Réseaux bayésiens et analyse de survie pour l'estimation de courbes de pénétrance du cancer broncho-pulmonaire lié à des prédispositions génétiques

Lucas Ducrot

Sorbonne Université
Laboratoire de Probabilités, Statistique et Modélisation - LPSM, CNRS 8001
École doctorale 386 : Sciences Mathématiques de Paris Centre

THÈSE DE DOCTORAT
Discipline : Mathématiques appliquées

présentée par Lucas Ducrot

Réseaux bayésiens et analyse de survie pour l'estimation de courbes de pénétrance du cancer broncho-pulmonaire lié à des prédispositions génétiques

sous la direction de Grégory Nuel et Patrick Benusiglio

Rapporteuses :

Agathe GUILLOUX	INRIA Paris
Catherine LEGRAND	Université Catholique de Louvain

Soutenue le 13 mai 2024 devant le jury composé de :

Christophe AMBROISE	Université d'Évry Val d'Essonne	Président du jury
Patrick BENUSIGLIO	Sorbonne Université	Co-encadrant
Camille CHARBONNIER	Université de Rouen Normandie	Examinatrice
Agathe GUILLOUX	INRIA Paris	Rapporteuse
Catherine LEGRAND	Université Catholique de Louvain	Rapporteuse
Grégory NUEL	Sorbonne Université	Directeur
Maud THOMAS	Sorbonne Université	Examinatrice

LPSM
Sorbonne Université
4 place Jussieu, Paris

Remerciements

A l'approche de ma soutenance marquant la fin d'un projet de plus de trois ans, je m'engage dans le traditionnel mais finalement logique exercice des remerciements. Logique car en traversant cette période à la fois longue et intense, on se rend compte de l'importance des gens qui nous entourent.

Ainsi je souhaite remercier tout d'abord mes encadrants de thèse, Grégory et Patrick pour m'avoir fait confiance et guider sur ce projet. Plus particulièrement, Grégory, merci pour tes conseils et retours toujours pertinents et Patrick, merci d'être un médecin chercheur aussi passionné et ouvert à de nouvelles méthodes. J'ai appris énormément de choses en trois ans et c'est en grande partie grâce à vous.

Je remercie mes deux rapporteuses Agathe Guilloux et Catherine Legrand pour avoir, d'abord, acceptées cette tâche chronophage de relecture et d'évaluation mais également de l'avoir fait de manière très méthodique. Leurs nombreux conseils et remarques sont tous pertinents, ont nourri mon analyse et permis d'affiner mes recherches ainsi que cette présentation de soutenance.

Je remercie le reste de mon jury de thèse Christophe Ambroise, Camille Charbonnier et Maud Thomas, d'avoir accepté d'être présents aujourd'hui pour évaluer mon travail. Notamment, Christophe Ambroise, merci d'avoir initialement accepté d'être un potentiel rapporteur sans la moindre hésitation et avec humour ; Camille, j'ai été ravi de te rencontrer lors d'EMGM et je te remercie pour nos discussions à cette occasion qui m'ont éclairé sur mon sujet. Enfin Maud, merci d'avoir accepté d'être d'abord dans mon comité de suivi de thèse et aujourd'hui la représentante locale de mon jury. C'est un honneur pour moi de vous avoir tous, avec Agathe Guilloux et Catherine Legrand, présents pour ma soutenance.

Je remercie vivement Hugues Moretto pour son efficacité et son dévouement au bon fonctionnement des systèmes informatiques du LPSM, tu m'as énormément aidé tout le long de ma thèse. De la même manière, je remercie l'ensemble de l'équipe administrative (Valérie, Nathalie, Marouane, Imene et anciennement Louise) pour leur accompagnement au quotidien et sans qui nous ne pourrions pas nous concentrer sur la recherche. Je remercie Jasmine pour son travail de numérisation de données si important pour les projets appliqués. Je remercie également Nadia Nathan et Marie Legendre pour nous avoir apporter un projet très intéressant sur lequel nous avons pu travailler.

Durant ma thèse, j'ai eu la chance de rencontrer deux personnes que je qualifierai d'anges gardiennes, Anna Bonnet et Charlotte Dion, qui m'ont accompagné et donné des conseils à de nombreuses reprises. Pour l'ensemble du soutien qu'elles m'ont apporté, je les remercie chaleureusement. J'ajoute également à ces remerciements Laurent Mazliak avec qui j'ai beaucoup échangé à la fois sur des sujets de mathématiques et d'autres plus généraux, ça a toujours été un plaisir de discuter avec toi.

Si j'ai mis en avant trois membres permanents qui ont particulièrement compté pour moi, ce serait oublier que le labo est en fait rempli de personnes géniales qui contribuent à une ambiance bienveillante et conviviale. Je n'ai commencé à traîner du côté de la salle café des statistiques qu'au milieu de ma thèse, je m'y suis pourtant senti tout de suite à mon aise, notamment en essayant de contribuer (à la hauteur de mes moyens) aux mots-croisés. De ce côté-là, je remercie la team élite composé d'Arnaud, Anna Be, Anna Bo, Claire et Antoine de m'avoir laissé suggérer des mots même s'ils étaient nuls ! Ce temps passé dans cette salle café m'a également permis de connaître Erwan, Maxime, Étienne, Christophe puis plus récemment Margaux et Rafael, tous aussi accueillants et intéressants que les premiers

cités. Rencontrés de manière similaire, je remercie sincèrement Laure et Sylvain qui m'ont conseillé et proposé leur aide que ce soit pour de la relecture ou pour mon avenir à plusieurs occasions. Je remercie enfin Stéphane, dont je n'avais entendu dire que du bien, et qui surpasse ce que l'on dit de lui. Ayant eu mon bureau du côté probabilités, j'ai également pu échanger avec Quentin, Camille, Thierry, Romain, Piet, Damien et Catherine qui ont tous contribué à ce que je me sente à l'aise dans ces couloirs loin des statistiques. J'ajoute à ces remerciements Gabriel Lang, espiègle et cultivé, qui ne m'en a jamais voulu de l'avoir confondu avec Eric Moulines et avec qui les moments partagés sont toujours accompagnés de rires.

J'arrive maintenant à ceux que j'ai le plus côtoyés durant cette thèse, mes chers collègues doctorants. J'ai eu la chance d'arriver en cours d'année dans un groupe bien formé qui m'a accueilli à bras ouverts. Tous dans le même bateau à naviguer sur les eaux tumultueuses de la thèse, sur des sujets différents mais avec un véritable esprit d'équipe(age). C'est grâce à cette entraide, que ce soit pour de l'administratif ou des questions plus fondamentales, que nous avons progressé dans nos projets et je veux donc remercier chacun d'entre eux en essayant de n'oublier personnes.

Pour ce faire, je vais reprendre le découpage habituel par bureaux en commençant par le "bureau des absents". Je remercie ainsi Lucas I, pour avoir été la première personne à m'avoir parlé et intégré au groupe lors de mon arrivée, Pierre, pour m'avoir appris des choses sur la théorie des jeux et Loic, pour ta gentillesse et ton sourire constant.

Je passe ensuite au "bureau du ficus", le plus petit des bureaux pourtant rempli de personnes géniales. Je remercie donc Yoan, pour ta bienveillance et ton originalité permanente, Guillaume, pour ton rire et avoir partagé avec moi la peur de ne pas être réinscrit en thèse, Sonia, pour m'avoir battu au gainage, Robin, pour ton rhum arrangé et tes problèmes mathématiques posés au moment du café, Tristan, pour ta sympathie et ta capacité de vulgarisation et Antonio, d'être quelqu'un d'aussi intense et engagé sur un grand nombre de problématiques importantes.

Vient ensuite le tour du "bureau des anciens" ou "bureau de l'ambiance", pièce centrale de la vie doctorale du LPSM. Je remercie tout d'abord Emilien, d'être celui dont je suis le plus proche au labo, un ami fidèle avec qui j'ai toujours été sur la même longueur d'onde, David, thank you for bringing people together and being a wonderful man, Lucas B, pour avoir partagé des moments sports avec moi que ce soit à l'escalade ou à la salle, Nicolas, pour ton humour et nous avoir initié au jeu de rôle papier, Thomas, pour ta zénitude permanente, Irene et Rémi, que je connais moins mais avec qui j'ai toujours partagé des moments sympa.

Je passe maintenant à mon bureau, le "bureau des isolés" mélangeant des statisticiens appliqués, des probas et des dynamiciens. Je remercie Alexandra, ma grande soeur de thèse qui m'a accueilli et m'a aidé régulièrement, Bastien, pour ta bonhomie et ton honnêteté à toute épreuve, Jérôme, pour avoir été celui avec qui j'ai finalement passé le plus temps et qui m'a fait comprendre les liens entre mes intérêts de recherches et les siens, Nadège, pour sa gentillesse perpétuelle, William, pour l'ensemble de nos discussions et notre intérêt commun pour les gâteaux et Sergi, le plus stylé du labo.

Le tour des bureaux côté probabilités étant terminé, j'en viens aux doctorants de statistiques. Je remercie Ariane, pour m'avoir suivi dans des aventures épiques à l'escalade ou en ligne et de ne pas m'en avoir trop voulu qu'en nous l'avons oublié en route avec Emilien, Iqraa, pour avoir été la première statisticienne à qui j'ai parlé, j'espère toujours te rapatrier à l'escalade, Camila, pour ta jovialité et toutes les fêtes auxquelles tu m'as invité, Miguel, pour ton énergie débordante, Alexis, que j'ai appris à connaître aux JdS. Je remercie également Adeline, pour tous tes conseils, Ludovic, pour nos échanges de clins d'oeil, Yazid, pour avoir égayé nos pauses cafés, Francesco, pour avoir vécu avec moi les difficultés de faire une thèse très appliquée dans un labo de maths, on a ainsi pu se serrer les coudes, Grace,

pour avoir partagé nos TDs ensemble, Romain, pour sa sympathie naturelle, Nathan, pour nos discussions autour d'une bière à Bruxelles, Paul E, pour son intérêt excessif pour BG3 (mais je comprends moi-même pourquoi) et Paul L, pour son leadership naturel.

Pour finir ces remerciements aux doctorants, je vais maintenant parler des nouveaux arrivants, qui représentent la "nouvelle âme" du labo. Je remercie Alexandre, pour l'ensemble de ton œuvre (escalade, gastronomie, esprit), on a passé pas mal de temps ensemble ces derniers temps, toujours avec plaisir, Ulysse, pour ta culture (générale et mathématique) phénoménale et ta bienveillance naturelle, Maxence, pour ta gentillesse et pour m'avoir aidé à trouver un postdoc, Elias, pour ton originalité et ton incompréhension constante des phrases de Bastien, Vladimir, d'être une crème, Roland, d'être toujours souriant et attentionné, Stan, d'être mon acolyte handballeur au LPSM ; Jean-Luc, Sobihan et Abdoulaye, pour la cool-attitude qui rayonne dans votre bureau, Ferdinand, pour nos discussions chez Camila. Je rajouterai une pensée pour Maya et Valentin, deux anciens stagiaires partis vers d'autres horizons pour leur doctorat et qui ont largement participé à des moments cool de la vie du labo.

Je veux maintenant remercier quelques personnes qui m'ont donné des opportunités à diverses moments de mes études et de ma carrière. Je pense notamment à Yrjö Gröhn, Philipp Messer et Jérôme Dejardin. J'ajouterai Magalie Fromont pour m'avoir aidé dans ma quête de postdoc et je terminerai par François Petitet-Gosgnach qui est l'enseignant qui m'a le plus marqué dans ma scolarité.

Je remercie tous mes amis qui m'accompagnent depuis des années et qui me permettent d'évoluer positivement. Je pense donc à Guillaume, Léo, Édouard, une team formée au lycée et même avant, puis aux nombreux autres, arrivés depuis, Amaury, Bleuenn, Claire, Bharath, Jocelyn, Maud, Théo, Hugo, Colette. J'ajouterai l'ensemble du Handball de Sorbonne Université, je ne peux citer de noms, n'en donner que quelques uns serait injuste vis-à-vis d'un groupe aussi homogène de gentillesse. J'ai pu reprendre mon sport de prédilection dans une ambiance géniale grâce à vous et pour cela, je vous en remercie. Hâte de trinquer avec vous tous !

Enfin je remercie l'ensemble de ma famille et plus particulièrement mes parents, Monique et Christian, ainsi que ma soeur, Clio, pour le soutien inconditionnel qu'ils m'ont donné tout au long de ma vie. Papa et maman, merci d'être des parents exceptionnels, si j'en suis là aujourd'hui c'est grâce à vous et j'ai pleinement conscience de la chance que j'ai de vous avoir. Clio, merci d'être un modèle de compassion et de persévérance depuis mon plus jeune âge et même bien plus que ça. Je remercie finalement Laura qui partage ma vie depuis plusieurs années, qui m'a aidé à traverser l'ensemble de ma thèse et plus encore, merci d'être à mes côtés !

Résumé

Cette thèse se concentre sur l'estimation de courbes de pénétrance de maladies génétiques à partir de données de pédigrée, avec un intérêt particulier pour la prédisposition génétique au cancer broncho-pulmonaire. Dans ce contexte, elle vise à proposer des résultats à la fois cliniques et épidémiologiques ainsi que des résultats méthodologiques.

Les consultations en génétique sont proposées aux patients ayant des antécédents familiaux sévères de maladies génétiques. Les médecins généticiens doivent sélectionner, parmi ces patients, lesquels se voient proposer un test génétique, ainsi qu'évaluer les risques de maladie pour ces patients et leurs familles. La progression des connaissances en génétique est rapide et le nombre de variants pathogènes identifiés pour différentes maladies augmentent chaque année. Cela entraine un besoin d'outils de prédiction et d'évaluation de risque important, en particulier dans le cadre du cancer broncho-pulmonaire. En effet, les liens entre ce dernier et des variants pathogènes (*SFTPA1*/*SFTPA2* et sur les gènes *TP53* et *EGFR*) sont connus mais encore peu décrits.

Les méthodes existantes pour évaluer le risque de survenue de maladies reposent sur les courbes de pénétrance, mais leur estimation présente des défis en raison du faible nombre de patients et du biais de sélection omniprésent dans les jeux de données collectés en génétique. Pour surmonter ces obstacles, la thèse explore l'utilisation de données familiales, en utilisant un ensemble d'outils statistiques dont les réseaux bayésiens, les modèles de mélange et l'analyse de survie, ainsi que des modèles existants, pour lesquels elle tente d'affaiblir certaines hypothèses.

Le chapitre 1 propose une présentation du contexte médical de la thèse, introduisant les notions de maladies génétiques et de conseil en génétique. Le chapitre 2 est une introduction méthodologique présentant, et illustrant sur des exemples, les concepts d'analyse de survie, de réseaux bayésiens, d'algorithme somme-produit, de modèles de mélanges et d'algorithme EM. Il propose également un état de l'art de l'estimation de courbe de pénétrance pour des maladies génétiques et une mise en évidence du biais de sélection en génétiques. Il se conclue par un récapitulatif des questions de recherche abordées.

Cette thèse s'est, ensuite, orientée autour de quatre projets. Les deux premiers projets, correspondant aux chapitres 3 et 4, proposent des résultats plutôt cliniques et épidémiologiques. La premier projet, décrit au chapitre 3, porte sur une comparaison de différentes méthodes de prédiction de variants pathogènes pour les cancers sein/ovaire (Score de Manchester et modèles familiaux, type *BOADICEA*). Le second projet, abordé au chapitre 4, propose une estimation des pénétrances de la pneumopathie interstitielle et du cancer broncho-pulmonaire pour les porteurs de variants pathogènes *SFTPA1* et *SFTPA2*.

Les deux derniers projets, correspondant aux chapitres 5 et 6, sont plus méthodologiques. Le chapitre 5 est consacré au développement d'une nouvelle méthode d'estimation de courbe de pénétrance de maladie génétique à partir de données de pédigrée lorsque la maladie présente des cas sporadiques. Elle se base sur une contrainte d'incidence de la maladie en population générale et une paramétrisation du ratio de risques instantanés entre les porteurs et les non-porteurs de variants pathogènes. Le chapitre 6 se consacre, lui, à la mise en évidence du biais introduit par la sélection en génétique et ses conséquences sur les résultats de la méthode développée au chapitre 5. Des méthodes de corrections connues, comme la *Proband's phenotype Exclusion Likelihood* (PEL) et la *Genotype-Restricted Likelihood* (GRL), combinées à notre méthode, sont appliquées à des données simulées.

Mots-clés : Analyse de survie, Réseau bayésien, Modèle de mélange, Prédisposition génétique, Cancer broncho-pulmonaire, Cancer sein/ovaire, Biais de sélection (ascertainment), Maladie mendélienne, Algorithme EM, Algorithme somme-produit, Raking.

Abstract

This thesis focuses on estimating penetrance curves of genetic diseases from pedigree data, with a particular interest in the genetic predisposition to bronchopulmonary cancer. In this context, it aims to provide clinical and epidemiological results, as well as methodological findings.

Genetic counselling is offered to patients with severe family histories of genetic diseases. Geneticists must select from these patients who should be offered genetic testing, and assess the disease risks for these patients and their families. Advances in genetics are rapid, and the number of identified pathogenic variants for different diseases increases each year. This necessitates significant predictive and risk assessment tools, especially in the context of bronchopulmonary cancer. Indeed , the links between the latter and pathogenic variants (such that *SFTPA1*/*SFTPA2* and the genes *TP53* and *EGFR*) are known but still poorly described.

Existing methods for assessing the risk of disease occurrence rely on penetrance curves, but their estimation faces challenges due to the small number of patients and the omnipresent selection bias in genetics-collected datasets. To overcome these obstacles, the thesis explores the use of familial data, employing a set of statistical tools including Bayesian networks, mixture models, survival analysis, as well as existing models, for which it attempts to weaken certain assumptions.

Chapter 1 provides an overview of the medical context of the thesis, introducing the concepts of genetic diseases and genetic counseling. Chapter 2 serves as a methodological introduction, presenting and illustrating concepts such as survival analysis, Bayesian networks, sum-product algorithm, mixture models and EM algorithm, using examples. It also offers a state-of-the-art review of penetrance curve estimation for genetic diseases and highlights the selection bias in genetics. The chapter concludes with a summary of the addressed research questions.

Then, the thesis revolves around four projects. The first two projects, corresponding to chapters 3 and 4, offer predominantly clinical and epidemiological results. The first project, described in Chapter 3, compares different methods for predicting pathogenic variants for breast/ovarian cancers (Manchester Score and family models like BOADICEA). The second project, addressed in Chapter 4, provides estimates of penetrance for interstitial lung disease and bronchopulmonary cancer for carriers of pathogenic variants *SFTPA1* and *SFTPA2*.

The last two projects, corresponding to chapters 5 and 6, are more methodological. Chapter 5 is dedicated to developing a new method for estimating the penetrance curve of a genetic disease from pedigree data when the disease presents sporadic cases. It is based on an incidence constraint of the disease in the general population and a parameterization of the relative hazard between carriers and non-carriers of pathogenic variants. Chapter 6 focuses on highlighting the bias introduced by selection in genetics and its consequences on the results of the method developed in Chapter 5. Known correction methods, such as Proband's Phenotype Exclusion Likelihood (PEL) and Genotype-Restricted Likelihood (GRL), combined with our method, are applied to simulated data.

Keywords: Survival analysis, Bayesian network, Mixture model, Genetics predisposition, Lung cancer, Breast/Ovarian cancer, Selection bias (ascertainment), Mendelian disease, EM algorithm, Sum-product algorithm, Raking.

Résumé long

Cette thèse porte sur l'estimation de courbes de pénétrance de maladies génétiques à partir de données de pédigrée et, plus généralement, sur les problématiques de prédiction de prédisposition génétique et de calcul de risques de maladies, pour la pratique clinique en consultation de génétique. Elle est financée par l'Institut des Sciences du Calcul et des Données (ISCD) de Sorbonne Université et ses principaux objectifs sont le calcul de courbes de pénétrance du cancer broncho-pulmonaire pour des individus porteurs de variants pathogènes, notamment sur les gènes *TP53*, *EGFR*, *SFTPA1* et *SFTPA2*.

Les consultations de génétique sont généralement proposées aux patients présentant, dans leur famille, des cas jeunes et/ou nombreux de maladies pour lesquels des prédispositions génétiques sont connues. Lors de ces consultations, les deux principaux objectifs du médecin généticien sont de décider si un patient doit se voir proposer un test génétique et d'évaluer les risques de survenue de la maladie pour le patient et les membres de sa famille. En effet, actuellement, il n'est pas encore possible de tester génétiquement l'ensemble des patients adressés en génétique. Pour les aider dans ces tâches, les médecins généticiens sont demandeurs d'outils précis, robustes et facilement utilisables. Ainsi de nombreuses méthodes statistiques ont été développées pour un certain nombre de maladies et variants pathogènes associés. Par exemple, pour la prédiction des variants pathogènes *BRCA1* et *BRCA2* dans le cadre du cancer sein/ovaire, il existe, entre autre, un score de régression (*Manchester Scoring System*) et un modèle s'appuyant sur les données familiales (*BRCAPRO*). De la même manière, il existe des modèles permettant d'évaluer le risque de cancer sein/ovaire à partir de données familiales, comme *BOADICEA* ou *IBISRisk*. Les prédispositions génétiques au cancer broncho-pulmonaire étant encore méconnues, il existe, pour le moment, peu ou pas d'outils de ce type, ni pour la prédiction de variants pathogènes, ni pour l'évaluation du risque.

Ces outils d'évaluation du risque de maladies se basent, en réalité, sur des courbes de pénétrance. Ces courbes sont issues de l'analyse de survie, un domaine des statistiques s'intéressant au temps avant la survenue d'un évènement d'intérêt, décrit comme une variable aléatoire. Ainsi, pour que les outils d'évaluation du risque soient précis et robustes, il est primordial d'estimer ces courbes de pénétrance. Cependant, l'estimation de ces courbes présente, dans le cadre des maladies génétiques, plusieurs difficultés. Le premier problème est la taille des jeux de données permettant l'estimation des courbes. En effet, pour une maladie et un variant pathogène donnés, le nombre de patients génotypés est généralement faible, ce qui implique des larges intervalles de confiance. Le deuxième problème est le biais de sélection. En effet, les patients sont adressés en consultation de génétique car leur famille vérifie certains critères (cas jeunes/plusieurs cas, etc). Ainsi, les familles porteuses de variants pathogènes vues en consultation ne sont pas représentatives des familles porteuses dans la population générale. Cette sélection implique un biais dans l'estimation des courbes de pénétrance.

Pour répondre à ces problématiques, notamment dans le cadre de maladies ayant une forte composante monogénique présentant une transmission mendélienne, plusieurs méthodes ont été développées. Ainsi, pour augmenter la taille des données utilisables, les biostatisticiens et les généticiens utilisent généralement les données familiales, considérant que si un porteur de variant pathogène a été identifié, il y a, a priori, plusieurs porteurs dans la famille. La famille est modélisée par un réseau bayésien où les génotypes des enfants sont dépendants seulement du génotype de leurs parents et les génotypes des fondateurs de la famille suivent l'équilibre d'Hardy-Weinberg. L'idée est donc d'estimer les courbes de pénétrance grâce à des estimateurs permettant d'inclure des pondérations. Ainsi chaque individu contribue à l'estimation des courbes selon sa probabilité d'être porteur d'un variant. Le modèle

permettant de faire ces estimations est un modèle de mélange de survie/pénétrance. Un certain nombre de méthodes reprennent ce modèle et proposent également diverses stratégies pour corriger le biais de sélection.

Les méthodes publiées dans ce cadre se basent sur un certain nombre d'hypothèses concernant la maladie étudiée et les variants associés. L'ensemble des méthodes existantes considèrent, par exemple, que la fréquence allélique du variant pathogène en population générale est connue. Lorsque la maladie ne présente pas de cas sporadiques (*i.e.* seuls les porteurs de variants pathogènes peuvent être affectés), il existe une méthode d'estimation non-paramétrique mais la majeure partie des maladies génétiques présentent des cas sporadiques. Lorsque la maladie présente des cas sporadiques, des méthodes paramétriques supposent que l'incidence de la maladie est connue en population générale et l'utilisent comme approximation pour l'incidence des non-porteurs de variants pathogènes. Dans le cadre du cancer broncho-pulmonaire, les découvertes de variants pathogènes sont relativement récentes et leurs fréquences alléliques peuvent ne pas être connues. De plus, le cancer broncho-pulmonaire est une maladie présentant un important nombre de cas sporadiques (étant donné que de nombreux facteurs environnementaux favorisent son développement). Ainsi, si l'incidence en population générale est effectivement connue, l'approximation de l'incidence chez les non-porteurs peut être fausse.

Cette thèse s'inscrit dans ce contexte pluridisciplinaire, à l'interface des statistiques et de l'épidémiologie génétique. Elle se propose de répondre à plusieurs problématiques, à la fois médicales et méthodologiques :

- Estimer des courbes de pénétrance de maladies mendéliennes liées au poumon (notamment le cancer broncho-pulmonaire) pour des porteurs de variants pathogènes récemment découverts, grâce aux méthodes existantes.
- Étendre les méthodes existantes d'estimation de maladies mendéliennes présentant des cas sporadiques à partir de données de pédigrée, en relaxant certaines hypothèses.
- Prendre en compte le biais de sélection dans les études de génétiques.

Cette thèse est divisée en six chapitres suivis d'un chapitre de conclusion et perspectives. Les chapitres 1 et 2 constituent l'introduction. Les chapitres 3 et 4 sont des chapitres indépendants proposant des résultats cliniques et épidémiologiques. Les chapitres 5 et 6 sont des projets méthodologiques.

Chapitres 1 & 2

Ces deux chapitres sont consacrés à l'introduction de la thèse. Le chapitre 1 aborde le contexte médical détaillant, dans un premier temps, la notion de maladie génétique puis, dans un second temps, les objectifs et problématiques du conseil en génétique. Le chapitre 2 contient l'ensemble des notions et outils statistiques utilisés dans le reste de la thèse, notamment sur l'analyse de survie, les réseaux bayésiens, l'algorithme somme-produit, les modèles de mélange et l'algorithme EM. Il contient également une section détaillant le modèle de mélange de survie utilisé tout au long de cette thèse et une section mettant en évidence le biais de sélection en génétique. Il se termine par un résumé des questions de recherche abordées dans cette thèse.

Chapitre 3

Ce chapitre est consacré à la question du choix des patients se voyant proposer un test génétique dans le contexte du cancer sein/ovaire.

Les recommandations récentes aux États-Unis préconisent un dépistage généralisé des anomalies génétiques germinales chez les patientes atteintes de cancer sein/ovaire (BC/OC), tandis que les oncogénéticiens européens font face à des contraintes de ressources, nécessitant la sélection des cas à tester. Notre étude, menée au laboratoire d'oncogénétique de l'AP-HP Sorbonne Université à Paris, évalue le score de Manchester (score de régression) adapté aux panels multigènes et aux directives françaises (MSS-F) en tant qu'outil de prise de décision pour la sélection des patients se voyant proposer des tests génétiques.

L'étude a porté sur 1220 patientes atteintes de cancer du sein et d'ovaire de 2016 à 2020. Elle était rétrospective de 2016 à 2017 et prospective de 2018 à 2020, le MSS-F étant utilisé pour guider les décisions de test au cours de cette dernière période de 2 ans. Étonnamment, les performances du MSS-F dans la prédiction des variants pathogènes dans cette population étaient inférieures aux attentes, soulevant deux questions :

- Le MSS-F devrait-il être utilisé plus tôt dans le parcours de soins pour sélectionner les patients qui devraient être adressés aux généticiens ?
- Le MSS-F devrait-il être remplacé par des modèles familiaux tels que *BOADICEA*, *BRCApro* ou *Claus-Easton* pour prédire les variants pathogènes ?

Pour comparer le MSS-F avec les modèles familiaux, les antécédents familiaux de 210 patientes ont été numérisés. Pour imiter une utilisation du MSS-F plus tôt dans le parcours de soins, la méthode du raking a été utilisée afin que la population pondérée corresponde à une population ciblée de patientes non sélectionnées atteintes de cancer sein/ovaire.

D'une part, les résultats ont révélé une faible performance du MSS-F et des modèles familiaux dans la prédiction des variants pathogènes des panels multigènes, avec *BOADICEA* présentant une AUC de 0.61 et 0.57 pour le MSS-F. Même en prédisant uniquement *BRCA1/BRCA2*, les performances restent moyennes. D'autre part, après l'utilisation du raking, l'AUC du MSS-F s'est nettement améliorée à 0.74, suggérant que le MSS-F fonctionne mieux sur une population non sélectionnée de femmes atteintes de cancer sein/ovaire entrant dans le parcours de soin.

En conclusion, dans le contexte du conseil génétique, les modèles familiaux n'ont pas surpassé le MSS-F, ce qui signifie que les outils d'aide à la décision pour la prédiction de variants pathogènes ne sont pas actuellement suffisants pour le cancer sein/ovaire. L'amélioration des performances du MSS-F sur des cas de cancer du sein et d'ovaire non sélectionnés suggère qu'il pourrait être plus efficace en amont pour orienter les patientes vers l'oncogénétique. Par conséquent, l'étude suggère la nécessité d'améliorer l'accès aux tests génétiques germinaux pour tous les cas de cancer du sein et d'ovaire adressés aux généticiens.

Chapitre 4

Ce chapitre est consacré à l'estimation de pénétrance de la pneumopathie interstitielle (ILD) et du cancer broncho-pulmonaire pour des porteurs de variants pathogènes *SFTPA1* et *SFTPA2*.

L'ILD est une affection chronique qui affecte les poumons, provoquant des lésions progressives pouvant conduire à une insuffisance respiratoire au fil du temps. Des recherches récentes ont identifié un lien entre des variants dans les gènes codant les protéines du

surfactant (SP)-A1 (*SFTPA1*) et SP-A2 (*SFTPA2*) et l'ILD ainsi que le cancer du poumon. La pénétrance de ces maladies pour les porteurs de ces variants rares *SFTPA1/2* est encore inconnue, mais cruciale pour le suivi des patients et le conseil génétique. Nous avons identifié ces variants pathogènes dans 27 familles indépendantes où au moins un individu présentait une ILD et/ou un cancer du poumon. L'objectif de notre étude est d'estimer la pénétrance de l'ILD et du cancer du poumon chez les individus porteurs des variants *SFTPA1* ou *SFTPA2*, dont la pathogénicité a été confirmée par des études fonctionnelles in vitro.

Sur la base de pedigrees étendus regroupant 744 individus sur 27 familles parmi lesquels 59 porteurs, des données phénotypiques ont été recueillies auprès de 328 d'entre eux. La pénétrance pour l'ILD et le cancer du poumon a été évaluée en utilisant une méthode existante basée sur un algorithme EM dont l'étape d'espérance est réalisée par un algorithme somme-produit (dans les réseaux bayésiens formés par les arbres généalogiques) et l'étape de maximisation par un estimateur de Kaplan-Meier.

Les résultats montrent une pénétrance du premier événement (ILD ou cancer broncho-pulmonaire) de 50% à l'âge de 60 ans. La pénétrance du premier événement est forte mais incomplète, atteignant 89,3% à l'âge de 80 ans. Le premier événement est le plus souvent l'ILD, le cancer du poumon survient généralement plus tard. La pénétrance du cancer du poumon est moindre que celle de l'ILD mais forte par rapport à la pénétrance du cancer broncho-pulmonaire en population générale avec une pénétrance de 50 % à l'âge de 84 ans et aucun cas avant 30 ans. Cela confirme la pathogénicité des variants *SFTPA1* et *SFTPA2* pour ces deux maladies pulmonaires.

Chapitre 5

Ce chapitre est consacré au développement d'une nouvelle méthode d'estimation de pénétrance pour des maladies génétiques présentant des cas sporadiques à partir de données de pédigrée.

Dans le cadre de maladies génétiques avec faible fréquence allélique en population générale et forte pénétrance (maladies mendéliennes par exemple), les approches familiales sont souvent intéressantes. En effet, les patients ont généralement une famille sévèrement touchée par la maladie et sont donc adressés au généticien. Dans ce contexte , l'estimation du risque de survenue des maladies dépendantes de l'âge (grâce aux courbes de survie/pénétrance) est requise pour la mise en place de protocoles médicaux et le suivi des patients.

Le problème principal pour effectuer ces estimations réside dans le fait que les génotypes sont souvent non-observés et doivent être traités comme une variable latente. Dans le cadre spécifique où la maladie ne présente pas de cas sporadique (*i.e.* seuls les porteurs de variants pathogènes peuvent être affectés par la maladie), le problème est plus simple à traiter car un malade est, de fait, un porteur d'un variant. L'incertitude sur les génotypes repose donc sur les personnes non-affectées. Dans ce scenario, une méthode utilisant des algorithmes d'espérance-maximisation et somme-produit a déjà été publiée.

Cependant, la plupart des maladies affectent à la fois les porteurs et non-porteurs de variants pathogènes à des taux différents. Les méthodes existantes dans ce cas supposent généralement que l'incidence de la maladie est connue dans la population générale ainsi que la proportion de porteurs de mutation. Elles approximent également l'incidence pour les non-porteurs par l'incidence pour la population générale. Cela se rapproche de la réalité dans les cas où la mutation présente une très faible fréquence allélique et une pénétrance très élevée, mais cette hypothèse s'effondre dans des scénarios plus modérés.

La méthode proposée dans ce chapitre vise à généraliser les méthodes précédentes d'estimation de survie des maladies génétiques. Elle repose sur deux hypothèses : l'incidence

de la population générale est constante par morceaux et connue, le risque relatif entre les porteurs et les non-porteurs est également constant par morceaux mais inconnu.

Le modèle est un mélange de survie paramétré par le risque relatif et la proportion de porteurs. À paramètres fixés, les risques instantanés (incidences) des porteurs et des non-porteurs peuvent être calculés sous contrainte du risque instantané en population générale grâce à une méthode de point fixe. Avec les données de pédigrée, la vraisemblance du modèle peut être calculée avec un algorithme somme-produit et ainsi, la vraisemblance étant une fonction d'un nombre de paramètres finis, les paramètres du maximum de vraisemblance sont estimés à l'aide d'un algorithme d'optimisation BFGS.

La méthode a été testée sur 2000 jeux de données simulés, chacun comprenant 744 personnes (reparties sur 28 familles). Les simulations standard suivent le modèle avec une proportion de porteurs de 0.0975, un risque relatif (RH1=20, RH2=10) avec une coupure à l'âge de 50 ans. Une analyse de robustesse est également effectuée où les jeux de données sont générés avec un risque relatif suivant une fonction de Weibull.

La méthode estime sans biais les différents paramètres du modèle sur les exemples proposés. Les intervalles de confiance sont calculé par la méthode de la Hessienne (mais une version bootstrap est également proposée). Les performances de la méthode sont améliorées lorsqu'elle est appliquée sur des jeux de données de taille supérieure et/ou ayant moins de données manquantes.

Chapitre 6

Ce chapitre est consacré à la correction du biais de sélection pour la méthode développée au chapitre 5.

Les données de pédigrée collectées en génétique sont généralement soumises à un biais de sélection important. En effet, les patients sont adressés en conseil génétique selon un certain nombre de critères médicaux (cas jeunes ou nombre de cas important de la maladie dans la famille), ce qui introduit un premier biais de sélection. Ensuite, dans les familles des patients testés génétiquement et effectivement porteurs de variants pathogènes, ce sont les membres ayant la plus forte probabilité d'être porteur qui se voient généralement offrir un test génétique. Ceci induit un second biais de sélection où les individus génotypés sont souvent porteurs de variants pathogènes.

Ces biais de sélection se cumulent et présentent un problème majeur pour toute méthode statistique utilisant ces données et faisant l'hypothèse qu'elles sont représentatives d'une population générale. C'est, entre autres, le cas de la méthode développée au chapitre 5 pour l'estimation de courbe de pénétrance de maladie mendélienne présentant des cas sporadiques à partir de données de pédigrée. En effet, la méthode a été développé et testée sur des données de pédigrée simulées qui ne présentent pas de biais de sélection. D'une part, les familles n'étaient pas sélectionnées sur un critère particulier et, d'autre part, les porteurs et les non-porteurs avaient la même probabilité d'être testés.

Dans ce chapitre, les biais introduits dans les estimations par la méthode développée sont mis en évidence, par des simulations de jeux de données biaisés. Par la suite, plusieurs méthodes de correction de biais de sélection sont présentées et adaptées au contexte des données simulées. L'étude s'intéresse notamment à la *Genoptype-Restricted Likelihood* (GRL) et la *Proband's phenotype Exclusion Likelihood* (PEL) pour les méthodes publiées, ainsi qu'à une correction simple, où la seule hypothèse utilisée est que la fréquence allélique du variant pathogène est connue.

Sur les données simulées, la combinaison de notre méthode avec la GRL semble ne pas fonctionner. Cependant, la simple hypothèse "fréquence allélique connue" permet à notre méthode de corriger une partie du biais de sélection, notamment le biais induit par les tests

génétiques, les résultats présentent cependant toujours un biais. La combinaison de notre méthode avec la correction PEL, toujours sous l'hypothèse "fréquence allélique connue", semble prometteuse. En effet, les résultats présentent un biais inférieur qu'avec l'hypothèse sur la fréquence allélique seule.

En appliquant la combinaison de notre méthode avec la PEL, la courbe de pénétrance du cancer broncho-pulmonaire pour les porteurs de variants pathogènes *SFTPA1* et *SFTPA2* est recalculée.

Listes des valorisations

Articles

Articles soumis :

- Ducrot L, Nathan N, Benusiglio P, Borie R, Nuel G, Legendre M. Penetrance of interstitial lung disease and lung cancer in carriers of *SFTPA1* or *SFTPA2* pathogenic variants. Soumis à PLOS Genetics avril 2024, travail associé au Chapitre 4.

Articles proches d'une soumission :

- Benusiglio P R, Ducrot L, Hasnaoui J, Coulet F, Desseignés C, Canlorbe G, Gueye D, Uzan C, Guillerm E, Nuel G. The Manchester Scoring System in 2022 : performances before and after raking in breast and ovarian cancer patients undergoing multigene panel testing. Le but est de publié un short report dans Journal of Medical Genetics. Le travail a déjà fait l'objet d'un rapport de fin de projet à l'attention de l'INSERM. Travail associé au Chapitre 3.
- Ducrot L, Nuel G. Estimation of penetrance in age-dependent genetic disease with sporadic cases from pedigree data. Journal visé Mathematical Medicine and Biology, travail associé aux Chapitres 5 et 6, disponible sur HAL.

Présentations orales

Titre : Estimation of penetrance in age-dependent genetic disease with sporadic cases from pedigree data. Travail associé au Chapitres 4, 5 et 6.

- Avril 2024, Smilinaire, Collège de France, Paris
- Février 2024, Séminaire MIA, AgroParisTech, Saclay
- Octobre 2023, Journées des Jeunes Probabilistes et Statisticiens, Ile d'Oléron
- Juillet 2023, Journées des Statistiques, Bruxelles
- Mars 2023, Groupe de Travail des Thésards LPSM, Paris

Titre : The Manchester Scoring System in 2022. Travail associé au Chapitre 3.

- Juin 2021, ISCD workshop, Paris

Posters

- Ducrot L, Nuel G. Estimation of penetrance in age-dependent genetic disease with sporadic cases from pedigree data. European Mathematical Genetics Meeting 2023, travail associé au Chapitre 5.
- Benusiglio P R, Ducrot L, Hasnaoui J, Coulet F, Desseignés C, Canlorbe G, Gueye D, Uzan C, Guillerm E, Nuel G. The Manchester Scoring System in 2022 : performances before and after raking in breast and ovarian cancer patients undergoing multigene panel testing. ESHG annual meeting 2022, travail associé au Chapitre 3.

Table des matières

Chapitre 1

Introduction - Contexte médical

1.1 Motivation

Lors de consultations en génétique, les médecins ont généralement deux objectifs principaux à l'égard de leurs patients : déterminer si le patient doit être testé génétiquement sur un panel de gènes prédisposant à certaines maladies et quantifier les risques de survenue de ces maladies à différentes échelles de temps (5 ans, 10 ans, etc). Pour cela, ils disposent d'informations comme l'état clinique du patient, des données sur ses apparentés, son pédigrée et parfois des données de génotypage du patient en question et/ou de certains membres de sa famille.

De nombreux modèles exploitant ce types de données ont vu le jour pour répondre généralement à l'une ou l'autre des deux problématiques. Ces modèles sont souvent spécifiques à une maladie ou un groupe de maladies et à un panel de gènes d'intérêt. Par exemple, pour le cancer sein/ovaire, prenant en compte un certain nombre de mutations pathogènes (BRCA1, BRCA2, etc), le score de Manchester est un outil d'aide à la décision pour déterminer quel patient doit être génotypé, et les modèles familiaux Boadicea, BRCApro ou encore IBIS sont des outils de prédiction du risque de survenue d'un cancer.

Dans ces modèles, les calculs de risques et de probabilité de génotype s'appuient, en fait, sur des courbes de pénétrance/survie des maladies considérées. Pour que ces modèles puissent exister et aient un intérêt clinique, il est donc primordial d'être capable d'estimer ces courbes en amont.

Les généticiens sont confrontés à plusieurs problèmes pour calculer ces courbes de survie. Le premier est le faible nombre de patients généralement génotypés pour une maladie et un panel de gènes donnés ; cela implique notamment d'importantes incertitudes sur le calcul des courbes. Le deuxième problème majeur est le biais de sélection (nommé biais d'*ascertainment* en génétique). Cela se traduit par le fait que les individus présentant des prédispositions génétiques à des maladies, venant en consultation de génétique, ne sont souvent pas représentatifs de ces mêmes individus en population générale (l'ensemble de la population) ; cela implique un biais dans les courbes de survie dont la sévérité est surestimée.

Le but de cette thèse est d'essayer de répondre à ces problématiques en utilisant les données à disposition des généticiens et notamment les pédigrées des patients, c'est-à-dire leur arbre généalogique comportant également les antécédents cliniques des apparentés. Ceux-ci sont, en effet, porteurs d'informations riches et exploitables, à la fois génétiques et cliniques.

1.2 Maladies génétiques

Le but de cette section est de présenter le contexte médical de cette thèse. Il est important de savoir ce que sont les maladies génétiques, quels sont leurs origines et modes de transmission pour mieux cerner les problématiques spécifiques qui seront rencontrées par la suite. Pour cela, sont brièvement présentées les notions, de génotype (gènes et allèles), de prédisposition génétique à une maladie et de pénétrance. Les différentes classes de maladies génétiques sont détaillées avec des exemples concrets de maladies et leurs gènes de prédisposition associés.

1.2.1 Génotype : gènes et allèles

Un être vivant (animal, végétal, unicellulaire, etc) est un organisme qui est composé d'une ou plusieurs cellules, les unités de base de la vie. Ces cellules contiennent dans leur noyau des chromosomes, qui sont des structures constituées d'ADN (acide désoxyribonucléique). L'ADN est une molécule contenant les instructions génétiques qui déterminent les caractéristiques et le fonctionnement de l'organisme. C'est une double chaîne de bases azotées, appelées nucléotides, structurée en hélice. Il existe quatre types de nucléotides : adénine (A), thymine (T), cytosine (C) et guanine (G). L'ADN est donc une suite spécifiquement codée grâce à ces quatre bases azotées. Ainsi, un être vivant est une entité biologique qui utilise l'information génétique stockée dans son ADN (code génétique) pour se développer, se reproduire et fonctionner.

Au coeur de l'ADN, les gènes sont des segments spécifiques de la molécule qui contiennent des instructions pour la synthèse de protéines, les acteurs clés du fonctionnement cellulaire. Chaque gène code pour une ou plusieurs protéines ou fonctions particulières dans l'organisme ; ce qui signifie que les gènes sont responsables, en partie car d'autres mécanismes interviennent également (*i.e.* épigénétique, etc), de la détermination des caractéristiques et des traits héréditaires d'un être vivant. En résumé, les gènes sont les unités fonctionnelles de l'ADN qui dirigent la production de protéines et influencent ainsi le développement et le fonctionnement de l'organisme. La Figure 1.1 montre les différentes échelles du vivant.

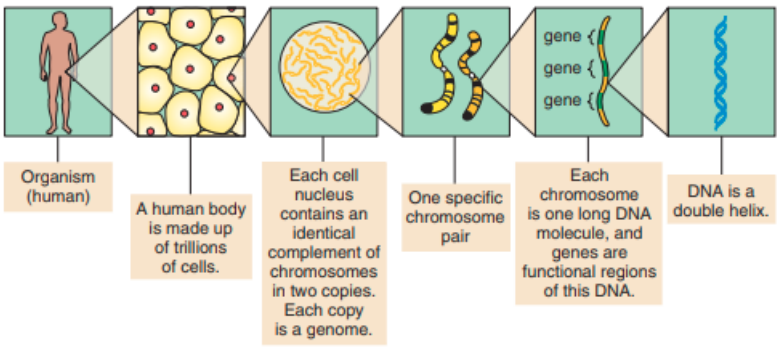


FIGURE 1.1 : Différents niveaux d'observation du vivant, de l'organisme à l'ADN, schéma publié dans *An introduction to genetic analysis* [29]

Chaque gène peut exister sous différentes versions appelées allèles. Dans une population donnée, il y a souvent une version d'allèle qui est considérée comme "normale", courante ou encore commune, et d'autres versions, appelées allèles mutés, qui diffèrent de l'allèle normal par des variations génétiques spécifiques. Ces allèles mutés peuvent avoir un impact sur la fonction du gène et, dans certains cas, être responsables de traits ou de conditions médicales particuliers. Ainsi, les allèles représentent les variations génétiques qui contribuent à la diversité au sein d'une population et jouent un rôle clé dans la génétique de l'hérédité

et des caractéristiques individuelles. Les individus présentant deux allèles identiques pour un gène sont dits homozygotes (pour le gène en question), les autres présentent donc deux allèles différents et sont dits hétérozygotes (voir Figure 1.2).

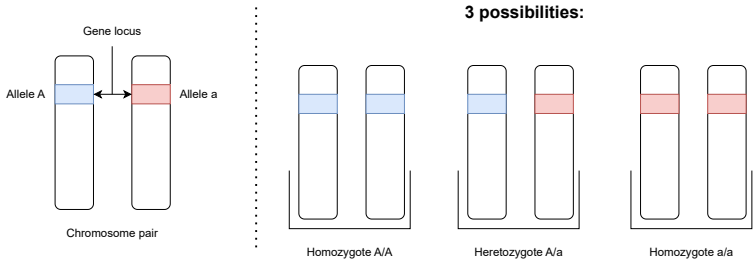


FIGURE 1.2 : Schéma pour un gène et deux allèles A (allèle commun) et a (allèle rare)

La génétique est donc la branche de la biologie qui étudie l'hérédité et l'expression de l'information génétique d'une génération à l'autre.

1.2.2 Phénotype

Phénotype et prédisposition à une maladie

Le phénotype fait référence à l'ensemble des caractéristiques physiques, comportementales et fonctionnelles observables d'un organisme. C'est la manifestation concrète des gènes d'un individu, déterminée par son génotype (l'ensemble de ses allèles). Le phénotype englobe des aspects tels que la couleur des yeux, la taille, la susceptibilité à des maladies, et bien d'autres traits. Il résulte de l'interaction complexe entre les facteurs génétiques et des facteurs environnementaux. En résumé, le phénotype est ce qui est visible, mesurable chez un individu, et il reflète la combinaison de son patrimoine génétique et de son expérience environnementale comme le montre la Figure 1.3.

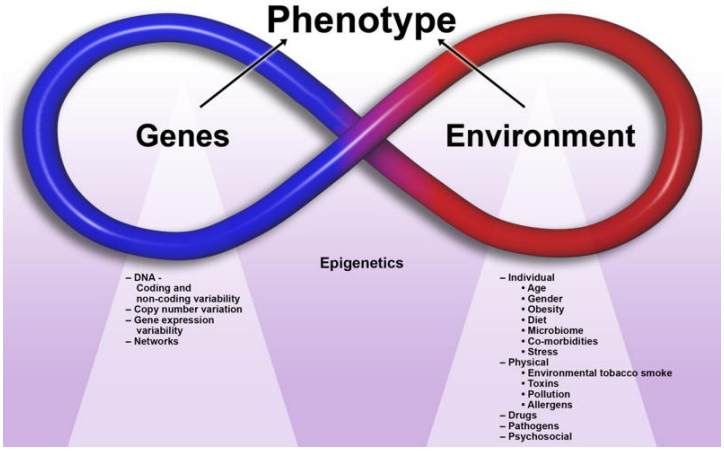


FIGURE 1.3 : Phenotype resultant de l'interaction entre génétique et environnement, illustration de Tesfaye M. Baye [7]

Ainsi, la prédisposition à certaines maladies fait partie intégrante du phénotype d'un individu, car elle influence sa probabilité de développer certains problèmes de santé.

Maladie génétique

Une maladie génétique est donc une maladie dont la survenue est favorisée par des variations génétiques et, possiblement, par des facteurs environnementaux. Il est donc important de noter que, lorsque l'on parle d'une maladie génétique particulière, il s'agit en fait d'un trio composé d'une maladie, d'un ensemble de variants génétiques et d'un ensemble de facteurs d'environnement.

Cette définition est cependant à nuancer d'un point de vue pratique car si la majorité des maladies génétiques résultent de la combinaison de tous ces facteurs, il est parfois difficile d'identifier ces trios, par manque de données, de connaissances scientifiques ou tout simplement car la modélisation est trop complexe pour être exploitable d'un point de vue épidémiologique. Dans les cas complexes, il est parfois nécessaire de simplifier la modélisation qui peut rester néanmoins, toujours informative. Cependant, il existe un certain nombre de cas où l'origine génétique est connue et où la maladie en est une conséquence directe.

Pénétrance

Comme précisé auparavant, une maladie génétique suppose l'existence de variations génétiques et un environnement favorisant sa survenue. Cela ne signifie pas forcément que l'ensemble des individus prédisposés vont être atteints par la maladie. De la même manière, cela ne signifie pas que les individus considérés comme normaux ne peuvent pas être atteints par la maladie. Cette question de la proportion d'individus (prédisposés ou non) allant développer la maladie introduit naturellement le concept de pénétrance qui est décrit dans la section analyse de survie 2. Pour introduire brièvement la notion, la pénétrance d'une maladie dans un groupe d'individus est définie comme le pourcentage d'individus susceptibles d'être affectés par la maladie qui la développent effectivement. Cette pénétrance peut être constante ou être une fonction du temps dans les modèles où l'age au diagnostic est important à prendre en compte.

- Dans le cas où la pénétrance est constante par rapport au temps, il est naturel d'introduire F_0, F_1 et F_2 qui sont respectivement les pénétrances pour les individus homozygotes normaux, hétérozygotes et homozygotes mutés.
- Si on s'intéresse à l'age au diagnostic d'une maladie (modélisé par une variable aléatoire T), la fonction de pénétrance représente la probabilité que la maladie ait été diagnostiquée au temps t :

  $$F(t) = \mathbb{P}(t > T)$$

  De la même manière, il est possible de définir la survie au diagnostic d'une maladie $S(t)$ qui donne son nom à l'analyse de survie et représente la probabilité que la maladie n'ait pas été diagnostiquée au temps t :

  $$S(t) = \mathbb{P}(t < T) = 1 - F(t)$$

  Ces fonctions $F(t)$ et $S(t)$ peuvent, comme dans le cas constant par rapport au temps, être différentes pour les individus homozygotes normaux, hétérozygotes et homozygotes mutés. Il est donc naturel d'introduire des notations similaires comme $F_0(t)$, $F_1(t)$, $F_2(t)$, etc.

La pénétrance peut varier considérablement d'une maladie génétique à une autre et d'un groupe d'individus à un autre (typiquement entre un groupe d'individus prédisposés contre un groupe d'individus considérés normaux). Il est important de définir quelques éléments qui serviront dans la suite du manuscrit.

- Une maladie génétique pour laquelle seuls les individus prédisposés peuvent développer la maladie est dite sans cas sporadiques.
- A l'inverse, si des individus considérés comme normaux peuvent développer la maladie génétique. Ce sont des cas sporadiques.
- Pour les maladies génétiques présentant des cas sporadiques, il est naturel de s'intéresser aux pénétrances, d'une part, du groupe des individus prédisposés et, d'autre part, du groupe des individus dits normaux (typiquement les $F_0(t)$, $F_1(t)$, $F_2(t)$ définis précédemment). Le groupe des individus prédisposés présente, a priori, une pénétrance plus forte que le groupe d'individus normaux.
- Pour les maladies génétiques ne présentant pas de cas sporadiques, par définition, la pénétrance pour le groupe des individus dits normaux est constante et vaut 0.
- Si la pénétrance d'un groupe d'individus vaut 1, cela signifie que l'ensemble des individus du groupe ont développé la maladie, on dit que la maladie présente une pénétrance complète pour ce groupe.
- Certaines maladies génétiques présentent une pénétrance constante. Pour certains syndrômes par exemple, les nouveaux-nés porteurs de mutations présentent ou non la maladie. La pénétrance se calcule alors directement comme le rapport du nombre de nouveaux-nés malades sur le nombre de nouveaux-nés porteurs.
- Les maladies génétiques où les porteurs développent ou non la maladie après un certain nombre d'années, où l'exposition potentielle à certains facteurs fait varier leur chance de la développer, présentent une pénétrance variable au cours du temps. C'est dans ce cadre que l'analyse de survie devient nécessaire.

Incidence

L'incidence d'une maladie en épidémiologie est une mesure qui évalue le nombre de nouveaux cas d'une maladie particulière qui apparaissent au sein d'une population pendant une période donnée. Elle est souvent exprimée en termes de taux d'incidence, c'est-à-dire le nombre de nouveaux cas de la maladie par unité de temps (par exemple par personne-année) rapport à une population donnée. Dans le contexte d'une maladie génétique dont on souhaite comprendre la distribution des ages au diagnostic, l'incidence est donc le nombre de nouveaux cas diagnostiqués par année de vie (à chaque age). En se rapportant à la section sur l'analyse de survie 2, l'incidence d'une maladie génétique est donc la fonction de risque instantané $\lambda(t)$.

Il est fréquent pour les maladies génétiques, lorsque la prévalence de la maladie varie considérablement avec l'âge, de décomposer l'incidence par groupes d'âge. Ce découpage permet alors une modélisation de l'incidence comme constante par morceaux (i.e. un risque instantané constant par morceaux). Ce sont les données que l'on trouve dans les registres médicaux, notamment pour les cancers (par exemple 20-25, 25-30, etc). Cette modélisation de la survie par un risque constant par morceaux, assez naturelle grâce aux données à disposition, est détaillée dans la section analyse de survie.

1.2.3 Classes de maladies génétiques

Comme décrit précédemment, l'expression des maladies génétiques résultent souvent d'une combinaison de variants génétiques et facteurs environnementaux. Ces maladies peuvent, néanmoins, dans un certain nombre de cas ne dépendre que de variants génétiques. Elles

sont généralement reparties dans trois grandes classes en fonction de leur origine : anomalie chromosomique, monogénique et multifactorielles.

Anomalies chromosomiques

Les anomalies chromosomiques, également appelées aberrations chromosomiques, sont des modifications dans la structure, le nombre ou l'organisation des chromosomes d'un individu. Elles peuvent avoir un impact significatif sur le développement, la santé et le fonctionnement d'un individu, et être responsables de diverses maladies génétiques. Elles surviennent généralement lors de la formation des gamètes (ovules et spermatozoïdes) comme illustrée par la Figure 1.4, ou pendant les premières divisions du développement embryonnaire. Quelques exemples d'anomalies chromosomiques connues sont :

- Trisomie : Il s'agit d'une anomalie chromosomique dans laquelle une cellule ou un individu a un chromosome supplémentaire par rapport au nombre normal. Par exemple, la trisomie 21, également connue sous le nom de syndrome de Down, est causée par la présence de trois chromosomes 21 au lieu des deux habituels.
- Monosomie : C'est l'inverse de la trisomie. Un individu présente une perte d'un chromosome. La monosomie X, également connue sous le nom de syndrome de Turner, est un exemple, où les individus ne possèdent qu'un seul chromosome X au lieu des deux (XX).
- Anomalies du nombre de chromosomes sexuels : Ces anomalies affectent les chromosomes sexuels (X et Y) et peuvent donner lieu à des variations du développement sexuel. Par exemple, le syndrome de Klinefelter résulte d'une trisomie XXY.

Maladies d'origine monogénique

Les maladies génétiques d'origine monogénique sont des maladies héréditaires causées par des mutations dans un seul gène. A noter que, chez un individu, la maladie résulte d'une variation sur un seul gène, cependant plusieurs gènes peuvent être associés à la même maladie (*i.e.* cancer sein/ovaire pouvant résulter de mutations BRCA1, BRCA2 ou encore ATM). Ces mutations génétiques spécifiques entraînent des dysfonctionnements ou des défauts dans une protéine particulière, ce qui peut avoir des conséquences sur la santé de l'individu. On distingue les maladies mendéliennes et non-mendéliennes.

- **Maladies mendéliennes**

  Ce sont les maladies d'origine monogénique héritées selon un mode mendélien, où la mutation est transmise de génération en génération conformément aux lois de l'hérédité de Mendel.

  - La loi de la ségrégation : Selon cette loi, chaque individu possède deux copies (allèles) de chaque gène, une héritée de chaque parent. Lors de la formation des gamètes (spermatozoïdes et ovules), les allèles se séparent de manière aléatoire de telle sorte qu'un seul allèle est transmis à la descendance. En d'autres termes, les allèles se séparent (se "ségrègent") lors de la reproduction.
  - La Loi d'indépendance de la transmission des caractères : Cette loi stipule que l'hérédité d'un caractère génétique (comme la couleur d'un fruit) est indépendante de l'hérédité d'un autre caractère génétique (comme la texture du fruit) si les gènes associés à chaque trait sont sur des chromosomes différents ou éloignés

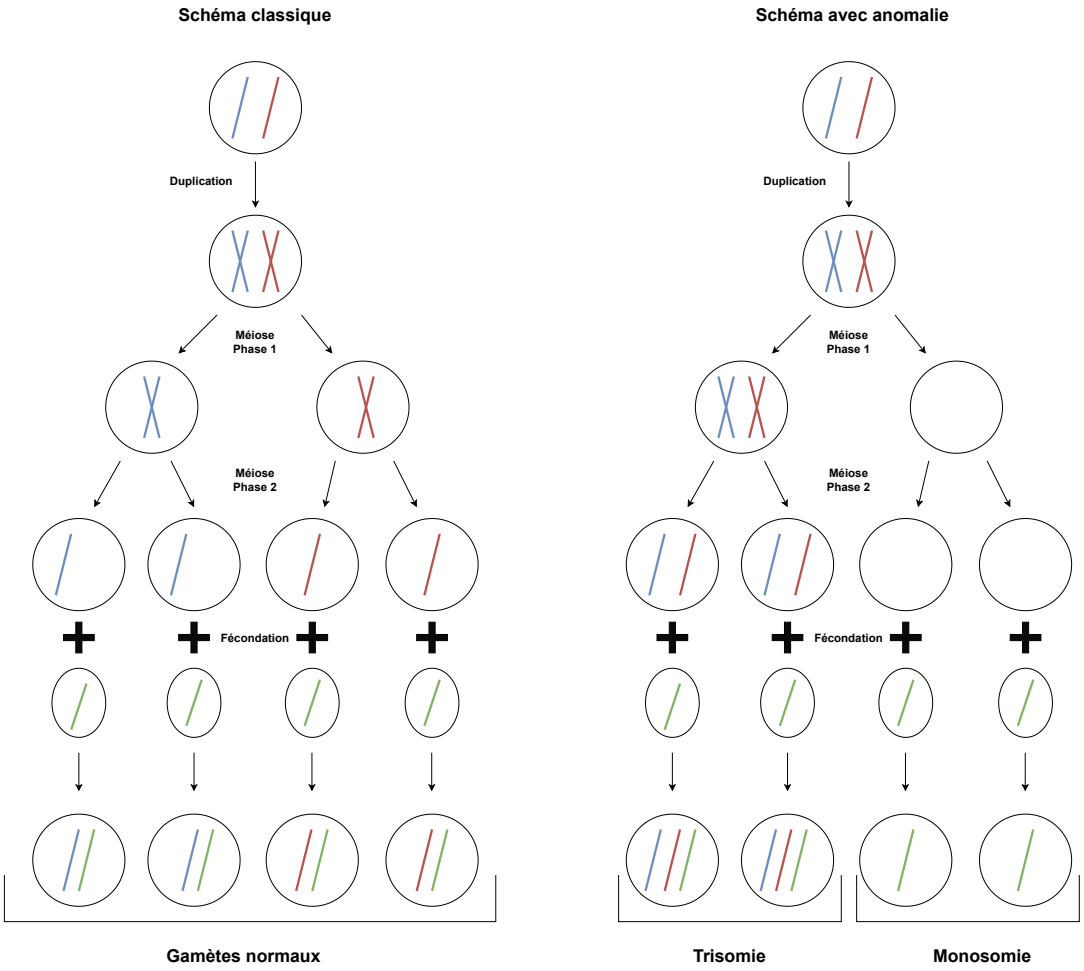


FIGURE 1.4 : Exemple d'apparitions d'anomalies chromosomiques par méiose

sur un même chromosome. Cela signifie que l'héritage de caractères distincts est aléatoire et ne dépend pas les uns des autres.

– La loi de la dominance : dans les cas où un individu hérite de deux allèles différents pour un gène (hétérozygotie), il existe un allèle dominant qui masque l'expression de l'autre allèle dit récessif.

On peut alors s'intéresser au mode de transmission du génotype pour les maladies mendéliennes, pour cela voir les Figures 1.5 et 1.6 :

– La transmission autosomique dominante se manifeste lorsque, sur un chromosome non-sexuel (autosomique), une seule copie du gène muté est suffisante à prédisposer à la maladie (dominante). Cela signifie que les individus hétérozygotes (ayant un allèle pathogène et un allèle normal) sont considérés comme porteurs et à risque. Les individus homozygotes (deux allèles identiques) porteurs d'un allèle pathogène sont généralement rares, voire inexistants, car la combinaison est associée à un phénotype sévère dès la naissance. Les individus hétérozygotes porteurs d'un allèle dominant muté ont 50% de risque de transmettre la prédisposition à chacun de leur enfant.

– Les maladies génétiques à transmission autosomique récessive nécessitent deux copies du gène muté, une provenant de chaque parent, pour se manifester. Les

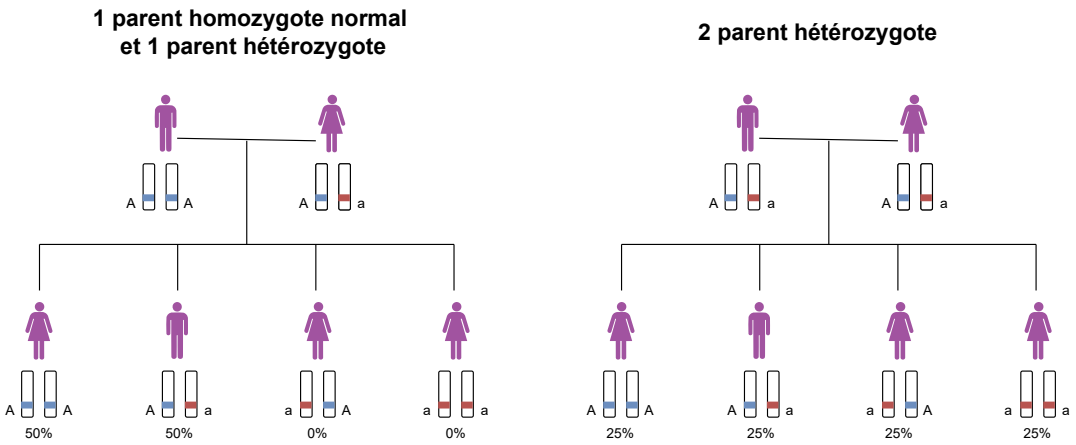


FIGURE 1.5 : Probabilités de transmission d'allèles dans le cas monogénique, allèle normal "A" et allèle muté "a". L'allèle paternel est représenté à gauche et l'allèle maternel à droite.

porteurs, qui ont une copie mutée et une copie normale du gène, ne montrent généralement pas de symptômes. Un exemple est la mucoviscidose, où les parents porteurs (hétérozygotes) n'ont pas la maladie, mais peuvent transmettre la mutation à leur descendance.

– Certaines maladies génétiques sont liées au chromosome X. Les hommes, ayant un chromosome X et un chromosome Y, sont plus susceptibles de manifester des maladies liées au chromosome X, car ils n'ont pas de deuxième chromosome X normal pour compenser une mutation. L'hémophilie est un exemple classique de maladie génétique liée au sexe, affectant principalement les hommes.

Modèle autosomal dominant

Génotypes des gamètes	A	a (dominant)
A	A/A Normal	A/a Prédisposé
a (dominant)	a/A Prédisposé	a/a Prédisposé

Modèle autosomal récessif

Génotypes des gamètes	A	a (récessif)
A	A/A Normal	A/a Normal
a (récessif)	a/A Normal	a/a Prédisposé

FIGURE 1.6 : Phénotype exprimé en fonction des allèles dans les modèles autosomaux récessif et dominant, allèle normal "A" et allèle muté "a".

- **Maladies monogéniques non-mendéliennes**

  Il existe des maladies monogéniques qui ne suivent pas strictement les lois de l'hérédité mendélienne. Elles sont plus complexes et ne se conforment pas parfaitement aux lois de l'hérédité de Mendel en raison de divers facteurs. La principale différence avec les maladies mendéliennes est la variabilité phénotypique importante : cela signifie que la même mutation peut provoquer des symptômes différents chez différents individus (même chez des membres d'une même famille).

Maladies d'origine multifactorielle

Les maladies génétiques d'origine multifactorielle, également appelées maladies complexes, résultent d'une combinaison de facteurs génétiques et environnementaux. Contrairement aux maladies monogéniques, où une mutation dans un seul gène est la principale cause de la maladie, les maladies multifactorielles sont influencées par de nombreux gènes et des facteurs environnementaux comme le montre la Figure 1.7.

Voici quelques caractéristiques de ces maladies :

- Polygénicité : Les maladies d'origine multifactorielle résultent de l'interaction de multiples gènes, chacun contribuant de manière modeste à la susceptibilité à la maladie. La présence de variations génétiques spécifiques dans plusieurs de ces gènes augmente le risque de développer la maladie.
- Facteurs environnementaux : Les facteurs environnementaux jouent un rôle clé dans le déclenchement ou la progression de ces maladies. Cela peut inclure des éléments tels que l'exposition à des toxines, le mode de vie, l'alimentation, le stress, etc.

Il est possible de citer quelques exemples de maladies complexes :

- Diabète de type 2 : le diabète de type 2 est influencé par des variations génétiques qui affectent la sensibilité à l'insuline et le métabolisme du glucose, combinées à des facteurs de risque tels que l'obésité et le mode de vie.
- Maladies auto-immunes : des maladies telles que la polyarthrite rhumatoïde et la maladie de Crohn résultent d'une combinaison de facteurs génétiques et d'expositions environnementales qui déclenchent une réponse auto-immune.
- Cancers : par exemple, le cancer broncho-pulmonaire qui peut résulter d'une combinaison de facteurs génétiques et de facteurs environnementaux (exemples : tabagisme, exposition au Radon).

Ces maladies sont particulièrement complexes à étudier en raison de la multitude de facteurs qui les influencent.

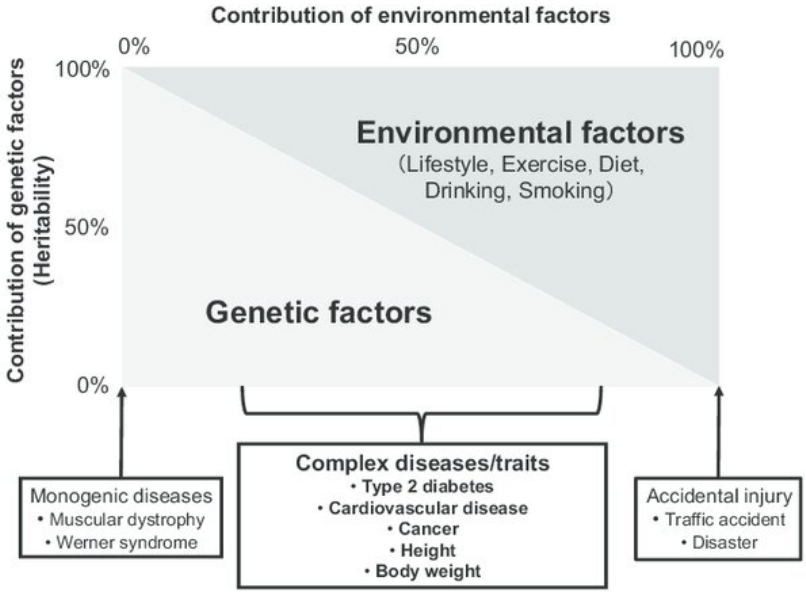


FIGURE 1.7 : Contributions des facteurs génétiques et environnementaux dans les maladies génétiques complexes, fait par Tanisawa [67].

Cette distinction monogénique/multifactorielle est une manière de définir les grandes caractéristiques de chaque maladie génétique. La réalité observée est souvent plus floue, certaines maladies génétiques complexes pouvant avoir une composante monogénique forte, et certaines maladies considérées comme monogéniques pouvant être influencées par des facteurs génétiques et environnementaux multiples. Le cas du cancer broncho-pulmonaire lié à la mutation EGFR est un exemple de maladie complexe dont la composante monogénique est majeure.

1.2.4 Identification des maladies génétiques

Les maladies génétiques ont des origines multiples et il est intéressant de comprendre comment ces mutations et ces facteurs environnementaux sont identifiés en recherche génétique.

Étant donnée la variété des origines des maladies génétiques, leur identification fait appel à diverses méthodes selon leurs caractéristiques (monogénique, polygénique, multifactorielle), notamment certaines méthodes majeures : le *Genome-Wide Association Study* (GWAS) et les scores de risque polygénique pour les maladies multifactorielles, l'analyse de ségrégation et le déséquilibre de liaison pour les maladies plus monogéniques. Cependant, il est important de noter que ces distinctions ne sont pas absolues, et il peut y avoir des chevauchements. Comme expliqué dans la partie présentant les classes de maladies génétiques, certaines maladies complexes peuvent avoir des composantes monogéniques fortes, et certaines maladies monogéniques peuvent être influencées par des facteurs génétiques et environnementaux multiples. Le choix de la méthode dépend donc de la nature spécifique de la maladie étudiée et des objectifs de la recherche génétique.

Analyse de ségrégation et déséquilibre de liaison

Le déséquilibre de liaison fait référence à la non-randomisation des allèles de deux gènes situés sur le même chromosome au cours de la transmission génétique d'une génération à l'autre [61]. En d'autres termes, il s'agit de la tendance de certains allèles de gènes à être transmis ensemble plus fréquemment que, ce à quoi on s'attendrait, si les allèles se séparaient de manière totalement aléatoire. C'est un concept clé dans l'identification de gènes associés à des maladies génétiques ou à des caractères complexes, car il permet de détecter des associations entre gènes et de comprendre comment ils sont hérités ensemble en raison de leur position sur le même chromosome.

Pour les maladies génétiques ayant une forte composante monogénique, le schéma de découverte est principalement le suivant :

- Recueil des données : les généticiens récoltent les informations sur un certain nombre de familles présentant une accumulation de cas d'une maladie particulière. L'information collectée est composée des arbres généalogiques des différentes familles ainsi que les statuts "malades" ou "non malades" des individus.

- Analyse de ségrégation : en utilisant ces données, le but est de vérifier si la maladie suit un potentiel schéma mendélien (monogénique récessif, dominant, etc). L'idée est de tester la validité des modèles et estimer les probabilités de transmission du trait ou de la maladie. Ces analyses statistiques permettent de déterminer la probabilité qu'un individu particulier soit porteur de la mutation causale. Il est important de noter qu'à ce stade, les individus ne sont pas génotypés, seule l'hypothèse d'une origine mendélienne est testée [65].

- Génotypage des familles : si l'hypothèse mendélienne est probable, il est nécessaire de collecter des informations génétiques, généralement des marqueurs génétiques, pour

chaque membre de la famille. Ces marqueurs sont des séquences d'ADN situées à différents endroits du génome et permettent de suivre la transmission des régions chromosomiques d'une génération à l'autre. Les marqueurs sont analysés pour détecter des régions chromosomiques où le gène muté, responsable de la maladie, est susceptible de se trouver. La progression des méthodes de génotypage fait que cette étape est de moins en moins réalisée de cette manière, le séquençage du génome/exome entier étant, maintenant, beaucoup plus simple.

- Analyse de déséquilibre de liaison : une fois les données génétiques recueillies, une analyse des liaisons génétiques est effectuée pour identifier les régions du génome où le gène responsable de la maladie est potentiellement localisé. Cette analyse se base sur la comparaison des marqueurs génétiques avec la distribution de la maladie dans la famille. Les régions chromosomiques qui montrent un déséquilibre de liaison, c'est-à-dire une association significative entre la maladie et des marqueurs spécifiques, sont considérées comme des régions candidates.

- Identification des gènes candidats : les régions chromosomiques présentant un déséquilibre de liaison sont alors explorées pour identifier les gènes candidats. Cela implique de rechercher les gènes situés dans ces régions et qui sont connus pour être associés à la maladie ou qui ont des fonctions biologiques qui suggèrent un rôle potentiel dans la maladie.

- Validation : les gènes candidats identifiés doivent ensuite être validés en étudiant leur rôle dans la maladie à l'aide de techniques moléculaires et cellulaires. Cela peut inclure des analyses fonctionnelles, des études d'expression génique et des études de séquençage pour rechercher des mutations spécifiques, sur des modèles animaliers par exemple (*in vitro* ou *in vivo*).

Genomewide Association Studies

Le *Genomewide Association Studies* (GWAS) se base sur l'étude des polymorphismes nucléotidiques (SNP = *Single Nucleotide Polymorphism*) qui sont des variations du génome sur une seule base de l'ADN qui est alors remplacée par une autre base (substitution) [42, 55]. Le GWAS est principalement utilisé pour l'identification des facteurs génétiques impliqués dans les maladies génétiques d'origine multifactorielle. Ces maladies résultent de l'interaction de multiples gènes et de facteurs environnementaux et le GWAS est particulièrement adapté pour ces maladies. En effet, le principe du GWAS est de tester un très grand nombre (centaines de milliers) de SNPs, en association avec un maladie particulière, chez des milliers de personnes, notamment grâce aux nouvelles banques de données compilant l'ensemble de ces informations. Chaque variant est associé à une augmentation de risque mineure et la combinaison d'un grand nombre de variants est associée à un risque cliniquement significatif de développer la maladie.

1.2.5 Modélisation des maladies génétiques

Modèle mendélien

Les modèles mendéliens sont les modèles utilisés lorsque la maladie génétique en question suit les règles énoncées dans la section sur les maladies mendéliennes. Ils sont donc utilisés pour les maladies monogéniques, suivant des règles de transmission mendélienne récessive ou dominante, sans facteurs environnementaux ni polygéniques.

Modèle à fragilité familiale

Les modèles à fragilité familiale sont une extension des modèles mendéliens [33] qui cherchent à expliquer par une variable cachée, la variation des phénotypes d'une famille à une autre, pour une mutation génétique et une maladie données. Chaque famille se voit attribuer un risque particulier lié à elle-même. Cette méthode est une manière de prendre en compte statistiquement une variable non observée représentant potentiellement une combinaison de facteurs environnementaux et/ou un facteur polygénique familial dans le cadre mendélien. Elle permet de s'éloigner légèrement du cadre mendélien sans perdre la plupart des hypothèses qui lui sont liées.

Score de risque polygénique

Dans le cas des maladies complexes, résultant de combinaisons de mutations sur différents gènes, le score de risque polygénique (PRS = *Polygenic Risk Score*) est un outil permettant de représenter le risque de survenue de la maladie chez un individu donné [14].

Le principe des scores de risque polygénique est d'utiliser les SNPs, identifiés comme associés à une maladie (généralement détectés par GWAS), pour calculer un risque de développer la maladie pour des nouveaux patients. Un score est attribué à chaque SNP en fonction de son association statistique avec la maladie. Les SNPs qui sont plus fortement corrélés avec la maladie reçoivent un score plus élevé, tandis que ceux avec une corrélation plus faible reçoivent un score plus bas. Ainsi, lorsqu'un nouvel individu est génotypé, la somme des scores associés à ses propres SNPs lui confère un score global lié au risque de développer la maladie. BOADICEA [39] inclut dans son calcul du risque des cancers sein/ovaire une part polygénique.

Il est important de noter qu'il n'y a pas de notions de causalité entre les SNPs identifiés et la maladie. Cependant les SNPs trouvés ainsi peuvent identifier de bons gènes candidats grâce à des analyses de déséquilibre de liaison.

1.2.6 Maladies génétiques étudiées

Dans cette thèse, plusieurs maladies génétiques sont abordées dans les différents projets développés. Le but de cette section est de donner quelques informations sur les maladies, les mutations génétiques associées et les potentiels facteurs de risque pour une meilleure compréhension des chapitres suivants. Les exemples présentés sont des cancers car il est estimé qu'entre 5 et 10% des cancers en France seraient liés à la présence d'une altération génétique héréditaire d'après l'Institut National du Cancer (INCA 2020). Cependant bien d'autres maladies présentent des origines génétiques connues.

Cancer sein/ovaire

Le cancer du sein est un des cancers les plus courants en France chez les femmes, en 2020 environ 58 000 cas ont été recensés. Le cancer des ovaires est plus rare mais a touché, toujours en 2020, environ 5 000 femmes [66]. Les cancers sein/ovaire sont deux types de cancer qui ont des liens génétiques importants. Les mutations dans les gènes BRCA1 et BRCA2 sont les mutations génétiques les plus connues, associées à ces deux cancers, et sont héritées de manière autosomique dominante. Un certain nombre d'autres gènes ont également été identifiés comme étant des facteurs de prédisposition dont PALB2, TP53, CDH1, PTEN, RAD51C, RAD51D, MLH1, MSH2, MSH6, PMS2 et EPCAM [30]. Il est estimé qu'environ 5% des cancers du sein seraient liés à une prédisposition génétique (Fondation ARC).

Le facteur de risque le plus important est le sexe, pour le cancer des ovaires évidemment (puisque les hommes n'y sont pas sujets) mais également pour le cancer du sein (entre

0.5 et 1% des cancers du sein touchent des hommes). Chez la femme, l'histoire hormonale (apparition des premières règles, âge à la ménopause etc) est un des facteurs principaux du risque de développement de la maladie.

Cancer broncho-pulmonaire

Le cancer broncho-pulmonaire, également connu sous le nom de cancer du poumon, est l'un des types de cancer les plus répandus en France, avec 48 000 cas en 2020 [66]. Il existe deux principaux types de cancer du poumon : le carcinome non à petites cellules (CNPC) et le carcinome à petites cellules (CPC). Le CNPC est le type le plus courant, tandis que le CPC est généralement plus agressif.

Un certain nombre de facteurs environnementaux sont reconnus comme facteurs de risque pour le cancer du poumon notamment le tabagisme, les expositions à l'amiante, à la pollution de l'air, au radon.

Il existe également un certain nombre de facteurs génétiques, commençant à être décrits, comme les mutations sur les gènes *EGFR* et *TP53* (syndrôme de Li-Fraumeni), ou encore *SFTPA1/SFPTA2* notamment étudiés dans cette thèse. Actuellement, 1 à 2 % des adénocarcinomes sont identifiés comme liés à une prédisposition génétique [51], ce qui pourrait représenter plusieurs centaines de cas par an en France. Cependant, le domaine est encore méconnu et des recherches génétiques sont rarement lancées.

Le caractère fortement monogénique des cancers broncho-pulmonaires liés à certains variants comme *EGFR*, ainsi que le manque d'études à leur sujet font de cette maladie la motivation principale de cette thèse.

1.2.7 Focalisation de la thèse

Cette section a donc exposé une vision d'ensemble de la génétique médicale, des différentes origines et méthodes d'étude des maladies génétiques ainsi que quelques exemples comme le résume la Figure 1.8.

Cependant cette thèse va s'intéresser à un type bien spécifique de maladies génétiques : les maladies d'origine monogénique à mode de transmission mendélienne présentant ou non des cas sporadiques. Plus précisément, les maladies que nous étudions sont des maladies se développant au cours du temps, donc à pénétrance non constante, ce qui implique l'utilisation de méthodes d'analyse de survie.

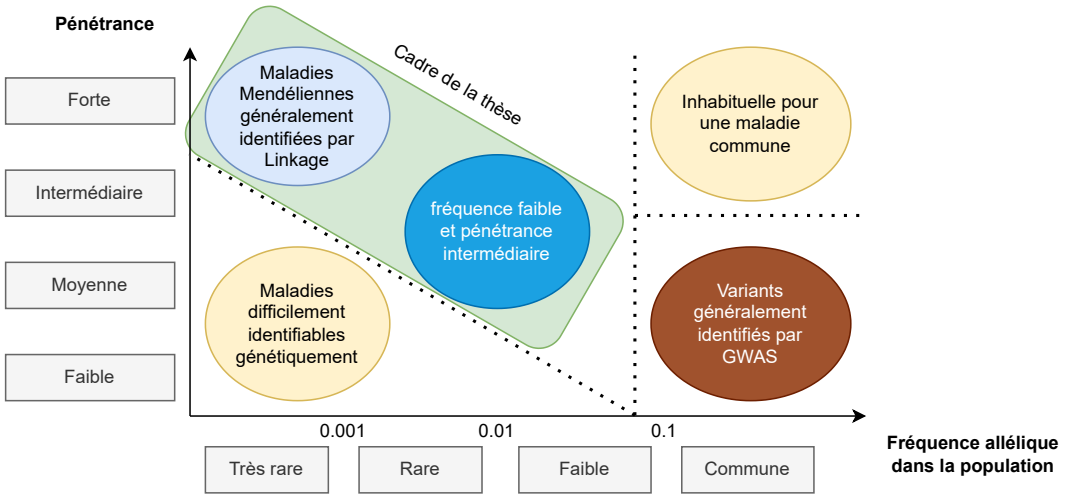


FIGURE 1.8 : Diagrame présentant les maladies génétiques en fonction de leur fréquence allélique et de leur pénétrance. Adapté d'un schéma fait par McCarthy [43].

1.3 Conseil génétique

Lorsqu'un patient fait partie d'une famille présentant un certain nombre de cas d'une maladie particulière, il est souvent invité à faire une consultation de génétique. Il est appelé le *proband* (il est le point d'entrée de la famille dans le parcours de soin génétique). Cette consultation avec un médecin généticien a plusieurs objectifs qui seront décrits dans la suite de cette section. Il est important de comprendre les enjeux de ces consultations car elles sont les bases de la récolte de données utiles à la fois pour la recherche en génétique et pour l'amélioration des parcours de dépistage/soin des patients présentant des prédispositions génétiques.

1.3.1 Données

Lorsqu'un *proband* arrive en consultation de génétique, le médecin collecte un certain nombre de données. Ces dernières sont reparties dans trois grandes classes : les données cliniques, les données de pédigrées et les données génétiques.

- données cliniques : ce sont l'ensemble des données relatives à la maladie et à l'état du patient. Cela comprend les antécédents médicaux, les examens cliniques, des informations spécifiques à la maladie (par exemple type histologique pour un cancer), les habitudes et facteurs de risque, etc.
- données de pédigrées : ce sont l'ensemble des données relatives à la famille du patient, comme la structure familiale (arbre généalogique) et les antécédents familiaux de la maladie. Le but est d'avoir un maximum d'informations sur chaque membre de la famille notamment les facteurs de risque connus, les ages et ages au diagnostic de la maladie, si décès la cause, etc.
- données génétiques : ce sont l'ensemble des données relatives à la génétique, notamment tout test de dépistage déjà effectué pour des panels de gènes connus, chez le patient ou des membres de sa famille.

Les données de pédigrées et génétiques peuvent être facilement représentées sous forme d'arbres généalogiques annotés (voir Figure 1.9), ce qui donne une vision plus globale de la

famille. Cette représentation permet de supputer, avant même tout calcul, si une branche particulière de la famille semble plus touchée. Elle permet, également, d'identifier un ou des membres pouvant être considérés comme centraux et potentiellement porteurs de beaucoup d'informations, qu'il serait intéressant de recevoir en consultation également.

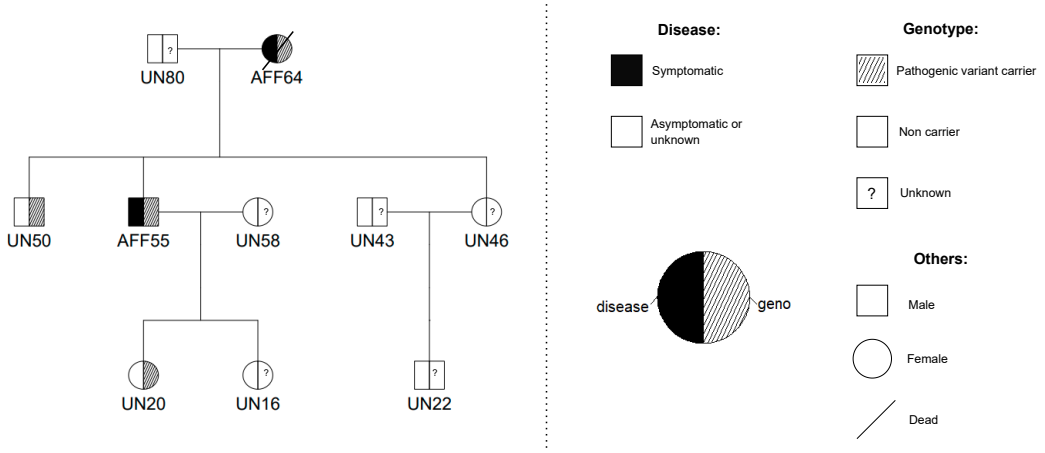


FIGURE 1.9 : Pédigrée et sa légende.

1.3.2 Objectifs

Les objectifs du conseil en génétique sont multiples, allant du choix de proposer un test génétique, au soutien psychologique pour les patients, en passant par l'organisation du plan de suivi si nécessaire, ou du choix de traitement adapté. L'ensemble de ces objectifs ne sont pas détaillés ici, à l'exception de deux objectifs primordiaux qui sont centraux à cette thèse : la décision de tester génétiquement le *proband* et l'évaluation du risque de maladie pour le *proband* et les membres de sa famille.

Décision de test génétique pour le *proband*

Lorsqu'un patient arrive pour la première fois en consultation de génétique, il est souvent adressé car sa famille présente une accumulation de cas d'une maladie particulière. A la vue des données dont il dispose, détaillées dans la section précédente - généralement les données cliniques sur le patient mais surtout l'histoire familiale liée à la maladie (ages au diagnostic des apparentés, arbre généalogique/liens familiaux, facteurs de risque, etc) - la première question que le généticien se pose est de savoir s'il doit proposer un dépistage génétique au patient. A l'heure actuelle, les génotypages coûtent de moins en moins chers et sont de plus en plus rapides mais il reste impossible, dans un contexte de contraintes budgétaires et de ressources, de tester tous les patients arrivant en consultation de génétique. Le généticien doit donc faire un choix justifié par des recommandations médicales.

Ces recommandations varient souvent d'un hôpital ou d'un pays à l'autre. C'est pourquoi, un certain nombre de modèles ont été développés pour essayer de proposer des aides à la décision dans ce cadre spécifique. Étant donnée la problématique (dépistage ou non), les modèles proposés sont souvent des régressions logistiques, faites sur une grande collection de données, comme le score de Manchester dans le cadre du cancer sein/ovaire et des mutations BRCA1/BRCA2 [25], ou des ensembles de critères cliniques comme les critères d'Amsterdam pour le syndrome de Lynch [73].

L'avantage principal de ces méthodes est leur simplicité d'utilisation dans la pratique clinique car il suffit d'avoir accès à des informations généralement présentes dans les dossiers cliniques des patients et le résultat se calcule instantanément. Si les critères sont vérifiés ou

si, pour les régressions logistiques, le score est au dessus d'un certain seuil, alors le choix du dépistage est automatique. Le généticien peut toujours dans les cas limites, arbitrer en faveur d'un dépistage.

Il existe aussi des méthodes familiales pour la détection de variants pathogènes comme BRCApro [9][50] pour les mutations BRCA1 et BRCA2 liées au cancer sein/ovaire. Ces modèles s'appuient sur la transmission familiale d'allèles pathogènes et s'avèrent puissant mais généralement moins pratiques à utiliser en consultation. Les modèles familiaux comme BOADICEA [5] et le modèle de Claus-Easton [28] ont été développés initialement comme des outils d'évaluation du risque de cancer sein/ovaire (comme discuté dans la section suivante). Ils peuvent néanmoins être utilisés pour la prédiction de mutations pathogènes également, sachant qu'ils calculent des probabilités d'être porteur de mutations pour l'estimation du risque.

Évaluation du risque de maladie

Lorsqu'un patient a été dépisté et présente effectivement une prédisposition génétique à une ou plusieurs maladies, le deuxième objectif principal du généticien est d'évaluer le risque de développement de la maladie pour le patient en question ainsi que les membres de sa famille. Le but est de proposer un parcours de soin et de prévention adapté à chaque individu en fonction de son propre risque de développer la maladie. Cette démarche s'inscrit toujours dans une balance subtile entre réussir une détection précoce de la maladie (ce qui améliore dans la majorité des cas les taux de guérison) et éviter un excès de prise en charge qui aurait deux effets : un coût important et une potentielle augmentation d'un risque de pathologie supplémentaire pour les personnes ayant un suivi important. En effet, si la prévention inclut des radiographies fréquentes par exemple, cela impliquerait une forte exposition à des radiations pour le patient.

Pour calculer ces risques de développer la maladie, les généticiens ont plusieurs outils à disposition. Le premier est l'utilisation des courbes de survie/pénétrance associées à la maladie et au variant génétique en question. Si ces courbes sont disponibles, il est possible d'évaluer le risque de maladie pour le patient. Cependant il est primordial de pouvoir l'estimer également pour les membres de sa famille. C'est pourquoi, un grand nombre de modèles familiaux ont été développés pour calculer le risque de maladies génétiques, notamment les plus courantes, comme le cancer sein/ovaire pour lequel il existe BOADICEA [5], IBIS [27] et le modèle de Claus-Easton [28] par exemple, ou pour le syndrome de Lynch avec MMRpro [13]. Ces modèles se basent sur des courbes de survie connues, la transmission d'allèles souvent de manière mendélienne et peuvent prendre en compte des fragilités familiales (pour tenir compte de composantes polygéniques ou environnementales non connues).

Ces modèles utilisent en variables d'entrée des données de pédigrées et calculent, à la fois, des probabilités a posteriori d'être porteur de mutations et un risque de développer la maladie à 5 ans, 10 ans etc, pour l'ensemble des membres de la famille. C'est avec l'aide de ces résultats que le médecin prend des décisions de prévention, de suivi et de soin.

1.3.3 Exemples de méthodes pour le cancer sein/ovaire

Score de Manchester

Le score de Manchester (*Manchester Scoring System*) est un outil d'aide à la décision dans le cadre du cancer sein/ovaire, développé à l'Université de Manchester, au Royaume-Uni, pour aider les professionnels de la santé à identifier les femmes qui pourraient bénéficier de conseils génétiques et de tests génétiques [25]. Le score est basé sur une régression logistique sur plus de 4000 familles qui prédit la présence de mutations *BRCA1* et *BRCA2* à partir de 22 variables comprenant des informations de cancers du patient et de ses apparentés et des

informations sur l'histologie du cancer sein/ovaire du patient. Ces variables sont présentées dans le tableau 1.10 Chaque variable est pondérée et le score final est la somme pondérée de toutes ces variables. Si l'individu présente un score supérieur à 15, il présente une probabilité d'être porteur de mutations *BRCA1* ou *BCRA2* supérieure à 10 % d'après l'étude.

Table 1 The Manchester Scoring System with pathology adjustment[9 15 16]

Cancer, age at diagnosis	*BRCA1*	*BRCA2*
FBC, <30	6	5
FBC, 30–39	4	4
FBC, 40–49	3	3
FBC, 50–59	2	2
FBC, >59	1	1
MBC, <60	5	8
MBC, >59	5	5
Ovarian cancer, <60	8	5
Ovarian cancer, >59	5	5
Pancreatic cancer	0	1
Prostate cancer, <60	0	2
Prostate cancer, >59	0	1
Breast cancer path adjustment in index case		
Grade 3	2	0
Grade 1	−2	0
ER positive	−1	0
ER negative	1	0
Triple negative	4	0
HER2+	−6	0
Ductal carcinoma in situ	−2	0
Lobular	−2	0
Ovarian cancer adjustment – any case in family*		
Mucinous germ cell or borderline tumours	No score, that is, score as 0	No score, that is, score as 0
High grade serous <60	+2	0
Adopted no known status in blood relatives	+2	+2

FIGURE 1.10 : Manchester Scoring System table from the 2017 reviewed publication [25]. F for female, M for male and BC for breast cancer, the following number is an age threshold.

BOADICEA

Le modèle BOADICEA (*Breast and Ovarian Analysis of Disease Incidence and Carrier Estimation Algorithm*) est un modèle de prédiction de risque de cancer sein/ovaire à partir de données de pédigrées [5]. Le modèle incorpore les effets connus des variants pathogènes *BRCA1*, *BRCA2*, *CHECK2* et *ATM* ; un score de risque polygénique basé sur 313 SNPs expliquant, d'après l'article, la variance polygénique des cancers du sein ; une variable supplémentaire de type fragilité familiale pour tenir compte d'autres effets familiaux ou génétiques inconnus ; des variables supplémentaires cliniques et environnementales (comme le style de vie, densité mammographique, prise ou dosage hormonal etc).

Le modèle a été développé à l'université de Cambridge et se base principalement sur une population britannique. Un outil numérique est disponible en ligne (CanRisk [11][6]) et permet l'utilisation rapide du modèle à la portée de tous mais principalement pour les professionnels de santé. Cet outil est très intéressant pour le conseil en génétique car s'il requière plus de temps et d'informations qu'un score de régression, il pourrait tout de même être utilisable en consultation.

BRCAPRO

Le modèle BRCApro [50] est également un modèle calculant la probabilité pour un individu d'être porteur d'un variant pathogène BRCA1 ou BRCA2 à partir de données familiales de cancer sein/ovaire chez les apparentés au premier et deuxième degrés.

Le modèle est disponible dans un package R BayesMendel [12], ce qui le rend plus difficile d'utilisation en clinique car il requiert des connaissances en codage pour être effectivement utilisé.

Modèle de Claus-Easton

Le modèle de Claus-Easton [28] est un modèle de prédiction du risque de cancer sein/ovaire. Il a été construit à la suite d'une série d'articles par Claus et Easton sur la modélisation mendélienne du risque associé aux variants pathogènes BRCA1 puis BRCA2 [21][20]. Il se base sur des calculs de risques cumulés pour le cancer du sein, notamment décrits dans la Figure 1.11.

Le modèle ne présente pas d'implémentation standard et il faut donc l'implémenter soi-même pour l'utiliser.

Age (Years)	Cumulative Probability for Genotype		
	AA	Aa	aa
20–29	.0167	.0167	.0002
30–39	.1444	.1444	.0027
40–49	.3758	.3758	.0138
50–59	.5477	.5477	.0275
60–69	.6743	.6743	.0497
70–79	.9452	.9452	.0798
80+	1.0000	1.0000	.1254

Figure 1.11 : Risques cumulés de cancer du sein par tranche d'âge pour les différents génotypes AA, Aa et aa [15]

1.3.4 Besoins et recherche

Comme expliqué précédemment, les généticiens doivent à la fois pouvoir sélectionner les patients à qui un test génétique est proposé mais également calculer les risques de développement de la maladie à différentes échelles de temps pour les patients et leur famille. Certes, de plus en plus d'outils d'aide à la décision sont développés mais les nouvelles connaissances ne cessent de mettre en évidence de nouvelles maladies génétiques (nouveaux variants pour des maladies connues notamment) et les médecins sont demandeurs de méthodes toujours plus fiables et faciles d'utilisation. Il est donc primordial de proposer de nouveaux outils.

L'ensemble des modèles familiaux qui proposent à la fois un calcul de la probabilité d'être porteur de mutations et un calcul du risque de maladie, reposent sur la connaissance de courbes de survie/pénétrance des maladies associées aux variants pathogènes. L'estimation des courbes de survie/pénétrance est donc un enjeu majeur pour la recherche et pour la consultation génétique.

Dans le chapitre suivant, l'ensemble des notions statistiques importantes à l'estimation de courbes de survie dans le cadre particulier des maladies mendéliennes à partir de données de pédigrées seront présentées.

Chapitre 2

Introduction - Statistiques

2.1 Préambule

Le but de ce chapitre est d'introduire l'ensemble des notions statistiques qui seront utiles dans la suite de ce manuscrit. Dans un premier temps, les notions abordées sont :

- l'analyse de survie qui est le cadre statistique classique pour la modélisation du risque de survenue d'un évènement au cours du temps (comme une maladie par exemple) ;
- les réseaux bayésiens qui sont des modèles graphiques probabilistes et permettent de représenter les liens familiaux (notamment pour la transmission génétique de parents à enfants) ;
- les modèles de mélanges et l'algorithme EM (Espérance-Maximisation) qui permettent respectivement de modéliser des populations mixtes (individus porteurs et non-porteurs de prédisposition génétique par exemple) et de calculer des maximums de vraisemblance dans ce type de modèles lorsque les classes de individus sont latentes.

Ensuite, dans un second temps, ce chapitre aborde les notions spécifiques à cette thèse :

- le modèle de base d'estimation de courbe de survie pour des maladies génétiques à partir de données de pédigrée, ainsi que l'état de l'art sur ce modèle ;
- la notion de biais de sélection (*ascertainment*) en génétique.

Enfin le chapitre détaille les questions de recherche auxquelles cette thèse aborde.

2.2 Analyse de survie

L'analyse de survie est une branche importante de la statistique qui traite de la modélisation et de l'analyse du temps jusqu'à l'occurrence d'un événement spécifique et, dans le cadre de cette thèse, le temps avant le diagnostic d'une maladie. Le but de cette partie est d'exposer brièvement plusieurs concepts de la survie [35] qui seront utilisés tout au long du manuscrit.

2.2.1 Définition

Le temps de survie, noté T, est une variable aléatoire qui représente la période écoulée jusqu'à l'occurrence d'un événement d'intérêt. Par exemple, en médecine, T peut représenter le temps jusqu'au diagnostic d'une maladie. Ce sera le cas dans toute la suite de ce manuscrit. Étant donné que T est une variable aléatoire, plusieurs fonctions d'intérêt peuvent être définies.

- La fonction de répartition $F(t)$ donne la probabilité que l'évènement se soit déjà produit à l'instant t :

$$F(t) = \mathbb{P}(T \leqslant t)$$

  Dans le reste du manuscrit, sachant que l'évènement d'intérêt est le diagnostic d'une maladie, cette fonction sera nommée la pénétrance de la maladie.

- La function de survie $S(t)$ donne la probabilité que l'évènement ne se soit pas produit à l'instant t :

$$S(t) = \mathbb{P}(T > t)$$

  Elle est directement liée à la pénétrance puisque $S(t) = 1 - F(t)$.

- Le risque instantané $\lambda(t)$ mesure la probabilité que l'évènement se produise au cours d'un temps infinitésimal dt sachant qu'il ne s'est pas encore produit au temps t :

$$\lambda(t) = \lim_{dt \to 0} \frac{\mathbb{P}(t < T < t + dt | T > t)}{dt}$$

- La fonction de densité $f(t)$ mesure la probabilité que l'évènement se produise au cours d'un temps infinitésimal dt :

$$f(t) = \lim_{dt \to 0} \frac{\mathbb{P}(t < T < t + dt)}{dt}$$

  La densité est directement liée à la survie et au risque instantané car $f(t) = S(t)\lambda(t)$.

- Le risque cumulé $\Lambda(t)$ tel que :

$$\Lambda(t) = \int_0^t \lambda(u)du$$

Il est important de noter qu'il existe un lien entre la fonction de survie $S(t)$ et le risque instantané $\lambda(t)$ puisque :

$$S(t) = \exp(-\int_0^t \lambda(u)du) = \exp(-\Lambda(t))$$

2.2.2 Données de survie, notion de censure

La censure est un concept fondamental en analyse de survie et réfère à la situation où l'on ne dispose pas d'informations complètes sur le temps de survie d'un individu dans une étude. Cela se produit lorsque le suivi d'un individu est arrêté avant l'occurrence de l'événement d'intérêt (par exemple, le décès d'un patient pour une raison indépendante de la maladie étudiée), ou lorsque la période de suivi est limitée et que certains individus n'ont pas encore subi l'événement à la fin de la période de suivi. La censure est généralement classée en trois types principaux :

- Censure à droite : il s'agit de la situation la plus courante en analyse de survie. La censure à droite se produit lorsque le temps de survie d'un individu n'est pas complètement observé, car le suivi s'arrête avant que l'événement se produise. Cela signifie que le temps observé est une borne inférieure pour du temps de survie.

- Censure à gauche : la censure à gauche se produit lorsque le temps de survie d'un individu n'est pas complètement observé parce que le suivi débute après que l'événement d'intérêt s'est déjà produit. Dans ce cas, le temps observé est une borne supérieure pour le temps de survie.

- Censure par intervalle : la censure par intervalle se produit lorsque les moments d'étude sont discrets (par exemple, chaque rendez-vous chez le dentiste) et que l'événement d'intérêt se produit entre deux instants (par exemple, une carie est apparue), le temps de survenue de l'événement est alors compris entre ces deux instants mais n'est pas connu.

La censure est un défi majeur en analyse de survie, car elle peut biaiser les estimations si elle n'est pas correctement prise en compte. Cependant, des méthodes statistiques telles que l'estimateur de Kaplan-Meier sont spécialement conçues pour traiter les données censurées et permettent d'estimer la fonction de survie.

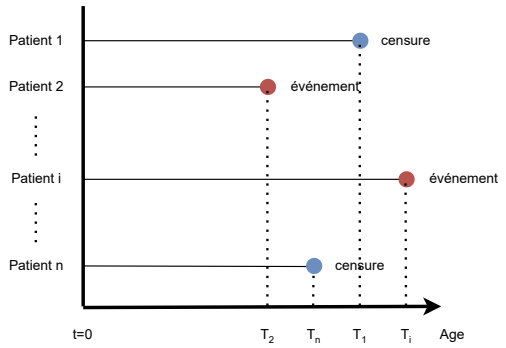
Censure à droite

Dans la suite de ce manuscrit, la censure observée est une censure à droite. En effet, les temps de diagnostic présentent ce type de censure, un patient présente ou non la maladie, dans ce dernier cas, l'événement d'intérêt ne s'est pas encore produit et la seule information disponible est la suivante : à un certain age, l'individu n'est pas affecté par la maladie.

D'un point de vue mathématique, la censure à droite peut s'écrire ainsi : soit $\tilde{T}$ le temps de survie avant le diagnostic d'une maladie et C le temps de censure, alors T le temps observé et δ le statut (diagnostiqué ou non) sont définis comme suit.

$$\begin{cases} T = \min(\tilde{T}, C) \\ \delta = \mathbb{1}_{\tilde{T} \leqslant C} \end{cases}$$

Les données de survie observées sont donc T et δ comme montrées sur la Figure 2.1. Ainsi, si n patients sont dans une étude, les données sont recueillies sous forme de table où chaque ligne représente un patient. Pour un patient i ($i \in \{1, .., n\}$), son age est donc T_i et son statut (diagnostiqué ou non) $\delta_i \in \{0, 1\}$. Pour pouvoir faire des estimations en prenant en compte la censure, il est souvent important de vérifier que $\tilde{T}$ et C sont indépendantes.



	T	δ
Patient 1	T_1	δ_1
Patient 2	T_2	δ_2
...	...	...
Patient i	T_i	δ_i
...	...	...
Patient n	T_n	δ_n

FIGURE 2.1 : Exemple de données de survie.

2.2.3 Estimation de la fonction de survie

A partir de ces données de survie, il est naturel de vouloir estimer les fonctions associées à la loi ou aux lois de survie qu'elles suivent. Pour cela, il existe un ensemble de méthodes permettant de faire soit de l'estimation paramétrique (en supposant que les fonctions de survie prennent des formes spécifiques paramétrées), soit de l'estimation non paramétrique, notamment avec l'estimateur de Kaplan-Meier.

Estimateur de Kaplan-Meier

L'estimateur de Kaplan-Meier est utilisé pour estimer empiriquement la fonction de survie $S(t)$ à partir de données censurées. Cet estimateur est utilisé lorsque les données de survie ne semble pas suivre une distribution particulière (comme celles décrites par la suite), ce qui peut être fréquent en médecine et en épidémiologie. L'idée est d'utiliser un estimateur du risque instantané, en se servant des pas de temps donnés par les données (chaque temps observé distinct donne un pas de temps) et d'estimer, à chaque pas de temps, la proportion de survivants parmi les personnes à risque. Cela se traduit par une fonction décroissante en escalier.

Voici une description de l'estimateur de Kaplan-Meier : soient n individus dont les temps observés sont $\{T_1, ..., T_n\} \in \mathbb{R}_+^n$ et les statuts (affectés ou non) $\{\delta_1, ..., \delta_n\} \in \{0, 1\}^n$.

- les individus sont réordonnés par temps observés dans l'ordre croissant tels que :

$$T_{(1)} < ... < T_{(N)}$$

  avec $N \leqslant n$ (si il y a des égalités de temps observés).

- le risque instantané est estimé à chaque pas de temps $T_{(i)}$ par $\frac{d_i}{R_i}$ où
  - d_i est le nombre d'évènements observés au temps $T_{(i)}$ (i.e. parmi tous les individus ayant pour temps observé $T_{(i)}$, le nombre d'individus pour lesquels $\delta = 1$) ;
  - R_i est le nombre d'individus encore à risque au temps $T_{(i)}$ (i.e. nombre d'individus ayant pour temps observés $T \geqslant T_{(i)}$)

- l'estimateur de Kaplan-Meier s'écrit alors :

$$S_{KM}(t) = \prod_{k=1}^{i} \left(1 - \frac{d_k}{R_k}\right) \text{ quand } T_{(i)} \leqslant t < T_{(i+1)}$$

Modèles exponentiels et modèle de Weibull

Les modèles paramétriques, tels que les modèles exponentiels et le modèle de Weibull, supposent une forme spécifique pour la fonction de risque. Ils peuvent être utilisés pour ajuster des données de survie lorsque la forme de la fonction de risque est connue ou supposée.

- **Modèles exponentiels avec risque constant :**

  Le modèle exponentiel est l'un des modèles de survie les plus simples et repose sur l'hypothèse que la fonction de risque (le taux de défaillance) est constante dans le temps. Il est adapté lorsque le risque de l'événement est approximativement constant dans le temps. La fonction de risque instantané dans le modèle exponentiel est donnée par :

  $$\lambda(t) = \lambda$$

  où λ est la constante qui représente le taux de défaillance constant. Le risque cumulé s'écrit donc :

  $$\Lambda(t) = \int_0^t \lambda du = \lambda t$$

  La fonction de survie correspondante est exponentielle :

  $$S(t) = \exp(-\lambda t)$$

  Le modèle exponentiel est largement utilisé pour analyser des données de survie lorsque le risque est constant. Il est souvent utilisé dans des contextes tels que la fiabilité des composants électroniques, où l'on suppose que les composants échouent à un taux constant.

- **Modèles exponentiels avec risque constant par morceau :**

  Il s'agit d'un modèle très similaire au précédent où le risque n'est plus constant mais constant par morceau. En supposant que le temps est découpé en k morceaux ($[0, t_1[$, .. , $[t_{k-1}, +\infty[$) la fonction de risque s'écrit :

  $$\lambda(t) = \begin{cases} \lambda_1 & \text{si } t < t_1 \\ \lambda_2 & \text{si } t_1 \leqslant t < t_2 \\ \dots & \\ \lambda_k & \text{si } t_{k-1} \leqslant t \end{cases}$$

  où les λ_1, ..., λ_k sont constants. Le risque cumulé s'écrit donc :

  $$\Lambda(t) = \begin{cases} -\lambda_1(t - t_0) & \text{si } t < t_1 \\ -\lambda_1(t_1 - t_0) - \lambda_2(t - t_1) & \text{si } t_1 \leqslant t < t_2 \\ \dots & \\ -(\sum_{i=1}^{k-1} \lambda_i(t_i - t_{i-1})) - \lambda_k(t - t_k) & \text{si } t_{k-1} \leqslant t \end{cases}$$

  La fonction de survie correspondante est (en posant $t_0 = 0$) :

$$S(t) = \begin{cases} \exp(-\lambda_1(t - t_0)) & \text{si } t < t_1 \\ \exp(-\lambda_1(t_1 - t_0) - \lambda_2(t - t_1)) & \text{si } t_1 \leqslant t < t_2 \\ \dots & \\ \exp(-(\sum_{i=1}^{k-1} \lambda_i(t_i - t_{i-1})) - \lambda_k(t - t_k)) & \text{si } t_{k-1} \leqslant t \end{cases}$$

Le modèle exponentiel avec risque constant par morceau est une extension du modèle exponentiel qui prend en compte des segments de temps où le risque de l'événement d'intérêt est constant, tout en permettant la variation du risque à différents moments. C'est une méthode puissante pour modéliser des données de survie complexes avec des variations temporelles dans le risque.

- **Modèle de Weibull :**

  Le modèle de Weibull est un modèle paramétrique plus flexible que le modèle exponentiel. Il permet de modéliser des risques qui peuvent varier dans le temps. La fonction de risque instantané dans le modèle de Weibull est donnée par :

  $$\lambda(t) = \frac{\beta}{\alpha} \times \left(\frac{t}{\alpha}\right)^{\beta-1}$$

  où α est un paramètre d'échelle, et β est un paramètre de forme. La fonction de survie correspondante est :

  $$S(t) = \exp(-\left(\frac{t}{\alpha}\right)^{\beta})$$

  Le modèle de Weibull est couramment utilisé pour modéliser des risques qui augmentent ou diminuent au fil du temps. Le paramètre β permet de déterminer la forme de la courbe de risque instantané : si $\beta < 1$, la courbe est décroissante (le risque diminue avec le temps), si $\beta > 1$, la courbe est croissante (le risque augmente avec le temps), et si $\beta = 1$, le modèle se réduit au modèle exponentiel. Le modèle de Weibull est utile pour ajuster des données de survie lorsque le risque n'est pas constant et peut varier de manière exponentielle ou non linéaire.

- **Estimation des paramètres du maximum de vraisemblance :**

  Les différents modèles présentés précédemment se résument à un nombre fini de paramètres : λ pour le modèle exponentiel avec risque constant, les λ_1, ..., λ_k (voire les t_1,...,t_k) pour le modèle exponentiel avec risque constant par morceau et enfin α et β pour le modèle de Weibull.

  L'estimation des paramètres d'une loi de survie paramétrique avec la méthode du maximum de vraisemblance est une approche courante en analyse de survie. Cette méthode permet d'ajuster un modèle de survie paramétrique à des données de survie en cherchant à maximiser la vraisemblance des données sous le modèle choisi.

  Voici les étapes générales pour estimer les paramètres d'une loi de survie paramétrique :

  – Choix du modèle de survie paramétrique convenant le mieux aux données.

– Écriture de la fonction de vraisemblance pour le modèle de survie choisi : on considère n données collectées, pour chaque donnée i ($i \in \{1, .., n\}$), le temps observé est T_i et le statut (censuré ou non) est $\delta_i \in \{0, 1\}$, en posant θ l'ensemble des paramètres du modèle considéré, $S(t, \theta)$ la fonction de survie et $\lambda(t, \theta)$ le risque instantané, la fonction de vraisemblance du modèle s'écrit :

$$\mathcal{L}(\theta) = \prod_{i=1}^{n} f(T_i, \theta) = \prod_{i=1}^{n} S(T_i, \theta) \times \lambda(T_i, \theta)^{\delta_i}$$

– Maximisation de la log-vraisemblance en utilisant des techniques d'optimisation numérique. Les méthodes couramment utilisées incluent la méthode de Newton-Raphson, la méthode de quasi-Newton (comme BFGS), et d'autres algorithmes d'optimisation.

– Estimation des paramètres du maximum de vraisemblance : ce sont les valeurs des paramètres qui maximisent la log-vraisemblance.

$$\hat{\theta}_{\text{ML}} = arg \max_{\theta} \mathcal{L}(\theta)$$

2.2.4 Modèle de Cox

Le modèle de régression de Cox est l'un des modèles les plus couramment utilisés, en analyse de survie, pour modéliser l'effet des covariables sur le risque de survie, en utilisant une fonction de risque proportionnelle. L'équation de Cox est la suivante :

$$\lambda(t, X_1, .., X_p) = \lambda_0(t) \times \exp(\beta_1 X_1 + \beta_2 X_2 + .. + \beta_p X_p) = \lambda_0(t) \times \exp(\beta^T X)$$

où $\lambda(t, X_1, .., X_p)$ est la fonction de risque instantané, $\lambda_0(t)$ est la fonction de risque de base, les β_i sont les coefficients de régression (réunis dans le vecteur β pour la notation matricielle) et X_i sont les covariables (réunis dans le vecteur X). Le but de ce modèle n'est pas d'estimer la fonction de risque instantané de base (*i.e.* $\lambda_0(t)$) mais plutôt les coefficients de régression β_i. Pour cela, le modèle de Cox utilise une maximisation de la vraisemblance partielle.

2.2.5 Modèle à fragilité

Les modèles à fragilité en analyse de survie sont utilisés pour modéliser la variation individuelle dans le risque de survenue d'événement au cours du temps. Ils sont particulièrement adaptés lorsque les individus présentent des caractéristiques hétérogènes qui ne peuvent pas être entièrement expliquées par les covariables observées permettant une extension du modèle de Cox. La notion de "fragilité" capture cette hétérogénéité non observée entre les individus.

La fragilité est une variable aléatoire qui capture les différences individuelles non observées ou non mesurées qui influencent le risque d'événement. Le modèle à fragilité combine le modèle de survie standard avec une distribution de probabilité pour la fragilité. En d'autres termes, il suppose que le risque individuel est affecté potentiellement par des covariables mesurées et une fragilité individuelle non mesurée. Mathématiquement, cela peut s'écrire comme :

$$\lambda(t, Z, X) = Z \times \lambda_0(t) \times \exp(\beta^T X)$$

où $\lambda(t, Z, X)$ est la fonction de risque instantané à un temps t, X les covariables mesurées, β les coefficients associés aux covariables, $\lambda_0(t)$ la fonction de risque de base, Z la variable de fragilité (positive).

La distribution de la fragilité spécifie comment la fragilité individuelle est répartie dans la population. Les distributions couramment utilisées incluent la distribution gamma, log-normale, ou Weibull, entre autres. L'estimation des paramètres dans un modèle à fragilité peut typiquement être effectuée par la méthode du maximum de vraisemblance.

2.2.6 Simulations et implémentation sous R

Le but de cette partie est de présenter quelques outils d'analyse de survie sous R (notamment avec le package *survival*) sur quelques exemples simulés. Les méthodes développées dans cette thèse ont notamment été testées sur des jeux de données simulés et les méthodes de simulation mises en œuvre sont aussi présentées.

Simulations de modèles exponentiels avec risque constant par morceaux par méthode de la transformée inverse

Le but est de simuler des nombres aléatoires suivant une loi de survie exponentielle avec risque constant par morceaux. La méthode utilisée est celle de la transformée inverse. Cette méthode permet d'échantillonner une variable aléatoire T suivant une loi de probabilité donnée grâce sa fonction de répartition F et une variable uniforme sur $[0, 1]$. Pour cela, en simplifiant, il faut :

- Connaître la fonction de répartition de la variable aléatoire souhaitée F.
- Être capable de calculer la fonction de répartition inverse F^{-1}, i.e. la fonction qui prend une probabilité cumulée en entrée et renvoie la valeur correspondante de la variable aléatoire. En d'autres termes, elle transforme une probabilité cumulée en une valeur aléatoire.
- Générer des nombres aléatoires uniformément distribués entre 0 et 1 (généralement obtenus à partir d'une fonction de génération de nombres pseudo-aléatoires) et les utiliser comme probabilités cumulées en les donnant en entrée à la fonction de répartition inverse. Les valeurs aléatoires obtenues suivent la loi de probabilité d'intérêt.

Dans le cas des modèles exponentiels avec risque constant par morceaux comme définis auparavant (avec t_0,...,t_k et λ_1,..., λ_k), la fonction de répartition $F(t) = 1 - e^{-\Lambda(t)}$ avec $\Lambda(t) = \int_0^t \lambda(u)du$.

En suivant la formulation de

$$\lambda(t) = \begin{cases} \lambda_1 & \text{si } t < t_1 \\ \lambda_2 & \text{si } t_1 \leqslant t < t_2 \\ \dots & \\ \lambda_k & \text{si } t_{k-1} \leqslant t \end{cases}$$

il est possible d'écrire

$$\Lambda(t) = \begin{cases} \lambda_1 t & \text{si } t < t_1 \\ \lambda_1 t_1 + \lambda_2(t - t_1) & \text{si } t_1 \leqslant t < t_2 \\ \dots & \\ \sum_{i=1}^{k-1} \lambda_i(t_i - t_{i-1})) + \lambda_k(t - t_k) & \text{si } t_{k-1} \leqslant t \end{cases}$$

Il suffit de générer un nombre aléatoire $u \sim U[0,1]$, poser $x = -\ln(1-u)$ et ainsi calculer

$$t = \Lambda^{-1}(x) = \begin{cases} \frac{x}{\lambda_1} & \text{si } x < \lambda_1 t_1 \\ t_1 + \frac{x - \lambda_1 t_1}{\lambda_2} & \text{si } \lambda_1 t_1 \leqslant x < \lambda_1 t_1 + \lambda_2 (t_2 - t_1) \\ \dots & \\ t_{k-1} + \frac{x - \sum_{i=1}^{k-1} \lambda_i (t_i - t_{i-1})}{\lambda_k} & \text{si } \sum_{i=1}^{k-1} \lambda_i (t_i - t_{i-1}) \leqslant x \end{cases}$$

Ainsi la valeur aléatoire t générée suit une loi de survie exponentielle avec risque constant par morceaux.

Simulations de données de survie avec censure sous R

L'ensemble du code est disponible en appendice. Le but est de simuler un jeu de données comprenant $2n$ individus repartis dans 2 groupes de même taille, dont n individus suivent une loi de survie exponentielle avec risque constant par morceaux et les n autres individus suivent la même loi mais avec un risque proportionnel supplémentaire β.

Dans un premier temps, il faut définir les différents paramètres, les t_1,...,t_k (cuts), les λ_1,...,λ_{k+1} (lambda) et β (beta). La Figure 2.2 montre les fonctions de risque instantané et de survie avec :

- des cuts à $\{20, 40, 60, 80\}$
- les risques instantanés constants du 1er groupe $\{\lambda_1, \lambda_2, \lambda_3, \lambda_4, \lambda_5\} = \{0, 0.002, 0.006, 0.009, 0.006\}$
- le risque proportionnel $\beta = 0.3$
- une censure uniforme $U[40, 80]$

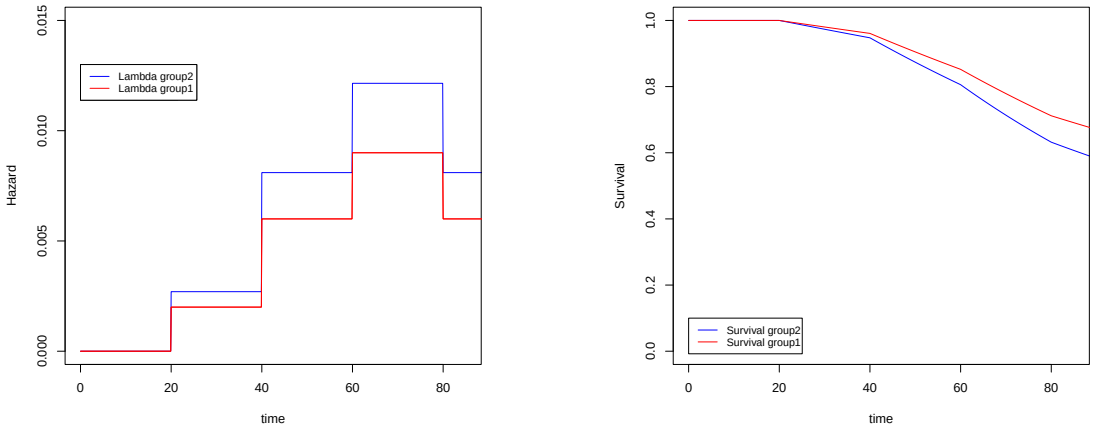


FIGURE 2.2 : Fonctions de risque instantané et de survie pour les deux groupes.

Les données de survie peuvent maintenant être générées à partir de ces paramètres.

Estimateur de Kaplan-Meier sous R

Maintenant que les données de survie sont générées et que les distributions de survie des deux groupes sont connues, il est intéressant d'essayer d'estimer ces survies grâce aux méthodes

existantes comme Kaplan-Meier. Le résultat de l'estimateur de Kaplan-Meier sur notre exemple est présenté en Figure 2.3.

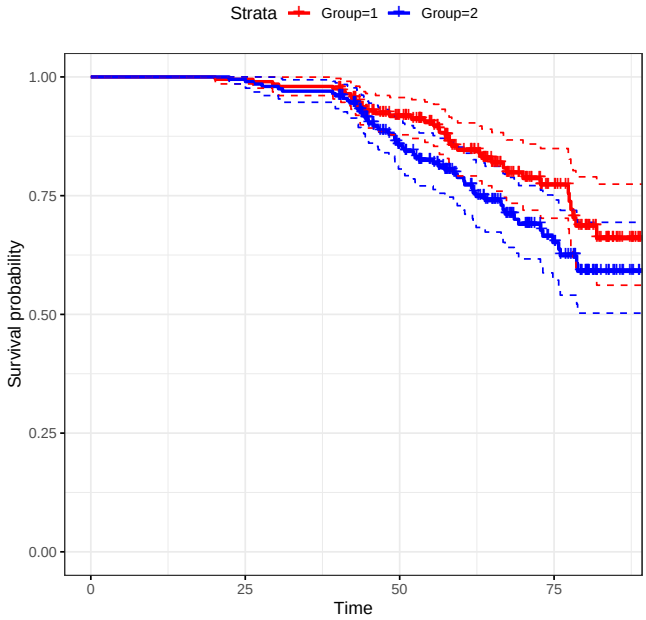


FIGURE 2.3 : Estimation des fonctions de survie avec Kaplan-Meier.

Modèle de Cox sous R

Étant donné que les données générées présentent 2 groupes distincts donc le risque relatif est une constante β, le modèle de Cox est tout indiqué pour estimer ce coefficient. On obtient une estimation de $\beta = 0.3856$.

Pour rappel, l'ensemble du code est disponible en appendice du chapitre 2.

2.3 Réseaux bayésiens

Cette section s'intéresse aux réseaux bayésiens et leur utilisation dans l'étude de la transmission génétique familiale. Dans ce but, la section présente une définition de ce que sont des réseaux bayésiens et les méthodes utilisées pour faire de l'inférence sur ces derniers, notamment la propagation de croyance avec des algorithmes somme-produit. L'idée est de montrer, sur un exemple, comment un arbre familial peut être modélisé par un réseau bayésien et comment calculer des vraisemblances et des lois de génotype a posteriori à partir de celui-ci.

2.3.1 Définition

Un réseau bayésien est un modèle probabiliste graphique qui représente des dépendances probabilistes entre un ensemble de variables aléatoires. Ces dépendances sont représentées sur un graphe acyclique orienté. Un réseau bayésien est composé de deux éléments principaux :

- Nœuds (ou nœuds aléatoires) : Chaque nœud du réseau représente une variable aléatoire. Ces variables peuvent être des événements, des caractéristiques, des états, ou tout autre phénomène dont on souhaite modéliser les interactions.

- Arêtes (ou arcs) : Les arêtes entre les nœuds représentent les relations de dépendance probabiliste entre ces variables. Une arête qui va d'un nœud A à un nœud B indique que la variable aléatoire A a une influence sur la variable aléatoire B.

Les réseaux bayésiens utilisent la théorie des probabilités bayésiennes pour quantifier ces dépendances. Chaque nœud est associé à une distribution de probabilité conditionnelle qui décrit comment la variable aléatoire du nœud dépend de ses parents (les nœuds reliés par des arêtes entrantes). Ainsi, dans un réseau bayésien représenté par un graphe acyclique orienté (orienté pour les relations causales et acyclique pour l'interprétation, les boucles impliquant des rétroactions) dont les n nœuds sont $X = \{X_1, ...X_n\}$ un ensemble de variables, en exploitant les propriétés d'indépendance conditionnelle, il est possible de factoriser la loi de probabilité $\mathbb{P}(X)$ de la manière suivante :

$$\mathbb{P}(X) = \mathbb{P}(X_1, ..., X_n) = \prod_{k=1}^{n} \mathbb{P}(X_k | X_{\mathrm{pa}(k)})$$

où pour tout $k \in \{1, ..., n\}$, $X_{\mathrm{pa}(k)}$ est l'ensemble des parents ($\mathrm{pa}(k) \in \{1, ..., n\} \setminus \{k\}$, potentiellement vide) de la variable X_k dans le graphe.

Les réseaux bayésiens sont utilisés dans plusieurs domaines dont trois grands champs d'application peuvent être mis en avant :

- Inférence de variables latentes : les variables latentes sont des variables non observées dans un modèle, qu'on cherche généralement à inférer à partir des variables observées.

- Apprentissage de structure : L'apprentissage de structure concerne la détermination des relations (arêtes) entre les variables dans un réseau bayésien, en se basant sur les données observées.

- Apprentissage de paramètres : sur une structure du réseau bayésien déterminée, l'apprentissage de paramètres consiste à estimer les probabilités conditionnelles associées à chaque variable en fonction de ses parents.

Dans le cadre de cette thèse, la structure des réseaux bayésiens considérés sera connue (arbre familial), tout comme les lois de dépendance entre les variables (transmission mendélienne). Tout l'enjeu sera donc l'inférence de variables latentes (les génotypes) à partir de variables observées (certains génotypes observés et des données de phénotypes). En effet, comme expliqué dans le chapitre précédent, dans les données de pédigrée, les génotypes sont rarement observés hormis celui du *proband* et, parfois, certains autres membres de la famille. Le but est donc d'inférer les probabilités marginales des génotypes pour chaque membre de la famille à partir des données connues.

2.3.2 Modélisation de pédigrées par réseaux bayésiens

Comme le montre la Figure 2.4, il est possible de représenter une famille sous forme de réseau bayésien. Dans l'exemple, en posant $X = \{X_1, ..., X_6\}$ les génotypes des différents membres de la famille, on considère que le génotype de chaque individu est indépendant du génotype des autres membres de la famille, conditionnellement aux génotypes de ses parents. Ainsi on peut alors écrire la loi de probabilité des génotypes :

$$\mathbb{P}(X_1, ..., X_6) = \mathbb{P}(X_1) \times \mathbb{P}(X_2) \times \mathbb{P}(X_3|X_1, X_2) \times \mathbb{P}(X_4|X_1, X_2) \times \mathbb{P}(X_5) \times \mathbb{P}(X_6|X_4, X_5)$$

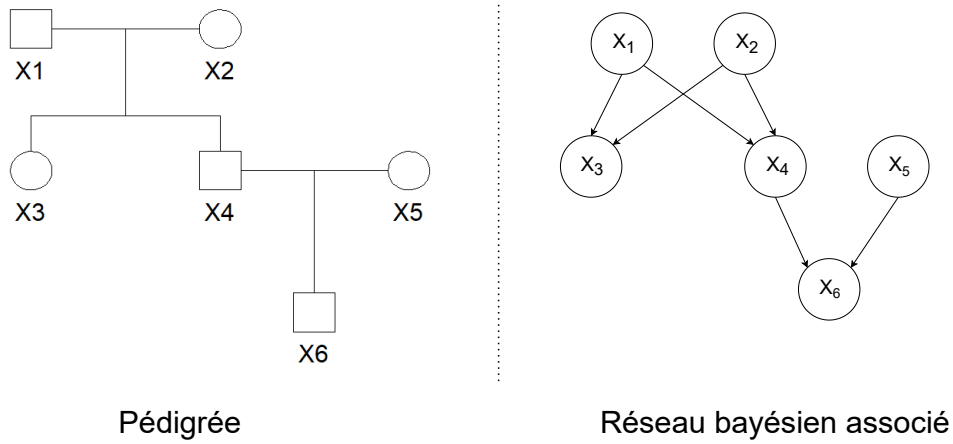


FIGURE 2.4 : Exemple de pédigrée pour une famille avec $n = 6$ personnes et le réseau bayésien associé.

Exemple : modèle mendélien autosomal

En utilisant la famille exemple de la Figure 2.4, on peut s'intéresser à un exemple de modélisation de $X = \{X_1, ..., X_6\}$.

- On se place dans un cadre monogénique avec deux allèles (un gène d'intérêt, un allèle normal "0" et un allèle muté "1").
- La transmission sera mendélienne autosomale dominante. Ainsi pour tout membre de la famille $i \in \{1, 2, 3, 4, 5, 6\}$, son génotype réel appartient à $\{00, 01, 10, 11\}$ (à gauche, allèle paternel et, à droite, allèle maternel), mais on considère, pour simplifier le modèle, qu'il n'est pas possible de distinguer les hétérozygotes $\{01, 10\}$. On pose donc les génotypes $X_i \in \{00, 01, 11\}$. La transmission de parents à enfant se calcule donc selon la table 2.1.
- Les génotypes des membres fondateurs de la famille (c'est-à-dire ceux qui n'ont pas de parents $\{1, 2, 5\}$) suivent l'équilibre de Hardy-Weinberg pour une mutation présentant une fréquence allélique $f = 0.2$ en population générale, comme présenté en Figure 2.2.

Parents X_i, X_j	$00, 00$	$00, 01$	$00, 11$	$01, 00$	$01, 01$	$01, 11$	$11, 00$	$11, 01$	$11, 11$
$\mathbb{P}(X_k = 00 \mid X_i, X_j)$	1.0	0.5	0.0	0.5	0.25	0.0	0.0	0.0	0.0
$\mathbb{P}(X_k = 01 \mid X_i, X_j)$	0.0	0.5	1.0	0.5	0.5	0.5	1.0	0.5	0.0
$\mathbb{P}(X_k = 11 \mid X_i, X_j)$	0.0	0.0	0.0	0.0	0.25	0.5	0.0	0.5	1.0

TABLE 2.1 : Loi du génotype pour un enfant $X_k \in \{X_3, X_4, X_6\}$ sachant le génotype des parents

	$X_i = 00$	$X_i = 01$	$X_i = 11$
$\mathbb{P}(X_i)$	0.64	0.32	0.04

TABLE 2.2 : Loi de probabilité pour les fondateurs suivant l'équilibre de Hardy-Weinberg X_1,X_2,X_5 avec $f = 0.2$

Notion d'évidence

On introduit ici la notion d'évidence (notée ici ev) qui représente les connaissances a priori disponibles sur les données. Dans le modèle présenté, on peut considérer que le génotype de l'individu 4 (voir Figure 2.4) est observé et qu'il est homozygote pour l'allèle pathogène, ainsi $X_4 = 11$. L'information pour les autres membres de la famille est inconnue. Il est possible d'écrire l'évidence de la manière suivante :

$$\text{ev} = \{X_4 = 11, X_{1,2,3,5,6} \in \{00, 01, 11\}\}$$

Ainsi on s'intéresse au calcul, à la fois, de la vraisemblance $\mathbb{P}(\text{ev})$ et aux lois marginales des génotypes sachant cette évidence $\mathbb{P}(X_k|\text{ev})$.

- La vraisemblance s'écrit :

$$\mathbb{P}(\text{ev}) = \sum_X \mathbb{P}(X, \text{ev}) = \sum_X \prod_{i=1}^{6} \underbrace{\mathbb{P}(X_i, \text{ev}|X_{\text{pat}_i}, X_{\text{mat}_i})}_{K_i(X_i, X_{\text{pat}_i}, X_{\text{mat}_i})}$$

  En reprenant l'exemple précédent :

$$\mathbb{P}(\text{ev}) = \sum_{X_1} \sum_{X_2} \sum_{X_3} \sum_{X_4} \sum_{X_5} \sum_{X_6} K_1(X_1)K_2(X_2)K_3(X_3, X_1, X_2)K_4(X_4, X_1, X_2)K_5(X_5)K_6(X_6, X_4, X_5)$$

- Les lois marginales s'écrivent pour tout $k \in \{1, ..., 6\}$:

$$\mathbb{P}(X_k|\text{ev}) = \frac{\mathbb{P}(X_k, \text{ev})}{\mathbb{P}(\text{ev})}$$

On peut remarquer ici que le calcul de la vraisemblance est nécessaire aux calculs des lois marginales. La première idée pour calculer cette vraisemblance pourrait être d'envisager toutes les combinaisons de génotypes possibles, en considérant un algorithme type *brute force*. Pour un modèle avec 3 génotypes possibles et d individus dans la famille, le nombre de configuration est 3^d. En reprenant l'exemple ci-dessus avec 6 individus, le nombre de configurations est $3^6 = 729$. Or il n'est pas rare que des familles comprennent entre 10 à 20 individus. En prenant une moyenne de 15, le nombre de configurations est $3^{15} = 14\,348\,907$. Ainsi l'algorithme *brute force* n'est pas une solution viable pour le calcul de la vraisemblance, mais il existe d'autres manières de simplifier les calculs, notamment via l'algorithme somme-produit.

2.3.3 *Belief propagation* dans les réseaux bayésiens

La *belief propagation* (propagation de croyance ou de conviction) est un algorithme qui a été introduit par Pearl [52] [53] puis généralisé par différents articles [38] [60] [57] et est utilisé dans les réseaux bayésiens pour effectuer des calculs d'inférence. Il est également connu sous le nom d'algorithme *sum-product* (Algorithme somme-produit).

L'objectif principal de la *belief propagation* est de calculer les probabilités marginales des nœuds du réseau bayésien, c'est-à-dire la probabilité d'une variable aléatoire étant données les observations, en tenant compte des relations de dépendance probabiliste entre les variables.

Algorithme somme-produit

L'idée de base derrière l'algorithme somme-produit est de propager de l'information le long des arêtes du réseau bayésien, en utilisant des messages pour mettre à jour les probabilités des nœuds, en fonction des probabilités de leurs voisins. L'algorithme se déroule en deux étapes principales : la propagation ascendante (somme) et la propagation descendante (produit).

- Propagation ascendante (*inward message*, étape de somme) : dans cette étape, chaque nœud envoie un message à ses nœuds parents en sommant les informations des nœuds descendants et en prenant en compte sa propre distribution de probabilité conditionnelle. Ces messages partent des nœuds feuilles du réseau et remontent progressivement vers la racine.
- Propagation descendante (*backward message*, étape de produit) : une fois que les messages ont atteint la racine, on commence la propagation descendante. Les nœuds parents envoient des messages à leurs nœuds enfants en utilisant la règle du produit de Bayes, mettant à jour les probabilités marginales des nœuds enfants.

L'algorithme somme-produit est particulièrement utile pour les réseaux bayésiens avec de nombreuses variables et des structures complexes, car il permet d'effectuer des calculs d'inférence, de manière efficace, en évitant les calculs coûteux mis en évidence précédemment.

Application sur des données de pédigrée

Les travaux de Elston et Stewart sur le calcul de vraisemblance à partir d'arbres familiaux ont donné lieu à l'algorithme d'Elston-Stewart [22][23]. Cet algorithme est largement utilisé en génétique et il est, en fait, l'équivalent de la phase ascendante de l'algorithme somme-produit. En plus de la vraisemblance, calculée grâce à la phase ascendante, la phase descendante permet de récupérer les lois marginales a posteriori pour l'ensemble des nœuds du réseaux [70].

Pour appliquer l'algorithme somme-produit aux arbres familiaux, il est nécessaire de suivre un certain nombre d'étapes, notamment transformer le réseau bayésien formé par l'arbre familial en un arbre utilisable par l'algorithme.

- Graphe moral : tout d'abord il faut passer de la structure de réseaux bayésiens à un graphe moral. Pour cela il suffit de reprendre les nœuds du réseau, rajouter les arêtes existantes entre les nœuds sans les orienter et, enfin, pour chaque nœud, rajouter des arêtes entre chacun de ses parents si elles n'existent pas déjà.
- Triangulation : l'objectif principal de la triangulation est de convertir le graphe moral en un graphe triangulé, c'est-à-dire un graphe dans lequel chaque cycle de longueur supérieure à trois a une corde (une arête ajoutée entre deux nœuds non adjacents du cycle).
- Graphe de Jonction : à partir du graphe triangulé, il faut construire un graphe de jonction. Chaque triangle (formé de trois variables) du graphe triangulé devient une clique (un nœud du graphe de Jonction). Il faut ensuite ajouter un arête (non orientée) entre chaque clique ayant au moins une variable commune.
- Arbre de Jonction : il faut réduire le graphe de jonction en un arbre de jonction. Pour cela, il faut supprimer des arêtes du graphe en respectant certaines règles :
  - Pour toutes cliques C_i, C_j, il existe un unique chemin connectant C_i et C_j que l'on note $path(i, j)$.

– Pour tous i, j, k et pour toute variable X_p dans la clique C_k, si $X \in C_i \cap C_j$ alors $C_k \in path(i, j)$.

– Pour toute variable aléatoire X_p, il existe au moins une clique C_i contenant X_p et ses parents $X_{parents_p}$.

Ces règles sont mises en œuvre sur l'arbre proposé en exemple. Le réseau bayésien donné par l'arbre étant simple, l'étape de triangulation est inutile (car le graphe est déjà triangulé) et l'étape de passage du graphe de jonction à l'arbre de jonction est instantanée comme le montre la Figure 2.5.

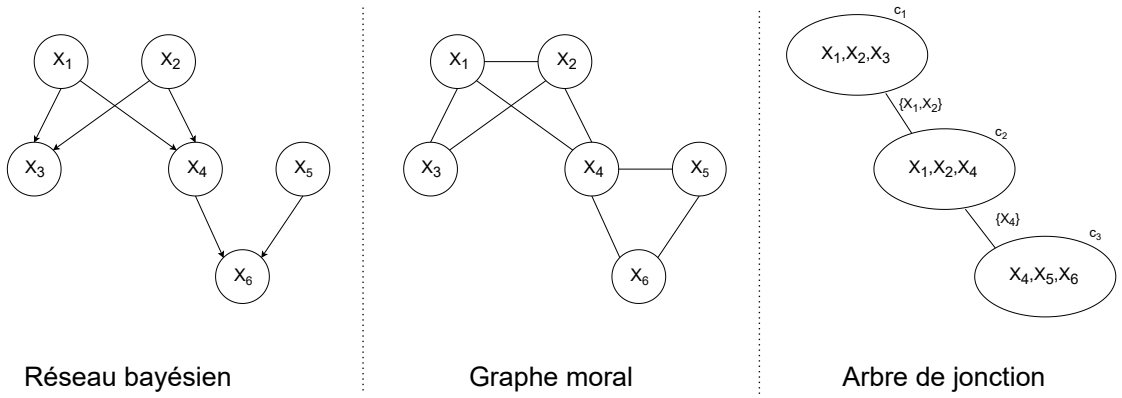


FIGURE 2.5 : Passage du réseau bayésien au graphe moral puis à l'arbre de jonction pour l'exemple présenté. Les variables présentes sur les arêtes de l'arbre de jonction sont les variables communes entre les deux cliques reliées.

A partir de l'arbre de jonction, il est possible d'appliquer l'algorithme somme-produit (voir Figure 2.6) :

- Choisir une racine (*root*) à l'arbre de jonction. On choisit ici C_1.

- Propager les messages ascendants des feuilles de l'arbre jusqu'à la racine qui sont des fonctions dépendantes des variables communes entre les cliques reliées. Ici, il n'y a qu'une seule feuille, donc il faut calculer les messages $M_{C_3 \to C_2}(X_4)$ et $M_{C_2 \to C_1}(X_1, X_2)$.

- Propager les messages descendants de la racine vers les feuilles de l'arbre qui sont également des fonctions dépendantes des variables communes entre les cliques reliées. Ici, à nouveau, il n'y a qu'une seule feuille, donc il faut calculer les messages $M_{C_1 \to C_2}(X_1, X_2)$ et $M_{C_2 \to C_3}(X_4)$.

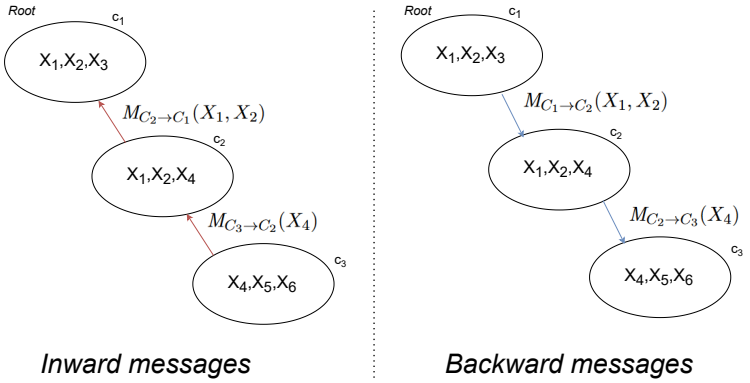


FIGURE 2.6 : Choix de la racine (*root*) et direction des messages ascendants (*inward*) et descendants (*backward*).

On peut alors calculer les messages :

- ascendants :
  - $M_{C_3 \to C_2}(X_4) = \sum_{X_5} \sum_{X_6} K_5(X_5) \times K_6(X_6, X_4, X_5)$
  - $M_{C_2 \to C_1}(X_1, X_2) = \sum_{X_4} K_4(X_4, X_1, X_2) \times M_{C_3 \to C_2}(X_4)$
- descendants :
  - $M_{C_1 \to C_2}(X_1, X_2) = \sum_{X_3} K_1(X_1) \times K_2(X_2) \times K_3(X_3, X_1, X_2)$
  - $M_{C_2 \to C_3}(X_4) = \sum_{X_1} \sum_{X_2} K_4(X_4, X_1, X_2) \times M_{C_1 \to C_2}(X_1, X_2)$

Étant donnée l'évidence $\text{ev} = \{X_4 = 11, X_{1,2,3,5,6} \in \{00, 01, 11\}\}$, on peut donc donner les valeurs des différents messages (voir tables 2.3 pour $M_{C_3 \to C_2}$ et $M_{C_2 \to C_3}$ et 2.4 pour $M_{C_2 \to C_1}$ et $M_{C_1 \to C_2}$).

X_4	00	01	11
$M_{C_3 \to C_2}(X_4)$	0.0	0.0	1.0
$M_{C_2 \to C_1}(X_4)$	0.0	0.0	0.04

TABLE 2.3 : Table des valeurs de $M_{C_3 \to C_2}$ et $M_{C_2 \to C_3}$ selon X_4.

X_1, X_2	$00, 00$	$01, 00$	$11, 00$	$00, 01$	$01, 01$	$11, 01$	$00, 11$	$01, 11$	$11, 11$
$M_{C_2 \to C_1}(X_1, X_2)$	0.0	0.0	0.0	0.0	0.25	0.5	0.0	0.5	1.0
$M_{C_1 \to C_2}(X_1, X_2)$	0.4096	0.2048	0.0128	0.2048	0.1024	0.0064	0.0128	0.0064	0.0004

TABLE 2.4 : Table des valeurs de $M_{C_2 \to C_1}$ et $M_{C_1 \to C_2}$ selon X_1, X_2.

Maintenant que l'ensemble des messages sont calculés, il est possible de les utiliser pour calculer à la fois la vraisemblance $\mathbb{P}(\text{ev})$ et les lois marginales pour chaque variable. En effet, pour chaque intersection entre deux cliques, on peut écrire :

- $C_2 \cap C_3 = \{X_4\}$, $\mathbb{P}(X_4, \text{ev}) = M_{C_2 \to C_3}(X_4) \times M_{C_3 \to C_2}(X_4)$.
- $C_1 \cap C_2 = \{X_1, X_2\}$, $\mathbb{P}(X_1, X_2, \text{ev}) = M_{C_1 \to C_2}(X_1, X_2) \times M_{C_2 \to C_1}(X_1, X_2)$.

De même, dans chaque clique, il est possible d'écrire :

- Dans la clique C_1, $\mathbb{P}(X_1, X_2, X_3, \text{ev}) = K_1(X_1)K_2(X_2)K_3(X_3, X_1, X_2)M_{C_2 \to C_1}(X_1, X_2)$.
- Dans la clique C_2, $\mathbb{P}(X_1, X_2, X_4, \text{ev}) = K_4(X_4, X_1, X_2)M_{C_1 \to C_2}(X_1, X_2)M_{C_3 \to C_2}(X_4)$.
- Dans la clique C_3, $\mathbb{P}(X_4, X_5, X_6, \text{ev}) = K_5(X_5)K_6(X_6, X_4, X_5)M_{C_2 \to C_3}(X_4)$.

La vraisemblance peut donc s'écrire de plusieurs manières à partir des messages et, par exemple, en reprenant l'intersection $C_2 \cap C_3$, on peut donc calculer :

$$\mathbb{P}(\text{ev}) = \sum_{X_4} M_{C_2 \to C_3}(X_4) \times M_{C_3 \to C_2}(X_4) = 0.04$$

Et grâce au calcul de la vraisemblance, on peut calculer l'ensemble des lois marginales, par exemple celles de X_1 et X_2 grâce à l'intersection $C_1 \cap C_2$ (voir table 2.5 et 2.6) :

$$\mathbb{P}(X_1|\text{ev}) = \frac{\mathbb{P}(X_1, \text{ev})}{\mathbb{P}(\text{ev})} = \frac{\sum_{X_2} M_{C_1 \to C_2}(X_1, X_2) \times M_{C_2 \to C_1}(X_1, X_2)}{\mathbb{P}(\text{ev})}$$

$$\mathbb{P}(X_2|\text{ev}) = \frac{\mathbb{P}(X_2, \text{ev})}{\mathbb{P}(\text{ev})} = \frac{\sum_{X_1} M_{C_1 \to C_2}(X_1, X_2) \times M_{C_2 \to C_1}(X_1, X_2)}{\mathbb{P}(\text{ev})}$$

	$X_1 = 00$	$X_1 = 01$	$X_1 = 11$
$\mathbb{P}(X_1\|\text{ev})$	0.0	0.8	0.2

TABLE 2.5 : Loi marginale de X_1

	$X_2 = 00$	$X_2 = 01$	$X_2 = 11$
$\mathbb{P}(X_2\|\text{ev})$	0.0	0.8	0.2

TABLE 2.6 : Loi marginale de X_1

Implémentation de l'algorithme somme-produit pour des données de pédigrée : *Bped*

L'algorithme somme-produit est détaillé sur un exemple dans le section précédente mais, de manière générale, il est implémenté dans un programme C++ (aussi développé en R) nommé *Bped*, disponible sur demande à Grégory Nuel (Gregory.Nuel@math.cnrs.fr).

Bped utilise en variables d'entrée, un fichier de pédigrée qui représente la structure familiale (voir la structure du fichier de l'exemple proposé en Table 2.7), un fichier d'évidence qui comprend l'évidence du génotype pour chaque individu de la famille (voir l'évidence de l'exemple en Table 2.8) et une fréquence allélique f. En sortie, *Bped* calcule la (log) vraisemblance et les lois marginales a posteriori pour chaque membre de la famille (voir résultats pour l'exemple en Table 2.9).

Familial ID	Individual ID	Paternal ID	Maternal ID
F_0	1	0	0
F_0	2	0	0
F_0	3	1	2
F_0	4	1	2
F_0	5	0	0
F_0	6	4	5

TABLE 2.7 : Fichier de pédigrée, les individus fondateurs ont pour parents $\{0, 0\}$

Familial ID	Individual ID	00	01	10	11
F_0	1	1	1	1	1
F_0	2	1	1	1	1
F_0	3	1	1	1	1
F_0	4	0	0	0	1
F_0	5	1	1	1	1
F_0	6	1	1	1	1

TABLE 2.8 : Fichier d'évidence

Familial ID	Individual ID	$\mathbb{P}(X = 00\|\text{ev})$	$\mathbb{P}(X = 01\|\text{ev})$	$\mathbb{P}(X = 10\|\text{ev})$	$\mathbb{P}(X = 11\|\text{ev})$
F_0	1	0.0	0.4	0.4	0.2
F_0	2	0.0	0.4	0.4	0.2
F_0	3	0.16	0.24	0.24	0.36
F_0	4	0.0	0.0	0.0	1.0
F_0	5	0.64	0.16	0.16	0.04
F_0	6	0.0	0.8	0.0	0.2

TABLE 2.9 : Fichier de résultats, $\mathbb{P}(\text{ev}) = 0.04$

2.4 Modèles de mélange et algorithme Espérance-Maximisation

Le but de cette section est de présenter l'algorithme EM et les modèles de mélanges. Ces derniers sont des modèles où on suppose que les données observées proviennent de plusieurs groupes ou classes, chacun étant associé à une distribution de probabilité spécifique. Cependant, on n'observe pas directement à quelle classe chaque donnée appartient. Ainsi, la classe des individus est une variable latente du modèle. L'algorithme EM est une méthode de calcul de maximum de vraisemblance utilisée lorsque la vraisemblance est difficilement calculable, notamment à cause de variables latentes. L'algorithme EM est très souvent utilisé dans le cadre des modèles de mélange car ce sont des modèles qui présentent naturellement des variables latentes.

Dans la suite de cette section, les modèles de mélanges seront donc introduits, notamment avec la présentation d'un exemple. Ensuite l'algorithme EM sera détaillé et mis en application sur l'exemple présenté précédemment.

2.4.1 Modèles de mélange

Un modèle de mélange est un modèle probabiliste qui suppose que les données observées proviennent d'un ensemble de plusieurs distributions de probabilités (composantes) plutôt que d'une seule distribution. Mathématiquement, un modèle de mélange peut être défini comme suit.

On dit qu'une variable aléatoire X suit une loi de mélange s'il existe une variable aléatoire $S \in \{1, ..., K\}$ telle que pour tout $i \in \{1, ..., K\}$, $X|S = i$ suit une loi de probabilité particulière de la classe i. On note :

- K le nombre de classes dans le mélange.
- π_k la probabilité d'appartenir à la k-ième classe, avec $\sum_{k=1}^{K} \pi_k = 1$. On pose $\pi = \{\pi_1, ..., \pi_K\}$.
- $p(x|\theta_k)$ la densité de probabilité conditionnelle de la k-ième classe, où θ_k représente les paramètres de cette classe. On pose $\theta = (\theta_1, ..., \theta_K)$.

La variable aléatoire X admet pour densité de probabilité, exprimée en fonction des paramètres du modèle :

$$p(x|\theta,\pi) = \sum_{k=1}^{K} \pi_k \times p(x|\theta_k)$$

Si on observe un ensemble de données $\{x_1, ..., x_n\}$ où x_i représente une observation. Selon le modèle, chaque observation x_i est générée à partir d'une des K classes du mélange, avec sa probabilité associée. Les observations sont considérées indépendantes de sorte que la vraisemblance totale pour l'ensemble des données observées peut s'écrire comme suit :

$$\mathcal{L}(\theta,\pi) = \sum_{i=1}^{n} p(x_i|\theta,\pi)$$

L'objectif, dans un modèle de mélange, est généralement d'estimer les paramètres $\{\pi, \theta\}$ ainsi que les lois a posteriori de chaque classe. En d'autres termes, on souhaite, généralement, estimer les lois de probabilité de chaque classe (via leur paramètres) et savoir pour chaque observation, à quelle classe elle appartient (via la loi a posteriori de la classe)

Exemple :

On s'intéresse dans cet exemple à une population composée de personnes obèses et non-obèses, on considère la variable obésité comme latente. Ainsi on pose :

- y le poids observé d'un individu
- X une variable observée composée du sexe (0 : homme, 1 : femme) et la taille (en cm)
- S le statut non-observé d'obésité (0 : non-obèse, 1 : obèse)

On s'intéresse donc au modèle de mélange où le poids de chaque groupe (obèse et non-obèse) est donné par les lois suivantes :

Pour les obèses :

$$y|X, S = 0 \sim \mathcal{N}(\alpha_0 + \alpha_{\text{sex}} X_{\text{sex}} + \alpha_{\text{height}} X_{\text{height}}, \sigma^2)$$

Pour les non-obèses :

$$y|X, S = 1 \sim \mathcal{N}(\beta_0 + \beta_{\text{sex}} X_{\text{sex}} + \beta_{\text{height}} X_{\text{height}}, \tau^2)$$

De plus on pose la probabilité d'être obèse telle que :

$$\mathbb{P}(z = 1) = p \quad \mathbb{P}(z = 0) = 1 - p$$

On fixe les valeurs des paramètres comme suit :

- On considère 2000 individus $n = 2000$
- La probabilité d'être obèse est $p = 0.25$
- La variance de la loi normale pour les non-obèses est $\sigma = 7$
- La variance de la loi normale pour les obèses est $\tau = 10$
- Les paramètres pour les non-obèses $\alpha_0 = 20$, $\alpha_{\text{sex}} = 10$, $\alpha_{\text{height}} = 0.25$
- Les paramètres pour les non-obèses $\beta_0 = 40$, $\beta_{\text{sex}} = 30$, $\beta_{\text{height}} = 0.35$

On obtient ainsi une population qui est le résultat d'un mélange comme le montre la Figure 2.11. Pour rappel le code R permettant de générer la population et le graphique est disponible en appendice.

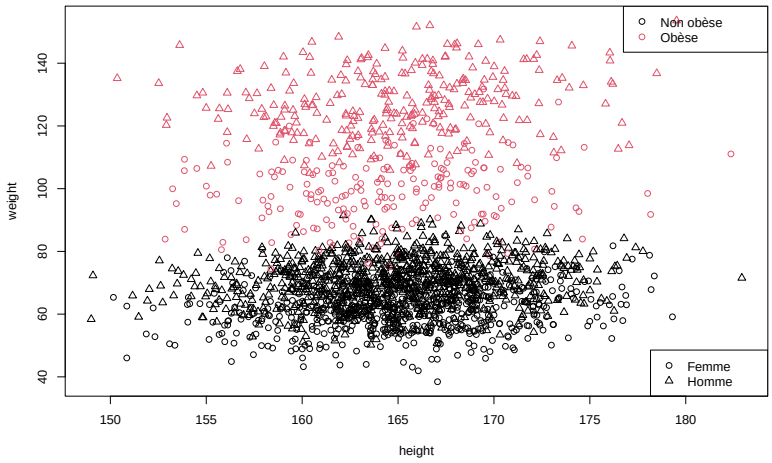


FIGURE 2.7 : Population simulée selon un modèle de mélange entre une population obèse et une population non-obèse, dont le statut n'est pas observé.

2.4.2 Algorithme EM

L'algorithme *Expectation-Maximization* (EM) est un algorithme itératif introduit par Dempster et al [17] et utilisé pour calculer le maximum de vraisemblance et estimer les paramètres d'un modèle probabiliste lorsque certaines variables sont cachées ou non observées. Il est donc couramment utilisé dans le contexte des modèles de mélange. On considère donc un modèle avec deux variables aléatoires X et S, suivant une distribution paramétrique de paramètre θ. On suppose que X est observable mais S est latente. L'estimateur du maximum de vraisemblance du modèle s'écrit donc :

$$\hat{\theta} = \arg\max_{\theta} \underbrace{\sum_{S} \mathbb{P}(X, S|\theta)}_{\mathbb{P}(X|\theta)}$$

Dans le contexte où l'espace latent est trop vaste, une des difficultés principales réside dans le fait que la vraisemblance n'est pas calculable. Le principe de l'algorithme EM est d'alterner deux étapes, E (espérance) et M (maximisation), qui garantissent que la log-vraisemblance augmente à chaque itération et, ainsi, converger ainsi vers une solution optimale.

Pour cela, il faut d'abord définir pour toute valeur du paramètre θ_{old}, la fonction auxiliaire $Q(\theta|\theta_{\text{old}})$:

$$Q(\theta|\theta_{\text{old}}) = \int_{S} \mathbb{P}(S|X;\theta_{\text{old}}) \log \mathbb{P}(X, S|\theta) dS$$

Cette dernière est égale à l'espérance $\mathbb{E}\left[\log \mathbb{P}(X, S|\theta)|X;\theta_{\text{old}}\right]$. Ensuite il suffit d'appliquer l'algorithme suivant :

Algorithme EM :

- initialiser de manière arbitraire θ_0.
- répéter jusqu'à convergence pour $i \in \mathbb{N}_+^*$
  - Etape E : calcul de l'espérance $Q(\theta|\theta_{i-1})$
  - Etape M : maximisation de la fonction $Q(\theta|\theta_i)$ pour obtenir $\theta_i = \arg\max_\theta Q(\theta|\theta_{i-1})$

Exemple :

En revenant à l'exemple proposé de modèle de mélange, pour n observations, il y a deux classes latentes $\{0,1\}$ où 0 est le groupe non-obèse et 1 le groupe obèse. On pose $S = \{S_1, ..., S_n\} \in \{0,1\}^n$ regroupant les classes de chaque observation, S n'étant pas observée. Les données observées sont $y = \{y_1, ..., y_n\}$ les poids et $X = \{X_1, ..., X_n\}$ comprenant les tailles (en cm) et les sexes (0 : homme, 1 : femme). On s'intéresse donc à la fois au problème de mélange et de régression et l'on souhaite estimer les paramètres pour la population non-obèse $\alpha = \{\alpha_0, \alpha_{sex}, \alpha_{height}\}$ et σ, pour la population obèse $\beta = \{\beta_0, \beta_{sex}, \beta_{height}\}$ et τ, et enfin la proportion d'obèse dans la population p.

On regroupe l'ensemble des paramètres du modèle en posant :

$$\theta = \{\alpha, \sigma, \beta, \tau, p\}$$

Dans un premier temps, on s'intéresse pour chaque observation $i \in \{1, ..., n\}$ à sa probabilité de faire partie de l'un ou l'autre des groupes $k \in \{0,1\}$ pour un ensemble de paramètres quelconque $\theta_{\text{old}} = \{\alpha_{\text{old}}, \sigma_{\text{old}}, \beta_{\text{old}}, \tau_{\text{old}}, p_{\text{old}}\}$ que l'on va noter

$$\eta_i(k) = \mathbb{P}(S_i = k|y_i, X_i; \theta_{\text{old}})$$

En posant le calcul, on peut facilement voir que :

$$\mathbb{P}(S_i = k|y_i, X_i; \theta_{\text{old}}) = \frac{\mathbb{P}(S_i = k, y_i, X_i|\theta_{\text{old}})}{\mathbb{P}(y_i, X_i|\theta_{\text{old}})} = \frac{\mathbb{P}(S_i = k|\theta_{\text{old}})\mathbb{P}(y_i, X_i|S_i = k, \theta_{\text{old}})}{\mathbb{P}(y_i, X_i|\theta_{\text{old}})}$$

Ainsi on note que :

- $\eta_i(0) \propto (1 - p_{\text{old}}) \times \texttt{dnorm}(y_i, \texttt{mean} = X_i\alpha_{\text{old}}, \texttt{sd} = \sigma_{\text{old}})$
- $\eta_i(1) \propto p_{\text{old}} \times \texttt{dnorm}(y_i, \texttt{mean} = X_i\beta_{\text{old}}, \texttt{sd} = \tau_{\text{old}})$

où $\texttt{dnorm}(\cdot, \texttt{mean}, \texttt{sd})$ est la densité d'une loi normale de moyenne `mean` et d'écart-type `sd`.

On s'intéresse maintenant la fonction auxiliaire $Q(\theta|\theta_{\text{old}})$ qui, par définition, vaut :

$$Q(\theta|\theta_{\text{old}}) = \sum_{i=1}^{n} \mathbb{P}(S_i = 0|y_i, X_i, \theta_{\text{old}}) \log \mathbb{P}(y_i, X_i, S_i = 0|\theta) + \mathbb{P}(S_i = 1|y_i, X_i, \theta_{\text{old}}) \log \mathbb{P}(y_i, X_i, S_i = 1|\theta)$$

En développant $\mathbb{P}(y_i, X_i, S_i|\theta)$ et en reformulant les différents termes, on peut réécrire la fonction telle que :

$$Q(\theta|\theta_{\text{old}}) = \sum_{i=1}^{n} \eta_i(0) \left(\log \mathbb{P}(y_i|X_i, S_i = 0; \theta) + \log \mathbb{P}(S_i = 0|\theta)\right) + \eta_i(1) \left(\log \mathbb{P}(y_i|X_i, S_i = 1; \theta) + \log \mathbb{P}(S_i = 1|\theta)\right)$$

Etape E :

On calcule $\eta_i(0)$ et $\eta_i(1)$ étant donné le paramètre θ_{old} :

- $\eta_i(0) = \frac{(1-p_{\text{old}}) \times \texttt{dnorm}(y_i, \texttt{mean}=X_i\alpha_{\text{old}}, \texttt{sd}=\sigma_{\text{old}})}{(1-p_{\text{old}}) \times \texttt{dnorm}(y_i, \texttt{mean}=X_i\alpha_{\text{old}}, \texttt{sd}=\sigma_{\text{old}}) + p_{\text{old}} \times \texttt{dnorm}(y_i, \texttt{mean}=X_i\beta_{\text{old}}, \texttt{sd}=\tau_{\text{old}})}$

- $\eta_i(1) = \frac{p_{\text{old}} \times \texttt{dnorm}(y_i, \texttt{mean}=X_i\beta_{\text{old}}, \texttt{sd}=\tau_{\text{old}})}{(1-p_{\text{old}}) \times \texttt{dnorm}(y_i, \texttt{mean}=X_i\alpha_{\text{old}}, \texttt{sd}=\sigma_{\text{old}}) + p_{\text{old}} \times \texttt{dnorm}(y_i, \texttt{mean}=X_i\beta_{\text{old}}, \texttt{sd}=\tau_{\text{old}})}$

Etape M :

On maximise la fonction $Q(\theta|\theta_{\text{old}})$ en calculant :

$$p_{\text{new}} = \arg\max_p \sum_{i=1}^{n} \eta_i(0) \log(1-p) + \eta_i(1) \log p$$

$$(\alpha_{\text{new}}, \sigma_{\text{new}}) = \arg\max_{\alpha,\sigma} \sum_{i=1}^{n} \eta_i(0) \mathbb{P}(y_i|X_i, z_i = 0; \alpha, \sigma)$$

$$(\beta_{\text{new}}, \tau_{\text{new}}) = \arg\max_{\beta,\tau} \sum_{i=1}^{n} \eta_i(1) \mathbb{P}(y_i|X_i, z_i = 1; \beta, \tau)$$

On peut remarquer que les nouveaux paramètres α_{new}, σ_{new}, β_{new}, τ_{new} sont estimés grâce à des régressions linéaires pondérées par les probabilités d'être dans chaque classe. Le paramètre p_{new} est atteint en $\frac{\sum_{i=1}^{n} \eta_i(1)}{n}$.

En implémentant sous R cet algorithme EM (disponible en appendice) et après 50 itérations, on trouve les résultats présentés dans la table suivante :

•	p	α_0	α_{sex}	α_{height}	σ	β_0	β_{sex}	β_{height}	τ
True parameter	0.25	20	10	0.25	7	40	30	0.35	10
Estimation	0.22	25	8	0.22	6	47	26	0.32	5

2.5 Estimation de courbe de survie pour les maladies mendéliennes

Le but de cette section est de présenter les enjeux liés à l'estimation de courbe de survie pour des maladies mendéliennes. On s'intéressera d'abord aux objectifs et problématiques associées. On présentera ensuite l'état de l'art concernant les méthodes d'estimation de courbe de survie dans ce cadre spécifique. On présentera ensuite le modèle général auquel cette thèse s'intéresse, un modèle de mélange de survies et les questions ouvertes sur ce modèle.

2.5.1 Objectifs et problématiques

Comme expliqué dans le chapitre d'introduction du contexte médical et génétique de cette thèse, un objectif majeur des médecins et généticiens est de pouvoir estimer des courbes de survie associées à différentes maladies mendéliennes (résultant de mutations sur un gène particulier) lorsque ces maladies peuvent se déclarer au cours du temps (par exemple des cancers d'origine génétique). En effet, ces courbes de survie sont généralement utilisées, ensuite, en consultation pour estimer des risques de survenue de la maladie pour les patients. Elles sont également utilisées dans des modèles, souvent basés sur des données familiales, pour estimer des risques de survenue de la maladie pour le patient et les membres de sa famille, ainsi que pour calculer des probabilités d'être porteur de mutations.

Les deux problématiques les plus importantes à prendre en compte pour faire ces estimations sont :

- la taille des jeux de données, car les mutations liées aux maladies mendéliennes sont rares en population générale, il y a donc peu de patients suivis ;
- le biais de sélection (appelé biais d'*ascertainment*) qui est dû aux règles définies (protocoles, recommandations du domaine médical) filtrant les personnes suivies.

Parallèlement, une autre problématique à prendre en compte est que, malgré le faible nombre de patients suivis pour une maladie et mutation données, il reste impossible de proposer des dépistages génétiques massivement pour des raisons économiques et matérielles. Ainsi, dans une famille touchée par une maladie mendélienne, seuls certains membres de la famille sont effectivement testés génétiquement. Le génotype dans les données de pédigrées est donc majoritairement une variable non-observée, déduite via la structure familiale.

2.5.2 État de l'art

Approche naïve

Pour estimer la courbe de survie d'une maladie mendélienne, les médecins généticiens attendent généralement d'avoir collecté des informations chez un certain nombre de patients génotypés, ce qui leur permet d'utiliser des estimateurs comme Kaplan-Meier. Cette méthode est une première approche mais elle se confronte aux différentes problématiques énoncées précédemment. Généralement le biais d'*ascertainment* est traité via des méthodes utilisées régulièrement en médecine et qui seront abordées dans la section suivante. Le nombre de patients suivis et génotypés étant souvent faible, les courbes de survie résultant de ces estimations peuvent présenter de larges intervalles de confiance.

Utilisation des données de pédigrées

Une façon d'apporter plus d'informations pour estimer ces courbes de survie est de comprendre que, dans une famille touchée par une maladie mendélienne, étant donné le mode de transmission, plusieurs personnes vont être porteuses de mutations pathogènes et peuvent être informatives au sens du calcul de la survie. Sur l'exemple donné dans la Figure 2.8, on peut remarquer que dans la famille représentée, en plus du *proband* (entouré en rouge), les membres entourés en vert sont porteurs de la mutation et pourraient donc être utilisés dans le calcul de la survie. Cependant, dans les faits, le génotype de ces porteurs n'est pas observé. Il faut donc passer par une estimation de leur probabilité d'être porteur, sachant le génotype connu (celui du *proband*), pour pouvoir l'utiliser dans une estimation.

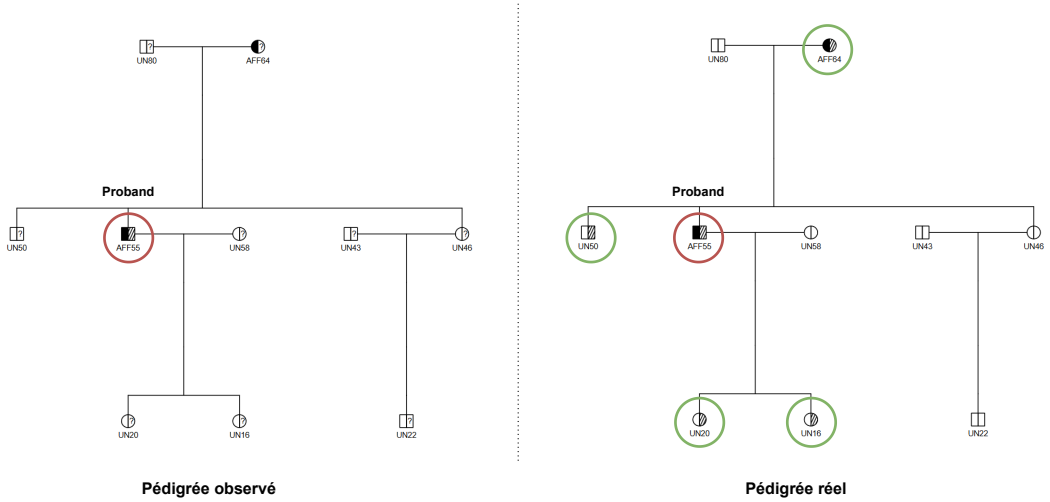


FIGURE 2.8 : Exemple de pédigrée : à droite le pédigrée effectivement recueilli par le médecin généticien, à gauche le pédigrée avec l'ensemble des génotypes. Le proband est entouré en rouge, les membres porteurs de la mutation sont entourés en vert.

Cette démarche a été utilisé par Alarcon [2] et se base sur un modèle de mélange de survie dans le cadre particulier où la maladie mendélienne considérée ne présente pas de cas sporadique, c'est-à-dire que les non-porteurs de mutation ne peuvent pas développer la maladie. La méthode développée s'appuie sur un algorithme EM (voir Section 2.4) dont l'étape E est assurée par une propagation de croyance dans le réseau bayésien représentant la structure familiale (voir Section 2.3) et l'étape M est assurée par l'estimateur de Kaplan-Meier pondéré par les probabilités d'être porteur de mutations (voir Section 2.2). Cette méthode a notamment été utilisé dans un article calculant la pénétrance de la maladie d'Alzheimer pour des variants de *SORL1* [58]. De manière similaire, plusieurs articles utilisent les données de pédigrée mais proposent une approche plus paramétrique de la fonction de pénétrance (notamment en prenant pour risque instantané une fonction de Weibull) [10, 1, 26].

2.5.3 Modèle de mélange de survie

L'utilisation des données de pédigrée entraine une augmentation de la taille des données utilisables mais au détriment d'une loi de survie qui serait unique à l'ensemble de la population observée. Ainsi, la population est composée de porteurs et de non-porteurs et l'on cherche à estimer à la courbe de survie pour les porteurs. On remarque que l'on se trouve dans le cadre d'un modèle de mélange présentant deux classes (porteurs et non-porteurs) et les lois de chaque classe sont des lois de survie.

En termes mathématiques, on considère une population composée de n individus. Les données observées sont les ages ou ages au diagnostic de la maladie $T = \{T_1, ..., T_n\} \in \mathbb{R}_+^n$ et les statuts affectés ou non $\delta = \{\delta_1, ..., \delta_n\} \in \{0, 1\}^n$ (0 : non-affecté, 1 : affecté). Étant dans un cadre de survie, on observe des évènements de censure, ici de censure à droite (voir Section 2.2).

On considère une maladie mendélienne à transmission autosomale dominante (il serait tout à fait possible de considérer, aussi, une transmission gonosomale et/ou récessive), donc monogénique et on supposera deux allèles pour ce locus (un allèle "normal" 0 et un allèle pathogène 1 dominant). On pose donc $X = \{X_1, ..., X_n\} \in \{00, 01, 10, 11\}^n$ les génotypes des individus, X est majoritairement non-observée.

Ainsi, on peut donc exprimer le modèle avec la formule suivante :

$$\mathbb{P}(T, \delta, X) = \underbrace{\mathbb{P}(X)}_{\text{Genetic Part}} \times \underbrace{\mathbb{P}(T, \delta | X)}_{\text{Survival Part}}$$

Il est intéressant de noter qu'en conditionnant le modèle selon les génotypes X, le modèle se structure en deux parties assez naturelles :

- **Partie génétique :** les données étant des pédigrées, la partie génétique a donc une structure de réseaux bayésiens (voir la section associée). En effet le génotype de chaque individu X_i est indépendant des autres individus conditionnellement à celui de ses parents que l'on note X_{pat_i} pour le père et X_{mat_i} pour la mère. Les fondateurs de la famille, c'est-à-dire les individus n'ayant pas de parents dans les données sont regroupés dans un ensemble F, suivent l'équilibre de Hardy-Weinberg et sont considérés comme indépendants. Ainsi la partie génétique se réécrit :

$$\mathbb{P}(X) = \prod_{i \in \mathrm{F}} \mathbb{P}(X_i) \times \prod_{i \notin \mathrm{F}} \mathbb{P}(X_i | X_{\text{pat}_i}, X_{\text{mat}_i})$$

- **Partie de survie :** on se retrouve dans un modèle de mélange où on considère deux classes, les non-porteurs tels que $X_i = 00$ et les non-porteurs $X_i \neq 00$. On considère que, conditionnellement au génotype, pour tout $\{i, j\} \in \{1, .., n\}^2$, les variables T_i et T_j ainsi que δ_i et δ_j sont indépendantes. Chaque classe suit donc sa propre loi de survie, définie par $S_0(t)$, $\lambda_0(t)$ pour les non-porteurs et $S_1(t)$, $\lambda_1(t)$ pour les porteurs. Ainsi la partie survie se réécrit :

$$\mathbb{P}(T_i = t, \delta_i | X_i) = \begin{cases} S_1(t)\lambda_1(t)^{\delta_i} & \text{if } X_i \neq 00 \\ S_0(t)\lambda_0(t)^{\delta_i} & \text{if } X_i = 00 \end{cases}$$

2.5.4 Question ouverte

Comme expliqué précédemment, une méthode a été proposé par Alarcon [2] pour estimer des courbes de survie à partir de ce modèle lorsque la maladie ne présente pas de cas sporadiques. D'un point de vue mathématique, l'absence de cas sporadiques se traduit par $S_0(t) = 1$, $\lambda_0(t) = 0$, c'est-à-dire une survie complète et un risque instantané de faire la maladie nul pour les individus non-porteurs de mutations. Ainsi le modèle se réécrit plus simplement :

- **Partie génétique :** elle est inchangée

$$\mathbb{P}(X) = \prod_{i \in \mathrm{F}} \mathbb{P}(X_i) \times \prod_{i \notin \mathrm{F}} \mathbb{P}(X_i | X_{pat_i}, X_{mat_i})$$

- **Partie de survie :** elle est simplifiée avec

$$\mathbb{P}(T_i = t, \delta_i | X_i) = \begin{cases} S_1(t)\lambda_1(t)^{\delta_i} & \text{if } X_i \neq 00 \\ 1 - \delta_i & \text{if } X_i = 00 \end{cases}$$

Dans ce cadre, la méthode profite du fait que les personnes affectées par la maladie sont considérés porteuses de la mutation, ce qui amène beaucoup d'informations sur les génotypes. L'incertitude se limite donc aux individus non-affectés.

Cependant, ce cas spécifique présente des limites, notamment dans le cadre des cancers d'origine génétique. En effet, les cancers ont des prévalences faibles en population générale mais potentiellement non négligeables, les cas sporadiques sont nombreux et il n'est plus

possible de considérer qu'un individu affecté est automatiquement porteur d'une mutation. Actuellement, cette méthode reste utilisée pour le calcul des courbes de survie de maladies présentant des cas sporadiques avec pour hypothèse qu'étant donnée la prévalence, le nombre de cas potentiellement sporadiques dans les données récoltées est négligeable.

Les méthodes paramétriques décrites précédemment [10, 1, 26] proposent une solution pour l'estimation de pénétrance pour des maladies génétiques présentant des cas sporadiques. Elles se basent sur l'hypothèse que l'incidence de la maladie est connue en population générale (ce qui est souvent le cas, par exemple pour les cancers) et approximent l'incidence pour les non-porteurs par l'incidence en population générale.

Chacune des méthodes développées se basent donc sur une ou plusieurs hypothèses pour prendre en compte les non-porteurs dans le calcul du risque (en les considérant comme non sujet à la maladie ou en approximant leur risque par celui de la population générale). Ces deux hypothèses étant une limite du modèle, il est nécessaire de proposer une méthode permettant de prendre en compte le risque "réel" des non-porteurs.

2.6 Biais de sélection en génétique : *ascertainment*

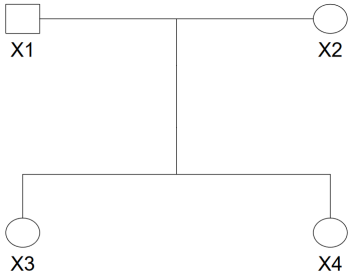
Cette section vise à présenter le biais d'*ascertainment*, à le mettre en évidence sur un exemple simple et à exposer différentes méthodes existantes utilisées pour le corriger.

2.6.1 Mise en évidence du biais d'*ascertainment*

En médecine, le biais d'*ascertainment* se produit lorsque les cas identifiés ou détectés dans une étude ne sont pas représentatifs de la population générale comme exposé dans l'étude de Sorscher sur le cancer du sein [63]. Cela peut entraîner des distorsions dans les résultats des études menées sur ces groupes et des conclusions non généralisables à l'ensemble de la population. Le terme *ascertainment* est largement utilisé en médecine ou en génétique et réfère donc à un biais d'échantillonnage. On peut noter qu'il y a différentes sources possibles à ce biais. Par exemple, l'utilisation de certains outils de diagnostic peut entraîner un biais, ou encore certaines populations peuvent être préférentiellement adressées aux cliniciens, ce qui aurait pour effet d'augmenter leur contribution dans de potentielles études cliniques. Dans cette thèse, la problématique principale est l'échantillonnage des patients vers le conseil en génétique, puis une fois le patient arrivé en consultation, l'échantillonnage vers un test génétique.

La sélection des patients auxquels on propose une consultation en génétique (puis un potentiel test) suit souvent des règles qui varient d'un pays à l'autre voir d'un hôpital à l'autre. Ces règles sont souvent dictées par des recommandations cliniques du type "famille présentant deux cas de la maladie chez des apparentés directs" ou "famille présentant un cas avant l'age de X années". Il est possible également que la sélection se fasse via l'utilisation de méthode d'aide à la décision comme le score de Manchester (voir section associée).

Pour mettre en évidence ce biais d'*ascertainment*, il est possible de simuler un jeu de données composé de familles simples dont la structure est présentée par la Figure 2.9 et de sélectionner à partir de ce jeu de données, les familles remplissant des critères donnés.



Age (Years)	Cumulative Probability for Genotype		
	AA	Aa	aa
20–29	.0167	.0167	.0002
30–39	.1444	.1444	.0027
40–49	.3758	.3758	.0138
50–59	.5477	.5477	.0275
60–69	.6743	.6743	.0497
70–79	.9452	.9452	.0798
80+	1.0000	1.0000	.1254

FIGURE 2.9 : Structure familiale utilisée pour les simulations, composée d'un père, un mère et deux filles. Risques cumulés de cancer du sein par tranche d'âge pour les différents génotypes AA, Aa et aa [15]

Les données sont générées de la manière suivante :

- 500000 familles de quatre individus : 1 père, 1 mère, 2 filles.
- Contexte d'une maladie mendélienne à transmission autosomale dominante : 1 gène, 2 allèle (allèle normal a et allèle pathogène A).
- Le génotype des parents X_1 et X_2 est déterminé par l'équilibre de Hardy-Weinberg avec une fréquence de l'allèle pathogène $f = 0.0033$
- Le génotype des filles est déterminé par transmission mendélienne, héritant aléatoirement un allèle de chaque parent.
- La maladie n'atteint que les femmes et ses lois de survie pour les porteurs et les non-porteurs de mutation suivent les risques cumulés du cancer du sein proposés par le modèle de Claus-Easton [15] rappelés en Figure 2.9, ce sont des survies exponentielles à risque instantané constant par morceaux.
- Chaque individu présente une censure suivant une loi uniforme $U[40, 80]$.
- On suppose l'ensemble des génotypes connus pour les estimations, la problématique étant de montrer le biais induit par la sélection et non pas l'estimation d'être porteur ou non.

On propose donc trois procédures de sélection qui pourraient tout à fait correspondre à des recommandations utilisées en pratiques cliniques :

- *Ascertainment* 0 : une famille est sélectionnée dans l'étude si elle présente au moins un cas de la maladie.
- *Ascertainment* 1 : une famille est sélectionnée dans l'étude si elle présente au moins un cas de la maladie survenu avant 45 ans.
- *Ascertainment* 2 : une famille est sélectionnée dans l'étude si la mère et au moins une des deux filles sont affectées par la maladie.

On se retrouve donc avec quatre populations (la population générale et trois sous-populations sélectionnées). On peut maintenant s'intéresser à différentes statistiques associées à ces populations. On peut, notamment, regarder à la fréquence de l'allèle pathogène (voir Table 2.10) et les lois de survie des porteurs et non-porteurs (voir Figure 2.10).

•	Nombre de familles	Fréquence allélique
Population générale	500000	0.0033
Ascertainment 0	52809	0.0237
Ascertainment 1	15660	0.0505
Ascertainment 2	2152	0.2105

TABLE 2.10 : Nombre de familles et fréquence de l'allèle pathogène dans la population générale et chaque *ascertainment*.

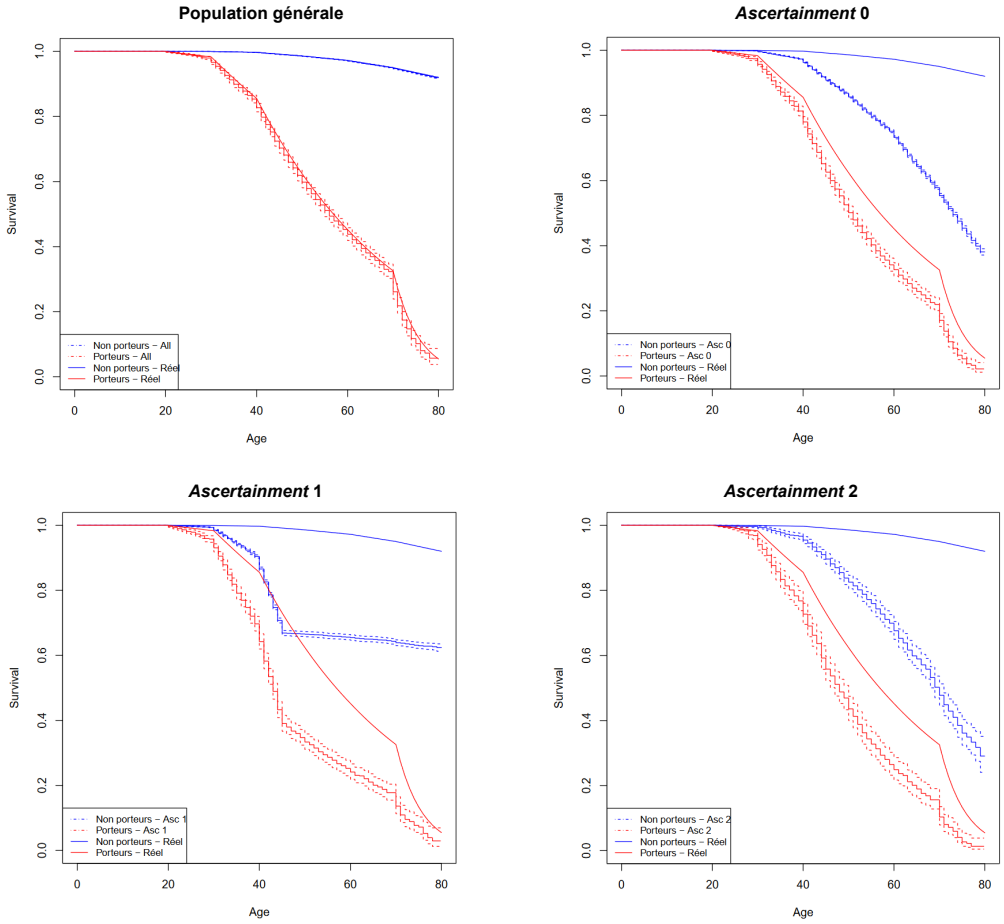


FIGURE 2.10 : Courbes de survies pour les porteurs et les non-porteurs estimées sur les *probands* via Kaplan-Meier pour la population générale et les différents *ascertainments*. Les courbes de survie réelles sont également présentes sur le graphique.

On remarque donc que la fréquence de l'allèle délétère est effectivement $f = 0.0033$ en population générale, ce qui est satisfaisant étant donné que les données sont générées à partir de cette valeur, cependant les différents *ascertainments* engendrent une augmentation de la fréquence dans les populations sélectionnées comme le montre la Table 2.10. De manière similaire, si on s'intéresse aux courbes de survie pour les *probands* porteurs et non-porteurs, on remarque qu'elles sont parfaitement respectées en population générale. Encore une fois, cela valide simplement les simulations. Cependant, pour les sous-populations sélectionnées, l'estimation est complètement biaisée avec une forte diminution de la survie pour les porteurs et non-porteurs dans les trois *ascertainments* et, dans l'*ascertainment* 1, une très grosse surreprésentation des cas avant 45 ans, comme le montre la Figure 2.10. Ces exemples

montrent donc à quel point le biais de sélection dans le cadre des données de génétique est important à considérer.

2.6.2 Méthodes de correction

Le biais d'*ascertainment* est un problème difficile car il suppose qu'une partie de la population n'est pas correctement échantillonnée. De fait, le problème est comment tenir compte d'une population peu ou pas observée dans les études. En génétique particulièrement, il existe notamment deux méthodes de correction, la *Proband's phenotype exclusion likelihood* (PEL) [1] et la *Genotype-restricted likelihood* (GRL) [10]. Le biais de sélection n'étant pas une problématique exclusive de la médecine ou de la génétique, il existe des méthodes utilisées dans d'autres domaines, par exemple les sondages. Le redressement statistique englobe un certain nombre de méthodes qui consistent à pondérer la population étudiée pour qu'elle se rapproche de la population cible en terme de statistiques. Dans le cas de cette thèse, on s'intéressera plus précisément à la méthode du *raking*. L'idée de cette section est simplement de présenter succinctement ces méthodes mais elles ne seront pas mises en application sur les données générées. Les chapitres 3 et 6 incluent des mises en applications de ces méthodes.

Proband's phenotype exclusion likelihood

La *Proband's phenotype exclusion likelihood* (PEL) [1] est une méthode de correction utilisée lorsque le critère d'*ascertainment* implique notamment une information sur l'âge de survenue le maladie (par exemple deux cas de la maladie avant 45 ans dans la famille). Le principe de la méthode s'appuie sur le fait que, dans ce type d'*ascertainment*, ce sont les phénotypes des *probands* qui ont permis d'échantillonner la famille et qui introduisent le biais. Pour le corriger, la PEL propose donc d'oublier le phénotype des *probands* lors des estimations faites sur la population sélectionnée.

En d'autres termes, le *proband* devient non informatif du point de vue de la maladie, mais ces caractéristiques, notamment son génotype s'il est disponible, reste utile.

Genotype-restricted likelihood

La *Genotype-restricted likelihood* (GRL) [10] est une méthode de correction utilisée lorsque le critère d'*ascertainment* implique plutôt une information génotypique (par exemple un porteur de variants pathogènes dans la famille). La correction vise à calculer la probabilité d'un génotype familial conditionnellement aux génotypes familiaux ayant permis échantillonner la famille. La problématique principale de cette méthode est que les règles d'*ascertainment* peuvent souvent être variables au sein d'un même jeu de données.

Redressement statistique : *raking*

Le *raking* n'est pas une méthode spécifiquement utilisée sur des données médicales et fait partie d'un ensemble de techniques appelées méthodes pondérées [34] ou redressement statistique. L'idée principale de la méthode est d'attribuer des poids à chaque observation de la population sélectionnée pour que ses statistiques pondérées se rapprochent le plus possible des statistiques d'une population cible (par exemple la population générale).

En prenant un exemple très simple : on interroge 100 personnes dont 30 sont des hommes et 70 des femmes et on leur demande "Buvez-vous du café ?". La question à laquelle on souhaite répondre est "combien de personnes boivent-elles du café ?" en population générale dans laquelle on sait qu'il y a 50% d'hommes et de femmes. Si on reprend les 100 observations, on remarque que la population observée n'est pas représentative de la population générale car elle présente 70% de femmes et 30% d'hommes. Les méthodes pondérées proposent donc

de corriger la population observée en donnant un poids aux 30 hommes leur permettant de représenter 50% de la population et de même pour les femmes.

L'exemple donné se limite à une seule variable mais le *raking* permet d'utiliser un grand nombre de variables pour stratifier au mieux la population étudiée. D'un point de vue mathématique, on considère que l'étude est faite sur n personnes pour lesquels on a collecté d variables dont on connait la valeur moyenne dans la population cible m_j (pour la variable $j \in \{1, ..., d\}$). On pose w_i le poids d'un individu $i \in \{1, ..., n\}$ de la population étudiée, x_{ij} la valeur de la variable j pour l'individu i, λ_j sont les multiplicateurs de Lagrange pour poser des contraintes affines et enfin G est une fonction permettant de poser des contraintes sur les poids (ici pour avoir des poids proches de 1). On peut poser le problème d'optimisation suivant :

$$\ell(w, \lambda) = \sum_{i=1}^{n} G(w_i) - \sum_{j=1}^{d} \lambda_j \sum_{i=1}^{n} (w_i x_{ij} - m_j) \quad \text{avec} \quad G(w) = w \log w - w + 1$$

dont on estime les solutions en initialisant les poids w_i à 1 puis en appliquant, par exemple, un algorithme de Newton-Raphson.

2.7 Questions de recherche

Cette introduction a permis de mettre en évidence d'une part le contexte médical de cette thèse et d'autre part d'introduire la plupart des outils mathématiques et méthodologiques utilisés par la suite. Pour résumer, cette thèse s'intéresse principalement aux maladies mendéliennes (c'est-à-dire d'origine monogénique) qui surviennent au cours du temps et sont donc diagnostiquées ou non au cours de la vie des patients. Il existe plusieurs exemples de maladies ayant ces caractéristiques mais cette thèse porte particulièrement sur les cancers broncho-pulmonaires ayant des origines génétiques (comme les variants pathogènes des gènes *SFTPA1* et *SFTPA2*, *TP53* - syndrôme de Li-Fraumeni - ou encore *EGFR*).

Dans le cadre des consultations de génétique, les médecins sont demandeurs d'outils d'aide à la décision, que ce soit pour le prédiction des variants pathogènes ou l'estimation de risque de survenue des maladies considérées pour les patients porteurs de ces variants et leur famille. Cette estimation du risque est souvent calculée à partir des courbes de survie/pénétrance. L'estimation de ces courbes de pénétrance est donc un enjeu primordial pour la médecine génétique et fait face à différentes problématiques. Tout d'abord, les maladies mendéliennes étant liées à des variants pathogènes dont la fréquence allélique est très faible en population générale, les patients suivis sont généralement peu nombreux. Ensuite les patients suivis le sont généralement car ils sont sélectionnés dans leur parcours de soin sur certains critères (nombre important de cas dans la famille, cas survenu à un jeune age, etc), cela implique un biais de sélection, nommé biais d'*ascertainment*, qui doit être pris en compte.

Dans le cadre des maladies génétiques, les données familiales (pédigrées) comprenant les liens de parentés entre les individus, les données de survie associées (age ou age au diagnostic pour chaque individu) et les génotypes disponibles sont utilisées pour augmenter la taille des jeux de données et permettre l'estimation de courbe de pénétrance. Elles sont relativement faciles et peu coûteuses à collecter d'une part et elles sont informatives au sens de la survie/pénétrance d'autre part car les apparentés d'un patient porteur d'un variant pathogène sont également susceptibles d'être porteurs. Il existe une méthode publiée pour estimer les courbes de pénétrance d'une maladie génétique à partir de données de pédigrée [2] dans le cas où la maladie ne présente pas de cas sporadique (c'est-à-dire que seuls les individus porteurs d'un variant pathogène peuvent être affectés), ce qui est une hypothèse

limitante dans le cas du cancer broncho-pulmonaire par exemple qui présente effectivement des cas sporadiques.

Cette thèse se place dans ce cadre de recherche et se propose de répondre à plusieurs questions à la fois médicales et méthodologiques :

- Estimer des courbes de pénétrance de maladies mendéliennes notamment liées au poumon pour des porteurs de variants pathogènes récemment découverts. Le chapitre 4 se consacre, par exemple, à l'estimation de courbes de pénétrance de la pneumopathie interstitielle et du cancer broncho-pulmonaire pour les porteurs de variants pathogènes *SFTPA1* et *SFTPA2* (en faisant l'hypothèse que les maladies considérées ne présentent pas de cas sporadiques).

- Estimer des courbes de pénétrance de maladies mendéliennes présentant des cas sporadiques à partir de données de pédigrée. Le chapitre 5 se consacre au développement d'une méthode d'estimation prenant en compte les cas sporadiques avec pour but de réestimer la courbe de pénétrance du cancer broncho-pulmonaire chez les porteurs de variants pathogènes *SFTPA1* et *SFTPA2*.

- Prendre en compte le biais d'*ascertainment* dans les études de génétiques. Le chapitre 3 se consacre à la correction du biais d'*ascertainment* dans une étude d'évaluation d'un score de prédiction de variants pathogènes pour les cancers sein/ovaire par la méthode du raking. Le chapitre 6 se consacre à l'étude de différentes méthodes pour intégrer une correction du biais d'*ascertainment* dans la méthode développée au chapitre 4.

Sur l'ensemble de ces projets, les outils mathématiques principaux sont :

- l'analyse de survie pour la modélisation de la pénétrance
- les réseaux bayésiens pour la modélisation des génotypes familiaux
- l'algorithme somme-produit pour les calculs de vraisemblance et de lois marginales dans les réseaux bayésiens
- l'algorithme EM pour le calcul de maximum de vraisemblance dans des modèles à variables latentes ou des méthodes d'optimisation tel BFGS si les modèles sont paramétriques.
- le raking, la PEL et la GRL comme méthodes de correction du biais d'ascertainment selon les problématiques considérées.

Appendices to Chapter 2

Cet appendice présente du code en R pour les différentes parties proposées.

2.A Analyse de survie

2.A.1 Simulations de données de survie avec censure sous R

Le but est de simuler un jeu de données comprenant $2n$ individus repartis dans 2 groupes de même taille, n suivent une loi de survie exponentielle avec risque constant par morceaux et les n autres suivent la même loi mais avec un risque proportionnel supplémentaire β.

```
# import survival
require(survival)

# define piecewise constant hazard generative functions
# Hazard function
pch.haz=function(time,cuts,alpha) stepfun(cuts,alpha)(time)

# Cumulative Hazard function
pch.cumhaz=function(time,cuts,alpha) {
  if (length(cuts)==0) return(time*alpha[1])
  k=length(alpha)
  I=as.numeric(cut(time,breaks=c(-1e-10,cuts,Inf),labels=1:k))
  cuts0=c(0,cuts)
  return(alpha[I]*(time-cuts0[I])+c(0,cumsum(alpha*c(diff(cuts0),0))[-k])[I])
}

# PCH random number generative function
pch.rsurv=function(n,cuts,alpha) {
  u=runif(n);
  k=length(alpha);
  if (length(cuts)!=(k-1)) stop("length(cp) must be equal to length(alpha)-1")
  cuts0=c(0,cuts)
  if (k>1) {
    thresh=exp(-cumsum(alpha[-k]*diff(cuts0)))
    if (n<=200) {
      seg=apply(matrix(rep(thresh,n),byrow=TRUE,nrow=n)>u,1,sum)+1
    } else {
      seg=rep(NA,n)
      for (i in 1:n)
        seg[i]=sum(thresh>u[i])+1
```

```
    }
  } else {
    seg=rep(1,n)
  }
  res=cuts0[seg]-(log(u)+cumsum(c(0,alpha[-k]*diff(cuts0)))[seg])/alpha[seg]
  return(res)
}
```

Dans un premier temps, il faut définir les différents paramètres, les t_1,...,t_k (cuts), les λ_1,...,λ_k (lambda) et β (beta) :

```
#parameters
cuts=c(20,40,60,80)
lambda_1=c(0,2,6,9,6)/1000
beta=0.3
lambda_2=lambda_1*exp(beta)

#survival and hazard functions for the 2 subgroups
S1=function(t) exp(-pch.cumhaz(t,cuts,lambda_1))
lambda1=function(t) pch.haz(t,cuts,lambda_1)

S2=function(t) exp(-pch.cumhaz(t,cuts,lambda_2))
lambda2=function(t) pch.haz(t,cuts,lambda_2)

#plot of survival and hazard functions
age=seq(from =0, to = 90, by =0.1)
plot(age,lambda2(age),type = "l",col = "blue",xlab="time",ylab = "Hazard",
        xlim = c(0,85),ylim = c(0,0.015) )
lines(age,lambda1(age),col = "red")
legend(0, 0.013, legend=c("Lambda␣group2", "Lambda␣group1"),
       col=c("blue", "red"), lty=1:1, cex=0.8)

plot(age,S2(age),type = "l",col = "blue",xlab="time",ylab = "Survival",
        xlim = c(0,85),ylim = c(0,1))
lines(age,S1(age),col = "red")
legend(0, 0.1, legend=c("Survival␣group2", "Survival␣group1"),
       col=c("blue", "red"), lty=1:1, cex=0.8)
```

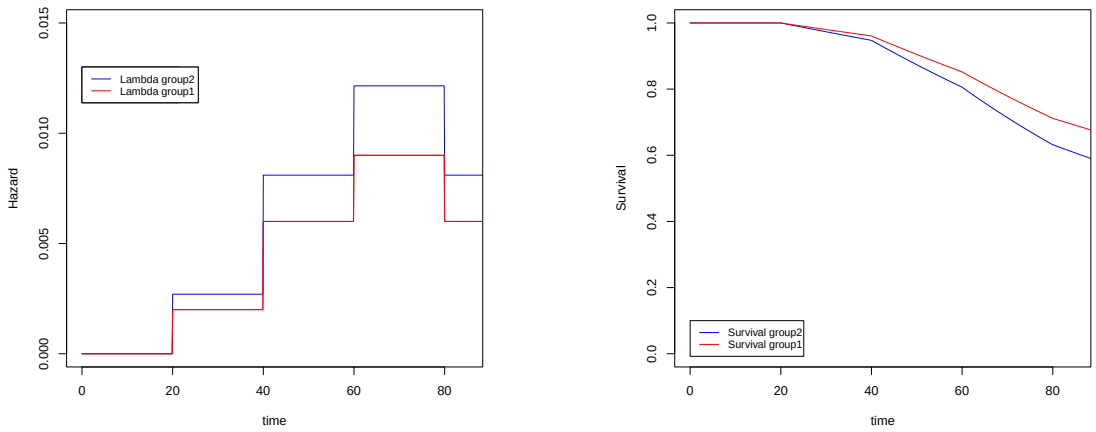


TABLE 2.11 : Fonctions de risque instantané et de survie pour les deux groupes.

Les données de survie peuvent maintenant être générées grâce au code suivant :

```
n=200

#forming 2 groups
Group<-c(rep(1,n),rep(2,n))

#Generate time to event for each individual
TrueT<-c(pch.rsurv(n,cuts,lambda_1),pch.rsurv(n,cuts,lambda_2))

#Generate a potential censored time for each individual as U(40,90)
Cens<-runif(n,40,90)

#Calculat the observed time and status
ObsT<-pmin(TrueT,Cens)
Delta<-as.numeric(TrueT<=Cens)
df<-data.frame(TrueT,Cens,ObsT,Delta,Group)
```

2.A.2 Estimateur de Kaplan-Meier sous R

Maintenant que les données de survie sont générées et que les distributions de survie des deux groupes sont connues, il est intéressant d'essayer d'estimer ces survies grâce aux méthodes existantes comme Kaplan-Meier.

```
#compute Kaplan-Meier estimator stratified for each group
fit=survfit(Surv(ObsT,Delta)~Group,data = df)

#import survminer for survival graph plots
library(survminer)

#plot KM estimator
ggsurvplot(fit, combine = TRUE, # Combine curves
           #risk.table = TRUE,                  # Add risk table
           conf.int = TRUE,                    # Add confidence interval
           conf.int.style = "step",            # CI style, use "step" or "ribbon
           #censor = FALSE,
           ggtheme = theme_bw(),# Remove censor points
           tables.theme = theme_cleantable(),  # Clean risk table
           palette =  c("red","blue"),
           linetype = 1,
           # color = c("red","blue"),
           ylim=c(0,1),xlim=c(0,85),
           xlab="Time")
```

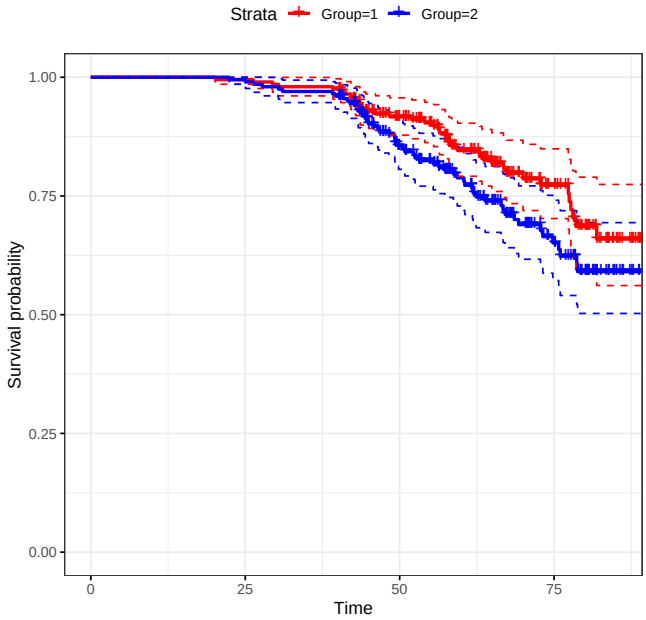


TABLE 2.12 : Estimation des fonctions de survie avec Kaplan-Meier.

2.A.3 Modèle de Cox sous R

Étant donné que les données générées présentent 2 groupes distincts donc le risque relatif est une constante β, le modèle de Cox est tout indiqué pour estimer ce coefficient.

```
#Cox proportional hazard
cox=coxph(Surv(ObsT,Delta)~Group,data=df)
cox


##Call:
##coxph(formula = Surv(ObsT, Delta) ~ Group, data = df)
##
##         coef exp(coef) se(coef)     z      p
##Group 0.3856    1.4705   0.2094 1.842 0.0655
##
##Likelihood ratio test=3.45  on 1 df, p=0.06341
##n= 400, number of events= 94
```

2.B Modèles de mélange et algorithme EM

2.B.1 Modèle de mélange

On fixe les valeurs des paramètres comme suit :

- On considère 2000 individus $n = 2000$
- La probabilité d'être obèse est $p = 0.25$
- La variance de la loi normale pour les non-obèses est $\sigma = 7$
- La variance de la loi normale pour les obèses est $\tau = 10$
- Les paramètres pour les non-obèses $\alpha_0 = 20$, $\alpha_{\text{sex}} = 10$, $\alpha_{\text{height}} = 0.25$

- Les paramètres pour les non-obèses $\beta_0 = 40$, $\beta_{\text{sex}} = 30$, $\beta_{\text{height}} = 0.35$

A partir de ces valeurs, il est possible de simuler des données selon ce modèle avec le code suivant :

```
n=2000
p=0.25
sigma=7
tau=10
# tau=10
set.seed(42)
# very simple covariate matrix
X=cbind(rep(1,n),
        sample(0:1,size=n,replace=TRUE,prob=c(0.45,0.55)),
        rnorm(n,165,5))
# arbitrary parameters
alpha=c(20,10,0.25)
beta=c(40,30,0.35)

z=sample(0:1,size=n,replace=TRUE,prob=c(1-p,p))
y=rep(NA,n)
y[z==0]=rnorm(sum(z==0),(X[z==0,]%*%alpha)[,1],sigma)
y[z==1]=rnorm(sum(z==1),(X[z==1,]%*%beta)[,1],tau)

plot(X[,3],y,pch=X[,2]+1,col=X[,2]+1,xlab="height",ylab="weight")

rbind(
  alpha,
  coefficients(lm(y[z==0]~X[z==0,]-1))
)
rbind(
  beta,
  coefficients(lm(y[z==1]~X[z==1,]-1))
)
```

On obtient ainsi une population étant le résultant d'un mélange comme le montre la Figure 2.11.

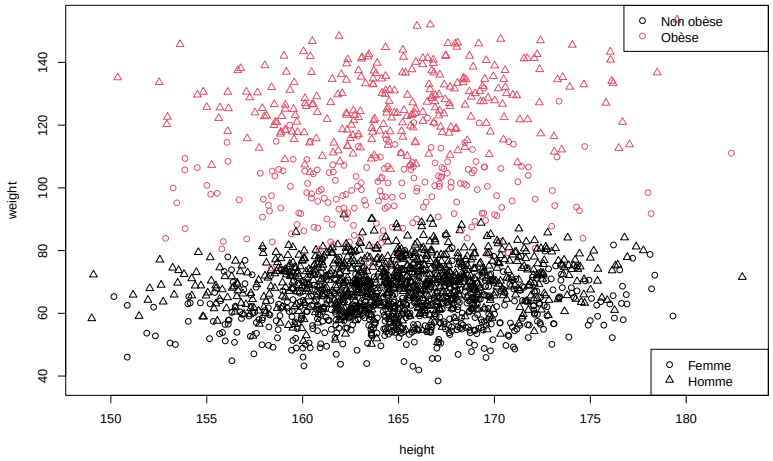


FIGURE 2.11 : Population simulée selon un modèle de mélange entre une population obèse et une population non-obèse, dont le statut n'est pas observé.

2.B.2 Algorithme EM

On rappelle la fonction Q et les étapes E et M :

$$Q(\theta|\theta_{\text{old}}) = \sum_{i=1}^{n} \eta_i(0)\left(\log \mathbb{P}(y_i|X_i, S_i = 0; \theta) + \log \mathbb{P}(S_i = 0|\theta)\right) + \eta_i(1)\left(\log \mathbb{P}(y_i|X_i, S_i = 1; \theta) + \log \mathbb{P}(S_i = 1|\theta)\right)$$

Etape E :

On calcule $\eta_i(0)$ et $\eta_i(1)$ étant donné le paramètre θ_{old} :

- $\eta_i(0) = \frac{(1-p_{\text{old}})\times \texttt{dnorm}(y_i, \texttt{mean}=X_i\alpha_{\text{old}}, \texttt{sd}=\sigma_{\text{old}})}{(1-p_{\text{old}})\times \texttt{dnorm}(y_i, \texttt{mean}=X_i\alpha_{\text{old}}, \texttt{sd}=\sigma_{\text{old}}) + p_{\text{old}}\times \texttt{dnorm}(y_i, \texttt{mean}=X_i\beta_{\text{old}}, \texttt{sd}=\tau_{\text{old}})}$
- $\eta_i(1) = \frac{p_{\text{old}}\times \texttt{dnorm}(y_i, \texttt{mean}=X_i\beta_{\text{old}}, \texttt{sd}=\tau_{\text{old}})}{(1-p_{\text{old}})\times \texttt{dnorm}(y_i, \texttt{mean}=X_i\alpha_{\text{old}}, \texttt{sd}=\sigma_{\text{old}}) + p_{\text{old}}\times \texttt{dnorm}(y_i, \texttt{mean}=X_i\beta_{\text{old}}, \texttt{sd}=\tau_{\text{old}})}$

Etape M :

On maximise la fonction $Q(\theta|\theta_{\text{old}}$ en calculant :

$$p_{\text{new}} = \arg\max_p \sum_{i=1}^{n} \eta_i(0)\log(1-p) + \eta_i(1)\log p$$

$$(\alpha_{\text{new}}, \sigma_{\text{new}}) = \arg\max_{\alpha,\sigma} \sum_{i=1}^{n} \eta_i(0)\mathbb{P}(y_i|X_i, z_i = 0; \alpha, \sigma)$$

$$(\beta_{\text{new}}, \tau_{\text{new}}) = \arg\max_{\beta,\tau} \sum_{i=1}^{n} \eta_i(1)\mathbb{P}(y_i|X_i, z_i = 1; \beta, \tau)$$

On peut remarquer que les nouveaux paramètres α_{new}, σ_{new}, β_{new}, τ_{new} sont estimés grâce à des régressions linéaires pondérées par les probabilités d'être dans chaque classe. Le paramètre p_{new} est atteint en $\frac{\sum_{i=1}^{n}\eta_i(1)}{n}$.

```
 cat("True parameters:","p=",p,"alpha=",alpha,"sigma=",
                                          sigma,"beta=",beta,"tau=",tau,"\n")

# slighty cheated initialization
eta=matrix(runif(2*n),n,2)
eta[z==0,1]=eta[z==0,1]+1
eta[z==1,2]=eta[z==1,2]+1
eta=eta/apply(eta,1,sum)

# main loop
max_loop=50
for (iter in 1:max_loop) {
  # M step
  p=mean(eta[,2])
  reg0=lm(y~X-1,weights=eta[,1])
  reg1=lm(y~X-1,weights=eta[,2])
  alpha=coefficients(reg0); sigma=summary(reg0)$sigma
  beta=coefficients(reg1); tau=summary(reg1)$sigma
  # E step
```

```
  eta=cbind((1-p)*dnorm(y,(X%*%alpha)[,1],sigma),
            p*dnorm(y,(X%*%beta)[,1],tau))
  eta=eta/apply(eta,1,sum)
  # verbose

}
cat("Estimations:␣")
cat("iter=",max_loop,"p=",p,"alpha=",alpha,"sigma=",
                          sigma,"beta=",beta,"tau=",tau,"\n")

## True parameters: p= 0.25 alpha= 20 10 0.25 sigma= 7
## True parameters: beta= 40 30 0.35 tau= 10
##
## Estimations:
## iter= 50 p= 0.2279974 alpha= 25.69071 8.183923 0.2235908
## sigma= 6.669284 beta= 47.60053 26.13588 0.3251815 tau= 4.629494
```

Chapitre 3

Comparaison de performance entre score de Manchester et modèles familiaux pour la détection de prédisposition génétique au cancer sein/ovaire

Résumé

Les recommandations récentes aux États-Unis préconisent un dépistage généralisé des anomalies génétiques germinales chez les patientes atteintes de cancer du sein et d'ovaire (BC/OC), tandis que les oncogénéticiens européens font face à des contraintes de ressources, nécessitant la sélection des cas à tester. Cette étude, menée au laboratoire d'oncogénétique de l'AP-HP Sorbonne Université à Paris, évalue le score de Manchester adapté aux panels multigènes et aux directives françaises (MSS-F) en tant qu'outil de prise de décision pour la sélection des patients se voyant proposer des tests génétiques.

L'étude a porté sur 1220 patientes atteintes de cancer du sein et d'ovaire de 2016 à 2020. Elle était rétrospective de 2016 à 2017 et prospective de 2018 à 2020, le MSS-F étant utilisé pour guider les décisions de test au cours de cette dernière période de 2 ans. Étonnamment, les performances du MSS-F dans la prédiction des variants pathogènes dans cette population étaient inférieures aux attentes, soulevant deux questions :

- Le MSS-F devrait-il être utilisé plus tôt dans le parcours de soins pour sélectionner les patients qui devraient être adressés aux généticiens ?
- Le MSS-F devrait-il être remplacé par des modèles familiaux tels que BOADICEA, BRCApro et Claus-Easton pour prédire les variants pathogènes ?

Pour comparer le MSS-F avec les modèles familiaux, les antécédents familiaux de 210 patientes ont été numérisés. Pour simuler une utilisation du MSS-F plus tôt dans le parcours de soins, la méthode du raking a été utilisée afin que la population pondérée corresponde à une population ciblée de patientes non sélectionnées atteintes de cancer du sein et d'ovaire.

D'une part, les résultats ont révélé une faible performance du MSS-F et des modèles familiaux dans la prédiction des variants pathogènes des panels multigènes, avec BOADICEA présentant une AUC de 0.61 et 0.57 pour le MSS-F. Même en prédisant uniquement *BRCA1/BRCA2*, les performances restent moyennes. D'autre part, après l'utilisation du raking, l'AUC du MSS-F s'est nettement améliorée à 0.74, suggérant que le MSS-F fonctionne mieux sur une population non sélectionnée de femmes atteintes de cancer du sein et ovaire entrant dans le parcours de soin.

En conclusion, dans le contexte du conseil génétique, les modèles familiaux n'ont pas surpassé le MSS-F, ce qui signifie que les outils d'aide à la décision pour la prédiction de variants pathogènes ne sont pas actuellement suffisants pour le cancer sein/ovaire. L'amélioration des performances du MSS-F sur des cas de cancer du sein et d'ovaire non sélectionnés suggère qu'il pourrait être plus efficace en amont pour orienter les patientes vers l'oncogénétique. Par conséquent, l'étude suggère la nécessité d'améliorer l'accès aux tests génétiques germinaux pour tous les cas de cancer du sein et d'ovaire adressés aux généticiens.

Valorisation : Ce projet est réalisé de manière collaborative par Patrick R. Benusiglio, Lucas Ducrot, Erell Guillerm, Jasmine Hasnaoui, Florence Coulet et Gregory Nuel. Il est financé par AAP Emergence, SIRIC CURAMUS, AP-HP (Sorbonne Université 2019ECUR07). Les résultats ont été présentés par un poster à ESHG annual meeting 2022 ainsi que dans un rapport INSERM. Pour une publication, il faudrait que les scores de Manchester (MSS3) soient calculés pour l'ensemble des 1220 patients afin de pouvoir réaliser une véritable étude comparative entre MSS-F et MSS3, ce qui n'a pas été possible au cours de cette thèse.

Contents

3.1 Abstract

Recent recommendations in the USA advocate for widespread germline testing in breast and ovarian cancer patients (BC/OC), whereas European oncogeneticists face resource constraints, necessitating the selection of cases for testing. This study, conducted at AP-HP Sorbonne University Oncogenetics laboratory in Paris, evaluates the Manchester Scoring System adapted for multigene panels and French guidelines (MSS-F) as a decision-making tool in selecting patients for germline testing.

The study encompassed 1220 breast cancer and ovarian cancer patients from 2016 to 2020. It was retrospective from 2016 to 2017 and prospective from 2018 to 2020 with MSS-F being used to guide testing decisions over this 2-year period. Surprisingly, MSS-F performance in predicting pathogenic variants on this population was lower than expected, prompting two questions :

- Should the MSS-F be used earlier in the care pathway to select patient that should be addressed to the geneticists ?
- Should the MSS-F be replace with family-based models like BOADICEA, BRCApro, and Claus-Easton in order to predict pathogenic variants ?

In order to compare MSS-F with familial models, the familial histories of 210 patients were digitized. To mimic a use of the MSS-F earlier in the care pathway, the raking method was used so the weighted population fit a targeted population of unselected breast and ovarian cancer patients.

On one hand the results revealed poor performance for MSS-F and familial models in predicting pathogenic variants of the multigene panels, with BOADICEA showcaseing an AUC of 0.61 and 0.57 for the MSS-F. Even when predicting only *BRCA1/BRCA2*, performances remain poor. On the other hand, after raking, MSS-F's AUC improved significantly to 0.74 suggesting that the MSS-F performs better on a unselected population of women with breast and ovarian cancer entering the medical care.

In conclusion, in the genetics counseling context, family-based models did not outperform MSS-F meaning that decision-making tools to predict pathogenic variants are not suffisant currently. MSS-F's improved performance on unselected breast/ovarian cases suggests it might be more effective upstream for patient referral to oncogenetics. Therefore the study suggests a need for an effort to enhance access to germline genetic testing for all breast/ovarian cases referred to geneticists.

3.2 Introduction

Recent studies from the United States suggest that germline genetic testing should be offered to all women with breast and ovarian cancer, as this might be the only way to identify all carriers of pathogenic variants (PV) in susceptibility genes [16, 75]. However universal genetic testing is yet to be widely implemented, especially in breast cancer. Indeed, there is a lack of trained professionals to provide genetic counseling, and limited resources and funding to carry out molecular analyses at such a large scale. In addition, there are concerns regarding psychosocial harms, as well as medical benefit, for example for carriers of variants of unknown significance (VUS), or of non-actionable variants associated with only low to moderate cancer risk [37, 31]. Patient selection is therefore very much still in place, using consensual criteria, logistic regression scores or PV prediction models [59, 44, 25, 32]. The most commonly cited tool is CanRisk, a mendelian model that takes into account personal and family history, and also family structure [11]. There have even been recent multidisciplinary efforts to develop more accurate selection tools [77, 54]. The situation is different regarding ovarian cancer, for which universal testing is feasible, given the lower number of new cases. The Manchester scoring system (MSS) is a model for assessing probability of a germline PV in *BRCA1* or *BRCA2* [25, 18]. MSS3, the latest version, was published in 2017 [25]. MSS3 integrates personal and familial clinical and pathological data. Each individual and family feature (from one side of the family) is given a numerical weight. These are then added to give a score that is converted into the probability of finding a *BRCA1/2* germline PV. Since 2018, we have been using MSS-F, a modified version of MSS3. We made empirical adjustments to account for multigene panels, and to fit with French clinical recommendations. In this study, we estimated the accuracy of MSS-F in predicting germline status in a large series of French breast and ovarian cases. The underlying question was whether it was suited to the demands of modern cancer genetics, considering the wide-ranging implications of a germline PV regarding treatment, surveillance and risk-reduction, or whether we should aim for universal testing in the coming years.

3.3 Material and methods

3.3.1 Patients

We included all female patients with breast or ovarian cancer in whom germline testing was performed between 01/2016 and 12/2020 at the Assistance Publique Hôpitaux de Paris – Sorbonne University (APHP-SU) Oncogenetics laboratory. Patients had been seen at the affiliated APHP-SU Cancer Genetics clinics, and by selected cancer specialists from regional hospitals taking part in our mainstreaming program [8]. The following genes were tested via a hereditary breast and ovarian cancer (HBOC) multigene panel : *BRCA1*, *BRCA2*, *ATM*, *CHEK2*, *CDH1*, *MLH1*, *MSH2*, *MSH6*, *PALB2*, *PMS2*, *PTEN*, *RAD51C*, *RAD51D*, and *TP53*.

3.3.2 Pathogenic variants probability estimation

MSS-F, an empirical, modified version of the MSS3 was used to determine germline testing indication (Table 3.1). MSS-F was based on the published, validated score but minor modifications were made, in order to account for multigene testing, and to fit with French national practice in place at the time. The following changes were made :

- Suppression of the minus points associated with HER2+++, as not to miss *TP53* PV since the majority of *TP53*-associated breast cancers are HER2+++ [49].
- Suppression of the minus points associated with the lobular breast cancer type, as not to miss CDH1 PV since CDH1-associated breast cancer are lobular [76, 19].
- Increase in the breast cancer diagnosis upper age limit associated with 11 points from 30 to 36 years, as all breast cancers diagnosed under this age were already offered germline testing.
- For ovarian cancer, suppression of the 60 year age limit for the high-grade serous type. The two extrapoints were thus awarded for this subtype regardless of age, since it is also predictive of genetic susceptibility in older patients [62]. In addition, germline testing was already considered in all ovarian cancer patients at the time of the study.
- Doubling of the score for a patient of Ashkenazi origin (Gareth Evans, personal communication)

For patients tested between 2018 and 2020, MSS-F was calculated prospectively. The recommendation was to offer genetic testing to all patients with an MSS-F $\geqslant$ 12 points, and to consider testing in patients with a score of 9-11 points. The 12-point threshold likely corresponds to a 5-10% *BRCA1/2* PV probability [25, 24]. These recommendations were however non-binding, and the clinician could decide to test patients with lower scores. For patients tested in 2016-1017, MSS-F was calculated retrospectively. For a random set of 210 patients, we also estimated PV probability in *BRCA1*, *BRCA2*, *ATM*, *CHEK2* and *PALB2* using CanRisk [11].

3.3.3 Statistics

MSS-F correlation with the identification of PV in one of the HBOC genes was assessed using AUC/ROC curves. Sensitivity, specificity, positive predictive value (PPV) and negative predictive value (NPV) were calculated at the 15-, 12- and 9-point testing threshold. We also ran CanRisk on a random selection of 210 patients and calculated the associated AUC [11]. Our case series consisted of selected high-risk patients referred to Oncogenetics. This

could have resulted in an underestimation of the performances of MSS-F. We used the raking method to correct for this selection bias [34]. With raking, one computes weight for every individual in the observed data so that the statistics (over multiple variables) match those of a target population. We therefore matched our population to an unselected series of 274 women with breast and ovarian cancer seen in the Gynecologic Oncology clinic, and in whom MSS-F had been previously calculated for research purposes [72]. We recalculated AUC using raking.

3.4 Results

We included 1219 patients, 1182 of them female, 37 male. Mean MSS-F score was 14 (range 1-46). AUC was 0.61 (0.56-0.66) (Figure 3.1). Sensitivity increased with decreasing testing thresholds. It was 0.95, 0.75 and 0.55 at the 9, 12 and 15-point threshold, respectively (Table 3.1). As for specificity, it was 0.14, 0.36 and 0.63 at these thresholds. NPV varied little ; it was 0.96 at the 9-point threshold, and 0.93 at the 12- and 15-point thresholds. PPV was 0.10, 0.11 and 0.13, at the 9-, 12-, and 15-point threshold respectively (Table 3.1). The CanRisk AUC for the 210 randomly selected cases was 0.61 (0.51-0.71), similar to the one obtained with MSS-F in these specific cases. Mean MSS-F in the 274 unselected patients from the Gynecologic Oncology clinic was 6 (range 0-35). We matched our Oncogenetics cases to these using raking. MSS-F performances were improved, with an AUC of 0.74 (0.67-0.88).

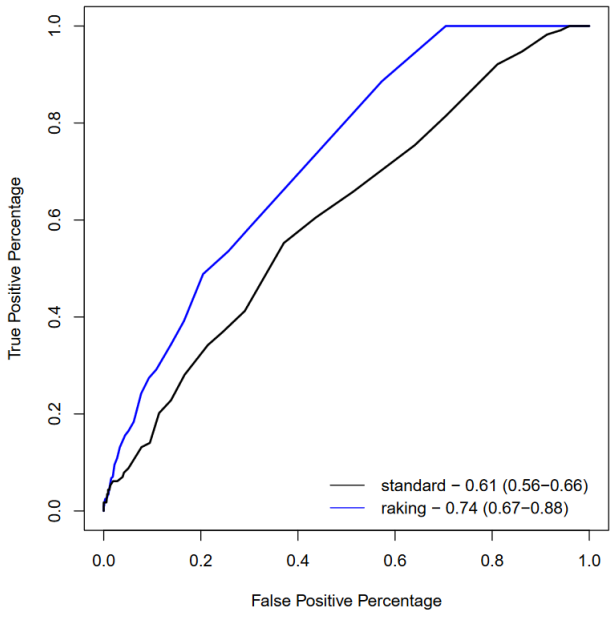


FIGURE 3.1 : MSS-F AUC before (standard) and after raking.

Testing threshold	9	12	15
Sensitivity	0.947	0.754	0.553
Specificity	0.138	0.359	0.629
PPV	0.102	0.108	0.133
NPV	0.962	0.934	0.932

TABLE 3.1 : MSS-F. Statistics at different testing thresholds without raking.

3.5 Discussion

Universal germline genetic testing for breast and ovarian cancer is yet to be widely implemented, mainly due to logistical, resource and financial constraints. Indeed, there is a lack of trained professionals to provide genetic counseling, and limited resources and funding to carry out molecular analyses at such a large scale. That would, however, not be a major problem, if there were accurate models to predict who is likely to carry a susceptibility PV. Commonly used models are MSS3, a logistical regression score, and CanRisk, a mendelian mathematical model [25, 11]. They take into account clinical and pathological criteria related to the patient's cancer, family history of cancer, and for CanRisk, also family structure. Of note, all models are likely to miss a proportion of carriers. In addition, recent studies have shown that family history was not predictive of germline status, suggesting that there is room for improvement in the genetic assessment of breast cancer cases [36].

In this study, we studied the performances of MSS-F in 1219 breast and ovarian cancer patients. MSS-F differs slightly from MSS3, as we made empirical modifications to account for multigene testing, and to fit with French national practice in place at the time. AUC was low 0.61 (0.56-0.66), and negative predictive value did not increase with decreasing testing thresholds. This means that a germline PV can never be ruled out, regardless of the score. Sensitivity was high, but only at the 9-point threshold (95%). Of note, the Manchester team recommends testing at the 12 to 15-point threshold. AUC did however increase to 0.74 (0.67-0.88) after raking, i.e. adjustment to a series of unselected cancer patients. This suggests that MSS-F would be more informative if used by gynecologists and oncologists to decide who to refer - or who to test in mainstreaming pathways -, and not downstream in already-selected patients referred to Clinical Cancer Genetics. Still, an AUC of 0.74 should be considered average. CanRisk also had average performances in a random subset of cases, with an AUC of 0.61 (0.51-0.71).

One limitation is that MSS-F is not validated as such. It is however based on the published MSS3 Modifications likely resulted in more liberal testing of patients. While this might explain the low AUC, we expected an increase in NPV with lower testing thresholds. A germline PV in a breast and ovarian cancer gene has major implications. It leads to long term personalized screening and risk reduction, for example annual breast magnetic resonance imaging or salpingo-oophorectomy. Relatives are then offered genetic testing, and those carrying the PV benefit from similar screening and risk-reduction measures. In addition PVs in BRCA1 and BRCA2 now have direct consequences regarding cancer treatment, as carriers meeting specific conditions are given olaparib, a drug that target *BRCA*-deficient cancer cells [56, 71]. In this context, one should aim at identifying all PV carriers among cancer patients in the short to medium term. Considering the limitations of current models, as illustrated in this study, universal germline testing looks like the only solution. There should therefore efforts to promote access to genetic counseling and germline genetic testing for all breast and ovarian cancer cases. The approval of olaparib in patients with germline *BRCA1/BRCA2* PV makes it even more urgent to remove barriers to testing.

Appendices to Chapter 3

The report detailed previously is the condensed result of a project with two major components which were :

- To understand why the MSS-F was performing worse than expected at predicting PV in the studied population.
- To compare its performance against classical predicting tools based on familial data.

The aim of this appendice is to provide some extra information on the methods and results.

3.A Introduction

3.A.1 Data collection

The data collected includes 1220 patients with breast and/or ovarian cancer case history referred between 2016 and 2020 to the AP-HP Sorbonne University Oncogenetics laboratory, Paris (a summary of the collected data is available on Figure 3.2). All of the patients have undergone germline panel testing for *BRCA1*, *BRCA2*, *PALB2*, *TP53*, *RAD51C*, *RAD51D*, *ATM*, *CHEK2*, *PTEN*, *CDH1*, *MMR* following French guidelines [45]. The study is prospective from january 2018 to june 2020 and retrospective before 2018. Indeed, the MSS-F score is calculated prospectively for patients included from 2018 onwards, with a recommendation to test at the 12 point threshold, and to consider testing at the 9-point threshold. The MSS-F is also calculated retrospectively for those who had genetic testing in 2016-2017, in those cases, the decision to test is based on clinical criteria. The Figure 3.3 shows the MSS-F repartition for the sub-groups (retrospective and prospective). The difference between the groups does not reach the typical statistical significance of 0.05 on the Wilcoxon rank-sum test (p-value=0.05087).

For 210 patients selected (200 totally randomly and 10 randomly amongst the remaining mutation carriers), the familial history is also digitized at the Canrisk format.

		Female	Male	Total
Centers	Tenon	318	8	326
	Pitié-Salpêtrière	614	18	632
	St-Antoine	89	4	93
	Est Francilien	61	2	63
	Sud Francilien	95	5	100
	CHG Longjumeau	5	1	6
Ages	Age < 45	320	0	320
	Age 45 – 60	486	12	498
	Age > 60	374	26	400
	Unknown age	2	0	2
Mutations	BRCA1	46	0	46
	BRCA2	38	4	42
	Others mutations	26	0	26
	No mutation	1072	34	1106

FIGURE 3.2 : Summary of collected data, by centers, ages and mutations.

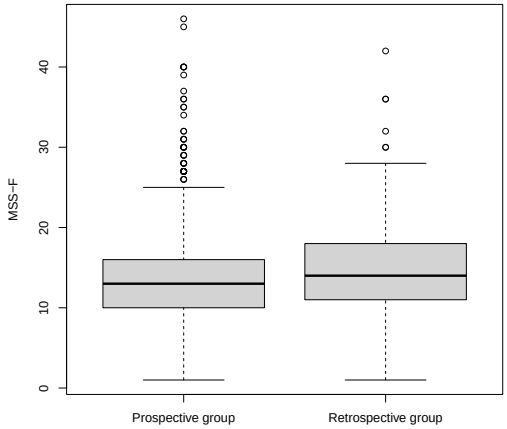


FIGURE 3.3 : Boxplot of MSS-F for the prospective and retrospective groups. Wilcoxon rank-sum test - p-value=0.05087.

3.A.2 MSS-F performance on AP-HP dataset

The performance of the MSS-F at predicting pathogenic variants on the studied population was lower in terms of AUC than the expected performance of the MSS as shown in Figure 3.4. Even though at 9, 12 and 15 points thresholds, the proportion of PV carriers was respectivily 10.1%, 10.9%, 13.5% which are concordant with the MSS litterature [25].

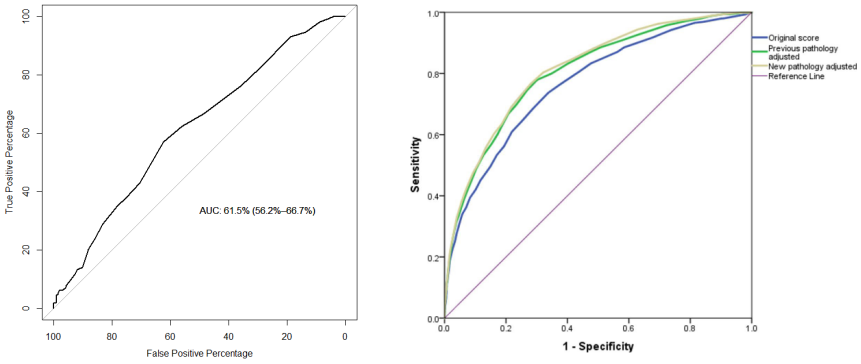


FIGURE 3.4 : ROC curve of MSS-F on collected data (AUC=0.61 [0.56-0.66]) vs ROC curve of MSS3 [25] (AUC=0.81 [0.79-0.83]).

Acknowledging this information, two questions arise :

- As patients seen in oncogenetic counselling already showcase high scores, it seems that MSS-F might not be relevant at this stage of the care pathway. Indeed most of the patients (1060/1220) reach the threshold at 9. A natural question is therefore, would the MSS-F perform better upstream in the care pathway to select patients that should be sent to geneticists ?

- Considering that geneticists still need to select patients which should be offered germline testing during counselling, how do models such as BOADICEA, BRCAPRO and Claus-Easton, which are family-based models, perform at predicting pathogenic variants compare to MSS-F ?

3.B Familial models performances comparison with MSS-F

3.B.1 Material and methods

The identification of pathogenic variants is a classification problem where the aim is to determine if a patient is carrier or non-carrier of a mutation. A standard metric used to assess the performance of a binary classifier (in this case carrier/non-carrier) is the area under the curve (AUC) of a receiver operating characteristic (ROC) curve. The ROC curve is the plot of the true positive rate (sensitivity) against the false positive rate (1-specifity) for different threshold of the classifier. Its area under the curve summarizes the overall performance of the classifier across all thresholds.

BOADICEA is a breast cancer risk prediction model which uses familial history of a patient [18]. Even though it is not its main objective, it is able to compute a probability of being carrier of a certain set of mutations including *BRCA1/2*, *PALB2*, *ATM*, *CHECK2*. BRCApro [50] is also a model used specifically to predict the probability of *BRCA1* and *BRCA2* mutations based on familial history of ovarian and breast cancer of first and second degree relatives. Like BOADICEA, Claus-Easton is a model based on familial history of breast and ovarian cancer used to predict the risk of cancer taking *BRCA1* and *BRCA2* mutations into account [28] but can also provide posterior probability of being a BRCA1/BRCA2 mutation carrier.

As 210 familial histories are digitized at the Canrisk format, results from BOADICEA can be directly compute from online API. In order to use BRCApro and Claus-Easton, files were modified thanks to a pipeline from the Canrisk format to each specific format. The

MSS-F performances at predicting mutations is then compared to BOADICEA's, BRCApro's and Claus-Easton's performances using AUC/ROC.

As most of the models are actually specific to *BRCA1* and *BRCA2* mutations, the performances are tested on the prediction of any pathogenic variant in the panel but, also specifically on the prediction of *BRCA1* and *BRCA2* mutations.

3.B.2 Results

On the subset of patients with digitized familial history, all familial models and MSS-F perform poorly at pathogenic variants as shown in Figure 3.5. BOADICEA is the most efficient with an AUC of 0.61 (0.51-0.71), followed by MSS-F 0.57 (0.47-0.67). BRCApro 0.53 (0.41-0.64) and Claus-Easton 0.54 (0.43-0.65) perform similarly.

When predicting only BRCA1 and BRCA2 (see Figure 3.6), the results remain mostly the same with very slight variations, BOADICEA showcases an AUC of 0.63 (0.62-0.74), MSS-F 0.55 (0.44-0.66), BRCApro 0.52 (0.39-0.64) and Claus-Easton 0.54 (0.42-0.67).

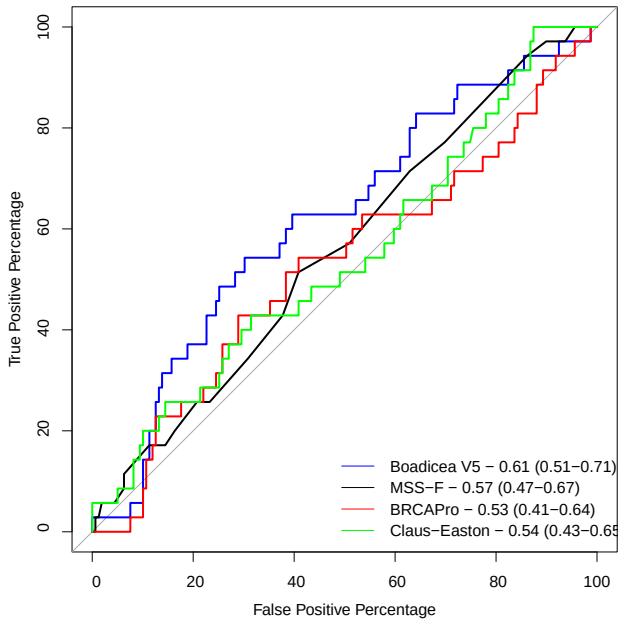


FIGURE 3.5 : ROC curves of BOADICEA, MSS-F, BRCAPro and Claus-Easton on the 210 selected patients for the prediction of any mutation of the panel.

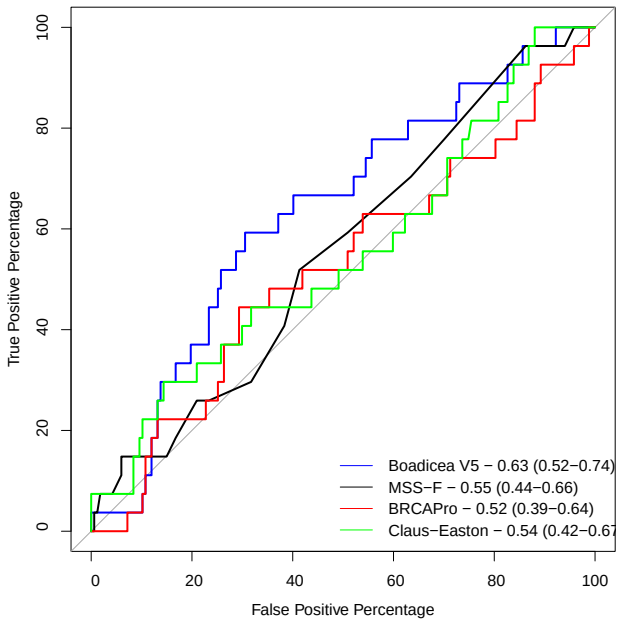


FIGURE 3.6 : ROC curves of BOADICEA, MSS-F, BRCAPro and Claus-Easton on the 210 selected patients for the prediction of BRCA1 or BRCA2 only.

3.C Raking for selection bias correction

3.C.1 Material and methods

The distribution of the MSS-F scores in the collected data is shifted toward the high risk (> 9) compared to the population used in MSS article [18]. This can be due to the selection bias as people referred to geneticist already showcase heavily affected pedigree and also because the prospective part of the study uses MSS-F ($\geqslant 9$) as selector for genetic testing.

A useful method to correct selection bias is the raking method [34]. The aim of raking is to compute weight for every individual in the observed data so the statistics (over multiple variables) of the weighted observed data match the statistics of a target population.

In this article, the target population is a dataset collected at AP-HP (Paris) which is composed of unselected women going to a gynaecological consultation for a breast or ovarian cancer diagnosis. The data includes MSS-F scores and other variables. The variables used to fit the statistics with the raking method are MSS-F categories, grade 1, grade 2, triple negative, ductal in situ. The AUC/ROC curves are re-calculated using the weighted data.

3.C.2 Results

The performance at predicting any pathogenic variant of the panel of MSS-F improves on the corrected population (after raking) with an AUC of 0.74 (0.67-0.88) as shown in Figure 3.7. The AUC of the MSS-F on raw data was 0.61 (0.56-0.66).

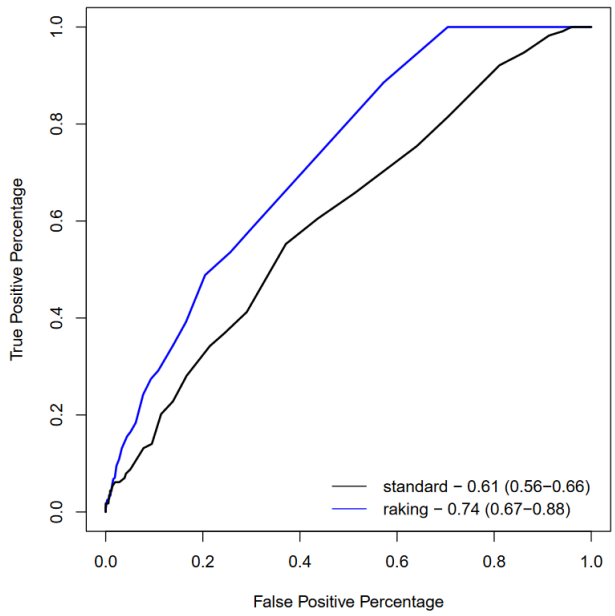


FIGURE 3.7 : MSS-F AUC before (standard) and after raking.

3.D Conclusion and perspectives

Once a patient is referred to the geneticist, it likely already showcases a high MSS-F, this could explain the modest performance of the MSS-F on raw data, it also suggests that MSS-F might not be necessary at this stage of the medical care. The improved performances of the MSS-F after raking show that, when applied to a more general population, the score performs as expected in the litterature. This suggests that the score should actually be used upstream in the medical care in unselected breast/ovarian cases as a selector to refer patients to the oncogenetics.

In the context of oncogenetics counselling where most of the patients showcases high MSS-F scores, it seems that models based on familial data do not perform especially better than MSS-F at predicting pathogenic variants.

To go further in this study, it would be necessary to compute the MSS3 in order to actually compare it with the MSS-F. For now, the results suggest that there should be a Europe-wide effort to promote access to germline genetic testing for the majority of BC/OC cases referred to oncogenetics counselling as decision-making tools to discrimate carriers and non-carriers at this stage do not perform well enough. The approval of PARP-inhibitors for BC patients with germline *BRCA1/BRCA2* pathogenic variants makes it even more urgent to remove barriers to germline testing.

Chapitre 4

Estimation de courbe de survie de la maladie du surfactant pour les porteurs de mutations *SFTPA1/SFTPA2*

Résumé

La pneumopathie interstitielle (ILD) est une affection chronique qui affecte les poumons, provoquant des lésions progressives pouvant conduire à une insuffisance respiratoire au fil du temps. Des recherches récentes ont identifié un lien entre des variants dans les gènes codant les protéines du surfactant (SP)-A1 (*SFTPA1*) et SP-A2 (*SFTPA2*) et l'ILD ainsi que le cancer du poumon. La pénétrance de ces maladies pour les porteurs de ces variants rares *SFTPA1/2* est encore inconnue, mais cruciale pour le suivi des patients et le conseil génétique. Nous avons identifié ces variants pathogènes dans 27 familles indépendantes où au moins un individu présentait une ILD et/ou un cancer du poumon. L'objectif de cette étude est d'estimer la pénétrance de l'ILD et du cancer du poumon chez les individus porteurs des variants *SFTPA1* ou *SFTPA2*, dont la pathogénicité a été confirmée par des études fonctionnelles in vitro.

Sur la base de pedigrees étendus regroupant 744 individus sur 27 familles parmi lesquels 59 porteurs, des données phénotypiques ont été recueillies auprès de 328 d'entre eux. La pénétrance pour l'ILD et le cancer du poumon a été évaluée en utilisant une méthode existante basée sur un algorithme EM dont l'étape d'espérance est réalisée par un algorithme somme-produit (dans les réseaux bayésiens formés par les arbres généalogiques) et l'étape de maximisation par un estimateur de Kaplan-Meier.

Les résultats montrent une pénétrance du premier événement (ILD ou cancer broncho-pulmonaire) de 50% à l'âge de 60 ans. La pénétrance du premier événement est forte mais incomplète, atteignant 89,3% à l'âge de 80 ans. Le premier événement est le plus souvent l'ILD, le cancer du poumon survient généralement plus tard. La pénétrance du cancer du poumon est moindre que l'ILD mais forte par rapport à la pénétrance du cancer broncho-pulmonaire en population générale avec une pénétrance de 50 % à l'âge de 84 ans et aucun cas avant 30 ans. Cela confirme la pathogénicité des variants *SFTPA1* et *SFTPA2* pour ces deux maladies pulmonaires.

Valorisation : Ce travail est une collaboration entre Lucas Ducrot (doctorant LPSM - Sorbonne Université), Nadia Nathan (MD, PhD - pneumologie pédiatrique AP-HP), Patrick Benusiglio (MD, PhD - oncogénéticien AP-HP), Raphaël Borie (MD, PhD - pneumologie AP-HP), Gregory Nuel (DR CNRS - LPSM - Sorbonne Université) et Marie Legendre

(PharmD, PhD - Génétique Moléculaire AP-HP). L'article est soumis à PLOS Genetics en avril 2024.

Contents

4.1 Abstract

Interstitial lung disease (ILD) is a chronic condition that affects the lungs, causing progressive damage that can lead to respiratory failure over time. Recent research has identified a link between heterozygous variants in the genes encoding surfactant proteins (SP)-A1 (*SFTPA1*) and SP-A2 (*SFTPA2*) and ILD and lung cancer. The penetrance of those two rare *SFTPA1/2*-associated clinical entities is still unknown while crucial for monitoring and genetic counseling. We have identified pathogenic variants in these two genes in 27 independant families in which at least one individual had ILD and/or lung cancer. The aim of this study was to estimate the penetrance of ILD and lung cancer in heterozygotes for *SFTPA1* or *SFTPA2* variants whose pathogenicity has been confirmed through in vitro functional studies.

Based on extended pedigrees gathering 744 individuals among whom 59 carriers, phenotypic data were retrieved from 328 individuals. Penetrance for ILD and for lung cancer have been assessed by using an existing method based on an EM algorithm of which the E-step is performed through sum-product algorithm (in the Bayesian networks formed by family trees) and the M-step by Kaplan-Meier estimator.

The results show a penetrance to the first event of 50% at the age of 60 years old. The penetrance to the first event is high but not complete reaching 89.3% at the age of 80. The first event is most of the time the ILD and lung cancer typically occurred later. The penetrance to lung cancer is lower than expected (as *SFTPA1* and *SFTPA2* pathogenic variants are linked to increased risks) with a penetrance of 50% at the age of 84 years old and no case before 30.

4.2 Introduction

Interstitial lung diseases (ILD) is a heterogeneous group of rare lung disorders that affect the distal parenchyma. This disease is associated with various degree of lung inflammation and lung remodeling, often leading to lung fibrosis. Monogenic causes represent around 20% of ILD etiologies, mainly including variants in telomerase- and surfactant-related genes. Among the latter, heterozygous variants of *SFTPA1* and *SFTPA2*, encoding the surfactant proteins (SP)-A1 and SP-A2, are associated with various phenotypes ranging from asymptomatic carriers to lung fibrosis and adenocarcinoma of the lung displaying a severe prognosis and leading to lung transplantation or death [46, 47, 40, 74, 41]. SP-A1 and SP-A2 are highly autologous proteins that assemble to form oligomers of SP-A. The penetrance of the disease in individuals carrying *SFTPA1* or *SFTPA2* variants, and the reasons for such variability in disease expression remain unknown. Understanding the penetrance of a disease for specific groups of patients has a significant impact on medical protocols especially for risk assessment of individuals who benefit from a pre-symptomatic diagnosis or follow-up, genetic counseling, medical monitoring and prevention.

Thus, this study aims at estimating the penetrance of ILD and lung cancer in *SFTPA1* or *SFTPA2* variant carriers.

4.3 Materials and methods

4.3.1 Patients and relatives

In the framework of the french national network for rare lung diseases *RespiFIL*, the families of patients carrying a *SFTPA1* or a *SFTPA2* missense pathogenic variant identified in Trousseau hospital clinical laboratory were included. Pathogenicity of the variants has been confirmed by in vitro functional studies [40]. Pedigrees were analyzed and the following data were collected : age at last follow-up, age at diagnosis of ILD and/or lung cancer, and genotype when available. The study was approved by the relevant ethics committee (*Comité de protection des personnes*) and written informed consent was obtained from all participants or their legal representatives. Clinical information was collected in a legally authorized database (CNIL No. 681248).

Data were retrieved from the standardized form sent by the clinician in charge of the patient. DNA was extracted from whole blood. *SFTPA1* and *SFTPA2* variants were diagnosed by Next Generation Sequencing (NGS) capture targeted panel (SeqCap EZ Choice, Roche diagnostics) or Sanger sequencing (Big Dye V3.1 sequencing kit and 3730XL sequencing machine, Thermo Fisher Scientific). Given the high homology between the *SFTPA1* and *SFTPA2* genes, following a double inhouse-pipeline analysis, NGS data was further analyzed in the IGV viewer by setting the VAF threshold to 5%. For Sanger sequencing, PCR primers were designed to avoid variations with an allelic frequency higher than 1% in the v2.1.1 gnomAD total population. In addition, PCR primers were designed with their most 3' base on a sequence difference between *SFTPA1* and *SFTPA2* to allow specificity.

4.3.2 Survival to an event

Survival to an event refers to the probability that the specified event has not happened up to a certain time (time to event). Especially, it can refer to the probability that a patient has not been diagnosed with a specific disease at a certain age which is called the survival function S(t). The function of interest that is usually considered instead is the penetrance function F(t) which is linked to the survival as $F(t) = 1 - S(t)$.

4.3.3 Model Description

The function of interest in this article was the penetrance $F(t)$, but for estimation and modeling purposes as well as the usage of specific packages, the estimated function is the survival $S(t)$ which is directly linked to the penetrance as $F(t) = 1 - S(t)$. The model, which is a direct implementation of the article of Alarcon [3] describes a group of individuals with potentially family links and the probabilities of their genotypes, ages or ages at diagnosis and status (affected by the disease or unaffected). This model can be conditioned on the genotypes and therefore can be decomposed in two subparts, a genetic one and a survival one.

For the genetic part, the joint distribution of the genotypes is given by a Bayesian network thanks to the family structure as the genotype of one individual only depends on the genotypes of its parents. These structures are informative as in pedigree data, most of the genotypes are unknown.

For the survival part, the age and status are independant conditionally to the genotypes, meaning that carriers and non-carriers do not share the same survival. The aim of the article is to estimate the survival of the mutation carriers.

The model is based the following assumptions :

1. One single locus of predisposition following autosomal dominant inheritance with two alleles (one "normal", one "pathogenic") was considered. *SFTPA1* and *SFTPA2* variants were considered in this article as a single locus since : i) they are only 55Mb apart on chromosome 10, ii) they lead to the same range of phenotypes, and iii) they never appear simultaneously in the carriers' families. This is due either to the low frequencies of variants in those genes or to the fact that the occurrence of pathogenic variants in both genes may be non-viable.

2. Genotypes of the families' founders follow Hardy-Weinberg equilibrium with an allelic frequency of the deleterious allele $f = 0.0005$ meaning 1 case over 2000 which is the upper limit for rare genetic disease.

3. Genotypes of descendants follow the Mendelian inheritance principle. The analysis is based on pedigree data, at least one individual of the included families carries a *SFTPA1* or *SFTPA2* variant and these variants run from parents to children with no *de novo* occurrence of the variants.

4. In the included families, only the carriers of the deleterious allele can be affected by the disease (no sporadic cases). ILD and lung cancer are both rare diseases. Therefore, the possibility of sporadic cases in the carrier's family is neglected.

5. The ascertainment bias is treated accordingly to the Proband's phenotype Exclusion Likelihood (PEL). [1, 4, 58], meaning that the probands are considered unaffected at age 0 in order to be uninformative toward the disease. However, their genotypes (carrier/non-carrier) were preserved and used in the analysis.

6. There is a possibility of false positives and false negatives throughout genetic testing. A genetic test is assumed to have a probability α (0.0001) of false positives and a probability β (0.02) of false negatives.

4.3.4 Model Fitting

In order to take into account the unknown genotypes in the data, the adopted framework is the same EM framework described by Alarcon [3]]. The objective is to estimate both the a

posteriori distribution of being carrier for each individual and the survival function of variant carriers. The EM algorithm is a well-known and used method to compute the maximum likelihood of a model in presence of incomplete data (in this case the genotypes). To do so, the EM algorithm starts with a random initialization of the parameters (i.e. survival) and then alternates two steps :

- Expectation-step : during this step, using the last computed survival function (M-step), the probabilities of being carriers are updated through belief propagation (sum-product algorithm) [70] using *Bped*, a C++ implementation of the algorithm (available on demand to Grégory Nuel). It is similar to Elston-Stewart algorithm [22, 23] with an additional backward propagation in order to compute the marginal distribution.

- Maximization-step : during this step, the new survival function is updated using a weighted Kaplan-Meier estimator [69, 68]. The weights used are the probabilities of being carrier computed for each individual during the E-step.

These two steps are iterated until convergence or up to 300 iterations.

4.3.5 Statistics

The 95% confidence intervals of the survival functions are computed using the R package Survival [69, 68]. The method is used both on an unstratified population and male/female stratified population in order to see if the survival is sex-dependent. The significance of the difference between male and female survival is estimated with a log-rank test. Two methods are used to perform the male/female stratification. The first method is a simple male/female stratification where each group has its own independent survival function. The second method uses Cox proportional hazard model to assess the association between the survival and the sex variable. Both methods are implemented with the R package survival [69, 68]. In order to quantify the role of each parameter (f frequency of the allele, α probability of False Positive in genetic testing and β probability of False Negative in genetic testing), a sensitivity analysis is performed. While one parameter is analysed, the others are set at their based values (i.e. $f = 0.0005$, $\alpha = 0.0001$, $\beta = 0.02$). Each parameter is tested over a particular set of values (i.e. $f \in \{0.05, 0.005, 0.0005, 0.00005\}$, $\alpha \in \{0.01, 0.001, 0.0001, 0.00001\}$, $\beta \in \{0.02, 0.002, 0.0002\}$).

4.4 Results

4.4.1 Genotypes and phenotypes

A total of 27 families have been included in this study. A *SFTPA1* pathogenic variant was identified in 10 families and a *SFTPA2* pathogenic variant was identified in 17 families. The pedigrees are provided in Supplemental Figure **??**. Among the families, the data of 27 index patients and 717 relatives were analyzed, accounting for a total of 744 included individuals (Table 4.7). A total of 22 and 37 individuals carried a *SFTPA1* or *SFTPA2* pathogenic variant respectively. An ILD was diagnosed in 64, a lung cancer in 23 and both in 20. Individuals were declared as asymptomatic in 221 cases and the clinical status was unknown in 416 cases. At the study time, 119 individuals were deceased, including 4 from ILD or lung cancer. The median age of the living individuals at the study time was 43 years. The median age at the disease onset was 49 years.

	SFTPA1 (n)	SFTPA2 (n)	Total (n)
Families	10	17	27
Individuals :	279	465	744
Males	143	241	384
Females	136	224	360
Dead	46	73	119
Median age at study time	49	41	43
Median age at death	56.5	52.5	54.5
Genotype:			
Heterozygotes for a pathogenic variant	22	37	59
Non-carriers	14	27	41
Unknown	243	401	644
Phenotype:			
Asymptomatic	80	141	221
ILD only	24	40	64
Lung cancer only	5	18	23
Both	6	14	20
Unknown	164	252	416
Median age at ILD and/or lung cancer diagnosis	45	49.5	49

TABLE 4.1 : Main characteristics, phenotype and genotype of the patients and relatives. Abbreviations : ILD, interstitial lung disease.

4.4.2 Survivals of ILD, lung cancer and to first event for *SFTPA1* or *SFTPA2* variants carriers

The method is applied to compute the survivals for ILD and lung cancer alone and survival to the first event. The survival functions are presented in Figure 4.5 and in Table 4.2. The survival to first event at 30 year-old was 0.93. ILD appeared before lung cancer in 15% (3 over 20 diagnosed both with ILD and lung cancer) of cases. The youngest age at lung cancer diagnosis was 30 years. *SFTPA1* and *SFTPA2* pathogenic variant carriers present a high risk of developing either ILD or lung cancer as the penetrance to first event at 80 year-old is 89.4% [74.1-95.7].

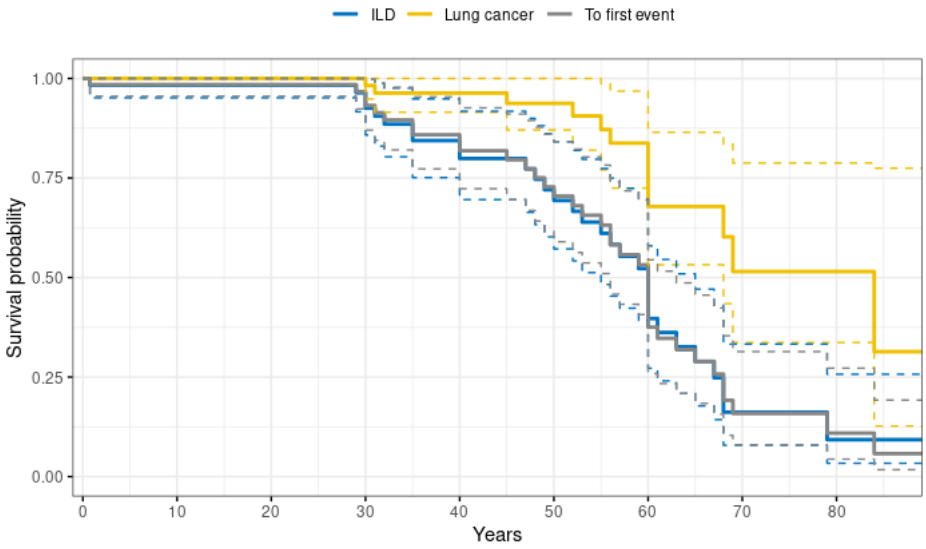


FIGURE 4.1 : Survival functions (solid lines) and 95% confidence intervals (dotted lines) are provided for interstitial lung disease (ILD) (blue lines), lung cancer (yellow lines) and first event (grey lines).

Age	ILD	Lung cancer	First event
10	0.983 [0.951 – 1.000]	1.000 [1.000 – 1.000]	0.983 [0.956 – 1.000]
20	0.983 [0.951 – 1.000]	1.000 [1.000 – 1.000]	0.985 [0.956 – 1.000]
30	0.924 [0.856 – 0.998]	0.982 [0.949 – 1.000]	0.933 [0.872 – 0.998]
40	0.795 [0.690 – 0.916]	0.964 [0.917 – 1.000]	0.820 [0.726 – 0.927]
50	0.688 [0.565 – 0.837]	0.939 [0.874 – 1.000]	0.707 [0.593 – 0.843]
60	0.379 [0.254 – 0.566]	0.686 [0.542 – 0.868]	0.374 [0.258 – 0.543]
70	0.126 [0.054 – 0.291]	0.530 [0.355 – 0.791]	0.153 [0.078 – 0.300]
80	0.071 [0.024 – 0.203]	0.530 [0.355 – 0.791]	0.106 [0.043 – 0.259]
90	0.071 [0.024 – 0.203]	0.330 [0.138 – 0.788]	0.056 [0.017 – 0.183]

TABLE 4.2 : Survival to ILD, lung cancer and to the first event.

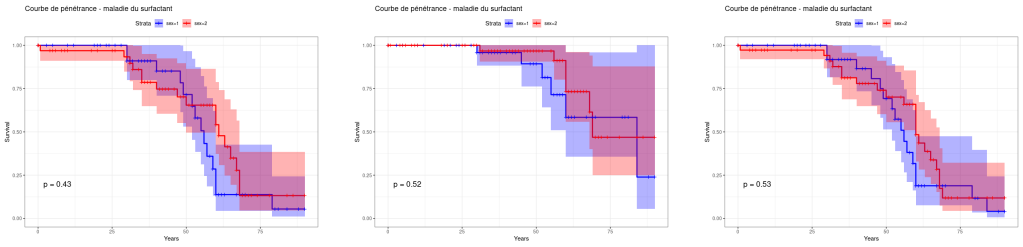
4.4.3 Sensitivity analysis

The sensitivity analysis showed very low variation dependencies to the different parameters f, α and β for all the computed survivals (Supplementary material and Supplementary Tables 4.4, 4.5 and 4.6 providing survivals to ILD or lung cancer alone and survivals to the first event at 30, 50 and 70 year-old). Considering f, the low dependency probably comes from the fact there is at least one carrier in each family. Therefore, the probability of being carrier relies more on being a relative of a variant carrier than the frequency of the allele in the general population. Considering α and β, the variations are low because there are many variant carriers showcase a disease which, in the model, consolidates the fact they are variant carriers.

4.4.4 Male/Female stratification

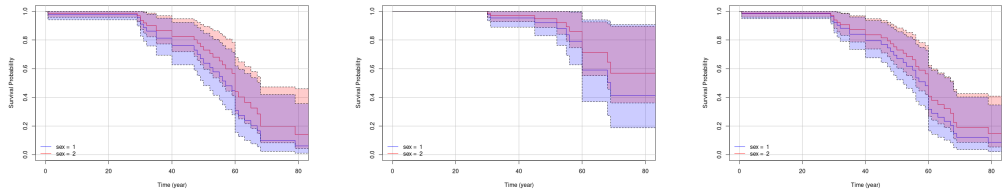
The method was also applied with a stratification male/female, as the previous results, to compute the survivals for both ILD and lung cancer alone and survival to the first event.

There is a trend for a better survival to the first event in male compared to female before 50 years old, the trend interchanged after 50 (Figures 4.2 and 4.3). The differences, however, do not reaching significance (p-values reported in Table 4.3). More details are presented in Supplementary materials.



ILD Lung cancer To first event

FIGURE 4.2 : Standard Male/Female stratification for ILD, Lung cancer and to the first event survival estimations for *SFTPA1* and *SFTPA2* mutation carriers



ILD Lung cancer To first event

FIGURE 4.3 : Male/Female Cox proportional hazard method for ILD, Lung cancer and to the first event survival estimations for *SFTPA1* and *SFTPA2* mutation carriers

Method	**ILD**	**Lung cancer**	**To First Event**
Standard	0.43	0.52	0.53
Cox	0.34	0.43	0.47

TABLE 4.3 : P-values for both standard stratification and Cox proportional hazard model and each disease

4.5 Discussion

The present study provides the first estimation of penetrance of ILD, lung cancer and of first of these two events for *SFTPA1* and *SFTPA2* variant carriers. These data were obtained using a large cohort of patients carrying *SFTPA1* and *SFTPA2* variants and their relatives. Based on the mathematical model, the computed risk showed that the risk of ILD or lung cancer in *SFTPA1* or *SFTPA2* carriers increases with age (mean age 60) to reach an important – but not complete – penetrance of 89.3 at 80 year-old. Interestingly, despite these variants being previously shown to be associated with a high risk of lung cancer, we prove herein that this event occurs mainly later than ILD, with a penetrance of 50% at 80 year-old, and with no case before 30 years.

Penetrance of dominant diseases is known to be heterogeneous. However, great variations are described depending on the involved gene and the functional consequences of the variants. *SFTPA1* and *SFTPA2* variants are associated with an impaired expression and secretion of the corresponding SP-A1 and SP-A2 proteins. However, the reason(s) why the age at onset

of the ILD symptoms vary from a few months to more than 80 years is currently unknown. Viral infections - especially in children - and environmental or occupational exposures in adults may have a role in triggering the disease. Unfortunately, these factors could not be retrieved for a relevant number of patients and relatives for study.

As expected in an autosomal disease, the study did not show any significant difference for penetrance between men and women.

Surfactant disorders are rare causes of ILD in adults. When a *SFTPA1* or *SFTPA2* variant is identified, the result is reported to the patient in the framework of a genetic counseling. In France, the patient has to inform his/her relatives about the genetic results in order to give them the opportunity to ask for a pre-symptomatic diagnosis. This analysis is offered to relatives over the age of majority who are at risk of carrying the variant.

Since 2018, the national network for rare lung disease (*RespiFIL*, www.respifil.fr) has launched multidisciplinary team meetings for genetic forms of ILD in adults and in children. Despite more and more cases being diagnosed, no guidelines are currently available for the pre-symptomatic management of surfactant diseases. The present study provides crucial information and could help to discuss the following management strategies : (i) the first CT-scan may not be useful before 30 year-old ; (ii) if the CT-scan is normal, the timeline between two CT-scans is 5 years (iii) *SFTPA1* or *SFTPA2* variant carriers should receive a clear information on environmental factors that could increase risk of lung fibrosis and cancer such as tobacco smoking or occupational exposures.

To validate the model and observe the natural history of the disease, a prospective study including patients with a surfactant-related disease and their adult relatives is currently in progress (*RaDiCo-ILD2*, ClinicalTrials.gov ID NCT06036719). Continuing data collection, including tobacco and occupational exposures is a perspective work to strengthen the results of the study.

The study displays some limits, especially due to the missing data in far relatives, but also because the model is based on the assumption that the disease (lung fibrosis and lung cancer) does not present sporadic cases. This assumption is factually not true but was chosen as the probability of sporadic cases in the observed families is very low and may be neglected. An improvement of the model could be to develop a mathematical model taking sporadic cases into account to compute the penetrance of such disease.

4.6 Conclusion

This study estimated the penetrance of interstitial lung disease (ILD) and lung cancer in individuals carrying *SFTPA1* or *SFTPA2* pathogenic variant. The investigation involved 27 independent families with at least one member being *SFTPA1* or *SFTPA2* pathogenic variant carrier.

The penetrance is estimated using an existing method based on an EM algorithm of which the E-step is performed through sum-product algorithm (in the Bayesian networks formed by family trees) and the M-step by Kaplan-Meier estimator.

The results show a penetrance to the first event of 50% at the age of 60 years old. The penetrance to the first event is high but not complete reaching 89.3% [74.0-95.6] at the age of 80. The first event is most of the time the ILD and lung cancer typically occurred later. The penetrance to lung cancer is lower than penetrance to ILD (as *SFTPA1* and *SFTPA2* pathogenic variants are linked to increased risks) with a penetrance of 50% at the age of 84 years old and no case before 30.

Following the guideline, ILD diagnosed after 50 are currently not considered as genetically related. This study shows that for *SFTPA1* and *SFTPA2* pathogenic variant carriers, the median age at ILD diagnosis is 60 [55-65] which means that genetic testing could be

appropriate for later forms of ILD.

While acknowledging certain limitations, such as missing data and assumptions about sporadic cases, the study sets the stage for further research. Ongoing prospective studies, like RaDiCo-ILD2, aim to validate the model and enhance understanding of the natural history of these diseases. Continued data collection, including environmental factors, is crucial for refining and strengthening the outcomes of this study.

Legal and ethical statement

Written informed consents were obtained from the patients.

Acknowledgments

We thank the French national networks for rare lung diseases : *Centre de référence des maladies respiratoires rares (RespiRare)*, *Centre de référence des maladies pulmonaires rares (OrphaLung)* and *Filière de soins pour les maladies respiratoires rares (RespiFIL)*. The ILD cohort has been developed in collaboration with the Rare Disease Cohort (RaDiCo)-ILD project (ANR-10-COHO-0003), the Clinical research collaboration for chILD-EU and the COST Innovative Grant OpenILD CIG16125.

Authors contribution

LD wrote the manuscript. NN, GN and ML reviewed the manuscript. NN and ML collected and computed the data. LD and GN performed the mathematical analyses.

Appendices to Chapter 4

4.A Supplementary Material

4.A.1 Sensitivity Analysis

Method

The hyperparameters of the model are α (False Positives probability of genetic testing) set at 0.0001, β (False Negatives probability of genetic testing) set at 0.02 and f (the frequency of the deleterious allele in the general population) set at 0.0005.

This sensitivity analysis investigates the variations of the results for multiple values of α, β and f.

In order to quantify the role of each parameter while one parameter is analysed, the other are set at their based values (i.e. $f = 0.0005$, $\alpha = 0.0001$, $\beta = 0.02$). Each parameter is tested over particular values (i.e. $f \in \{0.05, 0.005, 0.0005, 0.00005\}$, $\alpha \in \{0.01, 0.001, 0.0001, 0.00001\}$, $\beta \in \{0.02, 0.002, 0.0002\}$).

Results

The results are presented in the Tables 1, 2 and 3 which provide survivals to both ILD and lung cancer alone and survivals to the first event at 30, 50 and 70 years old.

The sensitivity analysis shows very low variation dependencies to the different parameters f, α and β for all the computed survivals. Some possible reasons can explain such low dependencies of the results to these parameters.

Considering f, the low dependency probably may come from the fact there is at least one carrier in each family. Therefore the probability of being carrier relies more on being a relative of a mutation carrier than the frequency of the allele in the general population.

Considering α and β, these parameters are here to take into account potential genetic testing errors, such as a possible non-carrier affected by the disease which is not possible in the model (consdering disease with no sporadic cases). The variations may be low because there is no such cases in the data.

4.A.2 Male/Female Stratification

Method

The method is used both on an unstratified population and male/female stratified population in order to see if survivals to both ILD and lung cancer alone and survival to the first

ILD			
α	**30**	**50**	**70**
1e-05	0.926 [0.859,0.998]	0.696 [0.576,0.842]	0.165 [0.081,0.336]
1e-04	0.925 [0.858,0.998]	0.693 [0.571,0.840]	0.161 [0.078,0.332]
1e-03	0.923 [0.853,0.998]	0.684 [0.560,0.836]	0.152 [0.071,0.322]
1e-02	0.920 [0.849,0.998]	0.677 [0.552,0.832]	0.146 [0.067,0.315]
Lung cancer			
α	**30**	**50**	**70**
1e-05	0.982 [0.948,1.000]	0.938 [0.871,1.000]	0.518 [0.340,0.790]
1e-04	0.982 [0.948,1.000]	0.937 [0.870,1.000]	0.515 [0.336,0.787]
1e-03	0.981 [0.946,1.000]	0.934 [0.864,1.000]	0.498 [0.320,0.775]
1e-02	0.979 [0.941,1.000]	0.927 [0.850,1.000]	0.466 [0.291,0.746]
To First Event			
α	**30**	**50**	**70**
1e-05	0.932 [0.871,0.998]	0.707 [0.593,0.842]	0.160 [0.081,0.316]
1e-04	0.932 [0.870,0.998]	0.704 [0.589,0.841]	0.158 [0.080,0.313]
1e-03	0.930 [0.867,0.998]	0.697 [0.580,0.837]	0.152 [0.075,0.306]
1e-02	0.928 [0.863,0.998]	0.690 [0.572,0.833]	0.144 [0.070,0.294]

TABLE 4.4 : Survivals at 30, 50 and 70 years old for different values of α with $\beta = 0.02$ and $f = 0.0005$ fixed.

ILD			
β	**30**	**50**	**70**
2e-04	0.924 [0.855,0.998]	0.688 [0.566,0.838]	0.154 [0.0739,0.322]
2e-03	0.924 [0.856,0.998]	0.689 [0.566,0.838]	0.155 [0.0746,0.323]
2e-02	0.925 [0.858,0.998]	0.693 [0.571,0.840]	0.161 [0.0786,0.332]
Lung cancer			
β	**30**	**50**	**70**
2e-04	0.981 [0.947,1.000]	0.936 [0.868,1.000]	0.507 [0.330,0.780]
2e-03	0.981 [0.947,1.000]	0.936 [0.868,1.000]	0.509 [0.331,0.781]
2e-02	0.982 [0.948,1.000]	0.937 [0.870,1.000]	0.515 [0.336,0.787]
To First Event			
β	**30**	**50**	**70**
2e-04	0.931 [0.868,0.998]	0.700 [0.585,0.839]	0.153 [0.076,0.306]
2e-03	0.931 [0.868,0.998]	0.701 [0.585,0.839]	0.154 [0.077,0.307]
2e-02	0.932 [0.870,0.998]	0.704 [0.589,0.841]	0.158 [0.080,0.313]

TABLE 4.5 : Survivals at 30, 50 and 70 years old for different values of β with $\alpha = 0.0001$ and $f = 0.0005$ fixed.

event are sex-dependant as it is standard to test in clinical statistics. For the male/female stratification, the significance of the difference between male and female survivals is estimated with a log-rank test.

To stratified the population into male and female, two methods are used :

- A standard stratification where male and female survivals are estimated separatively during the M-step of the EM aglorithm.
- A Cox proportional hazard model where the male and female survivals are estimated

ILD			
f	**30**	**50**	**70**
5e-05	0.924 [0.856,0.998]	0.690 [0.567,0.839]	0.163 [0.078,0.341]
5e-04	0.925 [0.858,0.998]	0.693 [0.571,0.840]	0.161 [0.078,0.332]
5e-03	0.924 [0.855,0.998]	0.686 [0.563,0.836]	0.136 [0.063,0.296]
5e-02	0.924 [0.856,0.998]	0.688 [0.565,0.837]	0.126 [0.054,0.291]
Lung cancer			
f	**30**	**50**	**70**
5e-05	0.981 [0.946,1.000]	0.934 [0.865,1.000]	0.502 [0.324,0.780]
5e-04	0.982 [0.948,1.000]	0.937 [0.870,1.000]	0.515 [0.336,0.787]
5e-03	0.982 [0.948,1.000]	0.937 [0.870,1.000]	0.515 [0.338,0.785]
5e-02	0.982 [0.949,1.000]	0.939 [0.874,1.000]	0.530 [0.355,0.791]
To First Event			
f	**30**	**50**	**70**
5e-05	0.931 [0.868,0.998]	0.701 [0.585,0.839]	0.158 [0.079,0.316]
5e-04	0.932 [0.870,0.998]	0.704 [0.589,0.841]	0.158 [0.080,0.313]
5e-03	0.932 [0.870,0.998]	0.704 [0.589,0.841]	0.155 [0.078,0.308]
5e-02	0.933 [0.872,0.998]	0.707 [0.593,0.843]	0.153 [0.078,0.300]

TABLE 4.6 : Survivals at 30, 50 and 70 years old for different values of f with $\alpha = 0.0001$ and $\beta = 0.02$ fixed.

jointly during the M-step and share an exponential coefficient.

Results

With the two stratification methods, the results show differences between male and female survivals but non-significant (P-value reported in Table 1).

With the two stratification methods, the results show differences between male and female survival functions but the log-rank test is not significant (P-value reported on the graphs). It is still possible that the sex plays a role in the survival to the ILD or lung cancer for *SFTPA1* or *SFTPA2* mutation carriers but the available data are not currently sufficient to assess that properly.

Method	**ILD**	**Lung cancer**	**To First Event**
Standard	0.43	0.52	0.53
Cox	0.34	0.43	0.47

TABLE 4.7 : P-values for both standard stratification and Cox proportional hazard model and each disease

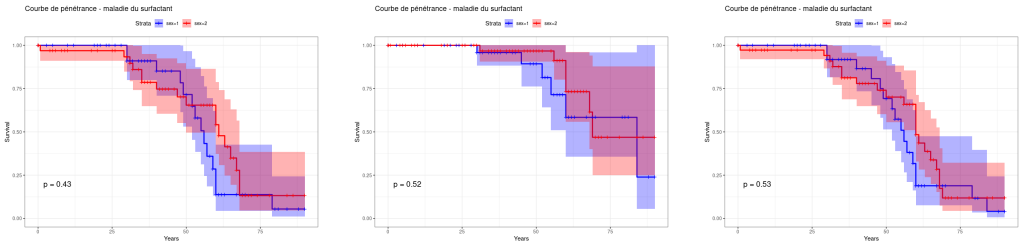
4.A.3 Model

Model description

In the context of genetic diseases with age dependencies, the function of interest is generally the penetrance :

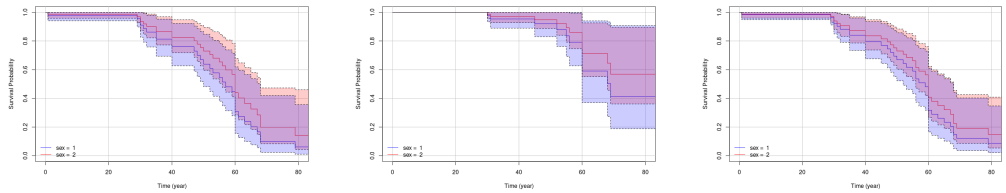
$$F(t) = \mathbb{P}(\text{disease diagnosed before age t})$$

,



ILD Lung cancer To first event

FIGURE 4.4 : Standard Male/Female stratification for ILD, Lung cancer and to the first event survival estimations for *SFTPA1* and *SFTPA2* mutation carriers



ILD Lung cancer To first event

FIGURE 4.5 : Male/Female Cox proportional hazard method for ILD, Lung cancer and to the first event survival estimations for *SFTPA1* and *SFTPA2* mutation carriers

but in this article, the estimated function is the survival function :

$$S(t) = \mathbb{P}(\text{disease not diagnosed before age t})$$

.

Though it is easy to retrieve the penetrance as it is directly linked to the survival :

$$S(t) = 1 - F(t)$$

.

The model, which is a direct implementation of the method of Alarcon [3], describes a group of individuals with potentially family links and the probabilities of their genotypes, ages or ages at diagnosis and status (affected by the disease or unaffected). In mathematical terms, let consider n individuals in set $\mathcal{I} = \{1, ..., n\}$, the set of founders which are individuals that have no parents in the data is noted $\mathcal{F} \subset \mathcal{I}$. The ages or ages at onset of the disease of all individuals is denoted $T = (T_1, ..., T_n) \in \mathbb{R}^n$ where T_i is the age for individuals i. The genotypes of individuals is denoted $X = (X_1, ..., X_n) \in \{00, 01, 10, 11\}^n$ where 0 represents wild-type allele and 1 the deleterious allele and first digit (respectively second) corresponds to the paternal (respectively maternal) allele (i.e. for example $X_i = 01$ means the individual i has a paternal allele 0 and a maternal allele 1). Also $\delta = (\delta_1, ..., \delta_n) \in \{0, 1\}^n$ denotes the status of individuals, δ_i is 1 if the individual i is affected and 0 if unaffected. Finally $G = (G_1, ..., G_n) \in \{0, 1\}^n$ represent the genetic test where G_i is the result of the genetic testing for individual i (0 for non-carriers and 1 for carriers). (T, δ) and G are considered independant conditionnally to X meaning that genetic testing process is independant from the age and status of individual. Therefore this model can be conditioned on the genotype X and decomposed in two subparts, a genetic one, a survival one and a genetic testing one as follows :

$$\mathbb{P}(T,\delta,X,G) = \mathbb{P}(X) \times \ \mathbb{P}(T,\delta,G|X) = \underbrace{\mathbb{P}(X)}_{\text{Genetic Part}} \times \underbrace{\mathbb{P}(T,\delta|X)}_{\text{Survival Part}} \times \underbrace{\mathbb{P}(G|X)}_{\text{Testing Part}} \tag{4.A.1}$$

- **Genetic Part :** the probability of the genotypes forms a Bayesian network thanks to the family structure as the genotype of one individual only depends on the genotypes of its parents. The set founders of the family $\mathcal{F}$ is the set of individuals that have not parents in the data, they follow Hardy-Weinberg equilibrium with allelic frequency f, the non-founders follow Mendelian transmission from parents :

$$\mathbb{P}(X) = \prod_{i\in\mathcal{F}} \mathbb{P}(X_i) \prod_{i\notin\mathcal{F}} \mathbb{P}(X_i|X_{\text{pat}_i}, X_{\text{mat}_i}) \tag{4.A.2}$$

- **Survival Part :** $S(t)$ and $\lambda(t)$ represent survival and hazard rate for mutation carriers

$$\mathbb{P}(T_i = t, \delta_i = 0|X_i) = \begin{cases} S(t) & \text{if } X_i \neq 00 \\ 1 & \text{if } X_i = 00 \end{cases} \tag{4.A.3}$$

$$\mathbb{P}(T_i = t, \delta_i = 1|X_i) = \begin{cases} S(t)\lambda(t) & \text{if } X_i \neq 00 \\ 0 & \text{if } X_i = 00 \end{cases} \tag{4.A.4}$$

- **Testing Part :** α represents the probability of False Positive (being genotyped as carrier while being non-carrier) and β represents the probability of False Positive (being genotyped as non-carrier while being carrier).
  - True negative : $\mathbb{P}(G_i = 0|X_i = 00) = 1 - \alpha$
  - False negative : $\mathbb{P}(G_i = 0|X_i \neq 00) = \beta$
  - False positive : $\mathbb{P}(G_i = 1|X_i = 00) = \alpha$
  - True positive : $\mathbb{P}(G_i = 1|X_i \neq 00) = 1 - \beta$

From this model description, the proposed model is based on some assumptions :

1. One single locus of predisposition following autosomal dominant inheritance with two alleles (one "wild-type", one "deleterious").
2. Genotypes of the families' founders follow Hardy-Weinberg equilibrium with an allelic frequency of the deleterious allele f.
3. Genotypes of descendants follow the Mendelian inheritance principle.
4. Only the carriers of the deleterious allele can be affected by the disease (no sporadic cases).
5. The ascertainment bias is treated accordingly to PEL standard (Proband's phenotype Exclusion Likelihood) [1].
6. Possibility of False Positives and False Negative throughout genetic testing. The hypothesis allows to better take into account genetic testing. A genetic test is assumed to have a probability α of False Positives and a probability β of False Negatives.

Model Fitting with EM algorithm

The model fitting is performed using the EM framework on the pedigree data as described in the article of Alarcon [3]. The objective is to estimate both the a posteriori distribution of being carrier for each individual and the survival function of mutation carriers. The EM algorithm is a well-known and used method to compute the maximum likelihood of a model in presence of incomplete data (in this case the genotypes). To do so, an auxiliary Q function need to be introduced. Here is a brief description of EM algorithm :

Idea : EM Algorithm is an iterative algorithm used to find parameters of the maximum log-likelihood of probabilistic models with latent variables

Model : T, X random variables following a distribution of parameter θ, X is unobserved

Maximum Likelihood Estimator : $\hat{\theta} = \arg\max_{\theta} \sum_{X} \mathbb{P}(T, X|\theta)$

Auxiliary function :

$$Q(\theta|\theta_{\text{old}}) = \int \mathbb{P}(X|T;\theta_{\text{old}}) \log \mathbb{P}(T, X|\theta) dX$$

Algorithm :

- **Expectation-step :** compute Expectation of $Q(\theta|\theta_{\text{old}})$
- **Maximization-step :** maximization of Q to find $M(\theta) = \arg\max_{\theta'} Q(\theta'|\theta)$

Application to the model

Applied to the model described in this article, the auxiliary function Q can be written as follows :

$$Q(\theta|\theta_{\text{old}}) = \text{cst.} + \sum_{i} \mathbb{P}(X_i \neq 00|\text{ev};\theta_{\text{old}}) \log \mathbb{P}(T_i, \delta_i|X_i \neq 00;\theta) \tag{4.A.5}$$

Starting from arbitrary θ, then the two steps of the EM algorithm are done as follows :

- Expectation-step : during this step, using the last computed survival function $\theta_{\text{old}} = \theta$ (M-step), the probabilities of being carriers $w_i = \mathbb{P}(X_i \neq 00|\text{ev}, \theta_{old})$ are updated through belief propagation.
- M-step : during this step, the new survival function θ is computed using a weighted Kaplan-Meier estimator [69, 68] which maximizes the Q function. The weights used are the probabilities of being carrier w_i computed for each individual during the E-step.

These two steps are iterated until convergence or up to 300 irations.

E-step

The E-step is performed using Bped, an implementation of belief propagation (sum-product algorithm) [70] in C++ (available on demand to Grégory Nuel), as used in Alarcon's article [3]. It is similar to Elston-Stewart algorithm [22, 23] with an additional backward propagation in order to compute the marginal distribution. Bped requires an evidence file to compute the a posteriori law of genotypes. The initial evidence can be written as follows :

- For individuals that are unaffected ($\delta = 0$) :

$$\mathbb{P}(T_i = t, \delta_i = 0|X_i) \propto \begin{cases} S(t) & \text{if } X_i \neq 00 \\ 1 & \text{if } X_i = 00 \end{cases} \tag{4.A.6}$$

- For individuals that are affected ($\delta = 1$) :

$$\mathbb{P}(T_i = t, \delta_i = 1|X_i) \propto \begin{cases} 1 & \text{if } X_i \neq 00 \\ 0 & \text{if } X_i = 00 \end{cases} \tag{4.A.7}$$

While taking into account the possibility of genotyping errors, the evidence can be modified as such (only for genotyped individuals which represent a fraction of the total population in the data) :

- For individuals that are genotyped as non-carriers ($G = 0$) :

$$\mathbb{P}(T_i = t, \delta_i, G_i = 0|X_i) = \begin{cases} \mathbb{P}(T_i = t, \delta_i|X_i \neq 00) \times \beta & \text{if } X_i \neq 00 \\ \mathbb{P}(T_i = t, \delta_i|X_i = 00) \times (1 - \alpha) & \text{if } X_i = 00 \end{cases} \tag{4.A.8}$$

- For individuals that are genotyped as carriers ($G = 1$) :

$$\mathbb{P}(T_i = t, \delta_i, G_i = 1|X_i) = \begin{cases} \mathbb{P}(T_i = t, \delta_i|X_i \neq 00) \times (1 - \beta) & \text{if } X_i \neq 00 \\ \mathbb{P}(T_i = t, \delta_i|X_i = 00) \times \alpha & \text{if } X_i = 00 \end{cases} \tag{4.A.9}$$

M-step

The M-step is performed with weighted Kaplan-Meier survival estimator [69, 68] which maximizes exactly the defined Q auxiliary function using as weights the $w_i = \mathbb{P}(X_i \neq 00|ev, \theta_{old})$.

$$Q(\theta|\theta_{\text{old}}) = \text{cst.} + \sum_i \underbrace{\mathbb{P}(X_i \neq 00|\text{ev}; \theta_{\text{old}})}_{\text{weights } w_i} \log \underbrace{\mathbb{P}(T_i, \delta_i|X_i \neq 00; \theta)}_{\text{survival}} \tag{4.A.10}$$

4.B Pedigrees

FIGURE 4.6 : Pedigree graph legend

FIGURE 4.7 : Family 1 (SFTPA1 pathogenic variant)

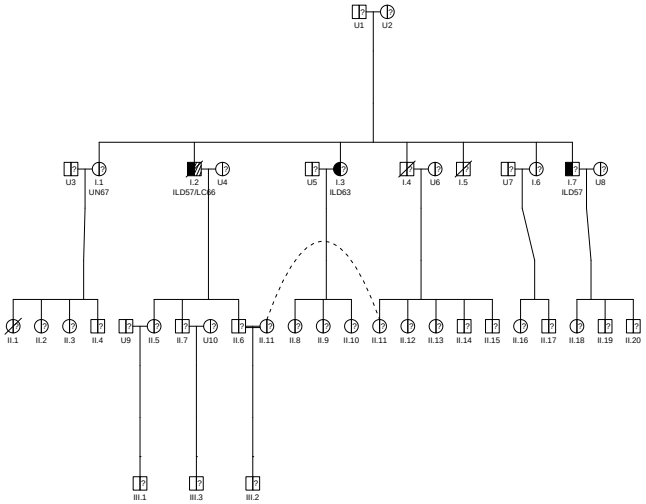


FIGURE 4.8 : Family 2 (SFTPA1 pathogenic variant)

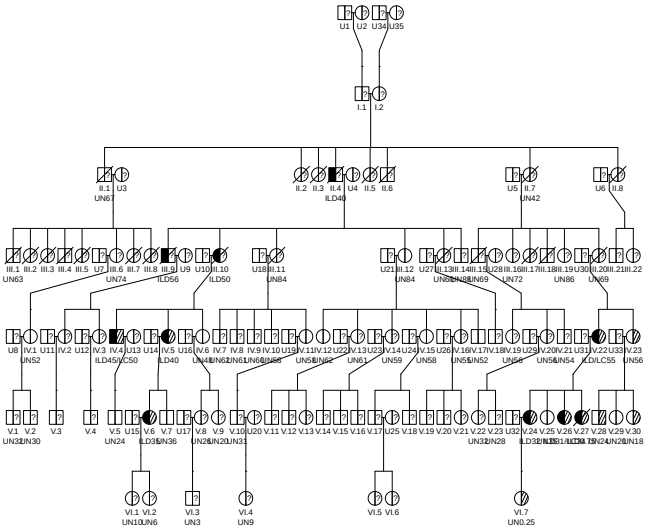


FIGURE 4.9 : Family 3 (SFTPA1 pathogenic variant)

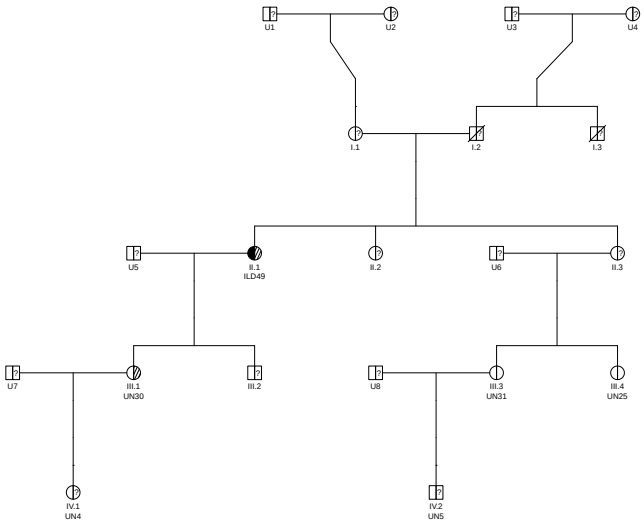


FIGURE 4.10 : Family 4 (SFTPA2 pathogenic variant)

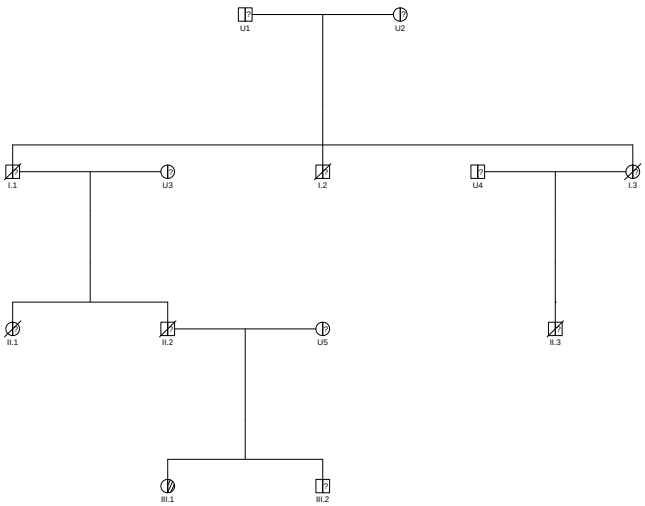


FIGURE 4.11 : Family 5 (SFTPA2 pathogenic variant)

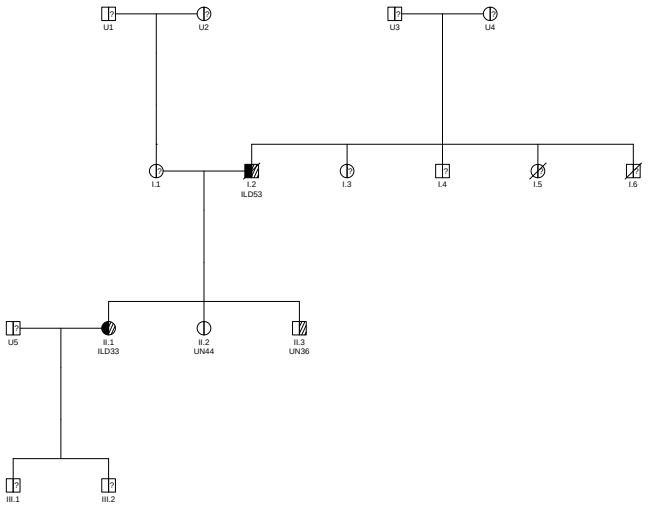


FIGURE 4.12 : Family 6 (SFTPA2 pathogenic variant)

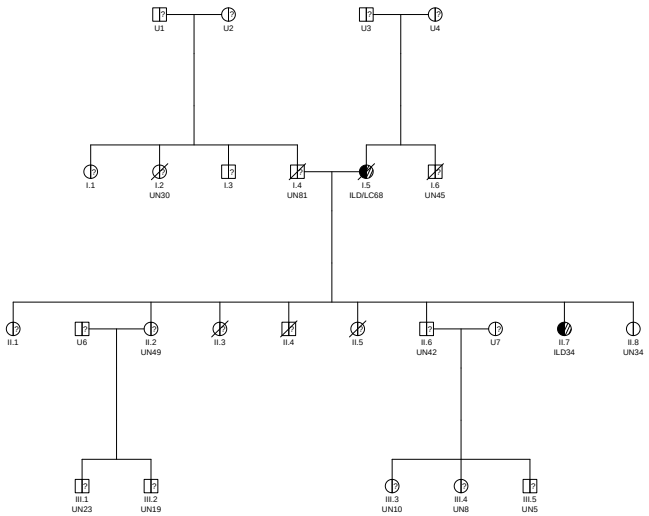


FIGURE 4.13 : Family 7 (SFTPA2 pathogenic variant)

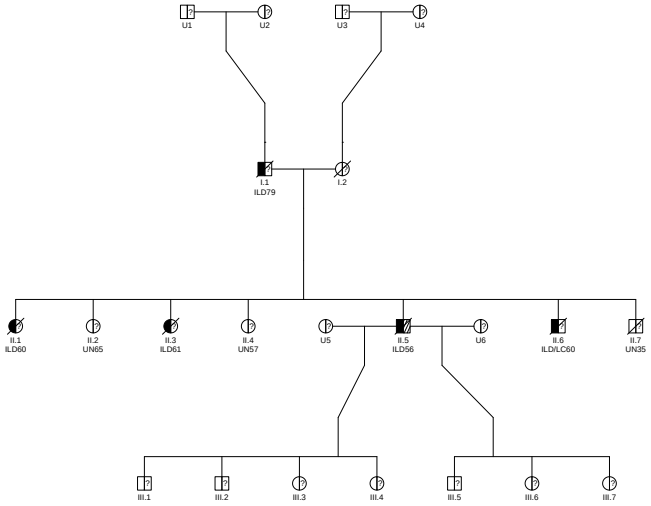


FIGURE 4.14 : Family 8 (SFTPA2 pathogenic variant)

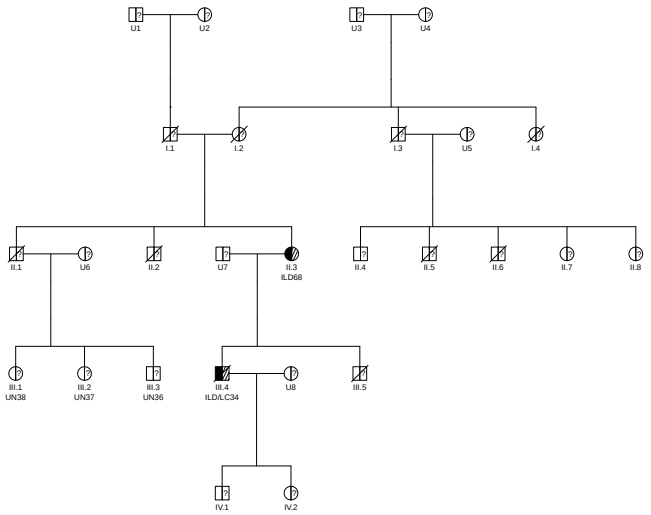


FIGURE 4.15 : Family 9 (SFTPA2 pathogenic variant)

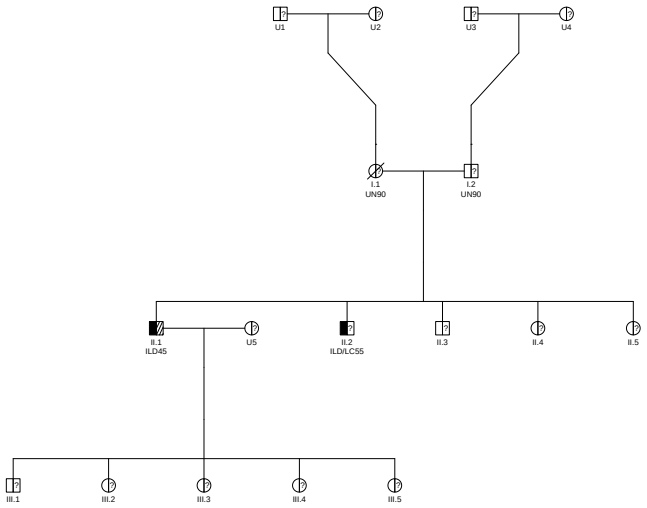


FIGURE 4.16 : Family 10 (SFTPA2 pathogenic variant)

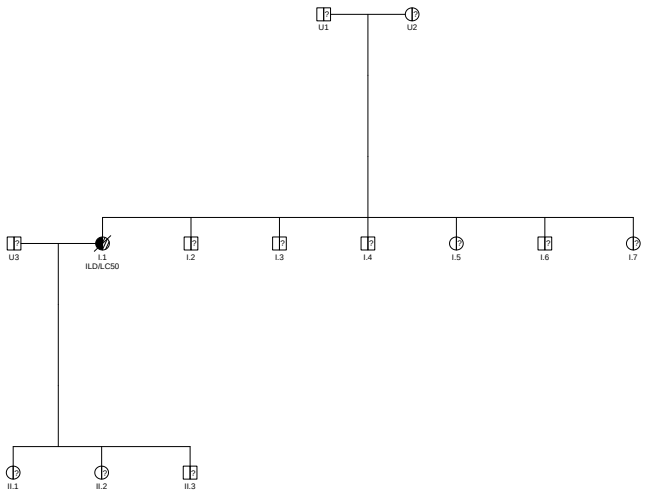


FIGURE 4.17 : Family 11 (SFTPA2 pathogenic variant)

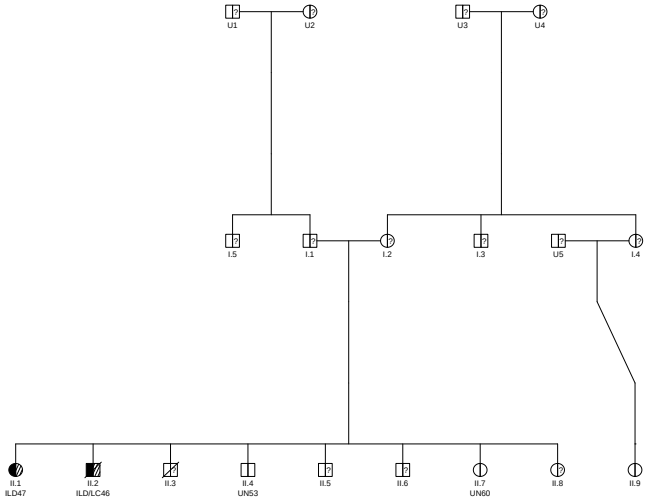


FIGURE 4.18 : Family 12 (SFTPA2 pathogenic variant)

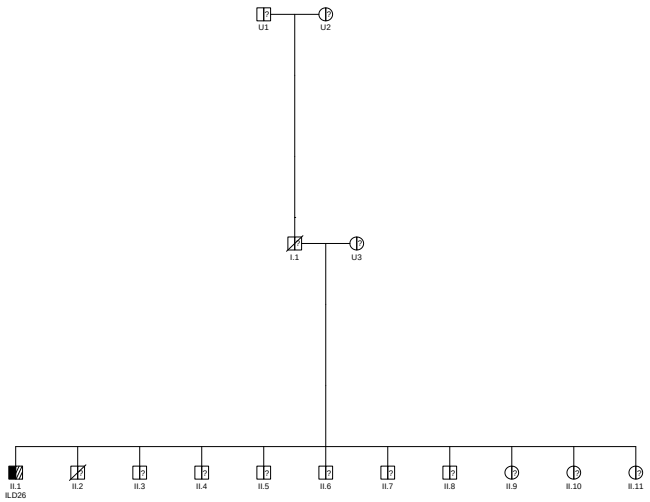


FIGURE 4.19 : Family 13 (SFTPA2 pathogenic variant)

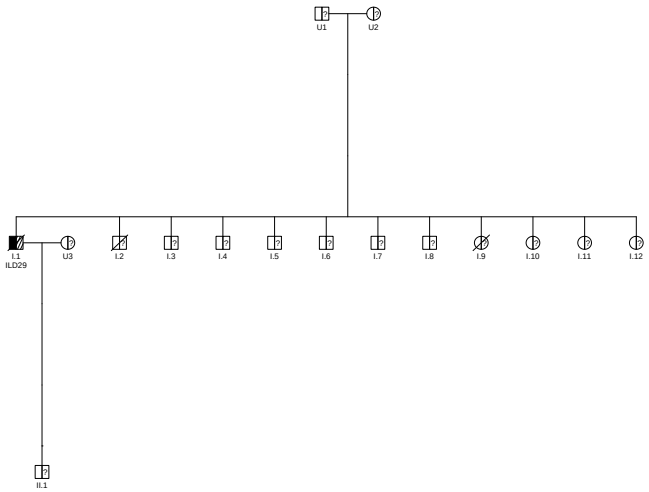


FIGURE 4.20 : Family 14 (SFTPA2 pathogenic variant)

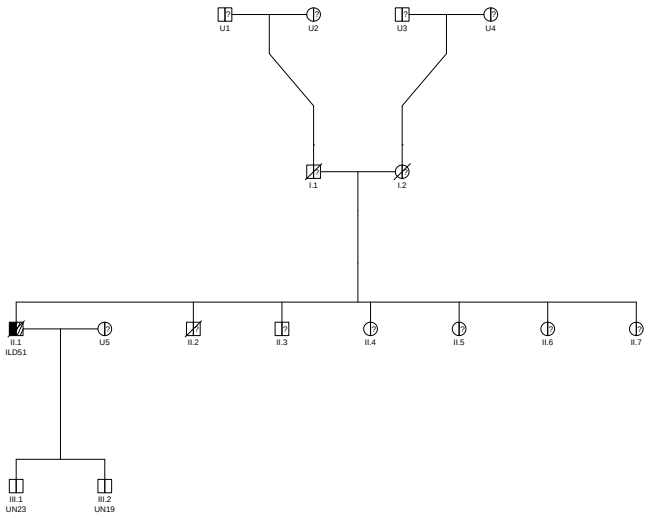


FIGURE 4.21 : Family 15 (SFTPA2 pathogenic variant)

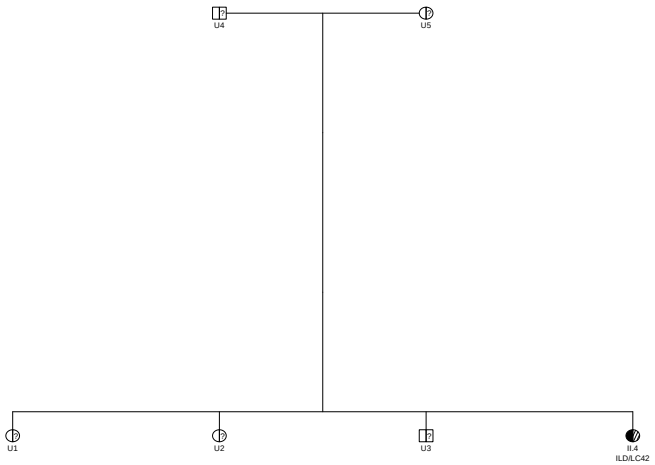


FIGURE 4.22 : Family 16 (SFTPA2 pathogenic variant)

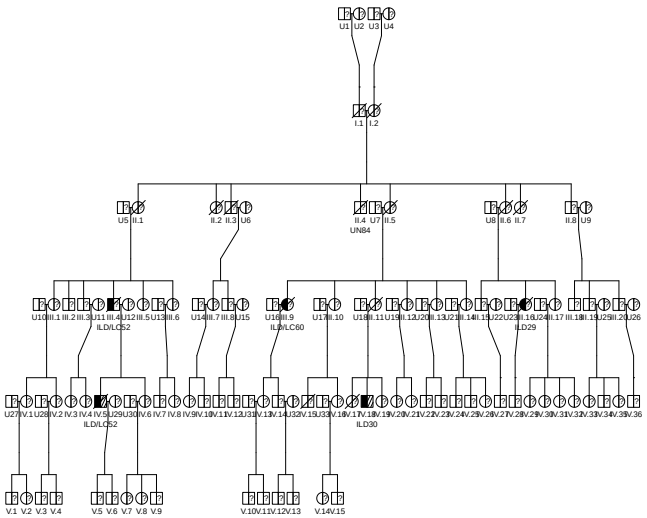


FIGURE 4.23 : Family 17 (SFTPA2 pathogenic variant)

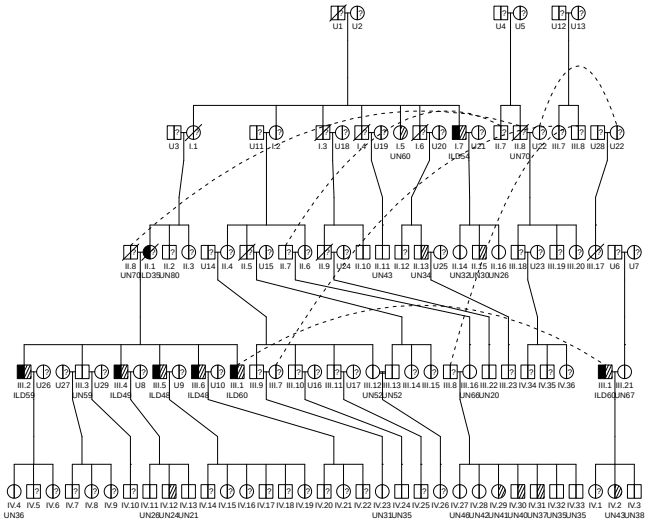


FIGURE 4.24 : Family 18 (SFTPA2 pathogenic variant)

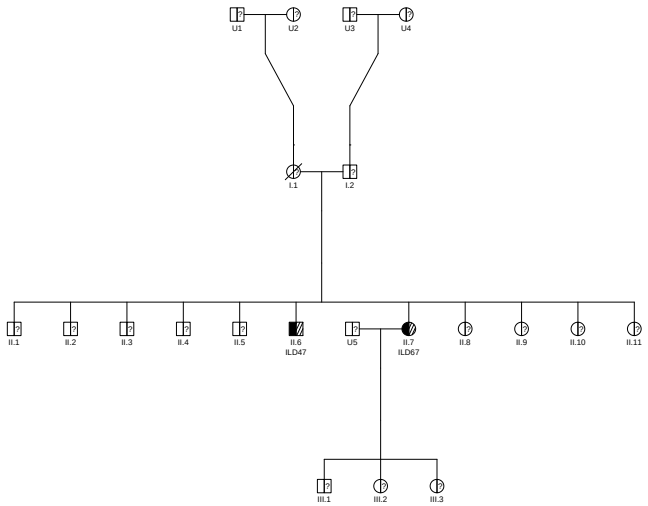


FIGURE 4.25 : Family 19 (SFTPA1 pathogenic variant)

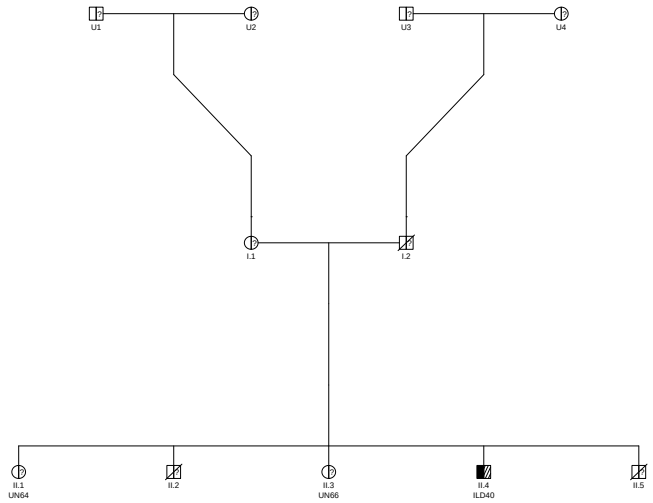


FIGURE 4.26 : Family 20 (SFTPA1 pathogenic variant)

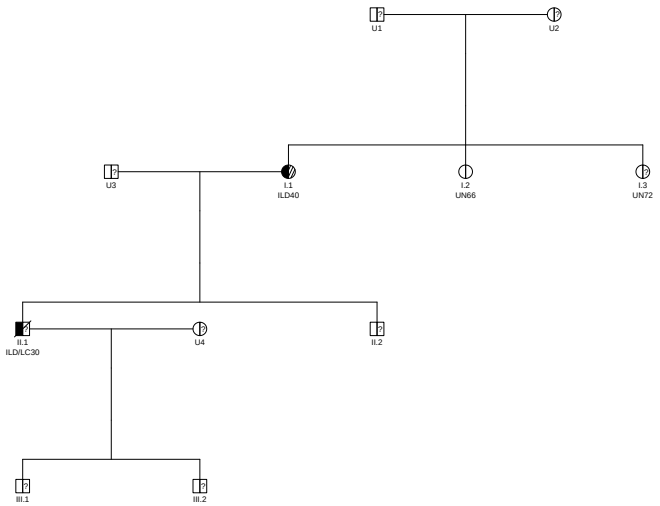


FIGURE 4.27 : Family 21 (SFTPA1 pathogenic variant)

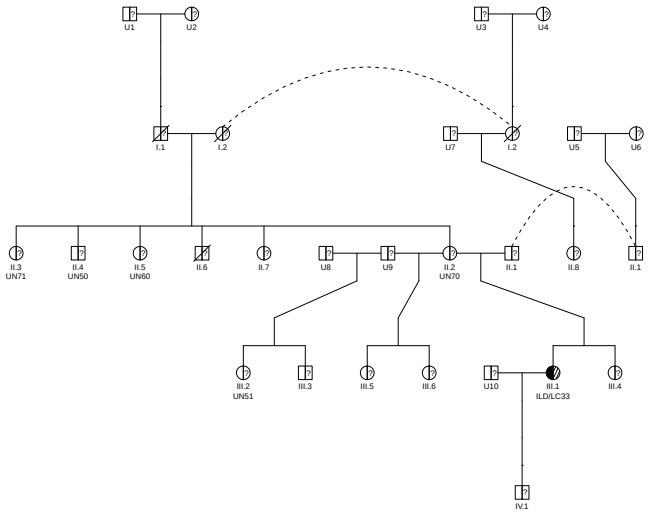


FIGURE 4.28 : Family 22 (SFTPA2 pathogenic variant)

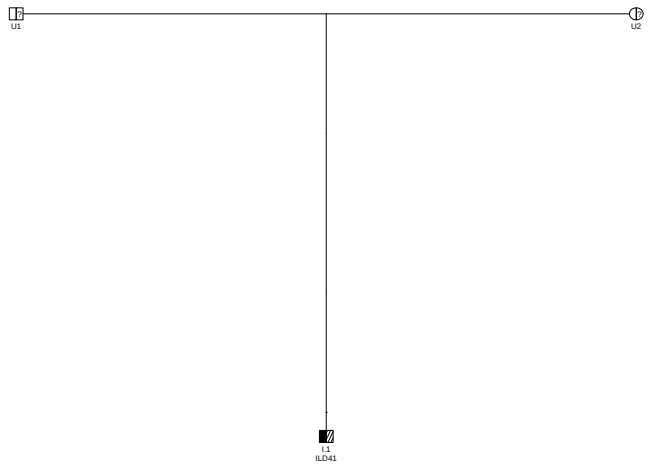


FIGURE 4.29 : Family 23 (SFTPA1 pathogenic variant)

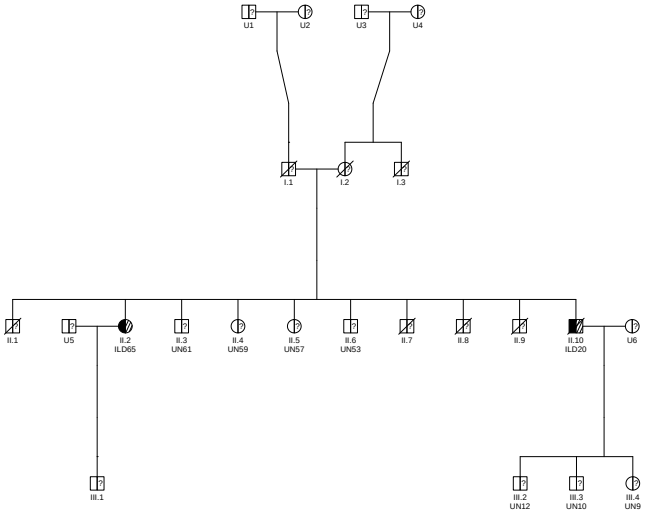


FIGURE 4.30 : Family 24 (SFTPA1 pathogenic variant)

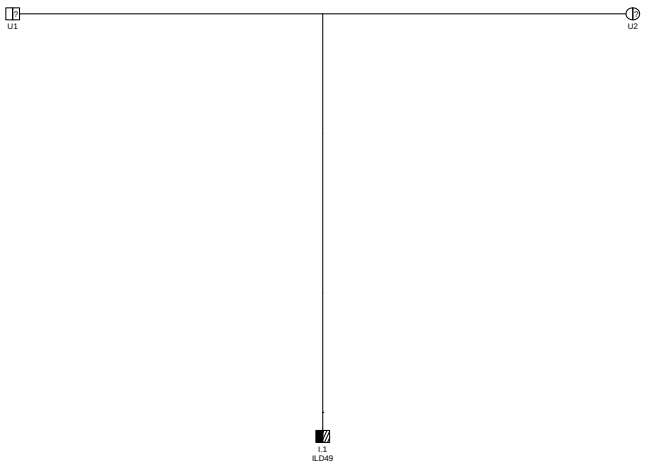


FIGURE 4.31 : Family 25 (SFTPA2 pathogenic variant)

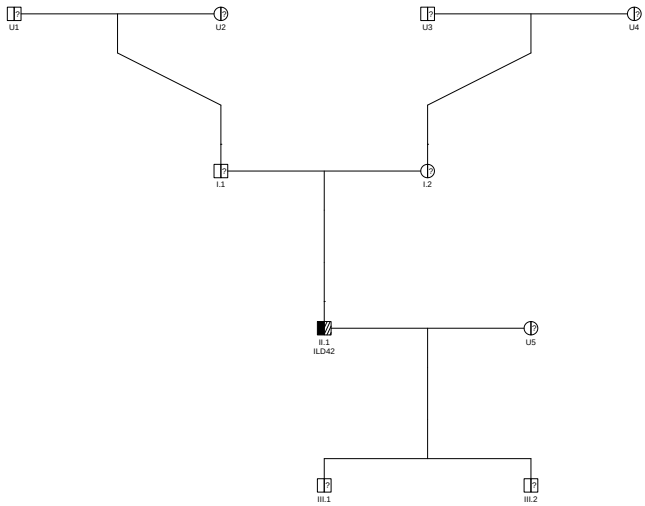


FIGURE 4.32 : Family 26 (SFTPA1 pathogenic variant)

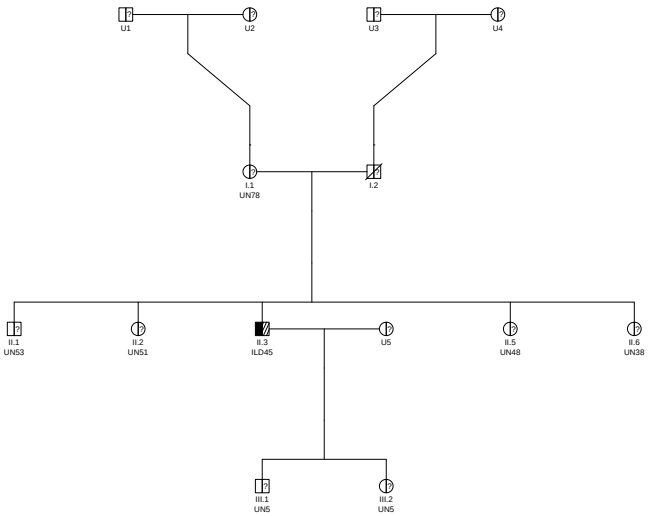
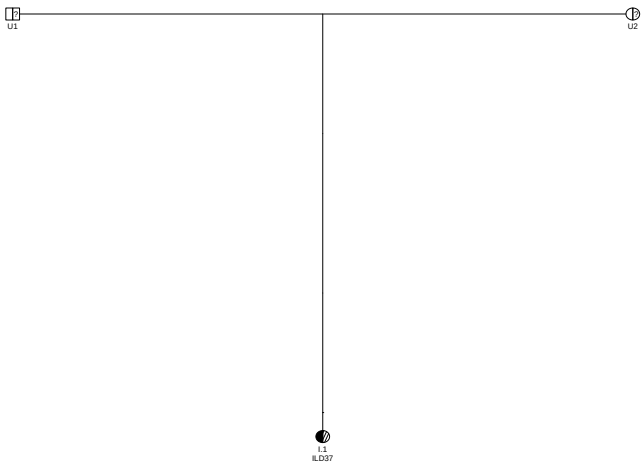


FIGURE 4.33 : Family 27 (SFTPA1 pathogenic variant)



Chapitre 5

Estimation de courbe de survie d'une maladie génétique présentant des cas sporadiques à partir de données de pédigrée

Résumé

Dans le cadre de maladies génétiques avec faible fréquence allélique en population générale et forte pénétrance (maladies mendéliennes par exemple), les approches familiales sont souvent pertinentes. En effet, les patients ont généralement une famille sévèrement touchée par la maladie et sont donc adressés au généticien. Dans ce contexte , l'estimation du risque de survenue des maladies dépendantes de l'âge (grâce aux courbes de survie/pénétrance) est requise pour la mise en place de protocoles médicaux et le suivi des patients.

Le problème principal pour effectuer ces estimations réside dans le fait que les génotypes sont souvent non-observés et doivent être traités comme une variable latente. Dans le cadre spécifique où la maladie ne présente pas de cas sporadique (*i.e.* seuls les porteurs de variants pathogènes peuvent être affectés par la maladie), le problème est plus simple à traiter car un malade est, de fait, un porteur d'un variant. L'incertitude sur les génotypes repose donc sur les personnes non-affectées. Dans ce scenario, une méthode utilisant des algorithmes d'espérance-maximisation et somme-produit a déjà été publiée.

Cependant, la plupart des maladies affectent à la fois les porteurs et non-porteurs de variants pathogènes à des taux différents. Les méthodes existantes dans ce cas supposent généralement que l'incidence de la maladie est connue pour la population générale ainsi que la proportion de porteurs de mutation. Elles approximent également l'incidence pour les non-porteurs par l'incidence pour la population générale. Cela se rapproche de la réalité dans les cas où la mutation présente une très faible fréquence allélique et une pénétrance très élevée, mais cette hypothèse s'effondre dans des scénarios plus modérés.

La méthode proposée dans ce chapitre vise à généraliser les méthodes précédentes d'estimation de survie des maladies génétiques. Elle repose sur deux hypothèses : l'incidence de la maladie pour la population générale est constante par morceaux et connue, le risque instantané relatif entre les porteurs et les non-porteurs est également constant par morceaux.

Le modèle est un mélange de survie paramétré par le risque relatif et la proportion de porteurs. À paramètres fixés, les risques instantanés (incidences) des porteurs et des non-porteurs peuvent être calculés sous contrainte du risque instantané en population générale grâce à une méthode de point fixe. Avec les données de pédigrée, la vraisemblance du modèle peut être calculée avec un algorithme somme-produit et ainsi, la vraisemblance étant une

fonction d'un nombre de paramètres finis, les paramètres du maximum de vraisemblance sont estimés à l'aide d'un algorithme d'optimisation BFGS.

La méthode a été testée sur 2000 jeux de données simulés, chacun comprenant 744 personnes (reparties sur 28 familles). Les simulations standard suivent le modèle avec une proportion de porteurs de 0,0975, un risque relatif (RH1=20, RH2=10) avec une coupure à l'âge de 50 ans. Une analyse de robustesse est également effectuée où les jeux de données sont générés avec un risque relatif suivant une fonction de Weibull.

La méthode estime sans biais les différents paramètres du modèle sur les exemples proposés. Les intervalles de confiance sont calculé par la méthode de la Hessienne (mais un version bootstrap est également proposé). Les performances de la méthode sont améliorées lorsqu'elle est appliquée sur des jeux de données de taille supérieure et/ou ayant moins de données manquantes.

Valorisation : Ce travail est effectué par Lucas Ducrot sous la direction de Grégory Nuel (DR CNRS - LPSM - Sorbonne Université). Il a été présenté sous forme de poster à EMGM 2023 (*European Mathematical Genetics Meeting*). Il a pour but d'être soumis à la revue *Mathematical Medicine and Biology* prochainement.

Contents

5.1 Abstract

In the context of genetic disease with low allele frequency in the general population and high penetrance (*i.e.* Mendelian disease), family-based approach is convenient as patients are often refered to geneticists due to their strongly affected pedigree. In this context, the estimation of survival in age-dependent genetic disease has direct applications in the medical protocol of patient care.

The main issue in these estimations is that genotypes are mostly unknown and must be treated as a latent variable. In the specific case where the disease does not present sporadic cases, the problem is easier as an affected individual is therefore a mutation carrier, the genotype incertainty leans on the unaffected population. In that simple case, methods already exist based on Expectation-Maximisation and sum-produt algorithm.

However, most diseases affect both people with and without known deleterious mutations at different rates. The few existing methods in this case generally assume that the incidence of the disease is known in the general population as well as the proportion of mutation carriers. They also assume that the incidence for non-carriers is equal to the incidence for the general population. This is close to reality for mutations with very low allele frequency and very penetrance but falls down in more moderate scenarios.

The proposed method aims to generalize previous estimation methods of genetic disease survival. It relies on two hypothesis : the hazard rate of general population is piecewise constant and known, the hazard ratio between carriers and non-carriers is also piecewise constant.

The model is a survival mixture parameterized by the hazard ratio and the proportion of carriers. At fixed parameters, the hazard rates (incidences) of carriers and non-carriers can be computed under the constrained hazard rate of general population through a fixed point method. With the pedigree data, the likelihood of the model can be computed with a sum-product algorithm and, therefore, the maximum likelihood parameters are estimated using a BFGS optimization algorithm.

The method is tested on 2000 simulated datasets of 744 people (28 families). Standard simulations followed the model with a proportion of carriers at 0.0975, hazard ratio is (RH1=20, RH2=10) with a cut-off at age 50. A robustness analysis is also performed where the dataset are generated with Weibull function as hazard ratio.

5.2 Introduction

In genetic counselling, risk estimations of genetic disease's onset is generally useful in order to guide medical protocols of patient care. In the context of genetic disease with low allele frequency in the general population and high penetrance (*i.e.* Mendelian disease), family-based approaches are generally used to evaluate that risk as patients are often selected through their strongly affected pedigree.

The main issue in these estimations is that genotypes are mostly unknown and must be treated as a latent variable. In the specific case where the disease does not present sporadic cases, the problem is easier as an affected individual is therefore a mutation carrier, the genotype incertainty leans on the unaffected population. In that simple case, methods already exist [2] based on Expectation-Maximisation [17] and Sum-product algorithm [70].

However, most diseases affect both people with and without known deleterious mutations at different rates. Typical example is breast cancer, as everyone is at risk but especially carriers of mutations (BRCA1/BRCA2 and others) which are affected at a much higher rate [21, 64]. The few existing methods [10, 58] in this case generally assume that the incidence of the disease is known in the general population as well as the proportion of mutation

carriers. They also assume that the incidence for non-carriers is equal to the incidence for the general population. This is close to reality for mutations with very low allele frequency and very penetrance but falls down in more moderate scenarios.

The proposed method aims to extend the previous estimation methods of genetic disease survival by relaxing some assumptions.

5.3 Objectives and notations

5.3.1 Notations

In this article, we consider the following context :

- an autosomal dominant disorder of one gene and two alleles ("normal" 0 and "pathogenic" 1), the genotype component $X \in \{00, 01, 10, 11\}$ ($X = 00$ for non-carrier, $X \neq 00$ for carrier) where the first is the paternal allele and the second the maternal allele ;
- the proportions of carriers in the population is denoted π_1 and non-carriers π_0 (with $\pi_0 = 1 - \pi_1$) ;
- the specific conditional hazard rates are $\lambda_1(t)$ for carriers and $\lambda_0(t)$ for non-carrier ;
- We denote the relative hazard between carriers and non-carriers $\mathrm{RH}(t)$ such as $\lambda_1(t) = \mathrm{RH}(t) \times \lambda_0(t)$;
- $S(t)$ (resp. $S_0(t)$ and $S_1(t)$) is the survival function (resp. conditional survival functions) associated with hazard $\lambda(t)$ (resp. $\lambda_0(t)$ and $\lambda_1(t)$) such as

$$S(t) = \exp\left(-\int_0^t \lambda(u)du\right); \quad S_0(t) = \exp\left(-\int_0^t \lambda_0(u)du\right); \quad S_1(t) = \exp\left(-\int_0^t \lambda_1(u)du\right).$$

- We consider as well a censorship event (which will not be needed) with a distribution function $g(t)$ and a repartition function $G(t)$ such as

$$G(t) = \int_0^t g(u)du\,.$$

5.3.2 Objective and Assumptions

The objective of this article is to estimate $S_0(t)$, $S_1(t)$ and π_1 from pedigree data with a constrained general population incidence $\lambda(t)$. In order to do so, we make two assumptions :

- the general population incidence $\lambda(t)$ is known and piecewise constant, which is often the case in medical registry (typically for cancer registry with 5-years bins) ;
- the hazard ratio between carriers and non-carriers $\mathrm{RH}(t)$ is unknown but piecewise constant (the piece-wise constant is not necessary, the main idea is to parameterize the hazard ratio, further extension of the method could study Weibull distribution as parameterization for example).

5.4 Model

The model describes a group of individuals with potentially family links and the probabilities of their genotypes, ages or ages at diagnosis and status (affected by the disease or unaffected). In mathematical terms, let consider n individuals in set $\mathcal{I} = \{1, ..., n\}$ distributed among

N families. The set of founders which are individuals that have no parents in the data is noted $\mathcal{F} \subset \mathcal{I}$. The ages or ages at onset of the disease of all individuals is denoted $T = (T_1, ..., T_n) \in \mathbb{R}^n$ where T_i is the age for individuals i. The genotypes of individuals is denoted $X = (X_1, ..., X_n) \in \{00, 01, 10, 11\}^n$ where 0 represents normal allele and 1 the pathogenic allele and first digit (respectively second) corresponds to the paternal (respectively maternal) allele (i.e. for example $X_i = 01$ means the individual i has a paternal allele 0 and a maternal allele 1). Also $\delta = (\delta_1, ..., \delta_n) \in \{0, 1\}^n$ denotes the status of individuals, δ_i is 1 if the individual i is affected and 0 if unaffected. Therefore this model can be conditioned on the genotype X and decomposed in two subparts, a genetic one, a survival one as follows :

$$\mathbb{P}(T, \delta, X) = \underbrace{\mathbb{P}(X)}_{\text{Genetic Part}} \times \underbrace{\mathbb{P}(T, \delta|X)}_{\text{Survival Part}},$$

- **Genetic Part :** the probability of the genotypes forms a Bayesian network thanks to the family structure as the genotype of one individual only depends on the genotypes of its parents. The set founders of the family $\mathcal{F}$ is the set of individuals that have not parents in the data, the genotypes of these individuals ($\mathbb{P}(X_i)$,$i \in \mathcal{F}$) follow Hardy-Weinberg equilibrium with allelic frequency $f = 1 - \sqrt{1 - \pi_1}$, the non-founders ($\mathbb{P}(X_i|X_{\text{pat}_i}, X_{\text{mat}_i})$,$i \notin \mathcal{F}$) follow Mendelian transmission from parents :

$$\mathbb{P}(X) = \prod_{i \in \mathcal{F}} \mathbb{P}(X_i) \prod_{i \notin \mathcal{F}} \mathbb{P}(X_i|X_{\text{pat}_i}, X_{\text{mat}_i})$$

- **Survival Part :** $\delta_i \in \{0, 1\}$ represents the status (affected or not) of individual i

  - if unaffected then

$$\mathbb{P}(T_i = t, \delta_i = 0|X_i) = \begin{cases} g(t)S_1(t) & \text{if } X_i \neq 00\,; \\ g(t)S_0(t) & \text{if } X_i = 00\,; \end{cases} \propto \begin{cases} S_1(t) & \text{if } X_i \neq 00\,; \\ S_0(t) & \text{if } X_i = 00\,; \end{cases}$$

  - if affected then

$$\mathbb{P}(T_i = t, \delta_i = 1|X_i) = \begin{cases} (1 - G(t))S_1(t)\lambda_1(t) & \text{if } X_i \neq 00\,; \\ (1 - G(t))S_0(t)\lambda_0(t) & \text{if } X_i = 00\,; \end{cases} \propto \begin{cases} S_1(t)\text{RH}(t) & \text{if } X_i \neq 00\,; \\ S_0(t) & \text{if } X_i = 00\,. \end{cases}$$

5.5 Developed method

5.5.1 Idea

Considering that the general population incidence $\lambda(t)$ (and by extension $S(t)$) is known, the model is parameterized by π_1 and $\text{RH}(t)$. The idea is that with this parametrization, $\lambda_0(t)$ and $\lambda_1(t)$ (as well as $S_0(t)$ and $S_1(t)$) can be computed under the constrained general population incidence $\lambda(t)$ through a fixed point method.

From there, the log-likelihood of the model can be computed with the pedigree data via Elston-Stewart algorithm [22, 23] or sum-product algorithm (belief-propagation) [70] in which evidence is based on the calculated $\lambda_0(t)$, $\lambda_1(t)$, $S_0(t)$ and $S_1(t)$.

Therefore the log-likelihood is a function of π_1 and $\text{RH}(t)$ and computable from pedigree data. The maximum likelihood parameters are estimated using BFGS algorithm [48]. The confidence intervals of the estimated parameters can be computed with the Hessian method.

5.5.2 Fixed point method

Idea

$\lambda(t)$ is assumed to be piecewise constant with known cuts (typically for cancer registry with 5-years bins), and $\mathrm{RH}(t)$ also is piecewise constant with known cuts (depend on the model and sometimes on X, e.g. bins $[0, 50]$ and $]50, +\infty[$).

For a given proportion π_1 and $\mathrm{RH}(t)$, we would like to compute $\lambda_0(t)$ such that :

$$S(t)\lambda(t) = \pi_0 S_0(t)\lambda_0(t) + \pi_1 S_1(t)\lambda_1(t)\,.$$

To solve this problem, $\lambda_0(t)$ which is supposed to be continuous, is discretized with a thin cutset. Therefore, it is assumed to be piecewise constant, with cuts every tenth of a year from 0 to 80, and these following fixed-point iterations are performed :

- initialize with $\lambda_0(t) = \lambda(t)$;
- repeat : compute $S_0(t)$ and $S_1(t)$ with current $\lambda_0(t)$ and update

$$\lambda_0(t) = \frac{\lambda(t)S(t)}{\pi_0 S_0(t) + \pi_1 S_1(t)\mathrm{RH}(t)}\,.$$

Simple Example

Let consider a general population incidence with cuts $20, 40, 60, 80$ and bin-specific yearly incidence $0.000, 0.003, 0.005, 0.010, 0.015$. Fixing the parameters at :

- $\pi_1 = 0.0975$;
- RH with cuts 50 and bin-specific values $20, 10$.

Then, λ_0 cuts are assumed to be every tenth of a year from 0 to 80. From this setup, it is possible to computed λ_0, λ_1, $S_0(t)$ and $S_1(t)$ after convergence to the fixed point as shown in figure 5.1.

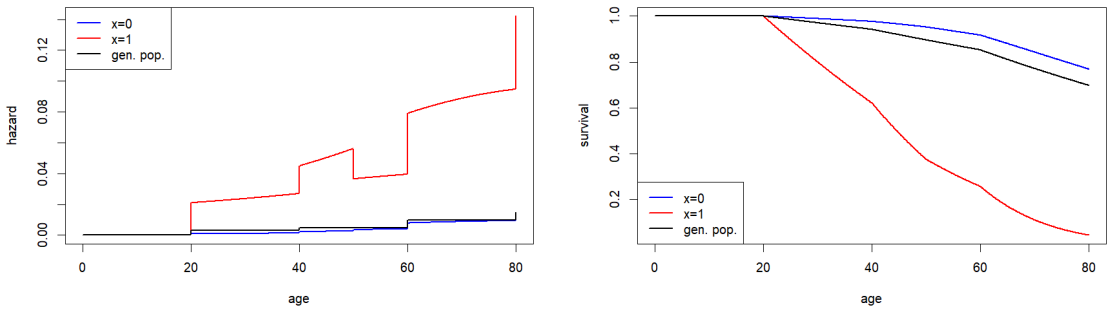


FIGURE 5.1 : Hazard rates and Survivals after fixed-point convergence in simple example.

5.5.3 Log-likelihood computation

For specific parameters $\theta = (\pi_1, \mathrm{RH}(t))$, $\lambda_0(t)$ and $\lambda_1(t)$ (as well as $S_0(t)$ and $S_1(t)$) are computed through the fixed point method. Now the log-likelihood of the model can be written as follows :

$$\mathcal{L}(\theta) = \sum_{\text{Families}} \log \left[\sum_{X} \prod_{i} \underbrace{\mathbb{P}(T_i, \delta_i | X_i; \theta)}_{\text{survival component}} \underbrace{\mathbb{P}(X_i | X_{\text{pat}_i}; X_{\text{mat}_i}; \theta)}_{\text{genetic component}} \right] .$$

This log-likelihood is computable using Elston-Stewart algorithm [22, 23] or sum-product algorithm [70] using $\lambda_0(t)$, $\lambda_1(t)$, $S_0(t)$ and $S_1(t)$ to calculate the evidence. In this article we use *bped*, an C++ implementation of the sum-product algorithm specifically designed for pedigree computation.

5.5.4 Maximum Log-likelihood estimation

As previously explained, the log-likelihood of the model can be computed as a function of the parameters π_1 and $\text{RH}(t)$. $\text{RH}(t)$ being actually a finite number of parameters, for instance in the simple example,

$$\text{RH}(t) = \begin{cases} \text{RH}_1 & \text{if t} \in [0, 50] \,; \\ \text{RH}_2 & \text{if t} \in]50, +\infty[\,. \end{cases}$$

The model comes down to a finite number of parameters, here only 3 $\theta = (\pi_1, \text{RH}_1, \text{RH}_2)$ which are estimated by maximizing the log-likelihood with Broyden–Fletcher–Goldfarb–Shanno (BFGS) algorithm [48] implemented in R with the function *optim*.

$$\hat{\theta}_{\text{ML}} = \arg \max_{\theta} \mathcal{L}(\theta)$$

5.5.5 Variables substitution

The estimated parameters are $\text{RH}(t)$ and π_1. Actually, it is possible to use a variable substitution in order to constrain the parameters of the model. For instance in the model, the genetic disorder has a low allele frequency in the general population. It is therefore interesting to set $\pi_1 \in [0, 0.2]$ meaning that the proportion of pathogenic variants carriers can not be higher that 0.2. With a similar thinking, the relative risk $\text{RH}(t)$ between carriers and non-carriers is expected to be higher than 1, the pathogenic variants carriers being a priori more at risk than the non-carriers. The rules set for the parameters can be find in the litterature or from knowledge from experts.

To apply these rules, the model includes a variable substitution such that :

- $\pi_1 = 0.2 \times \frac{e^{\theta_1}}{1+e^{\theta_1}}$
- $\text{RH}_1 = 1 + e^{\theta_2}$
- $\text{RH}_2 = 1 + e^{\theta_3}$

Therefore the parameters to estimate are $\{\theta_1, \theta_2, \theta_3\}$ which actually set $\pi_1 \in [0, 0.2]$, $\text{RH}_1 > 1$ and $\text{RH}_2 > 1$. But for the results, the article will present the estimated $\{\pi_1, \text{RH}_1, \text{RH}_2\}$ after substitution of estimated $\{\theta_1, \theta_2, \theta_3\}$ for better practical understanding.

5.5.6 Confidence intervals computation

The confidence intervals of the estimated parameters are computed using the Hessian method. The square roots of diagonal elements of the inverted Hessian matrix of the log-likelihood function estimate the standard deviations (SD) of the parameters. If the estimated parameters follow Gaussian distributions (discussed in the appendice), 95% confidence

intervals can be calculated adding and substracting $1.96 \times$ SD to the maximum-likelihood parameters.

The variables substitutions used to constrain parameters being strictly monotone, the confidence intervals for substituted variables are calculated by applying the substition to the border of the intervals.

The function *optim* implemented in R proposed an argument (`hessian=TRUE`) which returns an estimation of the Hessian matrix computed during optimization process.

5.6 Data simulations

5.6.1 Simulations

The developed method is tested on simulated data. The first set of data is simulated accordingly to the model. A second set of data is generated where the relative hazard is not piece-wise constant anymore and follows a Weibull function. This set allows to test the robustness of the method when the data do not follow strictly the model but are close enough (the Weibull function being close to a two part piecewise constant model). The familial structures are real and taken from an AH-HP dataset of families with *SFTPA1* and *SFTPA2* pathogenic variants carriers.

For both standard simulations and robustness analysis :

- 2000 datasets are generated.
- Each dataset is composed 744 inviduals over 28 families.
- Autosomal dominant transmission model with 1 gene and 2 alleles ("normal" and "pathogenic"). Genotypes of founders follow Hardy-Weinberg equilibrium.
- proportion of carriers is $\pi_1 = 0.0975$

Then the difference is on RH(t) as shown in figure 5.2 :

- standard analysis : $\text{RH}_1 = 20$, $\text{RH}_2 = 10$ with a cut at 50 years old ;
- robustness analysis : RH(t) is a Weibull function (shape = 3 and scale = 1.5).

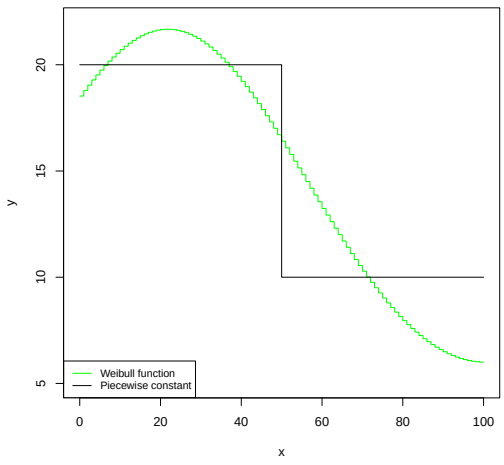


FIGURE 5.2 : RH(t) for standard (black) and robustness (green) simulations.

5.6.2 Missing data

Missing data are introduced in the generated dataset in order to mimic real data. Genotypes and phenotypes are randomly considered missing. Four levels of missingness are considered including the oracle :

- Oracle : all the data is known.
- 30% : about 30% of the data is missing. 20% of the phenotypes and 40% of genotypes are missing.
- 50% : about 50% of the data is missing. 35% of the phenotypes and 65% of genotypes are missing.
- 70% : about 70% of the data is missing. 50% of the phenotypes and 90% of genotypes are missing.

5.6.3 Augmented data

For the standard simulations, augmented datasets are generated to analyse how the developed method scales with dataset's size. To do so, 2000 datasets of 1488 individuals over 56 families (initial size $\times 2$) and 2000 datasets of 2976 individuals over 112 families (initial size $\times 4$) are generated.

5.7 Results on simulations

5.7.1 Results on standard and robustness data

The results presented are violin plots of the parameters estimated by the proposed method with Oracle, 30%, 50% and 70% of missing data. The parameters estimated from standard simulations are presented in Figure 5.3 and those estimated from robustness simulations in Figure 5.4.

The results show a great fit to the expected values of the parameters both from the standard simulations and the robustness ones. The median values of each parameter for every level of data missingness except the 70% level on robustness analysis which does not match exactly but remains very close. The variance increases with the level of missing data as expected.

There are few outliers in the estimations that seems to reach the boundaries fixed for our parameters (*i.e.* $\mathrm{RH}_1 > 1$, $\mathrm{RH}_2 > 1$ and $\pi_1 \in [0, 0.2]$).

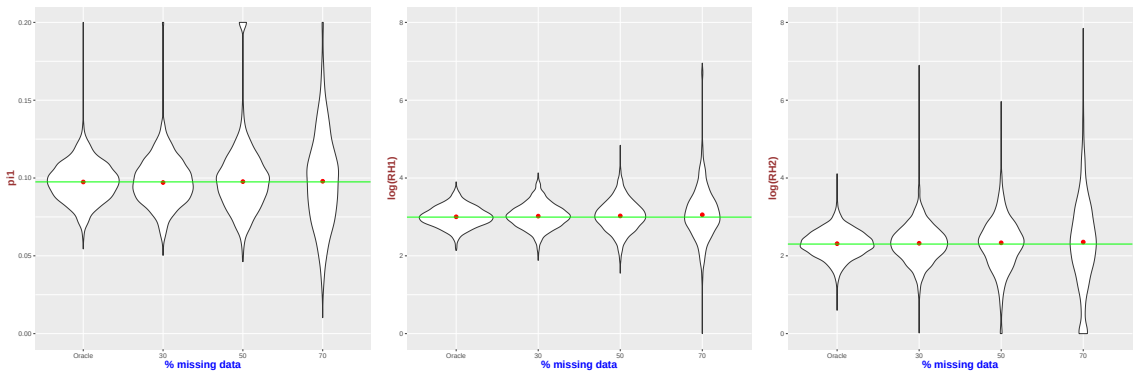


FIGURE 5.3 : Violin plots of π_1, $\log(\mathrm{RH}_1)$ and $\log(\mathrm{RH}_2)$ estimation on standard simulations. Green line represents the real parameter to estimate.

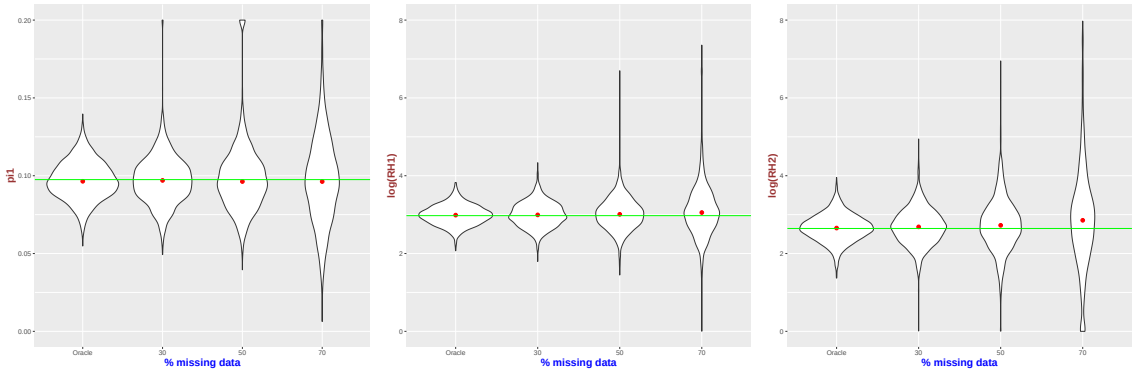


FIGURE 5.4 : Violin plots of π_1, $\log(\mathrm{RH}_1)$ and $\log(\mathrm{RH}_2)$ estimation on robustness simulations. Green line represents the real parameter to estimate.

5.7.2 Results on augmented data

The results presented are violin plots of the parameters estimated by the proposed method with Oracle, 30%, 50% and 70% of available data on augmented data. The parameters estimated from datasets of standard simulation size $\times 1$, $\times 2$ and $\times 4$ are presented on the same figures to have a better overview of the results. π_1 estimations are shown in Figure 5.5, $\log(\mathrm{RH}_1)$ in Figure 5.6 and $\log(\mathrm{RH}_2)$ in Figure 5.7.

The results show again a great fit to the expected values of the parameters. The bigger the size, the better the estimations as the variances decrease with datasets size. The variances still increases with the level of missing data as expected.

There are again few outliers in the estimations that seems to reach the boundaries fixed for our parameters on the datasets of size $\times 2$ for π_1.

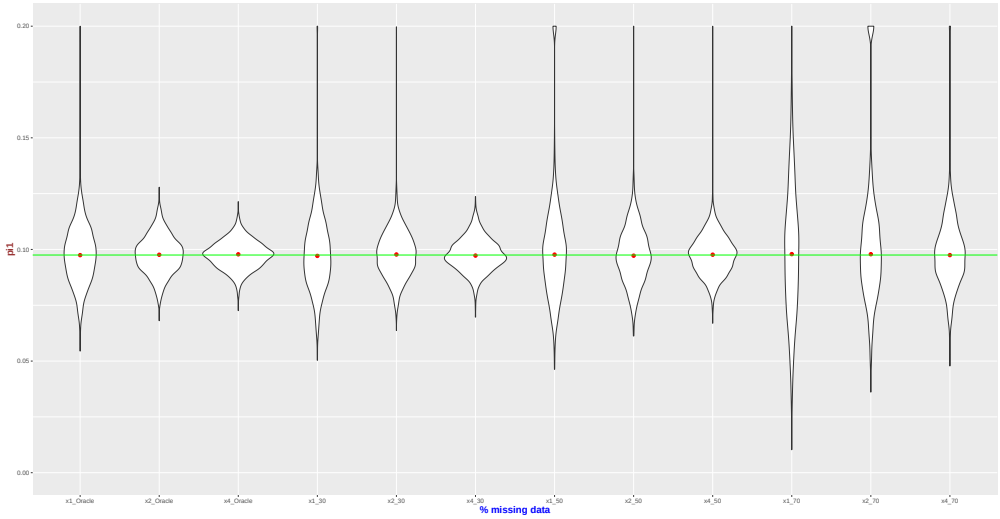


FIGURE 5.5 : Violin plots of π_1 for various datasets sizes and data missingness. Green line represents the real parameter to estimate.

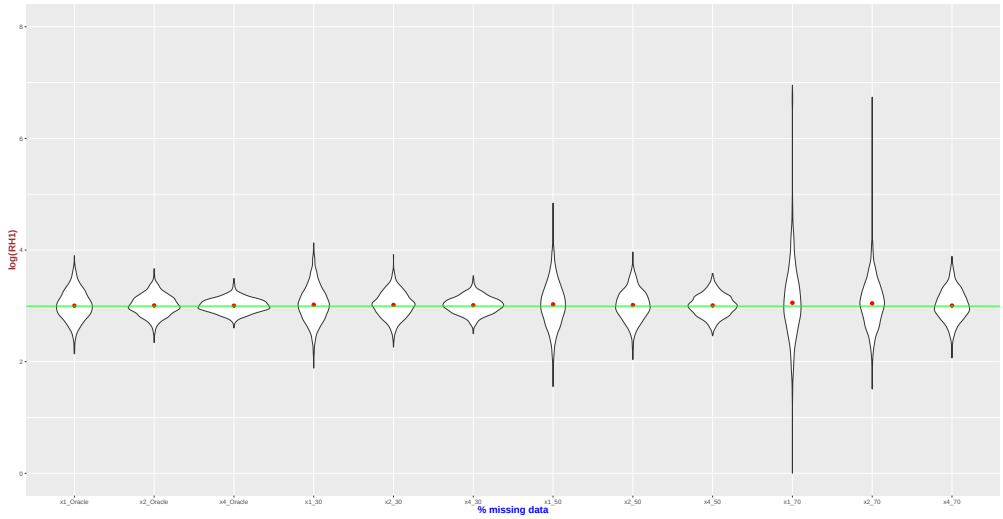


FIGURE 5.6 : Violin plots of $\log(RH_1)$ for various datasets sizes and data missingness. Green line represents the real parameter to estimate. Green line represents the real parameter to estimate.

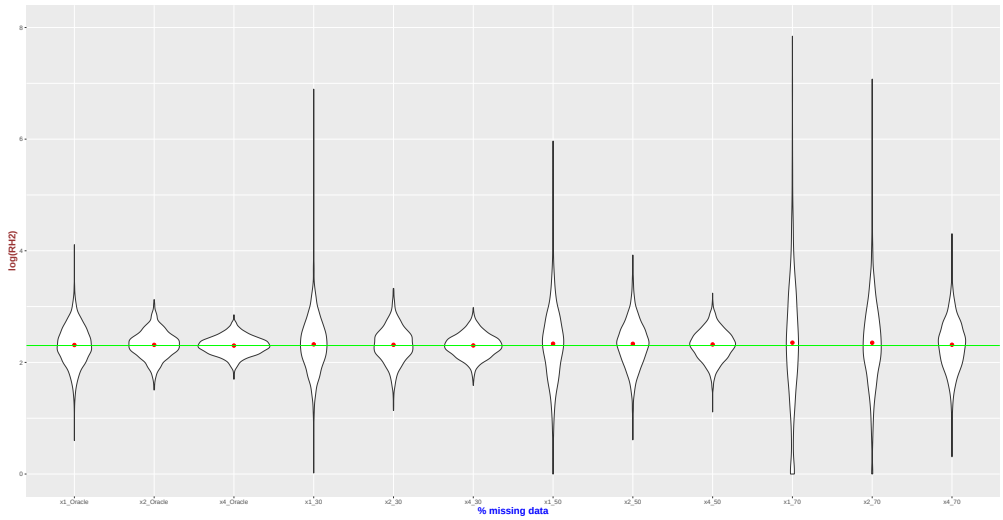


FIGURE 5.7 : Violin plots of $\log(RH_2)$ for various datasets sizes and data missingness. Green line represents the real parameter to estimate. Green line represents the real parameter to estimate.

5.7.3 Confidence intervals dependance on dataset's size

The coverage probability of confidence intervals computed for 100 datasets with the Hessian methods are presented in Table 5.1. The distributions of confidence intervals' sizes are presented for each parameter, level of missingness and datasets' size in Figure 5.8. Green represents the standard simulations, purple the datasets of size $\times 2$ and in brown the datasets of size $\times 4$.

The coverage probabilities decrease with higher level of missingness in the data. It seems that the coverage probability increases with the size of the datasets but it is not the case for the dataset size $\times 2$ at Oracle and 50% missing data level. The 30% level of missingness showcases the worst coverage probability overall for every parameter and datasets' size.

The size of the confidence intervals decreases the higher the dataset size is. Similarly the size of the confidence intervals decreases with low level of missingness. It is expected to perform better, the more information are known.

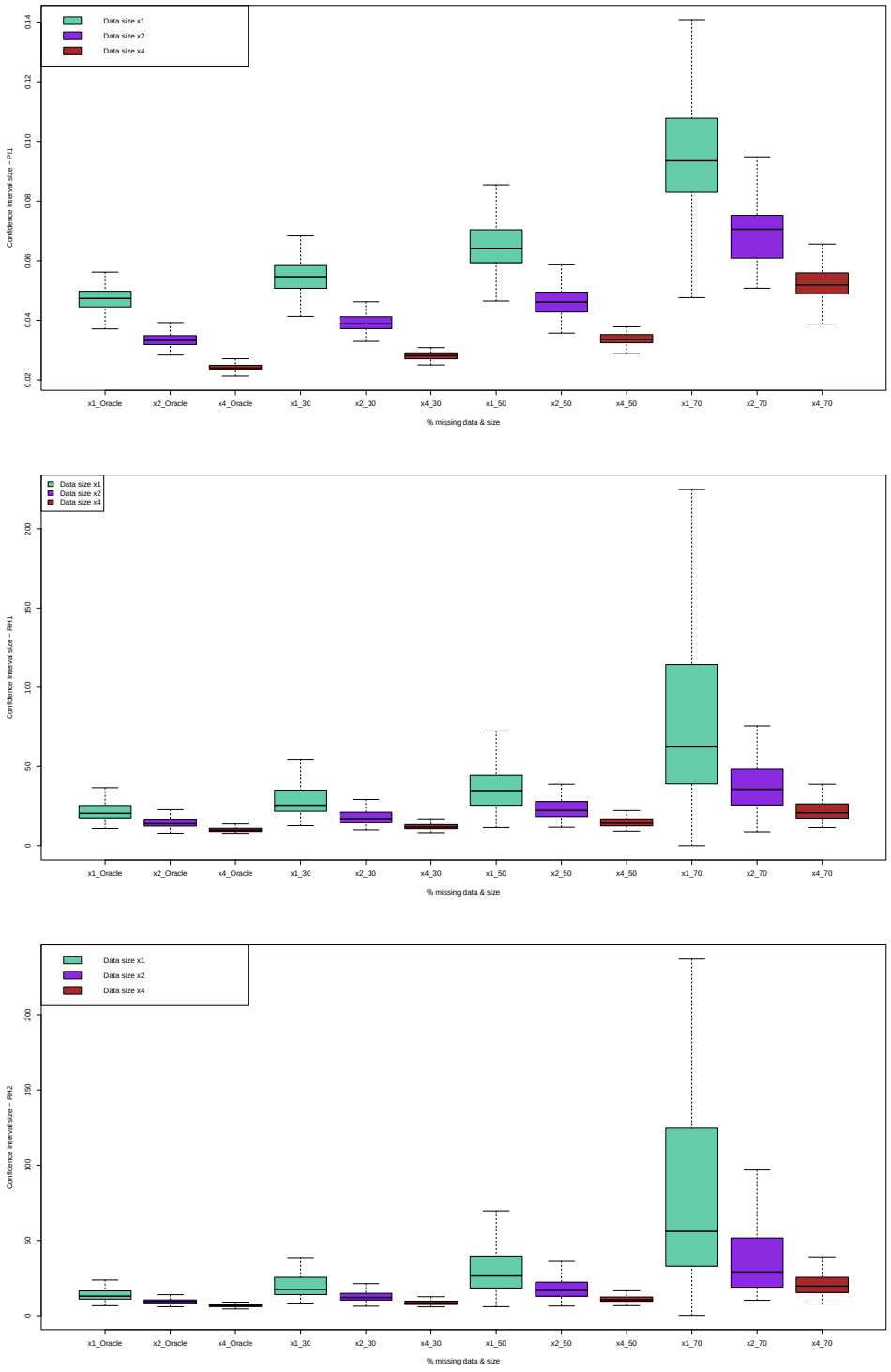


FIGURE 5.8 : Boxplots of the confidence intervals sizes for parameters π_1, RH_1 and RH_2.

	×1, Oracle	×2, Oracle	×4, Oracle	×1, 30%	×2, 30%	×4, 30%
π_1	0.95 [0.90,0.99]	0.96 [0.92,0.99]	0.96 [0.92,0.99]	0.94 [0.89,0.98]	0.95 [0.90,0.99]	0.96 [0.92,0.99]
RH_1	0.94 [0.89,0.98]	0.91 [0.85,0.96]	0.97 [0.93,1.0]	0.93 [0.88,0.98]	0.94 [0.89,0.98]	0.99 [0.97,1.0]
RH_2	0.95 [0.90,0.99]	0.90 [0.84,0.95]	0.91 [0.85,0.96]	0.96 [0.92,0.99]	0.89 [0.83,0.95]	0.94 [0.89,0.98]
	×1, 50%	×2, 50%	×4, 50%	×1, 70%	×2, 70%	×4, 70%
π_1	0.95 [0.90,0.99]	0.92 [0.86,0.97]	0.97 [0.93,1.0]	0.89 [0.83,0.95]	0.91 [0.85,0.96]	0.97 [0.93,1.0]
RH_1	0.94 [0.89,0.98]	0.91 [0.85,0.96]	0.96 [0.92,0.99]	0.90 [0.84,0.95]	0.91 [0.85,0.96]	0.95 [0.90,0.99]
RH_2	0.95 [0.90,0.99]	0.93 [0.88,0.98]	0.95 [0.90,0.99]	0.82 [0.74,0.89]	0.89 [0.83,0.95]	0.93 [0.88,0.98]

TABLE 5.1 : Coverage probability for each parameter and for each dataset size and missing data type.

5.8 Discussion

According to the results, the proposed method seems to estimate correctly the model parameters. The more data are available, the better are the estimations.

When applied to simulations generated from a sligthly different model, the method still estimates correctly the parameters (as observed according to the model on the simulated datasets).

The confidence intervals are very large for the 70% level of missingness but narrow with larger datasets and less missing data. However, the coverage probabilities do not fit to the expected 95% confidence intervals. The main reason probably being that the estimated parameters $\{\theta_1, \theta_2, \theta_3\}$ are mainly not Gaussian according the Shapiro-Wilk test as shown in the appendice. The confidence intervals remain useful with bigger datasets and less missing data. It is still possible to use a bootstrap method to estimate the confidence intervals also shown in the Appendice, the downside being increased computing cost.

5.9 Conclusion and perspectives

In conclusion, this article proposes a new method to estimate the penetrance/survival of a genetic disease with sporadic cases from pedigree data. Previous published methods generally make three assumptions. The first one is that the proportion of pathogenic variant carriers in the general population is known. The second one is that the incidence of the disease for the general population is also known. Finally these methods approximate the incidence of the non-carriers by the incidence of the general population. This last assumption is not far from reality in the case of very rare pathogenic allele and high penetrance but falls down as the allele is more and more common and the disease moderatly penetrant.

The proposed method generalises previous methods, relying only on the known incidence in general population assumption. To do so, the method incorporates the proportion of carriers in the population as a parameter of the model and use a fixed-point method to compute the incidences of carriers and non-carriers constrained by the incidence in the general population. The cost of this generalization is a parametrization of model.

The method performed well at estimating the parameters of the model on a simple example and on a robustness analysis. The method was also tested on a different set of simulations for which the results are presented in the appendice.

The main perspective of this work is to test the method on biased data which are the norms for collected pedigree in genetics. Indeed, patients are generally selected to be addressed to genetic counselling through specific sets of rules depending on countries/hospitals, this selection induced a first bias. Then, amongst these selected patients, only the carriers are generally followed and their family tested, which includes a second layer of selection.

Moreover, it would be interesting to relax the assumption on the relative hazard $\mathrm{RH}(t)$

which is currently piecewise constant. For instance, the robustness analysis is performed using a Weibull function as relative hazard, which is a function parameterized by only two parameters (scale and shape). It would interesting to implement more diverse relative hazard options. This perspective also leads to another question which would be to study model selection with this method. From an unspecify model, it would be interesting to determine the optimal numbers and positions of cuts in $\mathrm{RH}(t)$.

Finally this model makes the assumption that the phenotypes are independant conditionally to the genotypes. This is a standard assumption which has its limits especially when the genetic disease presents major environmental (smoking for lung cancer for instance) and/or polygenic risk factors. It would be interesting to add exposition variable or frailty to the model which it is not straightforward because of the general population incidence constraint.

Appendices to Chapter 5

5.A Normality of estimated parameters

5.A.1 Method

The normality of the estimated parameters $\{\theta_1, \theta_2, \theta_3\}$ is tested with a Shapiro-Wilk test which tests the null hypothesis "the tested distribution is Gaussian". Therefore, if the p-value is less than 0.05 (data is likely to occur less than 5% of the time under the null hypothesis), the null hypothesis is rejected and the tested distribution does not follow a Gaussian distribution.

5.A.2 Results

The Shapiro-Wilk test results for each parameter, each level of missingness are presented in Table 5.2 for standard dataset size, Table 5.3 for dataset size $\times 2$ and Table 5.4 for dataset size $\times 4$.

$\times 1$	Oracle	30%	50%	70%
θ_1	1.36e-63	1.73e-67	5.23e-66	1.30e-49
θ_2	0.269	0.000223	0.00161	2.19e-53
θ_3	2.99e-12	8.97e-31	1.31e-45	8.52e-48

TABLE 5.2 : Shapiro-Wilk test p-values for each parameter and each level of missingness on dataset of standard size.

$\times 2$	Oracle	30%	50%	70%
θ_1	0.842	9.12e-50	1.07e-68	2.75e-63
θ_2	0.360	0.761	0.0708	1.78e-26
θ_3	0.00105	8.27e-07	1.15e-08	4.45e-49

TABLE 5.3 : Shapiro-Wilk test p-values for each parameter and each level of missingness on dataset of size $\times 2$

×4	Oracle	30%	50%	70%
θ_1	0.417	0.000697	4.17e-64	5.89e-69
θ_2	0.0747	0.903	0.136	0.249
θ_3	0.131	0.509	0.000413	7.37e-14

TABLE 5.4 : Shapiro-Wilk test p-values for each parameter and each level of missingness on dataset of size ×4

According to the results, the estimated parameters $\{\theta_1, \theta_2, \theta_3\}$ are mostly not Gaussian. However, it seems that the bigger the data (increased dataset size), the closer to Gaussian the distributions are. Similarly, the p-values increase as the missingness decrease (with highest p-values for the Oracle).

5.B Results on modified simulations

This section contains the results obtained with the proposed method on a different set of simulations.

5.B.1 Simulations

The set of data is simulated accordingly to the model. The familial structures are real and taken from an AH-HP dataset of families with *SFTPA1* and *SFTPA2* pathogenic variants carriers. The families with more than 30 individuals are removed. The remaining data represent 206 individuals over 17 families, each family is then copied to obtain 412 individuals over 32 families. From this set of families' structures, the genotypes and phenotypes are generated the same way as the previous simulations.

- 100 datasets are generated.
- Each dataset is composed 412 individuals over 32 families.
- Autosomal dominant transmission model with 1 gene and 2 alleles ("normal" and "pathogenic"). Genotypes of founders follow Hardy-Weinberg equilibrium.
- proportion of carriers is $\pi_1 = 0.0975$
- the relative hazard $\mathrm{RH}_1 = 20$, $\mathrm{RH}_2 = 10$ with a cut at 50 years old ;

5.B.2 Missing data

In this scenario, only the genotypes can be missing, the phenotypes are all known. These data represent more condensed families, with less individuals but with more information on the phenotypes. Therefore, the data showcase only 412 people (compared to the 744 previously), over 32 families (28 on the other dataset).

Different levels of missingness are generated :

- Oracle : all the genotypes are known.
- 50% : 50% of the genotypes are missing completely at random.
- 66% : 66% of the genotypes are missing completely at random.
- 75% : 75% of the genotypes are missing completely at random.
- 90% : 90% of the genotypes are missing completely at random.

5.B.3 Results

The results on modified simulations showcase the same trends as the results on standard simulations. The Violin plots of the estimated π_1, RH_1 and RH_2 are presented on Figure 5.9.

The coverage probabilities for each parameter and each level of missingness are presented on Table 5.5 and the distribution of the size of confidence intervals are shown on boxplot in Figure 5.10.

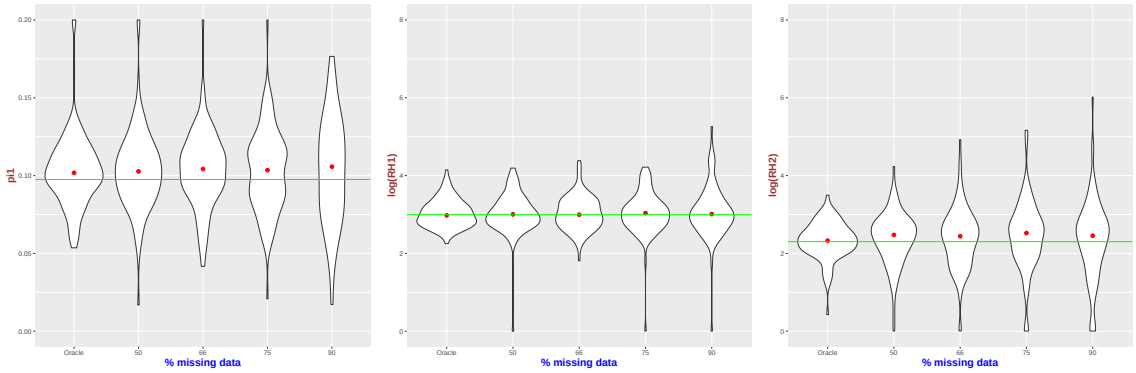


FIGURE 5.9 : Violin plots of π_1, $\log(RH_1)$ and $\log(RH_2)$ estimation on modified standard simulations. Green line represents the real parameter to estimate.

	Oracle	50%	66%	75%	90%
π_1	0.94 [0.89,0.98]	0.89 [0.83,0.95]	0.93 [0.88,0.98]	0.93 [0.88,0.98]	0.94 [0.89,0.98]
RH_1	0.96 [0.92,0.99]	0.89 [0.83,0.95]	0.95 [0.90,0.99]	0.91 [0.85,0.96]	0.93 [0.88,0.98]
RH_2	0.97 [0.93,1.0]	0.90 [0.84,0.95]	0.91 [0.85,0.96]	0.89 [0.83,0.95]	0.87 [0.80,0.93]

TABLE 5.5 : Coverage probability for each parameter and for each dataset size and missing data type.

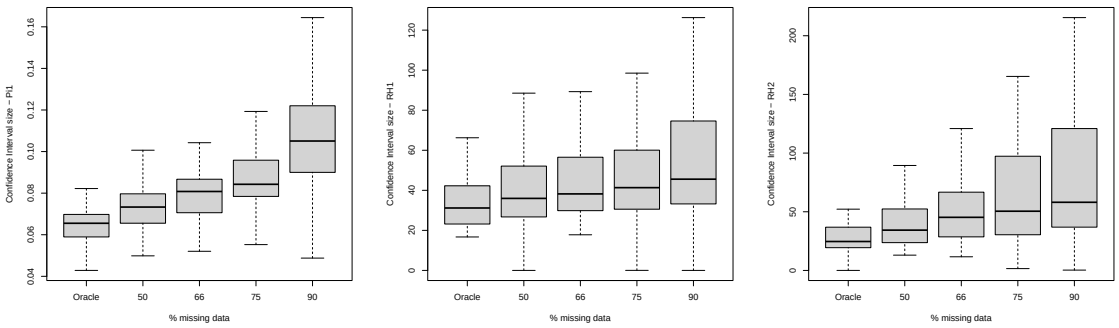


FIGURE 5.10 : Boxplots of the confidence intervals sizes for parameters π_1, RH_1 and RH_2 on modified simulations.

5.C Bootstrap for confidence interval estimation

5.C.1 Method

The bootstrap method is a resampling technique used in statistics to estimate the distribution of parameters by repeatedly resampling with replacement from the observed data.

Here is how it is performed in the context of this article :

- If the original dataset is composed of N families, randomly draw N families with replacement from the original dataset. This means that some observations may be repeated in the resampled dataset, while others may be omitted.
- Estimate the parameters with the method from the dataset generated with resampling.
- Repeat this procedure in order to generate 100 (for instance) values of estimated parameters.
- Take the 2.5 and 97.5 percentiles for the parameters which are the lower and upper bounds of the confidence interval for the estimated parameter.

5.C.2 Results

The bootstrap method is applied to 50 datasets from the standard simulations at all the levels of missingness. Each dataset is resampled 100 times. The coverage probability results are shown in Table 5.6.

	Oracle	30%	50%	70%
π_1	0.94 [0.86, 1.0]	0.92 [0.84, 0.98]	0.96 [0.90, 1.0]	0.98 [0.94, 1.0]
RH_1	0.94 [0.86, 1.0]	0.96 [0.90, 1.0]	0.90 [0.82, 0.98]	0.88 [0.78, 0.96]
RH_2	0.94 [0.86, 1.0]	0.90 [0.82, 0.98]	0.78 [0.66, 0.88]	0.92 [0.84, 0.98]

TABLE 5.6 : Coverage probability for each parameter and for each dataset size and missing data type with the bootstrap strategy.

Chapitre 6

Correction du biais d'ascertainment pour la méthode développée

Résumé

Les données de pédigrée collectées en génétique sont généralement soumis à un biais de sélection important. En effet, les patients sont adressés en conseil génétique selon un certain nombre de critères médicaux (cas jeunes ou nombre de cas important de la maladie dans la famille), ce qui introduit un premier biais de sélection. Ensuite, dans les familles des patients testés génétiquement et effectivement porteurs de variants pathogènes, ce sont les membres ayant la plus forte probabilité d'être porteur qui se voient généralement offrir un test génétique. Ceci induit un second biais de sélection où les individus génotypés sont souvent porteur de variants pathogènes.

Ces biais de sélection se cumulent et présentent un problème majeur pour toute méthode statistique utilisant ces données et faisant l'hypothèse qu'elles sont représentatives d'une population générale. C'est, entre autres, le cas de la méthode développée au chapitre 4 pour l'estimation de courbe de pénétrance d'une maladie mendélienne présentant des cas sporadiques à partir de données de pédigrée. En effet, la méthode a été développé et testée sur des données de pédigrée simulées qui ne présentent pas de biais de sélection. D'une part, les familles n'étaient pas sélectionnées sur un critère particulier et, d'autre part, les porteurs et les non-porteurs avaient la même probabilité d'être testés.

Dans ce chapitre, les biais introduits dans les estimations par la méthode développée sont mis en évidence, par des simulations de jeux de données biaisés. Par la suite, plusieurs méthodes de correction de biais de sélection sont présentées et adaptées au contexte des données simulées. L'étude s'intéresse notamment à la *Genoptype-Restricted Likelihood* (GRL) [10] et la *Proband's phenotype Exclusion Likelihood* (PEL) [1] pour les méthodes publiées, ainsi qu'à une correction simple, où la seule hypothèse utilisée est que la fréquence allélique du variant pathogène est connue.

Sur les données simulées, la combinaison de notre méthode avec la GRL semble ne pas fonctionner. Cependant, la simple hypothèse "fréquence allélique connue" permet à notre méthode de corriger une partie du biais de sélection, notamment le biais induit par les tests génétiques, les résultats présentent cependant toujours un biais. La combinaison de notre méthode avec la correction PEL, toujours sous l'hypothèse "fréquence allélique connue", semble prometteuse. En effet, les résultats présentent un biais inférieur qu'avec l'hypothèse sur la fréquence allélique seule.

En appliquant la combinaison de notre méthode avec la PEL, la courbe de pénétrance du cancer broncho-pulmonaire pour les porteurs de variants pathogènes *SFTPA1* et *SFTPA2* est recalculée.

Valorisation : Ce travail est effectué par Lucas Ducrot sous la direction de Grégory Nuel (DR CNRS - LPSM - Sorbonne Université). C'est un travail en cours qui n'a pour le moment pas fait l'objet de communication.

Contents

6.1 Abstract

The pedigree data collected in genetics are generally subject to significant selection bias (called ascertainment bias in genetics). Indeed, patients are typically referred to genetic counseling based on a set of medical criteria (such as young cases or a significant number of cases of the disease in the family), introducing a first selection bias. Then, in the families of patients genetically tested and found to carry pathogenic variants, it is usually the members with the highest probability of being carriers who are offered genetic testing. This introduces a second selection bias where genotyped individuals show a higher proportion of carriers of pathogenic variants compared to non-carriers.

These selection biases add up and pose a major problem for any method or statistic using such data, assuming they are representative of a general population. This is the case, among others, for the method developed in Chapter 4 for estimating the penetrance curve of a Mendelian disease with sporadic cases from pedigree data. The method is tested on simulated pedigree data that do not exhibit selection biases. Firstly, families are not selected based on a specific criterion, and secondly, carriers and non-carriers have an equal probability of being tested.

In this chapter, the biases introduced in the estimates by the developed method are highlighted through simulations of biased datasets. Subsequently, several methods for

correcting selection bias are presented and adapted to the context of the method. The study focuses on the Genotype-Restricted Likelihood (GRL) [10] and the Proband's Phenotype Exclusion Likelihood (PEL) [1] for published methods, as well as a simple correction, where the sole assumption is that the allele frequency of the pathogenic variant is known.

By applying the most promising correction methods, the penetrance curve for bronchopulmonary cancer for carriers of pathogenic variants in *SFTPA1* and *SFTPA2* is recalculated.

6.2 Introduction

Patients are typically referred to genetic counseling based on a set of medical criteria (such as young cases or a significant number of cases of the disease in the family). These rules introduce a selection bias. Then, in the families of patients genetically tested and found to carry pathogenic variants, it is usually the members with the highest probability of being carriers who are offered genetic testing. This introduces a second selection bias where genotyped individuals show a higher proportion of carriers of pathogenic variants compared to non-carriers.

These selection biases are a major problem for any method or statistic using such data, assuming they are representative of a general population. This is the case, among others, for the method developed in Chapter 4 for estimating the penetrance curve of a Mendelian disease with sporadic cases from pedigree data. The method is tested on simulated pedigree data that do not exhibit selection biases. Firstly, families are not selected based on a specific criterion, and secondly, carriers and non-carriers have an equal probability of being tested.

It is crucial to understand how this selection affects the results of the developed method and find tools to correct the introduced bias. To do so, we propose, in this article, to test our method on biased simulated datasets in order to highlight the bias results.

6.2.1 Context

Here is a reminder of the context and notations from which the data are generated :

- an autosomal dominant disorder of one gene and two alleles ("normal" 0 and "pathogenic" 1), the genotype component $X \in \{00, 01, 10, 11\}$ ($X = 00$ for non-carrier, $X \neq 00$ for carrier) where the first is the paternal allele and the second the maternal allele ;
- the proportions of carriers in the population is denoted π_1 and non-carriers π_0 (with $\pi_0 = 1 - \pi_1$) ;
- the pathogenic allele frequency f is directly linked to π_1 as :

$$f = 1 - \sqrt{1 - \pi_1}$$

- the specific conditional hazard rates are $\lambda_1(t)$ for carriers and $\lambda_0(t)$ for non-carrier ;
- We denote the relative hazard between carriers and non-carriers $\mathrm{RH}(t)$ such as $\lambda_1(t) = \mathrm{RH}(t) \times \lambda_0(t)$. It is piecewise constant with two components and a cut-off at age 50 such that :

$$\mathrm{RH}(t) = \begin{cases} \mathrm{RH}_1 & \text{if t} \in [0, 50] \,; \\ \mathrm{RH}_2 & \text{if t} \in]50, +\infty[\,. \end{cases}$$

- $S(t)$ (resp. $S_0(t)$ and $S_1(t)$) is the survival function (resp. conditional survival functions) associated with hazard $\lambda(t)$ (resp. $\lambda_0(t)$ and $\lambda_1(t)$) such as

$$S(t) = \exp\left(-\int_0^t \lambda(u)du\right);\ \ S_0(t) = \exp\left(-\int_0^t \lambda_0(u)du\right);\ \ S_1(t) = \exp\left(-\int_0^t \lambda_1(u)du\right).$$

- We consider as well a censorship event (which will not be needed) with a distribution function $g(t)$ and a repartition function $G(t)$ such as

$$G(t) = \int_0^t g(u)du\,.$$

- the general population incidence $\lambda(t)$ is known and piecewise constant, which is often the case in medical registry ;

The data we are dealing with are pedigree data which contain families' structures, censored ages or ages at diagnosis, status (affected or not) and genotypes if known (the genotypes are mostly unknown in pedigree data, therefore it is a latent variable of the model, but few genotypes are known because of genetic testing).

6.2.2 Simulations

The set of data is simulated accordingly to the described model. The familial structures are real and taken from an AH-HP dataset of families with *SFTPA1* and *SFTPA2* pathogenic variants carriers. The families with more than 30 individuals are removed. The remaining data represent 206 individuals over 17 families.

In this article, we propose to test the method on two scenario which represent sort of limit cases, the first one being a common pathogenic allele (allele frequency around 1/20) and the second one being a very rare pathogenic allele (allele frequency around 1/4000).

The general population incidence presents cuts at $20, 40, 60, 80$ and the bin-specific yearly incidences are $0.000, 0.003, 0.005, 0.010, 0.015$.

Common pathogenic allele scenario

- 100 datasets are generated.
- Autosomal dominant transmission model with 1 gene and 2 alleles ("normal" and "pathogenic"). Genotypes of founders follow Hardy-Weinberg equilibrium.
- proportion of carriers is $\pi_1 = 0.0975$
- standard analysis : $\mathrm{RH}_1 = 20$, $\mathrm{RH}_2 = 10$ with a cut at 50 years old ;

Rare pathogenic allele scenario

- 100 datasets are generated.
- Autosomal dominant transmission model with 1 gene and 2 alleles ("normal" and "pathogenic"). Genotypes of founders follow Hardy-Weinberg equilibrium.
- proportion of carriers is $\pi_1 = 0.0005$
- standard analysis : $\mathrm{RH}_1 = 10$, $\mathrm{RH}_2 = 5$ with a cut at 50 years old ;

6.2.3 Ascertainment

The first selection encountered is to be referred to the geneticist. Generally, the rules of ascertainment are defined at hospital or group of hospitals level. Typically, probands (patients entering the genetic counselling care) are selected if their family and themself showcase multiple and/or young cases of the disease. In order to model this selection, we propose the following rules of ascertainment :

- the proband is affected by the disease before the age of 45 and is carrier of pathogenic variant. It also has parents in the data (this makes the proband central in the family, even though it is not garanted).
- The family showcases at least another affected individual before the age of 45.

Each dataset is composed of 32 families. We generate thousands of families according to the model, then keep only the families that respect the ascertainment criteria and finally select randomly 32 families among them. Therefore each dataset may have varying number of individuals.

6.2.4 Genetic testing scenarios

The second selection encountered is the selection of individuals which are tested in the family after the ascertainment. In this article, we propose different testing scenarios :

- Oracle : all the genotypes are known.
- GRL-like : the genotypes of the proband and its parents are known (we call it GRL-like because this is the assumption made on the simulated data in the GRL article [10]).
- All under 45 : the genotypes of all family members affected by the disease under the age of 45 are known (which include the proband).
- Proband and another 45 : the genotypes of the proband and another family member affected by the disease under the age of 45 are known (meaning only two genotypes are known).
- Proband : only the genotype of the proband is known.

6.2.5 Results on selected datasets

The method is applied on all scenarios (common and rare pathogenic alleles, different genetic testings) in order to highlight potential bias in the estimations due to selection bias. The results are presented as violin plots of the estimated parameters (calculated from 100 estimations each). The common pathogenic allele scenario is shown in Figure 6.1. The rare pathogenic allele scenario is shown in Figure 6.2.

It is clear the selection process induces bias in the estimations in all the scenarios as shown on both Figures 6.1 and 6.2. The affected parameters are especially π_1 and RH_1, which is coherent with the selection process. Indeed, the proportion of pathogenic variant carriers is the selected data is much higher than in the real population, because of the ascertainment. Similarly, selecting families with multiple young cases (two under the age of 45) induces a natural bias in RH_1 leading to over-estimation. Finally, the estimation of RH_2 does not seem to be affected by the selection process which is interesting to notice.

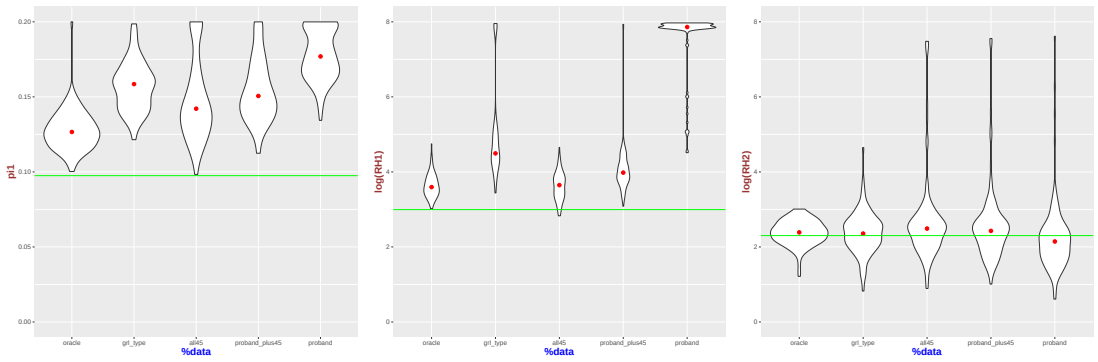


FIGURE 6.1 : Estimation of parameters π_1, $\log(\mathrm{RH}_1)$ and $\log(\mathrm{RH}_2)$ on selected datasets with common pathogenic allele, without correction.

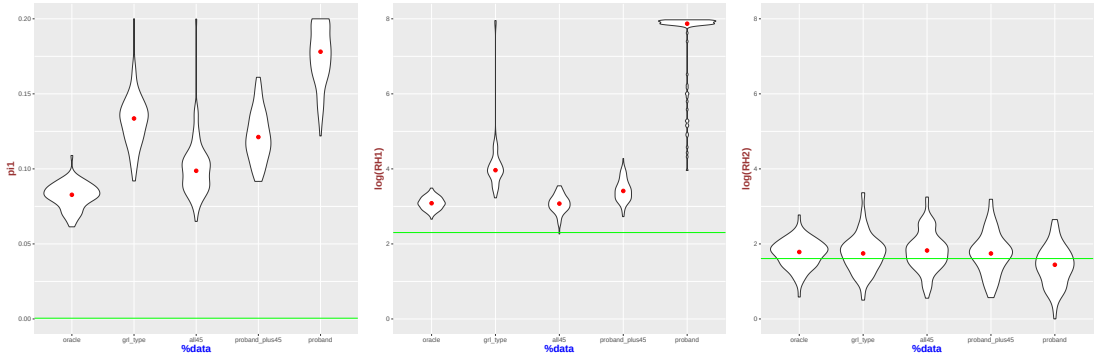


FIGURE 6.2 : Estimation of parameters π_1, $\log(\mathrm{RH}_1)$ and $\log(\mathrm{RH}_2)$ on selected datasets with rare pathogenic allele, without correction.

6.2.6 Objective

Now that the bias in the estimations due to the selection process is hightlighted. The objective of this article is to apply different known corrections with the method and analyse their performances.

The corrections which are implemented are the Proband's phenotype Exclusion Likelihood (PEL) [1] and the Genotyped-Restricted Likelihood (GRL) [10]. Both methods are detailed in the Material and methods section.

Interestingly, both methods are based on a common assumption that the pathogenic allele frequency is known. Therefore, it is vain to expect a correction of the parameter π_1 with these methods. However, it was tested and the results are presented in the appendice.

Because both corrections imply a known pathogenic allele frequency, it is consider as so for the rest of the article, the only parameters to estimate are RH_1 and RH_2.

6.3 Material and methods

In this section, the different correction methods are presented. As explained in Objective part of the Introduction, π_1 is considered known although the methods are also implemented with π_1 unknown. The results when π_1 is unknown are presented in the appendice.

6.3.1 Developed method with known allele frequency

The first applied correction is known allele frequency. It is not a specifically a correction but it is the assumption made for the other corrections. Therefore, it is interesting to see how the method behaves under this assumption alone.

6.3.2 Proband's phenotype Exclusion Likelihood

The second correction applied is the Proband's phenotype Exclusion Likelihood (PEL) [1]. It is used in various genetic studies [4, 58] to correct the ascertainment bias. The correction is to consider the probands uninformative toward the disease, *i.e.* unaffected at age 0. However, their genotypes (carrier/non-carrier) were preserved and used in the analysis. Then, the method can be performed on the corrected dataset (by maximising the likelihood).

The idea behind the method is that the probands are reflective of the ascertainment criteria and hold information about the potential of their parents to have affected children. The rest of the siblings in the family then offer an unbiased estimate of the ratio between affected and unaffected individuals.

6.3.3 Genotype-Restricted Likelihood

The third correction applied is the Genotype-Restricted Likelihood (GRL) [10]. The correction is to consider the model conditionned on the proband being a carrier. Therefore the function to maximize becomes a ratio of Likelihoods with the numerator being the likehood of the model when the known genotypes are indeed known and the denominator is the likelihood of the model when only the genotype of the proband is known.

6.4 Results

6.4.1 Developed method with known allele frequency

The results of the method when π_1 is known are presented in Figure 6.3 for the common pathogenic allele and in Figure 6.4 for the rare pathogenic allele.

Interestingly, this simple simplification seems to greatly correct the bias due to genetic testing as all the scenarios showcase very similar results. However, there is still over-estimation of RH_1 both in the common and rare pathogenic allele cases. RH_2 seems well estimated but that was already the case without correction.

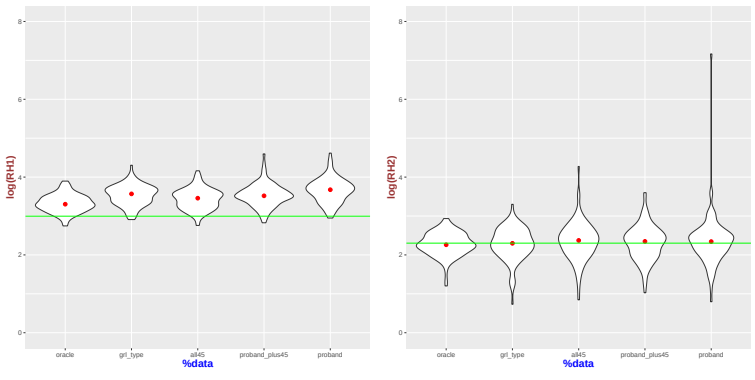


FIGURE 6.3 : Estimation of parameters $\log(RH_1)$ and $\log(RH_2)$ on datasets with common pathogenic allele and π_1 known.

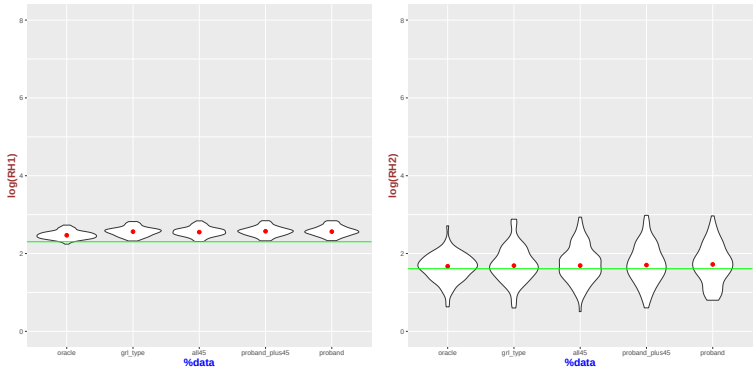


FIGURE 6.4 : Estimation of parameters $\log(RH_1)$ and $\log(RH_2)$ on datasets with rare pathogenic allele and π_1 known.

6.4.2 Proband's phenotype Exclusion Likelihood

The results of the method corrected with PEL are presented in Figure 6.5 for the common pathogenic allele and in Figure 6.6 for the rare pathogenic allele.

Having the same assumption on π_1, the PEL greatly correct the bias due to genetic testing as, again, all the scenarios showcase very similar results. However, contrary to the overestimation of RH_1 seen on the previous method, there is here a slight under-estimation of RH_1 both in the common and rare pathogenic allele cases. RH_2 seems well estimated but that was already the case without correction.

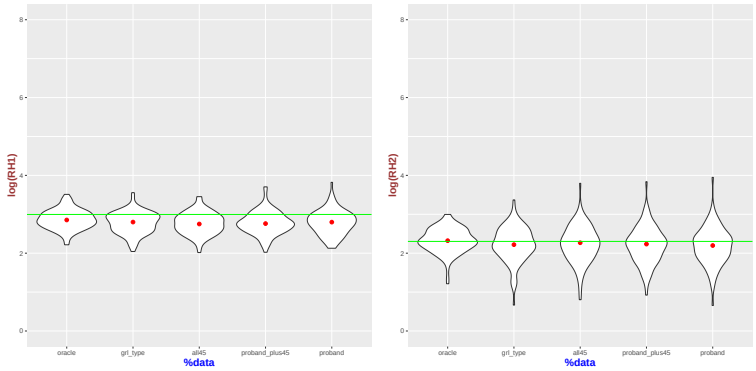


FIGURE 6.5 : Estimation of parameters $\log(RH_1)$ and $\log(RH_2)$ on datasets with common pathogenic allele with PEL correction.

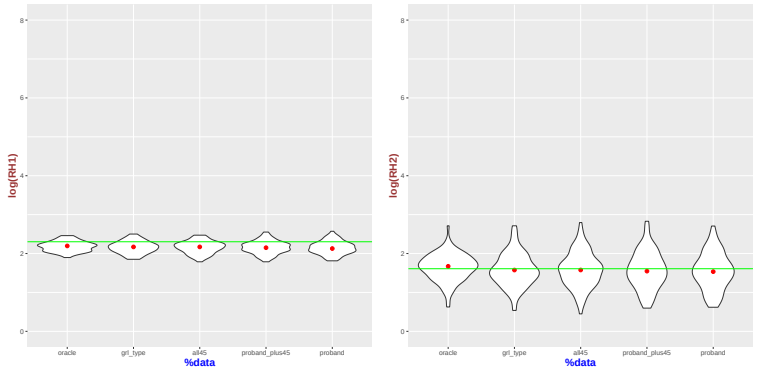


FIGURE 6.6 : Estimation of parameters $\log(RH_1)$ and $\log(RH_2)$ on datasets with rare pathogenic allele with PEL correction.

6.4.3 Genotype-Restricted Likelihood

The results of the method corrected with GRL are presented in Figure 6.7 for the common pathogenic allele and in Figure 6.8 for the rare pathogenic allele.

The GRL correction with our method does not perform well in all the scenarios considered underestimating both RH_1 and RH_2.

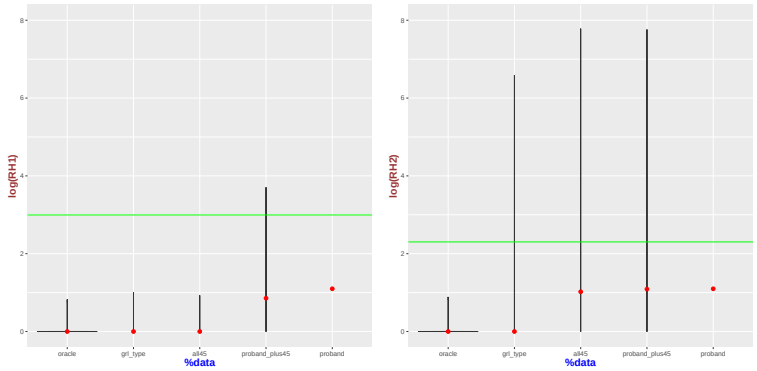


FIGURE 6.7 : Estimation of parameters $\log(RH_1)$ and $\log(RH_2)$ on datasets with common pathogenic allele with GRL correction.

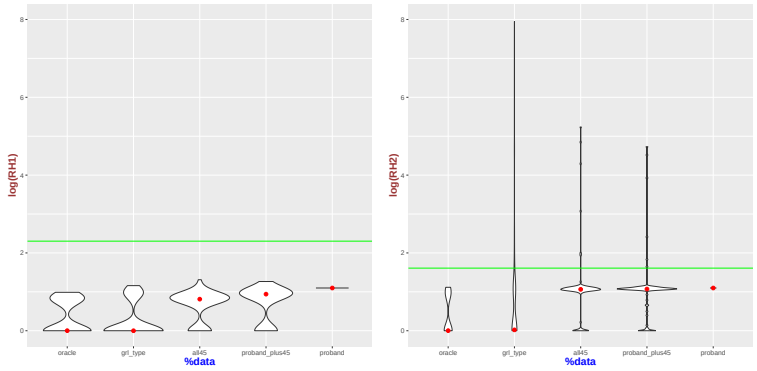


FIGURE 6.8 : Estimation of parameters $\log(RH_1)$ and $\log(RH_2)$ on datasets with rare pathogenic allele with GRL correction.

6.5 Discussion

Interestingly, setting π_1 as a known parameter allows the method to perform somewhat well wihtout any correction for the estimation of RH_1 and RH_2. It also seems to automatically correct the bias induced by the genetic testing. Because of the rules that select families showcases multiples cases of the disease before 45 years old, RH_1 is over-estimated by the method. RH_2 is estimated without bias. However, if the selection criteria was selecting families with numerous cases after 50 years old, it is possible that RH_2 estimations would be biased also.

While applying the PEL correction with π_1 known, the same results toward bias induced by the genetic testing is observed. Furthermore, the estimation of RH_1 seems to be less biased with the PEL correction. However, even if the estimation are better, the PEL correction seems to lead to under-estimation of RH_1. Finally, RH_2 is estimated without bias but, it would interesting to test it on other sets of ascertainment including a bias on RH_2 to see if the PEL correction still work like it does for RH_1. The results of the PEL correction are in accordance with the advertised results of the method.

The GRL correction does not perform well with our method in all the scenarios considered even on the GRL-like genetic testing scenario which follows the assumption made on the GRL article. Originally, the GRL was tested on generated data with the penetrance following a Weibull function which is quite far from the scenarios we are investigating in this article. This might explain the poor performance of this correction with our method.

6.6 Conclusion and perspectives

The main objective of this article is to present the bias induced by the ascertainment of the patients in genetic studies and to demonstrate it on our penetrance estimation method. To do so, datasets representing multiple scenarios were generated and the method applied on them. The scenarios include either a rare or common pathogenic allele, and multiple rules of genetic testing in the families (*i.e.* only the proband's genotype is known, the proband and its parents are genetically tested, etc).

Once the bias was pointed out, three correction methods were applied with our method in order to analyse which one was performing the best on the different scenarios. The first method was simply to set the proportion of pathogenic variant carriers (π_1) as known, which is a very standard assumption in genetic studies estimating penetrance. The second one is the Proband's phenotype Exclusion Likelihood (PEL) which assumes that π_1 is known and forget the phenotype of probands. The last one is the Genotyped-Restricted Likelihood (GRL) which also assumes that π_1 is known and performs the method by conditionning the model on probands' genotypes.

While setting π_1 known corrects the bias due to genetic testing, RH_1 remains over-estimated because of the ascertainment, it would be probably similar with RH_2 if the ascertainment rules were to include more severity after 50. The PEL method seems to correct quite well the bias induced by the ascertainment, which is very interesting. Finally GRL seems not fitted to our method.

This work proposes many perspectives. First of all, the fact that no correction method is designed to be used when π_1 is unknown (as shown in the appendice) shows that there is a need for such correction method. The fact that π_1 is a parameter of the model is a major component of our model that distinguishes it from previous methods. It would be very interesting to develop such correction method and we are currently working on methods such as raking to correct the bias with weighted likelihood. Secondly, the reality of ascertainment is that, most of the time, data are collected through different sets of rules, even in the same

dataset. Similarly, genetic testing might not be done following strict rules as we modeled in our scenarios. One perspective of this work would be to continue testing on simulations selected through a mix of ascertainments (selection and genetic testing) to analyse the robustness of the correction methods.

Finally the main purpose of this method is to be used on the real data we currently have and which are families with *SFTPA1* and *SFTPA2* pathogenic variants carriers. The next part of this work would be to estimate the penetrance of lung cancer for these carriers.

Appendices to Chapter 6

6.A Implementation of correction methods with unknown allele frequency

In this section, the results of the two ascertainment corrections are presented when all the parameters π_1, RH_1 and RH_2 are unknown. These methods are not designed to be used when π_1 is unknown but it is interesting to see how they perform.

6.A.1 Proband's phenotype Exclusion Likelihood

The results of the PEL are presented in Figure 6.9 for the common allele scenario and 6.10 for the rare allele scenario.

It is clear that π_1 is completly over-estimated in all scenarios and genetic testings. However, RH_2 seems to be estimated without bias, but as noted previously, it might be because the ascertainment isn't based on old cases of the disease (after 50). RH_1 estimations are clearly biased in two scenarios, the GRL-type genetic testing (where the proband and its parents are tested) and when only the proband is tested. The other scenarios seem to present small estimation bias.

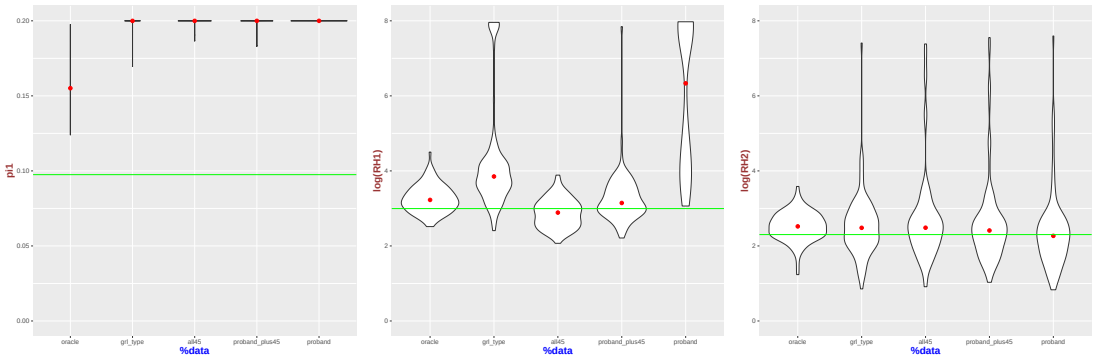


FIGURE 6.9 : Estimation of parameters $\log(RH_1)$ and $\log(RH_2)$ on datasets with common pathogenic allele with PEL correction and π_1 unknown.

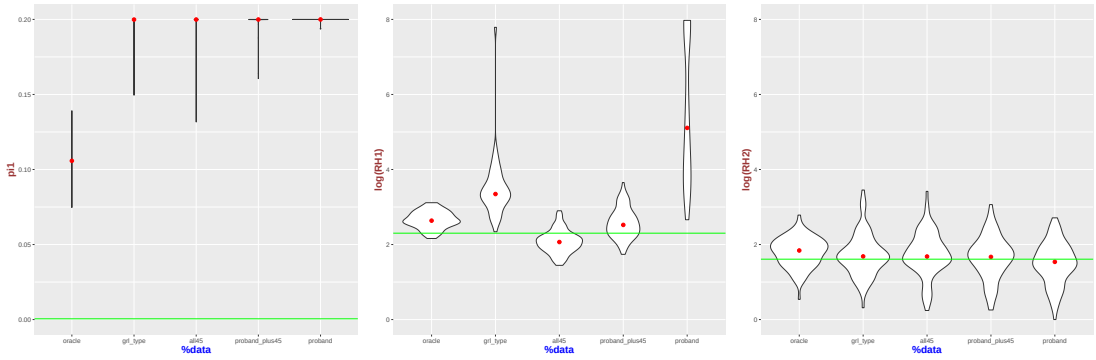


FIGURE 6.10 : Estimation of parameters $\log(RH_1)$ and $\log(RH_2)$ on datasets with rare pathogenic allele with PEL correction and π_1 unknown.

6.A.2 Genotype-Restricted Likelihood

The results of the GR are presented in Figure 6.11 for the common allele scenario and 6.12 for the rare allele scenario.

The GRL correction seems not to be efficient with our method as no parameter is correctly estimated, estimations showcasing no clear convergence to a specific value for many scenarios.

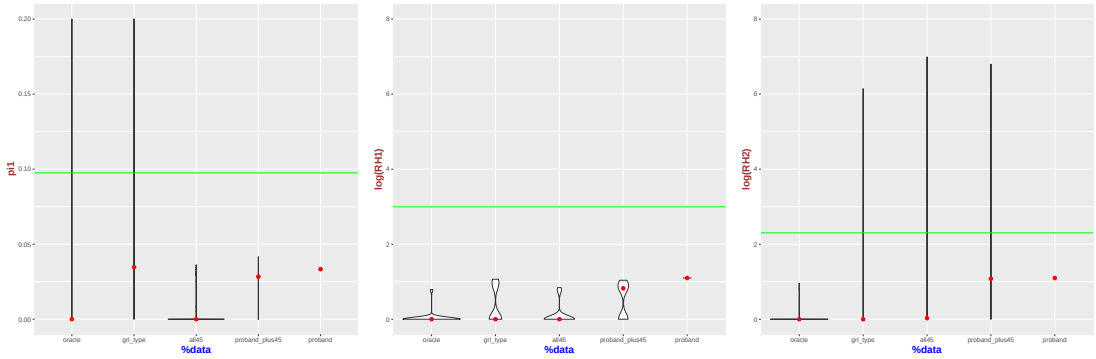


FIGURE 6.11 : Estimation of parameters $\log(RH_1)$ and $\log(RH_2)$ on datasets with common pathogenic allele with GRL correction and π_1 unknown.

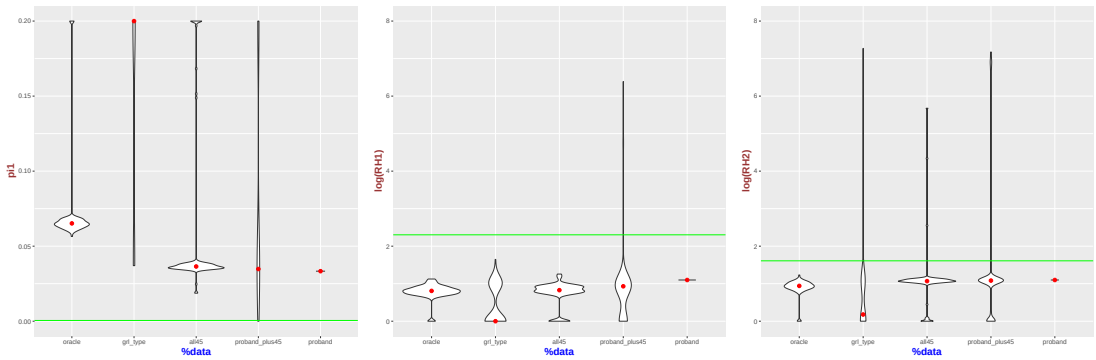


FIGURE 6.12 : Estimation of parameters $\log(RH_1)$ and $\log(RH_2)$ on datasets with rare pathogenic allele with GRL correction and π_1 unknown.

Raking correction with general population's incidence constrains

6.B Application to lung cancer for *SFTPA1* and *SFTPA2* pathogenic variants

This section showcases the results of the developed method combined with PEL correction on the dataset of used in Chapter 4. The aim is to estimate the survival function of *SFTPA1* and *SFTPA2* pathogenic variants carriers to lung cancer.

6.B.1 Material and methods

The set of data is from an AH-HP dataset of families with *SFTPA1* and *SFTPA2* pathogenic variants carriers. Based on extended pedigrees gathering 744 individuals (over 27 families), among whom 59 carriers, phenotypic data were retrieved from 328 individuals.

The French general population incidence of lung cancer is taken from Globocan in 2020 [66]. The cuts are every 5 years up to 70 years old. The bin-specific 5-year incidences are shown in Table 6.1. The general population incidence and survival are shown in Figure 6.13.

Age (year)	0-4	5-9	10-14	15-19	20-24	25-29	30-34	35-39	40-44	45-49	50-54	55-59	60-64	65-69	70+
Incidence	0.08	0.08	0.03	0.1	0.38	0.46	1.2	3.3	11.5	33.6	73.9	128.8	182.8	219.8	223.7

TABLE 6.1 : Incidence of lung cancer in France in 2020, for 100000 individuals by age-bin.

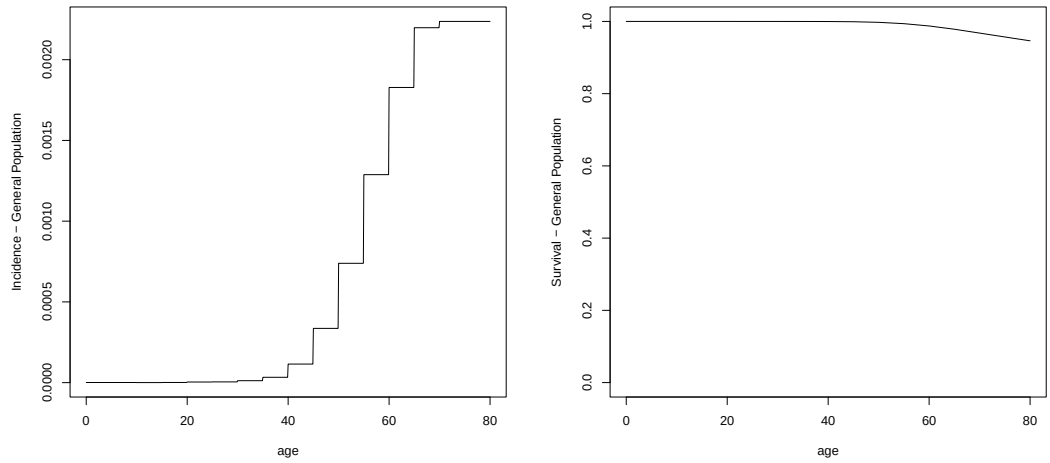


FIGURE 6.13 : Incidence and survival function of lung cancer for the French population in 2020 according to Globocan [66].

We consider the model of the article for carriers and non-carriers of *SFTPA1* and *SFTPA2* pathogenic variants. We assume the allelic frequency of *SFTPA1* and *SFTPA2* pathogenic variants to be $f = 0.0005$ as chosen in Chapter 4. We consider $\mathrm{RH}(t)$ to have two values $(\mathrm{RH}_1, \mathrm{RH}_2)$ with a cut at 50 years old.

6.B.2 Results

The estimated parameters $(\mathrm{RH}_1, \mathrm{RH}_2)$ and the 95% confidence intervals are reported in Table 6.2. The incidences and survival curves of lung cancer for *SFTPA1* and *SFTPA2* pathogenic variants carriers and non-carriers are presented in Figure 6.14.

	Value	Upper	Lower
RH_1	33.71	122.17	9.83
RH_2	16.69	35.44	8.15

TABLE 6.2 : Estimated values of the parameters and their upper and lower bounds for 95% confidence intervals.

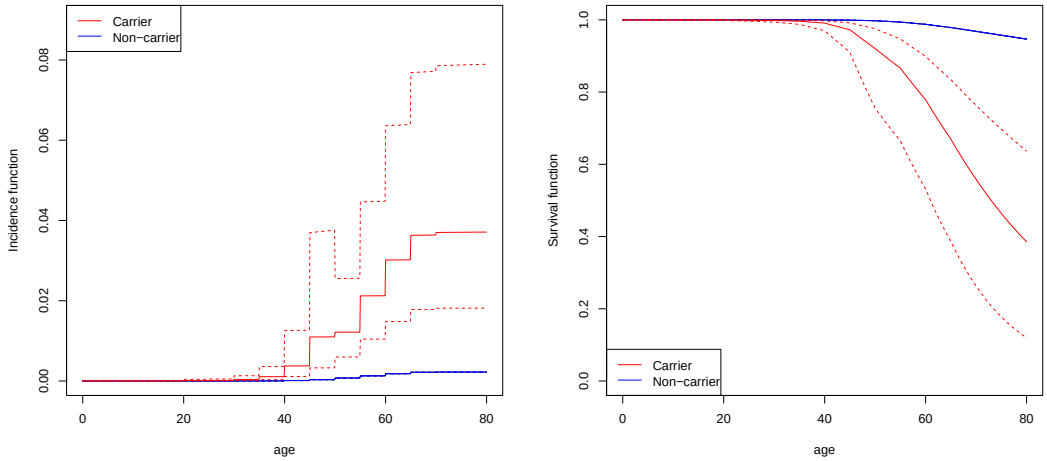


FIGURE 6.14 : Incidence and survival function of lung cancer *SFTPA1* and *SFTPA2* pathogenic variants carriers and non-carriers.

Chapitre 7

Conclusion et perspectives

7.1 Conclusion

Pour rappel, cette thèse traitent des problématiques suivantes, fondamentales en consultation génétique :

- Prédiction de variants pathogènes prédisposant à des maladies génétiques pour les patients vus en consultation de génétique.
- Évaluation du risque de survenue des maladies associées à ces variants pathogènes pour un patient porteur et les membres de sa famille.

Plus précisément, elle étudie le cas particulier des maladies à forte composante monogénique, survenant au cours de la vie (*i.e.* dont la pénétrance est dépendante de l'âge). Le cancer broncho-pulmonaire est un exemple d'une telle maladie pour laquelle de plus en plus de mutations prédisposantes sont découvertes (par exemple les variants pathogènes *SFTPA1* et *SFTPA2* ou des variants sur les gènes *TP53* et *EGFR*). Les cancers du sein et l'ovaire sont également des exemples mieux connus pour lesquels un grand nombre de variants pathogènes sont étudiés depuis plus de trente ans (*BRCA1*, *BRCA2* étant les gènes les plus connus). C'est donc dans ce cadre que plusieurs projets ont émergé.

Cette thèse est à l'interface entre les statistiques computationnelles et les applications médicales. Elle propose à la fois, des résultats cliniques et épidémiologiques dans les chapitres 3 et 4 et des résultats méthodologiques, dans les chapitres 5, et 6.

7.1.1 Conclusions cliniques et épidémiologiques

Le chapitre 3 souligne les défis persistants liés à la sélection de patients pour les dépistages génétiques dans le cadre des cancers du sein et de l'ovaire. Les modèles actuels, tels que MSS-F et CanRisk, présentent des performances faibles pour la prédiction de variants pathogènes lorsque ceux-ci sont utilisés en consultation de génétique. La correction de l'*ascertainment* par la méthode du raking montre que le MSS-F semble plus performant pour détecter les porteurs de mutations dans une population non-sélectionnée (permettant d'envoyer vers le conseil en génétique des personnes ayant une probabilité supérieure à 10% d'être porteurs). Une fois en consultation, les outils développés ne sont pas suffisamment performants pour distinguer les porteurs et les non-porteurs. Cela pose la question d'un

test génétique automatique à ce stade du parcours de soin, car l'identification des porteurs de variants pathogènes a une importance cruciale en raison des traitements actuellement disponibles.

Le chapitre 4 se consacre à l'estimation de la pénétrance de la pneumopathie interstitielle et du cancer broncho-pulmonaire pour les porteurs de variants pathogènes *SFTPA1* et *SFTPA2*. Cette estimation est une première pour ces variants et ces maladies dont le lien est connu mais pas encore suffisamment étudié. L'article met en évidence une pénétrance élévée de la pneumopathie interstitielle pour les porteurs. La pénétrance du cancer broncho-pulmonaire est moindre que la pneumopathie interstitielle mais reste très importante par rapport à la pénétrance de ce cancer en population générale.

7.1.2 Conclusion méthodologique

Le chapitre 5 propose donc une nouvelle méthode d'estimation de la pénétrance/survie pour des maladies génétiques présentant des cas sporadiques à partir de données de pédigrée, éliminant certaines hypothèses simplificatrices des méthodes existantes. En effet, ces méthodes sont basées sur l'hypothèse que la proportion de porteurs de variants pathogènes est connue et approximent la pénétrance des non-porteurs par la pénétrance de la population générale. Notre approche généralise un peu plus le modèle en intégrant la proportion de porteurs dans la population en tant que paramètre, ouvrant la voie à des applications plus flexibles. Elle se base, ensuite, sur une paramétrisation du ratio d'incidences entre les porteurs et les non-porteurs et utilise une méthode de point-fixe pour calculer ces incidences sous la contrainte d'incidence en population générale. Les paramètres du modèle sont estimés par une maximisation de la vraisemblance via une méthode BFGS.

Le chapitre 6 met en évidence les biais, introduits par la sélection des patients dans les études génétiques (*ascertainment*), sur les résultats de la méthode développée au chapitre 5. Les méthodes de correction étudiées dans ce chapitre ont été développées selon l'hypothèse d'une proportion de porteurs de variants pathogènes connue en population générale. Ainsi, nous avons testé notre méthode en considérant cette hypothèse uniquement, puis en appliquant respectivement deux méthodes de correction (la *PEL* et la *GRL*). Si la *GRL* semble ne pas fonctionner du tout avec notre méthode, la *PEL*, elle, montre des résultats très encourageants pour débiaiser les estimations.

7.2 Perspectives

L'ensemble de ces travaux ouvre un certain nombre de perspectives intéressantes pour la suite.

Dans l'étude des variants pathogènes *SFTPA1* et *SFTPA2*, la méthode d'estimation repose sur l'hypothèse que le cancer broncho-pulmonaire et la pneumopathie interstitielle ne présentent pas de cas sporadiques, ce qui est faux en réalité mais utile en première approximation. Il serait donc intéressant d'appliquer notre méthode développée au chapitre 5 à ces données en utilisant la PEL comme correction d'*ascertainment*. De plus les résultats actuels montrent une différence non-significative entre les pénétrances des hommes et des femmes, ainsi que des intervalles de confiance relativement importants ; il est donc important de continuer la collecte de données pour améliorer les résultats de l'étude. Concernant le modèle en lui-même, l'hypothèse d'indépendance des phénotypes conditionnellement aux génotypes présente des limites, notamment, dans les études familiales où il peut y avoir une composante de fragilité à prendre en compte ou lorsque la maladie étudiée est connue pour être associée à un facteur d'exposition impactant fortement le risque. C'est notamment le cas du cancer broncho-pulmonaire pour lequel de nombreuses expositions environnementales

accroissent le risque de survenue (tabagisme, pollution, radon, etc). Les données collectées ne permettent pas pour le moment d'inclure une variable d'exposition mais, la collecte de données se poursuit et s'attache à intégrer ce type de données. Une étude prospective (RaDiCo-ILD2) est actuellement en cours, ce qui devrait permettre de préciser nos résultats.

Le développement de notre méthode d'estimation de pénétrance pour des maladies génétiques présentant des cas sporadiques, qui introduit la proportion de porteurs de variants pathogènes comme paramètre du modèle, ouvre plusieurs perspectives. En effet, la combinaison de notre méthode et de la *PEL* semble prometteuse bien que toujours limitée par l'hypothèse d'une proportion de porteurs connue. Il serait donc intéressant, d'un point de vue médical, de l'utiliser pour l'estimation de la pénétrance du cancer broncho-pulmonaire pour d'autres variants pathogènes connus, notamment pour les gènes *EGFR*, dont une base de données est en cours de constitution à l'APHP, et *TP53* (syndrôme de Li-Fraumeni) dont une base de données est existante au CHU de Rouen.

Concernant les perspectives méthodologiques, il y en a plusieurs. La première perspective envisageable sur ce modèle est de relaxer la contrainte "constant par morceaux" en étudiant des formes paramétriques pour le ratio de risques instantanés, par exemple en le modélisant par une fonction de Weibull. De plus, comme pour le chapitre 4, la méthode s'appuie sur des données familiales, inclure dans le modèle une potentielle fragilité ou des variables d'exposition environnementale serait judicieux. Il serait également intéressant d'évaluer la robustesse de notre méthode combinée à la *PEL* sur des *ascertainments* plus complexes, notamment des mixtes de critères, ce qui se rapprocherait plus de ce qui est fait en pratique clinique. Enfin, la perspective la plus ambitieuse sur ce travail méthodologique serait probablement le développement d'une méthode de correction du biais de sélection qui lui serait spécifiquement adaptée. En effet, le fait qu'aucune méthode de correction ne soit conçue pour être utilisée lorsque la proportion de porteurs de variants pathogènes est inconnu montre qu'il existe un besoin réel. L'introduction de cette proportion comme paramètre du modèle est une composante majeure qui distingue notre méthode d'estimation des méthodes précédentes. Nous travaillons actuellement au développement d'une telle méthode qui s'appuierait sur du raking (méthode employée dans le chapitre 3) pour pondérer la vraisemblance du modèle.

Bibliographie

[1] F. Alarcon et al. “PEL : an unbiased method for estimating age-dependent genetic disease risk from pedigree data unselected for family history”. en. In : *Genetic Epidemiology* 33.5 (juill. 2009), p. 379-385. issn : 07410395, 10982272. doi : 10.1002/gepi.20390. url : https://onlinelibrary.wiley.com/doi/10.1002/gepi.20390 (visité le 03/05/2023).

[2] Flora Alarcon et al. “Non-parametric estimation of survival in age-dependent genetic disease and application to the transthyretin-related hereditary amyloidosis”. en. In : *PLOS ONE* 13.9 (sept. 2018). Sous la dir. de Wei Wang, e0203860. issn : 1932-6203. doi : 10.1371/journal.pone.0203860. url : https://dx.plos.org/10.1371/journal.pone.0203860 (visité le 14/03/2023).

[3] Flora Alarcon et al. “Non-parametric estimation of survival in age-dependent genetic disease and application to the transthyretin-related hereditary amyloidosis”. en. In : *PLOS ONE* 13.9 (sept. 2018). Sous la dir. de Wei Wang, e0203860. issn : 1932-6203. doi : 10.1371/journal.pone.0203860. url : https://dx.plos.org/10.1371/journal.pone.0203860 (visité le 13/06/2023).

[4] M. Anheim et al. “Penetrance of Parkinson disease in glucocerebrosidase gene mutation carriers”. en. In : *Neurology* 78.6 (fév. 2012), p. 417-420. issn : 0028-3878, 1526-632X. doi : 10.1212/WNL.0b013e318245f476. url : https://www.neurology.org/lookup/doi/10.1212/WNL.0b013e318245f476 (visité le 13/06/2023).

[5] A C Antoniou et al. “The BOADICEA model of genetic susceptibility to breast and ovarian cancer”. en. In : *British Journal of Cancer* 91.8 (oct. 2004), p. 1580-1590. issn : 0007-0920, 1532-1827. doi : 10.1038/sj.bjc.6602175. url : https://www.nature.com/articles/6602175 (visité le 31/10/2023).

[6] Stephanie Archer et al. “Evaluating clinician acceptability of the prototype CanRisk tool for predicting risk of breast and ovarian cancer : A multi-methods study”. en. In : *PLOS ONE* 15.3 (mars 2020). Sous la dir. d’Alvaro Galli, e0229999. issn : 1932-6203. doi : 10.1371/journal.pone.0229999. url : https://dx.plos.org/10.1371/journal.pone.0229999 (visité le 04/11/2023).

[7] Tesfaye M. Baye, Lisa J. Martin et Gurjit K. Khurana Hershey. “Application of genetic/genomic approaches to allergic disorders”. en. In : *Journal of Allergy and Clinical Immunology* 126.3 (sept. 2010), p. 425-436. issn : 00916749. doi : 10.1016/j.jaci.2010.05.025. url : https://linkinghub.elsevier.com/retrieve/pii/S009167491000850X (visité le 28/10/2023).

[8] Patrick R. Benusiglio et al. “Utility of a mainstreamed genetic testing pathway in breast and ovarian cancer patients during the COVID-19 pandemic”. eng. In : *European Journal of Medical Genetics* 63.12 (déc. 2020), p. 104098. issn : 1878-0849. doi : 10.1016/j.ejmg.2020.104098.

[9] Donald A. Berry et al. “BRCAPRO Validation, Sensitivity of Genetic Testing of *BRCA1/BRCA2* , and Prevalence of Other Breast Cancer Susceptibility Genes”. en. In : *Journal of Clinical Oncology* 20.11 (juin 2002), p. 2701-2712. issn : 0732-183X, 1527-7755. doi : 10.1200/JCO.2002.05.121. url : https://ascopubs.org/doi/10.1200/JCO.2002.05.121 (visité le 31/10/2023).

[10] Bernard Bonaïti et al. “Estimating penetrance from multiple case families with predisposing mutations : extension of the ‘genotype-restricted likelihood’ (GRL) method”. en. In : *European Journal of Human Genetics* 19.2 (fév. 2011), p. 173-179. issn : 1018-4813, 1476-5438. doi : 10.1038/ejhg.2010.158. url : https://www.nature.com/articles/ejhg2010158 (visité le 04/11/2023).

[11] Tim Carver et al. “CanRisk Tool—A Web Interface for the Prediction of Breast and Ovarian Cancer Risk and the Likelihood of Carrying Genetic Pathogenic Variants”. en. In : *Cancer Epidemiology, Biomarkers & Prevention* 30.3 (mars 2021), p. 469-473. issn : 1055-9965, 1538-7755. doi : 10.1158/1055-9965.EPI-20-1319. url : https://aacrjournals.org/cebp/article/30/3/469/264109/CanRisk-Tool-A-Web-Interface-for-the-Prediction-of (visité le 04/11/2023).

[12] Sining Chen et al. “BayesMendel : an R Environment for Mendelian Risk Prediction”. In : *Statistical Applications in Genetics and Molecular Biology* 3.1 (jan. 2004), p. 1-19. issn : 1544-6115. doi : 10.2202/1544-6115.1063. url : https://www.degruyter.com/document/doi/10.2202/1544-6115.1063/html (visité le 04/11/2023).

[13] Sining Chen et al. “Prediction of Germline Mutations and Cancer Risk in the Lynch Syndrome”. en. In : *JAMA* 296.12 (sept. 2006), p. 1479. issn : 0098-7484. doi : 10.1001/jama.296.12.1479. url : http://jama.jamanetwork.com/article.aspx?doi=10.1001/jama.296.12.1479 (visité le 31/10/2023).

[14] Shing Wan Choi, Timothy Shin-Heng Mak et Paul F. O’Reilly. “Tutorial : a guide to performing polygenic risk score analyses”. en. In : *Nature Protocols* 15.9 (sept. 2020), p. 2759-2772. issn : 1754-2189, 1750-2799. doi : 10.1038/s41596-020-0353-1. url : https://www.nature.com/articles/s41596-020-0353-1 (visité le 30/10/2023).

[15] E. B. Claus, N. Risch et W. D. Thompson. “Genetic analysis of breast cancer in the cancer and steroid hormone study”. eng. In : *American Journal of Human Genetics* 48.2 (fév. 1991), p. 232-242. issn : 0002-9297.

[16] Julie O. Culver et al. “Integration of Universal Germline Genetic Testing for All New Breast Cancer Patients”. eng. In : *Annals of Surgical Oncology* 30.2 (fév. 2023), p. 1017-1025. issn : 1534-4681. doi : 10.1245/s10434-022-12595-w.

[17] A. P. Dempster, N. M. Laird et D. B. Rubin. “Maximum Likelihood from Incomplete Data via the EM Algorithm”. In : *Journal of the Royal Statistical Society. Series B (Methodological)* 39.1 (1977), p. 1-38. issn : 0035-9246. url : https://www.jstor.org/stable/2984875 (visité le 14/03/2023).

[18] Evans Dg et al. “A new scoring system for the chances of identifying a BRCA1/2 mutation outperforms existing models including BRCAPRO”. en. In : *Journal of medical genetics* 41.6 (juin 2004). Publisher : J Med Genet. issn : 1468-6244. doi : 10.1136/jmg.2003.017996. url : https://pubmed.ncbi.nlm.nih.gov/15173236/ (visité le 21/11/2023).

[19] Laure Dossus et Patrick R. Benusiglio. “Lobular breast cancer : incidence and genetic and non-genetic risk factors”. eng. In : *Breast cancer research : BCR* 17 (mars 2015), p. 37. issn : 1465-542X. doi : 10.1186/s13058-015-0546-7.

[20] D. F. Easton, D. Ford et D. T. Bishop. “Breast and ovarian cancer incidence in BRCA1-mutation carriers. Breast Cancer Linkage Consortium”. eng. In : *American Journal of Human Genetics* 56.1 (jan. 1995), p. 265-271. issn : 0002-9297.

[21] D. F. Easton et al. “Genetic linkage analysis in familial breast and ovarian cancer : results from 214 families. The Breast Cancer Linkage Consortium”. eng. In : *American Journal of Human Genetics* 52.4 (avr. 1993), p. 678-701. issn : 0002-9297.

[22] R.C. Elston et J. Stewart. “A General Model for the Genetic Analysis of Pedigree Data”. en. In : *Human Heredity* 21.6 (1971), p. 523-542. issn : 0001-5652, 1423-0062. doi : 10.1159/000152448. url : https://www.karger.com/Article/FullText/152448 (visité le 14/03/2023).

[23] Robert C. Elston, Varghese T. George et Forrestt Severtson. “The Eiston-Stewart Algorithm for Continuous Genotypes and Environmental Factors”. en. In : *Human Heredity* 42.1 (1992), p. 16-27. issn : 1423-0062, 0001-5652. doi : 10.1159/000154043. url : https://www.karger.com/Article/FullText/154043 (visité le 14/03/2023).

[24] D. Gareth Evans et al. “Detection of pathogenic variants in breast cancer susceptibility genes in bilateral breast cancer”. eng. In : *Journal of Medical Genetics* 60.10 (oct. 2023), p. 974-979. issn : 1468-6244. doi : 10.1136/jmg-2023-109196.

[25] D Gareth Evans et al. “Pathology update to the Manchester Scoring System based on testing in over 4000 families”. en. In : *Journal of Medical Genetics* 54.10 (oct. 2017), p. 674-681. issn : 0022-2593, 1468-6244. doi : 10.1136/jmedgenet-2017-104584. url : https://jmg.bmj.com/lookup/doi/10.1136/jmedgenet-2017-104584 (visité le 30/10/2023).

[26] *First estimates of diffuse gastric cancer risks for carriers of CTNNA1 germline pathogenic variants | Journal of Medical Genetics*. url : https://jmg.bmj.com/content/59/12/1189 (visité le 28/12/2023).

[27] "First results from the International Breast Cancer Intervention Study (IBIS-I) : a randomised prevention trial". en. In : *The Lancet* 360.9336 (sept. 2002), p. 817-824. ISSN : 01406736. DOI : 10.1016/S0140-6736(02)09962-2. URL : https://linkinghub.elsevier.com/retrieve/pii/S0140673602099622 (visité le 31/10/2023).

[28] D. Ford et al. "Genetic Heterogeneity and Penetrance Analysis of the BRCA1 and BRCA2 Genes in Breast Cancer Families". en. In : *The American Journal of Human Genetics* 62.3 (mars 1998), p. 676-689. ISSN : 00029297. DOI : 10.1086/301749. URL : https://linkinghub.elsevier.com/retrieve/pii/S0002929707638488 (visité le 04/11/2023).

[29] Anthony J. F. Griffiths, éd. *An introduction to genetic analysis*. eng. 7. ed., 3. print. New York, NY : Freeman, 2002. ISBN : 978-0-7167-3527-4 978-0-7167-3520-5 978-0-7167-3771-1.

[30] Nancy Hamel, Marc Tischkowitz et William D. Foulkes. "PALB2/FANCN, acteur dans la prédisposition au cancer du sein ?" In : *médecine/sciences* 24.2 (fév. 2008), p. 120-121. ISSN : 0767-0974, 1958-5381. DOI : 10.1051/medsci/2008242120. URL : http://www.medecinesciences.org/10.1051/medsci/2008242120 (visité le 28/10/2023).

[31] Heather Hampel et Matthew B. Yurgelun. "Point/Counterpoint : Is It Time for Universal Germline Genetic Testing for All GI Cancers ?" eng. In : *Journal of Clinical Oncology : Official Journal of the American Society of Clinical Oncology* 40.24 (août 2022), p. 2681-2692. ISSN : 1527-7755. DOI : 10.1200/JCO.21.02764.

[32] Jan Henkel et al. "Diagnostic yield and clinical relevance of expanded germline genetic testing for nearly 7000 suspected HBOC patients". en. In : *European Journal of Human Genetics* 31.8 (août 2023). Number : 8 Publisher : Nature Publishing Group, p. 925-930. ISSN : 1476-5438. DOI : 10.1038/s41431-023-01380-2. URL : https://www.nature.com/articles/s41431-023-01380-2 (visité le 24/11/2023).

[33] Theodore Huang et al. "Practical implementation of frailty models in Mendelian risk prediction". en. In : *Genetic Epidemiology* 44.6 (sept. 2020), p. 564-578. ISSN : 0741-0395, 1098-2272. DOI : 10.1002/gepi.22323. URL : https://onlinelibrary.wiley.com/doi/10.1002/gepi.22323 (visité le 30/10/2023).

[34] Graham Kalton et Ismael Flores-Cervantes. "Weighting methods". en. In : *Journal of Official Statistics* 19 (2003), p. 81-97.

[35] David G. Kleinbaum et Mitchel Klein. *Survival Analysis : A Self-Learning Text*. en. Sous la dir. de M. Gail et al. Statistics for Biology and Health. New York, NY : Springer New York, 2005. ISBN : 978-0-387-23918-7 978-0-387-29150-5. DOI : 10.1007/0-387-29150-4. URL : http://link.springer.com/10.1007/0-387-29150-4 (visité le 28/10/2023).

[36] Allison W. Kurian et al. "Association of Family Cancer History With Pathogenic Variants in Specific Breast Cancer Susceptibility Genes". eng. In : *JCO precision oncology* 5 (2021), PO.21.00261. ISSN : 2473-4284. DOI : 10.1200/PO.21.00261.

[37] Morta Lapkus et al. "Exploring Breast Surgeons' Attitudes on Universal Genetic Testing : A Qualitative Study". eng. In : *Annals of Surgical Oncology* 30.10 (oct. 2023), p. 6108-6116. ISSN : 1534-4681. DOI : 10.1245/s10434-023-13895-5.

[38] S. L. Lauritzen et D. J. Spiegelhalter. "Local Computations with Probabilities on Graphical Structures and Their Application to Expert Systems". In : *Journal of the Royal Statistical Society. Series B (Methodological)* 50.2 (1988). Publisher : [Royal Statistical Society, Wiley], p. 157-224. ISSN : 0035-9246. URL : https://www.jstor.org/stable/2345762 (visité le 20/11/2023).

[39] Andrew Lee et al. "BOADICEA : a comprehensive breast cancer risk prediction model incorporating genetic and nongenetic risk factors". en. In : *Genetics in Medicine* 21.8 (août 2019), p. 1708-1718. ISSN : 10983600. DOI : 10.1038/s41436-018-0406-9. URL : https://linkinghub.elsevier.com/retrieve/pii/S1098360021015963 (visité le 30/10/2023).

[40] Marie Legendre et al. "Functional assessment and phenotypic heterogeneity of *SFTPA1* and *SFTPA2* mutations in interstitial lung diseases and lung cancer". en. In : *European Respiratory Journal* 56.6 (déc. 2020), p. 2002806. ISSN : 0903-1936, 1399-3003. DOI : 10.1183/13993003.02806-2020. URL : http://erj.ersjournals.com/lookup/doi/10.1183/13993003.02806-2020 (visité le 13/06/2023).

[41] Lv Liu et al. "Identification of a Missense Mutation in the *Surfactant Protein A2* Gene in a Chinese Family with Interstitial Lung Disease". en. In : *DNA and Cell Biology* 40.1 (jan. 2021), p. 126-131. ISSN : 1044-5498, 1557-7430. DOI : 10.1089/dna.2020.6045. URL : https://www.liebertpub.com/doi/10.1089/dna.2020.6045 (visité le 13/06/2023).

[42] Teri A. Manolio. “Genomewide Association Studies and Assessment of the Risk of Disease”. en. In : *New England Journal of Medicine* 363.2 (juill. 2010). Sous la dir. de W. Gregory Feero et Alan E. Guttmacher, p. 166-176. issn : 0028-4793, 1533-4406. doi : 10.1056/NEJMra0905980. url : http://www.nejm.org/doi/10.1056/NEJMra0905980 (visité le 28/10/2023).

[43] Mark I. McCarthy et al. “Genome-wide association studies for complex traits : consensus, uncertainty and challenges”. en. In : *Nature Reviews Genetics* 9.5 (mai 2008), p. 356-369. issn : 1471-0056, 1471-0064. doi : 10.1038/nrg2344. url : https://www.nature.com/articles/nrg2344 (visité le 02/11/2023).

[44] Kelly A. Metcalfe et al. “Genetic testing women with newly diagnosed breast cancer : What criteria are the most predictive of a positive test ?” eng. In : *Cancer Medicine* 12.6 (mars 2023), p. 7580-7587. issn : 2045-7634. doi : 10.1002/cam4.5515.

[45] Jessica Moretta et al. “Recommandations françaises du Groupe Génétique et Cancer pour l’analyse en panel de gènes dans les prédispositions héréditaires au cancer du sein ou de l’ovaire”. In : *Bulletin du Cancer* 105.10 (oct. 2018), p. 907-917. issn : 0007-4551. doi : 10.1016/j.bulcan.2018.08.003. url : https://www.sciencedirect.com/science/article/pii/S0007455118302248 (visité le 21/11/2023).

[46] Nadia Nathan et al. “Germline *SFTPA1* mutation in familial idiopathic interstitial pneumonia and lung cancer”. en. In : *Human Molecular Genetics* 25.8 (avr. 2016), p. 1457-1467. issn : 0964-6906, 1460-2083. doi : 10.1093/hmg/ddw014. url : https://academic.oup.com/hmg/article-lookup/doi/10.1093/hmg/ddw014 (visité le 03/05/2023).

[47] Nadia Nathan et al. “SFTPA mutations in interstitial lung disease (ILD) and lung cancer”. In : *Diffuse Parenchymal Lung Disease*. European Respiratory Society, sept. 2017, PA1516. doi : 10.1183/1393003.congress-2017.PA1516. url : http://erj.ersjournals.com/lookup/doi/10.1183/1393003.congress-2017.PA1516 (visité le 13/06/2023).

[48] Jorge Nocedal et Stephen J. Wright. *Numerical optimization*. 2nd ed. Springer series in operations research. New York : Springer, 2006. isbn : 978-0-387-30303-1.

[49] Kate Packwood et al. “Breast cancer in patients with germline TP53 pathogenic variants have typical tumour characteristics : the Cohort study of TP53 carrier early onset breast cancer (COPE study)”. In : *The Journal of Pathology : Clinical Research* 5.3 (mai 2019), p. 189-198. issn : 2056-4538. doi : 10.1002/cjp2.133. url : https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6648388/ (visité le 24/11/2023).

[50] Giovanni Parmigiani, Donald A. Berry et Omar Aguilar. “Determining Carrier Probabilities for Breast Cancer–Susceptibility Genes BRCA1 and BRCA2”. en. In : *The American Journal of Human Genetics* 62.1 (jan. 1998), p. 145-158. issn : 00029297. doi : 10.1086/301670. url : https://linkinghub.elsevier.com/retrieve/pii/S0002929707601323 (visité le 04/11/2023).

[51] Erin M. Parry et al. “Germline Mutations in DNA Repair Genes in Lung Adenocarcinoma”. en. In : *Journal of Thoracic Oncology* 12.11 (nov. 2017), p. 1673-1678. issn : 15560864. doi : 10.1016/j.jtho.2017.08.011. url : https://linkinghub.elsevier.com/retrieve/pii/S155608641730686X (visité le 26/10/2023).

[52] Judea Pearl. “Fusion, propagation, and structuring in belief networks”. In : *Artificial Intelligence* 29.3 (sept. 1986), p. 241-288. issn : 0004-3702. doi : 10.1016/0004-3702(86)90072-X. url : https://www.sciencedirect.com/science/article/pii/000437028690072X (visité le 20/11/2023).

[53] Judea Pearl. *Probabilistic reasoning in intelligent systems : Networks of plausible inference*. Probabilistic reasoning in intelligent systems : Networks of plausible inference. Pages : xix, 552. San Mateo, CA, US : Morgan Kaufmann, 1988. isbn : 978-0-934613-73-6.

[54] Pascal Pujol et al. “Clinical practice guidelines for BRCA1 and BRCA2 genetic testing”. eng. In : *European Journal of Cancer (Oxford, England : 1990)* 146 (mars 2021), p. 30-47. issn : 1879-0852. doi : 10.1016/j.ejca.2020.12.023.

[55] Shoba Ranganathan et al., éd. *Encyclopedia of bioinformatics and computational biology*. eng. OCLC : 1052465484. Amsterdam, Netherlands : Elsevier, 2019. isbn : 978-0-12-811432-2.

[56] Mark Robson et al. “Olaparib for Metastatic Breast Cancer in Patients with a Germline BRCA Mutation”. eng. In : *The New England Journal of Medicine* 377.6 (août 2017), p. 523-533. issn : 1533-4406. doi : 10.1056/NEJMoa1706450.

[57] Tuija Schmidt et Prakash P. Shenoy. “Some improvements to the Shenoy-Shafer and Hugin architectures for computing marginals”. In : *Artificial Intelligence* 102.2 (juill. 1998), p. 323-333. issn : 0004-3702. doi : 10.1016/S0004-3702(98)00047-2. url : https://www.sciencedirect.com/science/article/pii/S0004370298000472 (visité le 20/11/2023).

[58] Catherine Schramm et al. *Penetrance estimation of SORL1 loss-of-function variants using a family-based strategy adjusted on APOE genotypes suggest a non-monogenic inheritance*. en. preprint. Genetics, juill. 2021. doi : 10.1101/2021.06.30.450554. url : http://biorxiv.org/lookup/doi/10.1101/2021.06.30.450554 (visité le 13/06/2023).

[59] C. Sessa et al. "Risk reduction and screening of cancer in hereditary breast-ovarian cancer syndromes : ESMO Clinical Practice Guideline". eng. In : *Annals of Oncology : Official Journal of the European Society for Medical Oncology* 34.1 (jan. 2023), p. 33-47. issn : 1569-8041. doi : 10.1016/j.annonc.2022.10.004.

[60] Glenn R. Shafer et Prakash P. Shenoy. "Probability propagation". en. In : *Annals of Mathematics and Artificial Intelligence* 2.1 (mars 1990), p. 327-351. issn : 1573-7470. doi : 10.1007/BF01531015. url : https://doi.org/10.1007/BF01531015 (visité le 20/11/2023).

[61] Montgomery Slatkin. "Linkage disequilibrium — understanding the evolutionary past and mapping the medical future". en. In : *Nature Reviews Genetics* 9.6 (juin 2008), p. 477-485. issn : 1471-0056, 1471-0064. doi : 10.1038/nrg2361. url : https://www.nature.com/articles/nrg2361 (visité le 28/10/2023).

[62] Honglin Song et al. "The contribution of deleterious germline mutations in BRCA1, BRCA2 and the mismatch repair genes to ovarian cancer in the population". eng. In : *Human Molecular Genetics* 23.17 (sept. 2014), p. 4703-4709. issn : 1460-2083. doi : 10.1093/hmg/ddu172.

[63] Steven Sorscher. "Ascertainment Bias and Estimating Penetrance". In : *JAMA Oncology* 4.4 (avr. 2018), p. 587-587. issn : 2374-2437. doi : 10.1001/jamaoncol.2017.4573. url : https://doi.org/10.1001/jamaoncol.2017.4573 (visité le 20/11/2023).

[64] D. Stoppa-Lyonnet et al. "BRCA1 sequence variations in 160 individuals referred to a breast/ovarian family cancer clinic. Institut Curie Breast Cancer Group". eng. In : *American Journal of Human Genetics* 60.5 (mai 1997), p. 1021-1030. issn : 0002-9297.

[65] Xiangqing Sun. "Segregation Analysis Using the Unified Model". en. In : *Statistical Human Genetics*. Sous la dir. de Robert C. Elston, Jaya M. Satagopan et Shuying Sun. T. 850. Series Title : Methods in Molecular Biology. Totowa, NJ : Humana Press, 2012, p. 211-235. isbn : 978-1-61779-554-1 978-1-61779-555-8. doi : 10.1007/978-1-61779-555-8_12. url : https://link.springer.com/10.1007/978-1-61779-555-8_12 (visité le 28/10/2023).

[66] Hyuna Sung et al. "Global Cancer Statistics 2020 : GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries". en. In : *CA : A Cancer Journal for Clinicians* 71.3 (mai 2021), p. 209-249. issn : 0007-9235, 1542-4863. doi : 10.3322/caac.21660. url : https://acsjournals.onlinelibrary.wiley.com/doi/10.3322/caac.21660 (visité le 28/10/2023).

[67] Kumpei Tanisawa, Masashi Tanaka et Mitsuru Higuchi. "Gene-exercise interactions in the development of cardiometabolic diseases". en. In : *The Journal of Physical Fitness and Sports Medicine* 5.1 (2016), p. 25-36. issn : 2186-8123, 2186-8131. doi : 10.7600/jpfsm.5.25. url : https://www.jstage.jst.go.jp/article/jpfsm/5/1/5_25/_article (visité le 02/11/2023).

[68] Terry M Therneau. *A Package for Survival Analysis in R*. 2023. url : https://CRAN.R-project.org/package=survival.

[69] Terry M. Therneau et Patricia M. Grambsch. *Modeling survival data : extending the Cox model*. eng. 2. print. Statistics for biology and health. New York Berlin Heidelberg : Springer, 2001. isbn : 978-0-387-98784-2.

[70] Liviu R Totir, Rohan L Fernando et Joseph Abraham. "An efficient algorithm to compute marginal posterior genotype probabilities for every member of a pedigree with loops". en. In : *Genetics Selection Evolution* 41.1 (déc. 2009), p. 52. issn : 1297-9686. doi : 10.1186/1297-9686-41-52. url : https://gsejournal.biomedcentral.com/articles/10.1186/1297-9686-41-52 (visité le 14/03/2023).

[71] Andrew N. J. Tutt et al. "Adjuvant Olaparib for Patients with BRCA1- or BRCA2-Mutated Breast Cancer". eng. In : *The New England Journal of Medicine* 384.25 (juin 2021), p. 2394-2405. issn : 1533-4406. doi : 10.1056/NEJMoa2105215.

[72] Catherine Uzan et al. "Consultation personnalisée d'évaluation du risque de cancer du sein : premiers résultats". In : *Bulletin du Cancer* 107.10 (oct. 2020), p. 972-981. issn : 0007-4551. doi : 10.1016/j.bulcan.2020.08.003. url : https://www.sciencedirect.com/science/article/pii/S0007455120303660 (visité le 24/11/2023).

[73] H Vasen et al. “New clinical criteria for hereditary nonpolyposis colorectal cancer (HNPCC, Lynch syndrome) proposed by the International Collaborative Group on HNPCC”. en. In : *Gastroenterology* 116.6 (juin 1999), p. 1453-1456. issn : 00165085. doi : 10.1016/S0016-5085(99)70510-X. url : https://linkinghub.elsevier.com/retrieve/pii/S001650859970510X (visité le 30/10/2023).

[74] Yongyu Wang et al. “Genetic Defects in Surfactant Protein A2 Are Associated with Pulmonary Fibrosis and Lung Cancer”. en. In : *The American Journal of Human Genetics* 84.1 (jan. 2009), p. 52-59. issn : 00029297. doi : 10.1016/j.ajhg.2008.11.010. url : https://linkinghub.elsevier.com/retrieve/pii/S0002929708005958 (visité le 13/06/2023).

[75] Pat W. Whitworth et al. “Clinical Utility of Universal Germline Genetic Testing for Patients With Breast Cancer”. eng. In : *JAMA network open* 5.9 (sept. 2022), e2232787. issn : 2574-3805. doi : 10.1001/jamanetworkopen.2022.32787.

[76] Siddhartha Yadav et al. “Germline Pathogenic Variants in Cancer Predisposition Genes Among Women With Invasive Lobular Carcinoma of the Breast”. eng. In : *Journal of Clinical Oncology : Official Journal of the American Society of Clinical Oncology* 39.35 (déc. 2021), p. 3918-3926. issn : 1527-7755. doi : 10.1200/JCO.21.00640.

[77] Matthew B. Yurgelun et al. “Development and Validation of the PREMMplus Model for Multigene Hereditary Cancer Risk Assessment”. eng. In : *Journal of Clinical Oncology : Official Journal of the American Society of Clinical Oncology* 40.35 (déc. 2022), p. 4083-4094. issn : 1527-7755. doi : 10.1200/JCO.22.00120.

Table des figures