables.

– Cannings et Thompson [8] ont proposé une stratégie pour reconstruire la généalogie une fois le proposant identifié. Ils ont montré qu'en adoptant leur stratégie, le fait de corriger uniquement pour le recensement initial, permet de s'affranchir du biais de recensement de l'analyse.

1.1.2 Les méthodes dites "complexes"

Les méthodes dites "complexes" ont été développées afin de tester différentes hypothèses de transmission génétique dans le cas de maladies complexes. La méthodologie statistique est basée sur le maximum de vraisemblance. Le principe est de construire des modèles généraux qui vont emboîter les modèles génétiques que l'on veut tester. Pour cela, on a besoin de calculer la vraisemblance du modèle pour les généalogies.

1.1.2.1 Vraisemblance du modèle pour les généalogies

Avant de définir la vraisemblance du modèle pour une famille, nous allons définir précisément la fonction de pénétrance.

Si on note B l'allèle non muté, b l'allèle muté et $Phen$ le phénotype, on peut définir trois pénétrances :

$$f_{bb} = \mathbb{P}(Phen = \text{atteint} | Gen = bb)$$

$$f_{bB} = \mathbb{P}(Phen = \text{atteint} | Gen = bB)$$

$$f_{BB} = \mathbb{P}(Phen = \text{atteint} | Gen = BB)$$

Ces pénétrances permettent de définir différents modes de transmission de la maladie. On dira que l'allèle b a un effet dominant si $f_{bb} = f_{bB}$ et un effet récessif si $f_{bB} = f_{BB}$. Dans le cas d'un

phénotype avec un âge de début variable, la pénétrance est nécessairement une fonction de l'âge.

La vraisemblance du modèle pour une généalogie est la probabilité d'observer les phénotypes ($Phen$) sachant les paramètres du modèle, $\theta = (f, q)$, où f représente le vecteur des pénétrances $f = (f_{bb}, f_{bB}, f_{BB})$, q représente la fréquence de l'allèle muté et A représente le recensement des familles. Elle s'écrit :

$$\mathbb{P}(Phen|\theta, A).$$

Mettons de côté le problème du recensement pour le moment. On suppose l'indépendance entre les phénotypes des n individus de la généalogie conditionnellement à leur génotype.
De plus, le génotype étant une variable latente (les génotypes ne sont pas observés), la probabilité marginale des phénotypes peut être calculée en sommant sur l'ensemble des combinaisons génotypiques possibles :

$$\begin{aligned}\mathbb{P}(Phen|\theta) &= \sum_g \mathbb{P}(Phen, Gen = g|\theta) \\ &= \sum_g \mathbb{P}(Phen|Gen = g; \theta)\mathbb{P}(Gen = g|\theta) \\ &= \sum_g \mathbb{P}(Phen|Gen = g; f)\mathbb{P}(Gen = g|q),\end{aligned}$$

et,

$$\mathbb{P}(Gen) = \mathbb{P}(Gen_1)\mathbb{P}(Gen_2|Gen_1)...\mathbb{P}(Gen_n|Gen_1 \dots Gen_{n-1})$$

Si on suppose les lois de Mendel, les génotypes des enfants ne dépendent que des génotypes des parents. Pour les individus fondateurs, les génotypes dépendent des fréquences alléliques.
La difficulté majeure dans le calcul des probabilités génotypique est que l'ensemble des combinaisons génotypiques possibles peut être très grand. Pour parer à ce problème, Elston et Stewart ont proposé un algorithme récursif [21] de sorte que la complexité du calcul n'augmente pas exponentiellement mais linéairement avec la taille de la généalogie.

Cependant, les familles ne sont pas recensées aléatoirement en population mais à travers un individu atteint : le proposant. La prise en compte du recensement se fait alors à l'aide

d'un dénominateur représentant la probabilité de recensement de la famille. Dans le cas d'un trait binaire, on note π, la probabilité qu'un cas, choisi aléatoirement dans la population, soit sélectionné comme étant le proposant. La vraisemblance s'écrit alors :

$$\begin{aligned} L_A(\theta) &= \frac{\mathbb{P}(A|Phen;\theta;\pi)\mathbb{P}(Phen|\theta;\pi)}{\mathbb{P}(A|\theta;\pi)} \\ &= \frac{L(\theta)\mathbb{P}(A|Phen;\pi)}{\mathbb{P}(A|\theta;\pi)}, \end{aligned}$$

où, A représente l'évènement "la famille a été recensée". Le dénominateur peut s'écrire :

$$\mathbb{P}(A|\theta;\pi) = \sum_y \mathbb{P}(A|Phen = y;\pi)\mathbb{P}(Phen = y|\theta).$$

La difficulté est alors d'écrire cette probabilité et de modéliser l'évènement A.

1.1.2.2 Les modèles généraux

Les modèles généraux vont donc emboîter les modèles génétiques et vont nous permettre de les tester par la méthode du maximum de vraisemblance. Parmi ces modèles généraux, le **modèle des probabilités de transmission** (appelé aussi modèle des taux de transmission) permet de tester si la transmission est bien monogénique en fabriquant un sur-modèle où l'on supposera que la transmission des allèles d'une génération à l'autre ne suit pas nécessairement les lois de Mendel. Elston et Stewart [21] ont donc introduit les probabilités de transmission. De cette façon, le modèle où les probabilités de transmission sont mendéliennes sera comparé au modèle où ces probabilités sont libres, comprises entre 0 et 1. Mais ce modèle permet uniquement de tester le modèle monogénique. Pour pouvoir tester également le modèle polygénique, on utilise le **modèle mixte** proposé par Morton et MacLean [34] qui suppose que la susceptibilité à la maladie est la somme de trois composantes indépendantes qui sont la composante monogénique, la composante polygénique et la composante environnementale.
Enfin, le **modèle unifié**, proposée par Lalouel et al. [30], est la combinaison des deux modèles précédents. Ce modèle permet de se prémunir contre un modèle monogénique qui simulerait l'existence d'un gène majeur.

1.1.2.3 Prise en compte des âges de début variable

Dans les maladies à âge de début variable, on doit tenir compte du fait que les individus non atteints de la famille pourraient développer la maladie par la suite. Dans ce cas, la pénétrance est le risque cumulé à un âge donné.
Abel et al. [1] ont développé un modèle de régression logistique incluant une méthode d'analyse de survie afin de prendre en compte, dans l'analyse de données familiales, le problème d'âge de début variable ainsi que des covariables dépendantes du temps. Cette méthode est une extension du modèle de régression logistique introduit par Bonney [7]. La méthode est basée sur la modélisation de la probabilité d'être atteint à l'intérieur d'un certain intervalle de temps.
L'analyse de survie permet de modéliser la fonction de pénétrance. Si on note $F(t)$, la fonction de pénétrance variant avec l'âge pour les porteurs de la mutation, on a :

$$\mathbb{P}(Phen_i|Gen_i) = \begin{cases} 1 - F(t_i) & \text{Si } i \text{ n'est pas atteint à l'âge } t_i \\ F(t_i + 1) - F(t_i) & \text{Si } i \text{ est atteint entre } t_i \text{ et } t_{i+1} \end{cases}$$

où $Phen_i$ est le phénotype de l'individu i et Gen_i est son génotype.
$F(t_i)$ est alors le risque cumulé pour l'individu i à l'âge t_i. L'introduction d'une méthode d'analyse de survie ouvre des possibilités et le phénotype peut alors être exprimé par l'âge de début de la maladie. La fonction de pénétrance F peut être modélisée par une fonction du paramètre à estimer θ.

L'analyse de ségrégation a donc pour objectif principal de déterminer le modèle génétique sous-jacent à l'existence d'un risque familial, en particulier de tester si l'agrégation du trait ou de la maladie dans la famille semble être dûe ou non à un gène "majeur". Elle permet, par ailleurs, d'estimer la pénétrance, en particulier dans les modèles complexes, en tenant compte du recensement des familles et en prenant en compte les âges de début variables de maladie, mais cela augmente le nombre de paramètres à estimer et peut entraîner des problèmes de convergence. De plus, l'analyse de ségrégation n'utilise que des données phénotypiques et donc, elle ne permet d'estimer le risque associé à la mutation qu'en fonction de cette information. Pour

une maladie donnée, une fois que l'existence d'un gène a été détectée grâce à l'analyse de ségrégation, puis identifié par analyse de liaison, il est préférable d'estimer la fonction de pénétrance en utilisant l'information phénotypique mais également l'information génotypique. Ceci présente le double avantage d'avoir une meilleure information et de pouvoir mettre en place des stratégies de prise en charge et de suivie plus adaptées. Cependant, le problème du recensement des familles demeurent et doit toujours être pris en compte dans les modèles d'estimations de la fonction de pénétrance.

1.2 Données génétiques familiales pour l'estimation de la pénétrance

L'analyse de ségrégation utilise uniquement l'information phénotypique. Avec l'introduction de l'information génotypique des individus, certains auteurs, confondant biais de recensement et manque d'information, ont cru que cet apport de l'information génotypique leur permettrait de s'affranchir du biais de recensement [25]. Dans leur article [10], Carayol et al. ont montré que le biais de recensement demeurait si le recensement des familles n'était pas pris en compte dans la vraisemblance.

Une fois le modèle génétique connu, le problème est donc de parvenir à une estimation la plus précise possible de la fonction de pénétrance en tenant compte du recensement des familles.

Les données familiales pour l'estimation de la fonction de pénétrance sont toujours recueillies au travers d'un ou de plusieurs proposants. Cependant, les conditions de sélection des proposants dépendent du protocole sous jacent.

Dans le cadre d'un protocole de recherche, les critères de recensement des familles seront définis par le responsable de l'étude. Ces critères dépendent, en général, du modèle génétique de la maladie. Ainsi, dans le cas de maladies mendéliennes, le fait de recenser les familles sur l'existence d'un atteint (c'est-à-dire indépendamment de l'histoire familiale) fournira des familles comportant des individus porteurs d'une mutation prédisposante et permettra l'estimation de la pénétrance. Par contre, dans le cas de maladies complexes à sous-entités monogéniques, un tel recensement impliquerait la sélection de nombreuses familles dans lesquelles la mutation sera absente. On utilisera donc, en général, des critères familiaux pour augmenter la probabilité de trouver une mutation prédisposante dans les familles.

Par ailleurs, les données peuvent parfois avoir été recueillies dans le cadre d'une prise en charge. Dans ce cas, il n'y a pas véritablement de critères de recensement mais les membres des familles sont vus en consultation de génétique à la suite de recommandations médicales plus ou moins bien suivies. Dans une telle situation, le recensement n'est pas contrôlé.

La méthode utilisée afin d'estimer la pénétrance va donc dépendre du type de protocole qui a permis le recensement des familles mais également du modèle génétique de la maladie.

On peut cependant distinguer deux types de recensement pour les données familiales : le **recensement sur critères familiaux** et le **recensement sur critères indépendants de l'histoire familiale**.

1.2.1 Recensement sur critères indépendants de l'histoire familiale

Dans le cas de maladies mendéliennes, des critères simples suffisent à recruter des familles comportant des individus porteurs d'une mutation prédisposante (on dit alors que les familles sont informatives). Ainsi, le fait de recenser la famille sur la présence d'au moins un atteint est suffisant pour détecter la présence de la mutation dans cette famille. On parlera alors de recensement sur critères indépendants de l'histoire familiale.
Par ailleurs, il a été démontré que dans les maladies complexes à sous-entités monogéniques, les cas dus à une mutation prédisposante apparaissaient à des âges précoces par rapport aux cas sporadiques [11]. Dans ce cas, l'inclusion d'un critère d'âge à la sélection va augmenter la probabilité pour un atteint de porter la mutation et va nous permettre d'avoir des familles informatives. Ce mode de recensement est alors une alternative au mode de recensement sur critères familiaux dans le cas des maladies complexes à sous-entité monogénique.

1.2.2 Recensement sur critères familiaux

Dans les maladies complexes à sous-entité monogénique, aucune caractéristique clinique ne permet, en général, de différencier les formes héréditaires des cas sporadiques. C'est pourquoi on utilise des critères familiaux afin d'accroître la probabilité de trouver une mutation prédisposante chez un individu atteint car, dans ce cas, l'histoire familiale peut suggérer une prédisposition héréditaire à la maladie.
Par exemple, les critères d'Amsterdam I [10, 44] dans le syndrome du cancer colorectal héréditaire sans polypose (ou syndrome HNPCC pour "Hereditary Non Polyposis Colorectal Cancer") sont des critères familiaux reconnus pour la recherche de mutations prédisposantes. Ces critères sont basés sur la présence d'au moins trois apparentés atteints de cancer colorectal : l'un devant être un apparenté au premier degré des deux autres ; au moins deux générations successives

doivent être atteintes et au moins un des cancers doit être diagnostiqué avant 50 ans [10].
Lorsque l'on étudie des maladies complexes à sous-entités monogéniques, les critères utilisés seront donc plus généralement des critères familiaux. D'autres critères peuvent s'y ajouter. Par exemple l'expertise de 2003 [20] a proposé un certain nombre de critères évocateurs d'une prédisposition héréditaire au cancer du sein dans une famille :

- Le nombre de cas de cancer du sein chez une personne de premier ou deuxième degré dans la même branche parentale.
- La précocité de survenue du cancer du sein (40 ans ou moins).
- La présence de cancer de l'ovaire.
- L'existence de tumeurs primitives multiples sein-ovaire.
- La présence de cancer du sein chez l'homme.

La combinaison de ces critères sert à poser les indications de conseil génétique et de recherche de mutations des deux gènes actuellement identifiés, BRCA1 et BRCA2, selon un principe de score.

La section suivante présente différentes méthodes d'estimation de la fonction de pénétrance. Ces méthodes prennent en compte l'information génotypique des individus, l'âge de début variable de la maladie ainsi que le mode de recensement des familles.

1.3 Les méthodes d'estimation de la pénétrance

On se place toujours dans le cas où le modèle génétique est connu. Plusieurs méthodes ont été développées, selon le contexte, pour estimer la fonction de pénétrance en fonction de l'âge et en utilisant l'information génotypique. On remarquera que toutes ces méthodes sont basées sur une approche paramétrique.

1.3.1 La méthode "kin-cohort"

La méthode "kin-cohort" (the kin-cohort design) [45] a été développée afin d'estimer la pénétrance chez des individus volontaires, porteurs des mutations spécifiques BRCA1 et BRCA2 prédisposantes aux cancers du sein et de l'ovaire parmi la communauté Juifs Ashkenazes des Etats-Unis. Cette méthode consiste à comparer la proportion d'apparentés atteints entre les porteurs et les non-porteurs de la mutation dans une population dans laquelle la mutation est relativement fréquente.
Dans leur article [45], Wacholder et al. constituent, à partir de leurs données, deux échantillons distincts. Le premier échantillon, appelé les "Kin carriers", est constitué des apparentés du premier degré des volontaires qui sont porteurs de la mutation prédisposante. Le second échantillon, appelé les "kin non-carriers" est constitué des apparentés des volontaires qui sont non-porteurs. Une fois constitués ces deux sous-échantillons, le risque cumulé de développer la maladie avant t chez les "kin carriers" et les "kin non-carriers" s'écrit en fonction de la fréquence de l'allèle muté en population (p) et des proportions d'individus atteints avant t respectivement chez les "kin carriers" (R_+) et chez les "kin non-carriers"(R_-). La pénétrance est donc une fonction de (p, R_+, R_-).

Ce modèle est aussi appelé "genotype-proband design" (GPD) [23], qui insiste sur le fait que seul le proposant est génotypé. Ce modèle a été étendu au modèle GPDR, pour "Genotype Proband Design with supplemental genotype of Relatives", dans lequel un ou deux apparentés sont génotypés dans chaque famille nucléaire.

L'un des intérêts de ce type de méthode est qu'il ne nécessite aucune correction.

Cependant, ce type de méthode requiert de grands échantillons d'individus atteints et non atteints et il s'applique à des maladies fréquentes dans lesquelles la mutation prédisposante est elle aussi fréquente. De plus, ces méthodes font appel à des volontaires et cela peut entraîner un biais car les individus ayant une histoire familiale de la maladie étudiée peuvent être plus fréquemment volontaires que les individus sans histoire familiale de la maladie. Cette surreprésentation des individus avec une histoire familiale peut entraîner une surestimation de la pénétrance.

Lorsque les maladies sont rares ou lorsque la mutation prédisposante est rare, les données familiales doivent impérativement être sélectionnées sous d'autres critères. Dans ce cas, la méthode "kin-cohort" ne s'applique plus puisque une correction pour la sélection doit être prise en compte dans les méthodes.

1.3.2 Les vraisemblances proposées par Kraft et Thomas

Kraft et Thomas ont présenté quatre types de vraisemblance qui permettent d'estimer le risque en tenant compte du recensement des familles : les vraisemblances prospective, rétrospective, jointe et conditionnelle [28].

Nous noterons $Phen$ la variable aléatoire représentant le phénotype (atteint/non atteint) ; Gen la variable aléatoire représentant le génotype (porteur de la mutation/ non porteur de la mutation) et Sel représentant l'évènement "la famille est recensée". Les génotypes ne sont pas tous connus mais l'information sur le génotype d'un individu apporte de l'information sur les génotypes des membres de la famille.

Nous cherchons donc à estimer la fonction de pénétrance. Pour cela, nous modéliserons notre fonction de pénétrance par une loi connue, de paramètre θ inconnu.

La vraisemblance jointe est la probabilité des phénotypes et des génotypes conditionnellement au fait que la famille a été sélectionnée, $\mathbb{P}(Phen, Gen|Sel)$. Cette vraisemblance nécessite, comme la vraisemblance prospective, la modélisation explicite des critères de sélection. Pour l'estimation de la pénétrance, la vraisemblance jointe et la vraisemblance prospective sont équivalentes. Quant à la vraisemblance conditionnelle, d'après Kraft et Thomas, elle n'est pas adaptée à l'estimation du risque cumulé et ne permet d'estimer que le risque relatif. De plus, elle est

la moins efficace en terme de variance des paramètres estimés. La vraisemblance prospective et la vraisemblance rétrospective sont présentés dans les sections suivantes.

Kraft et Thomas ont étudié l'efficacité de ces vraisemblances lors de l'estimation du risque associé à une mutation à partir de données phénotypiques et génotypiques issues de fratries sélectionnées sur l'existence d'au moins un atteint et un non atteint, sans imposer l'existence d'un individu porteur. Le phénotype était donc binaire (atteint/non atteint) et le risque était indépendant du temps. Il est donc difficile d'extrapoler à partir de cette étude sur l'intérêt de ces vraisemblances sur des données issues de familles sélectionnées au moyen de critères phénotypiques et génotypiques dans le cas d'une maladie à âge de début variable. Carayol et Bonaïti-Pellié [9] ont choisi de développer la vraisemblance rétrospective car elle permet d'estimer le risque sans avoir à écrire explicitement les critères de sélection.
Outre la vraisemblance rétrospective, nous avons choisi de nous intéresser plus précisément à la vraisemblance prospective plutôt qu'à la vraisemblance jointe car elle a déjà été utilisée pour estimer la pénétrance sur des données de neuropathie amyloïde [38]. De plus, son efficacité pour l'estimation de la pénétrance dans l'étude de Kraft et Thomas est très proche de celle de la vraisemblance jointe, qui est la plus efficace.

1.3.3 La vraisemblance prospective

La vraisemblance prospective s'applique de préférence lorsque les familles sont recensées sur critères indépendants de l'histoire familiale car, dans ce cas, le recensement est moins complexe à modéliser. Elle correspond à la probabilité des phénotypes conditionnellement aux génotypes et aux critères utilisés pour la sélection des familles (Sel) : $\mathbb{P}(Phen|Gen, Sel)$. Le principe de la vraisemblance prospective a été utilisé dans la méthode ARCAD [31] qui a permis d'estimer le risque de cancers associés aux mutations du gène p53 dans le syndrome de Li-Fraumeni [12]. Elle a également été utilisée par Planté-Bordeneuve et al. [38] afin d'estimer la pénétrance chez des individus atteints de neuropathies amyloïdes héréditaires (NAH). La méthode prospective permet d'estimer la pénétrance et de corriger pour la sélection en modélisant explicitement cette sélection. Nous décrirons cette méthode précisément dans le quatrième cha-

pitre.

1.3.4 La vraisemblance rétrospective

La vraisemblance rétrospective s'applique dans le cas où les familles ont été recensées sur des critères complexes et lorsque ces critères sont difficilement modélisables. Elle est basée sur la modélisation de la distribution des génotypes conditionnellement aux phénotypes. Pour une famille f donnée, la Vraisemblance Rétrospective (VR) s'écrit :

$$VR_f = \mathbb{P}(Gen|Phen) = \frac{\mathbb{P}(Phen, Gen)}{\mathbb{P}(Phen)},$$

où $\mathbb{P}(phen)$ représente la probabilité des phénotypes observés.

La vraisemblance rétrospective permet de corriger le biais de sélection sans avoir à modéliser explicitement les critères de sélection puisque, en conditionnant sur l'ensemble des phénotypes observés, on conditionne implicitement sur la sélection si celle-ci ne dépend que de la distribution des phénotypes.

Ce principe a été utilisé par Clerget-Darpoux et al. [13] qui ont proposé le "Mod score" comme une extension de la fonction "Lod score" utilisée en analyse de liaison génétique [32] pour l'estimation des différents paramètres du modèle étudié, comme le vecteur de pénétrance, lorsque le taux de recombinaison et la fréquence de l'allèle muté étaient connus.

Kraft et Thomas [28] ont montré que cette vraisemblance manquait d'efficacité pour l'estimation de risque comparée aux autres vraisemblances telle que la vraisemblance prospective. Mais le cadre de leur étude se limitait à l'estimation d'un risque indépendamment de l'âge.

D'autre part Carayol et Bonaïti-Pellié ont montré que la méthode rétrospective fournissait des estimations biaisées dans le sens d'une surestimation du risque lorsque les critères de recensement impliquaient l'existence d'au moins un individu porteur d'une mutation (l'index) dans chaque famille [9]. Ils ont montré également que l'inclusion d'un critère d'âge accentuait le biais dû à la sélection sur le génotype particulièrement pour la pénétrance estimée aux âges jeunes.

Carayol et Bonaïti-Pellié ont donc développé une méthode utilisant le même principe de condi-

tionnement que la vraisemblance rétrospective mais en corrigeant sur le fait que la sélection n'est pas indépendant des génotypes observés. Ils ont appelé cette méthode la GRL pour "Genotype Restricted Likelihood" [9].

1.3.5 La Genotype Restricted Likelihood (ou GRL)

La Genotype Restricted Likelihood (ou GRL) a été développée à partir de la vraisemblance rétrospective par Carayol et Bonaïti-Pellié pour estimer la pénétrance dans le cas d'une sélection sur critères familiaux. Cette méthode s'applique dans le cas où les critères de sélection ne peuvent pas être facilement modélisés et dépendent du génotype [9]. Cette méthode a été développée à partir de la vraisemblance rétrospective [28] qui est basée sur la modélisation de la distribution des génotypes conditionnellement aux phénotypes. Cette dernière méthode permet de corriger le biais de sélection puisque, en conditionnant sur l'ensemble des phénotypes observés, on conditionne implicitement sur la sélection si celle-ci ne dépend que de la distribution des phénotypes.
Mais, la plupart du temps, les critères de sélection dépendent également des génotypes observés (puisque ne sont intégrées dans l'analyse que les familles dans lesquelles le cas index est porteur de la mutation étudiée).
La GRL prend en compte la sélection, même lorsque celle-ci dépend des génotypes. Une description détaillée de la GRL est donnée dans le troisième chapitre.

1.3.6 Récapitulatif

Le tableau 1.1 donne un récapitulatif des avantages et des inconvénients des différentes méthodes pour l'estimation de la pénétrance présentées dans ce chapitre.

TAB. 1.1 – Tableau récapitulatif des avantages et inconvénients de chacunes des méthodes

Méthode	Inconvénient(s)	Avantage(s)
La méthode kin-cohort	Elle s'applique dans le cas de mutation fréquente et nécessite de grands échantillons de données.	Elle ne nécessite aucune correction pour le recensement des familles.
La vraisemblance conditionnelle	Elle n'est pas adaptée à l'estimation du risque cumulé.	Elle ne nécessite pas la modélisation explicite des critères de recensement.
La vraisemblance prospective	Elle nécessite la modélisation explicite du mode de recensement des familles.	Elle est efficace selon l'étude de Kraft et Thomas.
La vraisemblance rétrospective	• Elle est peu efficace selon l'étude de Kraft et Thomas. • Elle ne tient pas compte du fait que le recensement dépend des génotypes.	Elle ne nécessite pas la modélisation explicite des critères de recensement
La GRL	Elle est probablement peu efficace car basée sur la vraisemblance rétrospective	• Elle ne nécessite pas la modélisation explicite des critères de recensement. • Elle tient compte du fait que le recensement dépend des génotypes.

Chapitre 2

Approche méthodologique du problème

Dans ce chapitre, nous allons mettre de côté le problème du recensement des familles pour exposer la méthodologie commune adoptée dans les différentes vraisemblances qui seront étudiées et/ou développées dans les chapitres 3 et 4. Nous garderons tout de même à l'esprit que les familles sont recensées au travers d'un individu atteint et génotypé que nous appelons "proposant".
Nous décrirons également la façon dont nous avons simulé nos familles. En effet, dans ce travail de thèse, nous avons étudié et comparé des méthodes d'estimation de la pénétrance en terme de biais relatif. Afin de contrôler la "vraie" fonction de pénétrance, nous avons donc simulé des familles sous différentes valeurs de la pénétrance, puis nous avons estimé la fonction de pénétrance à partir de nos échantillons simulés.

Nous commencerons, dans ce chapitre, par décrire les paramètres communs qui interviennent dans les vraisemblances ainsi que l'information utilisée à partir des données familiales. Nous décrirons ensuite la modélisation de la fonction de pénétrance et les paramètres à estimer. Enfin, nous exposerons le principe commun de nos simulations.

2.1 Écriture d'une vraisemblance à partir de données familiales

Les vraisemblances que nous avons étudiées ou développées dans ce travail (elles seront exposées dans les deux prochains chapitres) utilisent l'information phénotypique de tous les individus, même ceux dont le génotype est inconnu. Si l'individu est atteint, nous observerons son âge au diagnostic et s'il est non atteint, son âge aux dernières nouvelles. Il s'agit donc de données censurées.

L'utilisation de données familiales permet aux individus génotypés d'apporter une information sur les individus non génotypés au sein d'une même famille. Dans les différentes vraisemblances, l'introduction des individus non génotypés dans la vraisemblance est donc faite au travers d'une probabilité P_ω qui représente la probabilité de la configuration génotypique ω dans la famille f (i.e. la probabilité jointe des génotypes de la famille f dans la configuration ω). En notant $Gen_{i,\omega}$ le génotype de l'individu i dans la configuration ω et $Gen = (Gen_{1,\omega}, \ldots, Gen_{n_f,\omega})$, le vecteur génotypique de la famille f de taille n_f, on a :

$$P_\omega = \mathbb{P}(Gen_{1,\omega}, ..., Gen_{n_f,\omega}) = \prod_j \mathbb{P}(Gen_{j,\omega}) \prod_{\{l,m,n\}} \mathbb{P}(Gen_{i,\omega}|Gen_{m,\omega}, Gen_{n,\omega}),$$

où j représente le nombre d'individus fondateurs (i.e. les individus n'ayant pas de parents présents dans la généalogie) et $\{l, m, n\}$ représente le triplet d'un individu l et de ses deux parents m et n.

P_ω est donc une fonction de la fréquence de l'allèle délétère dans la population générale (nous faisons l'hypothèse de l'équilibre de Hardy-Weinberg chez les fondateurs), mais cette probabilité dépend également des génotypes des parents (nous faisons l'hypothèse d'une transmission mendelienne) ainsi que du taux de mutation *de novo*.

Pratiquement, P_ω peut se calculer informatiquement en utilisant l'algorithme de Elston et Stewart [21] à condition d'avoir au moins un génotype observé dans la famille (celui du proposant). L'algorithme de Elston et Stewart permet en effet de calculer la probabilité des phénotypes observés pour une généalogie famillial en partant des descendants et en remontant vers l'ancêtre. Pour ce faire, ils s'appuient sur une relation de récurrence entre la probabilité des parents et

celle des enfants.

Dans la suite, la fréquence de l'allèle délétère en population sera notée fq et le taux de mutation *de novo* sera noté pn. Les valeurs de fq et pn seront toujours supposés connus. Dans nos analyses, excepté dans l'étude de la sensibilité à une erreur sur ces paramètres, pn sera supposé nul et fq sera fixé à 10% ou 1% selon la vraisemblance étudiée.

2.2 Modélisation de la fonction de pénétrance et contribution des individus à la vraisemblance

La fonction de pénétrance $F(t_i)$ de l'individu i à l'âge t_i représente le risque cumulé d'être atteint à l'âge t_i conditionnellement au fait d'être porteur de la mutation prédisposante étudiée. La modélisation de la fonction de pénétrance utilise une approche de survie et la contribution de l'individu i à la vraisemblance est donc :

$$\mathbb{P}(Phen_i|Gen_i) = \begin{cases} 1 - F(t_i) & \text{Si } i \text{ n'est pas atteint à l'âge } t_i \\ F(t_i + 1) - F(t_i) & \text{Si } i \text{ est atteint entre } t_i \text{ et } t_i + 1 \end{cases}$$

où $Phen_i$ représente le phénotype de l'individu i et Gen_i son génotype.

Chez les porteurs de la mutation, nous avons modélisé la fonction de pénétrance par le modèle de Weibull étendu par J. Carayol pour l'analyse des données de neuropathie amyloïde héréditaire [38] :

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

où λ et α représentent les paramètres du modèle de Weibull. Les paramètres κ et δ ont été introduits afin d'améliorer la capacité d'ajustement du modèle aux données. Le paramètre κ correspond à la proportion d'individus mutés qui ne développeront jamais la maladie étudiée et δ correspond à l'âge en deçà duquel le risque d'être atteint est nul.

On suppose que la pénétrance chez les individus non porteurs est connue et fixée en fonction de l'incidence de la maladie en population générale.

Dans les deux chapitres suivants, nous avons donc utilisé une **approche paramétrique** qui consiste à faire une hypothèse de distribution (ou loi de probabilité) sur F(t), c'est la partie modélisation, puis dans un second temps à estimer les paramètres inconnus de cette distribution $(\alpha, \lambda, \kappa)$ en maximisant notre vraisemblance.

Cependant, les vraisemblances L que nous décrirons dans les deux prochains chapitres ne vérifient pas les conditions de régularité (c'est-à-dire qu'elles ne sont pas dérivables deux fois en le paramètre). En pratique, l'estimation des paramètres $(\kappa, \lambda, \alpha)$ a donc été obtenue par la maximisation de la log-vraisemblance globale grâce au programme GEMINI [29] qui utilise l'algorithme de Davidon-Fletcher- Powell qui est une méthode dite de quasi-Newton.

2.3 Principe de simulation des familles

Nous avons simulé nos familles sur trois générations de taille et de structure fixe. Pour simuler l'âge des individus au sein des familles, nous avons utilisé des paramètres démographiques obtenus auprès de l'INED (Institut National d'Études Démographiques) [37].
Dans nos simulations, nous avons supposé une transmission autosomique dominante du gène muté avec un risque cumulé variant avec l'âge. Les génotypes ont été attribués de façon aléatoire aux individus fondateurs en fonction de la fréquence de l'allèle délétère en population (i.e. pour les ancêtres et les conjoints). Pour les autres membres de la famille, les génotypes ont été affectés aléatoirement en fonction des lois de Mendel. Nous avons supposé que le taux de mutation *de novo*, pn, était nul. La fréquence de l'allèle délétère, fq, a été fixée à 0.1 ou 0.01 selon la méthode étudiée.
Les phénotypes ont été simulés selon une fonction de pénétrance basée sur le modèle de Weibull étendu (présenté dans la section précédente). Pour plus de simplicité, le paramètre α a été fixé à 3 et les paramètres κ et δ, spécifiques au modèle étendu, ont été fixés à 0 dans nos simulations. Chez les individus porteurs de l'allèle muté, nous avons considéré deux valeurs pour le λ du modèle de Weibull. La première valeur correspond à une pénétrance que nous appellerons **pénétrance basse** et la seconde correspond à une pénétrance que nous appellerons **pénétrance haute**. Les valeurs de ces pénétrances varieront selon la vraisemblance étudiée.

Pour les non porteurs de la mutation, nous avons considéré deux valeurs de risque selon le type de maladie étudiée. Dans le cas de maladies mendéliennes pour lesquelles tous les individus atteints sont porteurs, nous avons considéré logiquement un risque nul d'être atteint chez les non mutés. En revanche, dans le cas de maladies complexes à sous entités monogéniques, pour lesquelles seule une partie des individus atteints est porteur de la mutation prédisposante, nous avons considéré un risque cumulé égal au risque cumulé en population générale.

Dans la suite du manuscrit, nous noterons MM, le modèle de Maladies Mendéliennes et MCSM, le modèle de Maladies Complexes à Sous-entités Monogéniques.

Nous avons ensuite simulé le processus de sélection de nos familles en tentant de calquer la réalité des processus de recensement. La sélection étant dépendante du protocole sous-jacent ainsi que du modèle génétique de la maladie, nous décrirons donc la simulation de la sélection précisément dans les chapitres suivants.

2.4 Conclusion

La méthodologie décrite ci-dessus suffirait à estimer la fonction de pénétrance si les familles étaient recensées aléatoirement en population. Or, ce n'est pas le cas puisqu'elles sont recensées au travers d'un proposant lui-même sélectionné selon des schémas de sélections dépendant du modèle génétique et du protocole sous-jacent. Les vraisemblances utilisées pour estimer la fonction de pénétrance doivent donc impérativement prendre en compte le mode de recensement des familles pour s'affranchir du biais de recensement.

Dans les chapitres 3 et 4, nous avons donc comparé différentes vraisemblances tenant compte du mode de recensement des familles, en terme de biais relatif et d'efficacité relative selon la proportion de génotypes inconnus. Nous avons également développé une méthode d'estimation de la pénétrance avantageuse dans le cas de familles recensées sur des critères indépendants de l'histoire familiale.

Chapitre 3

Etude de la Genotype Restricted Likelihood (ou GRL)

Dans ce chapitre, nous nous intéressons à l'étude d'une méthode d'estimation dans le cas de données familiales recensées sur des critères familiaux. Ces critères sont utilisés lorsqu'aucune autre caractéristique clinique ne permet de différencier les formes héréditaires des autres. C'est le cas en particulier des données issues de familles dans lesquelles une mutation a été recherchée et identifiée chez un individu atteint à la suite d'une consultation de génétique. Cet individu est alors appelé le cas index, il s'agit du premier individu génotypé (il peut y avoir plusieurs proposants mais il n'y a qu'un seul cas index). Si une mutation génétique est trouvée chez le cas index, le test génétique est alors proposé aux membres de sa famille susceptibles de porter cette mutation. Ceux qui sont porteurs font alors l'objet d'une surveillance particulière pour dépister la maladie le plus tôt possible et améliorer le pronostic vital.
Dans la plupart des cas, les critères de recensement sont tels qu'il est impossible de les modéliser formellement, comme c'est le cas, par exemple, dans les recommandations aux cancers du sein que nous avons citées dans le premier chapitre.

L'estimation de la pénétrance à partir de ce type de familles nécessite l'utilisation d'une méthode adaptée permettant de s'affranchir du biais engendré par le recensement. Cette méthode devra tenir compte non seulement du recensement aux moyens de critères familiaux (recensement sur les phénotypes observés) mais aussi du recensement sur les génotypes observés

puisqu'un test génétique est proposé aux apparentés du cas index, uniquement si ce dernier est porteur de la mutation.
La Genotype Restricted Likelihood (ou GRL), proposée par Carayol et Bonaïti-Pellié [9], remplit ces critères et permet donc d'estimer la pénétrance en s'affranchissant complètement du biais de recensement.

Par ailleurs, l'information phénotypique sur une personne est plus facilement accessible que l'information génotypique pour laquelle se pose le problème des génotypes manquants. En effet, un test génétique est proposé aux apparentés du cas index, par l'intermédiaire de ce dernier, mais ceux-ci ne demandent pas toujours à en bénéficier.
Même si l'avantage de disposer d'individus apparentés est que la connaissance du génotype des uns apporte une information sur le génotype des personnes non génotypées, le fait d'avoir un grand nombre de génotypes inconnus peut diminuer significativement l'efficacité de la méthode d'estimation.

Dans ce chapitre, nous décrirons précisément la GRL qui est une méthode d'estimation paramétrique basée sur la modélisation des phénotypes conditionnellement aux génotypes. Puis nous présenterons l'étude d'efficacité relative que nous avons mené sur la GRL. Enfin nous discuterons les résultats.

3.1 La Genotype Restricted Likelihood (ou GRL)

3.1.1 Ecriture de la vraisemblance

La GRL est basée sur la modélisation de la distribution des génotypes conditionnellement aux phénotypes et à la sélection. On note $Phen$, le vecteur phénotypique pour la famille f, Gen le vecteur génotypique et Asc, l'évènement définissant le recensement de la famille f. Avec ces notations, la GRL s'écrit pour une famille f donnée :

$$\begin{aligned} GRL_f &= \mathbb{P}(Gen|Phen, Asc) \\ &= \frac{\mathbb{P}(Asc|Gen, Phen)\mathbb{P}(Phen|Gen)\mathbb{P}(Gen)}{\mathbb{P}(Asc, Phen)} \end{aligned}$$

En notant Ω, l'ensemble des configurations génotypiques possibles, compatibles avec la sélection de la famille f, on a :

$$\begin{aligned} GRL_f &= \frac{\mathbb{P}(Asc|Gen, Phen)\mathbb{P}(Phen|Gen)\mathbb{P}(Gen)}{\sum_{\omega\in\Omega}\mathbb{P}(Asc, Phen, Gen_\omega)} \\ &= \frac{\mathbb{P}(Asc|Gen, Phen)\mathbb{P}(Phen|Gen)\mathbb{P}(Gen)}{\sum_{\omega\in\Omega}\mathbb{P}(Asc|Phen, Gen_\omega)\mathbb{P}(Phen, Gen_\omega)} \\ &= \frac{\mathbb{P}(Asc|Gen, Phen)\mathbb{P}(Phen|Gen)\mathbb{P}(Gen)}{\sum_{\omega\in\Omega}\mathbb{P}(Asc|Phen, Gen_\omega)\mathbb{P}(Phen|Gen_\omega)\mathbb{P}(Gen_\omega)}, \end{aligned}$$

où $\mathbb{P}(Phen|Gen_\omega)$ correspond à la probabilité conditionnelle du vecteur des phénotypes sachant que la famille est dans la configuration génotypique ω.

Carayol et al. ont montré que, pour une famille donnée, la probabilité que la famille soit recensée conditionnellement aux phénotypes et aux génotypes des apparentés était la même quelle que soit la configuration génotypique ω [9] . Pour une famille f de s apparentés, la GRL s'écrit donc :

$$
\begin{aligned}
GRL_f &= \frac{\mathbb{P}(Phen|Gen)\mathbb{P}(Gen)}{\sum_{\omega\in\Omega}\mathbb{P}(Phen|Gen_\omega)\mathbb{P}(Gen_\omega)} \\
&= \frac{\sum_{z\in\Gamma}\prod_{k=1}^{s}\mathbb{P}(Phen_k|Gen_{k,z})\prod_{j=1}^{t}\mathbb{P}(Gen_{j,z})\prod_{h=1}^{u}\mathbb{P}(Gen_{h,z}|Gen_{p(h),z},Gen_{m(h),z})}{\sum_{\omega\in\Omega}\prod_{k=1}^{s}\mathbb{P}(Phen_k|Gen_{k,\omega})\prod_{j=1}^{t}\mathbb{P}(Gen_{j,\omega})\prod_{h=1}^{u}\mathbb{P}(Gen_{h,\omega}|Gen_{p(h),\omega},Gen_{m(h),\omega})}, \\
&= \frac{\sum_{z\in\Gamma}\prod_{k=1}^{s}\mathbb{P}(Phen_k|Gen_{k,z})\prod_{j=1}^{t}\mathbb{P}(Gen_{j,z})\prod_{h=1}^{u}\mathbb{P}(Gen_{h,z}|Gen_{p(h),z},Gen_{m(h),z})}{\sum_{\omega\in\Omega}\prod_{k=1}^{s}\mathbb{P}(Phen_k|Gen_{k,\omega})\ P_\omega},
\end{aligned}
$$

où P_ω est la probabilité décrite au chapitre 2 et dépend de la fréquence de l'allèle délétère fq et du taux de mutation de *novo* pn, supposés connus. Γ est l'ensemble des configurations génotypiques compatibles avec les génotypes des apparentés testés, t représente le nombre d'individus fondateurs (c'est-à-dire les individus dont les parents ne sont pas dans la généalogie) et u le nombre d'individus dont les parents sont connus. Pour un individu h donné, $Gen_{m(h)}$ et $Gen_{p(h)}$ représentent respectivement les génotypes de sa mère et de son père.
$\mathbb{P}(Gen_h|Gen_{p(h)}, Gen_{m(h)})$ correspond donc à la probabilité conditionnelle qu'un individu h soit de génotype Gen_h sachant le génotype de ses parents ($Gen_{m(h)}$ et $Gen_{p(h)}$).

Comme nous l'avons décrit dans le chapitre 2, la modélisation de la pénétrance utilise une approche d'analyse de survie basée sur le modèle de Weibull étendu. On rapelle que le risque chez les non porteurs de la mutation est supposé connu et fixé en fonction de l'incidence de la maladie en population générale.

3.2 Etude de la GRL et de son efficacité relative

Pour démontrer que la GRL est asymptotiquement sans biais, Carayol et Bonaïti-Pellié s'étaient placés dans une situation idéale où les génotypes de tous les apparentés étaient connus. Mais la GRL permet aussi l'utilisation de l'information phénotypique d'individus non génoty-

pés. Cependant, le nombre d'individus non génotypés peut être assez conséquent du fait même du protocole car les apparentés du cas index ne demandent pas toujours le test génétique.
Lorsque peu d'apparentés sont génotypés dans les familles, des problèmes d'efficacité peuvent alors être rencontrés, c'est-à-dire que la variance des estimations obtenues peut croître de façon importante avec la proportion de génotypes inconnus.

Nous avons donc entrepris d'étudier l'efficacité relative de la GRL, dans le cas où les génotypes sont manquants, au travers d'une étude de simulations. Nous avons également étudié la méthode ainsi que son efficacité relative en fonction des individus non génotypés introduits dans les familles afin de mesurer l'information apportée par ces derniers.

3.2.1 Simulations

Afin d'étudier l'efficacité relative de la GRL dans le cas où de nombreux individus n'ont pas été génotypés, nous avons simulé un grand nombre de familles sous différentes valeurs de pénétrance pour les individus porteurs. Nous avons utilisé le principe de simulation décrit dans le chapitre 2. La structure de nos familles était la suivante : un couple ancêtre ayant 4 enfants, chacun ayant eux-mêmes quatre enfants.
Pour obtenir des échantillons de familles sélectionnées suffisamment grands sans avoir à simuler un trop grand nombre de familles, la fréquence de l'allèle délétère a été fixée à 0.10 dans nos simulations. Nous avons supposé que le taux de mutation *de novo* était nul.
Pour la simulation des phénotypes, nous avons considéré la cas d'une **pénétrance haute**, qui correspond à un risque de 0.50 à 80 ans chez les porteurs ; et d'une **pénétrance basse**, qui correspond à un risque de 0.2 à 80 ans chez les porteurs. Le risque chez les non porteurs de la mutation ayant été fixé à 0.02 à 80 ans (correspondant au risque cumulé de cancer en population générale dans le syndrome HNPCC (cf. Annexe F).

Nous avons ensuite simulé le processus de sélection des familles. Pour cela, nous avons sélectionné les familles dans lesquelles au moins deux apparentés étaient atteints, dont l'index qui porte nécessairement la mutation.

La condition nécessaire pour qu'une famille entre dans l'analyse est donc qu'une mutation ait été identifiée chez le cas index. Cependant, pour que la famille soit informative, il faut qu'au moins un apparenté du cas index soit génotypé. En effet, dans la GRL, la somme au dénominateur s'effectue sur Ω qui est défini comme l'ensemble des configurations génotypiques dans lesquelles le cas index est porteur d'une mutation. Au numérateur, la somme s'effectue sur Γ qui définit l'ensemble des configurations génotypiques compatibles avec les génotypes des apparentés testés. Si aucun autre individu que le cas index n'a pu être testé, alors Ω et Γ sont identiques et la vraisemblance pour la famille vaut 1. Il est donc indispensable d'avoir un individu génotypé en plus du cas index.
Nous avons imposé des échantillons de 10 000 familles après la sélection afin de se placer dans les conditions asymptotiques.

3.2.2 Calcul de l'efficacité

La perte et le gain d'efficacité de la GRL ont été étudié en calculant l'Efficacité Relative Asymptotique (ARE) des pénétrances estimées, qui est l'inverse du rapport de la variance estimée dans une situation donnée, à la variance obtenue dans une situation de référence :

$$ARE = \frac{VAR_{\textit{Situation de référence}}}{VAR_{\textit{Situation étudiée}}}$$

Pour estimer la variance de la pénétrance dans chacune des situations, nous avons simulé 1000 réplicats d'échantillons de 10 000 familles, puis nous avons considéré l'estimation de la pénétrance à 70 ans, qui est le plus souvent utilisé comme référence dans la littérature, et avons calculé la variance empirique en ce point.
Soit $(p_1^{70}, ..., p_{1000}^{70})$, l'échantillon de taille 1000 des estimations de la pénétrance au point $t = 70$, on a :

$$\widehat{VAR} = \frac{1}{1000}\sum_{i=1}^{1000}(p_i^{70} - \overline{p}^{70}),$$

où $\overline{p}^{70}$ est la moyenne empirique de l'échantillon.

Nous nous attendons à ce que la distribution des génotypes observés dépende du nombre d'apparentés testés. De par sa définition, la GRL est dépendante de cette distribution. Nous avons donc étudié la GRL en fonction de cette variable. Nous avons constitué différents échantillons en affectant le statut "génotype connu" ou "génotype inconnu" aux apparentés du cas index à partir d'un même échantillon de familles sélectionnées où les génotypes étaient simulés. Puis, nous avons comparé l'efficacité relative asymptotique de la GRL dans les différentes situations considérées. Ainsi, si l'efficacité relative asymptotique (ARE) est supérieure à 1, cela signifie que la GRL est plus efficace sur l'échantillon considéré que sur l'échantillon de référence. Si l'ARE vaut 1, c'est que la GRL a la même efficacité sur les deux échantillons.

3.2.3 Résultats

Le tableau 3.1 présente l'efficacité de la GRL en fonction de la proportion de génotypes inconnus dans la famille. Quelle que soit la valeur de la pénétrance, l'efficacité relative de la GRL diminue avec le nombre d'individus génotypés par famille. Dans le cas où 75% des génotypes sont inconnus, l'efficacité relative de la GRL diminue de plus de la moitié, quelle que soit la valeur de la pénétrance. Dans le cas "minimal" où seul l'index et l'un de ses apparentés sont génotypés, l'ARE de la GRL est de seulement 7%, et ceci que la pénétrance soit basse ou haute.
La GRL est donc peu efficace dans le cas où tous les génotypes sont inconnus à part celui de l'index et de l'un des ses apparentés. La GRL étant une vraisemblance de type rétrospective basée sur la modélisation des génotypes sachant les phénotypes, nous nous attendions à une baisse significative de l'efficacité de la méthode avec l'augmentation des génotypes inconnus.

Nous avons également entrepris d'étudier les estimations de la pénétrance obtenues avec la GRL selon les individus inclus dans la famille. Pour cela, nous avons étudié différentes situations à partir du même échantillon sélectionné sur l'existence d'un individu atteint et muté (l'index) ainsi qu'un autre individu au moins atteint dans la famille nucléaire de l'index. La figure 3.1 représente la structure des familles simulées ainsi que les différentes situations d'ap-

Situation	Référence	ARE	
		Pénétrance basse	**Pénétrance haute**
Proportion de génotypes inconnu			
❖ 25%	Tous les génotypes sont connus	0,99	0,88
❖ 50%		0,74	0,70
❖ 75%		0,47	0,41
❖ Seuls deux genotypes sont connus (Celui du cas index et celui de l'un de ces apparenté)		0,07	0,07

TAB. 3.1 – Efficacité de la GRL selon la proportion de génotypes inconnus.

parentés introduits dans les familles pour étudier la GRL selon les données incluses dans la famille.

Les situations étudiées sont les suivantes : (A) Seule la famille nucléaire de l'index est incluse et seul l'index et l'un de ces apparentés sont génotypés ; (B) La famille nucléaire de l'index et le ou les autre(s) famille(s) nucléaire(s) avec au moins un atteint entrent dans l'analyse, mais sans apporter d'information génotypique supplémentaire ; (C) La famille nucléaire de l'index et le ou les autre(s) famille(s) nucléaire(s) avec au moins deux atteints sont incluses, mais n'apportent pas d'information génotypique supplémentaire par rapport à (A) ; (E) Tous les individus de la famille entrent dans l'analyse mais sans apporter d'information génotypique supplémentaire par rapport à la situation (A).

Pour comparer les estimations de la pénétrance obtenues selon les individus inclus dans la famille, nous avons calculé la moyenne de l'estimation obtenue à 80 ans pour 1000 réplicats d'échantillons de 10000 familles. Le tableau 3.2 présente les résultats obtenus.

On constate que les estimations obtenues sont différentes. Ce résultat est plus évident dans

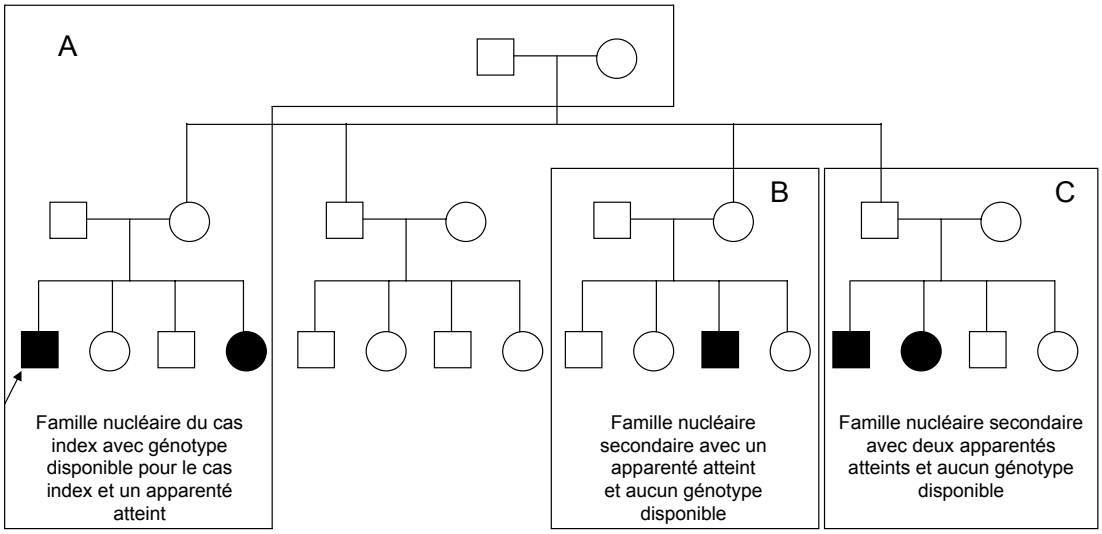


FIG. 3.1 – Structure des familles avec : (**1**) La famille nucléaire de l'index et le couple ancêtre (famille A) ; (**2**) famille A + les membres de la famille nucléaire secondaire avec un atteint et pas de génotype connu (famille de type A+B+C ou A+B) ; (**3**) famille A + la famille nucléaire secondaire avec deux apparentés atteints et pas de génotype connu (famille de type A+C) ; (**4**) tous les membres de la famille. Le cas index est marqué par une flèche.

le cas d'une pénétrance haute. L'inclusion de l'ensemble des individus de la famille donne une estimation proche de celle obtenue lorsque seule la famille nucléaire de l'index est incluse. Par contre, l'estimation apparaît biaisée dans le sens d'une surestimation du risque quand on inclut les familles nucléaires secondaires avec au moins un atteint (A+B+C dans la généalogie 3.1). Lorsque l'on inclut les familles nucléaires secondaires avec au moins deux atteints (A+C dans la généalogie 3.1), la moyenne obtenue est identique à celle obtenue dans le cas où tous les individus de la généalogie sont inclus.

Nous avons ensuite étudié l'efficacité relative de la GRL uniquement dans le cas où toute l'information des branches sans génotypes est inclus, en prenant comme référence la famille nucléaire de l'index. Dans le cas d'une pénétrance basse, nous avons trouvé une $ARE = 1,04$ et dans le cas d'une pénétrance haute, l'ARE était de 1. Quelle que soit la valeur de la pénétrance, l'information apportée par les branches de la famille sans génotype n'augmente donc pas l'efficacité de la GRL (ARE proche ou égale à 1).

Situation	Moyenne de l'estimation à 80 ans à partir de 1000 réplicats	
	Pénétrance basse	**Pénétrance haute**
Données incluses dans l'analyse (Fig. 2.1)		
❖ Famille nucléaire de l'index (fam. A)	0.18	0.48
❖ Toute l'information des branches sans génotype	0.18	0.49
❖ Famille nucléaire de l'index + famille nucléaire secondaire avec au moins 1 atteint (fam. A+B+C)	0.19	0.56
❖ Famille nucléaire de l'index + famille nucléaire secondaire avec au moins 2 atteints (fam. A+C)	0.18	0.49

TAB. 3.2 – Moyenne des estimations de la pénétrance obtenue à 80 ans selon les données incluses dans la famille

3.3 Conclusion et discussion

La GRL, développée par Carayol et Bonaïti-Pellié, est une méthode qui permet d'obtenir des estimations asymptotiquement sans biais de la pénétrance quels que soient les critères ayant permis le recensement des familles. Ceci est très utile, notamment dans le cas où les familles ont été recensées à l'aide de critères complexes, particulièrement difficiles à modéliser. Cette méthode a donc l'avantage de pouvoir être appliquée à l'estimation du risque de n'importe quelle maladie, pour laquelle des mutations ont été identifiées et quels que soient les critères de recensement, connus ou non.

Nous avons montré dans ce chapitre, à l'aide de simulations, que cette méthode était peu efficace lorsqu'une proportion de génotypes était inconnue. Pourtant, il arrive très fréquemment d'avoir un grand nombre de génotypes inconnus chez les apparentés du cas index. Par exemple, le cas où seul l'index et l'un de ses apparentés sont génotypés n'est pas rare. Nous avons vu dans ce cas que l'efficacité de la méthode était très faible (moins de 10%). Nous avons également montré que l'utilisation de toute l'information phénotypique disponible ne permettait pas d'augmenter l'efficacité de la GRL.
De plus, nous avons constaté que la pénétrance était surestimée dans le cas où les familles nucléaires secondaires, possédant au moins un atteint, étaient ajoutées à la famille nucléaire de l'index lors de l'estimation. En effet, dans ce cas, une sélection est opérée à l'intérieur des familles. Or, la GRL corrige sur une sélection des familles et non pas sur une sélection à l'intérieur des familles. Ceci peut donc expliquer la "surestimation" obtenue dans ce cas.
Avec ce raisonnement, la pénétrance des données devrait aussi être augmentée lorsque l'on inclut à la famille nucléaire de l'index, les familles nucléaires secondaires avec au moins deux atteints. Pourtant, les estimations sont proches de celles obtenues avec uniquement la famille nucléaire de l'index. Ce résultat peut s'expliquer par le fait que, avec une pénétrance haute de 0.5 à 80 ans et une pénétrance basse de 0.2 à 80 ans, les familles possédant au moins deux atteints dans une famille nucléaire secondaire de l'index sont assez rares. L'échantillon obtenu dans ce cas est alors relativement semblable à l'échantillon obtenu lorsque seule la famille nucléaire de l'index est pris en compte et ceci peut expliquer que les moyennes des estimations

obtenues soient proches elles aussi.

La GRL est basée sur la vraisemblance rétrospective dont Kraft et Thomas ont montré qu'elle était peu efficace. La GRL a été utilisée pour l'estimation de la fonction de pénétrance dans le syndrome HNPCC. Les intervalles de confiance qui ont alors été calculés illustrent également le manque d'efficacité de la méthode [2]. Par conséquent, lorsque les données sont recensées à partir de critères indépendants de l'histoire familiale et que les critères de recensement sont alors plus simples à modéliser, il est préférable de développer d'autres méthodes, corrigeant également pour le biais de recensement, et qui seraient plus efficaces que la GRL. Ceci a fait l'objet de la suite de notre travail dans les chapitres 3 et 4.

Le travail que nous venons de présenter dans ce chapitre a fait l'objet d'une publication (Annexe F) dans *European Journal of Human Genetics* en 2007.

Chapitre 4

Recensement sur critères indépendants de l'histoire familiale

Dans ce chapitre, nous nous intéressons aux données recensées sur l'existence d'au moins un individu atteint dans la famille. Dans le cas d'une maladie mendélienne, ce type de recensement suffit à détecter la mutation prédisposante dans la famille. Dans le cas des maladies complexes à sous-entités monogéniques, l'inclusion d'un critère d'âge au recensement peut suffire à détecter la mutation prédisposante puisque les sous-entités monogéniques surviennent souvent à des âges précoces.
Comme dans le cas du recensement sur critères familiaux, l'estimation de la pénétrance à partir de familles recensées sur critères indépendants de l'histoire familiale nécessite l'utilisation de méthodes prenant en compte ces critères afin d'obtenir des estimations sans biais. Toutefois, les critères indépendants de l'histoire familiale sont moins complexes et plus faciles à modéliser et il est, dans ce cas, préférable d'utiliser des méthodes plus efficaces que ne l'est la GRL et notamment des méthodes dont l'efficacité ne diminue pas de manière drastique lorsque un grand nombre de génotypes est inconnu.

La vraisemblance prospective est une méthode d'estimation de la pénétrance qui s'adapte bien à ce type de recensement. Elle a été utilisée par Planté-Bordeneuve et al. [38] pour estimer la pénétrance dans la neuropathie amyloïde héréditaire. Cette vraisemblance prend en compte le recensement en le modélisant mathématiquement. Cependant, elle repose sur certaines hy-

pothèses qui ne sont pas toujours vérifiées, ce qui peut entraîner un biais. C'est pourquoi nous avons développé une méthode d'estimation moins sensible à ces hypothèses, qui corrige également pour le recensement : la Proband's phenotype Exclusion Likelihood (ou PEL).

Nous commencerons, dans ce chapitre, par présenter ces deux vraisemblances : la Prospective et la PEL. Puis, nous présenterons les avantages de la PEL par rapport à la vraisemblance prospective ainsi que les différentes études de sensibilité que nous avons menées. Nous appliquerons ensuite les méthodes à deux jeux de données : l'un sur le cancer du sein associé aux mutations BRCA1 et BRCA2 à partir de familles recensées sur l'existence d'une femme atteinte avant 36 ans ; l'autre sur la neuropathie amyloïde héréditaire à partir de données françaises et portugaises. Enfin, nous présenterons l'étude que nous avons menée sur la NAH (Neuropathie Amyloïde Héréditaire) issue de la population suédoise.

4.1 La méthode prospective

La vraisemblance prospective correspond à la probabilité des phénotypes sachant les génotypes et le recensement. Comme la GRL, la méthode prospective est une méthode d'estimation basée sur le maximum de vraisemblance qui utilise une approche d'analyse de survie (cf. Chapitre 2). Cette méthode corrige pour le recensement des familles sur l'existence d'au moins un individu atteint en modélisant la probabilité de cet évènement.

4.1.1 Ecriture de la vraisemblance prospective

Pour une famille f donnée de n_f individus, la vraisemblance prospective s'écrit :

$$\begin{aligned} VP_f &= \mathbb{P}(Phen|Gen_{obs}, Asc) \\ &= \frac{\mathbb{P}(Phen|Gen_{obs})\mathbb{P}(Asc|Phen, Gen_{obs})\mathbb{P}(Gen_{obs})}{\mathbb{P}(Asc|Gen_{obs})\mathbb{P}(Gen_{obs})} \\ &= \frac{\mathbb{P}(Phen|Gen_{obs})\mathbb{P}(Asc|Phen, Gen_{obs})}{\mathbb{P}(Asc|Gen_{obs})} \end{aligned}$$

où,

* $Phen$ représente le vecteur phénotypique pour les n_f individus de la famille f, $Phen = (Phen_1, ..., Phen_{n_f})$.
* Gen_{obs} représente le vecteur des génotypes observés dans la famille f.
* Asc représente l'évènement "la famille vérifie les critères de sélection" (i.e. existence d'au moins un individu atteint dans la famille).

Comme une famille est recensée si elle possède au moins un individu atteint, la probabilité de recensement ne dépend que du phénotype des individus et non de leur génotype. On a donc :

$$\mathbb{P}(Asc|Phen, Gen_{obs}) = \mathbb{P}(Asc|Phen) = C$$

où C est une constante qui ne dépend pas des paramètres à estimer. Nous pourrons donc éliminer cette constante lors de la maximisation de la vraisemblance.

Si on suppose l'indépendance entre les n familles de l'échantillon, la vraisemblance totale s'écrit :

$$VP = \prod_{f=1}^{n} VP_f = \prod_{f=1}^{n} \frac{N_f}{D_f},$$

avec $N_f = C.\mathbb{P}(Phen|Gen_{obs})$

et $D_f = \mathbb{P}(Asc|Gen_{obs})$

Afin d'introduire les individus non génotypés dans la vraisemblance, on écrit :

$$N_f = C.\sum_{\omega=1}^{\Omega} P_\omega.\mathbb{P}(Phen|Gen_\omega),$$

où :

* P_ω représente la probabilité de la configuration génotypique ω, déjà décrite dans le chapitre 2.
* Ω représente l'ensemble des configurations génotypiques possibles et dépend donc du nombre de génotypes inconnus.
* Gen_ω représente le vecteur génotypique dans la configuration ω.

Si on note $p_{i,\omega}$, la probabilité pour un individu i d'être atteint dans la configuration ω et si on note π, la probabilité pour un individu atteint d'être sélectionné, la correction pour le recensement au dénominateur peut s'écrire :

$$\begin{aligned} D_f &= \mathbb{P}(Asc|Gen_{obs}) \\ &= \sum_{\omega=1}^{\Omega} P_\omega.\mathbb{P}(oneaf), \end{aligned}$$

où, $\mathbb{P}(oneaf)$ est la probabilité que, dans la configuration ω, au moins un individu atteint soit recensé dans la famille :

$$\begin{aligned} \mathbb{P}(oneaf) &= 1 - \mathbb{P}(\textit{aucun individu n'est sélectionné dans la famille}) \\ &= 1 - \prod_{i=1}^{n_f} \Big((1 - p_{i,\omega}) + p_{i,\omega}(1 - \pi) \Big) \end{aligned}$$

On rappelle que la fonction de pénétrance utilise une approche de survie basée sur le modèle de Weibull. On a donc

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

Le paramètre δ (i.e. l'âge en deçà duquel le risque d'être atteint est nul) est supposé connu. Les paramètres à estimer sont donc $(\kappa, \lambda, \alpha)$.

4.1.1.1 Estimation du paramètre π

Théoriquement, la probabilité π pour un individu atteint d'être sélectionné pourrait être estimée conjointement aux paramètres de la fonction de pénétrance $(\kappa, \lambda, \alpha)$. Cependant, l'estimation conjointe de ces quatre paramètres est difficilement faisable en raison du trop grand nombre de paramètres. Nous avons donc choisi d'estimer π au préalable, puis de le fixer dans l'équation de la vraisemblance, comme cela est le plus souvent fait en analyse de ségrégation.
En supposant que tous les individus sont indépendamment recensés, π peut être estimé par maximum de vraisemblance à partir de la distribution des proposants parmi les individus atteints dans les familles avec plusieurs atteints (les familles multiplexes). Soit $X_{proposant}$ la variable aléatoire représentant le nombre de proposants parmi r atteints, $X_{proposant}$ suit la loi Binomiale tronquée suivante [33] :

$$\mathbb{P}(X_{proposant} = a) = \frac{\binom{r}{a} \pi^a (1-\pi)^{r-a}}{1-(1-\pi)^r}, \forall 1 \leqslant a \leqslant r,$$

avec r, le nombre d'atteints.

4.1.1.2 Prise en compte d'un critère d'âge

Parfois, un critère d'âge est introduit dans le recensement afin d'augmenter la probabilité d'observer la mutation dans la famille. Dans ce cas, la famille est recensée s'il y a au moins un individu atteint avant un certain âge, noté $agesel$. Ce nouveau paramètre est alors pris en compte au dénominateur de la vraisemblance de la manière suivante. Pour un individu i d'age $age_{courant}$:

- Si $age_{courant} \geq agesel$, la probabilité pour i de ne pas être atteint avant $agesel$ est :

$$1 - F(agesel)$$

.

- Si $age_{courant} \leq agesel$, la probabilité pour i de ne pas être atteint avant $agesel$ est :

$$1 - F(age_{courant})$$

.

4.1.2 Étude de la méthode prospective

La méthode prospective repose sur l'hypothèse que tous les individus atteints ont la même probabilité d'être recensé. Or, ce n'est pas le cas dans la pratique puisque le recensement a lieu durant une période donnée et qu'un critère d'âge est parfois inclus pour la sélection. Nous avons donc simulé des familles sous différents modèles de maladie avec différents modes de sélection, "les plus réalistes possibles" et dans lesquels les individus peuvent avoir des probabilités différentes d'être sélectionnés. Puis, nous avons étudié la vraisemblance prospective en terme de biais relatif et d'efficacité relative.

4.1.2.1 Simulations

Nous avons simulé des familles de taille et de structure fixes comme le montre la figure 4.1 : un couple ancêtre avec quatre enfants, chacun ayant deux enfants.

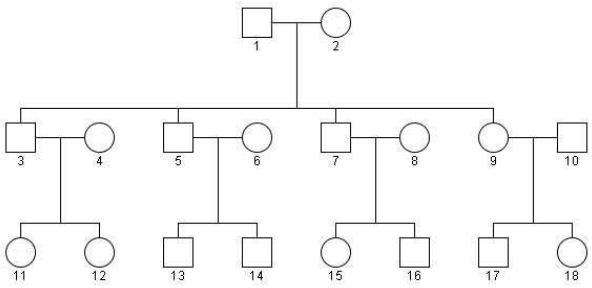


FIG. 4.1 – Structure des familles simulées

L'âge et le génotype des individus ont été simulés comme nous l'avons décrit au chapitre 2. Dans nos simulations, les paramètres fq et pn ont été fixé respectivement à 0.01 et 0.
Chez les individus mutés, les phénotypes ont été simulés selon deux valeurs de la pénétrance : la **pénétrance haute** correspondant à une pénétrance de 0.8 à 80 ans et la **pénétrance basse** correspondnt à une pénétrance de 0.5 à 80 ans. Pour les non porteurs de la mutation, nous avons considéré deux valeurs de risque selon le type de maladie étudiée. Dans le cas de maladies mendéliennes (MM) pour lesquelles tous les individus atteints sont porteurs, nous avons considéré

logiquement un risque nul d'être atteint chez les non mutés. En revanche, dans le cas de maladies complexes à sous-entités monogéniques (MCSM), pour lesquelles seule une partie des individus atteints est porteur de la mutation prédisposante, nous avons considéré un risque cumulé de 0.10 à 80 ans. Ce choix correspond à la valeur du risque cumulé de cancer du sein en population générale, qui fera l'objet de notre application concernant les maladies complexes à sous-entités monogéniques.

Afin de sélectionner les individus de la façon la plus réaliste possible, nous avons défini des périodes de temps, que nous avons notées T, pour la sélection du proposant. Ainsi, nous avons supposé que seuls les individus porteurs et atteints durant cette période T pouvaient être proposant avec une probabilité p_s. Ces périodes de temps traduisent le fait que les recueils de données se font pendant des périodes de temps données. La probabilité p_s traduit le fait que, dans ces périodes, tous les individus atteints ne seront pas nécessairement sélectionnés.
Nous avons considéré deux périodes de temps distinctes : une période de 20 ans, qui correspond donc à $T = 20$, pour laquelle un grand nombre d'individus atteints peuvent être proposants, et une période de 1 an, qui correspond à $T = 1$, pour laquelle la probabilité d'avoir plus d'un individu atteint qui soit proposant dans une famille est négligeable. Une famille sera donc sélectionnée dans l'échantillon si elle contient au moins un proposant.

Dans le cas des maladies complexes à sous-entités monogéniques les cas dus à la mutation prédisposante surviennent souvent beaucoup plus tôt que les cas sporadiques. C'est pourquoi un critère d'âge est souvent introduit lors de la sélection, comme nous l'avons déjà mentionné [5, 16]. Comme l'âge de 36 ans est le critère d'âge utilisé dans la sélection des familles de cancer du sein utilisées dans l'application qui suivra, nous avons pris le parti d'utiliser ce même critère d'âge dans nos simulations. Ainsi, dans le modèle MCSM, un individu atteint avant 36 ans, durant la période T, sera proposant avec une probabilité p_s et on sélectionne les familles avec au moins un proposant.

4.1.2.2 Calcul du biais relatif

Afin d'étudier la vraisemblance prospective selon le modèle génétique et le mode de recensement, nous avons calculé empiriquement le biais relatif à partir de 1000 réplicats d'échantillons de 10000 familles, et ceci, au point $t = 70$ ans car c'est l'âge le plus couramment utilisé dans les études [2, 17, 22, 24]. Le biais relatif s'écrit alors de la façon suivante :

$$\hat{B} = \frac{1}{1000} \sum_{i=1}^{1000} \left(\frac{\hat{R}_i - R_0}{R_0} \right),$$

où R_i est la pénétrance estimée à 70 ans pour le réplicat i et R_0 est la vraie valeur de la pénétrance à 70 ans.

4.1.2.3 Calcul de l'efficacité

Comme dans l'étude de la GRL dans le chapitre 3, l'efficacité de la méthode prospective a été étudiée à l'aide de l'ARE (Efficacité Relative Asymptotique) des pénétrances estimées :

$$ARE = \frac{VAR_{\textit{Situation de référence}}}{VAR_{\textit{Situation étudiée}}}$$

Pour estimer la variance de la pénétrance dans chacune des situations, nous avons simulé 1000 réplicats d'échantillons de 10 000 familles, puis nous avons conservé l'estimation de la pénétrance à 70 ans et nous avons calculé la variance empirique en ce point (cf. Section 3.2).

4.1.2.4 Résultats de l'étude de biais et d'efficacité de la vraisemblance prospective

Le tableau 4.1 montre le biais relatif obtenu avec la vraisemblance prospective sous différents modèles génétiques, divers paramètres de sélection, avec une pénétrance haute et une pénétrance basse et dans le cas où tous les individus sont de génotype connu. Nous avons vérifié que l'introduction de différentes proportions de génotypes inconnus, variant de 50% au cas extrême où seul le proposant est génotypé, n'influait pas sur les estimations de la pénétrance et ne modifiait donc pas le biais relatif. Dans le modèle MM, on constate que l'estimation de π varie assez peu et est de l'ordre de 50%. On constate également, dans le modèle MM, que la

méthode est quasiment sans biais, quel que soit le modèle de recensement et quelle que soit la valeur de la pénétrance. Par contre, dans le modèle MCSM, la méthode prospective a un biais non négligeable, en particulier lorsque la pénétrance est basse et lorsque $T = 1$, c'est-à-dire lorsque la probabilité de sélection est très faible.
La vraisemblance prospective, dans sa correction au dénominateur, fait l'hypothèse que tous les individus ont la même probabilité d'être sélectionnés. Or, cette hypothèse n'est pas toujours vérifiée, en particulier dans le modèle MCSM où la probabilité d'être sélectionné pour les individus atteints en dehors de la période T est nulle. Cet écart aux hypothèses de la vraisemblance prospective peut être une explication du biais de la méthode sous le modèle MCSM.

Lors de l'analyse avec la méthode prospective, la probabilité pour un individu atteint d'être sélectionné π a été estimée grâce à la distribution des proposants parmi les atteints dans les familles multiplexes (familles avec plusieurs atteints), comme nous l'avions décrit précédemment. Cependant dans le modèle MCSM, il n'y a qu'un unique proposant par famille et l'estimation du π est alors impossible. Dans un tel cas, nous avons choisi de fixer π à 0.01. Le tableau 4.2 montre la sensibilité de la vraisemblance prospective à des valeurs arbitraires du paramètre π. Nous avons étudié les mêmes modèles génétiques et les mêmes situations de sélections que lors de l'étude du biais relatif mais au lieu d'estimer π, nous l'avons fixé arbitrairement, respectivement à 1%, 10%, et 50% dans l'analyse. Nous avons indiqué, entre parenthèses, la différence obtenue entre le biais relatif calculé avec le π fixé et le biais relatif calculé avec le π estimé ($B_{\pi} - B_{\hat{\pi}}$). Dans le modèle MM, la méthode présente une sensibilité non négligeable à la valeur de π. Dans ce modèle, nous avons vu que les estimations du π étaient proches de 0.5. Quand π est fixé à 1%, le biais relatif obtenu quand $ps = 1$ et $T = 20$, dans le cas d'une pénétrance basse, est de -13% alors qu'il était de -3% lorsque π était estimée. La sensibilité apparaît moins importante pour une pénétrance haute et sous le modèle MM. Finalement, plus la valeur du π fixée arbitrairement est proche de sa valeur estimée, moins les estimations sont biaisées.

Le tableau 4.3 présente l'étude d'efficacité de la vraisemblance prospective en fonction de la proportion d'individus génotypés dans la famille. Comme dans l'étude d'efficacité de la GRL, la situation de référence est celle où tous les individus sont génotypés. Contrairement à la GRL, la

vraisemblance prospective reste efficace dans le cas où une proportion de génotype est inconnu, quelle que soit la valeur de la pénétrance. Le cas où seul le proposant est génotypé est un cas que l'on rencontre couramment. Même dans ce cas extrême, la méthode prospective demeure efficace avec une ARE variant de 59% à 95% selon les situations.

Modèle génétique (*)	p_s (*)	periode T (en année)	π	Bais relatif (en %)	
				Pénétrance haute	**Pénétrance basse**
MM	1	20	0.55	0	-3
		1	0.45	3	2
	0.5	20	0.40	1	0
		1	0.28	1	-2
MCSM	1	20	(**)	5	14
		1	(**)	9	21
	0.5	20	(**)	5	14
		1	(**)	10	21

(*) MM : maladie mendélienne; MCSM : maladie complexe à sous-entités monogéniques ; p_s : probabilité d'être sélectionné pour un individu atteint durant la période T (et avant 36 ans dans le cas de MCSM). Le paramètre π (probabilité pour un atteint d'être sélectionné) est estimé à partir d'une loi binomiale tronquée. (**) π est fixé à 0.01.

TAB. 4.1 – Biais relatif de la vraisemblance prospective selon le modèle génétique et la sélection

Modèle	p_s (*)	Periode T (en année)	Biais relatif (en %) avec une valeur fixée de π: B_π (difference avec le biais obtenu en utilisant π: $B_\pi - B_\pi$)	
			$\pi = 1\%$	
			Pénétrance haute	Pénétrance basse
MM	1	20	-5 (-5)	-13 (-10)
		1	-2 (-5)	-5 (-7)
	0.5	20	-3 (-4)	-6 (-6)
		1	-2 (-3)	-3 (-5)
MCSM	1	20	4 (-1)	12 (-2)
		1	9 (0)	21 (0)
	0.5	20	4 (-1)	12 (-2)
		1	9 (-1)	20 (-1)
			$\pi = 10\%$	
			Pénétrance haute	Pénétrance basse
MM	1	20	-4 (-4)	-12 (-9)
		1	0 (-3)	-2 (-4)
	0.5	20	-7 (-8)	-3 (-3)
		1	-1 (-2)	-1 (1)
MCSM	1	20	4 (-1)	13 (-1)
		1	9 (0)	21 (0)
	0.5	20	5 (0)	13 (-1)
		1	9 (-1)	21 (0)
			$\pi = 50\%$	
			Pénétrance haute	Pénétrance basse
MM	1	20	0 (0)	-1 (2)
		1	3 (0)	4 (2)
	0.5	20	2 (1)	5 (5)
		1	3 (2)	7 (9)
MCSM	1	20	5 (0)	14 (0)
		1	10 (1)	22 (1)
	0.5	20	6 (1)	15 (1)
		1	10 (0)	22 (1)

(*) p_s *: probabilité d'être sélectionné pour un individu atteint durant la période* T.

TAB. 4.2 – Sensibilité de la vraisemblance prospective à une erreur de la valeur de π

Situation	p_s (*)	Periode T	ARE	
			Pénétrance basse	**Pénétrance haute**
Proportion de génotypes inconnus	1	20		
❖ 50%			1,00	0,95
❖ 75%			0,96	0,93
❖ Seul le proposant est génotypé			0,88	0,88
	1	1		
❖ 50%			0,72	0,94
❖ 75%			0,71	0,94
❖ Seul le proposant est génotypé			0,59	0,88
	0.5	20		
❖ 50%			1,00	0,99
❖ 75%			0,99	0,94
❖ Seul le proposant est génotypé			0,90	0,88
	0.5	1		
❖ 50%			0,90	0,99
❖ 75%			0,89	0,98
❖ Seul le proposant est génotypé			0,85	0,95

TAB. 4.3 – Efficacité relative de la vraisemblance prospective dans le cas du modèle MM selon la proportion de génotypes manquants par rapport à la situation de référence où tous les génotypes sont connus.

La vraisemblance prospective apparaît donc comme une méthode efficace lorsque de nombreux génotypes sont manquants et quasiment sans biais dans le cas d'un modèle MM. Cependant, elle présente un biais non négligeable dans le cas d'un modèle MCSM. Afin de réduire le biais relatif, nous avons développé la Proband's phenotype Exclusion Likelihood (ou PEL) que nous présentons dans la section suivante.

4.2 La Proband's phenotype Exclusion Likelihood (ou PEL)

La Proband's phenotype Exclusion Likelihood est une méthode que nous avons développée afin de réduire le biais relatif de la vraisemblance prospective, notamment dans les modèles MCSM. Comme la vraisemblance prospective et comme la GRL, il s'agit d'une méthode d'estimation de la pénétrance basée sur le maximum de vraisemblance et qui utilise une approche d'analyse de survie. La PEL corrige pour le biais de recensement mais a l'avantage de ne pas nécessiter la modélisation explicite des critères de sélection et donc, de ne pas faire d'hypothèses sur la probabilité de sélection des individus. Après avoir décrit cette méthode, nous l'avons étudiée en termes de biais relatif, d'efficacité relative mais également de robustesse dans diverses situations.

4.2.1 Ecriture de la PEL

Le principe de la PEL est de corriger pour le recensement en retirant l'information phénotypique du proposant et en dupliquant les familles lorsqu'il y a plusieurs proposants dans la famille.
Ce principe de "retirer le proposant" fut introduit par Weinberg en 1912 [47] afin d'estimer le rapport de ségrégation pour la descendance de deux parents hétérozygotes sous une transmission récessive [14].
La PEL utilise donc la probabilité des phénotypes des membres de la famille autres que le proposant, $Phen^*$, conditionnellement à l'ensemble des génotypes observés. Avec les mêmes notations que dans le chapitre 2, la PEL s'écrit, pour une famille f donnée, de la façon suivante :

$$L_f = \mathbb{P}(Phen^* | Gen_{\text{obs}})$$

En introduisant les génotypes inconnus et en faisant l'hypothèse que les phénotypes des apparentés sont indépendamment distribués conditionnellement à leur génotype, on a alors :

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

Si on fait l'hypothèse que les n familles de l'échantillon sont indépendantes entre elles, la vraisemblance totale s'écrit donc :

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

La contribution des membres de la famille à la vraisemblance se fait exactement comme nous l'avons décrit précédemment. De même que dans la méthode Prospective et dans la GRL, nous avons choisi une approche paramétrique pour l'estimation des paramètres $(\kappa, \lambda, \alpha)$ de la fonction de Weibull.

4.2.2 Etude de la PEL

Plusieurs situations peuvent avoir un effet sur l'estimation par la PEL et peuvent ainsi entraîner un biais. Par exemple, on suppose que les proposants sont identifiés sans aucune ambiguïté. Pourtant, ce n'est pas toujours le cas et des proposants peuvent parfois être omis. Dans ce cas, les familles ne sont pas répliquées et cela peut entraîner un biais dans le processus d'estimation. La plupart du temps, la fréquence de l'allèle délétère ou encore le taux de mutation *de novo* ne sont pas connus dans la population et ces paramètres sont souvent fixés à des valeurs arbitraires. Or, cela peut être une autre source de biais de la PEL lorsqu'une proportion d'individus n'a pas été génotypé.

Dans un premier temps, nous avons étudié le biais relatif et l'efficacité relative de la PEL dans les deux modèles génétiques et les divers modes de sélection que nous avons présentés dans les sections précédentes. Puis, nous avons comparé notre méthode à la vraisemblance Prospective en terme de biais et d'efficacité en utilisant, comme précédemment, la mesure du biais relatif ainsi que la mesure de l'efficacité relative asymptotique (ARE). Pour étudier le biais relatif de la méthode, nous avons considéré que tous les génotypes étaient connus.
Ensuite, nous avons évalué les propriétés de la PEL dans ces différentes situations à l'aide des mêmes simulations que pour l'étude de la vraisemblance Prospective (cf. Section 4.1). Enfin, nous avons étudié l'intérêt du modèle de Weibull étendu pour modéliser la fonction de péné-

trance.

4.2.2.1 Étude de la PEL en termes de biais relatif et d'efficacité, et comparaison avec la vraisemblance Prospective

Le tableau 4.4 présente le biais relatif de la pénétrance estimée avec la PEL et avec la méthode Prospective au point $t = 70$ lorsque tous les individus sont génotypés. Nous avons vérifié que nous obtenons les mêmes résultats en incluant différentes proportions de génotypes inconnus. Dans le modèle MM, nous avions vu que la méthode Prospective était très peu biaisée et on remarque que la PEL a également un biais négligeable. En revanche, pour le modèle MCSM, la PEL reste très peu biaisée tandis que la méthode Prospective devient biaisée. Cette différence au niveau du biais entre les deux méthodes est davantage marquée lorsque la pénétrance est basse. Par exemple, pour une pénétrance basse, une période de 1 an et une probabilité p_s de 0.5, le biais relatif avec la vraisemblance Prospective est de 21% alors qu'il n'est que de 3% avec la PEL.

Le tableau 4.5 présente l'efficacité relative asymptotique (ARE) de la PEL dans le modèle MM. Dans ce cas, contrairement au cas du modèle MCSM, la méthode Prospective comme la PEL sont sans biais. Il est donc intéressant de comparer leur efficacité par l'efficacité relative asymptotique.
En se reportant aux tableaux 4.3 et 4.5, on constate que la PEL apparaît légèrement moins efficace dans le cas de génotypes manquants que la vraisemblance Prospective, notamment dans le cas d'une pénétrance haute et dans la situation où seul le proposant est génotypé. Cependant, la PEL reste tout à fait efficace avec une ARE variant de 74% pour la plus faible (dans le cas extrême où seul le proposant est génotypé) à 97% pour la plus élevée (dans le cas où la moitié des génotypes sont inconnus).

Dans le cas du modèle MCSM, il ne semble pas judicieux de comparer l'efficacité relative des deux méthodes selon la proportion de génotypes manquants, puisque la vraisemblance Prospective est biaisée. Dans ce cas, seul l'efficacité relative de la PEL a été étudiée. Le ta-

Modèle génétique(*)	p_s (*)	Période T (en année)	Biais relatif avec la PEL (en %)		Bias relatif avec la vraisemblance Prospective (en %)	
			Pénétrance haute	**Pénétrance basse**	**Pénétrance haute**	**Pénétrance basse**
MM	1	20	2	2	0	-3
		1	2	3	3	2
	0.5	20	2	3	1	0
		1	2	3	1	-2
MCSM	1	20	2	3	5	14
		1	2	4	9	21
	0.5	20	2	3	5	14
		1	3	3	10	21

(*) MM : maladie mendélienne; MCSM : maladie complexe à sous-entités monogéniques ;

p_s : probabilité d'être sélectionné pour les individus atteint durant la période T *(et avant 36 ans dans le cas de MCSM).*

TAB. 4.4 – Etude du biais relatif pour la pénétrance estimée à 70 ans pour la PEL et pour la vraisemblance Prospective

bleau 4.6 présente donc l'efficacité relative de la PEL dans le modèle des maladies complexes à sous-entités monogéniques (MCSM). Nous avons étudié l'efficacité relative dans le cas le plus défavorable où seul le génotype du proposant est connu, la situation de référence étant toujours celle pour laquelle tous les individus sont génotypés. La PEL perd de l'efficacité et l'efficacité relative calculée dans ce cas est inférieure à l'efficacité relative calculée dans le modèle MM avec une ARE variant de 56% pour la plus faible à 69% pour la plus forte.

4.2.2.2 Etude de la PEL à une erreur sur l'identification des proposants

Le tableau 4.7 montre le biais relatif pour l'estimation de la pénétrance avec la PEL dans le cas où un unique proposant à été identifié dans chaque famille. Ce proposant a été choisi aléatoirement, dans le processus de simulation, parmi les proposants "potentiels" dans les familles. Ces familles n'ont donc pas été répliquées alors que, dans le cas de plusieurs proposants, elles auraient dû l'être. Dans les modèles MCSM, la quasi-totalité des familles ne contient qu'un unique proposant "potentiel" et les familles n'ont donc pas besoin d'être répliquées. Dans ce

Situation	p_s (*)	Période T (en année)	ARE	
			Pénétrance basse	**Pénétrance haute**
Proportion de génotypes inconnus	1	20		
❖ 50%			0,97	0,92
❖ 75%			0,92	0,88
❖ Seul le proposant est génotypé			0,82	0,74
	1	1		
❖ 50%			0,96	0,96
❖ 75%			0,92	0,92
❖ Seul le proposant est génotypé			0,84	0,83
	0.5	20		
❖ 50%			0,96	0,92
❖ 75%			0,92	0,87
❖ Seul le proposant est génotypé			0,83	0,81
	0.5	1		
❖ 50%			0,98	0,95
❖ 75%			0,97	0,91
❖ Seul le proposant est génotypé			0,87	0,84

(*) p_s : Probabilité d'être sélectionné pour les individus atteint durant la période T (et avant 36 ans dans le cas de MCSM).

TAB. 4.5 – Efficacité de la PEL sous le modèle génétique MM selon la proportion de génotypes manquants par rapport à la situation de référence où tous les génotypes sont connus.

p_s (*)	Période T (en année)	Efficacité Relative Asymptotique (ARE)	
		Pénétrance basse	**Pénétrance haute**
1	20	0,64	0,68
1	1	0.60	0,67
0.5	20	0,66	0,69
0.5	1	0,56	0,66

(*) p_s : Probabilité d'être sélectionné pour les individus atteint durant la période T *(et avant 36 ans dans le cas de MCSM).*

TAB. 4.6 – Efficacité de la PEL dans le cas du modèle génétique MCSM dans le cas extrême où seul le proposant est génotypé par rapport à la situation de référence où tous les individus sont génotypés

cas, il paraît logique que les résultats soient proches des résultats du tableau 4.4 pour la PEL. Dans le modèle MM, le fait d'omettre la réplication peut être plus sensible car les familles avec plusieurs proposants sont plus courantes, notamment lorsque $p_s = 1$ et $T = 20$. Dans ce cas précisément où $p_s = 1$ et $T = 20$, on constate qu'un biais relatif est engendré avec la PEL lorsque les proposants sont mal définis et que, par conséquent, un seul proposant par famille a été identifié. Ce constat est davantage marqué dans le cas d'une pénétrance basse où le biais est de -13% alors qu'il n'est que de 2% lorsque tous les proposants ont été identifiés et que les familles sont répliquées autant de fois qu'il y a de proposants. Par contre, dans les autres cas de sélection, et notamment quand $T = 1$, le fait d'omettre des proposants n'influence quasiment pas l'estimation de la pénétrance et on obtient des résultats proches de ceux obtenus avec réplication. En fait, plus la probabilité d'être sélectionné est faible (c'est-à-dire lorsque les critères de sélection sont stringents) et moins le fait de répliquer les familles a une incidence sur les estimations de la pénétrance avec la PEL.

Modèle génétique	p_s (*)	Période T (en année)	Biais sans réplication (en %)	
			Pénétrance haute	**Pénétrance basse**
MM	1	20	- 4	- 13
		1	1	1
	0.5	20	- 1	-4
		1	2	2
MCSM	1	20	0	1
		1	2	3
	0.5	20	1	2
		1	2	3

(*) p_s *: Probabilité d'être sélectionné pour les individus atteints durant la période* T.

TAB. 4.7 – Etude du biais relatif dans l'estimation de la pénétrance à 70 ans avec la PEL sans réplication dans les familles comportant plusieurs proposants

4.2.2.3 Impact d'une mauvaise spécification des paramètres du modèle génétique

Dans cette section, nous avons étudié la sensibilité des estimations de la pénétrance avec la PEL à une mauvaise spécification de la fréquence de l'allèle délétère, fq et du taux de mutation *de novo*, pn sous un modèle MM, pour $p_s = 1$ et $T = 20$ et pour une vraie pénétrance basse. Afin d'étudier la sensibilité de la PEL à ces paramètres, nous avons introduit un taux de génotypes manquants de 80% (le génotype du proposant étant toujours connu). Le tableau 4.8 montre les résultats de cette étude. On constate que la méthode n'est pas sensible à une erreur sur le taux de mutation *de novo*, pn. Si pn est surestimé et fixé à 10^{-4} dans l'analyse alors que sa vraie valeur est de 10^{-5}, le biais engendré est de 2%. A l'inverse, si le paramètre pn est "sous-estimé" et fixé à 0 dans l'analyse alors que sa vraie valeur est de 10^{-5}, le biais engendré est de seulement 1%. Quant à fq, lorsque ce paramètre est "sous-estimé" et fixé à 10^{-6} alors que sa vraie valeur est 0.01, cela entraîne un biais négligeable de -1%. Par contre, le biais est plus important (-8%) quand le paramètre est surestimé (fixé à 0.1 quand sa vraie valeur est 0.01). Mais cette situation n'est pas très réaliste puisque nous étudions des mutations rares.

Paramètres fixes dans les simulations	Paramètre étudié	Valeur du paramètre dans les simulations	Valeur du paramètre fixé dans l'analyse	Biais relatif de l'estimation de la pénétrance à 70 ans (en %)
$pn = 0$	fq	$fq = 0.01$	$fq = 10^{-6}$	-1
			$fq = 0.1$	-8
$fq = 0.01$	pn	$pn = 10^{-5}$	$pn = 10^{-4}$	2
			$pn = 0$	1

TAB. 4.8 – Sensibilité de la PEL à une mauvaise spécification des paramètres fq et pn

4.2.2.4 Intérêt du modèle de Weibull étendu

La figure 4.2 montre l'estimation de la pénétrance dans le cas de familles simulées avec un paramètre $\kappa = 0.10$ dans le modèle de Weibull étendu, et analysées d'une part en estimant le paramètre κ conjointement aux autres paramètres (α et λ) du modèle de Weibull et d'autre part, en ignorant ce paramètre et en le fixant à 0. L'existence d'un $\kappa = 0.10$ dans les données signifie que, pour la maladie et la mutation étudiées, 10% des individus porteurs de la mutation ne seront jamais atteints de la maladie. On remarque que le fait de ne pas inclure de paramètre κ dans le modèle de Weibull, alors que les données ont été simulées avec un κ non nul, change l'allure de la courbe et provoque un biais non négligeable par rapport à la vraie courbe de pénétrance.

Par contre, lorsque le paramètre κ est estimé, la figure 4.2 montre bien que la courbe estimée et la vraie courbe sont très proches. Le fait d'étendre le modèle de Weibull en y ajoutant ce paramètre κ, permet donc un meilleur ajustement de la fonction de pénétrance aux données étudiées.

La figure 4.3 montre l'impact, sur l'estimation de la fonction de pénétrance par la PEL, d'une omission du paramètre δ qui représente l'âge minimum en-deçà duquel le risque chez les porteurs d'être atteints de la maladie étudiée est nul. La vraie fonction de pénétrance des données simulées (courbe pleine) suit un modèle de Weibull dans lequel δ a été fixé à 25 ans (nous avons posé $\kappa = 0$ dans les simulations). Avant 25 ans, la vraie fonction de pénétrance est donc égale à zéro.

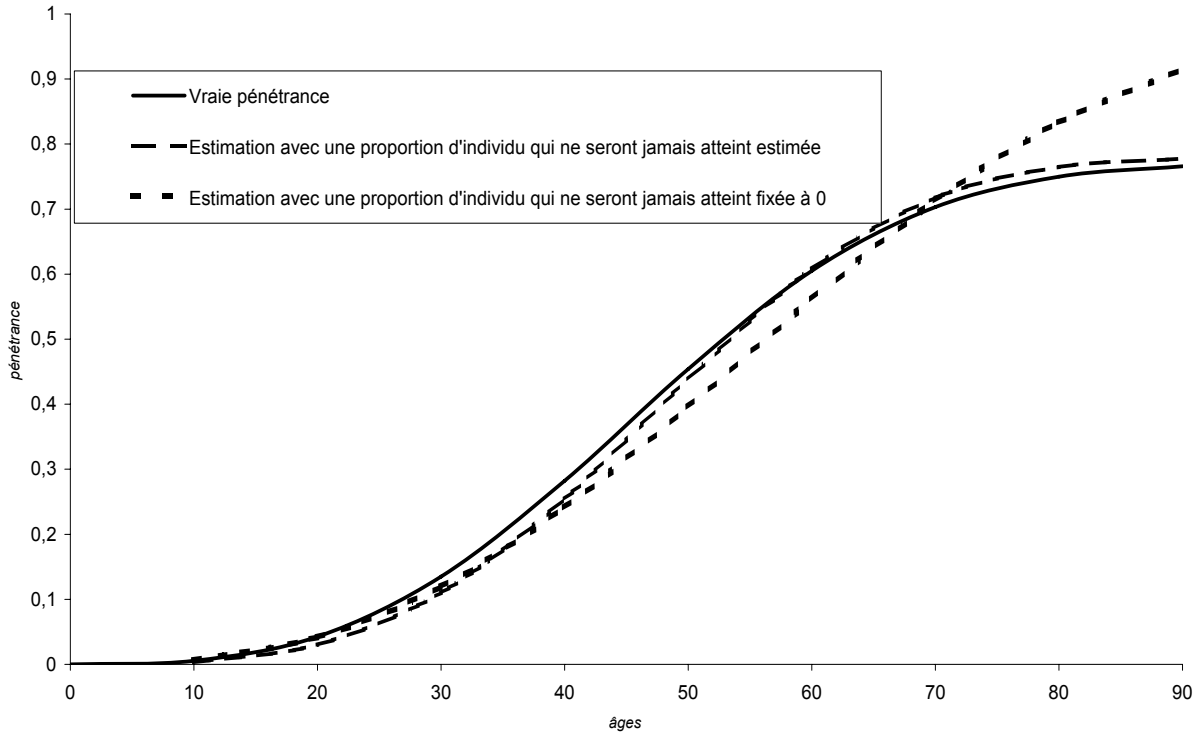


FIG. 4.2 – Intérêt du paramètre κ dans le modèle de Weibull étendu

La courbe en pointillés représente l'estimation de la fonction de pénétrance avec la PEL dans le cas où le paramètre δ a été fixé à sa vraie valeur de 25 ans dans l'analyse. Dans ce cas, la PEL fournit une bonne estimation de la fonction de pénétrance puisque la courbe en pointillées et la courbe pleine sont très proches.
La courbe en tirets représente l'estimation de la fonction de pénétrance avec la PEL dans le cas où le paramètre δ n'a pas été introduit dans le modèle de Weibull lors de l'analyse. Dans ce cas, l'estimation est assez peu biaisée. Le prise en compte de δ affûte donc l'estimation. Cependant, si aucune valeur de ce paramètre n'est fournie a priori dans la littérature, le fait de ne pas l'inclure dans l'analyse n'entraînera pas de biais conséquent.

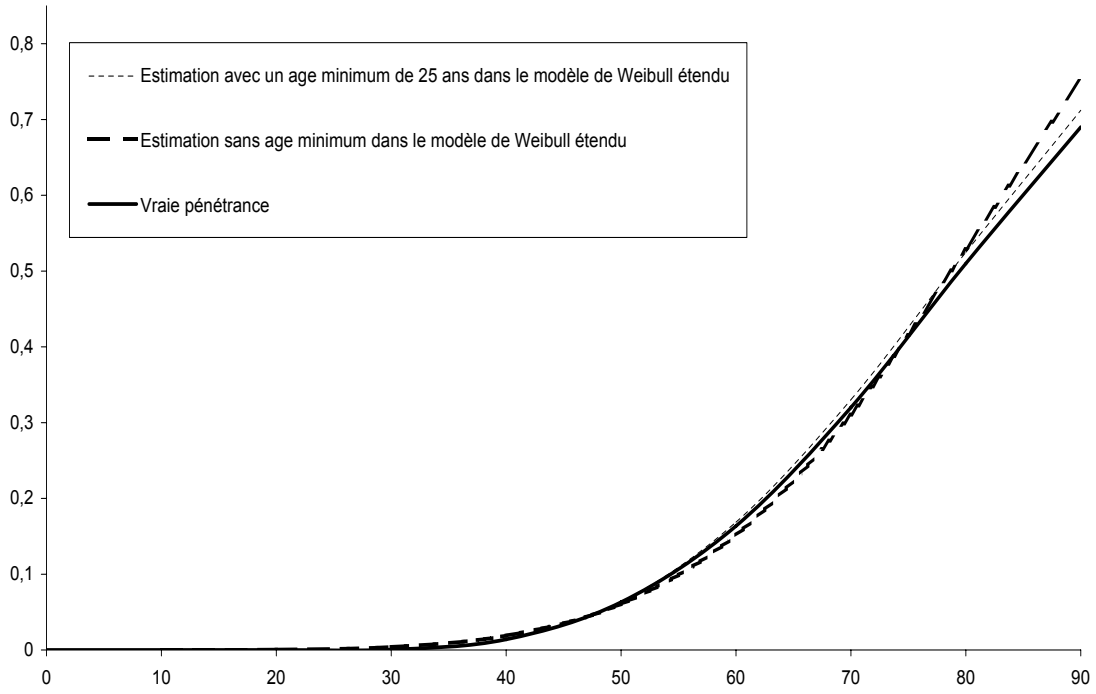


FIG. 4.3 – Intérêt du paramètre δ dans le modèle de Weibull étendu

4.3 Application à la NAH et au cancer du sein

Afin d'illustrer les deux modèles considérés dans ce chapitre, le modèle MM et le modèle MCSM, nous avons appliqué la vraisemblance Prospective ainsi que la PEL à deux maladies représentatives de ces deux modèles génétiques. Le modèle MM a été illustré par un jeu de données familiales sélectionnées sur l'existence d'individus atteints d'une maladie génétique : la neuropathie amyloïde héréditaire, que nous noterons NAH. Le modèle MCSM a été illustré par un échantillon de familles atteintes de cancers du sein et de l'ovaire avec les mutations BRCA1 et/ou BRCA2. Ces dernières familles ont été sélectionnées sur l'existence d'au moins un cas de cancer du sein diagnostiqué avant 36 ans.
Après avoir présenté le calcul des intervalles de confiance dans les deux applications, nous décrirons les deux échantillons de familles et le mode de sélection associée. Nous présenterons ensuite les estimations obtenues avec nos deux vraisemblances : la Prospective et la PEL.

4.3.1 Calcul des intervalles de confiance

Dans les deux applications qui suivent, nous présentons les intervalles de confiance à 95% aux points $t = 50$, $t = 60$ et $t = 70$ avec la PEL. Ces intervalles de confiance (IC) ont été estimés en utilisant une méthode basée sur le bootstrap [18].
Les échantillons de familles étudiés ont été reéchantillonés 10 000 fois et, pour chaque nouvel échantillon, nous avons estimé le risque à différents âges T'. Les intervalles de confiance aux âges T' ont été obtenus en utilisant la méthode des percentiles simple [19]. Les bornes inférieures et supérieures du risque estimé à l'âge T' sont donc données par les percentiles $\frac{\alpha}{2}$ et $\frac{1-\alpha}{2}$ de la distribution des 10 000 valeurs de risques estimés à l'âge T' avec $\alpha = 5\%$.

4.3.2 Les données de neuropathie amyloïde héréditaire (NAH)

Les neuropathies amyloïdes héréditaires sont des maladies génétiques à transmission autosomique dominante. Les NAH sont les plus graves des neuropathies héréditaires de l'adulte et

sont mortelles en 10 ans sans traitement après les premières manifestations. Elles sont caractérisées par des dépôts endoneuraux de substances amyloïdes constituées de variants de transthyrétine (TTR) et elles résultent d'une mutation ponctuelle du gène de la TTR. Ces caractéristiques concernent principalement le système nerveux périphérique et le coeur. Elles ont d'abord été décrites au Portugal et, bien que la maladie soit présente dans le monde entier, on la retrouve particulièrement dans des zones limitées comme le Portugal, le Japon ou encore la Suède, avec différentes variations génotypiques, phénotypiques et différents âges de début de la maladie.
En France, les patients d'origine portugaise représentent une part non négligeable des cas de NAH, et présentent certaines caractéristiques par rapport aux Français. De nombreux variants de la transthyrétine pathogène ont été découverts dans la population française, mais un seul variant a été détecté dans la population portugaise, le Val30Met.
Depuis 1993, la transplantation hépatique (TH) est proposée comme traitement pour supprimer la principale source de synthèse de la protéine amyloïdogène. La TH a permis une réduction de 98% du taux de TTR mutée circulante. La TH, pour être efficace, doit être proposée tôt, dès l'apparition des premières manifestations de la neuropathie et justifie un dépistage familial présymptomatique.
Dans la pratique, toutes les familles sont adressées au département de Neurologie de l'hôpital de Bicêtre.

Pour notre application, nous avons repris l'échantillon utilisé dans l'article de Planté-Bordeneuve et al. [38] provenant du département de neurologie de l'hôpital de Bicêtre qui est centre de référence pour la NAH. Une étude familiale systématique a été menée et une recherche de mutation était proposée aux apparentés du proposant. Par souci d'homogénéité, nous avons restreint cet échantillon aux porteurs de la mutation Val30Met. Finalement, nous avons donc un échantillon de 20 familles françaises et 33 familles portugaises.
Le génotype TTR était disponible pour 108 apparentés des familles françaises dont 47 porteurs de la mutation Val30Met, et pour 139 apparentés des familles portugaises dont 50 porteurs de la mutation Val30Met. Les proportions de génotypes inconnus parmi les apparentés étaient de 72% pour les Français et de 68% pour les Portugais.
Dans l'analyse, nous avons fixé la fréquence de l'allèle délétère à 0.001 (i.e. $fq = 0.001$) et le

taux de mutation *de novo* à 0 (i.e. $pn = 0$).

4.3.3 Résultats de l'application à la NAH

La figure 4.4 montre la fonction de pénétrance estimée d'une part avec la PEL (courbe avec des points et courbe avec des étoiles) et d'autre part avec la méthode Prospective (courbe pleine et courbe en tirets) pour les familles françaises et portugaises. Les IC sont donnés aux âges 50, 60 et 70 ans.

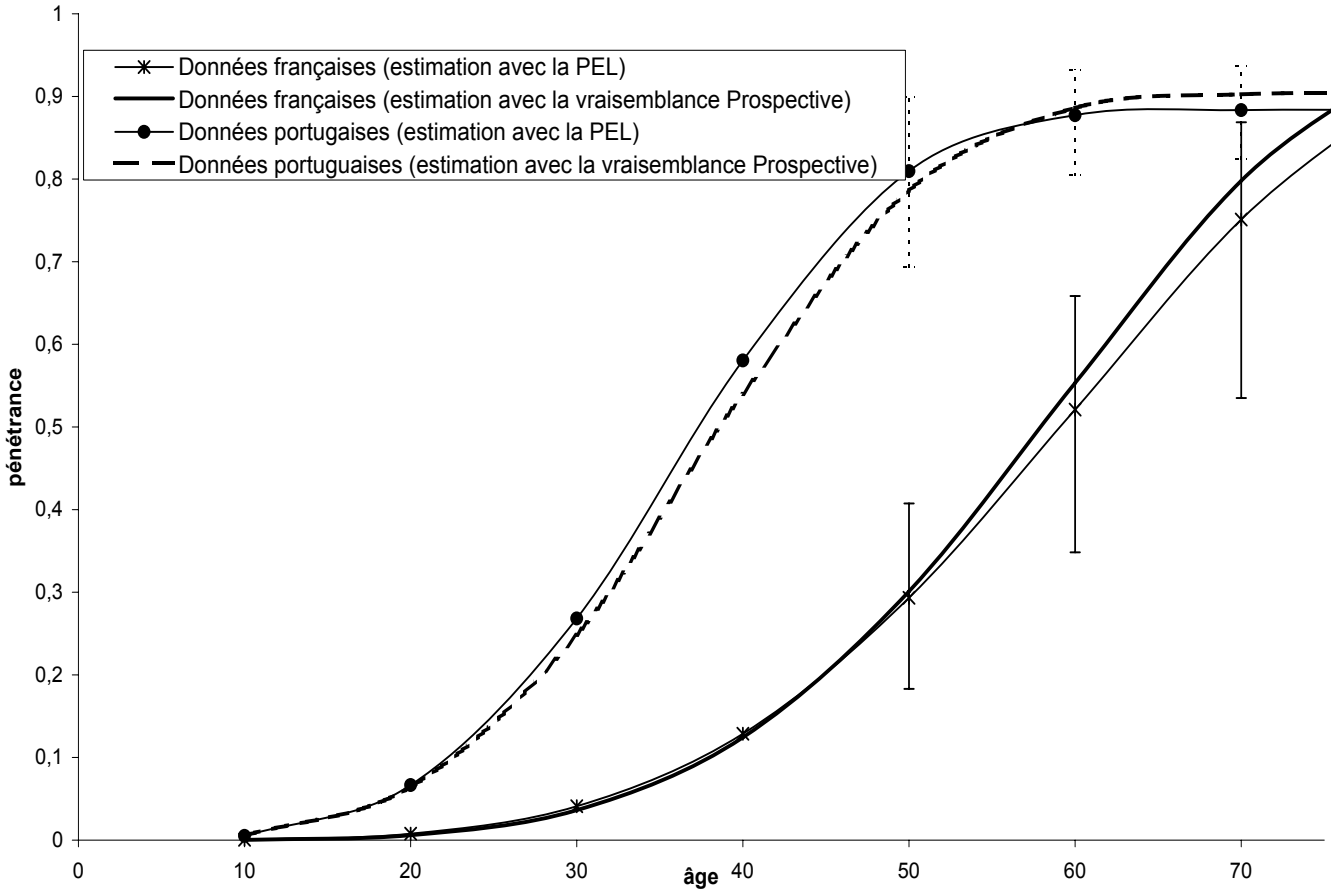


FIG. 4.4 – Application à des données de neuropathie amyloïde héréditaires

Dans l'échantillon portugais, le plateau montre l'existence d'une proportion d'individus porteurs qui ne seront jamais atteints. Le paramètre κ dans le modèle de Weibull a été estimé à 0.09 ($\hat{\kappa} = 0.09$) et a été trouvé significativement différent de zéro avec une $p_{\text{value}} < 0.001$. Ce plateau

"significatif" illustre bien l'importance de l'extension du modèle de Weibull par ce paramètre κ.

Par ailleurs, aussi bien dans l'échantillon portugais que dans l'échantillon français, les estimations de courbes obtenues avec les deux méthodes sont très proches et le choix de la méthode n'apparaît pas crucial. Ce résultat vient appuyer les résultats obtenus par simulations dans le cas de maladies mendéliennes pour lesquelles nous avions trouvé une équivalence en termes de biais et d'efficacité relative entre la PEL et la méthode Prospective.

4.3.4 Les données de cancer du sein

L'implication de facteurs génétiques dans le cancer du sein est aujourd'hui confirmée à travers un modèle monogénique expliquant une petite proportion de cas à forte pénétrance, au sein d'une majorité de cas sporadiques, et deux gènes de prédisposition ont été mis en évidence : BRCA1, sur le chromosome 17 et BRCA2, sur le chromosome 13, qui confèrent un risque très élevé de tumeur, survenant souvent à des âges précoces.

Les familles de cancer du sein que nous avons étudiées ici ont été sélectionnées parmi 317 femmes atteintes de cancer du sein, diagnostiquées avant 36 ans entre janvier 1990 et janvier 1998, et qui étaient suivies à l'Institut Curie. Un conseil génétique a été proposé à toutes les femmes et 153 d'entres elles sont venues en consultation à l'Institut Curie.

Un dépistage génétique des mutations BRCA1, BRCA2 et TP53 a été systématiquement proposé quelle que soit l'histoire familiale. Parmi ces femmes atteintes, 145 ont subi un test génétique [11].

L'ensemble de la séquence des deux gènes a été analysé par une combinaison de $DGGE$, $DHPLC$ et PTT [42, 46]. 16 patients avec une mutation BRCA1 et 14 patients avec une mutation BRCA2 ont été identifiés. Dans ces 30 familles, un test génétique a été proposé à tous les apparentés. Parmi les 30 familles, le génotype était disponible pour 33 apparentés parmi lesquels 17 étaient des individus porteurs. Dans 16 familles, le proposant était donc le seul individu de la famille ayant été génotypé et la proportion globale de génotypes inconnus parmi les apparentés était de 91%.

Pour l'analyse, nous avons fixé la fréquence de l'allèle délétère à 0.001 (i.e. $fq = 0.001$) et le taux de mutation *de novo* à zéro (i.e. $pn = 0$). Le risque cumulé chez les non-porteurs a été fixé égal au risque rencontré dans la population générale, c'est-à-dire à 0.10 à 70 ans [3].

4.3.5 Résultats de l'application au cancer du sein

La figure 4.5 montre l'estimation de la fonction de pénétrance pour les données de cancer du sein avec la PEL d'une part (courbe en tirets) et avec la vraisemblance Prospective d'autre part (courbe en pointillés). Comme l'échantillon était relativement petit, nous avons mis en commun les familles avec la mutation BRCA1 (16 familles) et les familles avec la mutation BRCA2 (14 familles). Nous avons donc analysé un échantillon de 30 familles avec les mutations BRCA1/2. Les intervalles de confiance ont été calculés par bootstraps, comme dans les données NAH, aux points $t = 50$, $t = 60$ et $t = 70$.

Les deux méthodes fournissent des courbes de pénétrance différentes. Cette différence était attendue au vu des résultats théoriques obtenus par simulations (cf. Section 4.2). En effet, l'application aux données de cancer du sein correspond au modèle MCSM que nous avons étudié. Or, dans ce modèle, la PEL donnait des estimations non biaisées, contrairement à la vraisemblance Prospective. Dans cette application, la PEL fournit donc probablement une courbe de pénétrance non biaisée tandis que la méthode Prospective est certainement biaisée. Toutefois, la différence entre les deux estimations n'est probablement pas significative car les intervalles de confiance contiennent les deux estimations. Ceci est probablement dû à la faible taille de l'échantillon.

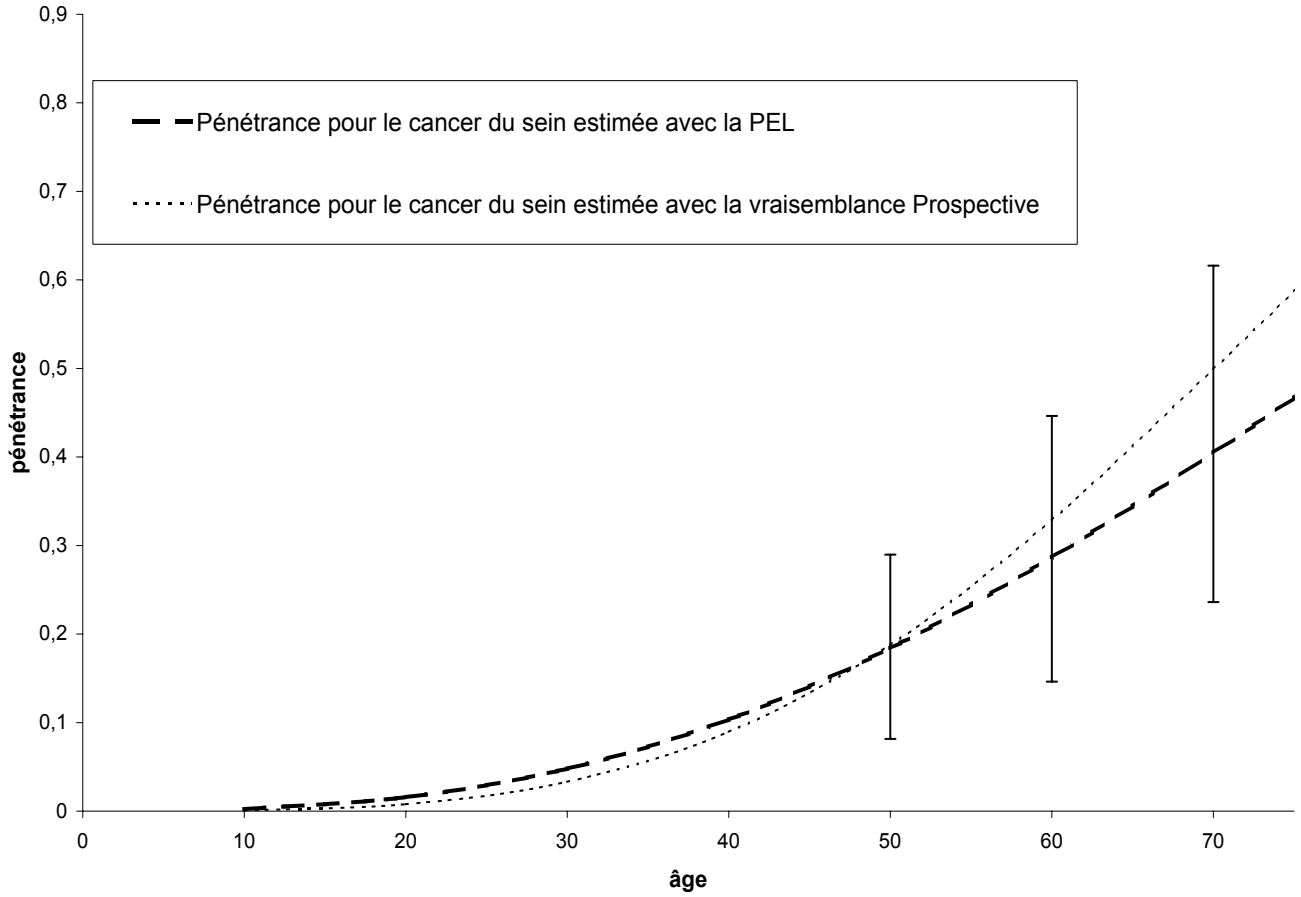


FIG. 4.5 – Application à des données de cancer du sein

4.4 Étude de l'hétérogénéité de la pénétrance de la NAH dans la population suédoise

La PEL nous a permis d'étudier des données de NAH pour une population suédoise. La neuropathie amyloïde héréditaire (NAH), déjà décrite dans la section précédente, ne semble pas également répartie selon les zones géographiques. En Europe, elle se retrouve plus souvent au Portugal et en Suède qu'en France. En Suède, des foyers ont été décrits dans les villes de Pitea et de Skelleftea. Comme pour les familles portugaises, la mutation Val30Met est, dans les familles suédoises, de loin le variant le plus courant du gène TTR [40]. Nous nous sommes intéressés plus particulièrement à cette mutation Val30Met et avons donc analysé un échantillon suédois pour la NAH.

Selon l'origine géographique, des âges de début de la maladie différents ont été décrits. En effet, alors que la moyenne d'âge du premier symptôme est de 33 ans au Portugal, elle est de 56 ans en Suède [41, 27].

Une analyse préalable des données suédoises avait suggéré que les âges de début étaient plus tardifs que dans la population portugaise et même française. De plus, les auteurs avaient démontré [15] que l'âge de début de la maladie était significativement plus élevé lorsque la mutation était transmise par le père que lorsqu'elle était transmise par la mère.

Après avoir décrit les données et leur mode de recueil, nous estimerons la fonction de pénétrance avec la PEL. Nous étudierons ensuite si les risques varient selon les différentes régions. Enfin, nous étudierons l'hypothèse selon laquelle la pénétrance dépendrait du sexe du parent transmetteur de la mutation.

4.4.1 Description des données

Parmi les 401 patients porteurs de la mutation Val30Met identifiés au département de génétique clinique de l'université de Umea en Suède depuis 1986, une recherche généalogique complète était disponible pour 122 patients. Comme certains de ces 122 individus étaient apparentés entre eux, ils ont été regroupés et l'échantillon final était donc constitué de 85 familles.

Le diagnostic de neuropathie amyloïde TTR était basé sur la recherche de dépôts amyloïde dans le prélèvement par biopsie ainsi que sur l'identification d'un variant TTR (la mutation Val30Met chez tous les patients).
Parmi les familles, 8 familles contenaient 9 patients homozygotes pour la mutation Val30Met. Ces patients homozygotes ont été retirés de l'analyse car ils étaient trop peu nombreux pour permettre une estimation fiable du risque associé à ce génotype.
Finalement, nous avons mené notre étude sur un échantillon de 77 familles.

4.4.2 Résultats

Nous avons utilisé la proportion d'homozygotes pour la mutation Val30Met parmi les atteints pour estimer la fréquence de la mutation à 0.04 (voir Annexe B). Dans nos analyses, nous avons donc fixé la fréquence de l'allèle délétère, fq, à cette valeur.
La prise en compte de l'effet du sexe du parent transmetteur n'étant pas prévu dans le programme, nous avons procédé à une manipulation de fichier.
L'hypothèse nulle que nous voulons tester est donc : H_0 : {La pénétrance est la même lorsque la mutation vient du père ou lorsque la mutation vient de la mère.} Pour ce faire, nous avons calculé les vraisemblance L_P, L_M, L_Cte, et L_T à partir de la méthode PEL, telles que :

- L_P est la vraisemblance du modèle pour l'échantillon dans lequel les phénotypes ont été notés comme inconnus lorsque la mutation venait de la mère, ou lorsque l'origine de la mutation était inconnue.
- L_M est la vraisemblance du modèle pour l'échantillon dans lequel les phénotypes ont été notés comme inconnus lorsque la mutation venait du père, ou lorsque l'origine de la mutation était inconnue.
- L_{Cte} est la vraisemblance du modèle pour l'échantillon dans lequel tous les phénotypes ont été notés comme inconnus (pour avoir la constante).
- L_T est la vraisemblance du modèle pour l'échantillon dans lequel les phénotypes ont été notés comme inconnus lorsque l'origine de la mutation était inconnue (échantillon de référence).

Puis nous avons fait le rapport de vraisemblance : $\frac{L_T}{L_{ST}}$,où L_{ST} est la vraisemblance du modèle pour l'échantillon dans lequel le sexe du parent transmetteur a été pris en compte. Afin que les individus dont l'origine de la mutation est inconnue ne soient pas pris en compte deux fois, nous avons calculé la constante L_{Cte}. Nous avons donc :

$$L_{ST} = L_P * L_M * L_{Cte}$$

Nous réalisons donc un test du χ^2 à 1 degré de liberté à partir de la statistique :

$$-2\log[L_T - (L_P + L_M - L_{Cte})]$$

.

La figure 4.6 montre la courbe de pénétrance estimée avec la PEL pour tout l'échantillon mais aussi pour les hommes et les femmes séparément. Nous n'avons pas trouvé d'effet significatif du sexe ($\chi^2 = 2,44$, $2ddl$, $p_{value} \cong 0.3$). Dans chacune des estimations que nous avons menées dans les différents "sous-échantillons", le paramètre κ était toujours estimé à 0, ce qui explique l'absence de plateau dans l'ensemble des courbes que nous présentons. Les valeurs du risque ainsi que les intervalles de confiance (obtenus par bootstraps) de 30 à 90 ans sont donnés dans le tableau 4.9.

L'échantillon global contenait des familles provenant de trois régions de Suède : Lycksele, Pitea et Skelleftea. Pour la région de Lycksele, seulement 10 familles étaient disponibles, ce qui est trop faible. Nous avons donc comparé uniquement les deux autres régions entre elles. La figure 4.7 montre les courbes de pénétrance estimées dans les deux régions. On constate que la pénétrance est plus basse dans la région de Pitea que dans celle de Skelleftea avec des pénétrances respectives de 32% et 65% à 80 ans avec un test d'homogénéité significatif ($\chi^2 = 26.60$, $2ddl$, $p < 0.001$).

Parmi les 77 familles, nous connaissions le sexe du parent transmetteur de la mutation pour 762 individus : pour 435 individus, le parent transmetteur était la mère et pour 327, le parent transmetteur était le père. La figure 4.8 montre la courbe de pénétrance estimée selon le sexe du

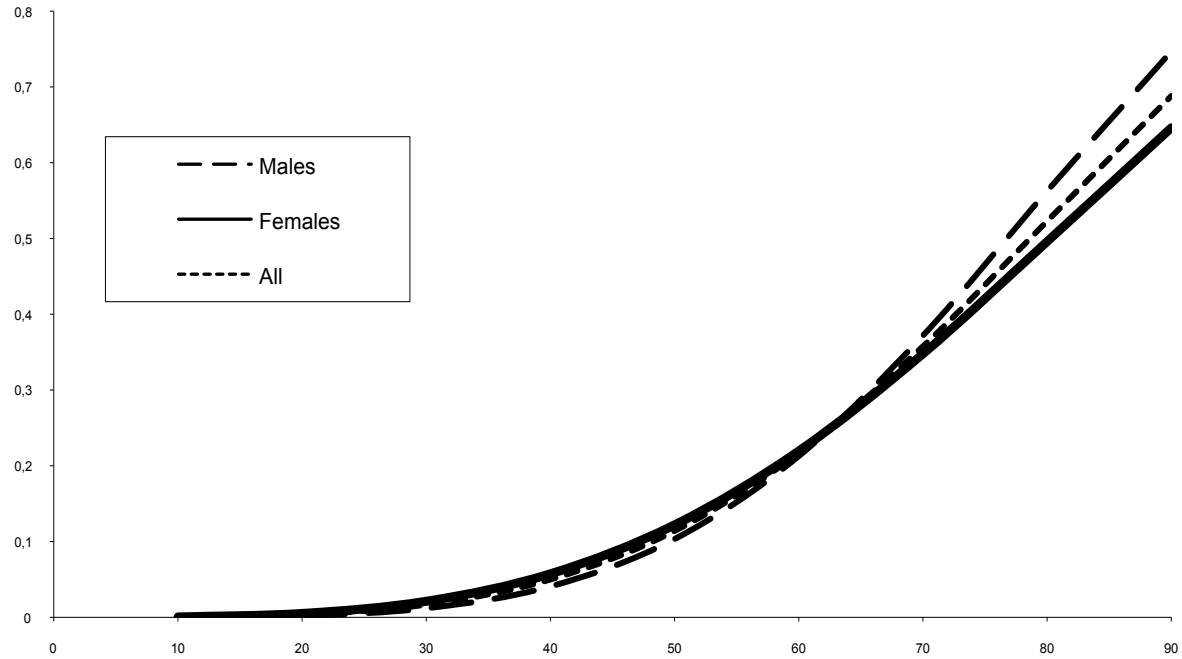


FIG. 4.6 – Estimation de la pénétrance selon le sexe

Age	Estimation de la pénétrance par la PEL	Intervalles de confiance à 95%
30	0.017	0.008 □0.032
40	0.05	0.029 □0.082
50	0.11	0.08 □0.16
60	0.22	0.16 □0.29
70	0.36	0.28 □0.45
80	0.52	0.42 □0.63
90	0.69	0.55 □0.79

TAB. 4.9 – Estimation de la pénétrance par la PEL dans les familles suèdoises

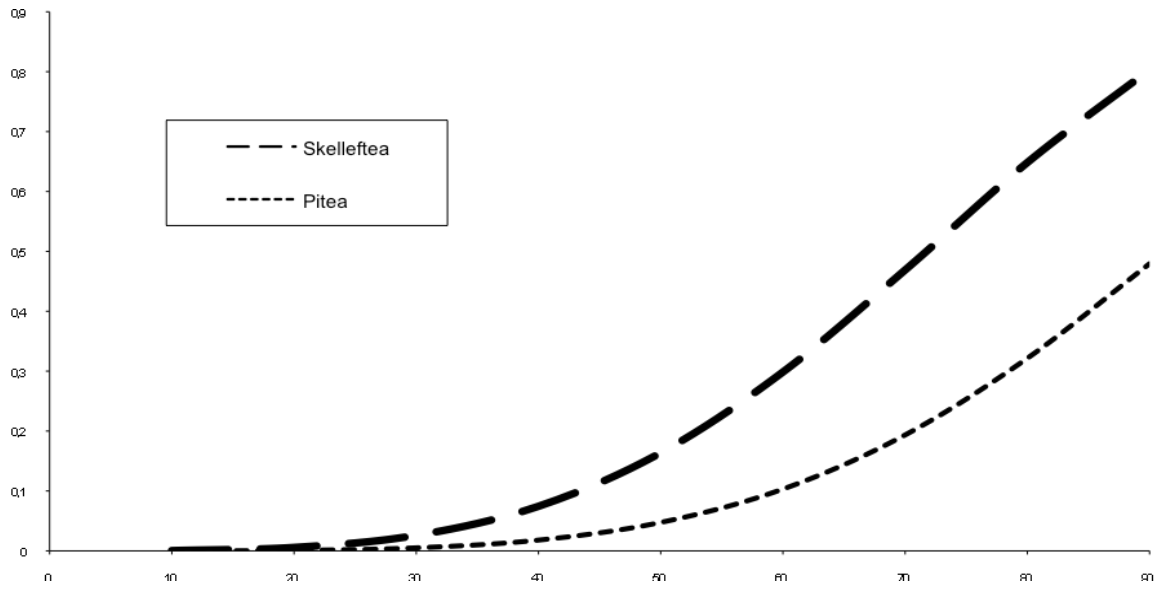


FIG. 4.7 – Estimation de la courbe de pénétrance selon la région Skelleftea et Pitea

parent transmetteur. On remarque une différence nette entre les deux courbes et la pénétrance est significativement plus élevée lorsque la mutation est héritée de la mère plutôt que du père ($\chi^2 = 7.84$, $2ddl$, $p < 0.02$).

Le tableau 4.10 montre les valeurs de l'estimation de la pénétrance en fonction de l'âge selon que la mutation ait été transmise par la mère ou par le père ainsi que les IC à 95% obtenus par bootstraps. Ainsi, à 60 ans, le risque est de 16% [13% - 25%] lorsque la mutation est transmise par le père alors que le risque est, au même âge, deux fois plus grand lorsque la mutation est transmise par la mère 30% [21% - 44%].

Cette différence de risque en fonction du sexe du parent transmetteur a également été analysée par Bonaïti B. et al. [6] dans les données NAH provenant de familles françaises et portugaises, que nous avons décrites dans la section précédente. Dans cette analyse, les pénétrances ont été estimées à l'aide d'une version modifiée de la PEL permettant de prendre en compte le sexe du parent transmetteur de la mutation, en ordonnant le génotype des hétérozygotes selon que la mutation provienne du père ou de la mère. Dans l'échantillon portugais, une différence signifi-

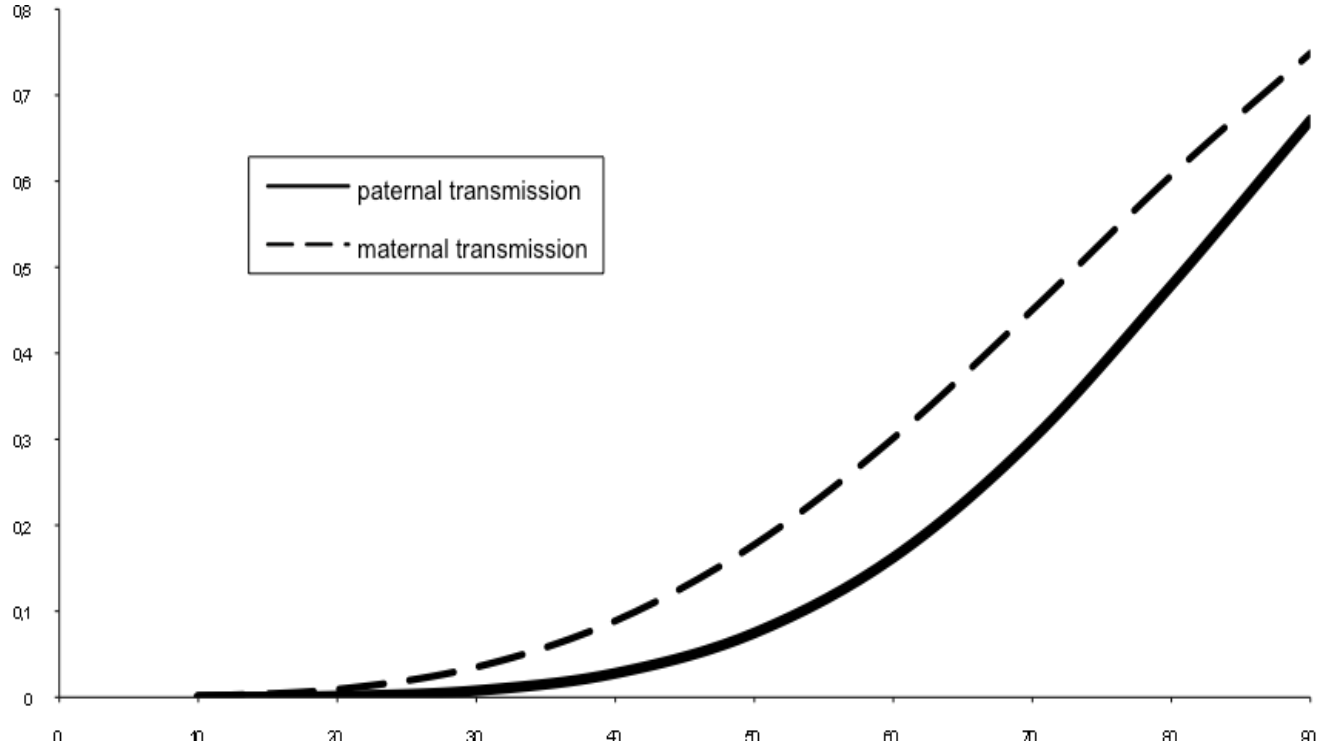


FIG. 4.8 – Estimation de la courbe de pénétrance selon le sexe du parent transmetteur de la mutation

Age	Estimation de la pénétrance (IC à 95%) quand la mutation est transmise par :	
	Le père	La mère
30	0.008 (0.003 – 0.016)	0.035 (0.013 – 0.059)
40	0.028 (0.016 – 0.041)	0.089 (0.043 – 0.14)
50	0.075 (0.052 – 0.11)	0.18 (0.10 – 0.27)
60	0.16 (0.13 – 0.25)	0.30 (0.21 – 0.44)
70	0.30 (0.21 – 0.44)	0.45 (0.34 – 0.61)
80	0.48 (0.30 – 0.66)	0.61 (0.47 – 0.76)
90	0.67 (0.40 – 0.85)	0.75 (0.57 – 0.89)

TAB. 4.10 – Estimation de la pénétrance selon le sexe du parent transmetteur de la mutation

cative du risque a été trouvée selon le sexe du parent transmetteur. Par contre, dans l'échantillon français, la différence n'était pas significative.

4.5 Conclusion et Discussion

Dans ce chapitre, nous nous sommes intéressés au cas de familles sélectionnées sur l'existence d'au moins un individu atteint dans la famille, en contraste à la sélection sur critères familiaux vue au chapitre 3.

Nous avons d'abord présenté la vraisemblance Prospective qui nécessite la modélisation mathématique des critères de sélection et qui fournit des estimations de la fonction de pénétrance. Nous avons montré que la vraisemblance Prospective était une méthode efficace dans le cas de génotypes manquants, sans biais dans le cadre d'un modèle MM, mais qu'elle pouvait conduire à des estimations fortement biaisées dans le cadre d'un modèle MCSM.

Le but de notre travail était donc de proposer une méthode simple qui estime la fonction de pénétrance à partir de familles sélectionnées sur l'existence d'un atteint, qui tienne compte du fait qu'un nombre variable d'individus pouvait ne pas être génotypés dans les familles et qui ne soit pas biaisée pour la sélection.

A l'aide de simulations sous différents modèles génétiques et avec des schémas de sélection différents, nous avons montré que la PEL fournissait des résultats très satisfaisants. Cette méthode a l'avantage d'être simple, de faire peu d'hypothèses sur le processus de recensement et de fournir des estimations non biaisées de la fonction de pénétrance, à condition que les familles avec plusieurs proposants soient comptées autant de fois qu'il y a de proposants. Lorsque les proposants ne sont pas bien identifiés et que les familles ne sont pas répliquées, nous avons montré que la fonction de pénétrance pouvait être sous-estimée avec néanmoins un faible biais relatif.

Nous avons montré également une faible sensibilité de la méthode à une mauvaise spécification des paramètres du modèle génétique et avons constaté que l'extension du modèle de Weibull permettait, en général, un meilleur ajustement aux données. Nous avons constaté que la PEL perdait de l'efficacité lorsqu'un grand nombre de génotypes était inconnu mais beaucoup moins que la GRL dans le même cas.

Dans le cas d'une sélection sur critères indépendants de l'histoire familiale, nous recommanderons donc d'utiliser la PEL préférentiellement à la vraisemblance Prospective, qui est

biaisée dans certains cas, et préférentiellement à la GRL, qui manque d'efficacité dans le cas de génotypes inconnus et dont l'utilisation doit être restreinte au cas de sélection sur critères familiaux pour lesquels l'utilisation d'une méthode simple telle que la PEL n'est pas adaptée.

Nous avons appliqué notre méthode ainsi que la vraisemblance Prospective à des données de NAH et de cancer du sein afin d'illustrer nos résultats théoriques. De plus, nous avons mené une étude sur la NAH avec des données suédoises.

Dans l'application à la NAH chez les Portugais et les Français, les fonctions de pénétrance estimées avec les deux méthodes étaient proches, comme nous l'attendions au vu des études de simulations. L'importance du paramètre κ dans le modèle de Weibull étendu était parfaitement illustré par l'estimation de la pénétrance dans l'échantillon portugais. Nous avons trouvé une pénétrance estimée à 70 ans chez les porteurs de Val30Met de 50% chez les Français, de 91% chez les Portugais et de 36% chez les Suédois.

Dans l'étude suédoise, nous avons montré une différence significative entre les pénétrances selon les régions. Cette différence de risque peut s'expliquer d'un point de vue environnemental : Skelleftea étant une région où les industries sont assez différentes de celles de Pitea ou de Lycksele.

Il existe à Skelleftea une grande industie métallurgique et il existait anciennement des usines de textile. Une étude entre ces activités et la maladie dans le nord de la Suède a révélé que ces industries semblaient être un facteur de risque de la maladie [26].

Un autre résultat intéressant a été révélé par notre analyse : la différence significative de risque selon que la mutation ait été héritée de la mère ou du père. Nos résultats vont dans le même sens que ceux d'une précédente analyse généalogique qui soulignait que l'âge de début de la maladie était significativement plus élevé lorsque la mutation était héritée du père [15]. De plus, une différence selon le sexe du parent transmetteur a été trouvée significative dans l'échantillon de familles portugaises atteinte de NAH. Cependant, la différence n'était pas significative pour l'échantillon de familles françaises atteinte de NAH. Ce résultat provient probablement du manque de puissance pour détecter un effet dans le cas de pénétrances faibles ainsi que de la petite taille de l'échantillon. Une hypothèse à ces résultats serait que l'hétérogénéité des pénétrances soit due à un polymorphisme d'un gène modificateur dans l'ADN mitochondrial, qui est

transmis uniquement par la mère [6, 35].

Dans l'application aux familles atteintes de cancer du sein, la probabilité de sélection était très petite, due à l'inclusion d'un critère d'âge pour la sélection. Nous nous attendions à des résultats différents selon la méthode d'estimation utilisée puisque, dans ce cas, la vraisemblance Prospective fournit des estimations biaisées tandis que les résultats que nous avons obtenus avec la PEL sont vraisemblablement corrects. Toutefois, toutes les familles remplissant les critères de sélection n'ont pas été sélectionnées et les patients ayant subi un test génétique ont pu être motivés par une forte histoire familiale de cancer. Ce biais potentiel est extrêmement difficile à corriger mais nous devons en tenir compte dans l'interprétation de nos résultats.
Cependant, les estimations que nous avons obtenues avec la PEL sont assez proches de celles obtenues par Bonadona et al. [5] qui estimaient un risque cumulé à 70 ans de 57% pour les porteurs de BRCA1 et de 31% pour les porteurs de BRCA2 (notre estimation à 70 ans était de 40% après avoir rassemblé les porteurs BRCA1 et BRCA2).

Les intervalles de confiance étaient obtenus par bootstraps, ce qui suppose une indépendance entre les familles. Or, ce n'est pas forcément le cas lorsque les familles sont répliquées, notamment dans l'application à la NAH pour laquelle certaines familles avaient plusieurs proposants et ont donc été répliquées. Cependant, nous avons vérifié, à l'aide de simulations (nous avons simulé 1000 réplicats d'échantillon de taille 100 et avons calculé la variance empirique dans les diverses situations), que la variance n'était pas modifiée selon que les familles avec plusieurs proposants étaient répliquées ou non.

Dans ce chapitre, nous avons développé une méthode d'estimation paramétrique utilisant un modèle de Weibull. Dans les études de simulations, les phénotypes étaient donc simulés sous une loi de Weibull. Cependant, même si le modèle de Weibull peut être caractérisé par sa flexibilité, on peut se demander si l'hypothèse d'une loi de Weibull est toujours validée par les données et si l'approche paramétrique que nous avons eue jusqu'à présent est toujours justifiée. Ces questions font l'objet du dernier chapitre dans lequel nous présentons une méthode d'estimation de la pénétrance que nous avons développée, basée sur une approche non-paramétrique.

Le travail sur le développement de la méthode PEL et de l'étude de son comportement a donné lieu à une publication (Annexe F) dans *Genetic Epidemiology* en 2008.

Le travail sur l'étude des données suédoises pour la NAH a donné lieu à une publication (Annexe F) dans *Amyloid* en 2008.

Une lettre a été soumise dans le Journal *Amyloid* concernant l'étude du sexe du parent transmetteur dans les données françaises et portugaises (Annexe F).

Chapitre 5

Estimation de la pénétrance par une méthode non paramétrique

Dans ce chapitre, nous nous plaçons à nouveau dans le cas d'échantillons recensés sur l'existence d'au moins un atteint dans la famille (i.e. recensés indépendamment de l'histoire familiale). Dans ce cas, nous avons développé une méthode, que nous avons présentée dans le chapitre précédent, la méthode PEL, et nous avons montré, par des études de simulation, qu'elle était sans biais dans de nombreuses situations. Dans nos simulations, nous avions simulé les phénotypes des familles à partir d'une loi de Weibull. Mais que se passe-t-il lorsque l'échantillon ne suit pas une loi de Weibull ? La PEL parvient-elle encore à estimer sans biais la fonction de pénétrance pour n'importe quel échantillon, même lorsque la distribution de ce dernier est très éloignée d'une loi de Weibull ?

La PEL utilise effectivement une approche d'analyse de survie basée sur la modélisation de la fonction de pénétrance par une loi de Weibull. Ce modèle a été étendu par l'ajout de paramètres pour le rendre plus robuste à une erreur de modélisation. Cependant, l'ajout de paramètres peut dégrader considérablement les performances de l'estimation.

Nous avons donc développé une méthode non-paramétrique d'estimation de la fonction de pénétrance à l'aide de la méthode de vraisemblance empirique. Nous l'avons appelée $IDEAL$ pour Index Discarding EuclideAn Likelihood.

Afin de bien comprendre cette méthode, nous commencerons par présenter la méthode de la vraisemblance empirique. Nous décrirons ensuite notre méthode, $IDEAL$, à partir de la mé-

thode de la vraisemblance empirique. Enfin, nous la comparerons à la PEL à l'aide de données simulées sous différentes lois.

5.1 Introduction générale à la méthode de Vraisemblance Empirique

Dans cette section, nous présentons la méthode de la vraisemblance empirique (que nous noterons EL pour Empirical Likelihood), sur laquelle est basée notre méthode $IDEAL$, et qui a été pensée par Owen [36] dans le but de se libérer des hypothèses sur le modèle de distribution. La EL est une méthode d'estimation non paramétrique, c'est-à-dire qu'aucune hypothèse de loi n'est faite sur les données. La EL repose sur la méthode des moments au travers d'une équation de moments qui relie le paramètre (ou la fonction) à estimer et les données. La méthode des moments est aussi une méthode d'estimation non paramétrique qui se justifie asymptotiquement. La EL fait intervenir les lois multinomiales pour palier au manque de précision de la méthode des moments qui est efficace pour un nombre de données suffisamment grand. Ceci permet d'obtenir l'estimateur du maximum de la vraisemblance empirique (noté MEL) comme étant l'argument minimum d'un critère (cf. équation 5.11).
Ensuite, pour faire le lien avec les résultats connus de la théorie de la vraisemblance classique (comme la distribution asymptotique de l'estimateur du maximum de vraisemblance, la distribution asymptotique du rapport de vraisemblance, etc.), le critère obtenu avec la méthode des moments est redémontré par la théorie de la vraisemblance classique (théorème 5.1.2.1), qui donne à nouveau l'estimateur du maximum de vraisemblance empirique (équation 5.13).
Enfin, les propriétés asymptotiques de cet estimateur sont rappelées dans le théorème 5.1.2.2.

5.1.1 Une méthode des moments

La méthode des moments est une méthode classique pour les problèmes d'estimation dans lesquels le paramètre d'intérêt, θ, est défini comme la solution d'une équation faisant intervenir les moments de la distribution. Cette équation peut s'écrire de la manière suivante :

Pour une fonction régulière (i.e. dérivable 2 fois) m définie par :

$$m : \begin{cases} \mathbb{R}^p \times \mathbb{R}^d \to \mathbb{R}^q \\ X, \theta \to m(X, \theta), \end{cases} \tag{5.1}$$

θ_0, la vraie valeur du paramètre, est la solution de :

$$\mathbb{E}_{P_0}[m(X, \theta_0)] = 0, \tag{5.2}$$

où

* X est une variable aléatoire
* m est une fonction
* P_0 représente la "vraie" distribution de X
* $\mathbb{E}[.]$ définit l'espérance sous P_0
* 0 désigne le vecteur nul de dimension q ($q \geq d$)
* p représente la dimension des observations X
* d est la dimension du paramètre d'intérêt θ.

On remarque que l'espérance est calculée sous la vraie distribution P_0.

L'équation qui suit illustre le cas d'un moment d'ordre 1 et d'un moment d'ordre 2 avec le paramètre inconnu μ et/ou la matrice de covariance Σ respectivement :

$$\mathbb{E}_{P_0}[X - \mu] = 0 \quad \text{et} \quad \mathbb{E}_{P_0}[(X - \mu)(X - \mu)^* - \Sigma] = 0 \tag{5.3}$$

où $(X - \mu)^*$ est le vecteur transposé du vecteur $(X - \mu)$.

La méthode des moments consiste à résoudre en θ la version empirique de l'équation 5.2. On a donc :

$$\frac{1}{n} \sum_{i=1}^{n} m(X_i, \theta) = 0, \tag{5.4}$$

et la solution de l'équation 5.4 sera l'estimateur des moments, $\hat{\theta}$.

Plus formellement, on peut l'écrire de la façon suivante :

$$\mathbb{E}_{P_n}[m(X,\theta)] = 0, \tag{5.5}$$

où

$$P_n(x) = \frac{1}{n}\sum_{i=1}^{n} \delta_{x_i}(x), \tag{5.6}$$

où δ_{x_i} est la mesure de Dirac au point x_i.

Cette approche très classique donne de bons résultats quand le nombre n de données est assez grand car elle repose essentiellement sur des conditions asymptotiques. Mais, lorsque le nombre de données est trop petit, la probabilité P_n peut devenir une approximation très médiocre de la vraie probabilité P_0. L'estimateur des moments sera alors très éloigné de la vraie valeur θ_0 du paramètre.
Pour combler cette lacune, on introduit la famille des distributions multinomiales Q de la façon suivante :

$$Q(x) = \begin{cases} q_i \ \text{ si } \exists i, x = x_i \\ 0 \ \ sinon \end{cases} \tag{5.7}$$

où $0 < q_i < 1$ et $\sum_{i=1}^{n} q_i = 1$.

On remarque qu'il s'agit du choix de dimension maximale que l'on peut prendre pour une famille de distribution puisque la dimension de la loi multinomiale est égale au nombre d'observations (en fait, q_i représente la probabilité d'avoir x_i). En remplaçant P_n par Q, l'équation des moments s'écrit :

$$\mathbb{E}_{Q}[m(X,\theta)] = 0 \tag{5.8}$$

Ce qui s'écrit aussi :

$$\sum_{i=1}^{n} q_i[m(X_i, \theta)] = 0. \tag{5.9}$$

Dans l'estimation du paramètre d'intérêt θ, les poids $(q_1, \ldots, q_n)$ introduits dans l'équation 5.7 (qui sont les paramètres de la multinomiale) sont considérés comme des paramètres de nuisance. Pour obtenir un critère "seulement en θ", on va optimiser par rapport aux autres paramètres. On va donc optimiser en Q en utilisant la distance de Kullback K (voir Annexe C). On obtient alors le **critère** suivant :

$$C(\theta) = 2n \min_{G}\{K(Q, P_n) \mid \mathbb{E}_Q[m(X, \theta)] = 0\}, \tag{5.10}$$

qui conduit à l'estimateur suivant :

$$\hat{\theta} = \arg\min_{\theta}\{C(\theta)\} \tag{5.11}$$

En fait, $C(\theta)$ est le rapport de vraisemblance empirique (ELR). C'est ce que nous allons voir dans la section qui suit en faisant ressortir le lien qui existe entre cette approche (que nous venons de décrire) et la méthode classique du maximum de vraisemblance.
Notons que, dans cette section, nous n'avons à aucun moment supposé de modèle de loi sur les données. Nous n'avons donc fait aucune modélisation et avons utilisé Q comme une estimation de la vraie loi des données.

5.1.2 Vraisemblance, Vraisemblance Empirique et rapport de Vraisemblance Empirique

Dans cette section, nous allons montrer comment la procédure décrite dans la section précédente peut être interprétée dans un contexte de vraisemblance classique, et comment le critère C peut être interprété comme un rapport de vraisemblance.
On va considérer la multinomiale Q comme un modèle paramétrique des données. Cette hypothèse de modèle paramétrique est considérée ici uniquement dans le but d'introduire formellement la méthode de la Vraisemblance Empirique (EL) comme une méthode de vrai-

semblance classique. Pour cela on définit, pour Q et θ vérifiant l'équation des moments (5.8) ($\mathbb{E}_Q[m(X,\theta)] = 0$), la densité de probabilité correspondant à Q de la manière suivante :

$$g_{(\theta,q_1,\ldots,q_n)}(x) = \begin{cases} q_i \ \text{ si } \exists i, x = x_i \\ 0 \ \ sinon \end{cases} \tag{5.12}$$

La fonction de vraisemblance correspondant à la densité de probabilité $g_{(\theta,q_1,\ldots,q_n)}$, telle que l'équation des moments (5.8) est vérifiée, est appelée Vraisemblance Empirique (EL, pour Empirical Likelihood) [36] et est définie de la façon suivante :

$$\begin{aligned} EL(\theta) &= \sup_{(q_1,\ldots,q_n)} \left\{ \prod_{i=1}^{n} g_{(\theta,q_1,\ldots,q_i)}(X_i) \middle| \mathbb{E}_Q[m(X,\theta)] = 0 \right\} \\ &= \sup_{(q_1,\ldots,q_n)} \left\{ \prod_{i=1}^{n} q_i \middle| \sum_{i=1}^{n} q_i m(X_i,\theta) = 0, \ \sum_{i=1}^{n} q_i = 1 \right\}. \end{aligned}$$

Comme dans le processus classique de maximum de vraisemblance, le rapport de log-vraisemblance correspondant à $EL(\theta)$ est défini par :

$$ELR(\theta) = -2\log\left(\frac{EL(\theta)}{\max_\theta \{EL(\theta)\}}\right) \tag{5.13}$$

Enfin, le théorème qui suit fait le lien entre $ELR(\theta)$ et $C(\theta)$.

Théorème 5.1.2.1

Si $q = d$, $C(\theta)$, défini par l'équation (5.10), est égal à $ELR(\theta)$:

$$C(\theta) = ELR(\theta)$$

La preuve de ce théorème est reportée dans l'annexe D.

On utilisera dorénavant la notation $ELR(\theta)$ au lieu de $C(\theta)$. Donc, l'estimateur $\hat{\theta}$ défini par l'équation (5.11) est l'estimateur du Maximum de Vraisemblance Empirique (MEL pour Maximum Empirical Likelihood), $\hat{\theta}_{MEL}$ et est donné par :

$$\hat{\theta}_{MEL} = \arg\max_\theta EL(\theta) = \arg\min_\theta ELR(\theta) \tag{5.14}$$

Le théorème suivant donne le comportement asymptotique de l'estimateur MEL.

Théorème 5.1.2.2

Soit $(X_1, \ldots, X_n)$ un échantillon indépendant et identiquement distribué sous P_0 tel que $\mathbb{E}[m(X_i, \theta_0)] = 0$. Sous certaines conditions de régularités (cf. [39]), l'estimateur du MEL donné par

$$\hat{\theta}_{MEL} = \arg\max_{\theta} EL(\theta), \tag{5.15}$$

est un estimateur asymptotiquement gaussien de θ_0 :

$$\sqrt{N}(\hat{\theta}_{MEL} - \theta_0) \xrightarrow[n\to\infty]{loi} \mathcal{N}(0, \sigma^2). \tag{5.16}$$

La Preuve du théorème 5.1.2.2 est dans [39].

Le théorème 5.1.2.2 donne le comportement asymptotique de l'estimateur MEL dans le contexte général. L'estimateur MEL est alors défini comme étant la solution d'une équation d'optimisation, résolue grâce à une méthode du Lagrangien. Un des avantages majeurs de cette estimation, est qu'il est possible d'ajouter des informations supplémentaires dans l'équation d'optimisation. Cela peut se traduire de la manière suivante :

$$\hat{\theta}_{MEL} = \arg\max_{\theta} \{EL(\theta) | \theta \in \mathcal{E}\}, \tag{5.17}$$

où $\mathcal{E}$ est un ensemble de contraintes sur le paramètre d'intérêt θ.

5.2 Présentation de $IDEAL$ (Index Discarding EuclideAn Likelihood)

Dans la section précédente, nous avons donné une description de la vraisemblance empirique. Dans son livre [36], Owen décrit très précisément la vraisemblance empirique et expose ses multiples applications.
La méthode que nous avons développée, $IDEAL$, est basée sur la théorie de la vraisemblance empirique.

$IDEAL$ est une méthode non paramétrique qui estime la fonction de pénétrance. Cette méthode s'applique à tous les modèles de maladies lorsque les familles ont été sélectionnées sur au moins un atteint.
Comme la PEL, $IDEAL$ corrige pour les critères de sélection en s'inspirant de la méthode décrite par Weinberg [14, 47]. De plus, $IDEAL$ fournit des bandes de confiance, c'est-à-dire deux fonctions qui bornent la pénétrance à chaque âge t.

Nous commencerons, dans cette section, par présenter la vraisemblance empirique appliquée à une fonction de répartition, puis nous présenterons sa version Euclidienne. Nous présenterons ensuite les modifications apportées afin de tenir compte de la sélection. Enfin, nous décrirons les bandes de confiance pour la fonction de pénétrance.

5.2.1 La vraisemblance empirique adaptée à l'estimation d'une fonction de répartition

Soit une suite de vecteurs aléatoires $X_1, ..., X_n \in \mathbb{R}^p$, indépendants et identiquement distribués ($i.i.d$) et qui suivent une loi P_0. $X_1, ..., X_n$ est donc un n échantillon. L'idée de Owen est de construire une vraisemblance en se donnant comme modèle l'ensemble des multinomiales qui ne chargent que l'échantillon (i.e. toute la probabilité est contenue dans l'échantillon : $\sum \mathbb{P}(X_i = x_i) = 1$), comme nous l'avions vu dans la section précédente. On rappelle

ici que la vraisemblance empirique (EL) peut être appliquée dès que le paramètre d'intérêt θ_0 est défini comme la solution de l'équation :

$$\mathbb{E}_{P_0}[m(X, \theta_0)] = 0$$

où

* X est une variable aléatoire
* m est une fonction
* P_0 représente la "vraie" distribution de X
* $\mathbb{E}[.]$ définit l'espérance sous P_0

Donc, selon les observations $X_1, ..., X_n$, pour estimer θ_0, on considère les valeurs de θ tel que $m(X_i, \theta)$ soit de moyenne nulle :

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

Dans le cas de l'estimation d'une fonction de répartition F, le problème s'écrit donc de la façon suivante : $\forall t > 0$,

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

où

* θ_0 correspond à $F(t)$
* $A =$ l'âge de début de la maladie. Par la suite, on notera $X = (Y, G, P, Pb)$ où X résume l'information pour un individu avec : , $Y =$ l'âge de l'individu, $G =$ le génotype, $P =$ le phénotype et $Pb = 1$ si l'individu est un proposant et $Pb = 0$ sinon.
* $m(x, y) = \mathbb{1}_{x \leq t} - y$ (où $\mathbb{1}_{x \leq t}$ est la fonction indicatrice de l'évènement $x \leq t$).

Dans la suite, θ_0 sera $F(t)$ et θ représente les valeurs possibles de θ_0.

La EL est donc construite au moyen de distributions multinomiales sur l'échantillon $(X_1, \ldots, X_n)$:

$$Q(x) = \begin{cases} q_i \text{ si } \exists i, x = x_i, \\ 0 \ sinon \end{cases}$$

avec $0 < q_i < 1$ et $\sum q_i = 1$.

On a la EL :

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

et l'estimateur MEL est donné par : $\hat{\theta} = \arg\max_\theta \{EL(\theta, t)\}$ qui est un estimateur asymptotiquement gaussien de θ_0 quelle que soit la distribution des données.

Le développement du rapport de la log-vraisemblance fait apparaître la divergence de Kullback (cf. section 5.3) :

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

où $K(Q, P_n) = -\int log\left(\frac{dQ}{dP_n}\right) dP_n$, où Q est la multinomiale et P_n est la probabilité empirique (qui est la meilleure multinomiale qui maximise la vraisemblance sans contrainte) :

$$P_n(x) = \begin{cases} \dfrac{1}{n} \text{ si } \exists i, x = x_i, \\ 0 \ sinon. \end{cases}$$

La méthode de la EL consiste donc à minimiser la divergence de Kullback entre Q et P_n. Cependant, d'autres divergences peuvent être utilisées comme par exemple la distance Euclidienne [4]. Dans notre méthode, nous utiliserons la distance Euclidienne (que nous noterons χ^2) au lieu de la divergence de Kullback dans l'expression du rapport de log-vraisemblance essentiellement pour des raisons de temps de calcul.

5.2.2 La Vraisemblance Euclidienne

La EL devient donc EAL pour (EuclideAn Likelihood) :

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

Comme pour la EL, la maximisation de $EAL(\theta,t)$ en θ donne un estimateur asymptotiquement gaussien de θ_0.

5.2.3 La Vraisemblance Euclidienne pour l'estimation de la fonction de pénétrance

Nous avons modifié la Vraisemblance Euclidienne à travers la mesure de probabilité P_n pour l'adapter à l'estimation de la fonction de pénétrance. Nous noterons W la nouvelle mesure de référence.

Pour estimer la pénétrance à un âge t, nous allons considérer uniquement l'ensemble des individus porteurs de la mutation et âgés de plus de t ans. Cela entraîne une variation de la mesure de référence W en fonction de t :

$$W(x) = \begin{cases} \dfrac{1}{n} \text{ si } \exists i, x = X_i, G_i = 1 \text{ et } Y_i \geq t, \\ 0 \ \ sinon. \end{cases}$$

Pour un individu i non atteint, l'âge de début de la maladie, A_i, n'existe pas. Formellement, l'information que nous allons utiliser est que, pour un tel individu i, A_i est plus grand que son âge Y_i. Dans ce cas, nous noterons donc que $A_i = +\infty$.

5.2.4 Prise en compte de la sélection

Les familles sélectionnées représentent un échantillon biaisé, dans lequel les individus atteints sont sur-représentés. Dans ce cas, $\theta_0 = F(t)$ ne vérifie pas l'équation d'estimation, et on a :

$$\mathbb{E}_{P_0}(\mathbb{1}_{A\leq t} - \theta_0) \neq 0,$$

où P_0 est la loi de probabilité sous laquelle les données (biaisées) sont distribuées.
Nous proposons donc ici de corriger pour le biais de sélection en pondérant les proposants avec un poids plus faible que les autres individus : si une famille possède k proposants potentiels, chacun sera pondéré par $1 - \frac{1}{k}$. Cette méthode est fortement inspiré de la méthode de Weinberg, comme dans le cas de la PEL [14, 47].

Avec cette nouvelle distribution, l'équation d'estimation $\mathbb{E}[\mathbb{1}_{A\leq t} - \theta_0]$ fournit une estimation non biaisée de θ_0. La mesure de référence W s'écrit alors :

$$W(x) = \begin{cases} \frac{1}{n} & \text{si } \exists i, x = X_i, Pb_i = 0, G_i = 1 \text{ et } Y_i \geq t, \\ \frac{1}{n} * \left(1 - \frac{1}{k}\right) & \text{si } \exists i, x = X_i, Pb_i = 1, G_i = 1 \text{ et } Y_i \geq t, \\ 0 \quad \text{sinon.} \end{cases} \tag{5.18}$$

Nous avons alors une nouvelle vraisemblance avec une nouvelle mesure de référence. Nous l'avons appelée $IDEAL$ pour Index Discarding EuclideAn Likelihood. On rappelle que k est le nombre de proposants dans la famille ; $Pb_i = 1$ si l'individu i est proposant et $Pb_i = 0$ sinon ; $G_i = 1$ si i est porteur de la mutation, 0 sinon et Y_i représente l'âge de l'individu i.
La démonstration de la correction pour la sélection est mise en annexe E.

5.2.5 Bandes de confiance

Une des propriétés importantes de la vraisemblance empirique mais aussi des vraisemblances qui s'y apparentent telle que la vraisemblance euclidienne, est qu'elle fournit des bandes de confiance pour la fonction de répartition F (cf. Chapitre 7 de [36]). Cela signifie que, quel que soit le niveau de confiance, par exemple 95%, on peut avoir deux fonctions de répartitions J et H telles que, avec une probabilité de 95% et pour tout t, on ait :

$$J(t) \leq F(t) \leq H(t)$$

La différence fondamentale avec l'intervalle de confiance, qui est donné t par t est que la bande de confiance est globale et s'applique à tout t : avec un probabilité de 95%, la fonction F se situe entre J et H pour tout t.

Intervalles de confiance : $\forall t > 0,\ \ \mathbb{P}(J(t) \leq F(t) \leq H(t)) = 95\%$,

Bandes de confiance : $\mathbb{P}(\forall t > 0, J(t) \leq F(t) \leq H(t)) = 95\%$.

De plus, les bandes de confiance ne sont pas asymptotiques, et on a :

$$J(t) = min\{\theta | IDEAL(\theta, t) \leq c_n\},$$
$$H(t) = max\{\theta | IDEAL(\theta, t) \leq c_n\},$$

où les valeurs critiques c_n sont répertoriées (cf. p159, du livre [36]).

5.3 Étude de $IDEAL$ et comparaison avec la PEL

Dans un premier temps, nous avons étudié la méthode $IDEAL$ en terme de biais relatif à l'aide de familles simulées sous une loi de Weibull, comme nous l'avions fait avec notre méthode paramétrique (la PEL) dans le chapitre précédent. Nous avons ensuite comparé les deux méthodes en terme de biais relatif dans le cas où les familles étaient simulées sous différentes lois, autre que la loi de Weibull. On reprendra la définition du biais relatif donné au chapitre 4.

5.3.1 Simulations

Les familles ont été simulées avec le même processus que dans le chapitre 4. Pour être dans les conditions asymptotiques, la taille des échantillons a été fixée à 5 000 familles après sélection, et donc à 90 000 individus puisque chaque famille contient un couple ancêtre avec 4 enfants, chacun d'eux ayant 4 enfants avec leur conjoint ou conjointe respectif. Afin de limiter le temps de calcul, nous avons fixé la fréquence de l'allèle délétère (noté fq) à 0.1. Nous avons fixé le taux de mutation *de novo* (noté pn) à 0 et nous avons considéré dans nos simulations que tous les génotypes étaient connus.
Comme dans le chapitre précédent, nous avons supposé deux types de modèle :

- Le modèle MM correspondant à la simulation d'une maladie mendélienne et dans lequel nous avons simulé un risque nul d'être atteint chez les non porteurs.
- Le modèle MCSM correspondant à la simulation d'une maladie complexe à sous-entités monogéniques et pour laquelle le risque d'être atteint chez les porteurs de la mutation a été fixé à 0.1 à 80 ans. Pour cette maladie, un critère d'âge de 36 ans pour la sélection a été introduit.

Afin d'étudier le comportement de la méthode $IDEAL$, nous avons simulé les phénotypes avec une fonction âge-dépendante basée sur le modèle de Weibull étendu (cf. Chapitre 2) correspondant à une valeur de risque cumulé de 0.5 à 80 ans.
Ensuite, pour comparer $IDEAL$ à la méthode PEL, nous avons simulé les phénotypes avec une fonction âge-dépendante basée sur d'autres modèles que celui de Weibull : Une loi uniforme (composée de lois uniformes) puis une loi de Cauchy. Nous avons choisi ces lois car elles sont

très différentes d'une loi de Weibull et peuvent être difficilement "approchables" par une telle loi.
Comme dans les chapitres précédents, nous avons introduit deux longueurs de périodes ($T = 20$ ans et $T = 1$ an) et deux probabilités ($p_s = 1$ et $p_s = 0.5$) pour la sélection des familles et nous avons fait l'hypothèse d'une transmission dominante.

5.3.2 Comportement de $IDEAL$ sous un modèle de Weibull

Afin de réduire le nombre de cas étudiés, nous avons réduit le nombre de situations à deux schémas de simulation et de sélection opposés :

* S_{Faible} : Simulation d'un modèle MCSM avec un critère d'âge de 36 ans et d'une probabilité de sélection faible ($T = 1$ et $p_s = 0.5$).
* S_{Forte} : Simulation d'un modèle MM et d'une probabilité de sélection forte ($T = 20$ et $p_s = 1$).

Nous avons donc simulé des familles sous le schéma S_{Faible} et la figure 5.1 montre la courbe de pénétrance estimée avec $IDEAL$. On constate que la méthode est sans biais et que la vraie courbe de pénétrance (courbe avec $*$) et la courbe de pénétrance estimée par $IDEAL$ (courbe en pointillés) sont pratiquement superposées. Les courbes en rouges représentent les bandes de confiance à 95% et on constate qu'elles sont très proches de la pénétrance estimée.
On rappelle que, dans ce même cas, la PEL donne également des estimations non biaisées de la pénétrance.

La figure 5.2 montre la courbe de pénétrance estimée avec $IDEAL$ sous le schéma S_{Forte}. Comme dans la figure 5.1, $IDEAL$ est sans biais et la courbe estimée et la vraie courbe sont superposées.

Dans les deux schémas de simulation S_{Faible} et S_{Forte}, $IDEAL$ donne donc des estimations sans biais. On peut donc logiquement supposer que la méthode est sans biais dans tous les autres shémas "intermédiaires" de simulations que nous avions décrit dans le chapitre 4. Nous pouvons conclure que, comme la PEL, $IDEAL$ est sans biais lorsque les familles sont simu-

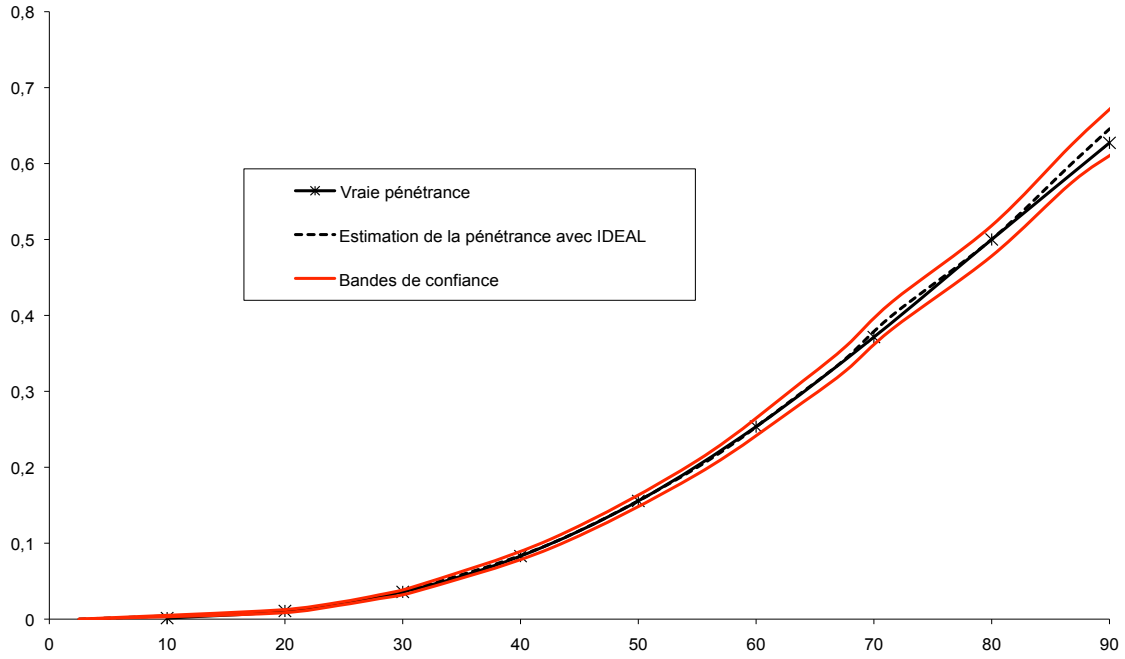


FIG. 5.1 – Biais relatif de $IDEAL$ dans le schéma de séléction S_{Faible}

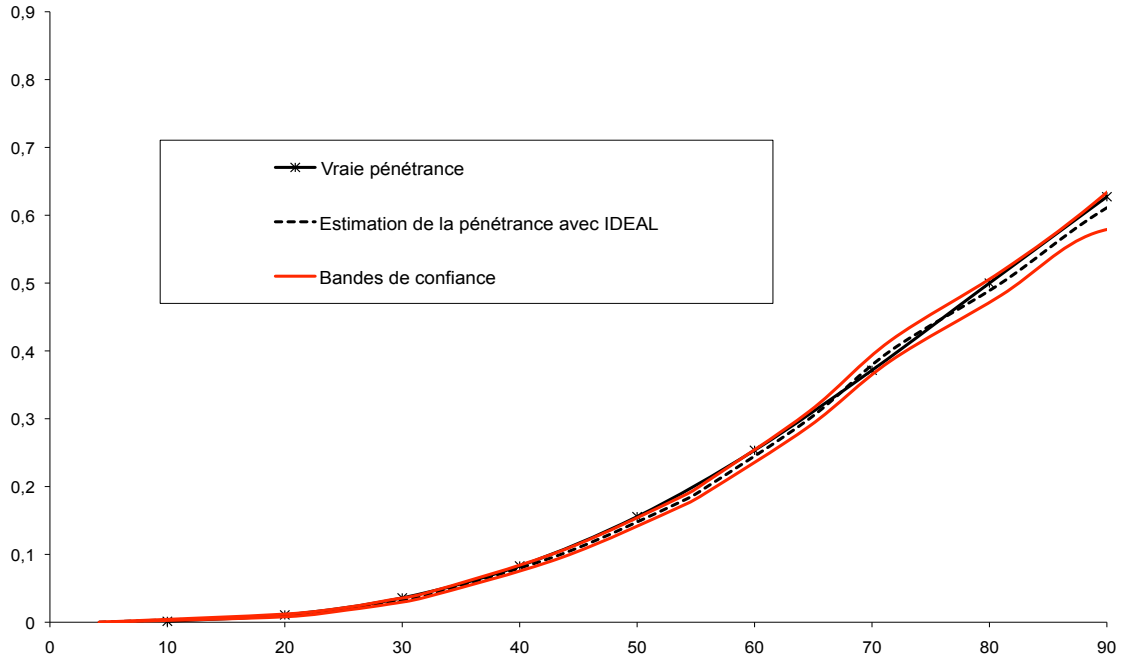


FIG. 5.2 – Biais relatif de $IDEAL$ dans le schéma de séléction S_{Forte}

lées sous une loi de Weibull. Nous allons maintenant comparer les deux méthodes dans le cas de familles simulées sous des lois très différentes de la loi de Weibull.

5.3.3 Comparaison de $IDEAL$ et de la PEL sous un modèle de loi uniforme et de loi de Cauchy

Nous avons choisi de montrer les résultats dans le cas de simulations sous un modèle MCSM avec un critère d'âge de 36 ans. Pour la sélection, nous avons étudié le cas où $T = 20$ ans et $p_s = 0.5$.
Les données ont été simulées sous une loi de Cauchy puis sous une loi uniforme.
La loi de Cauchy de paramètres ($\alpha, a > 0$), définit sur $\mathbb{R}$, s'écrit :

$$f_{\alpha,a}(x) = \frac{1}{\pi a \left(1 + \left(\frac{x-\alpha}{a}\right)^2\right)},$$

où α est appelé paramètre de location et a, le paramètre d'échelle.
Dans nos simulations, nous avons fixé les deux paramètres de la loi de Cauchy α et a à 0 et 5 respectivement.

La figure 5.3 montre l'estimation de la fonction de pénétrance avec $IDEAL$ et avec la PEL lorsque les phénotypes sont simulés sous un modèle âge-dépendant basé sur une loi uniforme. On constate que la PEL ne s'ajuste pas bien à la vraie courbe tandis que la pénétrance estimée par la méthode $IDEAL$ est parfaitement sans biais.

La figure 5.4 montre l'estimation de la fonction de pénétrance avec $IDEAL$ et avec la PEL lorsque les phénotypes sont simulés sous un modèle âge-dépendant basé sur une loi de Cauchy. Ici encore, l'estimation de la fonction de pénétrance avec $IDEAL$ est sans biais tandis que la courbe estimée avec la PEL est très biaisée.

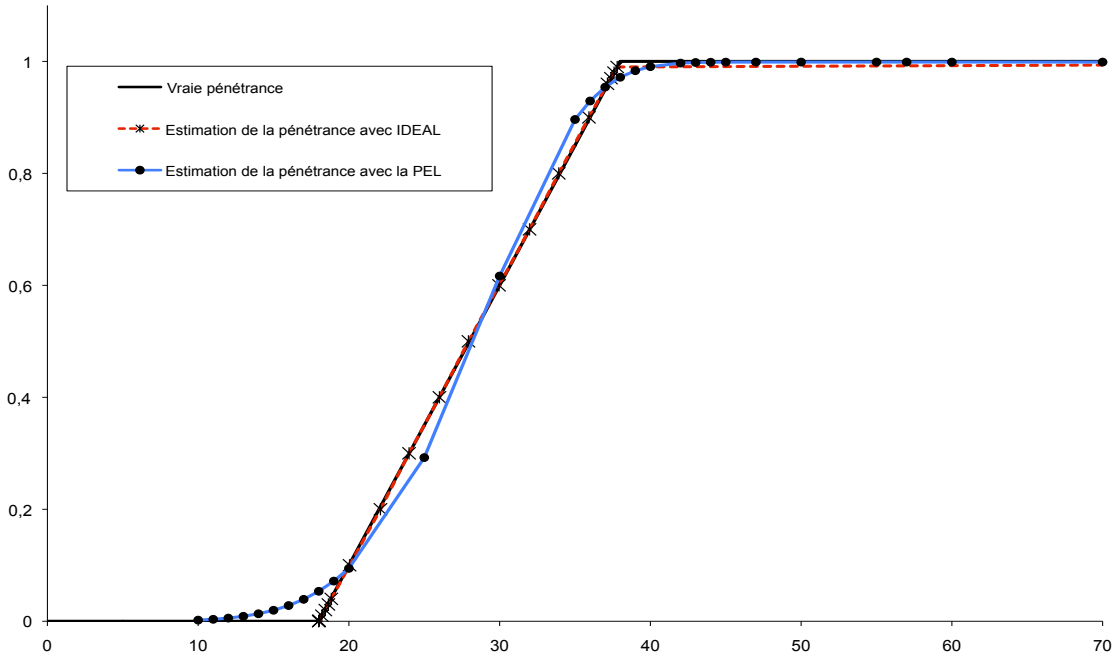


FIG. 5.3 – Comparaison de IDEAL et de la PEL sous une loi Uniforme

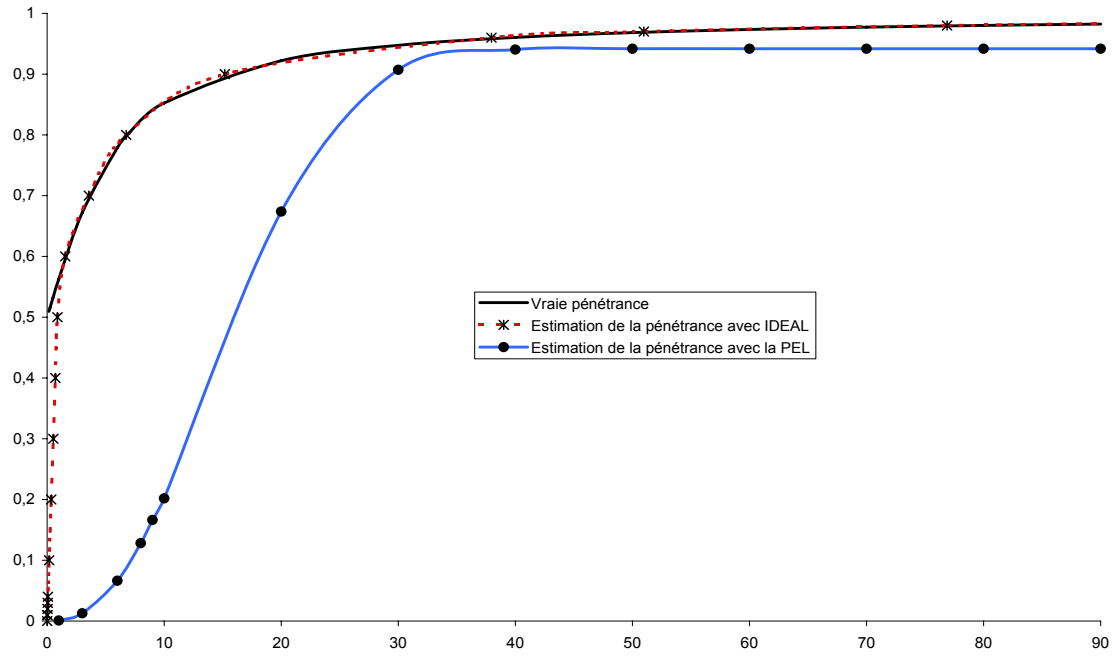


FIG. 5.4 – Comparaison de $IDEAL$ et de la PEL sous une loi de Cauchy

5.3.4 Sensibilité de $IDEAL$ et de la PEL à la taille de l'échantillon

Nous avons comparé la sensibilité de $IDEAL$ et de la PEL à la taille de l'échantillon. Pour cela, nous avons repris les différentes situations vues précédemment. Nous avons d'abord simulé des échantillons de 200 familles, sous une loi de Weibull, et dans les deux schémas opposés S_{Forte} et S_{Faible}. Puis nous avons simulé des échantillons de 200 familles, sous une loi uniforme et enfin, sous une loi de Cauchy.
Les figures 5.5 et 5.6 montrent les résultats obtenus respectivement avec un schéma de simulation S_{Forte} et S_{Faible}, lorsque les données suivent une loi de Weibull.

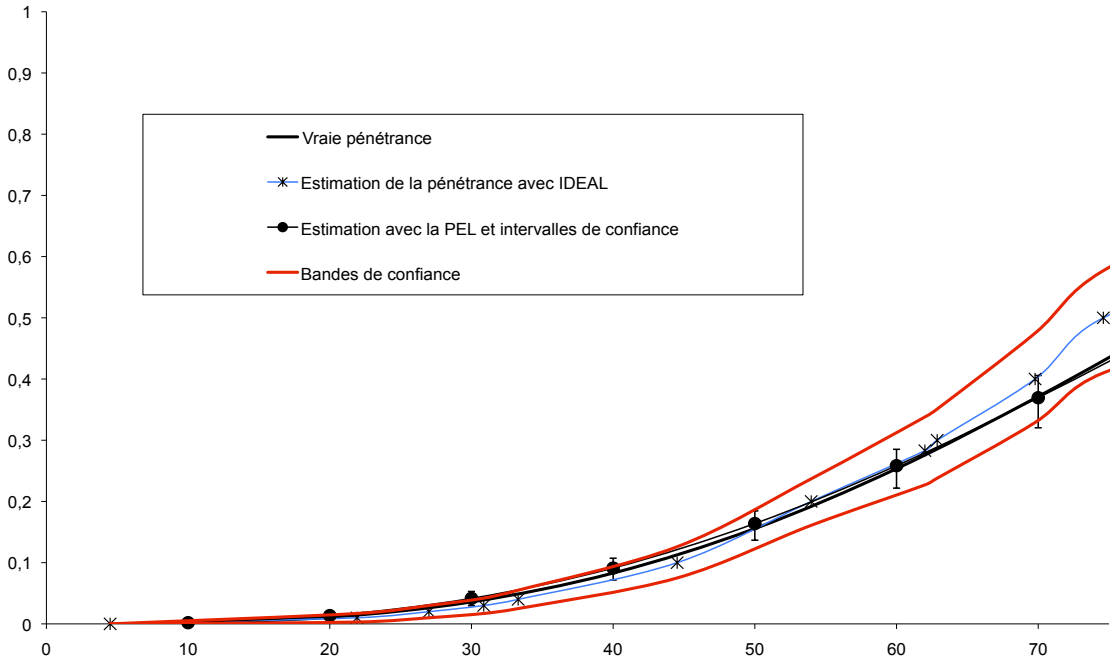


FIG. 5.5 – Estimations de la fonction de pénétrance dans un échantillon de 200 familles dans le cas d'un schéma S_{Forte}

On constate que dans ce cas, la PEL a un biais négligeable tandis que la méthode $IDEAL$ est sans biais dans les premiers âges, jusqu'à 60 ans puis, elle est biaisée à partir de 60 ans. De plus, les intervalles de confiance de la PEL sont beaucoup plus petits que les bandes de confiance obtenues avec $IDEAL$. Dans le cas du schéma S_{Faible}, la bande de confiance inférieure est même supérieure à la vraie pénétrance, mais ceci uniquement pour les âges au-delà de 65 ans.

Les figures 5.7 et 5.8 présentent les estimations de la pénétrance avec $IDEAL$ et la PEL

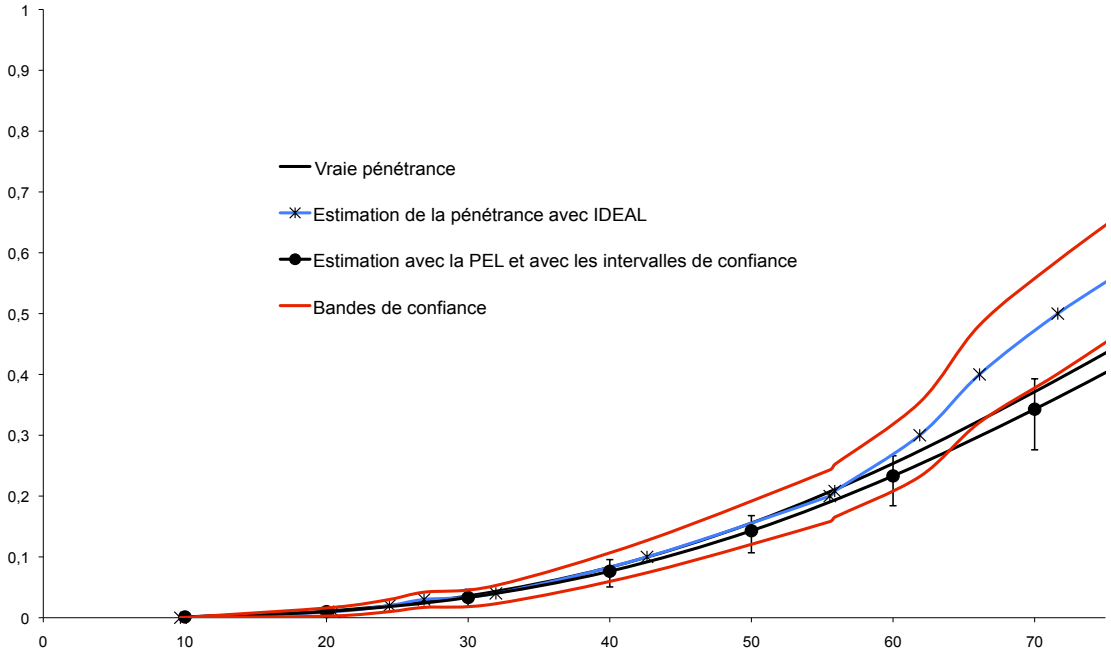


FIG. 5.6 – Estimations de la fonction de pénétrance dans un échantillon de 200 familles dans le cas d'un d'un schéma S_{Faible}

pour un échantillon simulé respectivement sous une loi uniforme et sous une loi de Cauchy. Pour chacune des figures, nous avons donné les intervalles de confiance pour la PEL aux points 30, 40 et 50 ans pour la loi uniforme et aux points 50, 60 et 70 ans pour la loi de Cauchy. Les bandes de confiance calculées par $IDEAL$ ont été données dans les deux figures. Dans les deux cas, on constate que $IDEAL$ donne une estimation meilleure que la PEL et que les bandes de confiance sont plus "étroites" que les intervalles de confiance.

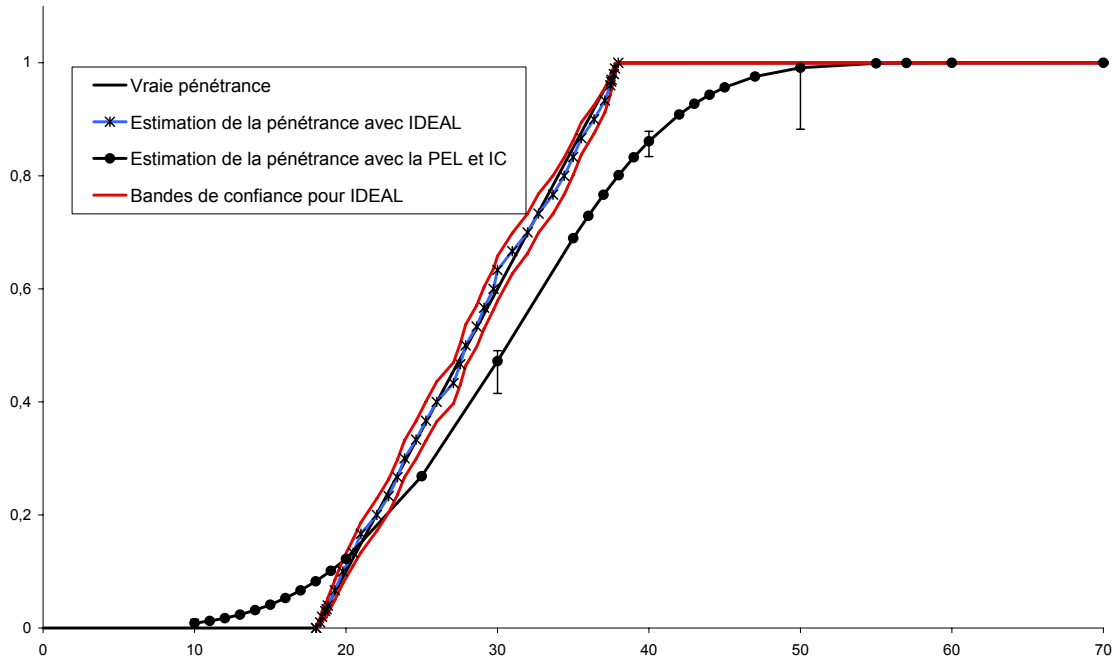


FIG. 5.7 – Estimation de la fonction de pénétrance un échantillon de 200 familles sous une loi uniforme

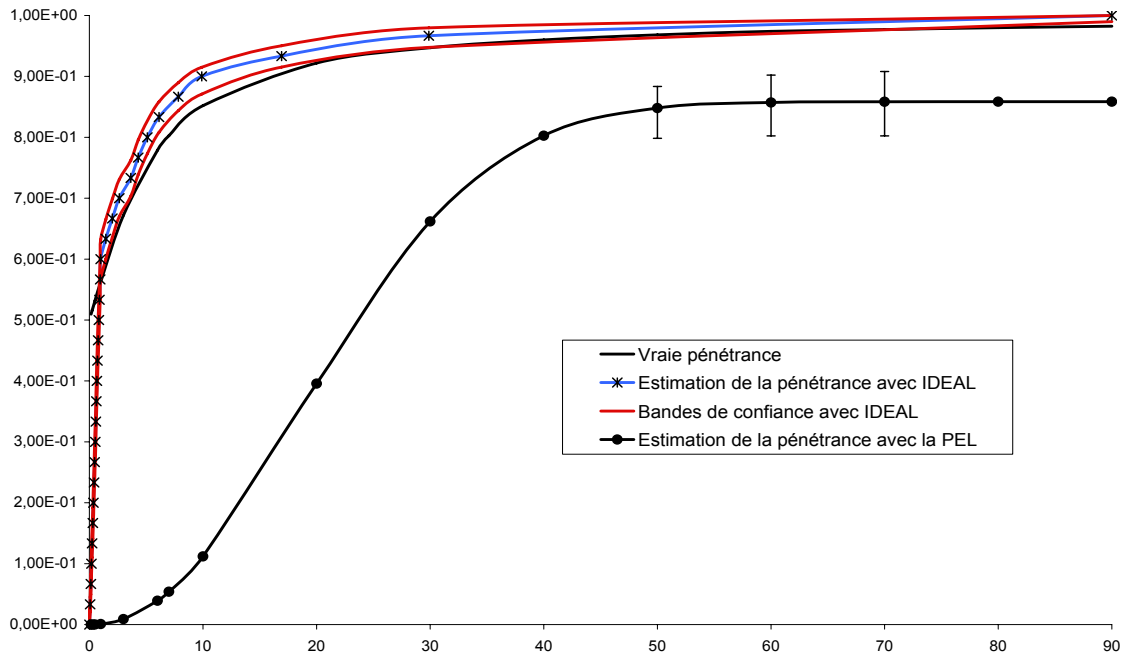


FIG. 5.8 – Estimation de la fonction de pénétrance dans un échantillon de 200 familles sous une loi de Cauchy

5.4 Conclusion et discussion

Dans ce chapitre, nous avons proposé une méthode d'estimation non paramétrique, $IDEAL$, adaptée pour la fonction de pénétrance et qui corrige pour le biais de sélection dans le cas où les familles sont sélectionnées par l'intermédiaire d'un atteint. Nous avons comparé cette méthode à la méthode PEL, décrite dans le chapitre 3, qui est une méthode paramétrique qui prend également en compte la sélection des familles. La différence fondamentale entre ces deux méthodes est que la PEL utilise une approche d'analyse de survie basée sur un modèle de Weibull pour modéliser la pénétrance alors que $IDEAL$ ne fait aucune hypothèse sur la loi de la pénétrance.

Nous avons d'abord démontré formellement dans l'annexe E que la correction pour la sélection utilisée dans $IDEAL$ fournissait des estimations sans biais. Puis, à partir de grands échantillons de données simulées sous une loi de Weibull, nous avons montré que $IDEAL$ fournissait des estimations non biaisées de la pénétrance. Nous avions montré dans le chapitre 4 que la PEL fournissait également des estimations sans biais dans ce cas.
En revanche, lorsque la loi des données s'écarte d'une loi de Weibull, nous avons montré que la PEL était biaisée et pouvait même être fortement biaisée (par exemple, lorsque les données sont simulées sous une loi de Cauchy) tandis que $IDEAL$ fournissait des courbes de pénétrance non biaisées.

Dans cette étude faite à partir de données simulées, nous avons considéré que l'ensemble des génotypes des individus était connu. Cependant, les génotypes inconnus peuvent être pris en compte dans la méthode $IDEAL$ par des pondérations. En effet, à partir des génotypes connus, on peut estimer la probabilité p_i de recevoir la mutation pour chaque individu i non génotypé. Dans la mesure de référence, on utilisera donc la valeur de p_i/n plutôt que $1/n$.

Nous avons aussi étudié le comportement de notre méthode dans le cas d'échantillons de taille non asymptotique. Dans ce cas, l'approche paramétrique (i.e. utilisation du modèle de Weibull étendu) est très utile pour "parer" au manque d'information en "forçant" la forme de la courbe. Dans ce cas, la PEL donne donc, logiquement, de meilleures estimations que $IDEAL$,

lorsque les données sont simulées sous une loi de Weibull. Cependant, lorsque la loi des données est éloignée d'une loi de Weibull, $IDEAL$ fournit alors de meilleures estimations que la PEL.

Une autre caractéristique de $IDEAL$ est le fait que la méthode fournisse des bandes de confiance plutôt que des intervalles de confiance. Notre étude a montré que la largeur de ces bandes de confiance augmente avec l'âge t. On peut expliquer cela par le fait que seuls les individus dont l'âge est supérieur à t sont considérés pour estimer la pénétrance en t. La population considérée diminue donc avec t et cela entraîne un élargissement des bandes de confiance puisque ces dernières fournissent l'information sur la précision de l'estimation de la fonction de pénétrance en t.

Enfin, nous avons remarqué dans la littérature que la fonction de pénétrance était souvent modélisée en utilisant une approche d'analyse de survie basée sur un modèle des données sans que cette approche soit formellement validée. Il peut donc être utile d'utiliser notre méthode non paramétrique, $IDEAL$, au préalable afin de valider ou non une approche paramétrique. Si l'estimation donnée par $IDEAL$ est proche de celle fournie par la PEL, l'utilisation de la seconde méthode sera préférentiellement utilisée. Dans le cas contraire, on préférera l'utilisation de la méthode non paramétrique.

Ce travail a donné lieu à une publication (Annexe F) dans *Genetic Epidemiology* en 2008.

Conclusion et Perspectives

L'estimation de risque de maladies dues à des mutations génétiques est indispensable non seulement à la prise en charge des individus porteurs de la mutation mais également à une meilleure compréhension de la maladie et de ses mécanismes sous-jacents. Les données qui permettent d'estimer ce risque sont des données familiales. Les familles ont pu être recensées dans un cadre de recherche (il s'agit de recenser des familles à travers un protocole établi au préalable) ou elles peuvent être recensées dans un but de prise en charge (il s'agit alors de détecter les individus porteurs de la mutation le plus tôt possible afin de les prendre en charge médicalement). Dans tous les cas, les maladies étudiées étant des maladies pour lesquelles la forme génétique est rare, les familles ne sont jamais recensées aléatoirement dans la population mais à partir de certains critères.
Lors de l'estimation des risques associés aux mutations, le fait de ne pas tenir compte du mode de recensement biaise fortement les résultats.

Dans ce travail, nous nous sommes intéressés au développement de méthodes prenant en compte le recensement des familles. Pour cela, nous avons distingué deux types de recensements : Le recensement sur des critères complexes et difficiles à modéliser d'une part et le recensement sur des critères indépendants de l'histoire familiale d'autre part.
Nous avons présenté une méthode, la GRL, développé par Carayol et Bonaïti-Pellié, qui permet d'estimer la fonction de pénétrance quels que soient les critères de recensement des familles. Cette méthode est basée sur la vraisemblance rétrospective qui est peu efficace. Nous avons également pu constater son manque d'efficacité à travers les intervalles de confiance calculés lors de l'application de la GRL à des données HNPCC. De plus, nous avons montré que la méthode

était très peu efficace lorsqu'un grand nombre d'individus n'était pas génotypés.
Lorsque les familles sont recensées sur des critères indépendants de l'histoire familiale, nous avons développé une méthode permettant d'estimer la fonction de pénétrance, que nous avons appelée PEL, et nous avons montré qu'elle restait efficace lorsqu'un grand nombre de génotypes était inconnu. La PEL est une méthode d'estimation paramétrique qui modélise la fonction de pénétrance par un modèle de Weibull. Nous avons choisi d'utiliser un modèle de Weibull étendu dans lequel des paramètres ont été ajoutés afin que le modèle s'adapte plus facilement aux données. L'avantage de cette approche paramétrique est, qu'en faisant une hypothèse de loi sur les données, nous réduisons la dimension du problème à l'estimation d'un nombre fini de paramètres. Si la loi des données est effectivement proche de la loi supposée, même un petit échantillon pourra conduire à de bonnes estimations. En revanche, si la loi des données est très éloignée de la loi supposée a priori, les estimations pourront être fortement biaisées.
De plus, dans les différentes études que nous avons pu trouver dans la littérature, la loi des données n'est jamais testée. Nous avons donc développé une autre méthode d'estimation de la fonction de pénétrance que nous avons appelée IDEAL. La particularité de cette autre méthode est qu'elle est non-paramétrique et ne fait, par conséquent, aucune hypothèse sur la loi des données. Cette approche non-paramétrique est par nature robuste aux erreurs de modèle (model mismatch) mais beaucoup plus "coûteuse" en termes de données (nécessaires à la procédure d'estimation pour garantir une bonne précision) et de temps de calcul.

Afin d'illustrer notre méthode PEL, nous l'avons appliquée à différents échantillons : un échantillon de données de cancer du sein ainsi qu'un échantillon de données NAH provenant de familles françaises et portugaises. Nous avons également analysé des données NAH provenant de familles suédoises.
On peut raisonnablement supposer que la distribution de ces données peut être approchée par une loi de Weibull. Cependant, il serait intéressant d'estimer les différentes fonctions de pénétrance pour ces applications, avec notre méthode IDEAL afin de valider ou de réfuter cette hypothèse.

Dans ce travail, nous nous sommes intéressées uniquement au risque associé à la muta-

tion prédisposante. Mais, outre la mutation prédisposante, le risque peut aussi être modifié par d'autres facteurs, génétiques ou environnementaux. Il serait donc intéressant de pouvoir inclure, notamment dans la GRL ainsi que dans la PEL des covariables environnementales et/ou génétiques afin de mettre en évidence l'effet de ces covariables individuelles sur le risque. Il serait également intéressant de prendre en compte une corrélation familiale qui ne soit due ni à la mutation, ni aux covariables.

Annexe A

La méthode du proposant de Weinberg

Crow [14] a montré que la réplication des familles autant de fois qu'il y a de proposants lorsque la familles contient plusieurs proposants, fournit un estimateur consistant, θ_s, du rapport de ségrégation p (c'est-à-dire que θ_s tend en probabilité vers p). Soit une fratrie de taille s, soit x, la probabilité qu'une famille avec r atteints soit recensée au moins une fois, on a :

$$x = 1 - (1 - \pi)^r,$$

où π est la probabilité de recensement d'un individu.
On remarque qu'ici, les familles sont en fait des fratries.

Soit a, le nombre de proposants dans une famille de r atteints. L'espérance du nombre de proposants a parmi les familles recensées au moins une fois est :

$$\mathbb{E}(a) = \frac{\pi r}{1 - (1 - \pi)^r} = \frac{\pi r}{x}.$$

En notant p, la probabilité pour un individu de la fratrie d'être atteint (i.e. p est la rapport de ségrégation) , l'estimation θ_s de p est donnée par le rapport de l'espérance du nombre d'individus atteints dans une fratrie sélectionnée et répliquée autant de fois qu'il y a de proposant, sur l'espérance du nombre d'individus dans une fratrie sélectionnée et répliquée autant de fois qu'il y a de proposants. θ_s s'écrit donc :

$$\theta_s = \frac{\displaystyle\sum_{r=1}^{s} a(r-1)x \binom{s}{r} p^r q^{s-r}}{\displaystyle\sum_{r=1}^{s} a(s-1)x \binom{s}{r} p^r q^{s-r}}$$

En remplaçant a par son espérance, $\mathbb{E}(a) = \pi r/x$, on a :

$$\begin{aligned}
\theta_s &= \frac{\pi \displaystyle\sum_{r=1}^{s} r(r-1)x \binom{s}{r} p^r q^{s-r}}{\pi(s-1) \displaystyle\sum_{r=1}^{s} \binom{s}{r} p^r q^{s-r}} \\
&= \frac{s(s-1)p^2 \displaystyle\sum_{r=2}^{s} \frac{(s-2)!}{(r-2)!(s-r)!}\ p^{r-2}q^{s-r}}{s(s-1)p \displaystyle\sum_{r=1}^{s} \frac{(s-1)!}{(r-1)!(s-r)!}\ p^{r-1}q^{s-r}} \\
&= \frac{s(s-1)p^2(p+q)^{s-2}}{s(s-1)p(p+q)^{s-1}} \\
&= p \qquad \text{car } p+q=1
\end{aligned}$$

θ_s est donc un estimateur consistant de p puisque Crow a montré que $\theta_s = p$.

Annexe B

Calcul de la fréquence allélique à partir de la proportion d'homozygotes

Considérons une maladie dans laquelle l'allèle délétère A est strictement dominant sur l'allèle normal a. Les individus atteints peuvent donc être homozygotes AA ou hétérozygotes Aa. Soit q la fréquence de l'allèle A et soit P, la proportion d'homozygotes parmi tous les individus atteints dans la population. En supposant Hardy-Weinberg, la fréquence théorique des homozygotes et des hétérozygotes dans la population est respectivement q^2 et $2q(1-q)$. Donc :

$$P = \frac{q^2}{q^2 + 2q(1-q)},$$

et finalement, on obtient :

$$q = \frac{2P}{1+P}.$$

Comme il y a 9 homozygotes parmi les 401 porteurs de la mutation, on a : $P = 9/401$ et $q = \dfrac{2(9/401)}{1 + (9/401)} = 0.04$.

Annexe C

La divergence de Kullback

Pour pouvoir classer les densités de probabilité des données, nous avons besoin d'une mesure sur ces densités de probabilité. Le choix le plus naturel est d'utiliser la divergence de Kullback qui permet de comparer un candidat Q avec la probabilité des données générée P_0 [43]. Pour Q et P, deux densités de probabilité distinctes, la divergence de Kullback est définie de la façon suivante :

$$K(Q,P) = \begin{cases} -\int log(\frac{dQ}{dP})\, dP & \text{si } Q \text{ est absolument continue par rapport à } P \\ +\infty \;\; sinon \end{cases} \tag{C.1}$$

Ceci donne une mesure plausible $K(Q, P_n)$. On remarque donc que si Q n'est pas absolument continue par rapport à P_n ($Q << P_n$), la divergence diverge. Donc, les seules densités de probabilité qui conviennent sont celles qui sont absolument continues par rapport à P_n

Annexe D

Preuve du théorème 5.1.2.1

La difficulté principale dans le calcul de $EL(\theta)$ pour tout θ se situe dans le problème d'optimisation, qui est résolue ici par la méthode du Lagrangien.

Lemme D.0.0.1

Si les hypothèses du théorème 5.1.2.1 sont vérifiées, le supremum qui apparaît dans la définition de $EL(\theta)$ est atteint et, il existe λ^* tel que les poids optimaux sont donnés par :

$$q_i^* = \frac{1}{n}(1 + \lambda^* m(x_i, \theta))^{-1},$$

et $EL(\theta)$ s'écrit donc :

$$EL(\theta) = \max_{\lambda} \left\{ \prod_{i=1}^{n} (n(1 + \lambda\, m(x_i, \theta)))^{-1} \right\}$$

Preuve D.0.0.1

Comme $-log$ est une fonction décroissante, on a :

$$\begin{aligned}
-log(EL(\theta)) &= -log \sup_{(q_1,\ldots,q_n)} \left\{ \prod_{i=1}^{n} q_i | \sum_{i=1}^{n} q_i m(x_i, \theta) = 0, \sum_{i=1}^{n} (q_i - 1/n) = 0 \right\} \\
&= \min_{(q_1,\ldots,q_n)} \left\{ -log \left(\prod_{i=1}^{n} q_i \right) | \sum_{i=1}^{n} q_i m(x_i, \theta) = 0, \sum_{i=1}^{n} (q_i - 1/n) = 0 \right\} \\
&= \min_{(q_1,\ldots,q_n,\lambda,\gamma)} \left\{ -\sum_{i=1}^{n} log(q_i) + n\lambda \sum_{i=1}^{n} q_i m(x_i, \theta) - \gamma \sum_{i=1}^{n} (q_i - 1/n) \right\}.
\end{aligned}$$

En dérivant, on a :

$$-1/q_i^* + n\lambda^* m(x_i, \theta) - \gamma^* = 0$$

En multipliant par q_i^* et en sommant sur i, on obtient $\gamma^* = -n$. On a donc :

$$q_i^* = \frac{1}{n}(1 + \lambda^* m(x_i, \theta))^{-1}$$

Finalement,

$$EL(\theta) = \min_{\lambda} \left\{ 2 \sum_{i=1}^{n} log\left(n(1 + \lambda\, m(x_i, \theta))\right) \right\}$$

Maintenant, on peut démontrer le théorème 5.1.2.1. Dans le cas pécifique où n, la dimension de l'espace d'arrivée de la fonction m est égale à la dimension d du paramètre d'intérêt, le maximum est atteint en θ tel que $\lambda^* = 0$.
Et donc, les poids maximums sont :

$$\hat{q_i^*} = \frac{1}{n}\Big(1 + 0\; m(x_i, \hat{\theta}_{MEL})\Big)^{-1} = \frac{1}{n} \tag{D.1}$$

et donc :

$$\sup_{\theta} EL(\theta) = n^{-n} \tag{D.2}$$

Le rapport de log vraisemblance s'écrit donc :

$$
\begin{aligned}
ELR(\theta) &= -2log\left(\frac{EL(\theta)}{\sup_\theta EL(\theta)}\right) \\
&= -2log\left(n^n \sup_{(q_1,\ldots,q_n)}\left\{\prod_{i=1}^{n} g(\theta,q_1,\ldots,q_i)(x_i)|\mathbb{E}_G[m(x,\theta)]=0\right\}\right) \\
&= 2\min_{(q_1,\ldots,q_n)}\left\{-\sum_{i=1}^{n} log(nq_i)|\mathbb{E}[m(x,\theta)]=0\right\} \\
&= 2n\min_{(q_1,\ldots,q_n)}\left\{-\frac{1}{n}\sum_{i=1}^{n} log\left(\frac{q_i}{1/n}\right)|\mathbb{E}[m(x,\theta)]=0\right\} \\
&= 2n\min_{(q_1,\ldots,q_n)}\left\{-\int log\left(\frac{dG}{dP_n}\right)dP_n|\mathbb{E}[m(x,\theta)]=0\right\} \\
&= C(\theta),
\end{aligned}
$$

Annexe E

Prise en compte de la sélection dans $IDEAL$

Weinberg propose une méthode corrigeant pour la sélection en "se débarrassant" du proposant : Pour une famille de k proposants, la famille est répliquée k fois et à chaque réplication, un proposant différent est retiré. Par exemple, s'il y a 3 proposants p_1, p_2, p_3 dans une famille f_e, cette famille sera répliquée 3 fois en 3 familles $f_{(e,1)}, f_{(e,2)}, f_{(e,3)}$. Dans $f_{(e,1)}$, c'est le proposant p_1 qui sera retiré, les autres seront comptés comme le reste des membres de la famille ; dans la famille $f_{(e,2)}$, c'est le proposant p_2 qui sera retiré et enfin, dans la famille $f_{(e,3)}$, le proposant p_3 sera retiré et les autres (p_1 et p_2) seront comptés comme le reste des membres de la famille.
La validité de cette méthode a été montrée par Crow dans [14] pour estimer le rapport de ségrégation A. Mais elle peut être transposée dans notre contexte d'estimation de pénétrance à partir de données familiales.
Le paramètre θ, qui représente la pénétrance au temps t, est donné par le rapport de l'espérance du nombre d'individus atteints dans une famille sélectionnée et répliquée, sur l'espérance du nombre d'individus porteurs de la mutation dans une famille sélectionnée et répliquée.
Dans ce qui suit, on note G_i et P_i respectivement le génotype et le phénotype de l'individu i.
On a :

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

où r représente le nombre d'individus de la famille, k représente le nombre de proposants et $Pb_i = 1$ si i est un proposant, 0 sinon. L'équation E.1 peut aussi s'écrire de la façon suivante :

$$\mathbb{E}\left[\sum_{i=1}^{r}(\mathbb{1}_{P_i=1} - \theta\mathbb{1}_{G_i=1})(k\mathbb{1}_{Pb_i=0} + (k-1)\mathbb{1}_{Pb_i=1})\right] = 0 \tag{E.2}$$

En divisant par k, on obtient :

$$\mathbb{E}\left[\sum_{i=1}^{r}(\mathbb{1}_{P_i=1} - \theta\mathbb{1}_{G_i=1})\left(\mathbb{1}_{Pb_i=0} + \left(1 - \frac{1}{k}\right)\mathbb{1}_{Pb_i=1}\right)\right] = 0 \tag{E.3}$$

Donc, en définissant W_0 comme suit :

$$W_0(x) = \begin{cases} 1 & \text{si } i \text{ n'est pas un proposant et } x = X_i, \\ 1 - \dfrac{1}{k} & \text{si } i \text{ est un proposant et } x = X_i, \\ 0 & \text{sinon.} \end{cases} \tag{E.4}$$

Alors θ est la solution de l'équation suivante :

$$\mathbb{E}_{W_0}\left[\sum_{i=1}^{r}(\mathbb{1}_{P_i=1} - \theta\mathbb{1}_{G_i=1})\right] = 0. \tag{E.5}$$

L'intégrale de W_0 ne vaut pas 1. Pour avoir une mesure de probabilité, W_0 est donc normalisée. Notre mesure de référence :

$$W(x) = \begin{cases} \dfrac{1}{n} & \text{si } \exists i, x = X_i, Pb_i = 0, G_i = 1 \text{ et } Y_i \geq t, \\ \dfrac{1}{n}(1 - \dfrac{1}{k}) & \text{si } \exists i, x = X_i, Pb_i = 1, G_i = 1 \text{ et } Y_i \geq t, \\ 0 & \text{sinon,} \end{cases} \tag{E.6}$$

converge, après normalisation, vers la version normalisée de W_0. L'estimation de la pénétrance $\hat{\theta}$, solution de : $\mathbb{E}_W\left[\sum_{i=1}^{r}(\mathbb{1}_{P_i=1} - \theta\mathbb{1}_{G_i=1})\right] = 0$, converge vers le paramètre d'intérêt θ. C'est pourquoi nous avons utilisé cette procédure de prise en compte de la sélection dans $IDEAL$.

Montrons maintenant que nous pouvons réécrire l'équation d'estimation comme nous l'avons présentée dans le corps du manuscrit.

Premièrement, sous W, G_i est constant et égal à 1. On peut donc s'en affranchir. Ensuite, pour t fixé, dire qu'un individu i est tombé malade (i.e. $P_i = 1$) est équivalent au fait de dire qu'il est tombé malade avant t (i.e. $A_i \leq t$). Finalement, on obtient donc :

$$\mathbb{E}_W \left[\sum_{i=1}^{r} (\mathbb{1}_{P_i=1} - \theta \mathbb{1}_{G_i=1}) \right] = \mathbb{E}_W \left[\sum_{i=1}^{r} (\mathbb{1}_{A_i \leq t} - \theta) \right]. \tag{E.7}$$

Annexe F

Articles

European Journal of Human Genetics (2007), 1–6

www.nature.com/ejhg

ARTICLE

Estimating cancer risk in HNPCC by the GRL method

Flora Alarcon[1,2], Christine Lasset[3,4], Jérôme Carayol[1,2,5], Valérie Bonadona[3,4], Hervé Perdry[1,2], Françoise Desseigne[3], Qing Wang[3] and Catherine Bonaïti-Pellié*[,1,2]

[1]*INSERM, U535, Villejuif, France;* [2]*Univ Paris-Sud, IFR 69, UMR-S535, Villejuif, France;* [3]*Centre Léon Bérard, Lyon, France;* [4]*Univ Lyon 1; CNRS, UMR5558, Lyon, France*

Hereditary nonpolyposis colorectal cancer (HNPCC) is an autosomal dominant syndrome caused by germline mutations of the mismatch repair (MMR) genes. Only a few studies have taken into account the selection of families tested for these mutations in estimating colorectal cancer (CRC) risk in carriers. They found much lower estimates of CRC risks than previous ones, but these estimates lacked precision despite the large number of families. The aim of this study was to evaluate the efficiency of the 'genotype restricted likelihood' (GRL) method that provides unbiased estimates of risks whatever the ascertainment process of families, and to estimate CRC and endometrial cancer risk for carriers of the MMR genes. Efficiency of the GRL method was evaluated using simulations. Risks were estimated from a sample of 36 families diagnosed with HNPCC and carrying a mutation of MSH2 or MLH1, ascertained through a cancer family clinic in Lyon (France). The efficiency of the GRL method was found to be strongly dependent on the proportion of family members tested. By age 70 years, CRC risk was estimated at 47% (95% confidence interval: 12–98%) for men and 33% (95% confidence interval: 24–54%) for women. The endometrial cancer risk was only 14% (confidence interval: 6–20%). As methods allowing for the selection of families lack efficiency, large-scale family studies should be undertaken and data should be pooled to provide reliable and precise estimates of risks for an optimal familial management.
European Journal of Human Genetics advance online publication, 2 May 2007; doi:10.1038/sj.ejhg.5201843

Keywords: ascertainment; penetrance; pedigree analysis; HNPCC; genetic counselling

Introduction

Hereditary nonpolyposis colorectal cancer (HNPCC) is an autosomal dominant syndrome that predisposes carriers to colorectal and endometrial cancer and cancers of other organs.[1] Mutations of the mismatch repair (MMR) genes (essentially MLH1, MSH2 and MSH6) have been shown to be responsible for a majority of families with this syndrome. Mutations are usually identified in families that fulfil the so-called Amsterdam criteria.[2,3] These criteria include having three close relatives with an HNPCC-associated cancer (of the colon, rectum, endometrium, small bowel, ureter or renal pelvis). If a mutation is identified in one family member (index case), genetic testing is offered to relatives. If they are found to be carriers, they may undergo intensive surveillance, which considerably improves the prognosis of the disease.

Most studies estimate the risk of colorectal cancer in families with HNPCC syndrome selected according to Amsterdam criteria without correcting for selection bias.[4–9] Estimates in these studies range from 0.68 to 0.82, but these values have been shown to be substantially overestimated.[10] Only a few studies have taken the ascertainment process into account,[11–13] and their estimates are lower than those of the other studies. Penetrance values for endometrial cancer range from 0.4 to 0.6.[4,6,8,11]

*Correspondence: Dr C Bonaïti-Pellié, INSERM U535 (Genetic epidemiology and structure of human populations), BP 1000, 94817 Villejuif Cedex, France.
Tel +33 1 45 59 53 49; Fax: +33 1 45 59 53 31;
E-mail: bonaiti@vjf.inserm.fr
[5]Current address: IntegraGen SA, 4 rue Pierre Fontaine, 91058 Evry Cedex, France
Received 12 July 2006; revised 4 April 2007; accepted 4 April 2007

Because the criteria used to select families did not include this tumour, these values should be unbiased.

We proposed an ascertainment-adjusted method for estimating the age-specific cumulative risk (penetrance) of a given disease associated with deleterious mutations in families in which these mutations have been identified.[14] This likelihood, called the 'genotype-restricted likelihood' (GRL), provides unbiased penetrance estimates, regardless of the criteria used to select the families and without modelling the ascertainment process. It also corrects for the bias that is introduced by selection according to genotype and which is inherent in this selection because genotypes are available in relatives only if a mutation is detected in the index case.

In the most recent study of Quehenberger *et al*,[13] endometrial and colorectal cancer risks were estimated for carriers of the MLH1 and MSH2 gene by using a maximum likelihood method that corrected for ascertainment by conditioning on all observed phenotypes, as in the GRL method. They confirmed that previous estimates of colorectal cancer risks were largely overestimated, as colorectal cancer risks by age 70 years were 26.7% for men and 22.4% for women. Despite the large number of families (84), the confidence intervals were quite large, suggesting a lack of efficiency of the method. Indeed, the retrospective likelihoods based on modelling genotypes as a function of given phenotypes are affected by a lack of efficiency.[15] This issue might be particularly crucial in case of missing genotypes, that is, the most usual situation. Using such methods, another question is whether or not to include parts of the pedigree in which the phenotypes of relatives are known but their genotypes are not available.

In this paper, we studied the efficiency of the GRL method according to the proportion of relatives tested in the families and to the amount of family information available for the analysis. We also evaluated this method in a sample of 36 families diagnosed with HNPCC.

Methods

Genotype restricted likelihood

The GRL is a function of observed genotypes (*Gen*), given observed phenotypes (*Phen*), and ascertainment (*Asc*) of families. It can be written as

$$P(Gen/Phen,\, Asc) = \frac{P(Asc/Gen,\, Phen)\, P(Gen/Phen)}{P(Asc/Phen)}$$

Let g denote the genotype of noncarriers of the mutated allele and G that of carriers, Gen_i is the genotype of individual i, $P(Gen_i)$ is the corresponding probability, and $P(Phen_i/Gen_i)$ is the probability of individual i phenotype given his/her genotype. Thus, the contribution of a given family f with s members can be written as (see Carayol and Bonaïti-Pellié[14] for a complete demonstration):

$$P(Gen/Asc,\, Phen) = \frac{\sum_{v\in\Gamma}\prod_{i=1}^{s} P(Phen_i/Gen_{i,v}) \prod_j P(Gen_{j,v}) \prod_{\{l,m,n\}} P(Gen_{l,v}/Gen_{m,v}, Gen_{n,v})}{\sum_{w\in\Omega C}\prod_{i=1}^{s} P(Phen_i/Gen_{i,w}) \prod_j P(Gen_{j,w}) \prod_{\{l,m,n\}} P(Gen_{l,w}/Gen_{m,w}, Gen_{n,w})}$$

where Γ corresponds to the set of genotypic configurations compatible with the genotypes of the individuals tested, Ω_C, to the set of genotypic configurations compatible with the selection criteria (ie, the index case carries the mutation), and $Gen_{i,v}$ and $Gen_{i,w}$, to the genotypes of individual i in genotypic configuration v and w, respectively. The product on j is taken over all individuals whose parents' status is unknown (grandparents and spouses) and the product on $\{l,m,n\}$ over all parent–offspring triplets.

For an individual i with genotype Gen_i, $P(Gen_i)$ is expressed as a function of the frequency of the mutated allele in the general population for a founder, assuming Hardy–Weinberg proportions. Otherwise, this probability depends on parental genotypes, assuming Mendelian transmission.

Finally, let $F_{Gen_i}(t)$ be the penetrance function at age t (cumulative risk by age t). If individual i is unaffected at age t_i, the contribution of i to the likelihood is

$$P(Phen_i/Gen_i) = 1 - F_{Gen_i}(t_i)$$

that is, the probability that individual i is still unaffected at age t_i (survival probability).

If individual i is affected at age t_i, the contribution of i to the likelihood is

$$P(Phen_i/Gen_i) = F_{Gen_i}(t_i + 1) - F_{Gen_i}(t_i)$$

that is, the probability of being affected at age t_i included in the 1-year interval $[t_i\,;t_i+1[$.

For the age-dependent penetrance function according to Gen_i, we chose a Weibull model with parameters λ_{Gen_i} (scale parameter) and a_{Gen_i} (shape parameter). This model is widely used in parametric survival analysis because of its ability to adjust to observed data.

To take into account the possibility that some carriers will never develop the disease, we introduced a third parameter, κ_{Gen_i}, corresponding to the fraction of individuals who will never be affected.[16,17] Finally, the penetrance function may be written as

$$F_{Gen_i} = (1 - \kappa_{Gen_i}).(1 - exp(-\lambda_{Gen_i}\, t^{a_{Gen_i}}))$$

Simulation of family data

We used simulations to study the efficiency of the GRL in cases where some family members had unknown genotypes. As in a previous paper,[14] samples of three-generation families with at least two affected members were simulated, with various penetrance values. The simulated pedigrees had a fixed structure: a couple of ancestors with four offsprings and their spouses, each with four offsprings. We simulated the genotypes of family members according to

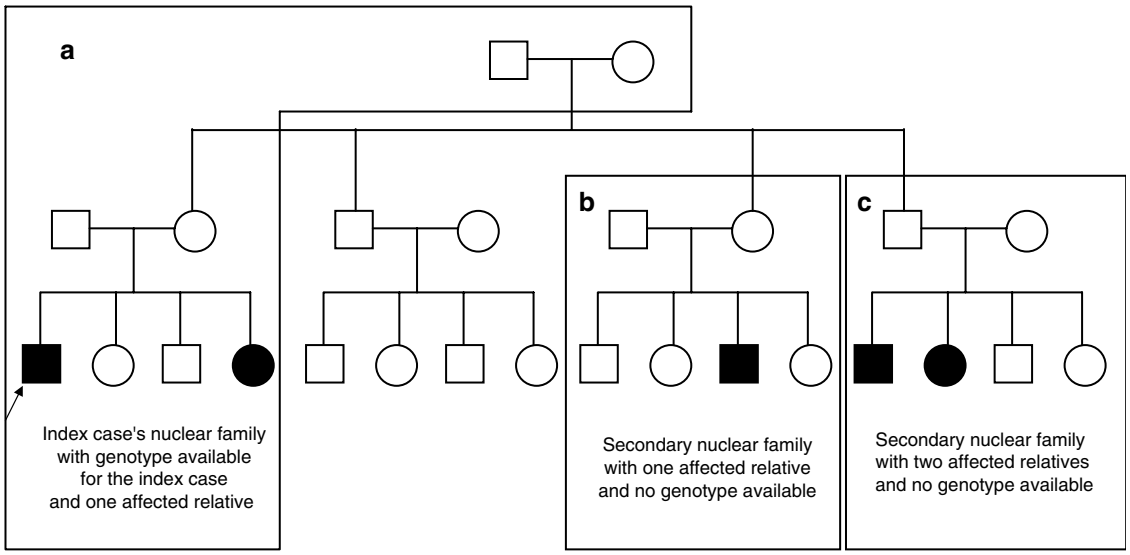


Figure 1 Pedigree structure with (**a**) index case's ancestors and nuclear family and (**b,c**) secondary nuclear families (index case marked with an arrow).

Mendel's laws for subjects whose parents were in the pedigree, ignoring the possibility of *de novo* mutation and according to the frequency of the mutated allele for founders. To obtain samples of sufficient size with at least one carrier individual (the index case), without simulating too many families, this frequency was set at 0.10. Phenotypes were simulated according to the age-dependent penetrance function, with the Weibull model. For noncarriers ($Gen = g$), the parameter κ_g was set at 0 and parameters λ_g and a_g at values corresponding to a cumulative risk of 0.02 by age 80. For carriers ($Gen = G$), we considered two different risk values, the first one corresponding to a cumulative risk of 0.2 (called 'low true penetrance') by age 80 and the second one to 0.5 (called 'high true penetrance') by the same age. We did not consider any gender differences in risks.

The families were selected if at least two members were affected. To keep sample fluctuations to a minimum, sample size was fixed to 10 000 families after selection.

The loss (or gain) of efficiency was investigated by computing asymptotic relative efficiencies (AREs) of penetrance estimates, that is, the inverse of the ratio of the variance estimate in a given situation to the variance obtained in a reference situation. To evaluate the variance of the penetrance, we simulated, in each situation, 1000 replicates of the family sample and computed the variance of the estimate by age 70.

The efficiency of the GRL according to the proportion of genotyped individuals in families was studied by comparing the variance of the cumulative hazard functions calculated with varying proportions of genotyped individuals (25, 50 and 75%) to the variance computed when all genotypes are known. We also considered the most extreme situation, where only two genotypes are known (the index case and one relative). Note that if only the index case is genotyped, Γ and Ω_c are identical, and the likelihood is a constant.

To study the information provided by family branches with no genotypic data, we selected families in which the index case's nuclear family included an affected relative tested for the mutation, and members of the secondary nuclear families of the third generation were not tested. We then compared the variance of the cumulative hazard function in four different situations, according to whether the sample included for each family (Figure 1): (1) only the ancestors and members of the index case's nuclear family (pedigree A); (2) pedigree A + members of secondary nuclear families with at least one affected (ie, family types B and C); (3) pedigree A + members of secondary nuclear families with at least two affected (family type C); and (4) all family members.

Parameters of the penetrance function were estimated by maximising the likelihood of simulated samples. We wrote a program that includes the maximisation procedure GEMINI as a subroutine[18] and provides maximum likelihood estimates of the parameters λ_G and a_G for carriers. Because κ_G was set at 0 in the simulation process, we did not estimate this parameter. We assumed that the penetrance was known for noncarriers and the three parameters were set at the same values as in the family simulation process.

HNPCC families

The index cases investigated in this study are patients referred by their physicians or self-referred for genetic counselling at the Centre Leon Bérard in Lyon (France)

from January 1994 to January 2004. MMR testing was offered when they fulfilled the Amsterdam criteria I, which include only colorectal tumours,[2] or II, which include extracolonic tumours associated with the syndrome,[3] or even less stringent criteria, when one of the classic criteria was missing. All the individuals included in this study signed an informed consent for genetic testing. As this study did not involve any additional intervention, it was exempt under French law from ethical review board approval. Blood samples were subjected to germline mutation screening of *MLH1* (NM_000249 for cDNA and NC_000003 for genomic DNA) and *MSH2* (NM_000251 for cDNA and NC_000002 for genomic DNA) genes using genomic DNA sequencing.[19] Of the 161 index cases meeting one of the selection criteria, 42 were found to carry a deleterious mutation of MLH1 or MSH2. Five families were not informative because none of the index case's relatives underwent mutation testing, and were therefore excluded. Another family was excluded because numerous consanguineous loops made the program unfeasible. Among the 36 mutated informative families, 22 index cases (61.1%) carried a mutation of MLH1 and 14 (38.9%) a mutation of MSH2**.** Genetic testing identified 129 mutation carriers (51 men and 78 women) and 59 noncarriers. Clinical information was available for 1185 family members (577 men and 608 women), 216 of whom were affected by colorectal cancer (97 men and 94 women). Age at diagnosis ranged from 20 to 89 years in men (mean: 44 years) and 18 to 82 years in women (mean: 45 years). Endometrial cancer was reported in 30 women. The youngest woman was diagnosed at age 32 years and the oldest one at age 88. Other tumours associated with the syndrome were observed (of the ovary, urinary tract, stomach and small intestine), but there were too few cases to allow estimation of penetrance. The cancer diagnosis was confirmed by medical and pathological reports in the great majority of affected relatives (85%).

The GRL method was used to estimate the parameters of the penetrance function. For each family member, the age t was taken as the age at last news or age of death if unaffected and the age at first diagnosis of colorectal cancer or endometrial cancer if affected. We assumed a frequency of 10^{-3} for the mutated allele and a *de novo* mutation frequency of 10^{-5}, after verifying that estimates of penetrance were not sensitive to errors in these values. Parameters for noncarriers were fixed at values that fit their incidence in the French population.[20] Maximum likelihood was used to estimate the three parameters of the penetrance function: λ_G, a_G and κ_G. Analyses were conducted separately for men and women.

Confidence intervals were calculated with the bootstrap method. One thousand samples were constructed by resampling the 36 HNPCC families, and the penetrance function was estimated for each new sample. We used the 2.5 and 97.5 percentiles of the distribution of estimated penetrance at different ages to determine the corresponding lower and upper bounds of the confidence interval of the risk for each cancer.

Results

Efficiency

As shown in Table 1, efficiency decreased with the percentage of relatives tested, whatever the penetrance value. This reduction was particularly marked when only one family member besides the index case was tested in which case efficiency fell to 7%.

Whatever the penetrance value, the information provided by family branches without genotypic data did not increase efficiency of penetrance estimate with an ARE of about 1.00 in all cases. This clearly indicates that the inclusion of family branches in the analysis provides no significant information when genotypes are not available.

We could check that, whatever the proportion of missing genotypes and the family branches included in the analysis, penetrance estimates using the GRL were unbiased.

Estimation of cancer risk in HNPCC

Figure 2 summarizes the penetrance functions of colorectal cancer estimated with the GRL from the 36 HNPCC families. Penetrance was negligible before 30 years. Although some cases of colorectal cancer were diagnosed before this age, most were index cases, which do not contribute to the likelihood. Penetrance was found to be higher in men than in women, with estimates of 0.47 and 0.34, respectively, at 70 years. Confidence intervals were rather large: 0.12–0.98 for men and 0.24–0.54 for women.

Estimated penetrance for endometrial cancer was very low before 40 years, because only three women developed this tumour at an earlier age, two of them being index cases (Figure 3). The cumulative risk at 70 years was estimated to be 0.14, with a confidence interval of 0.06–0.20.

Table 1 Efficiency of the GRL for estimating penetrance function according to the proportion of relatives tested for the mutation

	Asymptotic relative efficiency	
Situation	*Low true penetrance*	*High true penetrance*
Proportion of relatives tested		
●100%	Reference	Reference
●75%	0.99	0.88
●50%	0.74	0.70
●25%	0.47	0.41
●Minimal[a]	0.07	0.07

Abbreviation: GRL, genotype restricted likelihood.
[a]The most extreme situation, when only two genotypes are known (the index case and one relative).

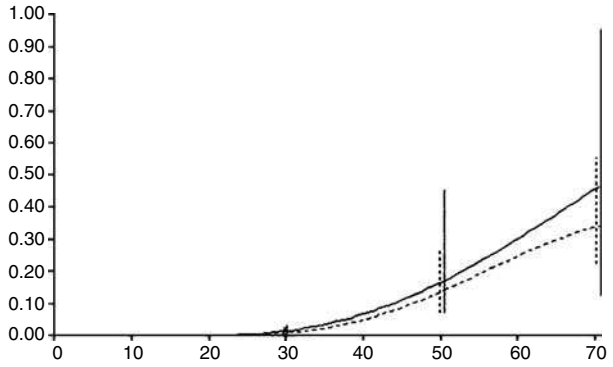


Figure 2 Penetrance function and confidence intervals at 30, 50 and 70 years of colorectal cancer risk in MSH2 and MLH1 mutation carriers for men (solid line) and women (dotted line), estimated from the 36 HNPCC families.

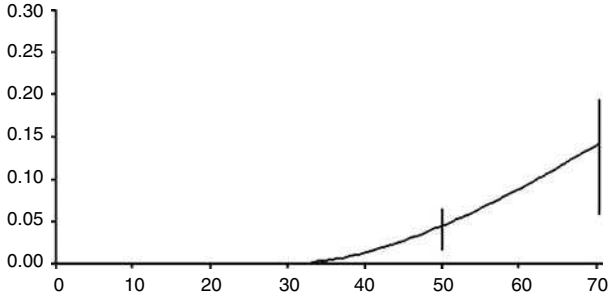


Figure 3 Penetrance function and confidence intervals at 30, 50 and 70 years of endometrial cancer risk in MSH2 and MLH1 mutation carriers, estimated from the 36 HNPCC families.

Discussion

The results reported here show that efficiency may be problematic when only a few individuals are tested. The proportion of relatives undergoing genetic testing in families with such a genetic mutation and associated disease appears quite low, despite the benefits of molecular screening and endoscopic surveillance. For example, in 32 Italian families with germline mutations of MSH2, MLH1 or MSH6, only 34% of the first-degree relatives of affected individuals underwent genetic testing.[21] In this study, only 24% of the 292 first-degree relatives of the 36 index cases were tested. This proportion may increase in the future, as families come to understand better the benefits of genetic testing.

We applied the GRL to estimate the risks of colorectal and endometrial cancer in families with HNPCC syndrome selected by familial criteria and identification of a MSH2 or MLH1 gene mutation. Lifetime penetrance of colorectal cancer was estimated at 47% for men and 33% for women. These risks were considerably lower than the first estimates reported in the literature, and were consistent with the values determined by studies taking into account the ascertainment bias.[11–13] Dunlop *et al*[11] selected subjects as a function of age at diagnosis of the index case (at or below 35 years of age) and presence of microsatellite instability (MSI) in the patient's tumour; MSI is characteristic of tumours due to MMR mutations, that is, independent of family history. They obtained risk estimates of 52% for colorectal cancer (CRC), and 42 % for uterine cancer by the age of 70 years. Parc *et al*[12] analysed data from families of patients referred to a cancer family clinic and satisfying at least one of the modified Amsterdam criteria.[3] To avoid ascertainment bias, they used a statistic based on the proportion of carriers among unaffected individuals, which allowed an estimation of the overall cancer risk (but not separate estimations for specific types). They obtained risk estimates of 43% by age 38 and 62% by age 51. Neither study provided confidence intervals but these intervals were probably large due to the small number of families in the first study and the relatively young ages of unaffected individuals tested for the mutation in the second one. Quehenberger *et al*[13] used a method based on the same principles as ours in that they conditioned the likelihood of the observed genotypes on the observed phenotypes and on the event that at least one cancer patient was a mutation carrier. We could expect that our estimates would be very close to theirs. Indeed, there was only a slight difference in that we found a higher risk of CRC and a smaller risk of endometrial cancer. However, because of the large confidence intervals in the two studies and of the absence of difference found by Quehenberger *et al*,[13] the penetrance values were not estimated separately for MLH1 and MSH2. Our results, combined to those of the three studies described above, confirm that most studies have overestimated the risks of colorectal cancer in HNPCC syndrome. Regarding the risk of endometrial cancer, we found a much lower estimate than previous studies but our results were not strongly different from those of Quehenberger *et al*,[13] who found a risk of 31.5% (confidence interval: 11.1–70.3%). In our study, the upper bound of the confidence interval was 20%, which enables us to conclude that previous studies might have overestimated this risk, probably because endometrial cancer, although not 'officially' included in the recommended criteria, has been known to be associated with the syndrome for a long time, and this factor might have played a role in referring patients from physicians to oncogeneticists.

A considerable advantage of the GRL, as well as other retrospective methods, is that it is valid regardless of the inclusion criteria. It can thus be applied to samples of families selected according to different criteria. This property should be used in the future to pool large amounts of data of HNPCC families from different studies, to obtain reliable and precise estimates of risks. This would also permit us to estimate the risk of other HNPCC-associated cancers, scarcely known at present, and help organising the management of families and the surveillance of carrier relatives. Such a study is presently ongoing

in France. It aims at collecting data from all the families tested for MLH1, MSH2 and MSH6 mutations. It will also allow us to detect a possible genetic heterogeneity among families according to the mutation involved, and to test for the role of other familial factors, either genetic or not, that could influence cancer risk in carriers.

Currently, carriers of MMR mutations in HNPCC families frequently undergo early colonoscopic screening at the age of 20 or 25 years. This should be considered when defining uninformative censoring events for unaffected relatives. Observation time was censored at age of first colonoscopy in the study by Quehenberger *et al*[13] and at current age in the present one. These procedures could lead, in the first case, to shorter observation times for most individuals and, in the second case, to overlooking removal of precancerous lesions such as polyps. However, the clinical events observed during colorectal surveillance should be taken into account. The age at first diagnosis of an adenomatous polyp or the age at last colonoscopy in the absence of polyp detection should be more appropriate censoring times as more complete surveillance information is used to define the observation times. This could increase the power of the studies and the accuracy of the estimations of cancer risks.

Acknowledgements

We acknowledge with gratitude the support of the Fondation de France and the Ligue Nationale Contre le Cancer through Jérôme Carayol's fellowship and the support of the Ligue Contre le Cancer du Rhône et de l'Ardèche for the management of the HNPCC families database. We thank Catherine Huber for helpful advice and Marie Dominique Reynaud for editing this paper.

References

1 Lynch HT, de la Chapelle A: Genetic susceptibility to non-polyposis colorectal cancer. *J Med Genet* 1999; **36**: 801–818.
2 Vasen HF, Mecklin JP, Khan PM, Lynch HT: The International Collaborative Group on Hereditary Non-Polyposis Colorectal Cancer ICG-HNPCC). *Dis Colon Rectum* 1991; **34**: 424–425.
3 Vasen HF, Watson P, Mecklin JP, Lynch HT: New clinical criteria for hereditary nonpolyposis colorectal cancer (HNPCC, Lynch syndrome) proposed by the International Collaborative Group on HNPCC. *Gastroenterology* 1999; **116**: 1453–1456.
4 Aarnio M, Mecklin JP, Aaltonen LA, Nyström-Lahti M, Järvinen HJ: Life-time risk of different cancers in hereditary non-polyposis colorectal cancer (HNPCC) syndrome. *Int J Cancer (Pred Oncol)* 1995; **64**: 430–433.
5 Voskuil DW, Vasen HF, Kampman E, Van't Veer P, the National Collaborative Group on HNPCC: Colorectal cancer risk in HNPCC families: development during lifetime and in successive generations. *Int J Cancer* 1997; **72**: 205–209.
6 Vasen HFA, Wijnen JT, Menko FH *et al*: Cancer risk in families with hereditary nonpolyposis colorectal cancer diagnosed by mutation analysis. *Gastroentology* 1996; **110**: 1020–1027.
7 Lin KM, Shashidharan M, Thorson AG *et al*: Cumulative incidence of colorectal and extracolonic cancers in MLH1 and MSH2 mutation carriers of hereditary nonpolyposis colorectal cancer. *J Gastrointest Surg* 1979; **2**: 67–71.
8 Aarnio M, Sankila R, Pukkala E *et al*: Cancer risk in mutation carriers of DNA-mismatch-repair genes. *Int J Cancer* 1999; **81**: 214–218.
9 Vasen HF, Stormorken A, Menko FH *et al*: MSH2 mutation carriers are at higher risk of cancer than MLH1 mutation carriers: a study of hereditary nonpolyposis colorectal cancer families. *J Clin Oncol* 2001; **19**: 4074–4080.
10 Carayol J, Khlat M, Maccario J, Bonaïti-Pellié C: Hereditary non-polyposis colorectal cancer: current risks of colorectal cancer largely overestimated. *J Med Genet* 2002; **39**: 335–339.
11 Dunlop MG, Farrington SM, Carothers AD *et al*: Cancer risk associated with germline DNA mismatch repair gene mutations. *Hum Mol Genet* 1997; **6**: 105–110.
12 Parc Y, Boisson C, Thomas G, Olschwang S: Cancer risk in 348 French MSH2 or MLH1 gene carriers. *J Med Genet* 2003; **40**: 208–213.
13 Quehenberger F, Vasen HFA, van Houwelingen HC: Risk of colorectal and endometrial cancer for carriers of mutations of the hMLH1 and hMSH2 gene: correction for ascertainment. *J Med Genet* 2005; **42**: 491–496.
14 Carayol J, Bonaïti-Pellié C: Estimating penetrance from family data using a retrospective likelihood when ascertainment depends on genotype and age of onset. *Genet Epidemiol* 2004; **27**: 109–117.
15 Kraft P, Thomas DC: Bias and efficiency in family-based gene-characterization studies: Conditional, Prospective, Retrospective, and joint likelihoods. *Am J Hum Genet* 2000; **66**: 1119–1131.
16 Sposto R: Cure model analysis in cancer: an application to data from the Children's Cancer Group. *Stat Med* 2002; **21**: 293–312.
17 Planté-Bordeneuve V, Carayol J, Ferreira A *et al*: Genetic study of transthyretin amyloid neuropathies: carrier risks among French and Portuguese families. *J Med Genet* 2003; **40**: e120.
18 Lalouel JM: *GEMINI – a computer program for optimization of general non linear functions*. Technical report no 14. Salt Lake City: university of Utah, Department of Medical Biophysics and Computing, 1979.
19 Wang Q, Lasset C, Desseigne F *et al*: Prevalence of germline mutations of hMLH1, hMSH2, hPMS1, hPMS2 and hMSH6 genes in 75 French kindreds with nonpolyposis colorectal cancer. *Hum Genet* 1999; **105**: 79–85.
20 Remontet L, Esteve J, Bouvier AM *et al*: Cancer incidence and mortality in France over the period 1978–2000. *Rev Epidemiol Sante Publique* 2003; **51**: 3–30.
21 Ponz de Leon M, Benatti P, Di Gregorio C *et al*: Genetic testing among high-risk individuals in families with hereditary non-polyposis colorectal cancer. *Br J Cancer* 2004; **90**: 882–887.

PEL: An Unbiased Method for Estimating Age-Dependent Genetic Disease Risk From Pedigree Data Unselected for Family History

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

[1]*University Paris-Sud, Villejuif, France*
[2]*INSERM, Villejuif, France*
[3]*Genetic Oncology, Institut Curie, Paris, France*
[4]*Department of Neurology, CHU Bicêtre, Le Kremlin Bicêtre, France*
[5]*University Paris-Descartes, Paris, France*

Providing valid risk estimates of a genetic disease with variable age of onset is a major challenge for prevention strategies. When data are obtained from pedigrees ascertained through affected individuals, an adjustment for ascertainment bias is necessary. This article focuses on ascertainment through at least one affected and presents an estimation method based on maximum likelihood, called the Proband's phenotype exclusion likelihood or PEL for estimating age-dependent penetrance using disease status and genotypic information of family members in pedigrees unselected for family history. We studied the properties of the PEL and compared with another method, the prospective likelihood, in terms of bias and efficiency in risk estimate. For that purpose, family samples were simulated under various disease risk models and under various ascertainment patterns. We showed that, whatever the genetic model and the ascertainment scheme, the PEL provided unbiased estimates, whereas the prospective likelihood exhibited some bias in a number of situations. As an illustration, we estimated the disease risk for transthyretin amyloid neuropathy from a French sample and a Portuguese sample and for *BRCA1/2* associated breast cancer from a sample ascertained on early-onset breast cancer cases. *Genet. Epidemiol.* 2008.

Key words: ascertainment bias; risk estimation; penetrance function; maximum likelihood method

*Correspondence to: F. Alarcon, INSERM U535, BP 1000, F-94817 Villejuif, France. E-mail: flora.alarcon@inserm.fr
PEL, Proband's phenotype Exclusion Likelihood; MD, monogenic diseases.
Received 20 October 2008; Accepted 21 October 2008
Published online in Wiley InterScience (www.interscience.wiley.com).
DOI: 10.1002/gepi.20390

INTRODUCTION

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

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

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

Other methods have been proposed for estimating penetrance, in a different context. The kin-cohort design [Wacholder et al., 1998] has been proposed to estimate penetrance for carriers of specific mutations of the *BRCA1* and *BRCA2* genes by comparing the proportions of affected relatives between mutation carriers and noncarriers in a population in which the specific mutations are frequent. This design is more generally referred to as the genotype-proband design (GPD) by Gail et al. [1999] to emphasize that the proband only was genotyped and extended to the so-called GPDR where one or two relatives are genotyped in each nuclear family. These population-based methods require very large samples of affected and unaffected individuals and do not need any correction for ascertainment.

The genotype restricted likelihood (GRL) has been proposed for analyzing pedigrees ascertained through familial criteria and for which a formal correction for ascertainment bias cannot easily be performed.

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

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

METHODS

THE PEL

The PEL is a maximum likelihood (ML) method that corrects for ascertainment bias when families, in which a deleterious mutation has been found, have been ascertained through at least one affected individual, i.e. regardless of family history. The penetrance function is estimated using the phenotypic information, conditioned on genotype, from all family members, including those with unknown genotype. Correction for ascertainment is performed by removing the phenotypic information of the individual who allowed the family to be detected (the proband) and by duplicating families when there are several probands in a family.

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

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

$\mathrm{Phen} = (\mathrm{Phen}_1, ..., \mathrm{Phen}_{n_f})$ with $\mathrm{Phen}_i = 1$ if i is affected and $\mathrm{Phen}_i = 0$ if i is unaffected, $\mathrm{Gen}_i = 0$ if i is not a carrier and $\mathrm{Gen}_i = 1$ otherwise.

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

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

$$L_f = P(\mathrm{Phen}^*/\mathrm{Gen}_{\mathrm{obs}}).$$

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

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

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

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

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

$$L_f = \sum_{\omega=1}^{\Omega} P_\omega \cdot P(\mathrm{Phen}^*/\mathrm{Gen}_\omega),$$

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

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

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

PENETRANCE FUNCTION AND CONTRIBUTION OF FAMILY MEMBERS TO THE LIKELIHOOD

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

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

$$P(\mathrm{Phen}_i/\mathrm{Gen}_i) = 1 - F(t_i).$$

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

$$P(\mathrm{Phen}_i/\mathrm{Gen}_i) = F(t_i + 1) - F(t_i).$$

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

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

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

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

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

In our computations, $(\kappa, \lambda, \alpha)$ were estimated by ML using the program GEMINI [Lalouel, 1979].

PROPERTIES OF THE PEL

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

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

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

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

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

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

Bias and efficiency. To study the behavior of the PEL according to the ascertainment scheme and the model considered, we first evaluated, in each case, the average relative bias B estimated by the average on 1,000 replicates of 100 pedigrees of the relative bias of risk estimate at age 70 years usually taken as the lifetime risk [Alarcon et al., 2007; Easton et al., 1995; Ford et al., 1998; Gong and Whittemore, 2003]:

$$B = \frac{1}{1,000} \sum_{i=1}^{1,000} \left(\frac{\hat{R}_i - R_0}{R_0} \right),$$

where $\hat{R}_i$ is the penetrance estimated at 70 years for the replicate i and R_0 is the true one at the same age.

To study the loss in efficiency due to unknown genotypes, and therefore, the interest of genotyping probands' relatives, we evaluated the asymptotic relative efficiencies (AREs) of penetrance estimates, that is, the inverse of the ratio of the variance estimated with various values of the proportion of unknown genotypes to the variance estimated when all genotypes are known. In each situation, the variance by age 70 was evaluated by simulating 1,000 replicates of 100 families.

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

We compared the properties of the PEL with those of the prospective likelihood in terms of bias and efficiency in various situations.

RESULTS

BEHAVIOR OF THE PEL ACCORDING TO THE ASCERTAINMENT SCHEME AND THE GENETIC MODEL

Table I presents the relative bias obtained with the PEL and, for comparison, the relative bias obtained with the

TABLE I. Relative bias in penetrance estimate at 70 years using the PEL and the prospective likelihood

Genetic model[a]			Relative bias (%)			
			High true penetrance		Low true penetrance	
Method used	P_s[a]	Time period T (in years)	PEL	Prospective likelihood	PEL	Prospective likelihood
MD	1	20	2	0	2	−3
		1	2	3	3	2
	0.5	20	2	1	3	0
		1	2	1	3	−2
CDMS	1	20	2	5	3	14
		1	2	9	4	21
	0.5	20	2	5	3	14
		1	3	10	3	21

[a]MD, Monogenic disease; CDMS, complex disease with monogenic sub-entities.
P_s, probability of being ascertained for individuals affected during the period T. PEL, Proband's phenotype Exclusion Likelihood.

TABLE II. Relative efficiency of the PEL according to the proportion of unknown genotypes under the MD model

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

[a]Situation in which only the proband's genotype is known.
PEL, Proband's phenotype Exclusion Likelihood; MD, monogenic diseases.

prospective likelihood when all genotypes are known, under the two different models considered MD and CDMS with an age criterion of 36 years as described above, and the two time periods considered. We have checked that results were similar when various proportions of unknown genotypes were introduced (results not shown).

Under both models, the PEL provided unbiased or nearly so estimates in all the situations considered.

The prospective likelihood provided less satisfactory results, particularly under the CDMS model.

Table II shows the relative efficiencies (AREs) under the MD model obtained when the proportion of unknown genotypes among relatives varies from 50 to 100%, compared to the case where all genotypes are known, in the two extreme situations of ascertainment, i.e. high ($P_s = 1$, $T = 20$), and low ($P_s = 0.5$, $T = 1$) ascertainment probability. In all situations, the loss in efficiency is quite small, whatever the true penetrance value. Under the CDMS model, the effect is quite similar, although slightly more important, with AREs varying from 56 to 69% according to the ascertainment probability and the true penetrance value, in the extreme situation where only the proband's genotype is known. Therefore, genotyping the relatives brings some additional information on the risk estimate, compared to the situation where only their phenotype is known, but the gain in efficiency is expected to be modest.

Regarding the prospective likelihood efficiency, we evaluated AREs only under the MD model where the prospective likelihood provided unbiased estimated. The prospective method provided similar efficiency as the PEL in a wide range of situations, varying from 0.5 to 1.9 according to the ascertainment probability, with the highest relative efficiency when ascertainment probability was high (the reference here is the estimation with the PEL). This method appeared as robust as the PEL to a high proportion of unknown genotypes, with AREs varying from 59 to 100% according to the ascertainment model and the proportion of unknown genotypes (results not shown).

APPLICATION

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

TTR amyloid neuropathy. *TTR* amyloid neuropathy is an autosomal dominant condition characterized by deposition of amyloid substance made up of mutated *TTR*. This severe condition, firstly described in Portugal, involves mainly the peripheral nervous system and the heart. Although distributed worldwide, the disease is often clustered in limited areas like in Portugal, Japan and Sweden with different genotypic and phenotypic variation including the age of first symptoms. In France, we are dealing with two populations, i.e. of Portuguese and of French origins. Virtually all the families are referred to the department of Neurology of Bicêtre Hospital which is the national center of reference for this rare disease. Many pathogenic *TTR* variants have been detected among the French population, but only one variant, the Val30Met, was detected in the Portuguese population. In a previous article, we had analyzed a sample of 79 families (46 French

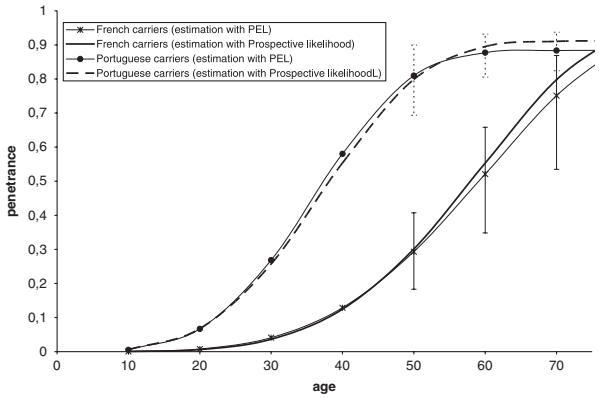


Fig. 1. Penetrance of the TTR mutation involved in the transthyretin amyloid neuropathy estimated using the PEL and the prospective likelihood in the French and in the Portuguese family sample.

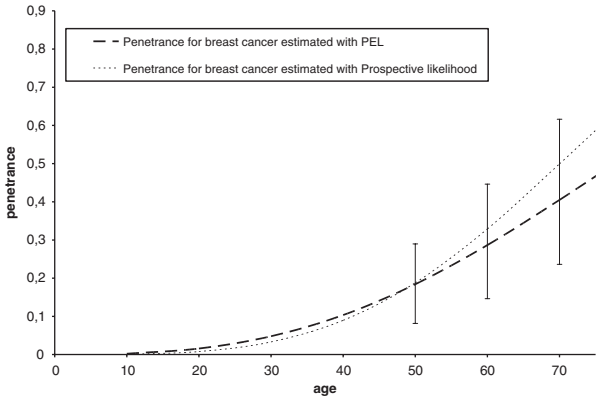


Fig. 2. Penetrance of BRCA1/2 carriers in the French breast cancer families.

and 33 Portuguese) investigated in the neurology department of Bicêtre Hospital and found a strong difference in penetrance function between these two populations [Plante-Bordeneuve et al., 2003]. In the present article, we restricted the analysis to Val30Met carriers (20 French and 33 Portuguese kindreds). *TTR* genotype was available for 108 and 139 relatives in French and Portuguese families of whom, respectively, 47 and 50 were carriers. The proportions of unknown genotypes among relatives were respectively 72 and 68%. The deleterious allele frequency was arbitrarily set to 0.001, and the de novo mutation was set to 0 in both analyses.

Figure 1 shows the penetrance functions estimated with the PEL in the French and Portuguese families. Confidence intervals (obtained by bootstraps) are given for ages 50, 60 and 70. The plateau in Portuguese estimation shows the existence of a proportion of carriers which will never be affected (i.e. $\kappa = 0.09$, significantly different from zero with a $P_{\text{value}} < 0.001$) which illustrates the importance of implementing κ in the Weibull model. In both samples, the penetrance curve estimated by the prospective likelihood was very close the one obtained with the PEL, which illustrates that the choice of the method is not crucial in such a situation.

Breast cancer due to BRCA1 or BRCA2 mutations. Families had been selected through 317 women suffering from invasive breast cancer, diagnosed before 36 years between January 1990 and January 1998, and followed up at the Institut Curie. Genetic counselling was proposed to all women and 153 of them came to the appointment to the Institut Curie cancer clinic. A *BRCA1* and *BRCA2* and *TP53* genetic screening was systematically proposed whatever the family history, and 145 of them underwent genetic testing [Chompret et al., 2001]. The entire coding sequence of both genes was analyzed by a combination of DGGE, DHPLC and PTT [Stoppa-Lyonnet et al., 1997; Wagner et al., 2000]. Sixteen and 14 patients with, respectively, a germline *BRCA1* or *BRCA2* mutation were identified. In these families, genetic testing was proposed to relatives as recommended by the French guidelines [Eisinger et al., 1998]. Among the 30 families, genotype was available for 33 relatives of whom 17 were found to be carriers. In 16 families, the proband was the only one individual tested in the family as no relative asked for genetic testing, and the overall proportion of unknown genotypes among relatives was 91%.

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

The cumulative risk for noncarriers was taken as the risk in the general population.

Figure 2 shows the estimations of the penetrance function using the PEL and the prospective likelihood. The sample being relatively small, families with BRCA1 (16 families) and BRCA2 (14 families) mutations were pooled. Confidence intervals were estimated for ages 50, 60 and 70 years. The two methods provide different penetrance curves, with smaller risks using the PEL, which illustrates that the two methods are not equivalent under such a model. However, the difference is not statistically significant, probably because of the small sample size.

DISCUSSION

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

Using simulations under various genetic models and various ascertainment schemes, the PEL, based on a principle of a classical method proposed in segregation analysis [Weinberg, 1912], turned out to generally provide very satisfactory results. This method has the advantage of being very simple and of leading to unbiased penetrance estimations in all the situations considered, and in particularly in situations where the prospective likelihood provided biased estimates.

We showed that efficiency of the PEL was moderately affected when a substantial proportion of genotypes was unknown under both models, and performed well even when a limited number of relatives have been genotyped.

Various situations may affect the estimation of penetrance by the PEL. The method assumes that probands are

always unambiguously identified, but this may not be the case under multiple ascertainment. However, we could check (results non shown) that method was robust to a misspecification of probands in a wide range of situations.

Another possible source of bias, when some genotypes are missing among relatives, may be a misspecification of the de novo mutation rate or of the deleterious allele frequency that are commonly fixed to arbitrary low values. We have checked that the PEL was quite robust to an error on these two parameters.

The other methods that have been mentioned in the introduction for estimating penetrance apply to a completely different context. The kin-cohort design [Wacholder et al., 1998], and more generally the GPD and the GPDR [Gail et al., 1999] cannot be applied to the same samples as the PEL. Indeed, the GPD and GPDR are population-based designs which require very large samples of affected and unaffected individuals and do not need any correction for ascertainment. Moreover, these methods have been designed for common mutations in common diseases, and can be applied neither to the MD model nor to the general situations of mutations predisposing to common cancers in which each mutation is very rare and carriers are difficult to identify at the population level.

The GRL has been developed for estimating penetrance from pedigrees ascertained on familial criteria, and corrects for ascertainment by conditioning the likelihood on all phenotypes of pedigree members [Carayol and Bonaiti-Pellie, 2004]. This method has the advantage of being unbiased, whatever the selection criteria, but as the retrospective likelihood [Kraft and Thomas, 2000] has the drawback of lacking efficiency, particularly when there are numerous unknown genotypes in the family [Alarcon et al., 2007]. Therefore, the GRL should be restricted to samples of families ascertained on multiple affected individuals in which ascertainment correction cannot easily be performed.

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

For breast cancer families, the inclusion of an age criterion implies a very low ascertainment probability, in which case the PEL is expected to provide a lower relative bias than the prospective likelihood. Thus, the results obtained by the PEL are more likely to be the correct estimates. We must however, keep in mind that not all families fulfilling the inclusion criterion were investigated and that the patients who underwent genetic testing might have been motivated by a stronger family history. Such a potential bias is difficult to correct but should be considered in the interpretation of results. Note that the risks obtained by the PEL are close to those obtained by Bonadona et al. [2005] from a population-based series of early-onset breast cancer cases (age at diagnosis less than 45), and slightly smaller than those obtained by Antoniou et al. [2003] from 22 studies unselected for family history, 10 of which were ascertained on early-onset breast cancer cases (limit 36–50 years), although risks estimated in these studies are not strictly comparable to our estimations because we pooled *BRCA1* and *BRCA2* mutations.

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

ACKNOWLEDGMENTS

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

REFERENCES

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

Alarcon F, Lasset C, Carayol J, Bonadona V, Perdry H, Desseigne F, Wang Q, Bonaiti-Pellie C. 2007. Estimating cancer risk in HNPCC by the GRL method. Eur J Hum Genet 15:831–836.

Antoniou A, Pharoah PD, Narod S, Risch HA, Eyfjord JE, Hopper JL, Loman N, Olsson H, Johannsson O, Borg A, et al. 2003. Average risks of breast and ovarian cancer associated with BRCA1 or BRCA2 mutations detected in case Series unselected for family history: a combined analysis of 22 studies. Am J Hum Genet 72:1117–1130.

Bonadona V, Sinilnikova OM, Chopin S, Antoniou AC, Mignotte H, Mathevet P, Bremond A, Martin A, Bobin JY, Romestaing P, Randrant D, Rudigoz RC, Léone M, Chauvin F, Easton DF, Lenoir GM., Lasset C. 2005. Contribution of BRCA1 and BRCA2 germline mutations to the incidence of breast cancer in young women: results from a prospective population-based study in France. Genes Chromosomes Cancer 43:404–413.

Cannings C, Thompson E. 1977. Ascertainment in the sequential sampling of pedigrees. Clin Genet 12:208–212.

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

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

Chompret A, Abel A, Stoppa-Lyonnet D, Brugieres L, Pages S, Feunteun J, Bonaiti-Pellie C. 2001. Sensitivity and predictive value of criteria for p53 germline mutation screening. J Med Genet 38:43–47.

Crow J. 1965. Problems of ascertainment in the analysis of family data. In: Neel JV, Shaw MW, Schull WJ, editors. Genetics and the Epidemiology of Chronic Disease. Washington DC: Public Health Source Publication.

Dunlop MG, Farrington SM, Carothers AD, Wyllie AH, Sharp L, Burn J, Liu B, Kinzler KW, Vogelstein B. 1997. Cancer risk associated with germline DNA mismatch repair gene mutations. Hum Mol Genet 6:105–110.

Easton DF, Ford D, Bishop DT. 1995. Breast and ovarian cancer incidence in BRCA1-mutation carriers. Breast Cancer Linkage Consortium. Am J Hum Genet 56:265–271.

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

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

Fisher R. 1934. The effects of methods ascertainment upon the estimation of frequencies. Ann Eugen 6:13–25.

Ford D, Easton DF, Stratton M, Narod S, Goldgar D, Devilee P, Bishop DT, Weber B, Lenoir G, Chang-Claude J et al. 1998. Genetic heterogeneity and penetrance analysis of the BRCA1 and BRCA2 genes in breast cancer families. The Breast Cancer Linkage Consortium. Am J Hum Genet 62:676–689.

Gail MH, Pee D, Benichou J, Carroll R. 1999. Designing studies to estimate the penetrance of an identified autosomal dominant mutation: cohort, case–control, and genotyped-proband designs. Genet Epidemiol 16:15–39.

Gong G, Whittemore AS. 2003. Optimal designs for estimating penetrance of rare mutations of a disease-susceptibility gene. Genet Epidemiol 24:173–180.

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

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

Le Bihan C, Moutou C, Brugieres L, Feunteun J, Bonaiti-Pellie C. 1995. ARCAD: a method for estimating age-dependent disease risk associated with mutation carrier status from family data. Genet Epidemiol 12:13–25.

Morton NE. 1959. Genetic tests under incomplete ascertainment. Am J Hum Genet 11:1–16.

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

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

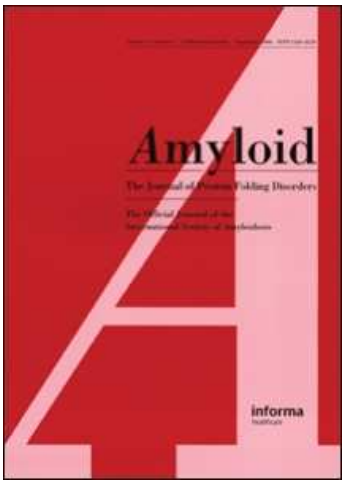
Stoppa-Lyonnet D, Laurent-Puig P, Essioux L, Pages S, Ithier G, Ligot L, Fourquet A, Salmon RJ, Clough KB, Pouillart P et al. 1997. BRCA1 sequence variations in 160 individuals referred to a breast/ovarian family cancer clinic. Institut Curie Breast Cancer Group. Am J Hum Genet 60:1021–1030.

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

Wacholder S, Hartge P, Struewing JP, Pee D, McAdams M, Brody L, Tucker M. 1998. The kin-cohort study for estimating penetrance. Am J Epidemiol 148:623–630.

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

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



Amyloid
Publication details, including instructions for authors and subscription information:
http://www.informaworld.com/smpp/title~content=t713668970

Heterogeneity of penetrance in familial amyloid polyneuropathy, ATTR Val30Met, in the Swedish population

Urban Hellman [a]; Flora Alarcon [bc]; Hans-Erik Lundgren [d]; Ole B. Suhr [d]; Catherine Bonaiti-PelliÉ [bc]; Violaine Planté-Bordeneuve [ef]
[a] Medical and Clinical Genetics, Umeå University, Umeå, Sweden [b] INSERM, U535, Villejuif, France [c] University Paris-Sud, Villejuif, France [d] Department of Internal Medicine, Umeå University, Umeå, Sweden [e] Department of Neurology, CHU Bicêtre, Le Kremlin Bicêtre, France [f] INSERM U488, Le Kremlin Bicêtre, France

Online Publication Date: 01 January 2008

To cite this Article Hellman, Urban, Alarcon, Flora, Lundgren, Hans-Erik, Suhr, Ole B., Bonaiti-PelliÉ, Catherine and Planté-Bordeneuve, Violaine(2008)'Heterogeneity of penetrance in familial amyloid polyneuropathy, ATTR Val30Met, in the Swedish population',Amyloid,15:3,181 — 186
To link to this Article: DOI: 10.1080/13506120802193720
URL: http://dx.doi.org/10.1080/13506120802193720

Amyloid, September 2008; 15(3): 181–186

informa
healthcare

Heterogeneity of penetrance in familial amyloid polyneuropathy, ATTR Val30Met, in the Swedish population

URBAN HELLMAN[1], FLORA ALARCON[2,3], HANS-ERIK LUNDGREN[4], OLE B. SUHR[4], CATHERINE BONAITI-PELLIÉ[2,3], & VIOLAINE PLANTÉ-BORDENEUVE[5,6]

[1]*Medical and Clinical Genetics, Umeå University, Umeå, Sweden,* [2]*INSERM, U535, Villejuif, France,* [3]*University Paris-Sud, IFR 69, Villejuif, France,* [4]*Department of Internal Medicine, Umeå University, Umeå, Sweden,* [5]*Department of Neurology, CHU Bicêtre, Le Kremlin Bicêtre, France, and* [6]*INSERM U488, Le Kremlin Bicêtre, France*

Keywords: *Transthyretin, amyloidosis, genetics, penetrance, Swedish families*

Abbreviations: *FAP = familial amyloidotic polyneuropathy; TTR = transthyretin*

Abstract

Transthyretin (TTR) familial amyloid polyneuropathies (FAP) are autosomal dominant devastating afflictions. They were first described in Portugal, later in Japan and Sweden and are now recognized worldwide. The TTR Val30Met mutation is the most common, and depending on the geographic origin, a wide variation in age at onset of the disease is observed. In Europe, northern Sweden is the second most prevalent area of the disease, and a late age of onset of 56 years has been reported. The present study aims to estimate the penetrance in TTR Val30Met Swedish families. Genealogical investigations, clinical data and genotyping were obtained in 77 TTR-Val30Met Swedish families. The penetrance in Val30Met carriers and variation within the endemic area, according to gender and transmitting parents were calculated by a newly developed bias-free method. The penetrance estimates were low, i.e. 1.7% and 22% at age 30 and 60 years, respectively, and far from complete (69%) by age 90 years. Differences between Piteå and Skellefteå regions were observed. Moreover, penetrance was significantly higher when the mutation was inherited from the mother than from the father. The low penetrance observed in TTR FAP kindreds and its variations is important information for the genetic counseling and treatment of Swedish FAP patients and their families.

Introduction

Transthyretin (TTR) amyloid neuropathy is the most severe neuropathy of adulthood with autosomal dominant transmission. The disease is characterized by extracellular deposition of amyloid composed of mutated TTR, of which the most common amyloidogenic mutation is the ATTR Val30Met. Its main clinical expression is a progressive sensory–motor neuropathy with autonomic manifestations frequently associated with cardiac dysfunction. Other manifestations include ocular symptoms related to vitreous amyloidosis and eventually renal failure. A fatal outcome occurs after an average duration of 10–13 years [1,2]. By removing the main source of mutated TTR, liver transplantation is the only therapeutic approach available, which can stabilize the disease when performed early in the course [3,4].

The condition was initially reported by Andrade (1952) in northern Portugal [5] in the area of Povoa de Varzim where the prevalence is 1 in 1000, and was subsequently recognized throughout the world including various European countries and two different foci in Japan [6,7]. Apart from Portugal, the most prevalent occurrence of the disease in Europe is in the north of Sweden where families are mainly clustered in limited areas around the towns of Skellefteå and Piteå [8].

Among 100 point mutations identified in the TTR gene, the Val30Met is by far the most common pathogenic variant and virtually the only one described in Portugal and Sweden [9]. Genotype–phenotype correlations are unclear. Moreover, phenotypic variations and striking differences in the age of onset are observed between populations of Val30Met patients. Hence, the mean age of first symptoms is 33 years in Portugal as well as in the main endemic areas in Japan [7,10], whereas in Sweden the onset is rather late at 56 years of age [11]. Screenings have not disclosed any protective variant of the TTR gene that could explain the older age of onset in Sweden.

Correspondence: Urban Hellman, Department of Medicine, Umeå University, 901 87 Umeå, Sweden. Tel: + 46 (0) 90 785 1621. Fax: + 46 (0) 90 130760. E-mail: urban.hellman@medbio.umu.se

ISSN 1350-6129 print/ISSN 1744-2818 online
DOI: 10.1080/13506120802193720

In Sweden, a population study suggested that disease risk in carrier individuals could be rather low [11]. The age of onset is variable among pedigrees and is significantly higher with paternal inheritance compared with that of maternal inheritance [12]. This finding needs to be confirmed by estimating the penetrance (cumulative risk of being affected by a given age) in mutation carriers, i.e. the risk that a carrier individual be affected given he/she has reached a given age.

Using a method to estimate penetrance that corrects for ascertainment bias toward affected individuals, significant differences in Portuguese and French families were noted [13].

The aim of the study was to assess the penetrance of Swedish TTR Val30Met, using a bias-free method, and to disclose differences in penetrance within the endemic area. We also estimated the impact of transmitting parents' gender on age at onset of the disease.

Patients and methods

Patients

Among the 401 patients treated at the Department of Clinical Genetics at Umeå University Hospital since 1986 and identified with a Val30Met mutation, a complete genealogical investigation was available for 122 families. As some families were related to each other pedigrees were merged, thus resulting in 85 pedigrees that could enter the analysis. Eight pedigrees contained nine homozygous patients for the Val30Met mutation, and were excluded from the analysis as these patients were too few to allow a reliable estimation of risk associated with this genotype. Finally, 77 pedigrees were available for penetrance analysis.

In patients, the diagnosis of TTR amyloid neuropathy was based on the finding of amyloid deposits in biopsy specimens; however the deposits were not always immunohistochemically diagnosed as TTR amyloid. The pathogenic TTR variant Val30Met mutation was identified by molecular genetic testing in all cases. Genealogical investigation was carried out for all patients since 1999, when possible, and data were entered using Cyrillic software (version 2.0). A detailed history of the disease was taken from the patient and/or from medical records. The date of first symptoms was determined as accurately as possible.

For the present study, a total of 1353 subjects were recorded and clinical information was obtained on 235 affected individuals, including the probands. Overall, DNA analysis was performed in 215 individuals of the families. Among them 49 were asymptomatic carriers and 19 were non-carriers. In addition, 35 asymptomatic individuals were found to be obligate carriers.

Genetic analysis

Over time different methods have been used to detect the TTR Val30Met mutation. Southern blots were first used to analyze the patients, but later patients were diagnosed by polymerase chain reaction (PCR)-based methods as previously described [11,14].

Estimation of penetrance

Statistical method. We used the same method as Planté-Bordeneuve *et al.* [13] which allows a correction for ascertainment bias when using family data. The method is based on the maximum likelihood principle, using a survival analysis approach. In addition to the correction for ascertainment, its main interest is to use phenotypic information on individuals whose genotype is unknown but whose probability of being a carrier is not negligible.

Briefly, the age-dependent penetrance function for mutation carriers was modeled by a Weibull distribution, defined by three parameters: λ (scale parameter), a (shape parameter) and κ (the fraction of individuals that would never be affected) [16]. The penetrance function can be expressed:

$$F(t_{\mathrm{m}}) = (1 - \kappa).(1 - exp\ [-(\lambda t_m^a)])$$

Following survival analysis principles, the contribution of pedigree members to the likelihood depends on, respectively, the age at diagnosis for affected and the age of unaffected relatives or their age at death.

The probability of unknown genotypes of untested individuals given observed genotypes in the families was computed from the frequency of the mutated allele in the general population assuming Hardy–Weinberg equilibrium for a founder (parents' status unknown). Otherwise, this probability depends on parental genotypes assuming Mendelian transmission. In this analysis, disease allele frequency was estimated using the proportion of homozygotes among all affected individuals in the population, as explained in the Appendix.

For the sake of simplicity, the probability of being homozygous was neglected for the reasons given above.

To test for a difference according to gender, geographic origin (Skellefteå, Piteå and Lycksele) or to the sex of the transmitting parent, we used a likelihood ratio test.

Family data used. All family data providing information on the risk of being affected for a relative

were retained. Thus, at-risk individuals or patients with reliable information on their date of birth, their age of onset or their date or age of death, if adapted, were included. In a minority of affected cases, the age of onset was extrapolated from the age of death, using the mean duration of the disease of 13 years in the Swedish population [2]. The branches of the families without information on the individual's phenotype or date of birth or on the age at first manifestations were excluded.

On the basis of the known age of onset in previous reports of TTR amyloid neuropathy in Sweden, individuals younger than 18 years were not considered in the present study [8,11,12].

Ethics

The project was approved by the Ethics Committee at Umea University.

Results

Using the proportion of homozygotes for the Val30Met mutation among affected patients, the frequency of this mutation was estimated to be 0.04 (9/401) in the population and was fixed at this value in all subsequent analyses.

Figure 1 shows the penetrance curves for the whole sample and for males and females separately. Gender effect was not significant ($\chi^2 = 2.44$, 2 df). In all of the analyses performed on the different samples, there was no evidence for a plateau; the parameter κ converged to its bound 0, so that only the two parameters a and λ were estimated. The risks and confidence intervals from age 30 to 90 years are shown in Table I. At age 30 and 60 years, the penetrance estimates were 1.7% and 22%, respectively. Even by age 90 years, penetrance was found to be far from complete (69%).

When subdividing the sample according to the three regions, Lycksele was too small to allow a valid estimate (10 families), thus, only Piteå and Skellefteå were compared. As shown in Figure 2, the penetrance was significantly lower for the Piteå region than for the Skellefteå region, with penetrance by age 80 of 32% and 65%, respectively. The homogeneity test was highly significant ($\chi^2 = 26.60$, 2 df, $p < 0.001$).

A clear difference according to the sex of the transmitting parent was found (Figure 3). Among the 77 families, the gender of the transmitting parent was known for 762 individuals; for 435 the transmitting parent was the mother, and for 327 the transmitting parent was the father. The penetrance was significantly higher when the mutation was inherited from the mother than from the father ($\chi^2 = 7.84$, 2 df, $p < 0.02$). Table II shows the penetrance estimates and confidence intervals according to the sex of the transmitting parent.

Discussion

The ATTR Val30Met mutation is by far the most frequent neuropathic form worldwide, and virtually the only one identified in endemic areas in Portugal, Japan, Sweden and Brazil. However, important differences in the disease expression, particularly regarding the age of onset and the penetrance seem to exist between countries. Significant differences between Portuguese and French populations have been noted [13], thus, cumulative disease risk in French carriers of the Val30Met mutation was estimated at 14% by age 50 and 50% by age 70 years, whereas the risk in Portuguese carriers was

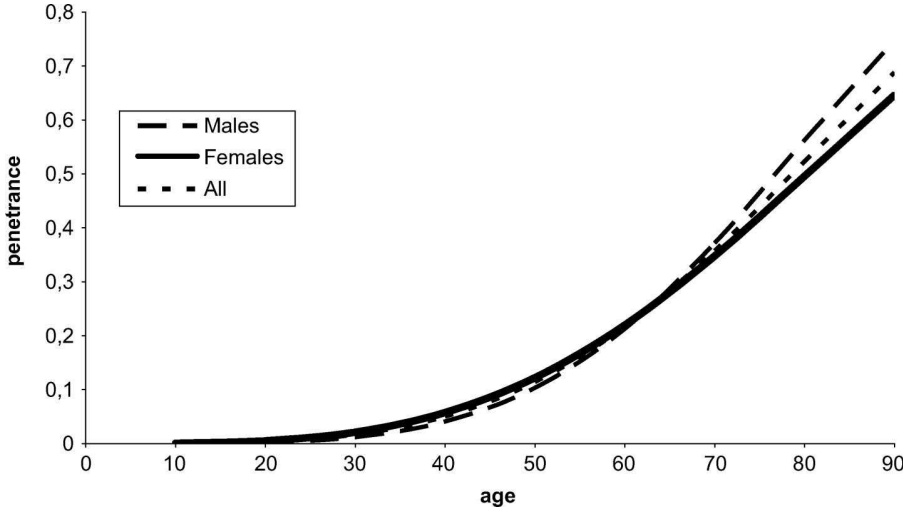


Figure 1. Estimated penetrance curve according to gender.

80% by age 50 and 91% by age 70 years. Swedish carriers appear to be more similar (although lower) to French than Portuguese carriers with a risk of 11% by age 50 and 36% by age 70 years. The penetrance increases with age in Sweden and does not display any plateau; a similar finding was noted in the French population, whereas a plateau was noted for the Portuguese. These results are consistent with the late age at onset averaging 56 years as previously reported in the Swedish population [11]. The reasons for the increase of penetrance with age can only be speculated on. The aging process involves a decreased ability to handle oxidative stress and also increases in the deposition of advanced glycation end products, of which both factors have been implicated in amyloid formation [17,18].

Table I. Estimation of penetrance in the 77 Swedish families (235 affected among 1353 family members).

Age (years)	Penetrance estimate	95% Confidence interval
30	0.017	0.008–0.032
40	0.05	0.029–0.082
50	0.11	0.08–0.16
60	0.22	0.16–0.29
70	0.36	0.28–0.45
80	0.52	0.42–0.63
90	0.69	0.55–0.79

The patients were all recruited from the Department of Medicine, Umeå University Hospital. This is a referral center for familial amyloid polyneuropathy (FAP) in Sweden. However, patients with the onset at an older age are generally not submitted for

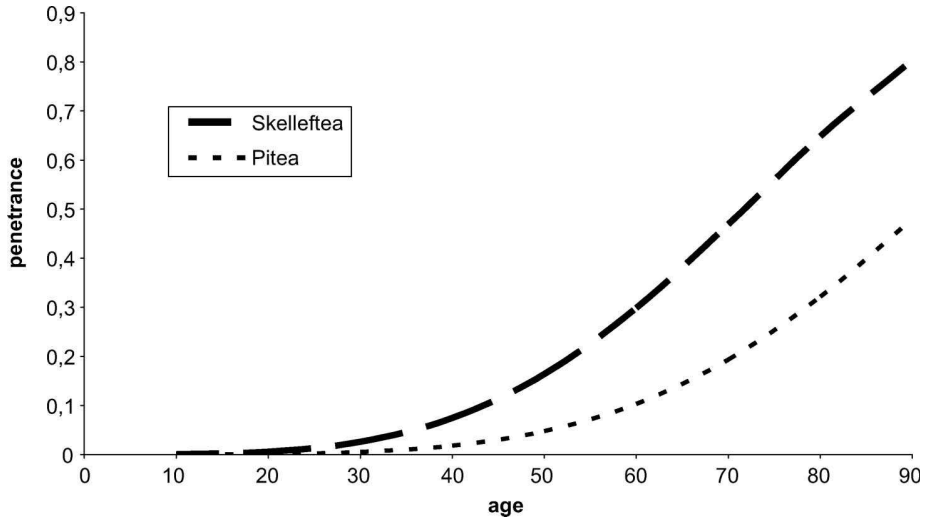


Figure 2. Estimated penetrance curve according to region.

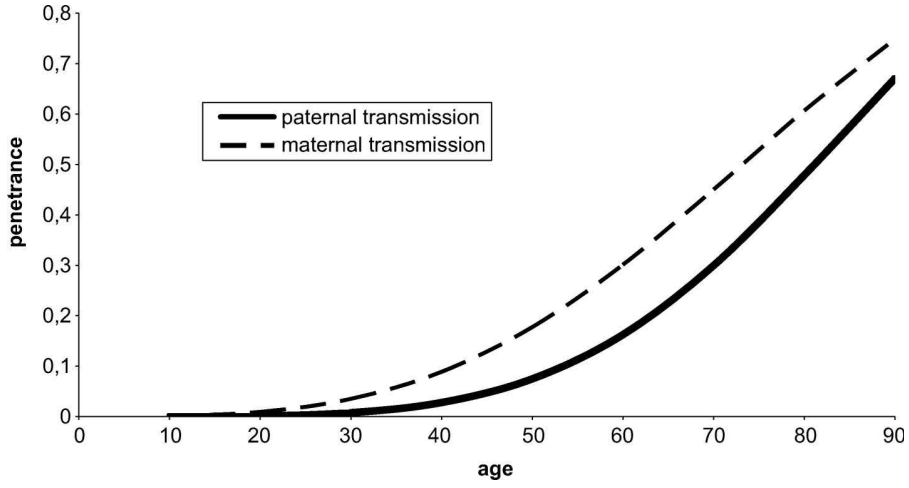


Figure 3. Estimated penetrance curve according to sex of the transmitting parent.

Table II. Penetrance estimates according to the sex of the transmitting parent.

Age (years)	Penetrance (95% confidence interval), when transmission by	
	Father (327)	Mother (435)
30	0.008 (0.003–0.016)	0.035 (0.013–0.059)
40	0.028 (0.016–0.041)	0.089 (0.043–0.14)
50	0.075 (0.052–0.11)	0.18 (0.10–0.27)
60	0.16 (0.13–0.25)	0.30 (0.21–0.44)
70	0.30 (0.21–0.44)	0.45 (0.34–0.61)
80	0.48 (0.30–0.66)	0.61 (0.47–0.76)
90	0.67 (0.40–0.85)	0.75 (0.57–0.89)

evaluation, especially not from hospitals with long experience in the disease (such as Skellefteå Hospital). It is therefore clear that the patient material has a bias against younger patients, and to some degree patients with rapidly progressing disease. Moreover, marked differences in penetrance of the trait were noted within the endemic area. Such differences might be due to genetic factors (other than the TTR mutation) but also to environmental factors. In particular, it is noteworthy that Skellefteå is an area with different industries from those of Piteå and Lycksele. In Skellefteå, a large metallurgic industry and previously also textile manufactures are located, and the latter turned out to be a risk factor for developing the disease in a study of occupation and disease in northern Sweden [19]. Discordant penetrance of FAP in two pairs of monozygotic twins have been reported [20]. These observations along with the observed variation of penetrance could support the role of environmental factors to explain differences in expression of the mutation.

The other remarkable finding is the marked differences of penetrance according to the gender of the transmitting parent. Indeed, the penetrance of the trait was significantly higher when inherited from the mother than from the father. These data are in line with a previous genealogical analysis of Swedish pedigrees underlying that the age of onset was significantly higher when the mutation was inherited from the father and that anticipation (higher penetrance in younger generations) was more marked in descendants of affected mothers [12]. Differences of age of onset according to the gender of the transmitting parent were also mentioned in Japanese and Portuguese families but such variation of penetrance has not been precisely analyzed in other populations [10,21]. It is tempting to link these differences to a parental imprinting phenomenon. In this setting, the occurrence of regions containing differential allele specific DNA methylation localized near the *TTR* gene might be interesting to explore. However, there is no evidence of other imprinted genes in the 18q11.2-12 region, to date (www.geneimprint.com). Another clue could be a possible role of the mitochondrial genome interacting with the expression or the pathogenesis of the TTR Val30Met mutation.

The penetrance estimates we found are higher than those previously reported. However, the previously reported frequency of symptomatic carriers at 2% was estimated from a study of healthy volunteers and the known number of affected individuals from the endemic area [11]. The difference between the penetrance estimates cannot be due to an ascertainment bias, since our method takes into account that families without any affected individuals will not be ascertained. Neither can it be attributed to the high gene frequency in the Swedish population, in comparison to other populations such as the Portuguese [10], as the method has been shown to be robust to a variation in this parameter [13]. Differences in reported prevalence could in part be due to an underestimation of affected individuals as probably is the case in the previous Swedish study [11], as it did not adjust for age. Indeed, the affected individuals were mostly older than 60 years, whereas those in the reference population were much younger.

In summary, different values of penetrance according to the geographic origin of the families within the endemic area were noted and substantiated the occurrence of anticipation related to female gender of the transmitting parent. The rather low penetrance observed in Swedish TTR FAP kindreds and its variations is important information for the genetic counseling and treatment of Swedish FAP patients and their families.

Declaration of interest: The authors report no conflicts of interest. The authors alone are responsible for the content and writing of the paper.

References

1. Plante-Bordeneuve V, Said G. Transthyretin related familial amyloid polyneuropathy. Curr Opin Neurol 2000;13:569–573.
2. Suhr O, Danielsson A, Holmgren G, Steen L. Malnutrition and gastrointestinal dysfunction as prognostic factors for survival in familial amyloidotic polyneuropathy. J Intern Med 1994;235:479–485.
3. Suhr OB, Holmgren G, Steen L, Wikstrom L, Norden G, Friman S, Duraj FF, Groth CG, Ericzon BG. Liver transplantation in familial amyloidotic polyneuropathy. Follow-up of the first 20 Swedish patients. Transplantation 1995;60:933–938.
4. Adams D, Samuel D, Goulon-Goeau C, Nakazato M, Costa PM, Feray C, Plante V, Ducot B, Ichai P, Lacroix C, Metral S, Bismuth H, Said G. The course and prognostic factors of familial amyloid polyneuropathy after liver transplantation. Brain 2000;123:1495–1504.

5. Andrade C. A peculiar form of peripheral neuropathy. Familiar atypical generalized amyloidosis with special involvement of the peripheral nerves. Brain 1952;75:408–427.
6. Reilly MM, Adams D, Booth DR, Davis MB, Said G, Laubriat-Bianchin M, Pepys MB, Thomas PK, Harding AE. Transthyretin gene analysis in European patients with suspected familial amyloid polyneuropathy. Brain 1995;118: 849–856.
7. Ikeda S, Nakazato M, Ando Y, Sobue G. Familial transthyretin-type amyloid polyneuropathy in Japan: clinical and genetic heterogeneity. Neurology 2002;58:1001–1007.
8. Sousa A, Andersson R, Drugge U, Holmgren G, Sandgren O. Familial amyloidotic polyneuropathy in Sweden: geographical distribution, age of onset, and prevalence. Hum Hered 1993;43:288–294.
9. Saraiva MJ. Transthyretin mutations in hyperthyroxinemia and amyloid diseases. Hum Mutat 2001;17:493–503.
10. Sousa A, Coelho T, Barros J, Sequeiros J. Genetic epidemiology of familial amyloidotic polyneuropathy (FAP)-type I in Povoa do Varzim and Vila do Conde (north of Portugal). Am J Med Genet 1995;60:512–521.
11. Holmgren G, Costa PM, Andersson C, Asplund K, Steen L, Beckman L, Nylander PO, Teixeira A, Saraiva MJ, Costa PP. Geographical distribution of TTR met30 carriers in northern Sweden: discrepancy between carrier frequency and prevalence rate. J Med Genet 1994;31:351–354.
12. Drugge U, Andersson R, Chizari F, Danielsson M, Holmgren G, Sandgren O, Sousa A. Familial amyloidotic polyneuropathy in Sweden: a pedigree analysis. J Med Genet 1993;30:388–392.
13. Plante-Bordeneuve V, Carayol J, Ferreira A, Adams D, Clerget-Darpoux F, Misrahi M, Said G, Bonaiti-Pellie C. Genetic study of transthyretin amyloid neuropathies: carrier risks among French and Portuguese families. J Med Genet 2003;40:e120.
14. Holmgren G, Bergstrom S, Drugge U, Lundgren E, Nording-Sikstrom C, Sandgren O, Steen L. Homozygosity for the transthyretin-Met30-gene in seven individuals with familial amyloidosis with polyneuropathy detected by restriction enzyme analysis of amplified genomic DNA sequences. Clin Genet 1992;41:39–41.
15. Fuchs U, Zittermann A, Suhr O, Holmgren G, Tenderich G, Minami K, Koerfer R. Heart transplantation in a patient with senile cardiomyopathy. Am J Transplant 2005;5:1159–1162.
16. Sposto R. Cure model analysis in cancer: an application to data from the Children's Cancer Group. Stat Med 2002;21:293–312.
17. Ando Y, Nyhlin N, Suhr O, Holmgren G, Uchida K, El Sahly M, Yamashita T, Terasaki H, Nakamura M, Uchino M, Ando M. Oxidative stress is found in amyloid deposits in systemic amyloidosis. Biochem Biophys Res Commun 1997;232:497–502.
18. Nyhlin N, Ando Y, Nagai R, Suhr O, El Sahly M, Terazaki H, Yamashita T, Ando M, Horiuchi S. Advanced glycation end product in familial amyloidotic polyneuropathy (FAP). J Intern Med 2000;247:485–492.
19. Hardell L, Holmgren G, Steen L, Fredrikson M, Axelson O. Occupational and other risk factors for clinically overt familial amyloid polyneuropathy. Epidemiology 1995;6:598–601.
20. Holmgren G, Wikstrom L, Lundgren HE, Suhr OB. Discordant penetrance of the trait for familial amyloidotic polyneuropathy in two pairs of monozygotic twins. J Intern Med 2004;256:453–456.
21. Yamamoto K, Ikeda S, Hanyu N, Takeda S, Yanagisawa N. A pedigree analysis with minimised ascertainment bias shows anticipation in Met30-transthyretin related familial amyloid polyneuropathy. J Med Genet 1998;35:23–30.

Appendix

Computation of allele frequency from the proportion of homozygous individuals in a dominantly inherited disease

Let us consider a disease where the deleterious allele A is strictly dominant on the normal allele a (affected individuals may be either homozygous AA or heterozygous Aa).

Let q be the frequency of the A allele and P the proportion of homozygotes among all affected individuals in the population.

Assuming that Hardy–Weinberg proportions are valid, the theoretical frequency of homozygotes and heterozygotes in the population are respectively q^2 and $2q(1-q)$.

Thus,

$$P = q^2/[q^2 + 2q(1-q)]$$

And finally,

$$q = 2P/(1+P)$$

Parent-of-origin effect in Transthyretin related amyloid polyneuropathy

B. Bonaïti[1,3], F. Alarcon[2,3,] C. Bonaïti-Pellié[3,2], V. Planté-Bordeneuve[2,4]

[1]INRA-SGQA, Jouy-en-Josas, France, [2]Univ Paris-Sud, IFR 69, Villejuif, France;

[3]INSERM, U535, Villejuif, France; [4]Department of Neurology, CHU Henri Mondor, Créteil, France.

Corresponding author
Dr V. Planté-Bordeneuve
Department of Neurology
CHU Henri Mondor
46 Av du M[al] Delattre de Tassigny
94000 Créteil
France
Tel : +00 33 (0)1 49 81 43 12
violaine.plante@hmn.aphp.fr / vplante@free.fr

Transthyretin (TTR) amyloid neuropathy is the most severe neuropathy of adulthood with autosomal dominant transmission [1]. Liver transplantation is the only therapeutic approach available, which can stabilize the disease when performed early in the course. Important differences in the disease expression, particularly regarding the age of onset and the penetrance, exist among countries. In a previous work [2], we found significant differences between Portuguese and French populations with cumulative risks of 18% by age 50 and 64% by age 70 years in French carriers, whereas this risk in Portuguese carriers was 80% by age 50 and 91% by age 70 years. Recently, Hellman and colleagues found that Swedish carriers seemed to be closer to French than to Portuguese carriers with an even lower risk of 11% by age 50 and 36% by age 70 years [3]. These results are consistent with the late age at onset, previously reported in the Swedish population [4].

The other remarkable finding of the study of Hellman et al. [3] is the marked differences of penetrance according to gender of the transmitting parent. Indeed, the risk of disease in carriers was significantly higher when the mutation was inherited from the mother than from the father. Differences of age of onset according to gender of the transmitting parent have also been mentioned in Portuguese families but such variation of penetrance has not been precisely analysed in other populations [5].

We investigated this effect in our samples of French and Portuguese families, using a modified version of the maximum likelihood method PEL [6], that takes into account the gender of the transmitting parent. In this aim, the genotype of heterozygous carrier individuals was ordered according to the parent (father or mother) who transmitted the mutation. In the 33 Portuguese families, the gender of the transmitting parent was known for 291 individuals: the mother in 153 and the father in 138 cases. Figure 1 shows the penetrance curves for French and Portuguese mutation carriers according to gender of the parent who transmitted the mutation. As in the Swedish families, the penetrance was found to be higher when the mutation was inherited from the mother (98% by age 50 and 99% by age 70) than from the father (60% by age 50 and 80% by age 70) and the difference was

highly significant (χ^2 = 37.6, 3 df, p<0.001). Among the 48 French families, there were 435 individuals for whom the gender of the transmitting parent was known: for 216 individuals the transmitting parent was the mother, and for 219 individuals the transmitting parent was the father.

Such a parent-of-origin effect on penetrance seems to be present in all population samples despite the heterogeneity of penetrance among these populations. Indeed, this result was highly significant in the Portuguese families alike those recently found in the Swedish kindred [3]. The reason why the French sample does not reach significance is probably due to the lack of power for detecting an effect when penetrance is low and to sample size. Indeed, penetrance estimates are much lower in the French sample than in the Portuguese sample, and the number of families is substantially smaller than in the Swedish sample. The mechanism underlying such maternal effect on penetrance remains to be elucidated. One might speculate on a genetically determined effect through an imprinting phenomenon. Interestingly, the human homologue *IMPACT* of the mouse *Impact* gene, which is an imprinted gene of unknown function, is located within the TTR region in 18q11.2-12. However, the human *IMPACT* is not imprinted probably because of different CpG islands distribution with differentially methylated region between both species [7]. On the other hand, given the maternal inheritance of the mitochondrial DNA, our results might fit with the recent finding of different mitochondrial haplogroup distribution in early and late onset Swedish and French cases of TTR-FAP. This work suggested that a genetic component affecting mitochondrial function could be one underlying mechanism of the phenotypic variation in the disease, through its impact on the amyloid generating process [8].

Finally, these results are of importance in clinical practice. They should be taken in account to perform the diagnosis in patients at the earliest stage of the affection and also for the management of asymptomatic carriers. In addition they will contribute to refine genetic counseling in the different populations of TTR-FAP.

References

[1] Said G, Ropert A, Faux N. Length-dependent degeneration of fibers in Portuguese amyloid polyneuropathy: A clinicopathological study. Neurology 1984;34:1025-1032.

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

[3] Hellman U, Alarcon F, Lundgren HE, Suhr OB, Bonaiti-Pellié C, Planté-Bordeneuve V. Heterogeneity of penetrance in familial amyloid polyneuropathy, ATTR Val30Met, in the Swedish population. Amyloid 2008;15:181-6.

[4] Holmgren G, Costa PM, Andersson C, Asplund K, Steen L, Beckman L, Nylander PO, Teixeira A, Saraiva MJ, Costa PP. Geographical distribution of TTR met30 carriers in northern Sweden: discrepancy between carrier frequency and prevalence rate. J Med Genet 1994;31:351-354.

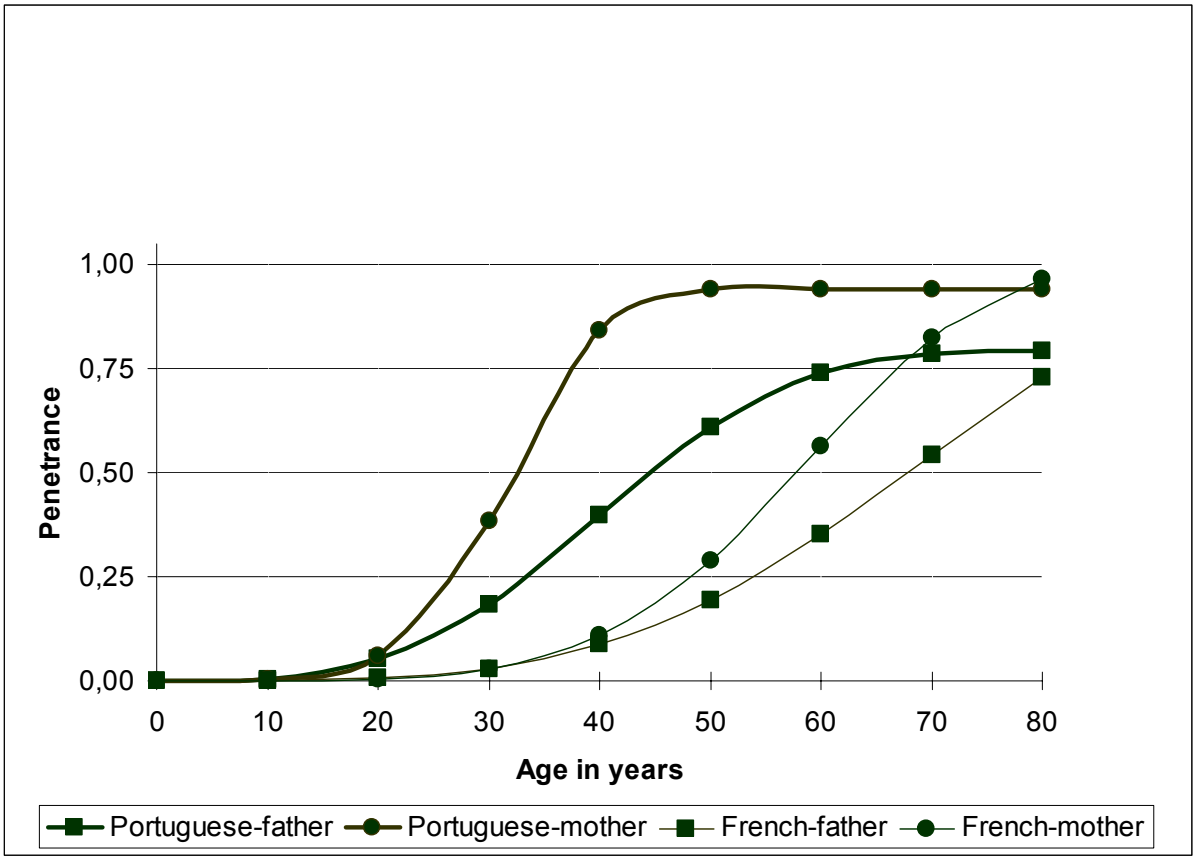
[5] Sousa A, Coelho T, Barros J, Sequeiros J. Genetic epidemiology of familial amyloidotic polyneuropathy (FAP)-type I in Povoa do Varzim and Vila do Conde (north of Portugal). Am J Med Genet 1995;60:512-521.

[6] Alarcon F, Bourgain C, Gauthier-Villars M, Planté-Bordeneuve V, Stoppa-Lyonnet D, Bonaïti-Pellié C. PEL: An unbiased method for estimating age dependent genetic disease risk from pedigree data unselected for family history. Genet Epidemiol (in press).

[7] Okamura K, Hagiwara-Takeuchi Y, Li T, Vu TH, Hirai M, Hattori M, Sakaki Y, Hoffman AR, Ito T. Comparative genome analysis of the mouse imprinted gene *Impact* and its nonimprinted human homolog *IMPACT*: toward the structural basis for species-specific imprinting. Genome Res 2000;10:1878-1889.

[8] Olsson M**,** Hellman U, Planté-Bordeneuve V, Jonasson J, Lång K, Suhr OB. Mitochondrial haplogroup is associated with the phenotype of familial amyloidosis with polyneuropathy in Swedish and French patients. Clin Genet (in press).

Figure 1. Penetrance according to gender of the transmitting parent in French and Portuguese mutation carriers



A Nonparametric Method for Penetrance Function Estimation

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

[1]*Univ. Paris-Sud, IFR69, UMR-S535, F-94817 Villejuif, France*
[2]*INSERM U535, BP 1000, F-94817 Villejuif, France*
[3]*CREST, Statistics Laboratory, France*
[4]*Université Paris-Dauphine, Ceremade, France*

In diseases caused by a deleterious gene mutation, knowledge of age-specific cumulative risks is necessary for medical management of mutation carriers. When pedigrees are ascertained through at least one affected individual, ascertainment bias can be corrected by using a parametric method such as the Proband's phenotype Exclusion Likelihood, or PEL, that uses a survival analysis approach based on the Weibull model. This paper proposes a nonparametric method for penetrance function estimation that corrects for ascertainment on at least one affected: the Index Discarding EuclideAn Likelihood or IDEAL. IDEAL is compared with PEL, using family samples simulated from a Weibull distribution and under alternative models. We show that, under Weibull assumption and asymptotic conditions, IDEAL and PEL both provide unbiased risk estimates. However, when the true risk function deviates from a Weibull distribution, we show that the PEL might provide biased estimates while IDEAL remains unbiased. *Genet. Epidemiol.* 2008.

Key words: risk estimation; ascertainment bias; nonparametric method

The supplemental materials described in this article can be found at http://www.interscience.wiley.com/jpages/0741-0395/suppmat
*Correspondence to: Flora Alarcon, INSERM U535, BP 1000, F-94817 Villejuif, France. E-mail: alarcon@vjf.inserm.fr
Received 6 February 2008; Accepted 16 May 2008
Published online in Wiley InterScience (www.interscience.wiley.com).
DOI: 10.1002/gepi.20354

INTRODUCTION

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

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

When families are selected through at least one affected, two methods taking into account the ascertainment bias have been proposed to estimate the penetrance function: the Proband's phenotype Exclusion Likelihood (or PEL) [Alarcon et al., 2008] and the Prospective Likelihood [Kraft and Thomas, 2000; Le Bihan et al., 1995; Plante-Bordeneuve et al., 2003]. Both are maximum likelihood (ML) methods implementing survival analysis. The Prospective Likelihood corrects for the ascertainment with an analytical expression of the ascertainment probability while PEL is a more intuitive method that corrects for the ascertainment by simply removing the individual (the proband) who allowed his family to be selected. It has been shown for various genetic models and selection schemes that PEL is practically unbiased while the Prospective method is biased in several situations [Alarcon et al., 2008]. However, the penetrance function implemented in the method is modeled with a Weibull distribution. Although this model is widely used in survival analysis because of its capacity to adjust to observed data, the assumption of a Weibull distribution for the penetrance can be a tricky limitation in some applications. A strategy to relax the constraints of the Weibull model is to extend it by adding new parameters. The model is then more general and can fit more situations. But, the complexity of the estimation procedure increases dramatically with the number of parameters (''the curse of dimensionality'') and may turn to be intractable.

In this paper, we propose a nonparametric method for penetrance function estimation, correcting for the ascertainment bias: the Index Discarding EuclideAn Likelihood (IDEAL). The method is applicable for all selection criteria and disease models with at least one affected. Instead of building a likelihood based on the Weibull model, we use a

nonparametric likelihood that does not assume any parametric family for the distribution. We use the Euclidean Likelihood, a fast version of Owen's Empirical Likelihood [Owen, 2001]. To the best of our knowledge, this paper is the first attempt to estimate a penetrance function by means of this approach.

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

METHODS

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

THE PEL

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

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

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

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

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

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

IDEAL

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

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

EMPIRICAL LIKELIHOOD

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

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

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

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

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

Empirical Likelihood is built by means of the multinomial distributions on the sample $X_1, \ldots, X_n$:

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

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

This leads to Empirical Likelihood [Owen, 2001]:

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

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

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

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

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

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

Thus, the Empirical Likelihood method consist in minimizing the Kullback discrepancy between $\mathbb{Q}$ and $\mathbb{P}_n$. Nevertheless, other choices of discrepancy can be used: the Hellinger distance, the Relative Entropy and the Euclidean distance are the more common, but the method can be generalized way beyond [see Bertail et al., 2007]. We propose here to use the Euclidean distance (denoted by χ^2) instead of the Kullback discrepancy in the expression of the log-likelihood ratio because it leads to a closed form for the likelihood that strongly reduces the computational time.

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

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

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

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

In order to estimate the value of the penetrance at an age t, one should consider only the population of carriers aged t or more years. Therefore, at any fixed t, the individual i is considered only if $G_i = 1$ (i.e. if i is affected) and if the current age Y_i of i is bigger than t. The technical effect of this remark is that the reference measure varies as a function of t:

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

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

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

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

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

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

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

$$\begin{aligned}
&\text{sequence of CIs}\ \ \forall t>0,\ \ \mathbb{P}(G(t)\le F(t)\le H(t)) = 95\%,\\
&\text{confidence band}\ \ \mathbb{P}(\forall t>0,\ G(t)\le F(t)\le H(t)) = 95\%.
\end{aligned}$$

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

$$\begin{aligned}
G(t) &= \min\{\theta|\text{IDEAL}(\theta,t)\le c_n\},\\
H(t) &= \max\{\theta|\text{IDEAL}(\theta,t)\le c_n\},
\end{aligned}$$

where critical values of c_n are tabulated [see Owen, 2001, p 159]. For a confidence level of 95% and $n \le 1,000$, c_n writes

$$\begin{aligned}
n \le 100,\ \ c_n &= 3.0123 + 0.4835\log(n)\\
&\quad - 0.00957\log(n)^2 - 0.001488\log(n)^3,\\
100 < n \le 1,000,\ \ c_n &= 3.0806 + 0.4894\log(n) - 0.02086\log(n)^2.
\end{aligned}$$

SIMULATION

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

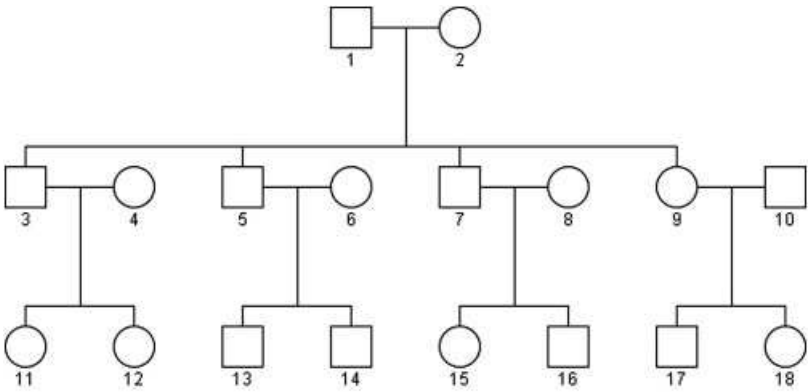


Fig. 1. Family structure.

size was fixed to 5,000 families and therefore to 90,000 individuals.

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

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

As in Alarcon et al. [2008], in order to model a realistic selection process, we defined a period length T to select the probands: only individuals affected during the last T years could be selected, and this with a probability p_s. We considered two different period lengths: a period of 20 years ($T = 20$) and a period of 1 year ($T = 1$). The probability p_s was introduced to simulate the fact that, in real situations, some individuals affected during the study are not detected and therefore do not become probands. A family was included in the sample as soon as one of its member was a proband. Under the CDMS model, we introduced an age criterion for selection (35 years) to increase the probability of detecting families with mutation carriers [Claus et al., 1990].

RESULTS

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

BEHAVIOR OF IDEAL UNDER A WEIBULL MODEL

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

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

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

COMPARISON OF IDEAL AND PEL UNDER UNIFORM AND CAUCHY MODELS

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

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

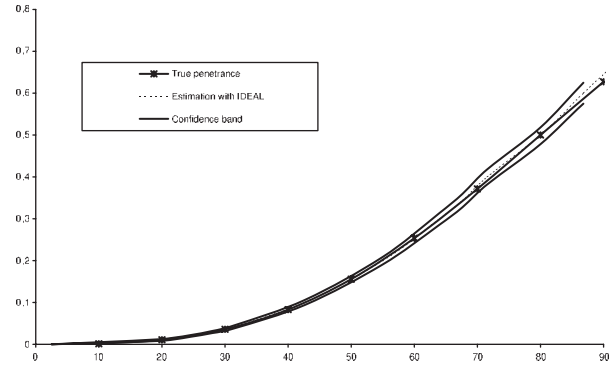


Fig. 2. Simulation in case of a low ascertainment probability.

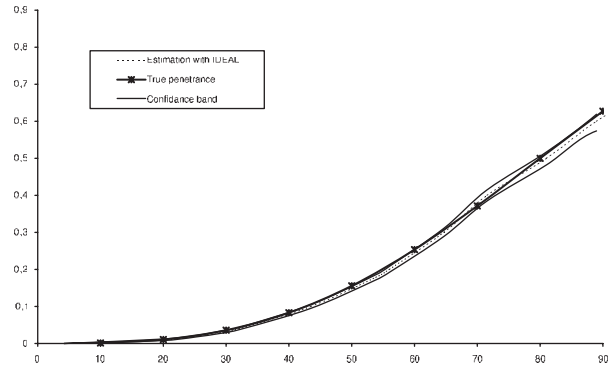


Fig. 3. Simulation in case of a high ascertainment probability.

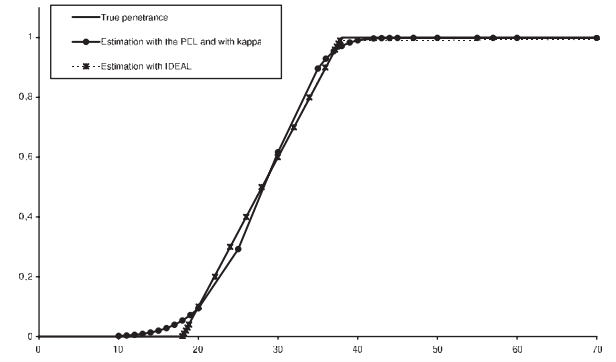


Fig. 4. Comparison of PEL and IDEAL under a Uniform distribution, in case of a CDMS model with an age criterion of 35 for the selection, $p_s = 0.5$ and $T = 20$ for the selection. PEL, Proband's Phenotype Exclusion Likelihood; IDEAL, Index Discarding EuclideAn Likelihood; CDMS, complex diseases with monogenic sub-entities.

that IDEAL is perfectly unbiased while PEL has a nonnegligible bias.

SENSITIVITY OF THE TWO METHODS TO THE SAMPLE SIZE

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

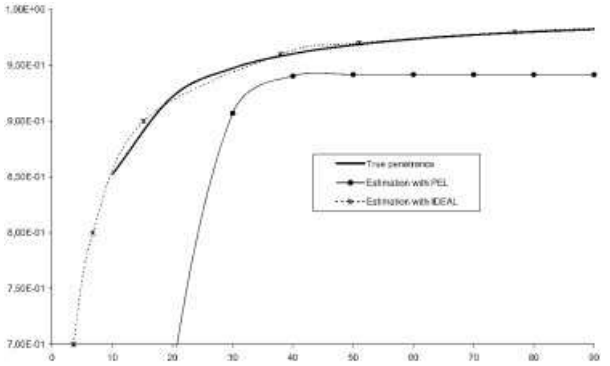


Fig. 5. Comparison of PEL and IDEAL under a Cauchy distribution with parameter 5, in case of a CDMS model with an age criterion of 35 for the selection, $p_s = 0.5$ and $T = 20$ for the selection. PEL, Proband's Phenotype Exclusion Likelihood; IDEAL, Index Discarding EuclideAn Likelihood; CDMS, complex diseases with monogenic sub-entities.

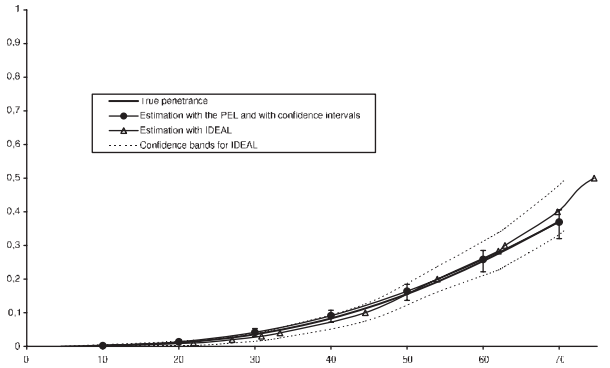


Fig. 6. Simulation in case of an MD model. MD, monogenic diseases.

DISCUSSION

In this paper, we have proposed an estimation method (IDEAL) that adapts to the penetrance function model and that corrects for the ascertainment bias when families are ascertained through at least one affected. We have compared this method with a parametric method (PEL) also designed to take into account the ascertainment bias. First, we have shown that IDEAL corrects for the ascertainment bias and that it leads to unbiased estimates. Then, we have shown through simulations on large samples that IDEAL performs as well as PEL when the true penetrance is Weibull distributed and significantly

better when this assumption fails. This adaptability of IDEAL allows to estimate penetrance functions in new contexts without risking a bias due to a model misspecification.

In the simulation part, we have only reported results for the theoretical situations where all genotypes are known. Unknown genotypes can easily be taken into account in an IDEAL method by means of weighting. For example, we can estimate a probability p_i of mutation for each individual from the known genotypes of the family and use p_i/n as a reference instead of $1/n$.

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

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

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

ACKNOWLEDGMENTS

We would particularly like to acknowledge the comments and corrections of Catherine Bourgain. We also acknowledge the reviewers for their comments.

REFERENCES

Alarcon F, Bourgain C, Gautier-Villars M, Planté-Bordeneuve V, Stoppa-Lyonnet D, Bonaïti-Pellié C. 2008. PEL: an unbiased method for estimating genetic disease risk from pedigree data unselected for family history, submitted.

Bertail P, Harari-Kermadec H, Ravaille D. 2007. **φ**-Divergence empirique et vraisemblance empirique généralisée. Ann Econ Stat 85:131–157.

Bonadona V, Sinilnikova OM, Chopin S, Antoniou AC, Mignotte H, Mathevet P, Bremond A, Martin A, Bobin JY, Romestaing P et al. 2005. Contribution of BRCA1 and BRCA2 germ-line mutations to the incidence of breast cancer in young women: results from a prospective population-based study in France. Genes Chromosomes Cancer 43:404–413.

Carayol J, Bonaïti-Pellié C. 2004. Estimating penetrance from family data using a retrospective likelihood when ascertainment depends on genotype and age of onset. Gen Epidemiol 27:109–117.

Claus EB, Risch NJ, Thompson WD. 1990. Age of onset as an indicator of familial risk of breast cancer. Am J Epidemiol 131:961–972.

Crow J. 1965. Problems of ascertainment in the analysis of family data. In: Neel JV, Shaw MW, Schull WJ, editors). Genetics and the Epidemiology of Chronic Disease. Washington, DC: Public Health 'Source' Publication.

Dunlop MG, Farrington SM, Carothers AD, Wyllie AH, Sharp L, Burn J, Liu B, Kinzler KW, Vogelstein B. 1997. Cancer risk associated with germline DNA mismatch repair gene mutations. Hum Mol Genet 6:105–110.

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

Lalouel JM. 1979. GEMINI: a computer program for optimization of general nonlinear functions. Technical Report No. 14, Department of Medical Biophysics and Computing, University of Utah, Salt Lake City.

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

Owen AB. 2001. Empirical likelihood. Boca Raton: Chapman & Hall, CRC.

Plante-Bordeneuve V, Carayol J, Ferreira A, Adams D, Clerget-Darpoux F, Misrahi M, Said G, Bonaiti-Pellie C. 2003. Genetic study of transthyretin amyloid neuropathies: carriers risks among French and Portuguese families. J Med Genet 40:793–796.

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

APPENDIX: INDEX DISCARDING

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

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

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

rewritten:

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

Dividing by k, we get

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

Therefore, if we set W_0 as follows:

$$W_0(x)=\begin{cases}1 & \text{if } i \text{ is not a proband and } x=X_i,\\ 1-\frac{1}{k} & \text{if } i \text{ is a proband,}\\ 0 & \text{otherwise,}\end{cases}$$

then θ is given as the solution of

$$\mathbb{E}_{W_0}\left[\sum_{i=1}^{r}(\mathbb{1}_{P_i=1}-\theta\mathbb{1}_{G_i=1})\right]=0.$$

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

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

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

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

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

Publications

Revues à comité de lecture, articles publiés ou sous presse

[J1] **Alarcon F**, Lasset C, Carayol J, Bonadona V, Perdry H, Desseigne F, Wang Q, Bonaïti-Pellié C. Estimating cancer risk in HNPCC by the GRL method. *European Journal of Human Genetics*. 15(2007), 831–836.

[J2] **Alarcon F**, Bonaïti-Pellié C, Harari-Kermadec H. A nonparametric method for penetrance function estimation. *Genetic Epidemiology*. 33(1). 2008. 38-44

[J3] **Alarcon F**, Bourgain C, Gauthier-Villard M, Planté-Bordeneuve V, Stoppa-Lyonnet D, Bonaïti-Pellié C. PEL : An unbiased method for estimating age-dependence genetic disease risk from pedigree data unselected for family history. *Genetic Epidemiology*. 2008. *(Online)*

[J4] Hellman U, **Alarcon F**, Lundgren HE, Suhr OB, Bonaïti-Pellié C, Planté-Bordeneuve V. Heterogeneity of penetrance in familial amyloid polyneuropathy, ATTR Val30Met, in the Swedish population. *Amyloid*. 15(3). 2008. 181-186.

Revues à comité de lecture, articles soumis

[J5] Bonaïti B, **Alarcon F**, Bonaïti-Pellié C, Planté-Bordeneuve V. Parent-of-origin effect in Transthyretin related amyloid polyneuropathy. *Soumis à Amyloid*.

Communications orales

ALARCON F, BOURGAIN C, BONAÏTI-PELLIÉ C. Quantifying the bias in disease risk estimates associated with a deleterious mutation using families ascertained through one affected individual. European Mathematical Genetics Meeting, Cardiff (Royaume Uni), 5-7 avril 2006.

HASSID S, NOGUES C, CARAYOL J, **ALARCON F**, LABBE M, REZVANI A, STOPPA-LYONNET D, BERTHET P, FRICKER JP, le Groupe Génétique et Cancer, ANDRIEU N, BONAÏTI-PELLIÉ C. Estimation du risque de cancer associé aux gènes BRCA : étude GENECAN. Congrès ADELF-EPITER, Dijon, 30 août-1er septembre 2006. Rev Epidemiol Sante Publique 2006 ;54 (HS2) : 2S16.

PLANTE-BORDENEUVE V, **ALARCON F**, HELLMAN U, FERREIRA A, ZAROS C, SUHR O, WADDINGTON CRUZ M, MISRAHI M, SAID G, BONAÏTI-PELLIE C. Difference of penetrance in transthyretin amyloid neuropathies across families from European and Brazilian origin. 16th Meeting of the European Neurological Society, Lausanne, 27-31 mai 2006. J Neurol 2006, 253(suppl) :38.

HELLMAN U, **ALARCON F**, LUNDGREN HE, SUHR OB, BONAÏTI-PELLIÉ C, PLANTÉ-BORDENEUVE V. Heterogeneity of penetrance in familial amyloid polyneuropathy, ATTR Val30Met, in the Swedish population. International Amyloidosis Symposium, Londres, 2-5 septembre 2008.

PLANTÉ-BORDENEUVE V, BONAÏTI B, **ALARCON F**, BONAÏTI-PELLIÉ C. Parent-of-origin effect in Transthyretin related amyloid polyneuropathy. International Amyloidosis Symposium, Londres, 2-5 septembre 2008.

Communications affichées

HASSID S, NOGUÈS C, CARAYOL J, **ALARCON F**, MOHAMDI H, LABBÉ M, REZVANI A, STOPPA-LYONNET D, BERTHET P, FRICKER JP, ANDRIEU N, BONAÏTI-PELLIÉ C. BRCA penetrance for breast and ovarian cancers : a heterogeneity study. Annual meeting of the International Genetic Epidemiology Society, Tampa (USA), 16-17 novembre 2006. Genet Epidemiol 2007, 31 : 450-507 (abstract 86).

ALARCON F, BOURGAIN C, PLANTÉ-BORDENEUVE V, STOPPA-LYONNET D, BONAÏTI-PELLIÉ C. Estimating genetic disease risk from family data : choosing the optimal method to correct for ascertainment. Annual meeting of the International Genetic Epidemiology Society, York, 7-10 septembre 2007. Genet Epidemiol 31(6) :615.

ALARCON F, BOURGAIN C, PLANTÉ-BORDENEUVE V, STOPPA-LYONNET D, BONAÏTI-PELLIÉ C. Importance du choix de la méthode dans l'estimation des risques associés à des mutations génétiques. 4èmes Assises de génétique humaine et médicale, Lille, 17-19 janvier 2008, Med Sci 2008 24(HS n°1) : 123-4.

PLANTÉ-BORDENEUVE V, BONAÏTI B., **ALARCON F**, BONAÏTI-PELLIÉ C. Parent-of-origin effect in Transthyretin related amyloid polyneuropathy. Meeting de l'American Academy of Neurology, Seattle, 25 avril-2 mai 2009.

Bibliographie

[1] ABEL, L., AND BONNEY, G. A Time-Dependent Logistic Hazard Function for Modeling Variable Age of Onset in Analysis of Familial Diseases. *Genetic Epidemiology 7* (1990), 391–407.

[2] ALARCON, F., LASSET, C., CARAYOL, J., BONADONA, V., PERDRY, H., DESSEIGNE, F., WANG, Q., AND BONAÏTI-PELLIÉ, C. Estimating cancer risk in HNPCC by the GRL method. *European Journal of Human Genetics 15* (2007), 831–836.

[3] BELOT, A., GROSCLAUDE, P., BOSSARD, N., JOUGLA, E., BENHAMOU, E., DELAFOSSE, P., GUIZARD, A., MOLINIÉ, F., DANZON, A., BARA, S., ET AL. Cancer incidence and mortality in France over the period 1980-2005. *Rev Epidemiol Sante Publique* (2008).

[4] BERTAIL, P., HARARI-KERMADEC, H., AND RAVAILLE, D. ϕ-Divergence empirique et vraisemblance empirique generalisee. In *Ann Econ Stat* (2007), vol. 85, pp. 131–157.

[5] BONADONA, V., SINILNIKOVA, O., CHOPIN, S., ANTONIOU, A., MIGNOTTE, H., MATHEVET, P., BRÉMOND, A., MARTIN, A., BOBIN, J., ROMESTAING, P., ET AL. Contribution of BRCA1 and BRCA2 germ-line mutations to the incidence of breast cancer in young women : results from a prospective population-based study in France. *Genes Chromosomes Cancer 43*, 4 (2005), 404–13.

[6] BONAÏTI, B., ALARCON, F., BONAÏTI-PELLIÉ, C., AND PLANTÉ-BORDENEUVE, V. Parent-of-origin effetc in Transthyretin related amyloid polyneuropathy. *Soumis à Amyloid* (2008).

[7] BONNEY, G. Regressive logistic models for familial disease and other binary traits. *Biometrics 42*, 3 (1986), 611–25.

[8] CANNINGS, C., AND THOMPSON, E. Ascertainment in the sequential sampling of pedigrees. *Clinical Genetics 12*, 4 (1977), 208–212.

[9] CARAYOL, J., AND BONAITI-PELLIE, C. Estimating penetrance from family data using a retrospective likelihood when ascertainment depends on genotype and age of onset. *Genetic Epidemiology 27*, 2 (2004), 109–117.

[10] CARAYOL, J., KHLAT, M., MACCARIO, J., AND BONAITI-PELLIE, C. Hereditary nonpolyposis colorectal cancer : current risks of colorectal cancer largely overestimated, 2002.

[11] CHOMPRET, A., ABEL, A., STOPPA-LYONNET, D., BRUGIERE, L., PAGES, S., FEUNTEUN, J., AND BONAÏTI-PELLIE, C. Sensitivity and predictive value of criteria for p53 germline mutation screening. *J Med Genet 38*, 1 (2001), 43–7.

[12] CHOMPRET, A., BRUGIERE, L., GARDES, M., DESSARPS-FREICHEY, F., ABEL, A., HUA, D., LIGOT, L., DONDON, MG. BRESSAC-DE PAILLERETS, B., FRÈBOURG, T., LEMERLE, J., BONAÏTI-PELLIE, C., AND FEUNTEUN, J. P53 germline mutations in childhood cancers and cancer risk for carrier individuals. *Br J Cancer 82*, 1 (2000), 1932–1937.

[13] CLERGET-DARPOUX, F., BONAITI-PELLIE, C., AND HOCHEZ, J. Effects of misspecifying genetic parameters in lod score analysis. *Biometrics 42*, 2 (1986), 393–9.

[14] CROW, J. Problems of ascertainment in the analysis of family data. *Genetics and the Epidemiology of Chronic Diseases* (1965), 23–44.

[15] DRUGGE, U., ANDERSSON, R., CHIZARI, F., DANIELSSON, M., HOLMGREN, G., SANDGREN, O., AND SOUSA, A. Familial amyloidotic polyneuropathy in Sweden : a pedigree analysis. *Journal of Medical Genetics 30*, 5 (1993), 388–392.

[16] DUNLOP, M., FARRINGTON, S., CAROTHERS, A., WYLLIE, A., SHARP, L., BURN, J., KINZLER, K., AND VOGELSTEIN, B. Cancer risk associated with germline DNA mismatch repair gene mutations. *Hum Mol Genet 6*, 1 (1997), 105–110.

[17] EASTON, D., FORD, D., AND BISHOP, T. Breast and ovarian cancer incidence in BRCA1-mutation carriers. Breast Cancer Linkage Consortium. *American Journal of Human Genetics 56*, 1 (1995), 265.

[18] EFRON, B. The bootstrap, jacknife and other resampling plans. *Philadelphia : SIAM* (1982).

[19] EFRON, B. Better Bootstrap Confidence Intervals. *Journal of the American Statistical Association 82*, 397 (1987), 171–185.

[20] EISINGER, F., AND AL. Identification et prise en charge des prédispositions héréditaires aux cancers du sein et de l'ovaire. *Bull Cancer 91* (2004), 219–237.

[21] ELSTON, R., AND STEWART, J. A general model for the genetic analysis of pedigree data. *Hum Hered 21*, 6 (1971), 523–42.

[22] FORD, D., EASTON, D., STRATTON, M., NAROD, S., GOLDGAR, D., DEVILEE, P., BISHOP, D., WEBER, B., LENOIR, G., CHANG-CLAUDE, J., ET AL. Genetic Heterogeneity and Penetrance Analysis of the BRCA1 and BRCA2 Genes in Breast Cancer Families. *The American Journal of Human Genetics 62*, 3 (1998), 676–689.

[23] GAIL, M., PEE, D., BENICHOU, J., AND CARROLL, R. Designing studies to estimate the penetrance of an identified autosomal dominant mutation : cohort, case-control, and genotyped-proband designs. *Genetic epidemiology 16*, 1 (1999), 15.

[24] GONG, G., AND WHITTEMORE, A. Optimal designs for estimating penetrance of rare mutations of a disease-susceptibility gene. *Genetic Epidemiology 24*, 3 (2003), 173–180.

[25] GREEN, J., DRISCOLL, M., BARNES, A., MAHER, E., BRIDGE, P., SHIELDS, K., AND PARFREY, P. Impact of Gender and Parent of Origin on the Phenotypic Expression of Hereditary Nonpolyposis Colorectal Cancer in a Large Newfoundland Kindred With a Common MSH2 Mutation . *Colon and Rectum 45*, 9 (2002), 1223–1232.

[26] HARDELL, L., HOLMGREN, G., STEEN, L., FREDRIKSON, M., AND AXELSON, O. Occupational and other risk factors for clinically overt familial amyloid polyneuropathy . *Epidemiology 6*, 6 (1995), 598–601.

[27] HOLMGREN, G., COSTA, P., ANDERSSON, C., ASPLUND, K., STEEN, L., BECKMAN, L., NYLANDER, P., TEIXEIRA, A., SARAIVA, M., AND COSTA, P. Geographical distribution of TTR met 30 carriers in northern Sweden . *J Med Genetic 31* (1994), 351–354.

[28] KRAFT, P., AND THOMAS, D. Bias and Efficiency in Family-Based Gene-Characterization Studies : Conditional, Prospective, Retrospective, and Joint Likelihoods. *The American Journal of Human Genetics 66*, 3 (2000), 1119–1131.

[29] LALOUEL, J. *Gemini : A Computer Program for Optimization of General Nonlinear Functions*. University of Hawaii, Population Genetics Laboratory, 1979.

[30] LALOUEL, J., RAO, D., MORTON, N., AND ELSTON, R. A unified model for complex segregation analysis. *American Journal of Human Genetics 35*, 5 (1983), 816.

[31] LE BIHAN, C., MOUTOU, C., BRUGIERES, L., FEUNTEUN, J., AND BONAITI-PELLIE, C. ARCAD : a method for estimating age-dependent disease risk associated with mutation carrier status from family data. *Genet Epidemiol 12*, 1 (1995), 13–25.

[32] MORTON, N. Sequential tests for the detection of linkage. *American Journal of Human Genetics 7*, 3 (1955), 277.

[33] MORTON, N. Genetic Tests Under Incomplete Ascertainment. *American Journal of Human Genetics 11*, 1 (1959), 1.

[34] MORTON, N., AND MACLEAN, C. Analysis of family resemblance. 3. Complex segregation of quantitative traits. *American Journal of Human Genetics 26*, 4 (1974), 489.

[35] OLSSON, M., HELLMAN, U., PLANTÉ-BORDENEUVE, V., JONASSON, J., LANG, K., AND SUHR, O. Mitochondrial haplogroup is associated with the phenotype of familial amyloidosis with polyneuropathy in swedish and french patients. *Clinical Genetics Sous presse* (2008).

[36] OWEN, A. *Empirical Likelihood*. Chapman & Hall/CRC, 2001.

[37] PENNEC, S. La place des familles à quatre générations en France. *Population 51*, 1 (1996), 31–59.

[38] PLANTE-BORDENEUVE, V., CARAYOL, J., FERREIRA, A., ADAMS, D., CLERGET-DARPOUX, F., MISRAHI, M., SAID, G., AND BONAITI-PELLIE, C. Genetic study of transthyretin amyloid neuropathies : carrier risks among French and Portuguese families. *J Med Genet 40*, 11 (2003), e120.

[39] QIN, J., AND LAWLESS, J. Empirical likelihood and general estimating equations. *Annals of Statistics 22* (1994), 300–300.

[40] SARAIVA, M. Transthyretin mutations in hyperthroxinemia and amyloid diseases. *Hum Mut 17* (2001), 493–503.

[41] SOUSA, A., COELHO, T., BARROS, J., AND SEQUEIROS, J. Genetic epidemiology of familial amyloidotic polyneuropathy (FAP)-Type I in Povoa do Varzim and Vila do Conde (North of Portugal) . *Am J Med Genet (Neuropsych Genet) 60* (1995), 512–521.

[42] STOPPA-LYONNET, D., LAURENT-PUIG, P., ESSIOUX, L., PAGÈS, S., ITHIER, G., LIGOT, L., FOURQUET, A., SALMON, R., CLOUGH, K., POUILLART, P., ET AL. BRCA1 sequence variations in 160 individuals referred to a breast/ovarian family cancer clinic. Institut Curie Breast Cancer Group. *American Journal of Human Genetics 60*, 5 (1997), 1021.

[43] VAN DER VAART, A. *Asymptotic statistics Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge University Press, 1998.

[44] VASEN, H., MECKLIN, J., MEERA KHAN, P., AND LYNCH, H. The International Collaborative Group on Hereditary Non-Polyposis Colorectal Cancer (ICG-HNPCC). *Diseases of the Colon & Rectum 34*, 5 (1991), 424–425.

[45] WACHOLDER, S., HARTGE, P., STRUEWING, J., PEE, D., MCADAMS, M., BRODY, L., AND TUCKER, M. The Kin-Cohort Study for Estimating Penetrance. *American Journal of Epidemiology 148*, 7 (1998), 623–630.

[46] WAGNER, T., STOPPA-LYONNET, D., FLEISCHMANN, E., MUHR, D., PAGÈS, S., SANDBERG, T., CAUX, V., MOESLINGER, R., LANGBAUER, G., BORG, A., ET AL. Denaturing High-Performance Liquid Chromatography Detects Reliably BRCA1 and BRCA2 Mutations. *Genomics 62*, 3 (1999), 369–376.

[47] WEINBERG. Method und Fehlerquellen der Untersuchung auf Mendleschen Zahlen Beim Menschen. *Arch. Rass.u.Ges.Biol 9* (1912), 165–174.

Résumé

Certaines maladies à âge de début variable sont dues à la présence de mutation(s) d'un gène. Pour ces maladies, l'estimation précise du risque cumulé d'être atteint à un certain âge chez les porteurs de la mutation (appelé fonction de pénétrance) permet une meilleure compréhension des mécanismes sous-jacents de la maladie et permet également de développer et d'améliorer des stratégies de prévention. L'estimation de la pénétrance se fait à partir de données familiales recensées sur certains critères plus ou moins complexes. Cependant, la plupart des études utilisent des méthodes d'estimation qui ne tiennent pas compte du biais que représente ce recensement, ce qui implique des fonctions de pénétrance fortement surestimées.

Au cours de cette thèse, nous nous sommes intéressés au développement de méthodes d'estimation de la fonction de pénétrance corrigeant pour le recensement des familles. Dans un premier temps, nous avons étudié une méthode permettant d'estimer la pénétrance quel que soit le mode de recensement des familles. Nous nous sommes ensuite intéressés plus particulièrement au cas de familles recensées sur l'existence d'au moins un atteint par famille. Dans ce cadre, nous avons développé une méthode d'estimation, que nous avons appelée la PEL, et l'avons comparée à une méthode déjà existante, la méthode prospective. Nous avons montré que la PEL était moins biaisée que la méthode prospective. Nous avons ensuite appliqué ces méthodes à deux jeux de données, l'un portant sur des familles françaises et portugaises, atteintes de neuropathie amyloide héréditaire ; l'autre portant sur des familles atteintes de cancer du sein. Nous avons également mené une étude sur des familles suédoises atteintes de neuropathie amyloide héréditaire.

La PEL est une méthode paramétrique basée sur un modèle de Weibull et nous avons montré qu'elle n'était pas adaptée lorsque la distribution des données s'éloignait fortement de ce modèle. Nous avons donc développé une méthode non-paramétrique, que nous avons appelée IDEAL, permettant l'estimation de la pénétrance en tenant compte du recensement des familles et l'avons comparée à la PEL. Nous avons montré que IDEAL était moins biaisée lorsque la loi des données était éloignée d'une loi de Weibul.

Mots-clés: estimation de risque, fonction de pénétrance, biais de recensement, PEL, IDEAL.