Modélisation statistique de l'intensité des expositions prolongées en étiologie du cancer : application au tabac, à l'amiante, au cancer du poumon, et au mésothéliome pleural

Emilie Leveque

université de BORDEAUX

Année 2018

Thèse présentée pour l'obtention du grade de

DOCTEUR DE L'UNIVERSITÉ DE BORDEAUX

École doctorale Sociétés, Politique, Santé Publique
Spécialité Santé Publique, option Biostatistique

Soutenue publiquement le 7 décembre 2018

Par Emilie LÉVÊQUE

Modélisation statistique de l'intensité des expositions prolongées en étiologie du cancer : application au tabac, à l'amiante, au cancer du poumon, et au mésothéliome pleural.

Sous la direction de Karen Leffondré et Aude Lacourt

Membres du jury

M. Brochard Patrick	Pr, Université de Bordeaux	Examinateur
Mme Guihenneuc Chantal	Pr, Université Paris Descartes	Rapporteure
Mme Jacqmin-Gadda Hélène	DR, Inserm U1219, Bordeaux	Présidente
M. Slama Rémy	DR, Inserm U1209, Grenoble	Rapporteur
M. Wild Pascal	Chargé de mission, INRS, Nancy	Examinateur
Mme Lacourt Aude	CR, Inserm U1219, Bordeaux	Membre invité Co-encadrante
Mme Leffondré Karen	MCF, Université de Bordeaux	Directrice de thèse

Je dédie ce manuscrit à mon parrain,

Remerciements

A Karen et Aude,
Il va être difficile d'être concise dans ce paragraphe, j'espère que vous m'en excuserez par avance. 5 ans que l'on travaille ensemble, 5 ans que vous me faites confiance.. Tellement confiance que je me suis lancée dans cette expérience de doctorat qui ne me paraissait pas envisageable il y a plus de 3 ans, moi la petite ingénieure math/info ! Je ne vous remercierai jamais assez pour tout ce que vous avez pu m'apporter tant professionnellement que personnellement. Une petite pensée pour toutes ces réunions passées, qui ont pu s'éterniser plus que de raison ! Votre enthousiasme communicatif face à tout ce projet scientifique, a été le moteur de tout ce travail de thèse réalisé ! Votre bienveillance à mon égard a pu me donner confiance, et l'envie de toujours faire le maximum pour ne pas vous décevoir ! Et enfin, je voudrais sincèrement vous remercier pour votre patience, votre écoute, et vos conseils qui ont largement contribué à me rassurer dans les moments où j'en avais le plus besoin.

A Chantal Guihenneuc, Florence Ménégaux, Patrick Brochard, Rémy Slama, Pascal Wild,
Je tiens sincèrement à vous remercier d'avoir accepté de juger ce travail en qualité de rapporteur et d'examinateur. Je suis convaincue que vos connaissances et vos regards respectifs vont beaucoup m'apporter.

A Hélène Jacqmin-Gadda,
Je tiens à vous remercier d'avoir acceptée d'être la présidente de ce jury. Je souhaite également vous remercier pour tout ce que vous pouvez faire pour l'équipe Biostatistique car cela contribue fortement à ce que l'on s'y sente bien tant professionnellement qu'humainement.

A l'ensemble des équipes Biostat et SISTM,
Je remercie l'ensemble des membres des deux équipes à la fois pour votre accueil, votre bienveillance, et pour les échanges partagés au fil des années à travers des séminaires ou des pauses cafés. Je voudrais souligner que l'environnement de travail a été des plus appréciables. Je peux vous certifier que vous mettez la barre assez haute pour les futures équipes que j'aurais l'opportunité d'intégrer par la suite.

Et plus particulièrement,
A Virginie,
Je tiens à te remercier de m'avoir fait confiance pour ce stage que j'ai pu réaliser avec toi en fin de cursus d'école ingénieur. Ce stage m'a permis de mettre un pied dans l'équipe Biostatistique et d'y être restée depuis. Sache que cette première expérience m'a beaucoup apportée.
A Pierre,
Merci pour tes petites visites au bureau Student, puis au bureau 45, que ce soit simplement pour discuter ou pour me déposer les petites relectures qui m'étaient dédiées. Je voudrais également souligner ta patience, ton écoute et ton dévouement dont tu fais preuve afin que le monitorat se passe le mieux possible pour tous. Cela représente beaucoup de travail à gérer et tu le fais très bien.

A Cécile,
Ce fut un réel plaisir de pouvoir travailler avec toi sur ce dernier travail de thèse. Merci pour ta patience, ton écoute, ta confiance, nos réunions efficaces, ton partage de connaissances. Je voudrais également insister sur ta bonne humeur et ton enthousiasme qui m'ont donné l'envie et la motivation d'avancer dans ce travail. Merci également d'avoir pris du temps pour les diverses relectures qui donnaient toujours lieu à des retours très constructifs.
A Sandrine,
Merci pour ta réactivité et ton efficacité dans tout ce qu'on peut te demander, que tu fais toujours avec un grand sourire et beaucoup de gentillesse. C'est toujours un plaisir d'échanger avec toi au détour d'un couloir.

A l'ensemble de l'équipe EPICENE,
Je vous remercie pour votre accueil et surtout votre écoute lors de mes présentations qui ont pu faire peur à certains au moment de l'affichage des fameuses formules statistiques. Un grand Merci car ce sont ces présentations qui m'auront appris à vulgariser mon travail pendant ma thèse, ce qui est très important dans la recherche et ce que j'apprécie beaucoup faire maintenant.

Plus particulièrement,
A Patrick Brochard, Je suis très touchée de vous avoir dans mon jury. Vous êtes la première personne avec qui j'ai eu une réunion professionnelle en arrivant sur Bordeaux pour mon stage. Apprenant par la même occasion que je serai dans votre bureau pendant ces 6 mois.. Ce fut un peu stressant pour l'étudiante ingénieur que j'étais et qui se retrouvait plongée dans le milieu de la santé publique et de la santé au travail. Cette expérience fut très enrichissante et m'a beaucoup apportée.
A Cécile et Christel,
Un très grand merci à toi Cécile pour avoir géré d'une main de maître tous mes déplacements depuis ces 5 ans avec efficacité et bienveillance. Des déplacements qui n'étaient jamais de simple Bordeaux-Paris à gérer pourtant !
Un énorme merci à toutes les deux pour votre accueil toujours chaleureux quand je passais vous voir et pour toutes ces discussions que l'on a pu partager.
A Céline,
Merci pour ta gentillesse et toutes tes explications éclairées qui ont pu m'aider dès que j'avais des questions sur les données.

A l'ensemble de l'équipe pédagogique de l'ISPED,
Merci pour votre disponibilité à tous et votre bienveillance qui ont permis de rendre cette expérience du monitorat très agréable.

Plus particulièrement,
A Fleur et Valérie,
Merci pour m'avoir accordé votre confiance pour les enseignements, d'avoir été à l'écoute et toujours disponibles pour répondre à mes questions. Vous avez contribué à ce que l'enseignement ait été à la hauteur de mes attentes. C'était également toujours un grand plaisir de vous croiser dans les couloirs et de pouvoir discuter avec vous.
A Marthe-Aline,
Merci pour tous ces échanges et ton partage sur tes expériences qui ont fortement contribué à me forger mon propre regard sur l'enseignement et la pédagogie.
Aux secrétaires pédagogiques,
Je tiens sincèrement à vous remercier pour votre réactivité, votre efficacité et votre bonne humeur dont vous avez pu faire part à chacune de mes sollicitations.

A Madame Danièle Luce, Madame Stücker Isabelle et Monsieur Pascal Guénel,
Je tiens à vous remercier pour tous vos conseils, votre réactivité ainsi que tous vos commentaires toujours très pertinents qui ont pu faire avancer mes travaux lors de ma thèse.

A Sébastien Marque, Louise Baschet, Marie Anne Caillaud et toute l'équipe Capionis,
Merci de m'avoir donné l'opportunité de pouvoir travailler avec vous sur des projets très intéressants. Un grand merci pour votre accueil, votre écoute et votre disponibilité.

A la bande, vous êtes devenus bien plus que des collègues au fil de toutes ces années. Nous avons énormément partagé tous ensemble à travers des soirées, des repas, des sorties plages et j'en passe. Tout cela nous menant parfois même à nous voir 7 jours sur 7 ! Merci pour ces moments forts que l'on a pu partager avec ce très beau mariage et ces naissances passées sans oublier celles à venir.
Alexandra, merci pour ton écoute, ta bienveillance et ton calme (non négligeable pour canaliser toute cette bande !). Un grand merci pour ta présence lors de ces dernières semaines de rédaction qui ont largement contribué à faire baisser la pression !
Anaïs & Boris, merci pour votre humour (un peu décalé parfois), pour votre bonne humeur toujours au top. Anaïs, quel bonheur d'avoir eu la chance de partager le même bureau pendant quelques années !
Astou, merci pour toutes ces ondes positives que tu peux transmettre, ton sourire, tes conseils et tes expressions toujours plus singulières les unes que les autres.
Chloé, tu as tellement pu m'apporter au fil de ces années bordelaises..Tous ces échanges, ces fous rires, ces restos, ces voyages, ces pâtisseries, ont contribué à renforcer ce lien entre nous. Mention spéciale pour la fin de la thèse..sans toi cela n'aurait clairement pas été la même chose, merci pour ton écoute plus que de raison et pour tous ces mots qui ont pu me booster jusqu'au bout.
Henri, merci pour toutes ces phrases inachevées ou ponctuées de ces zzittt légendaires. La bande ne serait pas la même sans notre Henri national !
Loïc, merci pour ce bel accent de Sète, ce bonjour si singulier (le fameux zzitt) et pour tes blagues qui ont le mérite de te faire d'abord rire. Une grosse pensée pour Matteo et son rire si communicatif.
Matmat, tu es la première personne que j'ai rencontrée en arrivant à Bordeaux, et tu es celle qui m'y a fait rester ! Tellement de choses partagées depuis ces 5 années. Je ne te remercierais jamais assez pour ton écoute, tes conseils toujours bienveillants, ton sourire, tes encouragements. Je suis heureuse d'avoir pu partager avec toi plusieurs moments importants qui se sont passés dans ta vie ces derniers temps. Une pensée à Andrea et Alice, avec qui il fait toujours bon passer un moment.
Pépette, merci pour ton grain de folie, pour toutes ces pauses thés, pour toutes ces soirées passées (avec modération !) dans la bonne humeur. Et une petite mention spéciale pour m'avoir fait découvrir quelques classiques Disney !
Robin, ô toi grand sage de cette belle bande, merci pour toutes ces ondes positives qui contribuent à rendre plus belle cette bande de petits fous

A Louis, Lucile, Clément,
On peut remercier ce fameux congrès EPICOH à Barcelone qui nous aura permis de nous rencontrer. Merci pour ses petites pauses cafés dans votre bureau qui étaient toujours un réel boost, pour ces petites soirées toujours très sympathiques à refaire le monde.

A la relève,
Maude, merci pour tous ces moments de partages, de discussions toujours plus enrichissantes les unes que les autres, ton écoute, ton enthousiasme débordant. Je suis très heureuse d'avoir pu partager avec toi ce petit weekend new-yorkais en espérant qu'il y en aura bien d'autres !
Corentin, merci pour toutes ses discussions toujours très sympathiques en ta compagnie et pour ta vision toujours sereine sur les choses.

A Camille,
Je tiens sincèrement à te remercier pour ton écoute, ta sincérité, ta bienveillance lors de tous nos repas hebdomadaires à la cafèt souvent ponctués de nos petites anecdotes respectives.

A Mathilde R,
Merci pour tous ces moments partagés depuis ton arrivée dans le bureau Student. Il s'en est passé des choses depuis ! Il est toujours agréable d'échanger avec toi. Merci pour ton écoute, ton sourire, tes conseils et ton regard sur les choses.

A mes collègues de bureau au fil de ces années, Lucie, Rémi, Marion, Myriam, Agnieszka, Florian, Casimir, sans oublier Viviane, la plus fidèle, qui m'a supportée quasiment depuis le début !

Aux anciennes doctorantes de l'ISPED que j'ai eu l'occasion de rencontrer depuis mon arrivée sur Bordeaux et avec qui il a toujours été agréable de discuter : Marion, Audrey, Mélanie, Julie, ...

A mes "co-doctorants", Mathilde R, Bénédicte et le 'petit nouveau' Hugues, prenez soin de Karen. Je vous souhaite d'avoir une si belle expérience de thèse que celle qui m'aura été permis de vivre.

A mon petit Alex,
Toi qui a fortement contribué à ce que ma vie bordelaise soit plus belle depuis toutes ces années. Merci pour tous ces moments de partage agrémentés de fous rires, de sushis, de goûters. Mille mercis pour tout ! Hâte de partager avec toi ce merveilleux voyage au Japon qui nous attend !

A Mathilde et Mathieu,
Je ne saurais jamais trop vous remercier pour toutes ces années d'amitié qu'il serait bien difficile de résumer. Merci pour votre écoute, toutes nos soirées jeux / karaokés, ces fous rires, ces week-ends... Merci de rendre tout si facile ! Merci infiniment pour tous vos messages et vos marques de soutien pendant ces dernières semaines. Je vous promets de faire honneur à mon futur rôle de marraine. Vous n'imaginez pas à quel point cela peut me rendre heureuse !

A Anne-so,
Je te remercie profondément pour tout ce qu'a pu m'apporter notre amitié depuis toutes ces années (que l'on ne comptera pas !). Tu as toujours été d'un soutien et d'une écoute sans faille. Notre voyage a été une très belle aventure et a fortement contribué à appréhender au mieux cette troisième de thèse.

A Charlotte,
Merci pour tout ma Blonde, même si l'océan Atlantique nous sépare, notre amitié n'en reste pas moins belle et ton soutien pas moins important à mes yeux. Merci à toi et Thomas pour votre accueil toujours des plus chaleureux dans ce beau pays qui est le Canada !

A la team Erasums : Martin (& Pauline), Aurélien (& Marie), Caroline,
Nul doute que vous avez contribué à rendre ce voyage de l'autre côté de l'Atlantique magnifique ! Merci pour tout ce que l'on peut partager depuis cette expérience finlandaise. Votre soutien et votre enthousiasme face tout ce qui peut m'arriver professionnellement me font toujours chaud au coeur. Merci pour vos accueils parisiens & luxembourgeois, il fait toujours bon vous retrouver même si c'est le temps d'une petite soirée !

A Amandine,
Toi, mon amie d'enfance retrouvée au hasard du tram bordelais un vendredi soir.. Quelle chance j'ai pu avoir que tu sois de nouveau à mes côtés. Merci pour toutes ces séances de pilates, ces discussions, cette bienveillance et cette écoute sans relâche. Sans oublier Paco avec qui il est toujours agréable d'échanger sur le monde et sur ce qui nous entoure.

A mes boys, Micka, Victor, Stéphane, Benoit
Vous êtes de très belles rencontres qui ont pu ensoleiller ma vie bordelaise. Merci à vous, de m'avoir permis de partager de beaux moments avec vous ! Mon rôle de fille de la bande n'était pas pour me déplaire, je dois bien l'avouer :)

A Elodie,
Merci pour toutes ces belles années d'amitiés. Merci à toi et Nicolas, pour ces petits moments bordelais que l'on a pu vivre.

A mes parents,
Je voudrais vous remercier pour votre soutien, pour votre écoute et de votre confiance face à tout ce que j'ai pu entreprendre ces dernières années. Je suis très heureuse de tout ce que j'ai pu accomplir jusque là et vous n'y êtes pas étrangers.

A mon grand père, mes tantes, ma marraine et mon tonton,
Je me suis sentie soutenue par toutes vos petites attentions, vos appels, vos messages depuis ces dernières années. Ce qui est très important à mes yeux. J'espère continuer à pouvoir partager avec vous les belles choses qui m'attendent.

A mon cousin Sébastien et Charlotte,
Merci pour votre écoute, votre soutien et vos rires. Je suis vraiment très heureuse de ce lien que l'on continue à renforcer au fil de tous ces merveilleux moments partagés.

A ma cousine Magali, David, Eva et mon filleul d'amour Hugo,
C'est toujours un plaisir de partager avec toute la petite famille dès que cela est possible. Quelle joie vous m'avez fait d'avoir pu venir sur Bordeaux tous les 4 ! Merci à vous, pour tous ces moments passés. Je ne vous remercierai jamais assez pour ce rôle de marraine dont je suis le plus fière !

A Alexandre,
La vie offre parfois de belles surprises, tu auras été celle de ces derniers mois. Je tiens à te remercier pour tout ce que tu auras pu m'apporter. Merci également pour ton soutien, ton écoute (que j'ai certainement mise à rude épreuve ces derniers temps !), tes conseils avisés et tes mots toujours très justes.

A ma soeur,
Merci pour la force que tu me donnes à travers tes marques d'affection, tes encouragements, tes messages et ton enthousiasme dans tout ce que je décide d'entreprendre. Le mot merci parait si dérisoire à côté de tout ce que tu peux m'apporter au quotidien.

Table des matières

Valorisation scientifique

Publications

Articles scientifiques publiés

Lévêque E, Lacourt A, Luce D, Sylvestre MP, Guénel P, Stücker I, Leffondré K. Time-dependent effect of intensity of smoking and of occupational exposure to asbestos on the risk of lung cancer : results from the ICARE case–control study. *Occup Environ Med* 2018 ;75 :586-592.

Lacourt A*, **Lévêque E*** (co-first), Guichard E,Gilg Soit Ilg A, Sylvestre M-P, Leffondré K. Dose-time-response association between occupational asbestos exposure and pleural mesothelioma. *Occup Environ Med* 2017 ;74 :691-697.

Lettre

Lacourt A, **Lévêque E**, Goldberg M, et al. Dose-time response association between occupational asbestos exposure and pleural mesothelioma : authors' response. *Occup Environ Med* 2018 ;75 :161-162.

Article soumis

Lévêque E, Lacourt A, Phillips V, Luce D, Guénel P, Stücker I, Proust-lima C and Leffondré K. Association between lung cancer and lifetime profiles of intensity of exposure to occupational asbestos and smoking : Results from the ICARE case-control study. *International Journal of Epidemiology.*

Article en préparation

Lévêque E, Lacourt A, Leffondré K and Proust-lima C. Zero-Inflated Poisson latent class mixed model for the identification of longitudinal profiles of environmental and occupational exposure intensities in epidemiological studies.

Résumé publié

Lévêque E, Lacourt A, Luce D, et al. O15-2 Dynamic longitudinal effects of increasing intensity of smoking and occupational exposure to asbestos on lung cancer : results from the icare case-control study.Occup Environ Med 2016 ;73 :A28. (EPICOH Barcelone 2016).

Communications scientifiques

Présentations orales

Dynamic longitudinal effects of increasing intensity of smoking and occupational exposure to asbestos on lung cancer : results from the icare case-control study. The 25th International Epidemiology in Occupational Health (EPICOH) Conference, Barcelona, September 5-7, 2016.

A Zero-Inflated Poisson latent class mixed model for the identification of longitudinal profiles of environmental and occupational exposure intensities in epidemiological studies. The 38th Annual Conference of the International Society for Clinical Biostatistics (ISCB), Melbourne, Australia, August 26-30, 2018.

Présentations affichées

Identification of lifetime profiles of smoking intensities and association with lung cancer risks : Results from the ICARE case-control study. The 51st Annual Meeting of Society for Epidemiological Research (SER), Baltimore, June 19-22, 2018.

Analyse de l'aspect temporel de la relation entre expositions prolongées et risque de cancer Application à 3 relations : amiante-mésothéliome pleural, amiante-cancer du poumon et tabac-cancer du poumon. Rencontres scientifiques de l'ANSES pour les 10 ans du Programme national de recherche environnement santé travail (PNREST) Paris 14 Novembre 2016.

Bourses

Subvention doctorale en "Sciences Humaines et Sociales, Epidémiologie, Santé Publique" par l'Institut National du Cancer (INCa) (2016-2018)

SPC Travel Scholarship for The 51st Annual Meeting of Society for Epidemiological Research (SER), Baltimore, June 19-22, 2018 : 400$

Travail en collaboration

Article en révision

A.Danjou, T.Coudon, D.Praud, **E.Lévêque**, E.Faure, P.Salizzoni, M.Le Romancer, G.Severi, F. Mancini, K.Leffondré, L.Dossus and B.Fervers. Long-term Airborne Dioxin Exposure and Breast Cancer Risk in a case-control study nested within the French E3N Prospective Cohort. *Environmental International.*

Liste des abréviations

AIC : Critère d'Information d'Akaike

BIC : Critère d'Information Bayésien

CBNPC : Cancer Bronchique Non à Petites Cellules

CBPC : Cancer Bronchique à Petites Cellules

cig/jour : nombre de cigarettes fumées en moyenne par jour au cours d'une année

CSI : Comprehensive Smoking Index

éq f/mL : équivalent en fibres par millilitre

IC : Intervalle de Confiance

ICARE : Investigation sur les CAncers Respiratoires et l'Environnement

ICE : Indice Cumulé d'Exposition

IME : Indice Moyen d'Exposition

JLCMM : Modèle Conjoint à Classes Latentes

LCMM : Modèle Mixte à Classes Latentes

MEE : Matrice Emploi Exposition

MP : Mésothéliome Pleural

OR : Odd Ratio

PNSM : Programme National de Surveillance du Mésothéliome

RR : Risque Relatif

VLEP : Valeur Limite d'Exposition Professionnelle

WCIE : Indice Cumulé pondéré d'Exposition

ZIP : modèle de Poisson à sur-représentation de Zéros

ZIP-LCMM : Modèle Mixte de Poisson à sur-représentation de Zéros à Classes Latentes

Introduction

Sommaire

Cette thèse porte sur l'étude des expositions prolongées, environnementales et professionnelles, en étiologie du cancer. Elle se concentre sur la problématique de la considération de l'aspect temporel de ces expositions, et plus particulièrement sur la modélisation de l'intensité des expositions au cours de la vie, pour évaluer l'association de ces expositions avec le risque de cancer. Pour illustrer cette problématique en épidémiologie environnementale, nous nous sommes intéressés en particulier à deux cancers ; le mésothéliome pleural et le cancer du poumon, et à deux expositions ; l'exposition professionnelle à l'amiante et la consommation tabagique.

0.1 La problématique de l'aspect temporel des expositions prolongées

Les expositions environnementales ou professionnelles sont, le plus souvent, des expositions prolongées. Elles peuvent en effet s'étendre sur une période assez longue de la vie (comme la carrière professionnelle), en comparaison avec les expositions de courte durée qui sont dues à des évènements ponctuels (comme par exemple les expositions aux rayonnements lors de l'explosion de Tchernobyl).

Pour estimer la relation entre des expositions prolongées et le risque de cancer, il est nécessaire de représenter au mieux l'exposition des sujets dans le modèle de régression. Cette exposition peut être caractérisée de manière très simple par une variable binaire en considérant le sujet comme exposé ou non exposé. Cette modélisation permet d'évaluer si le fait d'être exposé constitue un risque significativement plus élevé de développer le cancer. Cette simple analyse ne permet pas d'évaluer les aspects quantitatifs et temporels de l'exposition. En effet, les expositions prolongées sont multidimensionnelles. Elles sont définies par un ensemble de variables d'exposition telles que l'âge de début d'exposition, la durée totale d'exposition, le temps depuis l'arrêt de l'exposition, le temps depuis la première exposition ou encore l'intensité d'exposition. Dans les études cas-témoins en étiologie du cancer, l'histoire d'exposition est collectée de manière très détaillée. Néanmoins, pouvoir intégrer toute cette histoire d'exposition dans un modèle de régression pour évaluer l'association entre l'exposition prolongée et le cancer n'est pas trivial (Thomas [1988], Langholz et al. [1999], Thurston et al. [2005], Lubin and Caporaso [2006], Richardson et al. [2011], Vermeulen and Chadeau-Hyam [2012]).
Classiquement, la quantité d'exposition est représentée par un indice cumulé d'exposition (ICE) qui est la somme de toutes les intensités reçues au cours de cette histoire d'exposition. Par exemple, pour le tabac, l'ICE est généralement le nombre de cigarettes-années, obtenu en sommant chaque période de consommation représentée par le produit entre la durée de cette période et la consommation moyenne au cours de celle-ci. Cependant, l'ICE ne permet pas de prendre en compte le moment de l'exposition. Or, par exemple, pour le tabac, on sait que le risque de cancer du poumon diminue après l'arrêt du tabagisme (Peto et al. [2000], Leffondré et al. [2002], Rachet et al. [2004]). Ainsi, pour avoir une meilleure estimation de l'effet de la quantité fumée, d'autres variables d'exposition temporelles, telles que le temps depuis l'arrêt, peuvent être intégrées dans le modèle de régression. Cependant, se pose le problème de la multicolinéarité entre les variables temporelles incluses, comme, par exemple, entre l'âge de début d'exposition, la durée totale, le temps depuis l'arrêt et l'âge courant (Leffondré et al. [2002]). Pour éviter ce problème de multicolinéarité, une solution peut être de considérer un indice

agrégé intégrant les variables d'exposition considérées les plus caractéristiques de l'histoire de l'exposition. Un exemple d'un tel indice est le Comprehensive Smoking Index (CSI)(Leffondré et al. [2006]). Dans les études sur le cancer du poumon, le CSI permet d'ajuster finement sur l'histoire du tabagisme en tenant compte de l'intensité moyenne au cours de la vie, la durée et le temps de l'arrêt du tabagisme.

Cependant, tout comme l'ICE, le CSI ne permet pas de prendre en compte la dynamique de l'intensité d'exposition au cours du temps, ni d'identifier différents profils d'intensité d'exposition au cours de la vie et d'estimer leur association avec le risque de cancer. Dans la littérature, la variabilité de l'intensité d'exposition au cours de la vie reste assez peu étudiée. Pourtant, l'information sur l'intensité au cours de la vie est très souvent recueillie dans les études épidémiologiques car il est facilement concevable qu'une intensité forte en début de carrière, ou à un jeune âge, puisse avoir un effet différent sur le risque de cancer qu'une intensité forte en fin de carrière, ou à un âge plus avancé. Il apparait donc important de tenir compte de la variation de l'intensité de l'exposition au cours de la vie et d'évaluer son impact sur le risque de cancer. Dans le cadre de cette thèse, afin d'étudier cette problématique, nous nous sommes intéressés plus particulièrement à 3 relations dose-réponse :

- Intensité d'exposition professionnelle à l'amiante et risque de développer un mésothéliome pleural (MP) ;
- Intensité d'exposition professionnelle à l'amiante et risque de développer un cancer du poumon ;
- Intensité de la consommation de tabac et risque de développer un cancer du poumon

Bien que le rôle causal de l'exposition à l'amiante dans le MP ou du tabac dans le cancer du poumon ne fassent plus l'objet de controverse aujourd'hui, des questions demeurent sur le rôle de l'intensité de l'exposition au cours de la vie dans le développement de ces cancers. En particulier, l'intensité d'exposition a-t-elle un effet différent selon le moment où cette intensité est reçue au cours de la vie ? Existe-t-il des profils de trajectoires d'intensité d'exposition au cours de la vie associés à des risques de cancer différents ?

0.2 Mésothéliome pleural

0.2.1 Epidémiologie du MP

Le mésothéliome pleural malin est une tumeur maligne qui touche la plèvre. La plèvre est une membrane thoracique composée de deux feuillets ; un feuillet qui couvre la face externe des poumons (plèvre pulmonaire ou viscérale) et un qui est accolé à la cage thoracique (plèvre pariétale) (figure 1). Ces deux feuillets sont séparés par la cavité pleurale contenant habituellement du liquide qui sert au glissement du poumon dans la cage thoracique lors des mouvements respiratoires. La plèvre reste la localisation la plus fréquemment atteinte dans le cadre des mésothéliomes (plus de 85%). D'autres localisations peuvent être également atteintes comme le péritoine pour 10% des cas, on parlera de mésothéliome péritonéal, et plus rarement le péricarde (2% des cas). Dans la suite de ce manuscrit et par soucis de simplification, le terme mésothéliome fera systématiquement référence au mésothéliome pleural malin.

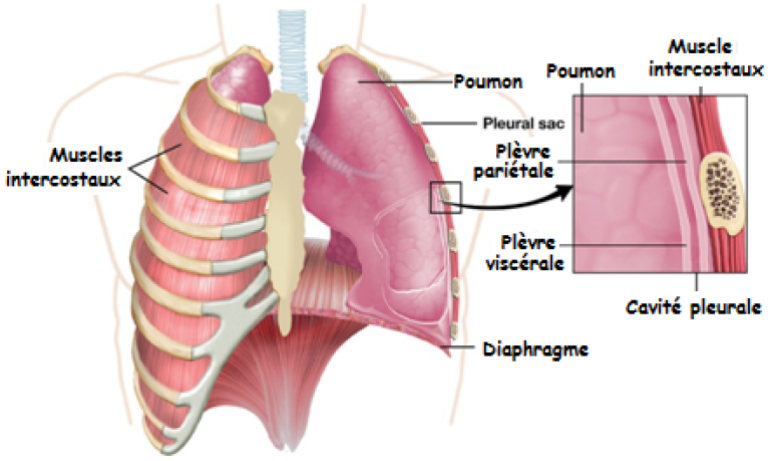

FIGURE 1 – Schéma des poumons et de la plèvre
http : //ressources.unisciel.fr/physiologie/co/grain1d$_1$*.html*

Le mésothéliome est un cancer rare. D'après Gilg Soit Ilg et al. [2009], entre 1998 et 2006 en France, le nombre moyen annuel de cas de mésothéliomes peut être estimé entre 535 et 645 chez les hommes et entre 152 à 210 chez les femmes. Les taux bruts d'incidence sont compris entre 1,85 et 2,23 pour 100 000 chez les hommes et entre 0,5 à 0,68 pour 100 000 chez les femmes. La France est considérée comme un pays avec une incidence moyenne contrairement aux Pays-Bas, la Belgique ou l'Australie, tous considérés comme des pays avec une forte incidence avec des taux d'incidence proches de 3 pour 100000 personnes (Bianchi and Bianchi [2014]). En France, Le Stang et al. [2010] ont montré une stabilisation de l'incidence chez les hommes entre 1998 et 2005. Quant à la mortalité, en 2005 en France, le mésothéliome pleural aurait été imputé à plus de 1090 décès (Belot et al. [2008]). Les projections en France prévoient un pic de décès annuels aux environs de 2020 (Le Stang et al. [2010]).

Le diagnostic est difficile à poser, comme pour tous les cancers rares. Il y a donc eu la mise en place d'une relecture des prélèvements tumoraux par des experts pathologistes pour avoir un diagnostic de certitude. Ces experts sont répartis sur toute la France et forment le réseau Mesopath. Comme les symptômes sont peu spécifiques (douleurs thoraciques, épanchement pleural (excès de liquide qui se répand dans la cavité pleurale) et altération de l'état général), le stade au diagnostic est souvent avancé. On distingue 5 stades de la maladie (IA, IB, II, III, IV, V) grâce à la classification TNM des tumeurs pleurales (T extension de la tumeur, N extension ganglionnaire, M extension métastatique) (Rusch [1995]).

L'âge médian au diagnostic est de 70 ans pour les hommes et 67 ans pour les femmes (Roos [2004]). Le pronostic du mésothéliome pleural est des plus sombres, la survie médiane est estimée entre 9 et 12 mois (Scherpereel and Astoul [2007]), avec une survie observée à 5 ans qui est de l'ordre de 5% selon les données des registres du réseau Francim (Le Stang et al. [2010], Galateau-Sallé et al. [2014]), sur la période entre 2005 et 2010. La survie à 5 ans varie peu entre les pays, celle observée en

France se situant dans les valeurs moyennes européennes (Siesling et al. [2012], De Angelis et al. [2014]).

Il n'existe aucun traitement curatif qui a été démontré efficace pour ce cancer rare. Pour diminuer la taille de la tumeur ou le nombre de métastases, le traitement standard est la chimiothérapie à base de sels de platine. La chirurgie curatrice est limitée à de rares stades précoces. Pour la prise en charge clinique, il existe le réseau Mésoclin constitué de centre d'experts cliniques du MP dont une des missions est d'assurer au mieux la prise en charge clinique des patients, partout en France.

0.2.2 Programme National de Surveillance du Mésohéliome (PNSM)

Afin de constituer un système de surveillance épidémiologique des effets de l'amiante sur la santé de la population française, un suivi des tumeurs primitives de la plèvre a été initié à partir de 1998 à la demande de la Direction Générale du Travail (DGT) et de la Direction Générale de Santé (DGS). Ce programme, dénommé Programme National de Surveillance du Mésothéliome (PNSM), est coordonné par la Direction Santé Travail (DST) de Santé publique France (SpF). Plus précisément, divers objectifs de ce programme de surveillance peuvent être mis en avant :

— estimer l'incidence nationale du mésothéliome en France et son évolution

— étudier la proportion des mésothéliomes en France attribuables à une exposition à l'amiante (notamment en milieu professionnel)

— contribuer à améliorer le diagnostic anatomopathologique du mésothéliome

— contribuer à la recherche d'autres facteurs étiologiques potentiels

— évaluer la reconnaissance du mésothéliome de la plèvre comme maladie professionnelle

Du fait de ces différents objectifs, le PNSM se compose de plusieurs volets : i) le volet "incidence", ii) le volet "expositions-étiologie", iii) le volet "confirmation anatomopathologique et clinique", et iv) le volet "médico-social".

Le PNSM repose sur l'enregistrement exhaustif depuis le 1er janvier 1998 de tous les cas incidents de mésothéliome pleural dans un nombre défini de départements français. Les caractéristiques socio-professionnelles et démographiques de la population couverte par ce programme sont similaires à celles de la population générale française. Les premiers résultats de ce programme sont publiés dans Goldberg et al. [2006].

0.2.3 Association entre l'amiante et le mésothéliome pleural

Même si d'autres particules minérales que l'amiante semblent impliquées dans l'étiologie du MP, comme par exemple ; les fibres d'érionite (Baris et al. [1987], Baris and Grandjean [2006]) et les fibres amphiboles de fluoro-édénite (Bruno et al. [2014], Conti et al. [2014]), l'exposition à l'amiante est clairement établie comme le principal facteur de risque du mésothéliome pleural.

0.2.3.1 Définition de l'amiante et circonstances d'exposition

L'amiante est une substance minérale naturelle fibreuse qui se distingue par deux grandes familles, les amphiboles et les serpentines. A l'intérieur de ces deux grandes familles, on retrouve une ou plusieurs espèces aux propriétés physico-chimiques différentes telles que la chrysotile pour la famille des serpentines et l'amosite, la crocidolite, l'anthophyllite, l'actinolite et la trémolite pour celle des amphiboles. Les fibres d'amiante sont définies comme des particules allongées dont le diamètre est inférieur à 3 μm, avec une longueur supérieure ou égale à 5 μm et un rapport longueur sur diamètre qui est égal ou supérieur à 3 (définition OMS). Seules les fibres respirables dont le diamètre est inférieur à 3.5 μm ont la capacité d'atteindre la plèvre.

L'amiante a été classé dans le groupe 1 "cancérigène certain" pour l'être humain par le Centre International de la Recherche sur le Cancer en 1977 (IARC [1977]). De par ses caractéristiques physico-chimiques (entre autres, résistance au feu, faible conductivité thermique, imputrescibilité), l'amiante a été très largement utilisé dans les secteurs de l'extraction et transformation de l'amiante, de l'isolation, des chantiers navals, etc.. En France, avant son interdiction en 1997, l'exploitation de l'amiante a atteint son apogée dans les années 70. Depuis, les fibres d'amiante restent présentes dans les bâtiments anciens. L'exposition à l'amiante est encore possible au travers de l'amiante en place et du désamiantage. Des Valeurs Limites d'Exposition Professionnelle (VLEP) aux fibres d'amiante ont pu être fixées par la législation. Depuis le 1er juillet 2015, suivant les différentes recommandations du rapport de l'Afsset [2009] (Agence française de sécurité sanitaire de l'environnement et du travail), la VLEP actuelle a été fixée à 10 fibres par litre (0,01f/ml) évaluée sur une moyenne de 8 heures de travail (équivalent à une journée classique de travail). Elle a été fixée en se basant sur une revue de la littérature des VLEP existantes en Europe.

On peut distinguer 3 types d'exposition à l'amiante :

— les expositions professionnelles liées à l'extraction, la transformation, la manipulation ou l'usinage de matériau contenant de l'amiante, et le désamiantage ;

— les expositions extra professionnelles regroupant les expositions para professionnelles (poussières par les vêtements de travail des conjoints), domestiques (par les objets ménagers) et de bricolage (activités réalisées en dehors du travail) ;

— les expositions environnementales définies par la pollution émise par une source naturelle ou industrielle.

Dans le cadre de cette thèse, nous nous focalisons seulement sur les expositions professionnelles à l'amiante même s'il existe des données indiquant une relation entre MP et expositions para-professionnelles, domestiques ou environnementales (Magnani et al. [2001], Goldberg and Luce [2009], Lacourt et al. [2014]).

Les expositions professionnelles à l'amiante peuvent être caractérisées de 3 manières différentes suivant le contact ou non avec les fibres par le travailleur :

— Directe : en lien avec les tâches du travailleur à son poste de travail qui manipule lui-même le matériau contenant l'amiante

— Indirecte : en lien avec la co-activité des autres travailleurs. Il y a des personnes dans l'environnement de travail qui manipulent directement le matériau contenant le polluant sans que le travailleur le manipule lui-même

— Passive : en lien avec la dégradation des matériaux en place dans l'environnement de travail

Certaines études ont mesuré l'impact global sur la population de l'exposition professionnelle à l'amiante sur le MP, en tenant compte de la proportion des personnes exposées. Ceci se traduit par le calcul d'une fraction de risque de MP attribuable à l'exposition professionnelle à l'amiante, qui a été évaluée proche de 90% pour les hommes dans des études européennes (Lacourt et al. [2014], Rushton et al. [2010], Ferrante et al. [2016]) et américaine (Spirtas et al. [1994]).

0.2.3.2 Association entre intensité d'exposition professionnelle à l'amiante et MP

La première étude montrant une association entre l'exposition professionnelle à l'amiante et le mésothéliome pleural a été menée chez les mineurs de crocidolite en Afrique du Sud (Wagner et al. [1960]). Par la suite, l'association a été largement confirmée à travers différentes études de cohorte et des études cas-témoins (Seidman et al. [1986], Selikoff and Seidman [1991], McDonald and McDonald [1980]). Le temps de latence, défini par le temps entre la première exposition professionnelle à l'amiante et le diagnostic ou décès, est très long, pouvant varier de 20 à 40 ans (Lanphear and Buncher [1992]), allant même parfois au-delà de 60 ans (Bianchi et al. [1997]). Plus précisement, Lanphear and Buncher [1992] ont montré un temps de latence moyen de 32 ans en se basant sur 21 articles scientifiques tandis que Marinaccio et al. [2007] et Bianchi et al. [1997] ont plutôt montré un temps de latence moyen proche de 45 ans en se basant sur des données de registres ou des données hospitalières italiennes. Dans les années 1980, Doll et Peto se sont intéressés à la relation entre l'exposition professionnelle à l'amiante et le mésothéliome pleural en utilisant des modèles de carcinogénèse (Peto et al. [1982], Doll and Peto [1985]). Ces modèles permettent d'évaluer le taux d'incidence du mésothéliome pleural en fonction de différentes variables d'exposition telles que l'intensité, la durée, le temps depuis la première exposition (Peto et al. [1985]). A partir de données de cohorte, Peto et al. [1982] a montré que le taux d'incidence du MP chez les travailleurs exposés à l'amiante semblait être proportionnel au temps depuis la première exposition à la puissance 3, reconnu comme le modèle cubique. L'âge à la première exposition semblait avoir peu ou pas d'influence sur le taux d'incidence. Cependant, une telle modélisation repose sur des hypothèses paramétriques fortes et ne permet pas d'estimer l'effet de l'intensité au cours du temps sans poser d'hypothèses à priori. D'après Langholz et al. [1999], le modèle de carcinogénèse est vu comme permettant de faire des prédictions spécifiques sur certains effets de l'exposition et non comme un outil descriptif pour étudier la relation entre l'exposition et le cancer.

Dans la littérature, la relation dose-réponse avec l'intensité d'exposition a largement été étudiée et a clairement été établie. Cette relation a pu être mise en évidence, principalement, avec l'exposition cumulée représentée par l'indice cumulé d'exposition (ICE) (Iwatsubo et al. [1998], Rödelsperger et al. [2001]), ou simplement avec l'intensité moyenne au cours de la vie (Agudo et al. [2000]). Généralement, la plupart des études cas-témoins en population générale qui s'intéressent à l'association entre l'intensité d'exposition et le MP ont considéré un effet non linéaire de la variable représentant l'intensité d'exposition, en la catégorisant. En plus d'induire une perte d'information importante, une telle catégorisation peut aussi influencer la forme de la relation dose-réponse estimée (Greenland [1995]). Récemment, l'exposition cumulée, représentée par l'indice moyen d'exposition (IME) dans Lacourt et al. [2012], a été considérée de manière plus flexible avec des splines, ainsi aucun a priori a été imposé sur la forme de la relation dose-réponse.

Dans la littérature, l'association entre l'exposition professionnelle à l'amiante et le MP a également pu être étudiée avec d'autres variables d'exposition, notamment la durée totale d'exposition qui a été retrouvée significativement associée avec le risque de MP (Spirtas et al. [1994], Rödelsperger et al. [2001], Rake et al. [2009]) ou encore l'âge à la première exposition (Spirtas et al. [1994]). Récemment, Reid et al. [2014] et Pira et al. [2016] ont pu retrouver une augmentation du risque de MP même après plusieurs années d'arrêt d'exposition (plus de 40 ans).

Il existe donc une association avec quasiment toutes les variables d'exposition. Néanmoins, à ce jour, il reste encore des questions à investiguer notamment en lien avec l'intensité d'exposition. En effet, comme on a pu le préciser en début de cette introduction, cette variable d'intensité est une variable dynamique qui peut varier au cours du temps. Et à ce jour, il n'existe aucune donnée étudiant l'effet de l'intensité d'exposition en fonction du temps et plus spécifiquement l'impact de la variation de l'exposition au cours du temps sur le risque de mésothéliome pleural avec des données provenant d'études cas-témoins. Notamment, l'évaluation de l'impact des intensités récentes sur le risque de MP reste assez peu étudiée. En effet, les intensités reçues quelques années avant le diagnostic ne sont généralement pas considérées dans l'évaluation de la relation dose-réponse en faisant l'hypothèse qu'elles n'ont pas d'impact sur le sur-risque de cancer. Or, à ce jour, il n'existe aucune donnée convaincante sur le nombre d'années à ne pas considérer, puisque d'une étude à l'autre cela peut être différent (Iwatsubo et al. [1998], Rödelsperger et al. [2001], Lacourt et al. [2014]). Cependant, dans le cadre de l'analyse des relations dose-réponse, certains auteurs ont pu mettre en évidence que cela pouvait amener à des résultats biaisés ou inconsistants suivant le choix considéré (Salvan et al. [1995], Richardson et al. [2011]). Par ailleurs, puisque l'exposition est majoritairement modélisée par l'indice cumulé d'exposition, cela pose l'hypothèse que quelque soit la dynamique de l'intensité d'exposition reçue au cours du temps, le risque de survenue de MP est le même. Ainsi, pouvoir considérer correctement cette dimension temporelle dans l'analyse de cette relation dose-réponse permettrait de répondre aux questions suivantes :

— les expositions récentes contribuent-elles au risque de MP ?
— divers profils de trajectoires d'intensité d'exposition conduisent-ils à des risques de MP différents, y compris s'ils mènent à une même dose cumulée totale ?

0.3 Cancer du poumon

0.3.1 Epidémiologie du cancer du poumon

Le cancer du poumon ou encore cancer broncho-pulmonaire est une maladie des cellules des bronches ou plus rarement des cellules qui tapissent les alvéoles pulmonaires (cavités qui sont un lieu d'échanges gazeux entre l'air respiré et le sang) (figure 2). Il existe deux principaux types de cancer selon l'origine des cellules des bronches dont elles sont issues : le cancer bronchique non à petites cellules (CBNPC) qui représentent 85% des cas de cancer du poumon et le cancer bronchique à petites cellules (CBPC). La différenciation de ces cellules est possible car elles n'ont pas le même aspect au microscope. De plus, il est très important de les distinguer lors du diagnostic puisqu'elles se comportent différemment au cours de la progression du cancer et n'ont pas la même sensibilité aux traitements. Pour les CBNPC, on distingue 3 formes distinctes assez fréquentes : adénocarcinome bronchique, carcinome épidermoïde et carcinome à grandes cellules.

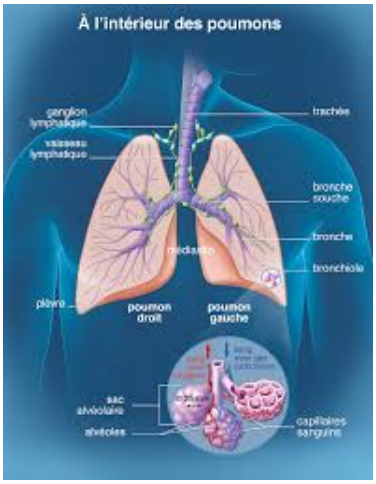

FIGURE 2 – Schéma des poumons
Les traitements du cancer du Poumon, INCA

Pour poser et confirmer le diagnostic, plusieurs examens sont nécessaires. Cela repose principalement sur un examen clinique qui peut être associé à une radiographie du thorax, un scanner thoracique et une biopsie. Les symptômes sont assez peu spécifiques de cette maladie. Ils combinent généralement des problèmes respiratoires (toux, expectorations sanguinolentes, infections pulmonaires à répétition,...) à une altération de l'état général du malade (fatigue inhabituelle, perte d'appétit, perte de poids, ...).

En France, en 2012, 28211 nouveaux cas de cancer du poumon ont été diagnostiqués chez les hommes, ce qui positionne le cancer du poumon à la 2ème place des cancers les plus fréquents chez les hommes, après le cancer de la prostate. Pour les femmes, avec 11284 nouveaux cas

en 2012, il se situe à la 4ème place. Les taux d'incidence standardisés sur l'âge de la population mondiale sont de 51.5 pour 100000 personnes-années pour les hommes et de 18.6 pour 100000 personnes-années pour les femmes (Binder-Foucard et al. [2013], Binder-Foucard et al. [2014]). D'après les données des registres Francim, pour les hommes, on constate des disparités importantes d'incidence en fonction des départements français sur la période 2009-2013 (taux d'incidence entre 40 et 66 pour 100000 personnes-années en France métropolitaine), ce qui n'est pas le cas pour les femmes à cette même période (entre 16 et 20 pour 100000 personnes-années). Chez l'homme, l'incidence du cancer du poumon est quasi stable depuis 1980, on constate une légère diminution en moyenne de 0.3% entre 2005 et 2012. Pour les femmes, depuis 1980, l'incidence ne cesse d'augmenter, ce qui s'explique principalement par une augmentation de la consommation de tabac chez les femmes ces dernières années. Ce phénomène a également été constaté aux Etats-Unis où le pic d'incidence maximum sur une période allant de 1999 à 2008 était en 2006 (CDC [2011]). En France, entre 2005 et 2012, on constate une augmentation moyenne de 5.4% de nouveaux cas de cancer du poumon chez les femmes.

En France, la tranche d'âge qui représente le plus grand nombre de cas de cancer du poumon est celle de 60 à 64 ans chez les hommes comme chez les femmes (avec 5053 cas pour les hommes, 1753 cas pour les femmes) (Binder-Foucard et al. [2013]).

Le cancer du poumon est la première cause de décès par cancer pour les hommes et la deuxième pour les femmes précédée par le cancer du sein. En France, en 2012, le taux de mortalité était de 37 pour 100000 personnes-années pour les hommes et 12.9 pour 100000 personnes-années pour les femmes. Comme pour l'incidence, entre 2005 et 2012, on constate une diminution annuelle moyenne de 2.2 % du taux de mortalité chez les hommes alors que l'on constate une augmentation annuelle moyenne de 4.6% chez les femmes (Binder-Foucard et al. [2013]).

Le pronostic vital est assez défavorable avec une probabilité de survie à 5 ans du diagnostic ne dépassant pas les 15% (Delva et al. [2016], Alberg et al. [2007], Janssen-Heijnen and Coebergh [2003]). Les CBNPC ont un meilleur pronostic vital que les CBPC (14.8% vs 5.1% à 5 ans pour les hommes).

Trois grands types de traitements peuvent être mis en place au moment du diagnostic :

— la chirurgie
— la radiothérapie
— les traitements médicamenteux (chimiothérapie conventionnelle (à base de sels de platine), les thérapies ciblées ou encore les immunothérapies)

Chacun de ces traitements médicamenteux n'agit pas de la même façon sur les cellules et n'est pas mis en place pour les mêmes raisons. Ainsi, suivant le stade de la maladie, certaines combinaisons de traitements pourront être mises en place pour être plus efficaces.

0.3.2 Association entre le tabac et le cancer du poumon

La consommation de tabac est le principal facteur de risque du cancer de poumon en France et dans le monde. A partir d'études réalisées en Europe, en Amérique du Nord et au Japon, 92% de décès par cancer du poumon sont attribuables au tabac pour les hommes dans les pays développés en 1990 contre 69% pour les femmes (Peto et al. [1992], Lopez et al. [1994]). En France en 2004, le taux de décès par cancer du poumon attribuable au tabac est de 91% pour les hommes et 59% pour les femmes (Hill and Doyon [2008]). L'effet cancérigène du tabac est reconnu depuis 1986 par le Centre International de la Recherche sur le Cancer (IARC [1986]).

0.3.2.1 Définition et prévalence du tabagisme

Dans cette thèse, nous entendons par tabagisme, la consommation de tabac. Le tabac peut être consommé de plusieurs manières ou sous différentes formes (cigarettes, pipes, cigares, ...). Dans les études épidémiologiques, comme ICARE (Luce and Stücker [2011]) par exemple, les autres types de consommations que la cigarette peuvent être convertis en équivalent grammes de tabac afin d'obtenir une équivalence en nombre de cigarettes.

La fumée de cigarettes comporte plus de 4000 substances dont près de 66 (IARC [2008]) ont été reconnues cancérigènes. Elle représente un danger pour les fumeurs mais aussi les non-fumeurs via le tabagisme passif (IARC [2004]). En effet, la fumée de tabac environnementale définissant le tabagisme passif a été reconnue comme cancérigène certain pour l'homme par le CIRC en 2004 (IARC [2004]). Cependant, dans le cadre de cette thèse, nous avons seulement étudié l'effet de l'intensité de la consommation active de cigarettes, et nous n'avons pas pris en compte le tabagisme passif.

D'après Beck et al. [2011] basé sur les chiffres de Baromètre santé, l'évolution du tabagisme chez les hommes a fortement diminué de 1975 jusque dans les années 2000 (passant de plus de 60% à près de 40%). Concernant les femmes, nous constatons une tendance différente avec une augmentation de 1975 jusqu'en 1990 (passant de 20% à plus de 30%) avant de diminuer dans les années 2000 pour repartir à la hausse entre 2005 et 2010, restant proche de 30%. En 2015, la prévalence pour hommes et femmes confondus reste assez stable (Guignard et al. [2015]) comparée à celle présentée dans Beck et al. [2011]. Chez les hommes et femmes confondus, une baisse du nombre de cigarettes fumées par jour est également constatée entre 2005 et 2010 (15.4 contre 13.9). La prévalence du tabagisme chez les 15-75 ans en activité reste assez stable entre 2000 et 2010 (près de 34%) alors que l'on peut constater une augmentation pour les personnes au chômage (40.9% en 2000 contre 51.1% en 2010). De même, la plus forte augmentation en terme de prévalence est constatée pour les sujets qui n'ont pas de diplôme (30.5% en 2000 et 34.0% en 2010).

0.3.2.2 Association entre intensité de consommation de tabac et cancer du poumon

Deux études cas-témoins dans les années 50 ont montré que le tabac était associé au cancer du poumon (Doll and Hill [1950], Wynder and Graham [1950]). Un temps de latence de 30 ans a été évoqué par Weiss [1997].

Dans les années 60-80, Armitage, Doll & Peto ont contribué à enrichir les connaissances sur la relation entre tabac et cancer du poumon en utilisant des modèles de carcinogénèse. Basée sur une étude prospective de docteurs britanniques masculins, ils ont pu déterminer l'incidence du cancer du poumon en fonction du nombre de cigarettes fumées par jour et de l'âge en modélisant l'effet de ces variables avec des polynômes de degrés différents, respectivement 2 pour l'intensité et 4 pour l'âge (Armitage and Doll [1961], Doll [1971], Doll and Peto [1978]). Ils ont également retrouvé que l'incidence de cancer du poumon chez les fumeurs courants était associée à la durée d'exposition à la puissance 4 ou 5, indépendamment de l'âge et de l'intensité.

Dans la littérature, la relation dose-effet avec l'intensité est clairement établie. L'intensité du tabac est souvent représentée soit par l'intensité moyenne au cours de la vie en nombre de cigarettes fumées par jour soit par l'indice cumulé d'exposition exprimé en paquet-année. Un effet plafond du risque pour les fortes intensités a été retrouvé dans certaines études (Vineis et al. [2000], Pesch et al. [2012]). Récemment, Vineis et al. [2000] ont montré un effet plateau pour des intensités moyennes journalières au cours de la vie à partir de 25 cigarettes fumées par jour lorsque le modèle était ajusté sur la durée d'exposition. Ce plateau n'était pas si nettement observé lorsque le modèle ne prenait pas en compte la durée. L'ajustement sur la durée d'exposition semble important pour évaluer au mieux la relation dose-effet. En effet, son effet sur le risque de cancer a été largement étudié dans la littérature, de manière séparée ou non de l'effet de l'intensité d'exposition (Thurston et al. [2005]).

D'autres variables d'exposition ont été étudiées pour évaluer leur association avec le risque de cancer du poumon. Dans la littérature, le rôle du temps depuis l'arrêt de consommation tabagique sur le risque de cancer du poumon a clairement été établi, avec un risque de cancer du poumon qui diminue après l'arrêt du tabac (Weiss [1997], Peto et al. [2000], Rachet et al. [2004]). Pesch et al. [2012] ont retrouvé que l'arrêt à de jeunes âges ($<$ à 60 ans) montre un plus grand bénéfice. Cependant, ils précisent que les hommes ne reviennent pas à un risque de base des non fumeurs même après un arrêt de plus de 35 ans, même si le rapport de côte est bien plus faible comparé à celui des fumeurs courants (OR=2.2, 95% IC : 1.8-2.8) contre un OR=23.6 (95% IC : 20.4-27.2) pour les fumeurs courants. L'effet de l'âge à l'initiation du tabagisme a été sujet à controverse dans la littérature, notamment avec Hegmann et al. [1993] et Benhamou and Benhamou [1994]. Plus récemment, Pesch et al. [2012] ont pu retrouver un effet de l'âge à l'initiation sur le risque de cancer du poumon à partir de l'analyse de données poolées. Tout ceci

a pu mettre en avant l'importance de la prise en compte de l'histoire d'exposition pour évaluer au mieux la relation dose-réponse. Ainsi, pour la relation tabac-cancer du poumon, certaines études ont modélisé de manière plus flexible les variables d'expositions continues au lieu de les catégoriser (Rachet et al. [2004]), mais également, ont essayé de caractériser plus finement l'histoire du tabagisme tout en tenant compte des problèmes de multicolinéarité entre les variables (Lubin and Caporaso [2006], Vlaanderen et al. [2013], Leffondré et al. [2006], Thurston et al. [2005], Chadeau-Hyam et al. [2014]). Notamment, le développement d'un indice agrégé comme le CSI, évoqué au début de cette introduction, a permis de montrer un meilleur ajustement sur l'histoire tabagique qu'avec le simple indice cumulé d'exposition représenté par le paquet-année pour des données provenant d'études cas-témoins. Le CSI est défini par la formule ci-après (Leffondré et al. [2006]).

$$\text{CSI} = (1 - 0.5^{dur^\star/\tau})(0.5^{tsc^\star/\tau})\ln(int + 1) \tag{1}$$

où

- int l'intensité moyenne au cours de la vie ;
- dur la durée totale d'exposition ;
- tsc le temps depuis l'arrêt ;
- $tsc^\star$=max(tsc-δ,0) ;
- $dur^\star$=max(dur+tsc - δ,0)-$tsc^\star$;
- τ est un paramètre de forme et δ est un paramètre de décalage

Néanmoins, la variabilité de l'intensité de la consommation de tabac au cours du temps n'a été que très peu étudiée pour évaluer son impact sur le risque de cancer du poumon. Hauptmann et ses collaborateurs ont proposé des méthodes permettant d'étudier cet effet au cours du temps de manière plus ou moins flexible (Hauptmann et al. [2000a], Hauptmann et al. [2000b]). Pourtant, elles restent peu utilisées bien que cela puisse mettre en avant l'importance de considérer la dynamique de l'intensité reçue au cours du temps dans l'analyse de relation dose-réponse à partir de données cas-témoins. Néanmoins, dans ces études, ni l'impact des expositions reçues à de jeunes âges, ni l'impact de la considération des expositions proches de la date index n'a été investigué sur le risque de cancer du poumon. De plus, cette méthode ne permettait pas d'identifier des profils d'intensité d'exposition directement à partir des données. A ce jour, aucune étude n'a permis d'identifier des profils longitudinaux de consommation de tabac vie entière directement à partir des données d'étude cas-témoins et d'en comparer les risques de cancer du poumon associés.

0.3.3 Association entre l'exposition professionnelle à l'amiante et cancer du poumon

L'association entre l'exposition professionnelle à l'amiante et le risque de cancer du poumon a été montrée pour la première fois par Doll [1955] chez des travailleurs exposés à l'amiante. D'après l'expertise collective de Inserm [1997], en dessous d'un temps de latence de 10 ans, aucun excès de risque peut être observé. Dans une revue critique de la littérature (Henderson et al. [2004]), il a été montré que le temps de latence était égal ou supérieur à 10 ans, pouvant même aller jusqu'à 44 ans. Alors qu'une autre revue de la littérature (Prazakova et al. [2014]) a retrouvé un temps de latence plutôt entre 15 à 20 ans.

La relation entre la dose cumulée d'exposition à l'amiante et le cancer du poumon est avérée depuis de nombreuses années et tous s'accordent à dire qu'il n'existe pas de seuil en dessous duquel l'exposition à l'amiante n'aurait aucun risque sur le cancer du poumon (Nielsen et al. [2014]). Cependant, dans la littérature récente, la forme de la relation est encore sujette à controverse. Certains l'ont considéré comme linéaire (Prazakova et al. [2014], Gustavsson et al. [2002]). En revanche, van der Bij et al. [2013] l'ont considéré non linéaire en utilisant des splines pour modéliser l'indice cumulé d'exposition (RR=1.028, 95% IC : 1.019-1.038) pour 4 f/mL-années et RR=1.322, 95 % IC : 1.208-1.446) pour 40 f/mL-années par rapport à un indice cumulé nul). D'après leur récente revue systématique de la littérature, Nielsen et al. [2014] ont retrouvé que la forme de la relation était approximativement linéaire mais elle pouvait se stabiliser pour des expositions élevées ($>$ 150 f/mL-années).

L'impact de l'arrêt de l'exposition professionnelle à l'amiante sur le risque de cancer du poumon a pu être étudié. Récemment, Nielsen et al. [2014] ont montré une diminution du risque après un arrêt de 7 à 15 ans voire jusqu'à plus de 40 ans pour certaines des études de la revue systématique. En revanche, une étude réalisée en Italie (Magnani et al. [2008]) a montré une diminution du risque relatif chez les sujets non exposés depuis 15 à 30 ans par rapport à ceux non exposés depuis 3 à 15 ans (RR= 0.7, 95% IC : 0.52-0.95)). Ils ont relevé une diminution du risque encore plus importante du risque relatif entre les sujets non exposés depuis plus de 30 ans et ceux non exposés depuis moins de 15 ans (RR=0.56, 95% IC : 0.35-0.92). Le rôle de l'âge à la première exposition reste assez peu étudié dans la littérature. Une récente revue systématique publiée par Kang et al. [2013] a montré que sur les 3 articles étudiant l'impact de l'âge à la première exposition professionnelle à l'amiante sur le risque de cancer du poumon (Chen et al. [2012], Pira et al. [2007], Luce et al. [2000]), aucun n'avait retrouvé d'association.

Dans la plupart des études sur cette relation, l'intensité d'exposition est représentée par un indice cumulé d'exposition. A ce jour, il existe peu d'études considérant l'effet de l'intensité au cours de la vie alors que certaines questions se posent encore. Pour tenter de répondre à cela,

Hauptmann et al. [2002] ont étudié l'effet de l'exposition (représentée par la durée d'exposition ou le nombre de fibres-années) en chaque temps mais en ne considérant pas les 5 dernières années avant le diagnostic. En effet, certaines études ne considèrent pas les expositions récentes, allant même jusqu'à 30 ans avant le diagnostic (Loomis et al. [2009]). Cela pose donc l'hypothèse que les intensités reçues peu de temps avant le diagnostic n'ont aucun impact sur le survenue ou le développement du cancer du poumon. Or, à ce jour, il n'existe aucune donnée probante permettant de vérifier cette hypothèse, la preuve en est, les études ne s'accordent pas toutes quant au nombre d'années d'exposition avant le diagnostic à ne pas considérer.

De plus, à notre connaissance, aucune étude n'a étudié l'association entre des profils de trajectoires d'intensité d'exposition professionnelle à l'amiante et le risque de cancer du poumon. Pourtant, nous pouvons faire l'hypothèse raisonnable que différents profils d'exposition existent dans les données et il est envisageable que ces profils d'exposition conduisent à des risques différents.

0.4 Objectifs de la thèse

Pour les 3 relations (amiante-MP, amiante-cancer du poumon, tabac-cancer du poumon), les questions épidémiologiques auxquelles nous voulions répondre étaient de deux types :

1. A quel moment de l'histoire de l'exposition, ou à quel âge, une augmentation de l'intensité de l'exposition est la plus délétère ? Les intensités d'exposition reçues les quelques années précédant le diagnostic contribuent-elles au risque de cancer ?

2. Existe-t-il différents profils de trajectoires d'intensité d'exposition et quelles sont leur association avec le risque de cancer ?

L'objectif général de la thèse était donc d'étudier l'association entre l'intensité du tabagisme et de l'exposition professionnelle à l'amiante, et le risque de MP et de cancer du poumon, à l'aide de méthodes statistiques permettant 1) de prendre en compte la variation de l'intensité au cours de la vie et d'estimer son effet, et 2) d'identifier les différents profils de trajectoires d'intensité d'exposition vie entière et comparer leur risque de cancer associé.

Le premier type de question épidémiologique a été abordé en estimant le poids relatif de l'intensité d'exposition à chaque âge et en chaque année de l'histoire d'exposition sur le risque de cancer grâce à un indice cumulé d'exposition existant (OBJECTIF 1, chapitre 2, articles 1 et 2). Le deuxième type de question sur les profils de trajectoires d'intensité d'exposition et les risques de cancer associés a été abordé en développant de nouveaux modèles mixtes à classes latentes (OBJECTIF 2, chapitres 3 et 4, article 3).

0.5 Plan du manuscrit

Le chapitre 1 concerne la description des données utilisées dans le cadre de cette thèse provenant de deux études cas-témoins françaises. Le chapitre 2 présente le premier objectif de thèse en incluant une description de la méthode statistique envisagée, les résultats obtenus avec les articles 1 & 2 sur les trois relations d'intérêt et une conclusion générale. Le chapitre 3 détaille le second objectif de thèse, concernant l'identification de profils de trajectoires d'exposition et la comparaison avec les risques de cancer associés, avec la méthode utilisée, l'article 3 présentant les résultats obtenus pour le cancer du poumon, des approches détaillées à titre de comparaison, ainsi qu'une discussion autour de l'application de tels modèles dans ce contexte épidémiologique. Le chapitre 4 met en avant le développement méthodologique envisagé pour mieux tenir compte de la particularité des données rétrospectives utilisées pour répondre au second objectif, il montre également les résultats préliminaires obtenus. Et enfin, une conclusion générale détaille les limites de ce travail de thèse tout en précisant quelques perspectives potentielles.

1 Chapitre 1 : Les données utilisées

Sommaire

1.1 Mésothéliome Pleural

1.1.1 Schéma d'étude

Les données utilisées pour l'étude de la relation entre les expositions professionnelles à l'amiante et le mésothéliome pleural (MP) proviennent d'une étude cas-témoins rétrospective en population générale constituée à partir de plusieurs sources.

Elle s'appuie sur des données d'enquêtes déjà réalisées provenant des sources suivantes :

— d'une étude cas-témoins réalisée entre 1987 et 1993 dont les détails se trouvent dans Iwatsubo et al. [1998]. L'objectif de cette étude était principalement de documenter la relation entre l'exposition professionnelle à l'amiante et le mésothéliome pleural. Elle a été menée en population hospitalière dont le recueil de cas incidents de MP fut réalisé entre 1987 et 1993 dans 5 régions françaises (Ile de France, Provence Alpes Côte d'Azur, Auvergne, Lorraine, Corse)

— du PNSM qui enregistre de manière exhaustive et permanente les cas incidents de mésothéliome pleural en population générale dans 25 départements français depuis le 1er janvier 1998 (voir figure . Ce programme de surveillance a précédemment été décrit plus en détail en introduction, en référence à Goldberg et al. [2006]

— des échantillons d'histoires professionnelles constitués par le DST-InVS de sujets interrogés soit en population générale en 2007 ou soit en tant que témoins dans quinze études épidémiologiques françaises conduites entre 1984 et 2000 (Goldberg et al. [2000])

L'objectif de cette étude était de permettre de documenter davantage la relation entre le mésothéliome pleural et les expositions professionnelles (à l'amiante mais aussi à d'autres facteurs comme les laines minérales par exemple). Plus précisément, en approfondissant la modélisation de la relation entre l'exposition à l'amiante et le mésothéliome pleural mais aussi en étudiant l'effet des co-expositions à d'autres fibres ou particules minérales (Lacourt [2010]).

1.1.2 Sélection des cas

Les cas sont issus de deux sources d'informations différentes et respectent les critères d'éligibilité détaillés ci-après pour chaque source. Nous allons distinguer les deux sources en nommant les cas A pour la première et les cas B provenant de la seconde :

— Les cas A :

- Cas issus du recueil passif de cas incidents de mésothéliome pleural réalisé dans le cadre de l'étude cas témoins conduite entre 1987-1993 (Iwatsubo et al. [1998])
- Domiciliation dans une des 5 régions françaises incluses dans l'étude (voir figure 1.1 (gauche))
- Consultation dans un des services hospitaliers participant
- Diagnostic posé durant la période d'étude (1/01/1987-31/12/1993) et histologiquement confirmé par les experts du panel national Mesopath
- Sujet vivant au moment du recrutement permettant de réaliser une enquête étiologique auprès du sujet lui-même (pas de proxy utilisé)

— Les cas B :

- Cas issus du PNSM
- Domiciliation dans un des 25 départements français inclus dans le programme (voir figure 1.1 (droite))
- Diagnostic posé entre le 01/01/1998 et 31/12/06 avec confirmation anatomopathologique par les experts du panel national Mesopath ou confirmation clinique par un groupe d'experts cliniques
- Sujet vivant au moment du recrutement permettant de réaliser une enquête étiologique auprès du sujet lui-même (pas de proxy utilisé)

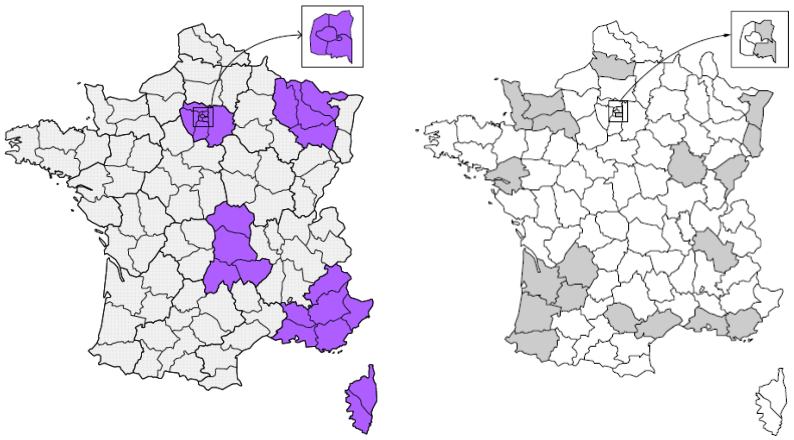

FIGURE 1.1 – Zone géographique de recrutement des cas A (à gauche) et des cas B (à droite). *Les zones colorées représentent les départements concernés.*

1.1.3 Sélection des témoins

Deux témoins par cas ont été tirés au sort aléatoirement à partir d'un échantillon d'histoires professionnelles constitué par le DST-InVS en 2007 (témoins A), et appariés en fréquence sur le sexe et l'année de naissance (± 5 ans). L'échantillon d'histoires professionnelles est composé de la carrière professionnelle de 10010 sujets représentatifs de la population générale française âgée de 25 à 74 ans en 2007. Les sujets de cet échantillon ont été sélectionnés par la méthode des quotas à partir de numéros de téléphone générés aléatoirement. Cette méthode permet de s'assurer de la représentativité d'un échantillon en utilisant des variables de stratification, ici, la région, le sexe, l'âge et la catégorie socio-professionnelle (Lacourt [2010]).

Néanmoins en 2007, le nombre de sujets suffisamment âgés était insuffisant pour satisfaire le critère d'appariement en fréquence sur l'année de naissance. Ainsi, un second échantillon d'histoires professionnelles a donc été utilisé pour compléter la sélection des témoins (témoins B). Ce second échantillon d'histoires professionnelles est composé de 8344 calendriers professionnels de témoins issus de 15 études épidémiologiques réalisées entre 1984 et 2006 dans la quasi-totalité des régions de la France. Dans Goldberg et al. [2000], la période d'interview et les zones géographiques associées à chacune des études considérées étaient détaillés dans le tableau 1 de l'article. Un extrait de ce tableau se trouve dans la figure 1.2 ci-après. Entre l'article publié et l'échantillon final constitué, il y a eu une augmentation de la taille d'échantillon. A notre connaissance, aucun article plus récent n'a été publié.

Author (reference)	Number and type of subjects included in the sample	Period of interview	Region
Boffetta et al, 1998 (16)	20 referents	1988—1994	Mainly Paris area
Clavel et al, 1995 (17)	649 cases and referents	1990—1992	18 different cities
Cordier et al, 1993 (18)	1375 cases and referents	1984—1987	6 different cities
Hours et al, 1996 (19)	255 cases and referents	1991—1993	7 different cities
Hours et al, 1994 (20)	289 cases and referents	1984—1987	Lyon area
Hours et al, 1991 (21)	459 referents	1984—1990	Lyon area
Iwatsubo et al, 1998 (9)	316 referents	1987—1995	5 different regions
Luce et al, 1993 (22)	320 referents	1986—1988	10 different cities
Richardson et al, 1992 (23)	361 cases and referents	1984—1988	Paris area
Stengel et al, 1995 (24)	362 cases and referents	1985—1990	Paris area
Stucker et al, 1995 (25)	294 referents	1990—1992	Paris and Besançon

FIGURE 1.2 – Caractéristiques des études épidémiologiques considérées pour la sélection des temoins B. *Tableau tiré de Goldberg et al. [2000]*

1.1.4 Recueil d'information

Les questionnaires standardisés, qui ont permis le recueil d'information, étaient différents entre les cas et les témoins ainsi que d'une source à l'autre. Cependant, quelle que soit la provenance des données, l'histoire professionnelle complète a été reconstituée au moyen d'un

entretien avec un enquêteur formé, en face à face pour les cas et par téléphone pour les témoins.

1.1.5 Calendrier professionnel

La carrière professionnelle complète correspondant à tous les emplois occupés au cours de la vie a été reconstituée grâce aux questionnaires standardisés. Dans le cadre de cette étude, un emploi est décrit par la combinaison d'un secteur d'activité et d'une profession.

Pour chaque emploi de plus de 6 mois, les informations suivantes ont été reportées :

- date de début et de fin de l'emploi occupé
- le secteur d'activité codé selon la Classification internationale type, par industrie, de toutes les branches d'activité économique (Citi édition 1975) et/ou la Nomenclature des activités françaises (Naf édition 1999)
- la profession codée selon la Classification Internationale Type des Professions (CITP édition 1968) et la classification des Professions et Catégories Socioprofessionnelles (PCS édition 1994)

1.2 Cancer du poumon

1.2.1 Schéma d'étude

ICARE (Investigation of occupational and environmental CAuses of REspiratory cancers) est une étude cas-témoins française multicentrique basée en population générale sur les cancers respiratoires. Elle inclut un groupe de cas de cancer du poumon et un groupe de cas de cancer des voies aérodigestives supérieures (regroupant les sites tels que pharynx, larynx,...) avec un groupe de témoins commun à tous les sites de cancer (Luce and Stücker [2011]).

Les objectifs de cette étude étaient 1) d'investiguer les facteurs de risque professionnels pour les cancers du poumon et des voies aérodigestives supérieures ; 2) d'identifier de nouveaux carcinogènes liés au milieu professionnel ; 3) d'étudier les effets conjoints entre les différents facteurs de risque professionnels ; 4) d'étudier les interactions entre les expositions professionnelles et les facteurs de risque avérés pour les sites de cancer (tabac, alcool) et 5) d'étudier les interactions entre les facteurs de risque génétiques et les carcinogènes liés au milieu professionnel.

Les cas incidents ont été identifiés avec la collaboration de 10 registres français du cancer. Ces registres couvrent 10 départements français qui représentent environ 13% de la population française (voir figure 1.3). Les distributions des emplois et des caractéristiques socio-démographiques de la population active de ces départements sont similaires à celle de la France. Le groupe commun de témoins est un échantillon aléatoire de la population générale de ces départements.

FIGURE 1.3 – Zone géographique de recrutement des cas de l'étude ICARE.
Figure tirée de Luce and Stücker [2011]

1.2.2 Sélection des cas

Les critères d'inclusion des cas étaient les suivants :

— tout patient résident dans l'un des départements français diagnostiqué entre 2001 et 2007 d'une tumeur maligne primaire de la cavité orale, pharynx, cavité sinonasal, bronches et poumon (Classification Internationale des Maladies (CIM),10ème version, C00-C14, C30-C34)

— cas histologiquement confirmés âgés de plus de 18 ans et de moins de 75 ans au diagnostic. Tous les types histologiques étaient inclus

Les cas ont été identifiés à partir de laboratoires anatomopathologiques et des hôpitaux du département. Une procédure a été mise en place pour réduire le temps entre le diagnostic et l'interview. A chaque cas identifié, le dossier médical était consulté afin de vérifier les critères d'inclusion et de recueillir le nom et l'adresse du médecin traitant. Après avoir recueilli l'accord du médecin traitant pour contacter les cas, une lettre pour informer les cas de l'étude leur été adressée, puis un rendez-vous était fixé pour un entretien en face-à face avec un enquêteur formé, à l'hôpital ou au domicile des patients selon leur préférence.

1.2.3 Sélection des témoins

Les témoins ont été sélectionnés par procédure de tirage au sort sur les listes téléphoniques dans les mêmes départements que les cas. Pour tenter d'accéder aux numéros sur la liste rouge, au dernier numéro des numéros tirés au sort était ajouté +1. Le recrutement s'est fait par téléphone par un institut de sondage expérimenté dans ce type de procédure par des enquêteurs formés. Ces appels téléphoniques ont été générés automatiquement par un système, qui permettait également de définir le nombre d'appels à un numéro donné et le moment. Chaque numéro était appelé 10 fois avant d'être abandonné. Les appels étaient effectués le soir en semaine et

la journée pour le samedi afin que chaque numéro ait la même probabilité d'être contacté.

Les témoins ont été appariés en fréquence aux cas suivant le sexe, l'âge (<40,40-54,55-64,>=65) et les départements. Une stratification supplémentaire a été réalisée pour que la distribution par statut socio-économique parmi les témoins soit comparable à celle de la population générale.

Les vagues de recrutement avaient lieu tous les 2 mois. Le nombre de vagues de recrutement était basé sur le nombre de cas estimés et le taux de participation attendu parmi les témoins. Les périodes de recrutement étaient différentes d'un département à l'autre, comme nous pouvons le voir dans la figure 1.4 ci-après.

	Lung cancer cases			Controls		
Study center	**N**	**%**	**Period of diagnosis**	**N**	**%**	**Recruitment period**
Bas Rhin	398	13.6	06.2004-09.2006	469	13.2	07.2004-09.2007
Calvados	354	12.1	03.2003-09.2006	462	13.0	04.2003-09.2007
Doubs	149	5.1	05.2004-06.2006	143	4.0	06.2004-09.2006
Haut Rhin	73	2.5	11.2004-09.2006	118	3.3	12.2004-01.2007
Hérault	334	11.4	08.2001-01.2006	450	12.7	09.2001-01.2007
Isère	476	16.3	08.2001-01.2006	501	14.1	09.2001-01.2007
Loire Atlantique	350	12.0	08.2004-09.2006	404	11.4	09.2004-09.2007
Manche	320	10.9	03.2003-09.2006	312	8.8	04.2003-09.2007
Somme	321	11.0	04.2002-06.2006	499	14.0	05.2002-01.2007
Vendée	151	5.2	08.2004-09.2006	197	5.5	09.2004-09.2007
Total	2926	100.0		3555	100.0	

FIGURE 1.4 – Périodes de diagnostic pour les cas et périodes de recrutement pour les témoins suivant les départements de résidence.
Tableau tiré de Luce and Stücker [2011]

1.2.4 Recueil d'information

Les sujets ont été interviewés en face-à-face grâce à des questionnaires standardisés qui ont permis de recueillir les caractéristiques socio-démographiques (sexe, âge, statut marital, niveau d'éducation, emploi des parents et du conjoint, le pays de naissance, le pays de naissance des parents) ; l'histoire résidentielle (les différentes adresses de résidence depuis la naissance) ; des informations relatives au mode de vie (tabac, alcool, ...) ; les caractéristiques anthropométriques (taille, poids à différents moments comme à l'interview, 2 ans avant l'interview et à 30 ans) ; les antécédants médicaux personnels et familiaux. Des informations cliniques concernant les cas par rapport à leur maladie ont également été recueillies par les registres de cancer. A la fin de l'entretien, un prélèvement salivaire était systématiquement proposé au sujet pour en obtenir son ADN.

Quand le sujet était trop malade pour répondre au questionnaire complet, une version simplifiée du questionnaire était administrée auprès du sujet lui-même ou auprès d'un proche qui répondait à sa place.

1.2.5 Tabac

L'historique de consommation de tabac vie entière a été recueilli. La consommation de cigarettes, cigares, cigarillos et pipe était détaillée en reportant pour chaque période de consommation homogène de tabac :

— les dates de début et de fin

— la quantité par jour (nombre moyen de cigarettes fumées par jour). A savoir que la quantité pour tous les autres types de consommations que la cigarette a été convertie en équivalent grammes de tabac afin d'obtenir une équivalence en nombre de cigarettes

— le type de cigarette (blonde ou brune, filtre ou non, marque, la façon d'inhaler, cigarettes roulées à la main ou non)

Tout nouvel épisode était synonyme d'un changement d'une de ces caractéristiques dans la consommation de tabac par rapport à l'épisode précédent. Les périodes d'arrêt d'au moins 1 an étaient également prises en compte dans l'historique de tabac.

Pour chaque sujet, nous pouvons donc lui attribuer une intensité annuelle moyenne journalière de tabac qui correspond au nombre moyen de cigarettes fumées par jour sur l'année.

1.2.6 Calendrier professionnel

L'histoire professionnelle complète a été documentée. Pour chaque emploi occupé au moins un mois, le nom et les activités de l'entreprise, la description des tâches (nature et fréquence du matériel utilisé), la description de l'environnement de travail étaient reportées. Des questionnaires spécifiques ont également pu être administrés pour détailler des tâches ou emplois fréquemment rencontrés (peinture, isolation, industrie du verre, construction, industrie chimique, soudage, travail de cuir...). Chaque emploi a été codé, suivant la CITP 1968 et la NAF 2000 à l'insu du statut cas-témoin du sujet, et par des personnes ayant suivi une formation dans le but d'homogénéiser et de standardiser le codage.

1.3 Évaluation de l'exposition professionnelle à l'amiante

1.3.1 Généralités sur une matrice emploi exposition

Une matrice emploi exposition (MEE) a été utilisée pour évaluer l'exposition professionnelle à l'amiante. Une MEE est un tableau qui va faire correspondre à chaque emploi, un certain nombre de paramètres d'expositions à la nuisance pour laquelle elle est spécifique.

Un emploi définit un groupe homogène d'exposition qui peut être caractérisé par l'intitulé de la profession seulement ou par la combinaison d'une profession et d'un secteur d'activité codé au moyen de nomenclatures standardisées. L'évaluation de l'exposition est plus précise puisque pour un même intitulé de profession, l'exposition peut varier d'un secteur à l'autre.

La MEE s'applique sur le calendrier professionnel des sujets à condition que les classifications de professions et d'activités soient les mêmes que celles utilisées dans la matrice. Ainsi, pour chaque emploi occupé au cours de la carrière professionnelle, l'exposition à la nuisance considérée peut être évaluée selon les divers paramètres fournis par la matrice. Ces paramètres d'expositions peuvent être simplement le statut exposé vs non exposé ou des paramètres semi quantitatifs comme la probabilité, la fréquence et l'intensité d'exposition.

De plus, pour tenir compte de la variabilité des expositions des emplois au cours du temps, notamment, du fait de l'évolution des techniques et des réglementations, les MEE peuvent être historisées. Ainsi, un même emploi va donc avoir une évaluation de l'exposition différente selon la période pendant laquelle il a été occupé, on parlera alors d'emploi-période.

1.3.2 La matrice employée

Pour évaluer l'exposition professionnelle à l'amiante des sujets inclus dans les deux études cas-témoins, une matrice emploi exposition relative aux fibres d'amiante a été utilisée. Cette matrice se veut exhaustive ; l'ensemble des couples professions-secteurs d'activité exposés à la nuisance en question est évalué. Cette évaluation a été réalisée par consensus d'experts.

Pour chaque emploi-période qui est défini selon une profession (CITP 1968), un secteur d'activité (Naf 2000) et une période historique, la MEE évalue l'exposition à l'amiante selon cinq paramètres :

1) la probabilité d'exposition (P_e) correspondant au pourcentage estimé des travailleurs exposés à l'amiante (de 0 pour aucune exposition jusqu'à 0.85 pour une exposition certaine)

2) la fréquence d'exposition liée aux tâches spécifiques (F_s) qui correspond à la proportion du temps moyen dédié à ces tâches spécifiques par rapport à l'ensemble du temps de travail au cours d'une journée de travail typique de 8 h (de 0.025 pour une exposition sporadique à 0.85 pour une exposition en continue)

3) la fréquence d'exposition liée à l'ambiance de travail ou « exposition de fond » (F_a) qui correspond à la proportion du temps moyen en présence de cette exposition de fond par rapport à l'ensemble du temps de travail au cours d'une journée de travail typique de 8 h (de 0.025 pour une exposition sporadique à 0.85 pour une exposition en continue)

4) l'intensité d'exposition liée aux tâches spécifiques (I_s) qui évalue la concentration moyenne journalière sur l'année à laquelle est soumis le sujet lors de ces tâches spécifiques (de 0.005 éq f/mL à 20 éq f/mL)

5) l'intensité d'exposition liée à l'ambiance de travail (I_a) qui évalue la concentration moyenne journalière sur l'année à laquelle est soumis le sujet à cette exposition de fond. Cette intensité est distinguée suivant deux types d'exposition ; exposition passive où le travailleur est exposé du fait de la contamination diffuse des locaux (de 0.0005 éq f/mL pour exposition très faible à 2 éq f/mL pour une exposition très forte) et une exposition indirecte où le travailleur est exposé via d'autres personnes qui manipule la substance (de 0.0025 éq f/mL pour exposition très faible à 10 éq f/mL pour une exposition très forte)

Chacun de ces paramètres étant évalué de manière semi-quantitative, il est habituel d'attribuer à chaque catégorie des valeurs numériques permettant par la suite de calculer divers indices d'exposition. Pour cette matrice amiante, la valeur correspondante au centre de l'intervalle a été attribuée à chaque catégorie. Le tableau 1.1 ci-après détaille de manière plus précise toutes les valeurs attribuées au sein des différentes classes définissant ces paramètres.

Paramètres d'expositions	Définitions	Valeurs numériques		
Probabilité d'exposition	P (%)	P_e		
0 : non exposé	0	0		
1 : peu possible	]0 – 5]	0,025		
2 : possible	]5 – 30]	0,175		
3 : probable	]30 – 70]	0,5		
4 : très probable	> 70	0,85		
Fréquence d'exposition	F (%) des jours travaillés	F_s & F_a		
0 : non exposé	0	0		
1 : sporadique	]0 – 5]	0,025		
2 : occasionnelle	]5 – 30]	0,175		
3 : fréquente	]30 – 70]	0,5		
4 : permanente	> 70	0,85		
Intensité d'exposition	I en éq f/mL	Type d'exposition		
		Exposition d'ambiance I_a		Exposition spécifique I_s
		Passive	Indirecte	Directe
0 : non exposé	0	0	0	0
1 : très faible	]0 - 0,01]	0,0005	0,0025	0,005
2 : faible	]0,01 – 0,1]	0,005	0,025	0,05
3 : moyen	]0,1 – 1]	0,05	0,25	0,5
4 : forte	]1 – 10]	0,5	2,5	5
5 : très forte	> 10	2	10	20

Tableau 1.1 – Définitions des paramètres d'exposition, matrice amiante "version 2007"

Pour tenir compte de l'évolution de la réglementation concernant l'usage de l'amiante et la protection des travailleurs en France, la MEE permet de distinguer l'évaluation de l'exposition suivant plusieurs périodes importantes.

Cinq périodes ont été retenues : 1945-1977, 1978-1997, 1998-2000, 2001-2005 et après 2005.

— 1945 est la date à laquelle la matrice commence à retracer les expositions survenues

— 1977 est la date à laquelle est définie une première valeur limite d'exposition qui va permettre un usage de l'amiante contrôlé et qui va diminuer à partir de cette date

— entre 1977 et 1997, les valeurs limites vont diminuer

— 1997 est la date qui signe l'interdiction totale de l'usage de l'amiante

— 2000 est une date qui permet de tenir compte du retrait progressif des fibres d'amiantes dans les installations et bâtiments

La matrice a été créée en 2007, ainsi au-delà de 2005 aucune évolution n'a pu être prise en compte. La seule exploitation minière d'amiante en France a été fermée en 1965, cette date apparait également pour le secteur de l'extraction.

A partir des paramètres précédemment définis (tableau 1.1), le niveau d'exposition (L) moyen sur une année donnée peut être défini comme le produit de l'intensité (I_s & I_a), de la fréquence (F_s & F_a) et de la probabilité (P_e) (équation 2). Ce niveau d'exposition L est exprimé en éq f/mL :

$$L = P_e \times [F_s \times I_s + F_a \times I_a] \tag{2}$$

La matrice permet donc d'attribuer un niveau moyen annuel pour tous les emplois occupés par le sujet. Lorsqu'un sujet occupait plusieurs emplois la même année, un niveau moyen d'exposition sur l'année au prorata de la durée de chacun des emplois occupés cette année-là a été calculé. Dans la suite de ce manuscrit pour plus de clarté, ce niveau d'exposition annuel sera appelé "intensité annuelle moyenne journalière de l'exposition à l'amiante" (en éq f/mL).

Dans la suite de ce manuscrit, un sujet a été considéré comme exposé s'il possède une probabilité non nulle d'être exposé à l'amiante durant un emploi, ce qui se traduit donc par au moins une intensité annuelle moyenne non nulle durant son histoire d'exposition.

1.4 Sélection de la population d'étude

Pour pouvoir mettre en place les méthodes statistiques envisagées dans le cadre de cette thèse, une sélection des sujets dans les deux études a été nécessaire.

Quelque soit la provenance des données (mésothéliome pleural ou cancer du poumon), seuls les hommes ont été considérés dans nos analyses. En effet, pour la relation entre le tabac et le cancer du poumon, les hommes et les femmes ne peuvent pas être analysés ensemble du fait de risques de cancer du poumon associé à la consommation de tabac potentiellement différents entre les deux sexes (Papadopoulos et al. [2014]). Pour l'exposition professionnelle à l'amiante, la MEE ne permet pas de repérer de façon précise les expositions chez les femmes qui sont majoritairement para-professionnelles. Ainsi, nous n'aurions pas eu la même qualité d'information par rapport à l'évaluation de l'exposition professionnelle entre les hommes et les femmes. De

plus, les effectifs de femmes étant assez limités, la mise en oeuvre des méthodes statistiques envisagées n'aurait pas forcément été possible.

De par l'objectif principal de la thèse qui se concentre sur la modélisation de l'intensité d'exposition au cours de la vie, seuls les sujets avec des histoires d'exposition complètes (reportées et codées) ont été considérés. Compte tenu de la complexité des méthodes statistiques utilisées dans le cadre de cette thèse, il était difficilement envisageable de mettre en oeuvre des techniques d'imputation multiple pour prendre en compte les sujets avec une histoire d'exposition incomplète. La figure 1.5 représente la sélection des sujets pour l'étude sur le mésothéliome pleural. La figure 1.6 représente la sélection des sujets pour l'étude ICARE.

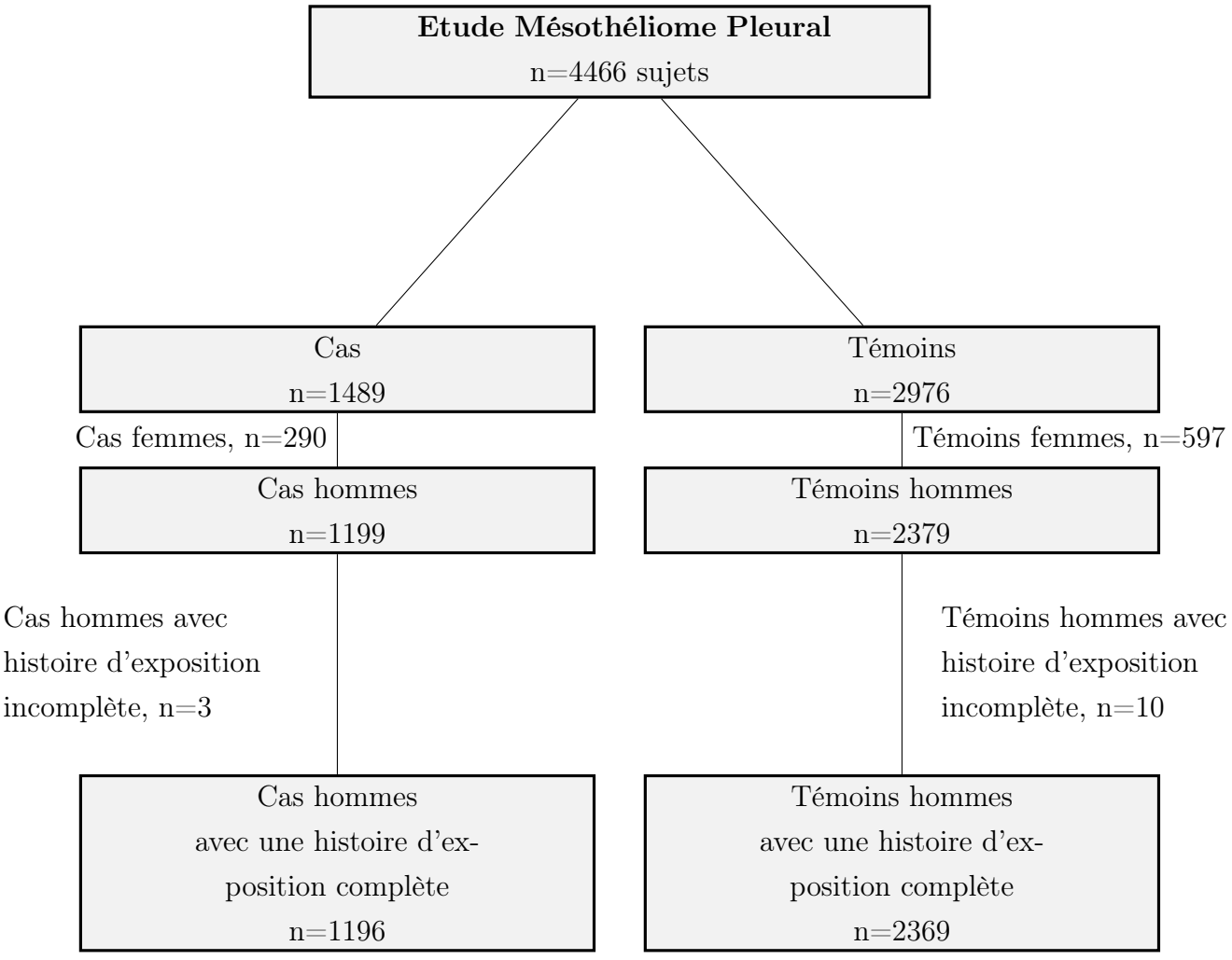

FIGURE 1.5 – Sélection des sujets pour l'étude Mésothéliome Pleural

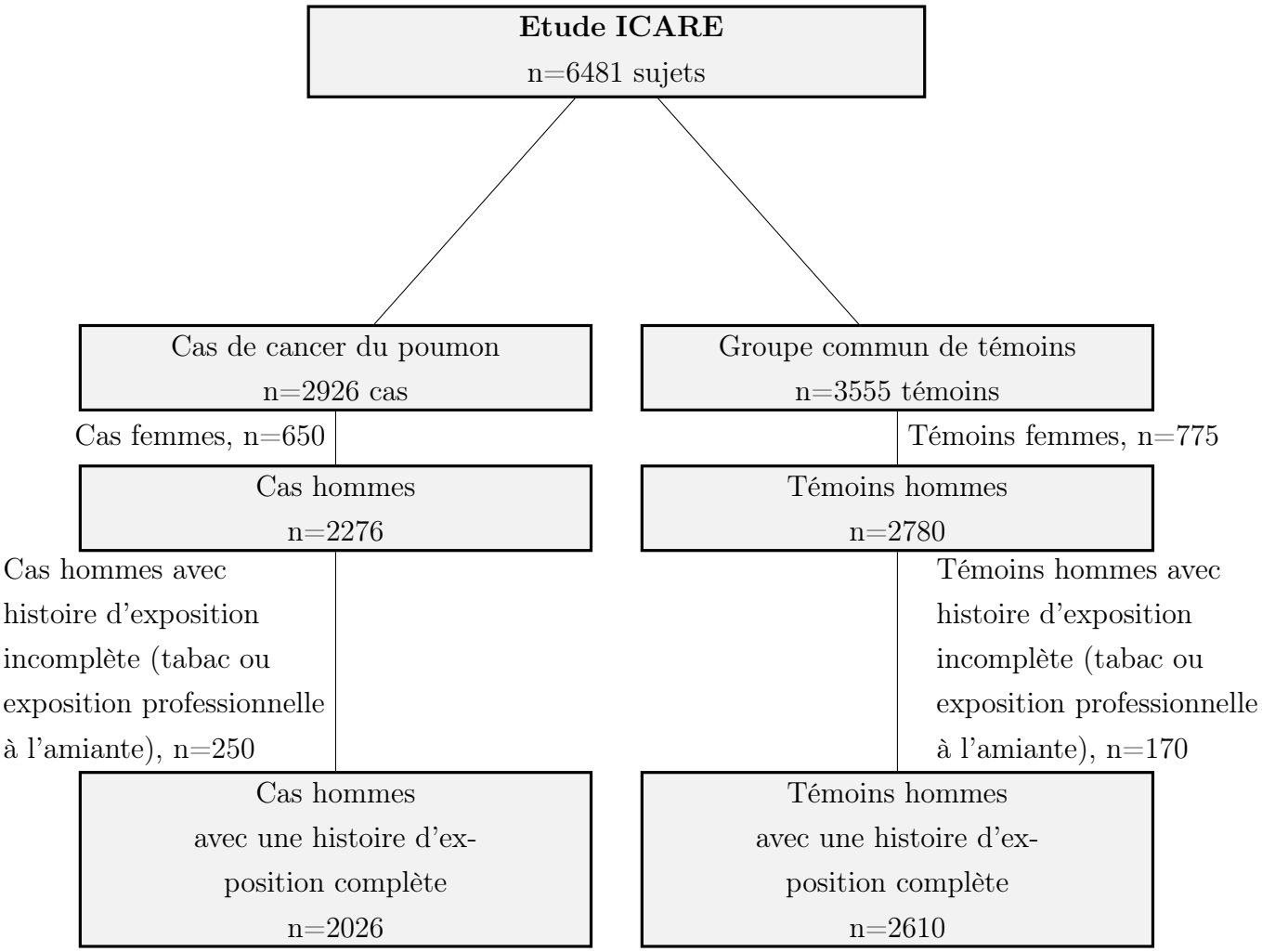

FIGURE 1.6 – Sélection des sujets pour l'étude cas-témoins ICARE, 2001-2007

2 Chapitre 2 : Estimation de l'effet de l'intensité en chaque temps de l'histoire d'exposition (WCIE)

Sommaire

2.1 Les indices cumulés d'exposition

Pour estimer la relation entre des expositions prolongées et le risque de cancer, on représente l'exposition des sujets dans le modèle de régression. Comme nous avons pu le voir en introduction, classiquement dans les études cas-témoins, elle est représentée au travers d'un indice qui résume l'exposition vie entière nommé indice cumulé d'exposition (ICE) qui représente la somme de toutes les intensités d'exposition reçues dans le passé. Cet indice comporte un certain nombre de limites. Notamment, en plus de ne pas dissocier l'effet de l'intensité de celui de la durée, cet indice cumulé ne prend pas en compte la variation de l'intensité au cours du temps ni le moment où cette intensité a été reçue. Cependant, dans le cadre de ce premier travail de thèse, on s'intéresse à l'effet de l'intensité reçue au cours du temps pour répondre aux questions suivantes :

— Quel est le poids relatif des expositions récentes ou anciennes ?
— L'effet de l'intensité de l'exposition dépend il de l'âge auquel cette intensité est reçue ?

Alors que l'ICE ne permet pas de répondre à ces questions, plusieurs approches ont été proposées pour considérer l'effet de l'intensité au cours du temps. Parmi elles, une approche consiste à découper la période d'exposition en différentes fenêtres de temps d'exposition (Finkelstein [1991]). C'est une approche simple à mettre en oeuvre qui permet d'étudier la relation dose-réponse dans le temps. Pour estimer plus précisément l'effet des intensités reçues au cours du temps, une pondération de l'indice cumulé d'exposition peut être utilisée, avec des hypothèses a priori plus ou moins fortes sur la forme de la fonction de poids incluse dans cet indice (Hauptmann et al. [2000b], Sylvestre and Abrahamowicz [2009], Langholz et al. [1999]). Enfin, une approche hiérarchique a été proposée par Richardson et al. [2011] dont le principe est de coupler une analyse par période d'exposition (similaire à la première approche) à une fonction de poids. Bien qu'elle permette d'obtenir des estimations précises, l'interprétation en terme d'excès de risque relatif est moins aisée par rapport à des rapports de côte classiques. Dans la suite de cette section, nous allons détailler l'approche proposée par Finkelstein [1991] et celle de la pondération de l'indice cumulé détaillé par Hauptmann et al. [2000b]. Ces deux approches paraissaient potentiellement les plus adaptées pour répondre au premier objectif de thèse.

2.1.1 Approche par fenêtres de temps d'exposition

L'approche appelée "Time-Windows" et développée par Finkelstein [1991] permet de considérer l'aspect temporel dans la relation dose-réponse. Dans la suite de ce manuscrit, elle sera nommée approche par fenêtres de temps d'exposition. Cette approche consiste à découper la période d'exposition en plusieurs fenêtres d'exposition définies à priori. La définition de ces fenêtres d'exposition peut fortement influencer les résultats épidémiologiques obtenus. Il est préférable de les définir à partir de connaissances de la relation étudiée. L'effet de l'exposition est estimé dans chacune de ces fenêtres de temps. Au sein d'une même fenêtre de temps, le risque

est considéré constant. Dans un modèle de régression logistique pour données cas-témoins, chacune de ces fenêtres de temps d'exposition est représentée par les intensités cumulées au cours du temps définissant ces fenêtres. Les exponentiels des paramètres associés à ces fenêtres de temps correspondent aux rapports de côte (ou Odd Ratio) pour l'augmentation d'une unité de cette intensité cumulée reçue dans cette fenêtre de temps.

Prenons un exemple fictif pour expliquer l'application de cette approche dans le cadre de données provenant d'une étude cas-témoins.
Considérons le temps avant la date index comme axe du temps t et Y le statut cas-témoin (Y=1 si cas, 0 si témoin). Supposons que pour chaque sujet i, on ait des intensités annuelles d'exposition $(x_i(t))$ de la date index jusqu'à 50 ans avant cette date, on choisit de découper l'axe du temps en 5 fenêtres de temps d'exposition de 10 ans chacune :

$$\text{logit}P(Y_i = 1|x_i(t)) = \beta_0 + \beta_1 \sum_{t=1}^{10} x_i(t) + \beta_2 \sum_{t=11}^{20} x_i(t) + \beta_3 \sum_{t=21}^{30} x_i(t) + \beta_4 \sum_{t=31}^{40} x_i(t) + \beta_5 \sum_{t=41}^{50} x_i(t) \tag{3}$$

L'interprétation des coefficients de ce modèle (équation 3) est la suivante : $\exp(\beta_1)$ est l'OR associé à l'augmentation d'une unité de l'intensité cumulée dans les 10 ans précédant la date index, ajusté sur l'exposition cumulée dans les autres fenêtres de temps. Grâce aux différents ORs estimés, il est possible d'identifier la période d'exposition où l'effet de l'intensité est maximum sur le risque de cancer.

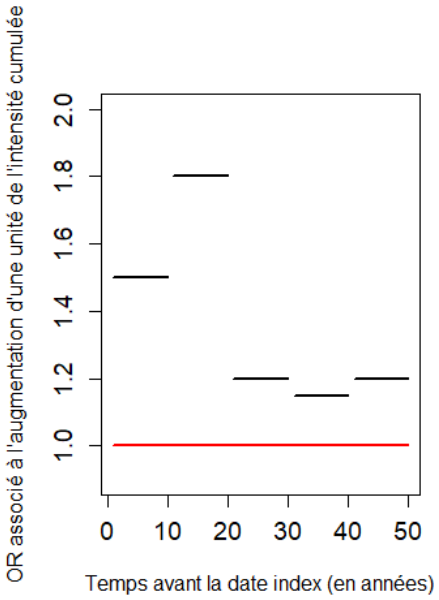

FIGURE 2.1 – Approche par fenêtres de temps d'exposition-Exemple fictif

De plus, ces ORs peuvent être représentés sur un graphe similaire à la figure 2.1 où en abscisse sera représenté l'axe du temps et en ordonnée les OR ou le ln OR. D'après la figure 2.1, on peut dire que ce sont les expositions reçues durant la deuxième fenêtre de temps d'exposition (11-20 ans avant la date index) qui contribuent le plus au risque de cancer. Pour évaluer l'incertitude de ces estimations, il est possible de calculer les intervalles de confiance à 95 % associés aux β_j, j=1...,5 par $(\exp(\hat{\beta}_j \pm 1.96\sqrt{var(\hat{\beta}_j)}))$. Grâce à ces ICs, on peut conclure si l'effet estimé est différent d'une fenêtre d'exposition à l'autre.

Cette approche a l'avantage d'être facile à mettre en œuvre, facile en terme d'interprétation des résultats et permet d'étudier l'effet de l'intensité à différentes périodes de temps de l'histoire d'exposition. Cependant, au travers de cet exemple, de nombreux inconvénients peuvent être mis en évidence :

— Les résultats sont très sensibles au nombre et aux bornes de ces fenêtres de temps d'exposition choisis à priori

— La forme en escalier obtenue semble peu réaliste pour la plupart des relations
Rappelons que cette approche considère que l'effet des expositions reçues au sein d'une même fenêtre de temps est constant sur le risque de cancer. Ceci semble peu réaliste de pouvoir considérer que l'effet évolue radicalement d'une année sur l'autre (dernière année de la fenêtre /première année de l'autre) avant d'être de nouveau constant tout au long de la fenêtre

— La corrélation entre les intensités cumulées des fenêtres de temps d'exposition adjacentes. Il faut éviter de choisir un nombre trop important de fenêtres au risque de créer de l'instabilité et de l'imprécision dans les estimations.

En conclusion, l'approche par fenêtres de temps d'exposition comporte de nombreuses limites mais elle permet de donner une idée générale de la forme de la relation dose-réponse dans le temps, ce qui n'est pas possible avec l'ICE classique.

2.1.2 Pondération de l'indice cumulé

L'ICE fait l'hypothèse que les intensités d'exposition reçues ont le même poids quelque soit le moment au cours de l'histoire d'exposition où elles sont reçues. Dans l'approche par fenêtres de temps d'exposition, même si elles peuvent avoir un poids différent d'une fenêtre d'exposition à l'autre, on a également l'hypothèse que les intensités reçues ont le même poids au sein d'une même fenêtre.

Pour relâcher cette hypothèse, il est possible de définir une fonction de poids dépendante du temps qui permet ainsi d'attribuer des poids différents à chaque intensité d'exposition reçue en fonction du moment au cours de l'histoire d'exposition où elle a été reçue. En faisant la somme de chaque intensité d'exposition pondérée, cela conduit à définir un indice cumulé pondéré de l'exposition qui sera noté WCIE dans la suite de ce manuscrit pour "Weighted Cumulative Index of Exposure". Il a été introduit par Vacek [1997] puis développé ensuite par plusieurs auteurs (Langholz et al. [1999], Hauptmann et al. [2000b], Richardson and Ashmore [2005], Sylvestre and Abrahamowicz [2009]).

Pour un sujet i,

$$WCIE_i = \sum_t w(t)x_i(t) \tag{4}$$

où

— $x_i(t)$ représente l'intensité annuelle moyenne journalière reçue l'année t pour le sujet i ;

— $w(t)$ représente la fonction de poids dépendante du temps. Elle peut être paramétrée de différentes manières. On peut utiliser des fonctions de poids dont la forme est définie a priori en utilisant des fonctions paramétriques connues (Langholz et al. [1999]) ou dont la forme est définie à partir des données grâce à des fonctions splines cubiques (Hauptmann et al. [2000b], Sylvestre and Abrahamowicz [2009], Richardson and Ashmore [2005]). Il est à noter que si, par exemple, $w(t)$=1 en tout temps t, cela revient à calculer l'ICE classique défini précédemment.

Comme nous souhaitions estimer l'effet de l'intensité d'exposition en chaque temps de l'histoire de l'exposition sur le risque de cancer, nous avons donc considéré l'utilisation d'un indice cumulé pondéré par une fonction de poids dépendante du temps pour répondre au premier objectif de la thèse. Afin de n'imposer aucune hypothèse a priori sur la forme de la fonction de poids, nous avons privilégié l'utilisation d'une fonction de poids avec des splines cubiques. La méthode est détaillée dans la section suivante. Nous avons considéré un tel indice pour les 3 relations d'intérêt dans cette thèse ; amiante-MP (article 1), amiante-cancer du poumon (article 2) et tabac-cancer du poumon (article 2). Nous allons expliciter les 3 applications avant d'insérer les articles 1 et 2 incluant les principaux résultats de ce premier travail de thèse. Il y aura également des discussions complémentaires aux articles avant de clore le chapitre par une conclusion générale résumant les principaux résultats de ce premier travail et détaillant deux des limites avec ses perspectives associées.

2.2 La méthode envisagée

L'intérêt du WCIE pour évaluer au mieux l'effet de l'intensité au cours du temps repose donc sur la caractérisation de la fonction de poids qui lui est associée. Hauptmann et al. [2000b] ont initialement proposé d'estimer cette fonction de poids à partir des fonctions splines cubiques. Cette approche a ensuite été reprise et développée pour des données de cohorte par Sylvestre and Abrahamowicz [2009]. Dans le cadre de cette thèse, nous avons donc considéré l'approche de Hauptmann et al. [2000b].

2.2.1 La fonction de poids considérée

Définition

La fonction de poids flexible dépendante du temps s'écrit comme suit :

$$w(t) = \sum_{j=1}^{m+4} \theta_j B_j(t) \tag{5}$$

où

— m est le nombre de noeuds intérieurs ;
— B_j représentent les fonctions de base Bsplines associées aux différents temps t ;
— θ_j sont les coefficients à estimer.

En effet, une fonction spline cubique est une combinaison linéaire des fonctions de base BSplines. Chacune de ces fonctions de base est associée à un coefficient à estimer. Ces fonctions de base sont définies à partir de polynômes (ici de degré 3) et de la position des noeuds (Atkinson [1991]). Par définition, elles assument des conditions de continuité, mettant en jeu des égalités sur les dérivées premières et secondes entre les noeuds ou en chaque noeud. De manière générale, le nombre de noeuds et la nature de la fonction spline utilisée déterminent le nombre de paramètres à estimer pour évaluer une telle fonction. Pour éviter des instabilités aux extrémités de l'axe du temps t considéré, Hauptmann et al. [2000b] a utilisé des splines naturelles cubiques (appelées également splines cubiques restreintes). De manière plus spécifique, la fonction de poids flexible considérée dans le cadre de nos applications sera détaillée en section 2.3.3.

A partir des équations (4) et (5), on peut réécrire l'indice cumulé pondéré (WCIE) :

$$WCIE_i = \sum_t \sum_{j=1}^{m+4} \theta_j B_j(t) x_i(t) \tag{6}$$

où $x_i(t)$ est l'intensité d'exposition reçue au temps t.

L'axe du temps

Pour répondre aux objectifs fixés, deux axes du temps ont été considérés :

i) un axe suivant le temps avant la date index en années, qui est une échelle rétrospective en accord avec le design des études cas-témoins ;

ii) un axe suivant l'âge courant.

Pour l'axe i), le WCIE peut être défini entre la date index (T_{min}=0) et le nombre d'années maximum avant la date index (T_{max}) observé parmi les sujets de la population d'étude :

$$WCIE_i = \sum_{t=T_{min}}^{T_{max}} w(t)x_i(t) \tag{7}$$

Pour l'axe ii), le WCIE est défini entre un âge minimum (Age_{min}) et un âge maximum (Age_{max}) :

$$WCIE_i = \sum_{t=Age_{min}}^{Age_{max}} w(t)x_i(t) \tag{8}$$

Nous avons fixé un minimum et un maximum par rapport à des contraintes techniques, à partir des données utilisées et non a priori. En effet, suivant les effectifs en chacun des temps définissant les extrémités de l'axe, nous avons dû réduire l'étendue de l'axe en ne considérant que des temps pour lesquels nous avions assez d'information afin d'assurer la convergence du modèle.

2.2.2 Le modèle de régression

Le WCIE est inclus tel quel dans un modèle de régression logistique quand les données proviennent d'une étude cas-témoins.
En utilisant l'équation (6), le modèle logistique s'écrit :

$$\text{logit}P(Y_i = 1|x_i, Z_i) = \ln \frac{P(Y_i|x_i, Z_i)}{1 - P(Y_i|x_i, Z_i)} = \beta_0 + \beta_1 \sum_{t=tmin}^{tmax} \sum_{j=1}^{m+4} \theta_j B_j(t) x_i(t) + \lambda Z_i \tag{9}$$

où

- Y_i est le statut cas/témoin du sujet i (1 pour cas, 0 pour témoin) ;
- Z_i est le vecteur des variables d'ajustement pour le sujet i ;
- λ est le vecteur des paramètres associés aux variables d'ajustement ;
- β_1 est le paramètre associé au WCIE ;
- $tmin$ et $tmax$ représentent l'étendue de l'axe du temps considéré. Pour le temps avant la date index, on peut aller de 0 (date index) à 60 ans avant. Pour l'âge courant, on peut estimer l'effet de l'intensité reçue de 12 ans à 65 ans par exemple.

Interprétation

L'effet global du WCIE est évalué par $\hat{\beta}_1$, qui représente le ln(OR) associé à l'augmentation d'une unité de l'indice cumulé pondéré d'exposition.

Il est également possible d'interpréter $\hat{\beta}_1\hat{w}(t)$. Avec les estimations de $\hat{\theta}_j$, il est possible de reconstruire $\hat{w}(t)$ en chaque temps t de l'axe du temps considéré ($\sum_j \hat{\theta}_j B_j(t)$).
Pour un temps k donné, l'exponentiel de $\hat{\beta}_1\hat{w}(k)$ peut être interprété comme l'OR associé à l'augmentation d'une unité de l'intensité d'exposition reçue au temps k, ajusté sur toutes les intensités d'exposition reçues aux autres temps. Cet OR associé à l'exponentiel de $\hat{\beta}_1\hat{w}(k)$ compare donc deux sujets qui ont seulement une différence d'une unité d'intensité d'exposition au temps k. Il y a donc un ajustement implicite sur l'histoire d'exposition. Illustrons ceci avec un exemple de deux sujets fictifs A et B, ayant des intensités d'exposition égales en tout temps sauf au temps k.

Sujet A, au temps k, x(k)=2 :

$$\begin{aligned} logit_{SujA} &= \text{logit}(P(Y=1|x(k)=0.2) \\ &= \beta_0 + \beta_1[w(0)*0 + w(1)*1 + \dots w(k)*2 + \dots w(T_{max})*1] \end{aligned}$$

Sujet B, au temps k, x(k)=3 :

$$\begin{aligned} logit_{SujB} &= \text{logit}(P(Y=1|x(k)=0.3) \\ &= \beta_0 + \beta_1[w(0)*0 + w(1)*1 + \dots w(k)*3 + \dots w(T_{max})*1] \end{aligned}$$

Par définition, un ln OR est calculé en faisant la différence entre les deux modèles.

$$\ln OR = (logit_{SujB} - logit_{SujA}) \tag{10}$$

Ainsi pour calculer l'OR associé à l'augmentation d'une unité d'intensité au temps k,

$$OR = \exp(\beta_1 w(k))(3-2)) \tag{11}$$

Pour calculer l'effet de l'intensité (via l'OR) en ce temps précis k, tout doit donc être similaire chez les deux sujets aux autres temps définissant l'axe du temps. Concrètement, ils ont des histoires d'exposition similaires sauf au temps k évalué. Ainsi avec le WCIE, il n'est pas nécessaire d'ajuster sur d'autres variables d'exposition telles que le temps depuis la première exposition ou la durée totale d'exposition. On s'affranchit donc de toute la problématique autour de la multicolinéarité entre ces variables, que l'on a pu détailler en introduction de ce manuscrit.

2.2.3 Estimation du modèle

Par définition des fonctions splines, l'estimation des poids $\hat{\theta}_j$ se fait à partir des données. La méthode d'estimation du modèle doit alors permettre d'estimer en même temps $\hat{\beta}_1$ et $\hat{\theta}_j$. Cette méthode est la vraisemblance sous contraintes où l'on détermine des conditions à travers des inégalités/égalités pour assurer l'identifiabilité du modèle (Hauptmann et al. [2000b]). Elle est plus complexe à mettre en place qu'une méthode de maximisation de vraisemblance classique implémentée dans n'importe quel logiciel statistique. Il existe une alternative plus simple pour l'estimation de ce modèle en utilisant une maximisation de vraisemblance linéaire grâce l'adaptation proposée par Sylvestre and Abrahamowicz [2009]. Cependant, elle ne permet pas d'obtenir les estimations séparées de $\hat{\beta}_1$ et $\hat{\theta}_j$. Nous avons donc utilisé la méthode d'estimation proposée par Hauptmann et al. [2000b] dans ce travail de thèse.

2.2.3.1 La vraisemblance

L'écriture analytique de la log vraisemblance (L) du modèle de régression logistique est :

$$L = \log l = \prod_{i=1}^{n_{tot}} Y_i logitP - \log(1 + \exp(logitP)) \tag{12}$$

où

- $logitP$ est le modèle de régression logistique défini par l'équation (9) ;
- n_{tot} est le nombre total de sujets.

2.2.3.2 Les contraintes fixées

Deux contraintes suffiront pour assurer l'identifiabilité du modèle :

- Une contrainte de positivité : les paramètres θ doivent être positifs ou non nuls : $\theta_j \geq 0$;
- Une contrainte de standardisation : $\sum_{t=1}^{T_{max}} w(t) = T_{max}$ où $T_{max} = \max_i T_i$.

  De manière pratique, on soustrait à la log vraisemblance le terme $\sum_{t=1}^{T_{max}} w(t) - T_{max}$, pour que ce terme soit le plus petit possible (voire soit nul) afin de maximiser au mieux la log vraisemblance définie par l'équation 12.

2.2.3.3 Calcul des intervalles de confiance

Les contraintes fixées pour estimer le modèle ont pour conséquence que les paramètres estimés $\hat{\theta}$ sont proches des frontières de l'espace restreint (supérieur ou égal à 0). On ne peut donc pas appliquer la théorie asymptotique pour calculer les intervalles de confiance de la fonction de poids. Ainsi, la technique du bootstrap (ou rééchantillonage) a été utilisée (Efron and Tibshirani [1994]).

Concrètement, B nouvelles bases de données ont été créées à partir du rééchantillonage de la population d'étude, en conservant la même proportion de cas et de témoins. Le modèle a été estimé à partir de ces B bases permettant d'obtenir B estimations pour β_1 et θ_j, entre autres. Après un tri par ordre croissant, les $2.5^{ème}$ et $97.5^{ème}$ percentile de la distribution des B estimations de chaque paramètre β_1 et θ_j sont utilisés pour reconstruire point par point l'IC de $\hat{\beta}_1\hat{w}(t)$ au niveau de confiance de 95%.

2.2.4 Test d'hypothèse

Pour évaluer l'apport de la modélisation de l'exposition par le WCIE par rapport à l'ICE classique pour les relations entre les expositions prolongées et le risque de cancer, un test de rapport de vraisemblance peut être réalisé en posant une hypothèse nulle permettant de considérer que les modèles sont réellement emboités.

Soit Z_p^T le vecteur des variables d'ajustement des deux modèles d'intérêt, la statistique de test peut être démontré comme suit :

Modèle logistique avec le WCIE :

$$logitP(Y_i|x_i, Z) = \beta_0 + \beta_1 \sum_t w(t)x_i(t) + \beta_p Z_p^T \tag{13}$$

Modèle logistique avec un ICE classique :

$$logitP(Y_i|x_i, Z) = \beta_0 + \beta_1 \sum_t x_i(t) + \beta_p Z_p^T \tag{14}$$

Pour que les modèles soient considérés emboités, il faut que

$$w(t) = 1 \forall t \Leftrightarrow \sum_{j=1}^{m+4} \theta_j B_j(t) = 1 \forall t$$

Or, on rappelle que les Bsplines ont une des propriétés suivantes :

$$\sum_{j=1}^{m+4} B_j(t) = 1$$

Ainsi, l'hypothèse nulle de ce test :

$$H0 = \theta_1 = \ldots = \theta_{m+4} = 1 \tag{15}$$

La statistique de test associé est :

$$(-2\hat{L}_{ICE}) - (-2\hat{L}_{WCIE}) \sim \chi_2(m+3) \tag{16}$$

où

— $\hat{L}_{ICE}$ est la log vraisemblance du modèle avec l'ICE ;
— $\hat{L}_{WCIE}$ est la log vraisemblance du modèle avec le WCIE ;
— $m+3$ est le nombre de degrés de liberté qui correspond à la différence du nombre de paramètres entre les deux modèles.

2.3 Applications aux données des deux études cas-témoins

Les deux études cas-témoins étant des études cas-témoins avec un appariement en fréquence, l'effet du WCIE a été estimé à partir d'un modèle de régression logistique non conditionnel ajusté sur les variables d'appariement et les facteurs de confusion potentiels.

2.3.1 Relation entre l'amiante et le MP (article 1)

L'effet de l'augmentation d'une unité de l'intensité de l'exposition professionnelle à l'amiante sur le risque de MP a été estimé en chaque année avant la date index puis en chaque âge avec le WCIE à partir des données provenant de l'étude cas-témoins PNSM présentée au chapitre 1. Les risques de MP associés à différents profils longitudinaux hypothétiques d'exposition à l'amiante ont ensuite été comparés.

2.3.1.1 Variables d'appariement

Les témoins ont été appariés en fréquence aux cas sur l'âge à la date index et la cohorte de naissance. Nous avons ajusté sur l'âge à la date index en le modélisant par des splines cubiques naturelles pour tenir compte de l'effet non linéaire de l'âge sur le logit, avec 3 noeuds placés aux $5^{ème}$, $50^{ème}$ et $95^{ème}$ percentiles de la distribution de l'âge de la population d'étude (54, 66, 79 ans). La variable "cohorte de naissance" était découpée selon les catégories suivantes :$<$ 1920, 1920- 1924, 1925-1929, 1930-1934, 1935-1939, 1940-1944, 1945-1949, $>$ 1949. La catégorie de référence est 1930-1934 car elle est la catégorie la plus représentée chez les témoins (22%).

2.3.1.2 Pour aller au-delà de la fonction de poids estimée

Au-delà d'estimer l'effet de l'intensité d'exposition en chaque temps de l'histoire d'exposition avec le WCIE, on peut ensuite comparer le risque de cancer d'un profil hypothétique de trajectoires d'intensité d'exposition par rapport à un autre (article 1).

Pour pouvoir comparer de tels profils avec l'ICE classique, il faut que ces profils aient des intensités cumulées différentes alors qu'avec le WCIE, on peut comparer des profils hypothétiques différents menant à une même valeur cumulée de l'ICE classique.

Prenons un exemple avec deux profils d'exposition professionnelle à l'amiante suivant l'axe du temps avant la date index menant à un ICE classique de 120 f/mL-années (figure 2.2) :
◇ Profil L :
2.5 f/mL de 0 à 20 ans avant la date index
0 f/mL de 21 à 40 ans avant la date index
3.5 f/mL de 41 à 60 ans avant la date index

◇ Profil C :
2 f/mL de 0 à 60 ans avant la date index

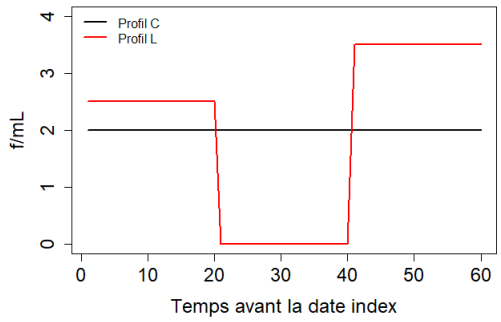

FIGURE 2.2 – Profils hypothétiques de trajectoires d'intensités d'exposition professionnelle à l'amiante

L'OR comparant la côte de mésothéliome pleural pour un sujet ayant le Profil L par rapport à celui qui a un Profil C se calcule comme :

$$\begin{aligned} OR =& \exp((2.5-2)[\hat{w}(0)+...+\hat{w}(20)]+(0-2)[\hat{w}(21)+...+\hat{w}(40)] \\ &+(3.5-2)[\hat{w}(41)+...+\hat{w}(60)]) \\ =& \exp((2.5-2)[\sum_j \hat{\theta}_j[B_j(0)+...+B_j(20)]]+(0-2)[\sum_j \hat{\theta}_j[B_j(21)+...+B_j(40)]] \\ &+(3.5-2)[\sum_j \hat{\theta}_j[B_j(41)+...+B_j(60)]]) \\ \neq& 1 \end{aligned}$$

Cet OR permet d'évaluer l'effet des intensités reçues à plus ou moins fortes intensités durant quelques années par rapport à des intensités reçues de manière constante tout au long de l'histoire d'exposition. En utilisant un ICE classique, l'OR entre ces deux profils est égal à 1 (=exp(120 − 120)) puisqu'il ne prend pas en compte quand a été cumulée l'exposition.

2.3.2 Cancer du poumon (article 2)

L'objectif du deuxième article de thèse était d'estimer l'impact de l'intensité de l'exposition à l'amiante ou au tabac sur le risque de cancer du poumon en chaque année avant la date index et en fonction de l'âge en utilisant le WCIE à partir des données provenant de l'étude cas-témoins ICARE (détail au chapitre 1).

2.3.2.1 Facteurs d'appariement

Dans l'étude ICARE, on rappelle que les témoins ont été appariés en fréquence aux cas sur l'âge à la date index et le département de résidence. Nous avons ajusté sur l'âge à la date index en utilisant des splines cubiques naturelles avec 3 noeuds placés aux $5^{ème}$, $50^{ème}$ et $95^{ème}$ percentiles de la distribution de l'âge de la population d'étude (42, 60, 73 ans). Il y a 10 départements de résidence : Calvados, Hérault, Isère, Loire-Atlantique, Manche, Bas-Rhin, Haut-Rhin, Somme, Vendée, Doubs-Territoire de Belfort. Nous avons donc ajusté sur 9 d'entre eux en prenant le département le plus représenté chez les témoins comme département de référence (l'Isère).

2.3.2.2 Facteurs de confusion

2.3.2.2.a Pour la relation tabac-cancer du poumon

Pour la relation entre le tabac (modélisé par le WCIE) et le cancer du poumon, un ajustement supplémentaire sur l'histoire de l'exposition professionnelle à l'amiante a été considéré. Il est représenté par l'ICE classique dont la formule est la suivante :

$$ICE_i = \sum_t x_i(t) \tag{17}$$

où $x_i(t)$ est l'intensité annuelle moyenne journalière en éq f/mL reçue l'année t par le sujet i.

L'effet de l'ICE n'était pas linéaire sur le logit de la probabilité de survenue du cancer du poumon, la variable a donc été catégorisée suivant les quartiles de la population des témoins (0, 0 - 0.3346, 0.3347 - 3.74, >3.74 f/mL-années).

2.3.2.2.b Pour la relation amiante-cancer du poumon

Pour la relation entre l'exposition professionnelle à l'amiante (modélisé par le WCIE) et le cancer du poumon, un ajustement supplémentaire sur l'histoire de l'exposition au tabac a été considéré. Il est représenté par le Comprehensive Smoking Index (CSI), précédemment défini dans l'introduction. Pour rappel, cet indice cumulé agrégé, combinant l'intensité (int), la durée totale d'exposition (dur) et le temps depuis l'arrêt (tsc) de manière paramétrique, s'écrit comme suit (Leffondré et al. [2006]) :

$$\text{CSI} = (1 - 0.5^{dur^\star/\tau})(0.5^{tsc^\star/\tau})\ln(int + 1) \tag{18}$$

où

- $tsc^\star$=max(tsc-δ,0) ;
- $dur^\star$=max(dur+tsc - δ,0)-$tsc^\star$;
- τ est un paramètre de forme et δ est un paramètre de décalage

Il a été mis en évidence que cet indice ajuste mieux que l'indice cumulé d'exposition classique représenté par le nombre de paquets-années par exemple, et notamment sur l'étude ICARE (Papadopoulos et al. [2014]) et sur deux études cas-témoins montréalaises (Leffondré et al. [2006]). Concernant les paramètres de forme et de décalage, nous avons utilisé les valeurs estimées par profil de vraisemblance sur les données de l'étude cas-témoins ICARE dans Papadopoulos et al. [2014]. Ainsi, le CSI a été calculé avec un τ de 26 ans et un δ de 1 an. Il a été inclus tel quel dans le modèle de régression.

2.3.2.3 Pour aller au-delà de la fonction de poids estimée

Avec le WCIE, on peut calculer un OR associé à l'augmentation d'une unité de l'intensité reçue à un temps k sur le risque de cancer avec $\exp(\hat{\beta}_1\hat{w}(k))$ (vu en section 2.2.2). Pour considérer l'effet dépendant du temps de l'intensité d'exposition, on peut également calculer l'OR associé à une différence constante annuelle d'intensité sur une période de temps spécifique, toutes les intensités étant égales sur les autres périodes (article 2). Il faut définir les périodes de temps sur lesquelles on veut évaluer cet impact sur le risque de cancer ainsi que la différence d'intensité pour laquelle on veut calculer l'OR associé.

Prenons un exemple avec le tabac, pour évaluer l'impact d'une différence constante annuelle de 10 cigarettes-années dans les 10 années précédant la date index, on calcule l'OR associé comme suit :

$$OR = \exp(10\hat{\beta}_1 \sum_{t=0}^{t=10} \sum_j \hat{\theta}_j B_j(t))$$

Pour calculer les IC associés, nous avons utilisé les valeurs des paramètres obtenues lors du rééchantillonage.

2.3.3 La fonction de poids flexible

Dans les trois applications, nous avons utilisé des splines naturelles cubiques de régression qui imposent des conditions en leurs extrémités. Les extrémités sont alors considérées comme des nœuds. On définit comme nœuds intérieurs, ceux qui se situent entre ces deux extrémités. Lorsque l'on estime une spline cubique naturelle à 1 noeud intérieur, 5 coefficients θ_j sont à estimer. Un nombre important de noeuds augmente donc considérablement le nombre de paramètres à estimer. De plus, la courbe reconstruite, à partir des paramètres associés aux fonctions de base Bsplines, risque d'être très bruitée et ne plus approcher de manière lisse les données en raison d'un trop grand nombre de noeuds. Il faut éviter d'évaluer cette courbe au-delà de 3 noeuds intérieurs. Le nombre de noeuds est choisi en utilisant celui qui va minimiser le critère d'Akaike (AIC). Au-delà du nombre de noeuds à définir, il faut également choisir leurs positions. Hauptmann et al les ont positionnés par rapport aux percentiles de la distribution des différents temps considérés de la population. N'ayant aucune information biologique sur la forme plausible de la fonction de poids, nous avons décidé de les placer de manière équidistante par rapport l'axe du temps comme Sylvestre and Abrahamowicz [2009].

Pour assurer la convergence de ce modèle, il faut également avoir un effectif suffisant en chaque temps de l'axe considéré. C'est ainsi que l'on a défini T_{max}, Age_{min} et Age_{max} comme précisé précédemment. Par exemple, pour le cancer du poumon, nous avions décidé de considérer seulement les expositions avant 70 ans pour le tabac et avant 65 ans pour l'amiante (qui peut aussi se justifier par le contexte professionnel de l'exposition). De même pour les expositions suivant la date index, le nombre d'années maximum depuis la date index était défini à 60 ans. Ainsi, les données d'exposition, au-delà de ces temps fixés, n'étaient pas prises en compte dans l'estimation de la fonction de poids.

2.3.4 Mise en oeuvre sous SAS

Nous avons utilisé la procédure NLP sous SAS avec un algorithme d'optimisation Double Dogleg (DBLDOG) (Dennis et al. [1981], Gay [1983]). Cette procédure permet une minimisation (ou maximisation) d'une fonction non linéaire continue suivant plusieurs contraintes (qui peuvent être linéaires, non linéaires, égalités, inégalités ou des bornes inférieures/supérieures) en utilisant différents algorithmes d'optimisation possibles.

Le choix de l'algorithme est à faire suivant :

— les types de contraintes fixées ;

Certains algorithmes vont accepter seulement certains types de contraintes.

— le nombre de paramètres à estimer ;
Les algorithmes vont être plus ou moins efficients suivant le nombre de paramètres en jeu.

— la nécessité d'utiliser ou d'expliciter des équations analytiques des dérivées pour le calcul de la fonction à minimiser ou maximiser.

Le choix s'est porté sur l'algorithme DBLDOG (Dennis et al. [1981], Gay [1983]) qui semblait être le plus adapté à notre problème de maximisation. Il combine les idées d'une méthode quasi-Newton avec un algorithme à régions de confiance. Il permettait d'avoir des résultats moins sensibles aux valeurs initiales et aux maxima locaux que l'algorithme NMSIMP (Nelder and Mead [1965]), qui paraissait, de prime abord, plus adapté aux contraintes de bornes tout en n'utilisant aucune dérivée pour le calcul. Cependant, nous avons constaté que i) les résultats semblaient peu convaincants puisque les estimations étaient instables d'un jeu de valeurs initiales pour les paramètres à l'autre et ii) la convergence était plus lente à atteindre (à raison de deux heures versus 10 min pour le même modèle avec les mêmes valeurs initiales avec DBLDOG). Nous devions choisir un algorithme qui prenait un temps de calcul raisonnable avec une bonne précision puisque nous devions ensuite ré-estimer le modèle 1000 fois pour obtenir les ICs de la fonction de poids par bootstrap.

Cette procédure nécessite de définir :

— le nombre maximum d'itération de l'algorithme ($MAXITER$) ;
— le nombre d'appel de la fonction ($MAXFU$) ;
— préciser si c'est une maximisation/minimisation (MAX/MIN) ;
— la liste des paramètres à estimer ($PARMS$) ;
— les contraintes (BC : contraintes bornées ; LIC : contraintes linéaires ; NLC : contraintes non linéaires).

Propre à l'utilisation de l'approche du WCIE, au sein de la procédure SAS NLP, nous avons :

— calculé les indices cumulés pour chaque individu sur sa période d'exposition ;
— écrit l'équation du modèle de régression ;
— explicité la log vraisemblance individuelle.

En sortie de la proc NLP, le critère de convergence peut apparaitre comme satisfait alors qu'il n'y a pas d'estimation des variances associées à certains paramètres. Cela signifie que le critère de convergence ne se base pas sur la validité de l'estimation de la matrice hessienne (nécessaire à l'estimation de la variance des paramètres) pour considérer que la convergence soit réellement atteinte. Il nous semblait important d'avoir l'estimation de ces variances pour considérer la convergence atteinte. Ainsi, nous avons, de nouveau, relancé la procédure avec des valeurs initiales différentes afin d'assurer l'estimation des paramètres et de leurs variances

associées pour assumer que la convergence était atteinte. Pour le calcul des IC par méthode de bootstrap, nous avons seulement conservé les modèles ayant bien toutes les estimations (paramètres et variances) du modèle. Dans le fichier de résultats des modèles, lorsqu'il y a un problème d'estimation avec un des paramètres, cela se traduit par une valeur "ACTBC" pour la variable X_TYPE du paramètre en question.

2.4 Application au Mésothéliome Pleural

2.4.1 Article 1 publié dans Occupational and Environmental Medicine (OEM) : co-premier auteur avec Dr Aude Lacourt

ORIGINAL ARTICLE

Dose-time-response association between occupational asbestos exposure and pleural mesothelioma

Aude Lacourt,[1] Emilie Lévêque,[1,2] Elie Guichard,[2] Anabelle Gilg Soit Ilg,[3] Marie-Pierre Sylvestre,[4,5] Karen Leffondré[2]

[1]University Bordeaux, INSERM, Bordeaux Population Health Research Center, team EPICENE, UMR 1219, Bordeaux, France
[2]University Bordeaux, ISPED, INSERM, Bordeaux Population Health Research Center, team Biostatistics, UMR 1219, Bordeaux, France
[3]Santé publique France, French national public health agency, Saint-Maurice, France
[4]Department of Social and Preventive Medicine, Montreal School of Public Health, University of Montreal, Montreal, Canada
[5]Research Center, University of Montreal Health Center (CRCHUM), Montreal, Canada

Correspondence to
Dr Aude Lacourt, Equipe EPICENE cancer et environnement, UMR U1219 - Bordeaux Population Health Center, Université Bordeaux, Bordeaux 33076, France; aude.lacourt@inserm.fr

AL and EL contributed equally.
AL and EL are co-first authors.

Received 10 October 2016
Revised 7 February 2017
Accepted 24 March 2017
Published Online First
13 May 2017

ABSTRACT

Objectives Early occupational exposure to asbestos has been shown to be associated with an increased risk of pleural mesothelioma (PM), which suggests that the timing of exposure might play a role in the dose–response relationship. However, none studies has evaluated the relative impact of increasing the annual intensity of occupational exposure to asbestos at each time of the whole exposure history. Yet such evaluation would allow the comparison of the risks of PM associated with different longitudinal profiles of occupational exposure to asbestos. Our objective was to estimate the time-dependent relative impact of asbestos exposure intensity over the whole occupational history and to compare the resulting estimated risks of PM associated with different profiles of exposure, using data from a large French case–control study.

Methods This study included 1196 male cases recruited in 1987–2006 and 2369 matched controls on birth year. Occupational exposure to asbestos was assessed using a job exposure matrix and represented in logistic regression models using a flexible weighted cumulative index of exposure.

Results Due to much stronger weights of early doses of asbestos exposure, subjects who accumulated 20 fibres/mL over their entire job history with high doses during the first years and low doses thereafter were at higher risk of PM than those who accumulated most of the doses later (OR=2.37 (95% CI 2.01 to 2.87)).

Conclusion This study provides new insights on the dose-time-response relationship between occupational asbestos and PM and illustrates the importance of considering timing of exposure in its association with cancer risk.

What this paper adds

- The key question is to compare the risk of pleural mesothelioma associated with different longitudinal profiles of asbestos exposure.
- Subjects highly exposed in early career were at higher risk than those who were exposed at a constant level over the entire career, for the same final cumulative asbestos exposure.
- The estimation of the time-dependent effect of asbestos exposure may help deciders in planning epidemiological surveillance program of workers occupationally exposed to asbestos.

INTRODUCTION

The role of occupational asbestos exposure in the occurrence of pleural mesothelioma (PM) is well known.[1] However, in order to elaborate adequate epidemiological surveillance or compensation programs, it is essential to precisely estimate the shape of the dose-time-response relationship and to identify the longitudinal profiles of exposure at higher risk of PM.

The timing of occupational asbestos exposure is likely to play an important role on the risk of PM. For example, in our previous study, we have shown that the risk of PM tended to decrease with increasing age at first exposure,[2] which was also retrieved in some other studies.[3–5] This may suggest that an increase of the intensity of exposure during early career may be particularly deleterious. By contrast, the doses of exposure received before diagnosis are often assumed not to contribute to the risk of PM and are just discounted in the statistical analyses. Some investigators even go so far as to discount the 20 years of exposure before diagnosis.[6–8] Yet, it is difficult to fix such a time window *a priori* since latency for PM may depend on the level of asbestos exposures.[9] In addition, it has been shown that this lagging procedure may produce inconsistent and biased results.[10–12] Finally, discounting recent exposure does not allow identification of any modification of the risk by recent exposures.

To our knowledge, no previous study attempted to assess the relative impact of increasing annual intensity of occupational exposure to asbestos on the risk of PM at any time over the whole occupational history, from early career and young ages to late career and older ages, without imposing any prior assumption. Yet, such evaluation of the time-dependent effect of exposure intensity would allow the comparison of the risks of PM associated with different longitudinal profiles of occupational exposure to asbestos.

To avoid any assumption on the time-dependent effect of each dose of exposure received throughout the entire exposure history, Hauptmann *et al*[13] and Sylvestre and Abrahamowicz[14] proposed a flexible weighted cumulative index of exposure (WCIE). More specifically, each dose received at a specific time (year in cancer studies) is weighted by a specific weight at that time that is estimated from the data, using a spline function

To cite: Lacourt A, Lévêque E, Guichard E, *et al*. *Occup Environ Med* 2017;**74**:691–697.

of time which requires no prior assumptions. The estimated weight function allows the estimation of the effect of increasing the dose of exposure at a given time of the exposure history on the risk of cancer. It also allows the comparison between subjects with different longitudinal profiles of exposure over lifetime, whether they have accumulated the same amount of exposure at the index date or not.

Because the weight function is estimated using a spline function, the use of the WCIE does not require making any assumption on the possible mechanisms of carcinogenesis due to the exposure. It thus provides an interesting complementary approach to multistage models of carcinogenesis. Indeed, multistage models also provide interesting insights on the dose-time relationship,[15] and have been developed successfully in the context of PM and occupational exposure to asbestos in heavily exposed industrial cohorts.[16 17] For example, these models suggest that the incidence of PM would be proportional to the intensity and duration of exposure as well as to the third or fourth power of time since first exposure.[16] However, such modelling usually relies on a number of assumptions regarding the form of the dose-time-response relationship, which might not be appropriate for low doses.[18] While the WCIE has already been used in the context of occupational asbestos exposure and lung cancer,[13] to the best of our knowledge, it has never been used in the context of occupational asbestos exposure and PM.

The overall objective of this study was to estimate the dose-time-relationship between occupational asbestos exposure and the risk of PM using the WCIE and data from a French case–control study. More specifically, we estimated the effect of an increase in the annual dose of exposure to asbestos on the risk of PM at each year before the index date and at each year of age, and we compared the risk of PM associated with different longitudinal profiles of exposure to asbestos.

MATERIALS AND METHODS

Study design

This study is based on a previous pooled case–control study described elsewhere.[2] Briefly, cases were all incident males with histologically confirmed diagnosis of PM either in five French regions between January 1987 and December 1993,[6] or in 22 French districts between January 1998 and December 2006.[19]

Controls were randomly selected from two data sets of the French Institute for Public Health Surveillance (InVS), which included a complete job history from a structured interview of each subject. Sample A (4758 males and 5252 females) was selected from the French general population aged 25–74 years in 2007 by a quota sampling method stratified by sex, age, region and socioeconomic status. Sample B was made of 8344 controls from 15 population-based case–control studies conducted in nearly all regions of France in 1984–2000. Controls were randomly selected from samples A and B with frequency matching to cases on sex and birth-year within 5-year groups. In total, the present data set consisted in 1196 male cases and 2369 male controls frequency matched to cases by year of birth (within 5-yeargroups).

Data collection

A different but standardised questionnaire was administered to cases and controls by trained interviewers. Common information recorded for each subject was age at diagnosis/interview, year of birth and complete job history. For each job (held for at least 6 months), the occupation and the industry were coded according to international classifications.[20 21]

Asbestos exposure assessment

To estimate the annual average intensity of occupational exposure to asbestos (called 'annual dose of exposure' thereafter) over the entire job history for each subject, we used a job exposure matrix (JEM) that has been previously used and described.[2] Each job (which defined rows of the JEM) was defined as a combination of an occupation (defined according to the International Standard Classification of Occupation edition 1968) and an industry (defined according to either the International Standard industry Classification revision 2 or the Nomenclature of French activities classification edition 1999). Industrial hygienists and occupational health experts assigned for each job, an estimation of a probability, a frequency and an intensity of exposure to asbestos on a semi-quantitative scale (table 1). This expertise was based on a combination of knowledge of occupational asbestos exposure in France and direct measurement in occupational settings. The probability of exposure was defined as the percentage of workers exposed to asbestos for that job (from 0 for no exposure, to 0.85 for definite exposure). The frequency of exposure was defined as the proportion of exposed work time on a typical 8-hour working day for that job (from 0.025 for sporadic exposure, to 0.85 for continuous exposure). The intensity of exposure was defined as the average annual intensity of exposure for that job and was expressed in equivalent fibres/mL (from 0.0005 equivalent fibres/mL for very low intensity to 20 equivalent fibres/mL for very high intensity). Different values were assigned to both frequency and intensity of exposure depending on whether the exposure to asbestos was due to specific tasks or to work environment contamination.

For each subject, the dose of occupational exposure to asbestos received in a given year was defined as the product of the probability, frequency and intensity of exposure during that year. As a result, the dose received in a given year represented an average estimated intensity of occupational exposure to asbestos (in equivalent fibres/mL) over the entire year.

Statistical analysis

All statistical analyses were performed using unconditional logistic regression including year of birth (less than 1920, 1920–1924, 1925–1929, 1930–1934, 1935–1939, 1940–1944, 1945–1949, up to 1949) and age at diagnosis/interview (in years). Age at diagnosis/interview was modelled using natural cubic regression splines to account for its non-linear effect on the logit of PM, using three knots at 5th, 50th and 95th percentile of the age distribution of all subjects.[22]

Occupational asbestos exposure was represented in the model by either the WCIE,[14 23] or the standard unweighted cumulative index of exposure to asbestos (CIE) for comparison purposes. More specifically, the CIE was simply the sum of the doses of exposure $x(t)$ received each year t of the occupational history, that is,

$$CIE = \sum_{t} x(t)$$

The CIE assumes that each annual dose of exposure to asbestos, $x(t)$, has the same weight over all times t. If time t is the year before diagnosis, this assumes that the dose received just before diagnosis (eg, $t=5$ years before diagnosis) has the same weight than the dose received at a long distance from diagnosis (eg, $t=40$ years before diagnosis). If the time t is the age in years, then the CIE assumes that the dose of exposure received at young ages (eg, at age $t=20$ years) has the same weight than the dose received at later ages (eg, at age $t=60$ years). Thus, in the CIE calculation, the time axis t (year before diagnosis or year of

Table 1 Numerical values of probability, frequency and intensity of asbestos exposure used in the Job Exposure Matrix (JEM) to derive all individual annual doses of exposure

Asbestos exposure parameters	Definition	Numerical values used to calculate annual doses		
Probability of exposure (% of workers exposed)		P		
Non exposed	0	0		
Possible	>0–5	0.025		
Probable	5–30	0.175		
Likely	30–70	0.5		
Definite	≥70	0.85		
Frequency of exposure (% of work time)		Fs and Fa*		
Sporadic	>0–5	0.025		
Occasional	5–30	0.175		
Frequent	30–70	0.5		
Continuous	≥70	0.85		
Intensity of exposure (equivalent fibres/mL)		Type of exposure		
			Ia†	Is†
		Passive	Indirect	Direct
Very low	>0–0.01	0.0005	0.0025	0.005
Low	0.01–0.1	0.005	0.025	0.05
Medium	0.1–1	0.05	0.25	0.5
High	1–10	0.5	2.5	5
Very high	≥10	2	10	20

*Frequency of exposure was defined as a combination of frequency of exposure due to specific tasks (Fs) and due to work environment contamination (Fa). The same numerical values were used for Fs and Fa.
†Intensity of exposure was defined as a combination of the intensity of exposure due to specific task (Is) and the intensity of exposure due to work environment contamination (Ia). As the asbestos JEM is based on experts judgement, intensity of exposure is expressed in equivalent fibre/mL. Three types of exposure were defined: passive exposure (workers were exposed according to diffuse contamination of buildings); indirect exposure (workers were exposed by other workers using asbestos materials) and direct exposure (workers used directly asbestos materials).

age) does not matter since it assumes that the effect of an annual dose does not depend on the year t when the dose is received in the occupational history.

By contrast the WCIE allows different weights of the annual dose depending on the year t when the dose is received, that is:

$$WCIE = \sum_t w(t)\, x(t)$$

The particularity of the WCIE that was used in the present study is that the weights $w(t)$ were not fixed a priori, as opposed to the standard lagging approach which *a priori* assigns a weight of zero to recent exposures and of 1 to all others. The weights $w(t)$ were actually estimated from the data without imposing any assumption on their values or form. To this end, we specified a cubic Bspline function for the weight function $w(t)$ in the logistic regression models where we included the WCIE as a covariate, as proposed in the studies by Hauptmann *et al*[23] and Sylvestre and Abrahamowicz.[14] Such a spline function just requires choosing a number of interior knots (in addition to the two knots placed at the extreme of the time window) and their position on the time window. We used one to three interior knots placed at equal distance and selected the number of knots according to the Akaike's information criterion (AIC). We used two different time axes: the time t (in years) before diagnosis/interview, which ranged from 1 to 61 years, and the age t (in years) which ranged from 12 to 79 years. It is important to note that both the regression coefficient β associated with the WCIE and the weight function $w(t)$ in the WCIE were estimated simultaneously in a single step via a constrained maximum of likelihood estimation of the logistic regression model including the WCIE. The weights $w(t)$ were constrained to be non-negative and standardised, as described in the study by Hauptmann *et al*.[23] For that purpose, we used the SAS PROC NLP.[24] Bootstrap percentiles were used to compute 95% pointwise CIs of the weight function.

In the logistic regression model including the WCIE, the exponential of the regression coefficient β associated with the WCIE can be interpreted as the OR associated with an increase of 1 fibre/mL in the total cumulative dose of occupational exposure to asbestos, accounting for potential time-dependent effect of the annual dose increase:

$$OR_{\textit{per one f/ml increase in the total cumulative dose}} = \exp(\beta)$$

From the same model, the exponential of $\beta w(t)$ for a specific year t can be interpreted as the OR associated with an increase of 1 fibre/mL in the dose of exposure received in year t, adjusted for all the doses received in all other years:

$$OR_{\textit{per one f/ml increase in the dose received in year t}} = \exp(\beta w(t))$$

In other words, the OR above compares two subjects with a difference of 1 fibre/mL in the average intensity of exposure to asbestos in the specific year t, the annual average intensity being equal between the two subjects in all other years. The adjustment for the doses received in the other years is indeed implicit. As a result, there is no need to further adjust for the time at which exposure started or stopped, or the total duration of exposure, since the two subjects have the same doses over all their occupational history, except in year t. By contrast, the standard CIE which does not account for the timing of exposure requires further adjustment for the time window of exposure. To obtain fair comparison between the results of CIE and WCIE, we thus further adjusted the effect of CIE for time since first exposure (in years).

Estimated regression coefficients β of CIE and WCIE, as well as the estimated weight function *w(t)* were then used to derive the estimated OR between different hypothetical longitudinal profiles of quantitative exposure to asbestos. We have arbitrarily selected some hypothetical profiles of subjects who have continuously been exposed to asbestos at different levels, but any other plausible longitudinal patterns of exposure, including intermittent ones, could be considered for illustrative purposes.

RESULTS

Cases were in average slightly older at diagnosis than controls at interview (67.0 vs 66.5 years). More than 40% of cases and controls were born between 1925 and 1934. Among ever exposed subjects, cases were first occupationally exposed to asbestos at an average age of 21.1 (±7.1) years for an average total duration of 28.5 (±13.0) years compared with 22.6 (±8.0) years and 25.6 (±14.1) years for controls, respectively (table 2).

For the WCIE model, the best AIC was obtained with one interior knot for the weight function *w(t)*, whether we used time before diagnosis (interior knot at 31 years, AIC=3888.1) or age as the time axis (interior knot at 46 years, AIC=3885.1). The OR associated with an increase of 1 fibre/mL in the total cumulative dose of occupational asbestos exposure over the entire job history was: $OR_{\text{per 1 fibre/mL increase in the total cumulative dose}} = 1.055$ (95% CI 1.044 to 1.067) when using the WCIE with time before diagnosis/interview as the time axis, 1.030 (95% CI 1.025 to 1.037) when using the WCIE with age as the time axis, and 1.025 (95% CI 1.019 to 1.031) when using standard CIE.

According to the WCIE model, the doses of occupational asbestos exposure received more than 40 years before diagnosis/interview had a stronger impact on the risk of PM than more recent doses (figure 1A). For example, an increase of 1 fibre/mL in the dose received in the fiftieth years preceding diagnosis/interview was associated with an OR of 1.091 ($OR_{\text{per 1 fibre/mL increase in the dose received in year 50 before diagnosis/interview}} = 1.091$, 95% CI 1.073 to 1.111), while it was 1.026 if the same increase occurred in the second year before diagnosis/interview ($OR_{\text{per 1 fibre/mL increase in the dose received in year 2 before diagnosis/interview}}=1.026$, 95% CI 1.019 to 1.031), all other doses being constant. Interestingly, the doses of asbestos exposure received within the 10 years preceding diagnosis have a significantly non-null impact on the risk of PM (figure 1A). Similarly, the doses of occupational asbestos exposure received at a young age contributed to a stronger risk of PM than doses received at an older age (figure 1B). For example, an increase of 1 fibre/mL in the dose received at the age of 20 years was associated with an OR of 1.064 ($OR_{\text{per 1 fibre/mL increase in the dose received at age 20 years}}=1.064$, 95% CI 1.053 to 1.079), while at the age of 50 years, it was 1.019 ($OR_{\text{per 1 fibre/mL increase in the dose received at age 50 years}}=1.019$, 95% CI 1.016 to 1.024), all other doses being constant.

The hypothetical longitudinal occupational asbestos exposure patterns that we compared are shown in figure 2 and the resulting OR are shown in table 3. As expected, subjects who accumulated 20 fibres/mL over their entire occupational history were at higher risk of PM than subjects who accumulated only 1 fibre/mL, using any model (table 3). However, the OR comparing different profiles of dose accumulation were substantially different according to the model used. Indeed, while the CIE conducted by definition to the same risk of PM (OR=1.00) for subjects who have accumulated exactly the same total dose of 20 fibres/mL over their entire occupational history, whatever their profile of exposure, the WCIE resulted in a much higher risk for subjects who accumulated most of the doses in early career than in subjects who accumulated them much later (OR=2.37 (95% CI 2.01 to 2.87)). The weights of the doses of asbestos exposure received in early occupational career were even so high that even if the subject has been exposed to very low doses of exposure thereafter, and has for example accumulated only 10 fibres/mL over his entire job history, he had a high risk of PM than a subject who has lately accumulated a total of 20 fibres/mL (OR=1.23 (95% CI 1.19 to 1.29)).

Table 2 Age, year of birth and occupational asbestos exposure characteristics of male subjects at the time of diagnosis/interview, French case–control study on pleural mesothelioma, France, 1987–2006

	Cases (1,196)		Controls (2,369)	
Age at diagnosis/ interview (years)				
Mean±SD	67.0 (10.3)		66.5 (6.5)	
Median (IQR)	68 (60–74)		66 (63–71)	
Range	25–93		29–89	
Year of birth	**(n)**	**(%)**	**(n)**	**(%)**
<1920	131	11.0	265	11.2
1920–1924	159	13.3	351	14.8
1925–1929	250	20.9	511	21.6
1930–1934	240	20.1	520	22.0
1935–1939	192	16.1	385	16.3
1940–1944	106	8.9	171	7.2
1945–1949	63	5.3	95	4.0
>1949	55	4.6	71	3.0
Exposed to asbestos	**(n)**	**(%)**	**(n)**	**(%)**
Yes	1046	87.5	1423	60.1
No	150	12.5	946	39.9
Total duration of exposure (years)*				
Mean±SD	28.5 (13.0)		25.6 (14.1)	
Median (IQR)	33 (19–38)		29 (12–38)	
Range	1–55		1–54	
Time since first exposure (years)*				
Mean±SD	45.9 (8.9)		43.3 (8.5)	
Median (IQR)	47 (41–53)		44 (40–49)	
Range	7–61		5–61	
Age at first exposure (years)*				
Mean±SD	21.1 (7.1)		22.6 (8.0)	
Median (IQR)	19 (16–24)		20 (17–26)	
Range	10–59		10–64	
Cumulative index of exposure* (fibres/mL)				
Mean±SD	18.0 (37.9)		6.1 (15.3)	
Median (IQR)	3.3 (0.3–21.4)		0.4 (0.01–4.1)	
Range	$1.2.10^{-5}$–551.1		$3.1.10^{-6}$–316.7	

*Distributions in ever exposed to asbestos subjects only (1046 cases and 1423 controls).

DISCUSSION

To the best of our knowledge, this is the first study that attempted to estimate the weight of each dose of asbestos received during the job history and to compare the risk of PM associated with different hypothetical longitudinal exposure profiles. We found

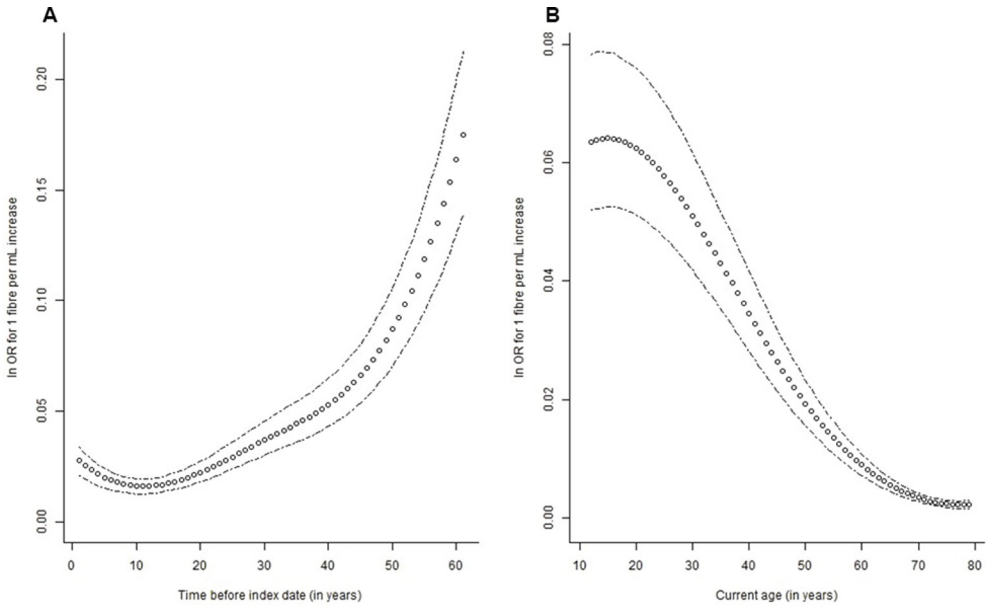

Figure 1 Estimated log OR of pleural mesothelioma associated with an increase of 1 fibre/mL in the dose of asbestos exposure received in years *t* before diagnosis/interview (A) or at age *t* (B), adjusted for the doses received in other years, birth year and age at diagnosis/interview. For example, in A, the estimated adjusted log OR associated with an increase of 1 fibre/mL in the dose of asbestos received the year 50 before diagnosis/interview equals 0.087, which gives OR per 1 fibre/mL increase in the dose received in year 50 before diagnosis/interview = exp(0.087)=1.091. For the year 2 before diagnosis/interview, OR per 1 fibre/mL increase in the dose received in year 2 before diagnosis/interview = exp(0.026)=1.026. Dotted curves are the point estimates derived from the WCE models, and dashed curves are approximate 95% pointwise CIs obtained from nonparametric bootstrap technique.

that an increase occurring more than 40 years before diagnosis/interview, or at 15–20 years old, had the strongest impact on the risk of PM. As a result, subjects who accumulated asbestos fibres in early career were at higher risk than those who accumulated them at the end of their career or at a constant level over the entire career.

From the graphical representation of the estimated time-dependent effect of the intensity of exposure to asbestos over the whole occupational history (figure 1), our results empirically showed that doses of asbestos exposure received in early occupational career likely contribute to early stages of the carcinogenesis process.[25] Moreover, our results also suggest that the doses of asbestos exposure received within the 10-years preceding diagnosis have a significantly non-null weight. While these exposures do not likely contribute to the initiation of cancer, they may play a role in later stages of the carcinogenesis process. These results suggest that recent asbestos exposure should not be discounted in studies on PM, even if some previous studies have found similar OR associated with an increase in cumulative dose, whether recent exposures were discounted or not.[7 8 26 27]

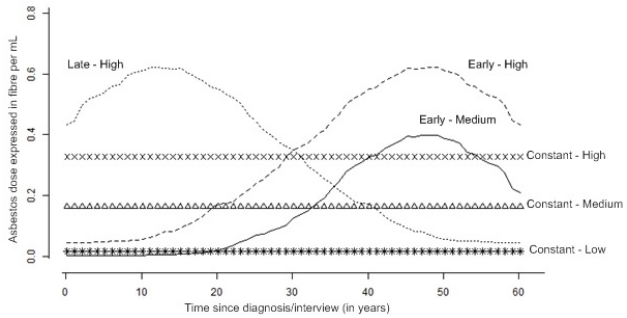

Figure 2 Six hypothetical asbestos exposure patterns according to time before diagnosis/interview or age at exposure. 'Constant-Low' represents subjects who are exposed to asbestos at a constant low annual dose of 0.0164 fibre/mL during 60 years, which conducts to a cumulative index of exposure (CIE) of 1 fibre/mL. 'Constant-Medium 'and 'Constant-High' represent subjects with constant doses, but at higher levels conducting to a CIE of 10 and 20 fibres/mL, respectively. 'Early-Medium 'represents subjects who accumulated 10 fibres/mL over lifetime but with much higher doses in early career. 'Early-High' has the same shape than 'Early-Medium' but with higher doses which result in a CIE of 20 fibres/mL. 'Late-High' represents subjects who accumulated a dose of 20 fibres/mL over lifetime, but with much higher doses in late career instead of early career.

Our findings are consistent with several previous studies that attempted to investigate temporal patterns of the association through the estimation of the effects of age at first exposure and time elapsed between diagnosis and either first or last exposure. In our previous study, we found that the risk of PM was lower when the first occupational exposure to asbestos occurred after the age of 20 years.[2] While they failed to find any exposure–response relationship with age of arrival at Wittenoom, Reid *et al* reported that exposure to blue asbestos in childhood (age of arrival at Wittenoom) was associated with an increased risk of mesothelioma.[4] From a case-control study conducted in Great Britain between 2001 and 2006, Rake *et al* reported an exposure-response relationship with duration of exposure for exposure occurring only before 30 years of age.[3] Despite the large CIs which did not permit to establish a clear exposure–response relationship with age at first exposure, other studies reported a significant effect of age at first exposure, with subjects first exposed at younger ages (before 25 years of age) being at higher risk than those first exposed later in life.[5 28] Our findings on the larger estimated weights for the doses received more than 40 years before the index date are also consistent with the mean latency periods comprised between 40 and 50 years.[9 29] Moreover, several studies have reported an increased risk of PM for subjects with longer time since first exposure.[4 28 30–34]

Table 3 Risk of pleural mesothelioma for different hypothetical occupational asbestos exposure patterns using time prior to diagnosis/interview as the time axis

Exposure pattern*	CIE (in fibres /mL)	WCIE (in fibres/mL)		Reference pattern*	CIE (in fibres /mL)	WCIE (in fibres /mL)	Model with CIE†		Model with WCIE‡	
							OR	95% CI	OR	95% CI
Early-High	20.0§	27.4§	vs	Constant-Low	1.0§	1.0§	1.60‖	1.44 to 1.77	4.09¶	3.13 to 5.52
Late-High	20.0	11.2	vs	Constant-Low	1.0	1.0	1.60	1.44 to 1.77	1.72	1.55 to 1.94
Early-High	20.0	27.4	vs	Late-High	20.0	11.2	1.00	–	2.37	2.01 to 2.87
Early-High	20.0	27.4	vs	Constant-High	20.0	20.0	1.00	–	1.48	1.38 to 1.62
Late-High	20.0	11.2	vs	Constant-High	20.0	20.0	1.00	–	0.63	0.56 to 0.68
Constant -Medium	10.0	10.0	vs	Late-High	20.0	11.2	0.78	0.74 to 0.82	0.94	0.92 to 0.95
Early-Medium	10.0	15.2	vs	Late-High	20.0	11.2	0.78	0.74 to 0.82	1.23	1.19 to 1.29

*Exposure patterns are illustrated in figure 2

†Logistic regression model including the CIE, age, year of birth, and time since first exposure to asbestos.

‡Logistic regression model including the WCIE, age, and year of birth

§For the subject with the 'Early-High' exposure pattern, who has been exposed at high doses of asbestos during early career, its CIE value equaled 20 fibres/mL, whereas its WCIE value equalled 27.42 fibres/mL (because doses received in early career had stronger estimated weights than more recent doses, as shown in figure 1A). For the subject with the 'Constant-Low' exposure pattern, who has been constantly exposed to 0.0164 fibres/mL during each year of the whole time window (1–61 years), both CIE and WCIE equalled to 1.0 because of constant doses and standardisation of the weights in the WCIE.

‖OR comparing the 'Early-High' with the 'Constant-Low' exposure pattern, estimated from the model with CIE:

$$\widehat{OR} = \exp\left(\widehat{\beta_{CIE}} \times (20 - 1)\right) = \exp(0.024658 \times 19) = 1.60$$

¶OR comparing the 'Early-High' with the 'Constant-Low' exposure pattern, estimated from the model with WCIE:

$$\widehat{OR} = \exp\left(\widehat{\beta_{WCIE}} \times (27.42 - 1)\right) = \exp(0.053296 \times 26.42) = 4.09$$

CIE, Cumulative index of exposure to asbestos; WCIE, weighted cumulative index of exposure to asbestos.

Our study has several strengths and limitations. An important strength of our study is related to the use of the weighted cumulative exposure. Indeed, this statistical approach allowed us to estimate the effect of each increase in the annual dose of exposure at any year of exposure history, without making any prior assumption on the possible mechanisms of carcinogenesis due to the exposure. The resulting estimates were automatically adjusted for the doses received in the other years, which also avoided us to further adjust for the timing of exposure. Such further adjustment, which is necessary when using a time-independent cumulative exposure, is not straightforward since it implies not only choosing the most appropriate adjusting covariate (eg, such as time since first exposure) but also to correctly model its effect. The large sample size is another strength of our study since it allowed us to obtain accurate estimates of the dose-time-response relationship. This sample size was obtained thanks to the pooling of two sets of cases and two sets of controls: cases were either included in a previous case–control study conducted between 1987–1993,[6] or identified through the French National Mesothelioma Surveillance programme (PNSM) from 1998 to 2006,[19] and controls were randomly selected from two population-based samples recording complete job history from the French Institute for Public Health Surveillance (InVS) in 1984–2000 and 2007. While such a design may conduct to some potential selection bias as discussed by Bonde *et al*,[35] it has been shown that both sets of cases were representative of all French cases in particular terms of socioeconomic status which is one of the principal components that determine occupational exposures.[6 19] The prevalence of occupational asbestos exposure in the two sets of controls was similar as in the general French population.[36 37]

A potential important limitation of our study is related to our measure of the annual dose of occupational exposure to asbestos. We chose to retrospectively assess occupational asbestos exposure by a job–exposure matrix which is recognised to be a good alternative to retrospective exposure assessment made by experts.[38] However, it is well known that job-exposure matrix may induce misclassifications of exposure. Indeed, job-exposure matrices are useful tools that create a standardised occupational exposure assessment for various homogeneous occupational groups, but do not account for intra-variability within each occupational group. Nonetheless, these misclassifications are non-differential,[38 39] and Hauptmann *et al* have shown that the method used to estimate the flexible time-dependent weight function seemed robust against non-differential measurement errors.[23] Our measure of exposure also suffers from the lack of information in the job-exposure matrix about the type of asbestos fibres. However, in France, almost all the asbestos used were imported and chrysotile seems to have been the most commonly used type. While all forms of asbestos fibres are classified in group 1 'carcinogen to humans with sufficient evidence' by the international agency for research on cancer, chrysotile asbestos is less potent than amphibole towards the induction of PM. Although the magnitude of the association between asbestos exposure and PM depends on fibres type,[40] there is no reason to think that the shape of the weight function, so the time-dependent effect of each increase in the annual dose of exposure, might differ from one fibres type to another. However, we still acknowledge that it would be of interest to assess this assumption by estimating the time-dependent weight function for each type of fibres separately, in studies which would have such a detailed exposure information and sufficient sample size.

In conclusion, by the mean of a flexible time-dependent weighted CIE that imposed no prior assumption on the possible mechanisms of carcinogenesis due to the exposure, we estimated the time-dependent effect of each increase in the dose of asbestos fibres received during the entire job history. In particular, our results indicate that increases in the annual intensity of exposure that occurred more than 40 years before the diagnosis of PM, or at ages 15–20 years, had the strongest impact on the risk of PM. Our study should encourage epidemiologists to consider such modelling of asbestos exposure in studies on PM, as well in other contexts. Our results may also help deciders in planning epidemiological surveillance programme of workers occupationally exposed to asbestos.

Acknowledgements The authors thank Y. Iwatsubo for allowing us to use data collected from the previous mesothelioma case–control study, the members of PNSM technical committee (P. Astoul, P. Brochard, S. Chammings, S. Ducamp, C. Frenay, F. Galateau-Salle, A. Gilg Soit Ilg, M. Goldberg, N. Le Stang and JC Pairon) for allowing us to use data collected from the PNSM.

Contributors AL performed literature review, drafted the first version of this manuscript and co-supervised all aspects of this manuscript. EL performed all statistical analyses presented in the manuscript and contributed to the drafting of the manuscript. EG performed preliminary statistical analyses. MPS contributed to the implementation of statistical methods and interpretation of results. AGS supervised data collection within the French National Mesothelioma Surveillance Program. KL supervised all aspects of this manuscript. All co-authors participated in the editing and correction of the final text.

Funding The French National Research Program for Environmental and Occupational health of Anses with support of the cancer TMOI of the French National Alliance for Life and Health Sciences (AVIESAN) – 2013/1/177.

Competing interests None declared.

Ethics approval French Data Protection Authority.

Provenance and peer review Not commissioned; externally peer reviewed.

REFERENCES

1 Wagner JC, Sleggs CA, Marchand P. Diffuse pleural mesothelioma and asbestos exposure in the north western cape province. *Br J Ind Med* 1960;17:260–71.
2 Lacourt A, Leffondré K, Gramond C, *et al*. Temporal patterns of occupational asbestos exposure and risk of pleural mesothelioma. *Eur Respir J*2012;39:1304–12.
3 Rake C, Gilham C, Hatch J, *et al*. Occupational, domestic and environmental mesothelioma risks in the british population: a case-control study. *Br J Cancer* 2009:100:1175–83.
4 Reid A, Franklin P, Olsen N, *et al*. All-cause mortality and cancer incidence among adults exposed to blue asbestos during childhood. *Am J Ind Med* 2013:56:133–45.
5 Spirtas R, Heineman EF, Bernstein L, *et al*. Malignant mesothelioma: attributable risk of asbestos exposure. *Occup Environ Med* 1994;51:804–11.
6 Iwatsubo Y, Pairon JC, Boutin C, *et al*. Pleural mesothelioma: dose-response relation at low levels of asbestos exposure in a french population-based case-control study. *Am J Epidemiol* 1998;148:133–42.
7 Lacourt A, Gramond C, Rolland P, *et al*. Occupational and non-occupational attributable risk of asbestos exposure for malignant pleural mesothelioma. *Thorax* 2014;69:532–9.
8 Rödelsperger K, Jöckel KH, Pohlabeln H, *et al*. Asbestos and man-made vitreous fibers as risk factors for diffuse malignant mesothelioma: results from a german hospital-based case-control study. *Am J Ind Med* 2001;39:262–75.
9 Marinaccio A, Binazzi A, Cauzillo G, *et al*; Italian Mesothelioma Register (ReNaM) Working Group. Analysis of latency time and its determinants in asbestos related malignant mesothelioma cases of the italian register. *Eur J Cancer* 2007:43:2722–8.
10 Richardson DB, Cole SR, Chu H, *et al*. Lagging exposure information in cumulative exposure-response analyses. *Am J Epidemiol* 2011;174:1416–22.
11 Checkoway H, Pearce N, Hickey JL, *et al*. Latency analysis in occupational epidemiology. *Arch Environ Health* 1990;45:95–100.
12 Salvan A, Stayner L, Steenland K, *et al*. Selecting an exposure lag period. *Epidemiology* 1995;6:387–90.
13 Hauptmann M, Pohlabeln H, Lubin JH, *et al*. The exposure-time-response relationship between occupational asbestos exposure and lung cancer in two german case-control studies. *Am J Ind Med* 2002;41:89–97.
14 Sylvestre MP, Abrahamowicz M. Flexible modeling of the cumulative effects of time-dependent exposures on the hazard. *Stat Med* 2009;28:3437–53.
15 Armitage P, Doll R. The age distribution of cancer and a multi-stage theory of carcinogenesis. *Br J Cancer* 1954;8:1–12.
16 Peto J, Doll R, Hermon C, *et al*. Relationship of mortality to measures of environmental asbestos pollution in an asbestos textile factory. *Ann Occup Hyg* 1985;29:305–55.
17 Peto J, Seidman H, Selikoff IJ. Mesothelioma mortality in asbestos workers: implications for models of carcinogenesis and risk assessment. *Br J Cancer* 1982;45:124–35.
18 Peto J. Problems in dose response and risk assessment. Castellani A, ed. *Epidemiology and quantification of environmental risk in humans from radiation and other agents. Italy: springer science & business media*, 2013:175–86.
19 Goldberg M, Imbernon E, Rolland P, *et al*. The french national mesothelioma surveillance program. *Occup Environ Med* 2006;63:390–5.
20 *International standard classification of occupations*. Genève International labour organization 1968.
21 *International standard industrial classification of all economic activities (Revision 2)*. New york United nation 1975.
22 Durrleman S, Simon R. Flexible regression models with cubic splines. *Stat Med* 1989;8:551–61.
23 Hauptmann M, Wellmann J, Lubin JH, *et al*. Analysis of exposure-time-response relationships using a spline weight function. *Biometrics* 2000;56:1105–8.
24 Chapter 4. The NLP Procedure. *SAS/OR® 9.22 user's Guide: constraint programming* : SAS Institute Inc.
25 Day NE, Brown CC. Multistage models and primary prevention of cancer. *J Natl Cancer Inst* 1980;64:977–89.
26 Pesch B, Taeger D, Johnen G, *et al*. Cancer mortality in a surveillance cohort of german males formerly exposed to asbestos. *Int J Hyg Environ Health* 2010;213:44–51.
27 Reid A, Heyworth J, de Klerk NH, *et al*. Cancer incidence among women and girls environmentally and occupationally exposed to blue asbestos at Wittenoom, western Australia. *Int J Cancer* 2008;122:2337–44.
28 Pira E, Pelucchi C, Piolatto PG, *et al*. First and subsequent asbestos exposures in relation to mesothelioma and lung cancer mortality. *Br J Cancer* 2007;97:1300–4.
29 Bianchi C, Giarelli L, Grandi G, *et al*. Latency periods in asbestos-related mesothelioma of the pleura. *Eur J Cancer Prev* 1997;6:162–6.
30 Ferrante D, Bertolotti M, Todesco A, *et al*. Cancer mortality and incidence of mesothelioma in a cohort of wives of asbestos workers in Casale Monferrato, Italy. *Environ Health Perspect* 2007;115:1401.
31 Hansen J, de Klerk NH, Musk AW, *et al*. Environmental exposure to crocidolite and mesothelioma: exposure-response relationships. *Am J Respir Crit Care Med* 1998;157:69–75.
32 Magnani C, Ferrante D, Barone-Adesi F, *et al*. Cancer risk after cessation of asbestos exposure: a cohort study of italian asbestos cement workers. *Occup Environ Med* 2008;65:164–70.
33 Pira E, Pelucchi C, Buffoni L, *et al*. Cancer mortality in a cohort of asbestos textile workers. *Br J Cancer* 2005;92:580–6.
34 Barone-Adesi F, Ferrante D, Bertolotti M, *et al*. Long-term mortality from pleural and peritoneal cancer after exposure to asbestos: possible role of asbestos clearance. *Int J Cancer* 2008;123:912–6.
35 Bonde JP. No indication that mineral wool causes mesothelioma. *Am J Respir Crit Care Med* 2013;188:873.
36 Goldberg M, Banaei A, Goldberg S, *et al*. Past occupational exposure to asbestos among men in France. *Scand J Work Environ Health* 2000;26:52–61.
37 Lacourt A, Gramond C, Audignon S, *et al*. Pleural mesothelioma and occupational coexposure to asbestos, mineral wool, and silica. *Am J Respir Crit Care Med* 2013;187:977–82.
38 Bouyer J, Dardenne J, Hémon D. Performance of odds ratios obtained with a job-exposure matrix and individual exposure assessment with special reference to misclassification errors. *Scand J Work Environ Health* 1995;21:265–71.
39 Kauppinen TP, Mutanen PO, Seitsamo JT. Magnitude of misclassification bias when using a job-exposure matrix. *Scand J Work Environ Health* 1992;18:105–12.
40 Hodgson JT, Darnton A. The quantitative risks of mesothelioma and lung cancer in relation to asbestos exposure. *Ann Occup Hyg* 2000;44:565–601.

2.4.2 Discussion complémentaire à l'article 1

2.4.2.1 Biais potentiels liés au design de l'étude

Des biais potentiels ont pu être introduits lors de la mise en place de cette étude cas-témoins un peu originale par rapport à une étude cas-témoins classique en population générale. En effet, la population d'étude a été constituée à partir de données déjà existantes. A l'origine, les données ont été recueillies pour un autre objectif que celui fixé pour cette étude. Considérer plusieurs sources pour les cas et témoins est un choix qui a été fait afin d'obtenir un échantillon de taille suffisante. Cependant, un tel choix devait assurer le respect des principes fondamentaux du design de l'étude cas-témoins. Les deux principaux piliers d'un tel design sont : 1) les cas et témoins doivent provenir de la même population source, 2) les informations recueillies doivent être similaires pour tous les sujets.

2.4.2.1.a Population source

Les cas inclus dans cette étude proviennent de deux sources distinctes (cas A et cas B). Concernant les cas A, ils ont été inclus grâce aux signalements réalisés par les services hospitaliers participant à l'étude dans 5 régions françaises entre 1987 et 1993. Étant donné la gravité de la maladie, les sujets malades se rendent tous à l'hôpital au moins une fois, la population source est donc bien celle de la population générale dans les 5 régions françaises entre 1987 et 1993. Les cas B ont été inclus s'ils étaient cas incidents vivants entre 1998 et 2006 dans les 25 départements du PNSM. Ce programme de surveillance permet un enregistrement exhaustif de tous les cas incidents du mésothéliome dans ces départements. La population source des cas B est donc celle de la population générale des 25 départements entre 1998 et 2006.
Pour étudier la comparabilité des cas A et B en terme d'exposition à l'amiante, ils ont été comparés entre eux sur différents indicateurs d'exposition (la probabilité maximale d'exposition rencontrée au cours de la vie, l'ICE et l'IME (Indice Moyen d'Exposition défini dans Lacourt et al. [2012])). Quel que soit l'indicateur considéré, il n'a pas été mis en évidence de différence significative de la distribution de l'exposition entre les cas A et B (Lacourt [2010]). On peut donc raisonnablement penser que la population source des cas est bien celle de la population générale entre 1987 et 2006 en France.

Les témoins inclus dans cette étude proviennent également de deux échantillons distincts (témoins A et B). Les témoins A ont été sélectionnés à partir d'un échantillon constitué de 10010 sujets âgés entre 25 et 74 ans en 2007. Cet échantillon a été constitué à partir de la méthode des quotas permettant de s'assurer au mieux de la représentativité vis-à-vis de la population générale âgée entre 25 et 74 ans en 2007. De par la définition des quotas (les variables de stratification ont été précisées au chapitre 1), nous pouvons raisonnablement penser que les témoins A sont bien issus de la population générale française vis-à-vis de l'exposition profes-

sionnelle à l'amiante. Le second échantillon de témoins (témoins B) a été constitué à partir de calendriers professionnels issus de différentes études épidémiologiques réalisées entre 1984 et 2000. La représentativité de cet échantillon vis-à-vis de la population générale française quant à l'exposition professionnelle à l'amiante a déjà été publiée et montrée dans Goldberg et al. [2000]. La population source des témoins B peut donc être considérée comme représentative de la population générale sur les différentes périodes de temps considérées. Comme pour les cas, au sein de sa thèse, Aude Lacourt a comparé les distributions des deux sources témoins suivant l'ICE et l'IME. Elle n'a pas mis en évidence de différence significative entre les témoins d'une source à l'autre (Lacourt [2010]). La population des témoins ainsi constituée peut être considérée comme représentative de la population française vis-à-vis de l'exposition professionnelle à l'amiante.

Même si les périodes de recrutement ne sont pas rigoureusement les mêmes pour les témoins par rapport aux cas, les sujets ont potentiellement été exposés durant la même période. De plus, même si les cas et témoins n'ont pas été identifiés dans les mêmes régions, celles-ci se veulent représentatives de la population générale française. Il est donc raisonnable de penser que les cas et témoins sont bien issus de la même population source et que cette dernière est représentative de la population française en ce qui concerne l'exposition professionnelle à l'amiante.

2.4.2.1.b Recueil d'information

Un biais d'information a pu être introduit par l'utilisation de plusieurs questionnaires suivant les différentes sources des cas et témoins. Cependant, seules les informations communes de l'ensemble de ces questionnaires standardisés ont été recueillies pour chaque sujet de cette étude. Au regard de l'exposition professionnelle à l'amiante, les cas et témoins ont eu à renseigner leur calendrier professionnel, qui est standardisé à travers les questionnaires. De plus, pour l'ensemble des sujets de l'étude et quelle que soit la source considérée, seules les informations sur les années de début et de fin d'emploi, les professions et les secteurs d'activité, ont été considérées des calendriers professionnels. Il est donc raisonnable de penser qu'il n'y a pas de différence d'information au regard de l'exposition professionnelle à l'amiante entre les cas et témoins de l'étude.

2.4.2.1.c Analyse complémentaire

Au-delà des détails apportés sur les biais potentiels liés à ces données, une analyse complémentaire a été réalisée sur d'autres données provenant d'une étude cas-témoins en population générale conduite entre 1998 et 2005. Cette analyse complémentaire a notamment permis de répondre à la lettre de Boffetta et al. (ci-après), qui remettait en question nos résultats obtenus dans l'article 1.

LETTERS

Response to: 'Dose–time–response association between occupational asbestos exposure and pleural mesothelioma' by Lacourt *et al*

The study by Lacourt reported in a recent issue of *Occup Environ Med*[1] does not satisfy two fundamental requirements of the design and conduct of a case–control study: these limitations invalidate its results and those of other reports based on the same dataset.[2]

First, all cases and controls should be sampled from the same population–time experience ('study-base').[3] An alternative, commonly used way to express this concept is that controls are subjects (or a random sample of theirs) who would have been selected as cases, had they developed the disease of interest.

Cases in the study by Lacourt *et al* were men who were diagnosed with mesothelioma 'either in five French regions between January 1987 and December 1993 or in 22 French districts between January 1998 and December 2006'.[1] The first series of cases was collected in a community-based case–control study,[4] and the second series was identified within a mesothelioma surveillance programme.[5]

Controls were identified from a sample of the national population selected in 2007 from all regions and from the controls selected in 15 community-based case–control studies conducted in nearly all regions of France in 1984–2000. No additional details are provided of these two sources of controls, for example, how were subjects selected for the 2007 'national sample', or whether the case–control study from which the first series of cases was selected[4] was among the 15 studies whose controls were used as second sources. Notably, no controls were selected between 2000 and 2006, when most of the cases were enrolled in the national surveillance programme. It is clear that cases and controls belong to different populations, both from a temporal and a geographic viewpoint, and that the principle of identifying cases and controls from the same study base was ignored. In particular, cases were selected from studies conducted in areas at high prevalence of asbestos exposure, while the same is not likely to be the case for studies used to select controls. Consequently, the prevalence of exposure to asbestos and to potential confounders and effect modifiers, in these latter studies, is hardly representative of that of the areas from which cases were drawn.

The second principle of case–control studies not met by the study by Lacourt *et al*[1] is that information on exposure, confounders and effect modifiers should be collected in the same way for all study subjects. The authors write that 'a different but standardised questionnaire was administered to cases and controls by trained interviewers', but report no results of sensitivity or validation analyses which might provide evidence on how much the 'standardisation' addressed differences in the assessment of exposure to asbestos, and other factors. This is particularly disturbing and likely to be a source of information bias, because the studies used to recruit cases were designed to assess the health effects of asbestos, while it is unclear whether the same applies to the studies used to select controls.

Prevention of selection and information bias is a pivotal methodological challenge in case–control studies,[6] and much effort is spent in well-conducted studies to try to reduce opportunity for bias in the design and conduct of the study. The authors briefly address selection and information bias in their study by stating that 'both sets of cases were representative of all French cases in particular terms of socioeconomic status which is one of the principal components that determine occupational exposures'[1]; however, no such evidence is provided in the original publications, for example, nowhere in the report of the multicentre case–control study there is a statement of the national representativeness of the cases or a comparison of socioeconomic status or any other characteristics with a national sample.[4] The authors also state that 'the prevalence of occupational asbestos exposure in the two sets of controls was similar as in the general French population',[1] but again they provide no evidence of that—not even a reference to published reports where more details on these populations could be found. Analyses aimed at detecting and quantifying possible bias have been recommended[7] and should be considered in case of studies conducted with unorthodox approaches to select study participants and gather exposure and confounder information. Caution has been expressed in the interpretation of quantitative estimates of risk based on expert assessment of retrospective exposures in the absence of actual measurements[8]; in the case of the study by Lacourt *et al*,[1] however, the limitations discussed above overshadow any consideration of other sources of bias.

Paolo Boffetta,[1] Enrico Pira,[2] Canzio Romano,[2] Francesco Saverio Violante,[3] Andrea Farioli,[3] Carlo Zocchetti,[4] Carlo La Vecchia,[5]

[1]Tisch Cancer Institute, Icahn School of Medicine at Mount Sinai, New York, New York, USA
[2]Department of Public Health and Pediatric Sciences, University of Turin, Turin, Italy
[3]Department of Medical and Surgical Sciences (DIMEC), Bologna University, Bologna, Italy
[4]Ricerche e Studi in Sanità e Salute (ReSiSS), Gallarate, Varese, Italy
[5]Department of Clinical Sciences and Community Health, Milan University, Milan, Italy

Contributors PB drafted the letter. All authors reviewed and edited it.

Competing interests None declared.

Provenance and peer review Not commissioned; internally peer reviewed.

To cite Boffetta P, Pira E, Romano C, *et al*. *Occup Environ Med* 2018;**75**:160.

Received 13 June 2017
Accepted 22 June 2017
Published Online First 20 August 2017

► http://dx.doi.org/10.1136/oemed-2017-104802

Occup Environ Med 2018;**75**:160.
doi:10.1136/oemed-2017-104570

REFERENCES

1 Lacourt A, Lévêque E, Guichard E, *et al*. Dose-time-response association between occupational asbestos exposure and pleural mesothelioma. *Occup Environ Med* 2017;74:691–7.
2 Lacourt A, Leffondré K, Gramond C, *et al*. Temporal patterns of occupational asbestos exposure and risk of pleural mesothelioma. *Eur Respir J* 2012;39:1304–12.
3 Miettinen OS. The "case-control" study: valid selection of subjects. *J Chronic Dis* 1985;38:543–8.
4 Iwatsubo Y, Pairon JC, Boutin C, *et al*. Pleural mesothelioma: dose-response relation at low levels of asbestos exposure in a French population-based case-control study. *Am J Epidemiol* 1998;148:133–42.
5 Goldberg M, Imbernon E, Rolland P, *et al*. The French National Mesothelioma Surveillance Program. *Occup Environ Med* 2006;63:390–5.
6 Rothman KJ, Greenland S, Lash TL, *et al*. *Modern Epidemiology*. 3rd ed. Philadelphia, PA: Lippincott, Williams & Wilkins, 2008.
7 Lash TL, Fox MP, MacLehose RF, *et al*. Good practices for quantitative bias analysis. *Int J Epidemiol* 2014;43:1969–85.
8 Siemiatycki J, Boffetta P. Invited commentary: is it possible to investigate the quantitative relation between asbestos and mesothelioma in a community-based study? *Am J Epidemiol* 1998;148:143–7.

Pour renforcer nos propos sur les biais potentiels et montrer que nos résultats obtenus n'étaient pas dû à la particularité de ces données, nous avons estimé l'effet de l'intensité d'exposition professionnelle à l'amiante sur le risque de mésothéliome pleural en chaque temps avant la date index en utilisant le WCIE sur des données provenant d'une étude cas-témoins multicentrique basée en population générale. Les cas étaient définis comme tous les cas incidents diagnostiqués avec un MP primaire entre 1998 et 2002 et résidant dans un des départements faisant parti du PNSM. Parmi les 750 cas identifiés durant cette période, 462 (371 hommes et 51 femmes) d'entre eux ont été interviewés entre 1998 et 2004. Parmi eux, près de 87% ont eu une confirmation pathologique et 13% une confirmation clinique. La figure 2.3 provenant de l'article de Rolland et al. [2010] montre la répartition de ces 462 cas au sein des différents départements français. Dans le cadre de nos analyses, nous avons seulement considérés les hommes. Les cas présents dans cette étude étaient tous des cas de ceux présents dans l'étude cas-témoins utilisée pour l'article 1.

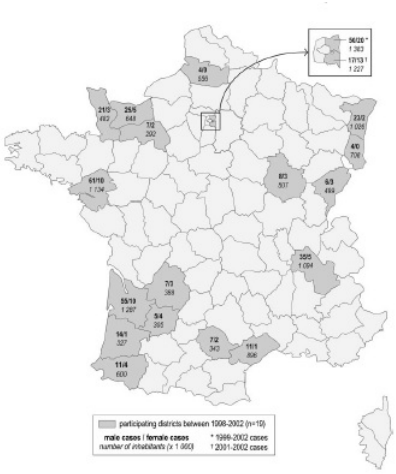

FIGURE 2.3 – Répartition géographique des 462 cas. *Figure tirée de Rolland et al. [2010]*

Deux témoins par cas, tirés au sort aléatoirement à partir des listes électorales, ont été appariés sur le sexe, l'âge (± 5 ans) et le département de résidence. Au total, 897 témoins (732 hommes et 165 femmes) ont été inclus dans cette étude. Les 732 témoins hommes concernés étaient tous différents de ceux présents dans l'étude cas-témoins utilisée pour l'article 1.
Après accord oral de participation des sujets, un auto-questionnaire permettant de reconstituer leur histoire résidentielle et professionnelle était envoyé par voie postale. Cet auto-questionnaire était repris et complété par un questionnaire complémentaire lors d'un entretien en face à face avec un enquêteur formé. Les emplois occupés pendant au moins 6 mois ont été codés selon des classifications standards nationales et internationales de profession et d'industrie (détaillées dans le chapitre 1).
Dans le cadre de cette analyse complémentaire, nous avons évalué l'exposition professionnelle à l'amiante avec la même MEE précédemment décrite. Seules les expositions ayant eu lieu moins

de 61 ans avant la date index étaient considérées afin de pouvoir comparer la fonction de poids estimée suivant le temps avant la date index avec celle de l'article 1.

	Cas (n=371)	Témoins (n=732)
Age à la date index		
Moyenne (SD)	67.8 (9.4)	66.9 (9.1)
Médiane (IQR)	69 (62-74)	68 (62-73)
Min-Max	41-93	41-89
Exposés à l'amiante		
Oui n(%)	336 (90,6%)	469 (64.7%)
Non n(%)	35 (9.4%)	260 (35.3%)
Durée totale d'exposition (années)		
Moyenne (SD)	32.6 (13.1)	27.2 (15.9)
Médiane (IQR)	37 (26-42)	31 (12-41)
Min-Max	1-54	1-53
Temps depuis la première exposition (années)		
Moyenne (SD)	49.3 (9.6)	48.0 (10.9)
Médiane (IQR)	51 (43-58)	50(41-57)
Min-Max	15-61	6-61
Age à la première exposition (années)		
Moyenne (SD)	18.4 (5.6)	18.6 (6.9)
Médiane (IQR)	17 (14-21)	17 (14-21)
Min-Max	12-49	12-57
Indice cumulé d'exposition (fibres/mL)		
Moyenne (SD)	19.6 (48.0)	6.0 (14.5)
Médiane (IQR)	4.9 (0.4-21.8)	0.2 (0.004-3.3)
Min-Max	1.252e-05-713.6	2.817e-05-111.6

Tableau 2.1 – Description des variables démographiques et d'exposition des sujets utilisés pour l'analyse complémentaire

D'après le tableau 2.1, les cas et témoins avaient en moyenne un âge proche de ceux présentés dans l'article 1 (cas / témoins : 67.0 ans / 66.5 ans) avec un âge minimum au diagnostic plus élevé pour les cas (41 ans vs 25 ans). La proportion de cas/témoins exposés professionnellement à l'amiante était plus élevée pour cette étude (article 1 : cas / témoins : 87.5% / 60.1%). Ils avaient également une durée et un temps depuis la première exposition en moyenne plus élevés. Les résultats de cette analyse complémentaire sont présentés dans la lettre ci-après.

Avec un effectif de la population plus faible par rapport à celui de l'article 1, on constate des IC plus larges pour la fonction de poids au delà de 50 ans avant la date index (figure 1B de la lettre). Cependant, on peut considérer qu'il y avait suffisamment d'information puisque la fonction de poids a pu être estimée en ces temps. Les résultats observés dans les deux analyses sont similaires, ce qui suggère que ceux obtenus dans l'article 1 ne s'expliquent pas par les biais invoqués par le design particulier de l'étude.

Dose–time–response association between occupational asbestos exposure and pleural mesothelioma: authors' response

We thank Bofetta *et al*[1] for giving us the opportunity to clarify the design of our case–control study on pleural mesothelioma (PM),[2] which has been yet described and discussed in more details elsewhere.[3 4]

We assembled the cases (n=496) from a French hospital-based case–control study conducted in 1987–1993 in five French regions (see Iwatsubo *et al*[5]) composed of 24 departments (over the 96 departments in metropolitan France) and incident cases (n=700) identified within the French National Mesothelioma Surveillance Programme (NMSP) in 1998–2006. The NMSP, which is coordinated by the French National Public Health Agency, has been established partly to estimate the national incidence of mesothelioma in France and thus included departments that were chosen and shown to be representative of France in terms of demographic, employment and economic activity characteristics as detailed in Goldberg *et al*.[6] In 2006, the NMSP included 25 departments including 18 that were different from the 1987–1993 case–control study. Overall, the 1196 cases of our study were recruited in 42 different departments that were well distributed throughout the French Metropolitan territory (see Lacourt,[3] p57). Moreover, the distribution of the occupational exposure to asbestos has been shown to be similar between the cases from the 1987–1993 study and the cases from the NMSP (see Lacourt,[3] p175). We thus believe that our cases were not selected from particular areas at high prevalence of asbestos exposure but more likely in areas representative of France regarding asbestos exposure.

The 2369 controls of our study were selected from two samples also elaborated by the French National Public Health Agency. A total of 751 controls were randomly selected from a representative national sample of the general population in 2007 including all French regions (see Fevotte *et al*[7]), and 1618 controls were selected from 15 epidemiological studies conducted in nearly all French regions in 1984–2000 (see Goldberg *et al*[8]). Overall, the 2369 controls were thus recruited in 1984–2000 and 2007 over the whole French territory. Controls have thus been selected partly during the period of cases' recruitment and partly at the end of that period. We acknowledge that controls who were recruited in 2007 had to survive without PM until that date and may thus not be representative of source population of the cases diagnosed before. However, in terms of exposure, they were potentially exposed during the same period of time, since all controls were frequency matched to cases on birth year within 5-year groups, and all analyses were carefully adjusted for age at the index date. For all these reasons, we believe that the distribution of occupational exposure to asbestos in our controls was very similar to that of the source population of the cases.

Regarding data collection, we acknowledge that questionnaires and interview procedures differed across the studies used to assemble all cases and controls, although occupational history was mostly collected using specific standardised questionnaires and face-to-face or telephone interviews with a trained interviewer (detailed information may be found in each specific reference mentioned earlier). However, the only information that we used from these questionnaires was birth date, index date and job history (start and end dates, occupation and industry for each job). Indeed, occupational exposure to asbestos was assessed using the same job exposure matrix (JEM), and we adjusted only for birth year and age.

To assess the robustness of our results, we reanalysed data from a single multicentre French population-based case–control study on PM conducted in 1998–2005. The study design is described in details in Rolland *et al*.[9] Briefly, incident cases were identified within the NMSP in 1998–2002 and interviewed in 1998–2004 (n=371, all also included in Lacourt *et al*[2]). Two population controls were randomly selected from electoral lists, individually matched to cases on sex, age (±5 years) and departments of residence and interviewed in 1998–2005 (n=732 controls, all different from controls in Lacourt *et al*[2]). Job history was provided in a standardised questionnaire administrated by trained interviewers. We used the same JEM as in Lacourt *et al*[2] to assess

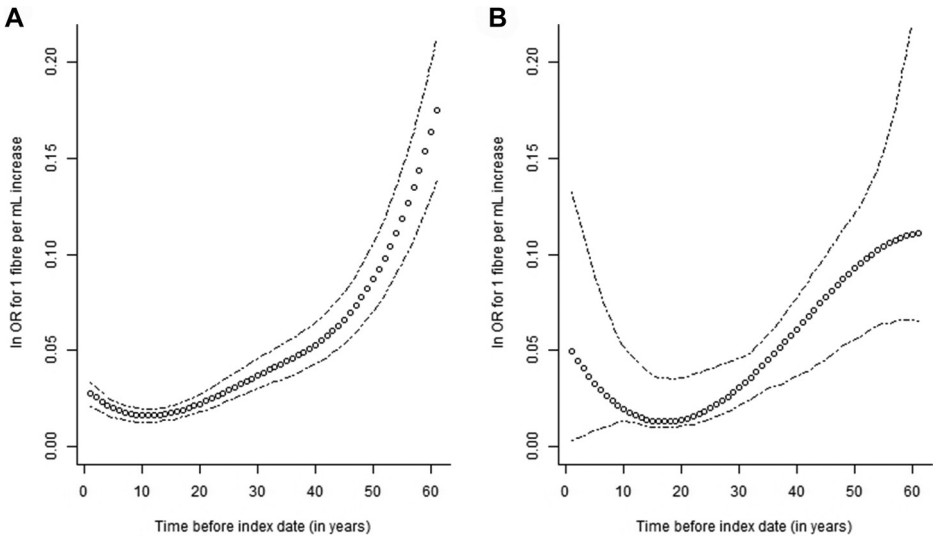

Figure 1 Estimated log OR of pleural mesothelioma associated with an increase of one fibre per millilitre in the average annual daily intensity of asbestos exposure received in years t before diagnosis/interview, adjusted for intensity received in other years, birth year and age at diagnosis/interview. (A) The results obtained with French case–control data assembled in Lacourt *et al*.[2] (B) The results obtained using the data from the French 1998–2005 case–control study.[9]

occupational asbestos exposure as well as the same statistical analysis based on the weighted cumulative index of exposure.[10] We obtained similar results as in Lacourt *et al*[2] (figure 1), with much larger CIs because of a much smaller sample size. The intensity of asbestos exposure received more than 35 years before diagnosis had the strongest contribution to the risk of PM, and recent intensity had a non-null weight despite very large CI.

Aude Lacourt,[1] Emilie Leveque,[1,2] Marcel Goldberg,[3,4] Karen Leffondre[2]

[1]University Bordeaux, INSERM, Bordeaux Population Health Research Center, Team EPICENE, UMR 1219, Bordeaux, France
[2]University Bordeaux, ISPED, INSERM, Bordeaux Population Health Research Center, Team Biostatistics, UMR 1219, Bordeaux, France
[3]UMS 11, INSERM, VILLEJUIF, France
[4]Paris Descartes University, Paris, France

Correspondence to Dr Aude Lacourt, Equipe EPICENE cancer et environnement, Bordeaux Population Health Center - Inserm U1219, Université Bordeaux - ISPED, 146 rue Leo Saignat, 33076 Bordeaux Cedex, France; aude.lacourt@inserm.fr

Contributors All authors contributed.

Competing interests None declared.

Patient consent Obtained.

Ethics approval Comission national informatique et libertés (CNIL).

Provenance and peer review Not commissioned; internally peer reviewed.

To cite Lacourt A, Leveque E, Goldberg M, *et al*. *Occup Environ Med* 2018;**75**:161–162.

Received 4 October 2017
Accepted 12 October 2017
Published Online First 23 October 2017

► http://dx.doi.org/10.1136/oemed-2017-104570

Occup Environ Med 2018;**75**:161–162.
doi:10.1136/oemed-2017-104802

REFERENCES

1 Boffetta P, Pira E, Romano C, *et al*. Response to: 'Dose-time-response association between occupational asbestos exposure and pleural mesothelioma' by Lacourt *et al*. *Occup Environ Med* 2018;75:160.
2 Lacourt A, Lévêque E, Guichard E, *et al*. Dose-time-response association between occupational asbestos exposure and pleural mesothelioma. *Occup Environ Med* 2017;74:691–7.
3 Lacourt A. Mésothéliome: étiologie professionnelle à partir d'enquêtes cas-témoins françaises [Online] PhD thesis. Université de Bordeaux: Bordeaux, France, 2010. (Accessed 19 September 2017).
4 Lacourt A, Leffondré K, Gramond C, *et al*. Temporal patterns of occupational asbestos exposure and risk of pleural mesothelioma. *Eur Respir J* 2012;39:1304–12.
5 Iwatsubo Y, Pairon JC, Boutin C, *et al*. Pleural mesothelioma: dose-response relation at low levels of asbestos exposure in a French population-based case-control study. *Am J Epidemiol* 1998;148:133–42.
6 Goldberg M, Imbernon E, Rolland P, *et al*. The French National Mesothelioma Surveillance Program. *Occup Environ Med* 2006;63:390–5.
7 Févotte J, Dananché B, Delabre L, *et al*. Matgéné: a program to develop job-exposure matrices in the general population in France. *Ann Occup Hyg* 2011;55:A78–78.
8 Goldberg M, Banaei A, Goldberg S, *et al*. Past occupational exposure to asbestos among men in France. *Scand J Work Environ Health* 2000;26:52–61.
9 Rolland P, Gramond C, Lacourt A, *et al*. Occupations and industries in France at high risk for pleural mesothelioma: a population-based case-control study (1998-2002). *Am J Ind Med* 2010;53:1207–19.
10 Hauptmann M, Wellmann J, Lubin JH, *et al*. Analysis of exposure-time-response relationships using a spline weight function. *Biometrics* 2000;56:1105–8.

2.4.2.2 Biais potentiels liés à l'usage d'une MEE

Dans les études cas-témoins, on a une information rétrospective de l'exposition. Lorsqu'il n'est pas possible d'avoir une évaluation objective de l'exposition à une nuisance, grâce à des mesures métrologiques par exemple, il est nécessaire de procéder à une évaluation rétrospective de l'exposition. Concernant l'exposition professionnelle, on reconstitue l'histoire d'exposition sur la base du calendrier professionnel avec des questions plus spécifiques sur les tâches réalisées. S'il y a suffisamment d'information, on peut réaliser une expertise individuelle des questionnaires qui est considérée comme le gold standard pour l'évaluation rétrospective des expositions. Tandis que l'expertise individuelle est longue et très coûteuse, l'évaluation rétrospective de l'exposition peut être automatisée au travers de l'utilisation de MEE. Pour évaluer l'exposition professionnelle à l'amiante à partir des calendriers professionnels de chacun de nos sujets, nous avons utilisé une MEE, que l'on a décrite au chapitre 1.

L'utilisation d'une MEE apparait comme une bonne alternative à l'expertise pour évaluer les expositions rétrospectives en milieu professionnel. Cet outil est de faible coût à l'utilisation, il est plus facile à mettre en place et surtout il permet d'avoir une évaluation standardisée et rapide de l'exposition à l'agent considéré. C'est un outil qui est plutôt transparent dans son utilisation permettant ainsi une reproductibilité des résultats plus aisée.

Néanmoins, elle comporte certaines limites. Une MEE donne une information au niveau d'un groupe homogène d'exposition défini par un emploi. Elle ne permet pas de prendre en compte l'hétérogénéité des activités au sein d'un même emploi, ni les activités non prévisibles par le simple intitulé mais qui ont pu être reportées par le sujet lors de l'interview. Ainsi, lors de la conception de la matrice, il est très important de définir des emplois les plus homogènes possibles ainsi que de garantir une bonne qualité de codage des emplois. C'est pourquoi un guide d'aide au codage est généralement élaboré afin de permettre un codage des emplois qui soit reproductible et homogène.
Une MEE fournit alors une évaluation de l'exposition moyenne au sein d'un emploi et ne permet pas de considérer la variabilité intra-individuelle. Lorsqu'elle est appliquée au niveau individuel, par principe, elle génère donc des erreurs de classification. Néanmoins, les erreurs de classification générées sont indépendantes du statut cas-témoin si 1) les sujets reportent de la même façon leur calendrier professionnel et 2) le codage est de même qualité entre cas et témoins.
Le fait de devoir se rappeler de son histoire professionnelle sur une longue période peut amener à des biais de mémorisation. Cependant, dans la littérature, le biais de mémorisation est considéré comme non différentiel au regard du statut cas-témoin (Teschke et al. [2002]). Concernant le codage des emplois, quelle que soit l'étude considérée, il a été réalisé à l'insu du statut cas-témoin du sujet. Ces erreurs de classification sont donc plutôt considérées comme non différentielles. Sans le démontrer clairement, Hauptmann et al. [2000b] ont, néanmoins, précisé que la méthode du WCIE est robuste aux erreurs de mesures non différentielles.

2.4.2.3 Limites statistiques du WCIE

2.4.2.3.a La fonction de poids utilisée

Utiliser des fonctions splines pour représenter la fonction de poids dépendante du temps dans le WCIE est un réel avantage. Contrairement aux fonctions paramétriques, les fonctions splines sont estimées directement à partir des données sans poser aucune hypothèse a priori sur la forme. Néanmoins, ces fonctions sont sensibles au nombre et à la position de ses noeuds. Pour éviter les problèmes de convergence des modèles et la sur-paramétrisation de ces fonctions, nous avons 1) limité le nombre possible de noeuds intérieurs et 2) choisi le nombre total de noeuds par minimisation de l'AIC. Il faut tout de même avoir conscience de cette flexibilité dans l'interprétation des résultats. Les fonctions splines peuvent également être instables aux extrémités de l'axe sur lequel elles sont modélisées. Cependant, en utilisant des splines cubiques naturelles avec des conditions sur les extrémités (dérivée seconde nulle en ces deux points extrêmes), nous avons limité ces instabilités.

2.4.2.3.b La linéarité

La fonction de poids telle qu'elle est considérée au sein du WCIE fait l'hypothèse qu'en chaque temps, l'intensité d'exposition a un effet linéaire sur le logit du risque de cancer. Dans la littérature, il a été montré que l'effet global du tabac pouvait atteindre un plateau pour de fortes intensités (Vineis et al. [2000]). L'effet de l'amiante sur le risque de cancer du poumon a été montré plutôt linéaire même s'il pouvait atteindre un plateau pour de fortes expositions (Nielsen et al. [2014]). De même, Lacourt et al. [2012] a montré que le risque de MP augmente fortement jusqu'à environ 1 f/mL puis augmente de façon moins importante au-delà de cette intensité moyenne cumulée (IME). Néanmoins, il est possible qu'en chaque temps, l'effet de l'intensité soit non linéaire. Par manque de littérature sur la prise en compte de la variabilité de l'intensité au cours de temps pour évaluer les relations entre expositions prolongées et cancer, nous n'avons pas de connaissance sur la linéarité de l'effet de l'intensité de ces expositions en chaque temps. Si cette forte hypothèse sur la linéarité de l'intensité en chaque temps semblait invalidée pour les relations étudiées, il faudrait donc pouvoir en étudier l'impact sur la forme de la fonction de poids.

2.5 Application au cancer du poumon

2.5.1 Article 2 publié dans Occupational and Environmental Medicine (OEM)

ORIGINAL ARTICLE

Time-dependent effect of intensity of smoking and of occupational exposure to asbestos on the risk of lung cancer: results from the ICARE case–control study

Emilie Lévêque,[1,2] Aude Lacourt,[2] Danièle Luce,[3] Marie-Pierre Sylvestre,[4,5] Pascal Guénel,[6] Isabelle Stücker,[6] Karen Leffondré[1]

► Additional material is published online only. To view please visit the journal online (http://dx.doi.org/10.1136/oemed-2017-104953).

[1]Université de Bordeaux, ISPED, INSERM, Bordeaux Population Health Research Center, Team Biostatistics, UMR 1219, Bordeaux, France
[2]Université de Bordeaux, INSERM, Bordeaux Population Health Research Center, Team EPICENE, UMR 1219, Bordeaux, France
[3]Université de Rennes, INSERM, EHESP, IRSET (Institut de recherche en santé, environnement et travail), UMR_S 1085, Pointe-à-Pitre, France
[4]Department of Social and Preventive Medicine, Montreal School of Public Health (ESPUM), University of Montreal, Montreal, Quebec, Canada
[5]Research Center, University of Montreal Health Center (CRCHUM), Montreal, Quebec, Canada
[6]INSERM, CESP, Cancer and Environment Team, Université Paris Saclay, Université de Paris-Sud, UVSQ, Villejuif, France

Correspondence to
Dr Aude Lacourt, Equipe EPICENE cancer et environnement, Bordeaux Population Health Center - Inserm U1219, Université Bordeaux - ISPED, 33076 Bordeaux Cedex - France; aude.lacourt@inserm.fr

Received 11 December 2017
Revised 28 March 2018
Accepted 27 April 2018

To cite: Lévêque E, Lacourt A, Luce D, *et al*. *Occup Environ Med* Epub ahead of print: [please include Day Month Year]. doi:10.1136/oemed-2017-104953

ABSTRACT

Objective To estimate the impact of intensity of both smoking and occupational exposure to asbestos on the risk of lung cancer throughout the whole exposure history.

Methods Data on 2026 male cases and 2610 male controls came from the French ICARE (Investigation of occupational and environmental causes of respiratory cancers) population-based, case–control study. Lifetime smoking history and occupational history were collected from standardised questionnaires and face-to-face interviews. Occupational exposure to asbestos was assessed using a job exposure matrix. The effects of annual average daily intensity of smoking (reported average number of cigarettes smoked per day) and asbestos exposure (estimated average daily air concentration of asbestos fibres at work) were estimated using a flexible weighted cumulative index of exposure in logistic regression models.

Results Intensity of smoking in the 10 years preceding diagnosis had a much stronger association with the risk of lung cancer than more distant intensity. By contrast, intensity of asbestos exposure that occurred more than 40 years before diagnosis had a stronger association with the risk of lung cancer than more recent intensity, even if intensity in the 10 years preceding diagnosis also had a significant effect.

Conclusion Our results illustrate the dynamic of the effect of intensity of both smoking and occupational exposure to asbestos on the risk of lung cancer. They confirm that the timing of exposure plays an important role, and suggest that standard analytical methods assuming equal weights of intensity over the whole exposure history may be questionable.

Key messages

What is already known about this subject?

- Occupational exposure to asbestos and smoking are well-known risk factors for lung cancer.
- However, certain questions on the specific dose–time–response relationship could still give rise to discussion, in particular on the contribution of recent intensities of exposure and on the relative impact of intensity of exposure at different ages.
- It remains unclear if recent exposure to asbestos or smoking contributes to the risk of lung cancer, and if the exposure intensity might have a different impact on the risk of lung cancer depending on the age at which the intensity applies.

What are the new findings?

- Intensity of smoking in the 10 years before lung cancer diagnosis had a much stronger association with the risk of lung cancer than more distant intensity of smoking.
- The intensity of occupational exposure to asbestos that occurred more than 40 years before diagnosis had a stronger association with the risk of lung cancer than more recent intensity, even if intensity in the 10 years preceding diagnosis also had a significant effect.

How might this impact on policy or clinical practice in the foreseeable future?

- The estimation of time-dependent effect of intensity of smoking and of occupational exposure to asbestos may help deciders in planning epidemiological surveillance and prevention programmes.

INTRODUCTION

Occupational exposure to asbestos and smoking are well-known risk factors for lung cancer. However, certain questions on the specific dose–time–response relationship could still give rise to discussion, in particular on the contribution of recent doses of exposure and on the relative impact of intensity of exposure at different ages. It remains indeed unclear if recent exposures to asbestos or smoking contribute to the risk of lung cancer, and thus if they should be fully ignored in the statistical analysis, and for how many months or years before diagnosis. Most often, and in particular for asbestos, epidemiological studies adopt a lagging exposure approach to discount the years of exposure before lung cancer diagnosis in the derivation of the cumulative dose of exposure. The duration of the lag may be a priori chosen.[1–3] It may also be the result of statistical criteria like the value that maximises the model fit to data,[4–6] although this approach may lead to CIs that are too narrow.[7] Such practices resulted in a large variability in the lag chosen for occupational exposure to asbestos in epidemiological studies on lung cancer, ranging

from 0,[4 8] 5 years,[3]10 years,[2 6 9] to 30 years.[10] In addition, the resulting estimate of the lag may be potentially biased if the cumulative dose of exposure does not account for the timing of exposure, in particular for intensity of exposure. Indeed, a strong intensity of asbestos exposure in early career may not conduct to the same risk of lung cancer than a strong intensity in a later period, even if the total cumulative dose remains equal.

Regarding the role of age at exposure, some studies have attempted to estimate the impact on lung cancer of age at first occupational exposure to asbestos[11] or at smoking initiation.[12–14] However, they did not investigate the impact of intensity of exposure at different ages. Yet we might assume that if there is an age-dependent susceptibility, then the exposure intensity might have a different impact on the risk of lung cancer depending on age at which the intensity applies.

The objective of this study was to estimate the impact of the intensity of both smoking and occupational exposure to asbestos in each year of the exposure history. In particular, we estimated the relative impact of exposure intensity at each year preceding lung cancer diagnosis and at each age of exposure history. To this aim, we used a flexible weighted cumulative index of exposure (WCIE)[15 16] and data from the ICARE (Investigation of occupational and environmental causes of respiratory cancers) case–control study.[17]

MATERIALS AND METHODS

Study design

ICARE is a large multicenter population-based case–control study on cancers of the respiratory tract (lung, oral cavity, pharynx, sinonasal cavities, larynx), conducted in 10 French territorial *départements* in 2001–2007.[17] In the present analysis, we focused on men and lung cancer only. The cases were aged 18–75 years and all with histologically confirmed primary lung cancer (C33–34, International Classification of Diseases for Oncology), which included all histological types. Each case was interviewed within 3 months of diagnosis. Controls were randomly selected through incidence density sampling every 2 months, using a random digit dialling procedure. Controls were frequency-matched to cases by sex, age strata (<40, 40–54, 55–64, >65 years) and residence area (*département*).[17] Random sampling of controls was also stratified on socioeconomic status within each sex–age–residence group to adequately represent the general population.[17]

Data collection and exposure assessment

Information on demographic characteristics, lifetime occupational history and smoking history was collected with a detailed standardised questionnaire during face-to-face interviews by trained interviewers.[17] For smoking history, subjects had to report their average number of cigarettes smoked per day by intervals of constant smoking habits. In the present analysis, we referred to the reported average number of cigarettes smoked per day in a given year as the 'annual average daily intensity of exposure to smoking' in that year.

Occupational exposure to asbestos was assessed by a job exposure matrix (JEM), which has been used elsewhere.[18 19] In the JEM, each job was defined as a combination of an occupation (defined according to the International Standard Classification of Occupation edition 1968) and an industry (defined according to either the International Standard Industrial Classification Revision 2 or the nomenclature of French activities classification edition 1999). Industrial hygienists and occupational health experts assigned for each job of the JEM an estimation of a probability, a frequency and an intensity of exposure to asbestos on a semiquantitative scale. Estimations were based on a combination of both the knowledge of occupational asbestos exposure in France and direct measurements in occupational settings, and accounted for asbestos exposure level changes in France over calendar time.[20] For each job, intensity of exposure was defined as the estimated annual average daily concentration of asbestos fibres in the air at workplace for that job (from 0.0005 equivalent fibres per mL for very low intensity, to 20 equivalent fibres per mL for very high intensity; online supplementary table 1S). Probability of exposure was the estimated percentage of workers exposed to asbestos for that job (from 0 for no exposure, to 0.85 for definite exposure). Frequency of exposure was the estimated proportion of exposed worktime on a typical 8-hour working day for that job (from 0.025 for sporadic exposure, to 0.85 for continuous exposure). For each job, we defined the level of exposure as the product of intensity, probability and frequency, which was thus expressed in equivalent fibers per mL. For each subject, we thus could derive the level of exposure for each year of his job history. When several jobs were occupied within a year, we calculated the mean level of exposure over the year prorated by the duration of each job occupied within that year. In the present analysis, we refer to this level of exposure in a given year as the 'annual average daily intensity of occupational exposure to asbestos' in that year.

Statistical analysis

The present analysis has been performed on subjects who had complete information on smoking and occupational asbestos history. Unconditional logistic regression models were used to estimate the dynamic effects of the intensity of occupational exposure to asbestos and smoking on the risk of lung cancer.

For occupational exposure to asbestos, the model included i) the WCIE for asbestos to estimate the dynamic effect over lifetime of the annual average daily intensity of occupational exposure to asbestos (in equivalent fibres per mL); ii) the Comprehensive Smoking Index (CSI), which is a single aggregate measure for each subject of his lifetime smoking history derived from his reported average number of cigarettes smoked per day over lifetime, total duration of smoking and time since smoking cessation at the index date (diagnosis for cases and interview for controls)[21 22]; iii) indicators of residential area (*département*); and iv) a spline function of age (in years) at the index date to adjust for its non-linear effect on the logit of lung cancer. The spline function was a natural cubic spline function with three knots (at 5th, 50th and 95th percentiles of the age distribution of the entire population). For smoking, the model included i) the WCIE for smoking to estimate the dynamic effect over lifetime of the annual average daily intensity of smoking (in number of cigarettes smoked per day); ii) indicators of the quartiles of the unweighted total cumulative dose of occupational exposure to asbestos[18]; iii) indicators of residential area; and iv) a spline function of age (in years) at the index date (the same as in the model for asbestos). Smoking and occupational exposure to asbestos were thus systematically mutually adjusted for, without interaction. In a sensitivity analysis, we further adjust all models for education (high school or university vs less).

In each model, the WCIE was written as follows:

$$WCIE = \sum_t w(t)\, x(t)$$

where $x(t)$ was the annual average daily intensity of the exposure in year t, and $w(t)$ its weight. In the WCIE for asbestos, $x(t)$ was the annual average daily intensity of occupational exposure

to asbestos (in equivalent fibers per mL) in year t. In the WCIE for smoking, $x(t)$ was the annual average daily intensity of exposure to smoking (in number of cigarettes smoked per day) in year t. Year t was either the year preceding the index date (to assess the relative impact of recent intensity compared with more distant intensity) or the year of age (to assess the relative impact of intensity at different ages). Overall, we thus estimated two models for asbestos (Model 1) and two models for smoking (Model 2), each of them using either the year before the index date (Models 1a and 2a) or the year of age (Models 1b and 2b) as the time axis.

As opposed to the standard lagging approach which usually assigns a weight $w(t)$ of 0 for the years t just before index date ($w(t)=0$ if $t<$lag) and 1 to all other years ($w(t)=1$ if $t\geq$lag), the weights $w(t)$ in the present study were estimated from the data without imposing any assumption on their values or functional form. To this end, we used cubic B-spline functions.[15 16] Such spline functions just require choosing a number of interior knots (in addition to the two knots placed at the extreme of the entire exposure time window) and their position throughout the entire exposure time window. We used one to three interior knots placed at equal distance and selected the number of knots that provided the best fit to data according to Akaike's information criterion (AIC). For each model, all regression coefficients (including β_1 associated with the WCIE) and the time-dependent weight function $w(t)$ in the WCIE were estimated simultaneously in a single step via constrained maximum likelihood estimation. The weights $w(t)$ were constrained to be non-negative and standardised, as described in Hauptmann *et al*.[15] For that purpose, we used the SAS PROC NLP with the double-dogleg algorithm. Bootstrap percentiles were further used to compute 95% pointwise CIs of the weight function $w(t)$.

In each model, the OR associated with a one-unit increase in $x(t)$, the annual average daily intensity of exposure in year t, equalled the exponential of $\beta_1 w(t)$. This time-dependent OR was inherently adjusted for the intensity of exposure in *all other years*, as well as for all other covariates included in the model. The implicit adjustment for the exposure history in *all other years* avoided the need to further adjust for the time at which exposure started or stopped, since the exposure intensity was by definition null outside the exposure time window.

RESULTS

Of 2276 male lung cancer cases and 2780 male controls participating in the ICARE case–control study, 250 cases and 170 controls were excluded because of either missing data on smoking history or an incomplete job history, or an unavailable evaluation of occupational asbestos for at least one job. Our present analysis thus included 2026 cases and 2610 controls. Distributions of sociodemographic variables were similar between the included and excluded subjects (online supplementary table 2S). As expected, the proportions of ever-smokers and ever-occupationally exposed to asbestos were higher in cases than controls (table 1), and ever-exposed cases were much more strongly exposed than ever-exposed controls (table 2).

Intensity of asbestos exposure received 40–60 years prior to diagnosis tended to have a stronger impact than more recent intensity (Model 1a in figure 1A and table 3), even if CIs were large beyond 50 years (figure 1A) because of fewer subjects contributing to that time window (Model 1a in table 3). In the same line, the intensity of asbestos exposure received in ages 15–25 years had a much stronger impact than the intensity received at later ages (Model 1b in figure 1B and table 3). Interestingly, recent intensity of asbestos exposure (received in the 10 years before lung cancer diagnosis) had a significant weight (Model 1a in figure 1A). As a result, when we compared two subjects who had exactly the same intensity of asbestos exposure over lifetime except within the 10 years before diagnosis, the subject who had a stronger exposure intensity in that time window had a significantly higher risk of lung cancer (OR=1.20, 95% CI (1.10 to 1.31) for a difference of 1 fiber/mL/day in each year of the last 10 years; Model 1a in table 3).

Intensity of smoking in the 10 years before lung cancer diagnosis had a much stronger association with the risk than more distant intensity (Model 2a in figure 2A and table 3). Our data do not suggest that intensity of smoking in ages 15–25 years would have a stronger impact than intensity at later ages (Model 2b in table 3 and figure 2B), even if the weights in that age window were significantly non-null. Indeed, when comparing two subjects who had exactly the same intensity of smoking over lifetime except from ages 15–25 years, the subject who smoked more during that age window had a significantly higher risk of lung cancer (OR=1.32, 95% CI (1.22 to 1.45) for a difference of 10 cigarettes per day in each year of ages 15–25 years; Model 2b in table 3).

All the results were similar when further adjusted for education (online supplementary figure 1S).

Table 1 Characteristics of cases and controls at the index date, ICARE case–control study, France, 2000–2007

Characteristics	Cases (n=2026)		Controls (n=2610)	
Age in years (mean, standard deviation)	60.3	9.0	58.2	9.9
Education level (n, %)				
Elementary school or less	600	29.6	489	18.7
Middle school	779	38.5	1028	39.4
High school	177	8.7	293	11.2
University	253	12.5	693	26.6
Other	21	1.0	18	0.7
Missing	196	9.7	89	3.4
Area of residence (*département*) (n, %)				
Calvados	240	11.8	336	12.9
Hérault	227	11.2	343	13.1
Isère	346	17.1	375	14.4
Loire-Atlantique	255	12.6	297	11.4
Manche	225	11.1	222	8.5
Bas-Rhin	247	12.2	331	12.7
Haut-Rhin	53	2.6	88	3.4
Somme	224	11.1	365	14.0
Vendée	106	5.2	144	5.5
Doubs-Territoire de Belfort	103	5.1	109	4.1
Smoking status (n, %)				
Ever-regular smoker*	1969	97.7	1838	70.4
Never smoker	57	2.9	772	29.6
Occupational exposure to asbestos (n, %)				
Ever-exposed†	1417	69.9	1520	58.2
Never exposed	609	30.1	1090	41.8

*Subjects who have smoked at least one cigarette per day during at least 1 year over their entire lifetime before the index date.
†Subjects who have been exposed to asbestos with a non-null probability of exposure in at least one job over their entire occupational history before the index date.

Table 2 Characteristics of occupational exposure to asbestos in ever-exposed to asbestos and smoking characteristics in ever-smokers, ICARE case–control study, France, 2001–2007

	Occupational exposure to asbestos		Smoking	
	Ever-exposed cases (n=1417)	Ever-exposed controls (n=1520)	Ever-smoking cases (n=1969)	Ever-smoking controls (n=1838)
Total duration of exposure (in years)				
Mean (SD)	27.2 (14.6)	24.6 (14.7)	37.8 (10.7)	25.9 (13.4)
Median (IQR)	31 (14–40)	26 (10–38)	39 (31–45)	26 (16–36)
Range	1–56	1–57	2–64	1–62
Average intensity over lifetime (in fibers/mL/day or number of cigarettes smoked per day)				
Mean (SD)	0.30 (0.64)	0.19 (0.43)	22.5 (10.5)	15.5 (9.8)
Median (IQR)	0.04 (0.002–0.38)	0.02 (0.001–0.17)	20 (17–28)	15 (9–20)
Range	2.0×10^{-6}–10.4	3.0×10^{-7}–6.9	0.1–82	0.07–68
Age at first exposure (in years)				
Mean (SD)	18.2 (5.9)	18.7 (6.2)	17.1 (3.4)	17.8 (4.1)
Median (IQR)	16 (14–20)	17 (14–21)	17 (15–19)	17 (15–20)
Range	12–57	12–56	12–50	12–58
Time since first exposure (in years)				
Mean (SD)	41.8 (10.8)	39.6 (12.2)	43.3 (9.1)	40.3 (10.1)
Median (IQR)	42 (35–50)	41 (30–49)	44 (37–50)	41 (33–48)
Range	1–61	2–61	5–64	5–63
Time since last exposure (in years)				
Mean (SD)	15.7 (14.7)	15.9 (15)	6.6 (9.2)	15.5 (13.6)
Median (IQR)	11 (4–27)	10.5 (3–26)	2 (1–10)	13 (1–26)
Range	0–59	0–57	0–55	0–55
Unweighted cumulative dose of exposure (in fibres/mL-years or cigarette-years)				
Mean (SD)	9.14 (22.7)	5.13 (13.5)	839 (474)	401 (354)
Median (IQR)	1.00 (0.05–9.5)	0.33 (0.02–3.74)	772 (522–1048)	330 (130–568)
Range	3.1×10^{-6}–372	9.4×10^{-7}–222	0.5–3275	0.2–2824

SD, standard deviation; Med, median; IQR, interquartile range

DISCUSSION

Our results indicate that intensity of smoking in the 10 years before diagnosis would have far the strongest contribution to the risk of lung cancer compared with more distant intensity of smoking. The intensity of recent occupational exposure to asbestos also contributed to the risk of lung cancer, although intensity that occurred more than 40 years before diagnosis had a stronger impact. These results suggest that discounting recent smoking, as well as recent exposure to asbestos, in the cumulative dose of exposure over lifetime, is questionable for lung cancer.

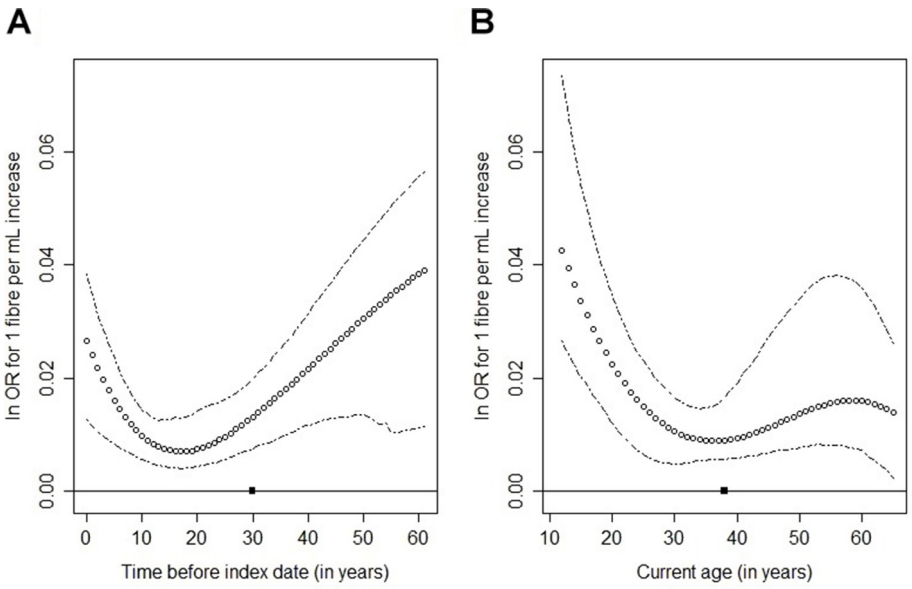

Figure 1 Estimated time-dependent effect of the annual average daily intensity of occupational exposure to asbestos using years before the index date (Model 1a, panel A) or age (Model 1b, panel B) as the time axis, adjusted for smoking (CSI), age at the index date (in years) and area of residence (*département*). The black dot on the time axis shows the position of the inner knot of the B-spline time-dependent weight function. ICARE case–control study, France, 2001–2007.

Table 3 Estimated time-dependent effects of the annual average daily intensity of occupational exposure to asbestos and smoking on the risk of lung cancer, ICARE case–control study, France, 2001–2007

	Occupational exposure to asbestos		Smoking	
Exposure time window	Mean annual number of subjects*	OR for a 1 fiber/mL difference in the annual average daily intensity each year of the specific exposure time-window (95% CI)†	Mean annual number of subjects*	OR for a 10 cigarettes/day difference in the annual average daily intensity each year of the specific exposure time-window (95% CI)‡
Years before index date		Model 1a		Model 2a
0–10	2384	1.20 (1.10 to 1.31)	2638	2.02 (1.96 to 2.36)
10–20	3387	1.09 (1.05 to 1.15)	3440	1.36 (1.31 to 1.46)
20–30	3551	1.11 (1.06 to 1.19)	3787	1.22 (1.18 to 1.32)
30–40	3151	1.21 (1.11 to 1.32)	3402	1.27 (1.22 to 1.37)
40–50	3000	1.33 (1.15 to 1.52)	2063	1.39 (1.21 to 1.51)
50–60	800	1.46 (1.21 to 1.73)	615	1.49 (1.17 to 2.30)
Years of age		Model 1b		Model 2b
15–25	3595	1.29 (1.15 to 1.49)	3645	1.32 (1.22 to 1.45)
25–35	3788	1.13 (1.06 to 1.21)	4204	1.34 (1.29 to 1.41)
35–45	3449	1.11 (1.07 to 1.24)	3626	1.44 (1.38 to 1.53)
45–55	2651	1.16 (1.09 to 1.44)	2569	1.58 (1.49 to 1.68)
55–65	1217	1.18 (1.08 to 1.46)	1261	1.58 (1.44 to 1.77)

OR, odds ratio; CI, confidence interval

* Mean annual number of subjects who contributed to the specific exposure time - window.

† Adjusted for asbestos exposure intensity in all other time-windows, smoking history (CSI), age at the index date (in years, spline) and area of residence (*département*).

‡Adjusted for smoking intensity in all other time - windows, cumulative occupational exposure to asbestos (quartiles of the standard cumulative index of occupational exposure to asbestos), age at the index date (in years, spline) and area of residence (*département*).

Regarding age-dependent susceptibility, our data do not suggest that the intensity of smoking before the age of 25 years would contribute more than the intensity at later ages, even if intensity in these young ages had a significant association with the risk of lung cancer. By contrast, the intensity of occupational exposure to asbestos before the age of 25 years had a stronger association with the risk of lung cancer than intensity at later age.

Few studies empirically assessed the time-dependent effect of annual average daily intensity of smoking or occupational exposure to asbestos over lifetime on the risk of lung cancer, without a priori imposing a lag period or a parametric function of the weight of intensity. Hauptmann *et al* used the WCIE,[15] and another similar statistical approach,[23] to assess the impact of pack-years at each year before lung cancer diagnosis in a German case–control study. As in the present study, they found that cigarettes smoked in the 10 years before diagnosis had the strongest contribution to the risk of lung cancer. These results are consistent with a previous study based on a multistage model and cohort data which found that promotion seems to be the dominant aetiological mechanism in smoking-related lung carcinogenesis.[24] For age-dependent susceptibility, Hauptmann *et al*[15] mentioned also that they did not find any evidence that intensity of smoking

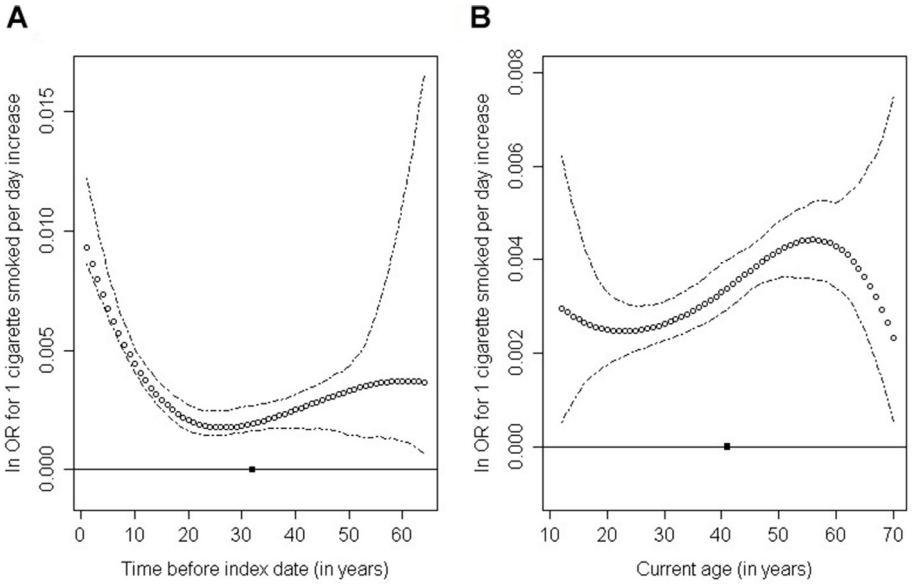

Figure 2 Estimated time-dependent effect of the annual average daily intensity of smoking (in number of cigarettes smoked per day) using years before the index date (Model 2a, panel A) or age in years (Model 2b, panel B) as the time axis, adjusted for occupational exposure to asbestos (quartiles of a standard cumulative index of occupational exposure to asbestos), age at the index date (in years) and area of residence (*département*). The black dot on the time axis shows the position of the inner knot of the B-spline time-dependent weight function. ICARE case–control study, France, 2001–2007.

at young ages was associated with a higher risk of lung cancer compared with intensity at later ages. These results are consistent with studies that did not find any association between age at smoking initiation and lung cancer risk after adjustment for both cigarette-years and time since smoking cessation in men.[12 25] For occupational exposure to asbestos, Hauptmann *et al*[3] also used the WCIE to analyse pooled data from two German case–control studies. They reported that doses of asbestos exposure that contributed the most to the risk of lung cancer were received 10–15 years before diagnosis. However, they acknowledged that their data were insufficient to draw conclusion beyond 25 years before diagnosis.[3] Similar results were obtained using a parametric time window approach imposing a lag of 10 years to analyse lung cancer mortality data among workers employed in a plant producing asbestos products in South Carolina in 1940–1965.[26] By contrast, our results suggest that the doses of asbestos received more than 40 years before diagnosis had the strongest significant impact on the risk of lung cancer, despite large CI of the weights beyond 50 years. Our results are in agreement with other studies that used more standard statistical methods.[27 28] The main reason that may explain the difference of results between our study and the studies by Hauptmann *et al*[3] or Richardson *et al*[26] is likely that, because of differences in contexts and study designs, our subjects were occupationally exposed to asbestos for a much longer duration than their subjects (mean duration of 27.2 and 24.6 years in our cases and controls, respectively, vs 7.1 and 5.6 years in Hauptmann *et al*,[3] and the median duration of employment in the plant of 1.1 years in Richardson *et al*[26 29]). We thus had potentially much more information on the exposure time window 0–60 years before diagnosis in our study. Another reason that may partly explain the discrepancies of results between studies on occupational exposure to asbestos and lung cancer is the method used to adjust for smoking. While Hauptmann *et al*[3] adjusted for time since smoking cessation and cigarette-years using separate variables, we used the CSI which accounted for their interaction and has been shown to provide a better fit to our data,[22 30] as well as to other case–control data on lung cancer.[21]

Our study has other strengths and limitations. As in most case–control studies, smoking and occupational histories were self-reported. We may concede that exposures that occurred many years in the past had stronger measurement error than more recent exposures. However, several previous studies have shown good validity of self-reported occupational histories.[31] Misclassification of exposures in JEM has also been shown to be non-differential.[32] Hauptmann *et al*[15] have shown that increasing measurement error with time before interview had a weak impact on the shape of the estimated weight function in the WCIE, and that it was robust to non-differential measurement errors.[15]

The use of cubic B-splines in the WCIE allowed us to estimate the time-dependent effect of intensity of both smoking and occupational exposure to asbestos over lifetime, without imposing any prior assumption on the shape of the time-dependent weight function. However, one important limitation of cubic B-splines is that the results may be sensitive to the number and position of the knots used. The results may be particularly sensitive for small sample sizes, and thus in exposure time window where only few subjects were exposed. This explains why we obtained large CIs for the effect of intensity received more than 50 years before lung cancer diagnosis for both exposures. The results obtained for this very distant exposure time window should thus be interpreted with caution. Moreover, while we did not impose any prior assumption on the shape of the time-dependent weight function, we assumed that the intensity of exposure at each time had a linear effect on the logit of lung cancer. Yet the effect of intensity of smoking has been observed to reach a plateau at high level of intensity in several studies.[33–35] This plateau could be explained by a saturation of metabolism or increasing DNA repair capacity with increasing intensity,[36] or increasing misclassification with increasing reported intensity.[37] Note that a plateau has also been suggested for the effect of cumulative dose of occupational exposure to asbestos on the risk of lung cancer.[9] To relax the linearity assumption of the effect of intensity at each time, it would be of interest to use a spline function in the WCIE for the time-dependent weight and for intensity itself, as for example proposed by Berhane *et al*,[38] although results might be difficult to interpret because these were expressed through a three-dimensional graph.

The use of two different time axes in the WCIE allowed us to illustrate the time-dependent effect of intensity with respect to time before diagnosis or age at exposure. However, we may wonder whether the estimated effect of intensity, for example 50–60 years before diagnosis, actually reflects the effect of intensity at ages 15–25 years. Because the two time axes are highly correlated and even perfectly collinear for a given attained age, it is very difficult to clearly disentangle the two effects.[39] We may also wonder whether the effect of recent intensity depends on attained age. This could be investigated by stratifying on age at index date, but this would require rather fine strata and sufficient data within each strata. Further studies with very large sample size are needed to address this question.

Finally, to reduce computational issues, we estimated the WCIE for smoking and the WCIE for asbestos in two separate regression models. The WCIE for smoking was adjusted for the unweighted cumulative dose of asbestos exposure, and the WCIE for asbestos was adjusted for the CSI. However, we did not investigate the potential interaction between smoking and asbestos. Previous studies show inconsistent results on the nature of their potential interaction,[9 40] and further studies are needed to investigate if the effect of one at a given time may depend on the level of the other.

CONCLUSION

Our results confirm that the timing of exposure to smoking and asbestos plays an important role in the risk of lung cancer, as for many other protracted exposures and chronic diseases.[41] They confirm that intensity of smoking in the 10 years before diagnosis has a much stronger association with lung cancer than more distant intensity, and suggest that intensity of occupational exposure to asbestos that occurred more than 40 years before diagnosis had a stronger effect than more recent intensity, even if intensity in the 10 years before diagnosis also had a significant effect. These results suggest that discounting recent exposures in lung cancer studies may be questionable for smoking and for asbestos exposure.

Contributors EL performed all statistical analyses and drafted the first version of the manuscript. AL contributed to the drafting of the manuscript and cosupervised all aspects of the manuscript. IS and DL supervised data collection within the ICARE study and contributed to the interpretation of the results. M-PS and PG contributed to the interpretation of the results in the final version. KL contributed to the drafting of the manuscript and supervised all aspects of the manuscript. All coauthors participated in the editing and correction of the final text.

Funding The French National Research Program for Environmental and Occupational Health of Anses with support of the Cancer TMOI of the French National Alliance for Life and Health Sciences (AVIESAN) – 2013/1/177.

Competing interests None declared.

Patient consent Obtained.

Ethics approval The Institutional Review Board of the French National Institute of Health and Medical Research (IRB-Inserm, n° 01-036) and by the French Data Protection Authority (CNIL n° 90120).

Provenance and peer review Not commissioned; externally peer reviewed.

REFERENCES

1 Clin B, Morlais F, Launoy G, *et al*. Cancer incidence within a cohort occupationally exposed to asbestos: a study of dose-response relationships. *Occup Environ Med* 2011;68:832–6.
2 Edwards JK, Cole SR, Chu H, *et al*. Accounting for outcome misclassification in estimates of the effect of occupational asbestos exposure on lung cancer death. *Am J Epidemiol* 2014;179:641–7.
3 Hauptmann M, Pohlabeln H, Lubin JH, *et al*. The exposure-time-response relationship between occupational asbestos exposure and lung cancer in two German case-control studies. *Am J Ind Med* 2002;41:89–97.
4 Olsson AC, Vermeulen R, Schüz J, *et al*. Exposure-Response Analyses of Asbestos and Lung Cancer Subtypes in a Pooled Analysis of Case-Control Studies. *Epidemiology* 2017;28:288–99.
5 Deng Q, Wang X, Wang M, *et al*. Exposure-response relationship between chrysotile exposure and mortality from lung cancer and asbestosis. *Occup Environ Med* 2012;69:81–6.
6 Elliott L, Loomis D, Dement J, *et al*. Lung cancer mortality in North Carolina and South Carolina chrysotile asbestos textile workers. *Occup Environ Med* 2012;69:385–90.
7 Richardson DB, Cole SR, Chu H, *et al*. Lagging exposure information in cumulative exposure-response analyses. *Am J Epidemiol* 2011;174:1416–22.
8 Gustavsson P, Nyberg F, Pershagen G, *et al*. Low-dose exposure to asbestos and lung cancer: dose-response relations and interaction with smoking in a population-based case-referent study in Stockholm, Sweden. *Am J Epidemiol* 2002;155:1016–22.
9 Nielsen LS, Bælum J, Rasmussen J, *et al*. Occupational asbestos exposure and lung cancer--a systematic review of the literature. *Arch Environ Occup Health* 2014;69:191–206.
10 Loomis D, Dement JM, Wolf SH, *et al*. Lung cancer mortality and fibre exposures among North Carolina asbestos textile workers. *Occup Environ Med* 2009;66:535.
11 Kang D, Myung MS, Kim YK, *et al*. Systematic Review of the Effects of Asbestos Exposure on the Risk of Cancer between Children and Adults. *Ann Occup Environ Med* 2013;25:10.
12 Benhamou S, Benhamou E. The effect of age at smoking initiation on lung cancer risk. *Epidemiology* 1994;5:560.
13 Frost G, Darnton A, Harding AH. The effect of smoking on the risk of lung cancer mortality for asbestos workers in Great Britain (1971-2005). *Ann Occup Hyg* 2011;55:239–47.
14 Hegmann KT, Fraser AM, Keaney RP, *et al*. The effect of age at smoking initiation on lung cancer risk. *Epidemiology* 1993;4:444–8.
15 Hauptmann M, Wellmann J, Lubin JH, *et al*. Analysis of exposure-time-response relationships using a spline weight function. *Biometrics* 2000;56:1105–8.
16 Sylvestre MP, Abrahamowicz M. Flexible modeling of the cumulative effects of time-dependent exposures on the hazard. *Stat Med* 2009;28:3437–53.
17 Luce D, Stücker I. ICARE Study Group. Investigation of occupational and environmental causes of respiratory cancers (ICARE): a multicenter, population-based case-control study in France. *BMC Public Health* 2011;11:928.
18 Lacourt A, Leffondré K, Gramond C, *et al*. Temporal patterns of occupational asbestos exposure and risk of pleural mesothelioma. *Eur Respir J* 2012;39:1304–12.
19 Lacourt A, Lévêque E, Guichard E, *et al*. Dose-time-response association between occupational asbestos exposure and pleural mesothelioma. *Occup Environ Med* 2017;74:691.
20 Févotte J, Dananché B, Delabre L, *et al*. Matgéné: a program to develop job-exposure matrices in the general population in France. *Ann Occup Hyg* 2011;55:865–78.
21 Leffondré K, Abrahamowicz M, Xiao Y, *et al*. Modelling smoking history using a comprehensive smoking index: application to lung cancer. *Stat Med* 2006;25:4132–46.
22 Papadopoulos A, Guida F, Leffondré K, *et al*. Heavy smoking and lung cancer: are women at higher risk? Result of the ICARE study. *Br J Cancer* 2014;110:1385–91.
23 Hauptmann M, Lubin JH, Rosenberg P, *et al*. The use of sliding time windows for the exploratory analysis of temporal effects of smoking histories on lung cancer risk. *Stat Med* 2000;19:2185–94.
24 Hazelton WD, Clements MS, Moolgavkar SH. Multistage carcinogenesis and lung cancer mortality in three cohorts. *Cancer Epidemiol Biomarkers Prev* 2005;14:1171–81.
25 Leffondré K, Abrahamowicz M, Siemiatycki J, *et al*. Modeling smoking history: a comparison of different approaches. *Am J Epidemiol* 2002;156:813–23.
26 Richardson DB, MacLehose RF, Langholz B, *et al*. Hierarchical latency models for dose-time-response associations. *Am J Epidemiol* 2011;173:695–702.
27 Hillerdal G, Henderson DW. Asbestos, asbestosis, pleural plaques and lung cancer. *Scand J Work Environ Health* 1997;23:93–103.
28 Selikoff IJ, Hammond EC, Seidman H. Latency of asbestos disease among insulation workers in the United States and Canada. *Cancer* 1980;46:2736–40.
29 Hein MJ, Stayner LT, Lehman E, *et al*. Follow-up study of chrysotile textile workers: cohort mortality and exposure-response. *Occup Environ Med* 2007;64:616–25.
30 Menvielle G, Franck JE, Radoï L, *et al*. ICARE study group. Quantifying the mediating effects of smoking and occupational exposures in the relation between education and lung cancer: the ICARE study. *Eur J Epidemiol* 2016;31:1213–21.
31 Teschke K, Olshan AF, Daniels JL, *et al*. Occupational exposure assessment in case-control studies: opportunities for improvement. *Occup Environ Med* 2002;59:575–94.
32 Bouyer J, Dardenne J, Hémon D. Performance of odds ratios obtained with a job-exposure matrix and individual exposure assessment with special reference to misclassification errors. *Scand J Work Environ Health* 1995;21:265–71.
33 Rachet B, Siemiatycki J, Abrahamowicz M, *et al*. A flexible modeling approach to estimating the component effects of smoking behavior on lung cancer. *J Clin Epidemiol* 2004;57:1076–85.
34 Lubin JH. Cigarette Smoking and Lung Cancer: Modeling Total Exposure and Intensity. *Cancer Epidemiology Biomarkers & Prevention* 2006;15:517–23.
35 Vlaanderen J, Portengen L, Schüz J, *et al*. Effect modification of the association of cumulative exposure and cancer risk by intensity of exposure and time since exposure cessation: a flexible method applied to cigarette smoking and lung cancer in the SYNERGY Study. *Am J Epidemiol* 2014;179:290–8.
36 Vineis P, Kogevinas M, Simonato L, *et al*. Levelling-off of the risk of lung and bladder cancer in heavy smokers: an analysis based on multicentric case-control studies and a metabolic interpretation. *Mutat Res* 2000;463:103–10.
37 Boffetta P, Clark S, Shen M, *et al*. Serum cotinine level as predictor of lung cancer risk. *Cancer Epidemiol Biomarkers Prev* 2006;15:1184–8.
38 Berhane K, Hauptmann M, Langholz B. Using tensor product splines in modeling exposure-time-response relationships: application to the Colorado Plateau Uranium Miners cohort. *Stat Med* 2008;27:5484–96.
39 Richardson DB, Cole SR, Langholz B. Regression models for the effects of exposure rate and cumulative exposure. *Epidemiology* 2012;23:892–9.
40 El Zoghbi M, Salameh P, Stücker I, *et al*. Absence of multiplicative interactions between occupational lung carcinogens and tobacco smoking: a systematic review involving asbestos, crystalline silica and diesel engine exhaust emissions. *BMC Public Health* 2017;17:156.
41 Vermeulen R, Chadeau-Hyam M. Dynamic aspects of exposure history-do they matter? *Epidemiology* 2012;23:900–1.

Supplementary materials

Table 1S Numerical values of probability, frequency, and intensity of asbestos exposure used in the Job Exposure Matrix (JEM) to derive individual average annual daily intensity of exposure.

Asbestos exposure characteristics	Definition	Numerical values used to calculate annual doses		
Probability of exposure (% of workers exposed)				
Non exposed	0	0		
Possible	> 0 - 5	0.025		
Probable	5 - 30	0.175		
Likely	30 - 70	0.5		
Definite	≥ 70	0.85		
Frequency of exposure (% of work time)				
Sporadic	> 0-5	0.025		
Occasional	5-30	0.175		
Frequent	30-70	0.5		
Continuous	≥ 70	0.85		
Intensity of exposure (equivalent fibres/ml)*		Passive exposure	Indirect exposure	Direct exposure
Very low	> 0 - 0.01	0.0005	0.0025	0.005
Low	0.01 - 0.1	0.005	0.025	0.05
Medium	0.1 - 1	0.05	0.25	0.5
High	1 - 10	0.5	2.5	5
Very high	≥ 10	2	10	15

* Intensity of exposure was defined as a combination of the intensity of exposure due to specific task and work environment contamination. Since the asbestos JEM was based on expert judgment, intensity of exposure was expressed in equivalent fibres/ml. Three types of exposure were defined: Passive exposure (workers were exposed according to diffuse contamination of buildings); indirect exposure (workers were exposed by other workers using asbestos materials); direct exposure (workers used directly asbestos materials).

Table 2S Socio-demographics characteristics of cases and controls that were included or excluded* from the present analysis, ICARE case control study, France, 2000-2007.

	Cases				Controls			
	Included (n=2026)		Excluded* (n=250)		Included (n=2610)		Excluded* (n=170)	
Age in years (mean, standard deviation)								
	60.3	9.0	60.4	9.6	58.2	9.9	56.1	9.9
Area of residence (*département*) (n, %)								
Calvados	240	11.8	32	12.8	336	12.9	22	12.9
Hérault	227	11.2	25	10.0	343	13.1	17	10.0
Isère	346	17.1	25	10.0	375	14.4	32	18.8
Loire Atlantique	255	12.6	18	7.2	297	11.4	14	8.2
Manche	225	11.1	37	14.8	22	8.5	25	14.7
Bas-Rhin	247	12.2	55	22.0	331	12.7	29	17.1
Haut-Rhin	53	2.63	3	1.2	88	3.4	1	0.6
Somme	224	11.1	45	18.0	365	14.0	22	12.9
Vendée	106	5.2	7	2.8	144	5.5	5	2.9
Doubs-Territoire de Belfort	103	5.1	3	1.2	109	4.1	3	1.8
Highest education level (n, %)								
Elementary school or less	600	29.6	75	30.0	489	18.7	32	18.8
Middle school	779	38.5	90	36.0	1028	39.4	53	31.2
High school	177	8.7	8	3.2	293	11.2	17	10.0
University	253	12.5	20	8.0	693	26.6	59	34.7
Other	21	1.0	4	1.6	18	0.7	1	0.6
Missing	196	9.7	53	21.2	89	3.4	8	4.7

* Subjects were excluded because of incomplete information on smoking history and/or occupational exposure to asbestos over lifetime.

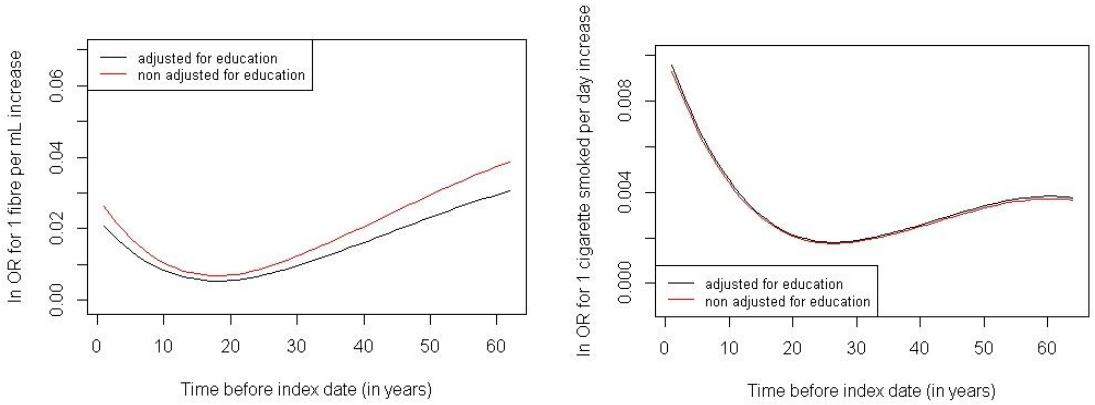

Figure 1S. Impact of adjustment for education (high school or university versus less) on estimated relationships. Left panel: Estimated time-dependent effect of the annual average daily intensity of occupational exposure to asbestos, adjusted for smoking (CSI), age at the index date (in years), and area of residence (*département*), with or without education. Right panel: Estimated time-dependent effect of the annual average daily intensity of smoking (in number of cigarette smoked per day), adjusted for occupational exposure to asbestos (quartiles of a standard cumulative index of occupational exposure to asbestos), age at the index date (in years), and area of residence (*département*), with or without education. ICARE case-control study, France, 2001-2007

2.5.2 Discussion complémentaire à l'article 2

Pour évaluer l'exposition professionnelle à l'amiante recueillie dans ICARE, nous avons utilisé la matrice emploi exposition précédemment décrite dans le chapitre 1. Les biais liés à l'usage de cet outil ont pu être détaillés précisément dans la discussion complémentaire à l'article 1 (section 2.4.2.2). Ils ne seront donc pas abordés dans cette discussion car ils sont similaires pour cette analyse, de même pour les limites statistiques évoquées dans la section 2.4.2.3.

2.5.2.1 Ajustement sur le niveau d'éducation (figure 1S de l'article 2)

Dans la littérature sur les relations entre l'exposition professionnelle à l'amiante et le cancer du poumon, ainsi que sur celle entre le tabac et le cancer du poumon, on peut se poser la question de l'ajustement sur le niveau d'éducation (comme reflet du niveau socio-économique) bien que cela ne fasse pas consensus dans la littérature (Richiardi et al. [2008]). Néanmoins, en raison de données manquantes sur le niveau d'éducation, il n'a pas été possible de considérer cette variable dans nos analyses. En effet, sur les 2026 cas et 2610 témoins inclus dans l'analyse de l'article 2, 196 cas (9.7%) et 89 témoins (3.4%) n'avaient aucune information sur leur niveau d'éducation (Table 2S du matériel supplémentaire de l'article 2). De plus, d'un point de vue computationnel, le recours à des méthodes d'imputation multiple (Rubin [1987]) aurait compromis la convergence des modèles.

Afin d'évaluer le biais de confusion résiduel potentiel dû au non ajustement sur le niveau d'éducation et son impact sur les fonctions de poids estimées, nous avons réalisé une analyse de sensibilité en ajustant sur le niveau d'éducation, représenté par le niveau scolaire le plus élevé. Nous avons exclu les sujets qui avaient un niveau d'éducation associé à la catégorie "Autre" (39 sujets) ou s'il était manquant (285 sujets).
Cette analyse complémentaire a été réalisée sur un échantillon de 1809 cas et 2503 témoins.
Pour n'ajouter qu'un seul paramètre dans le modèle, même si cela a engendré une perte d'information, le niveau d'éducation a été considéré par une variable binaire qui était égale à 1 pour les sujets qui étaient allés au moins jusqu'au Lycée & Université et 0 pour ceux qui n'étaient pas allés au-delà du collège. Les témoins avaient un niveau d'éducation plus élevé que les cas (tableau 2.2).

Le WCIE associé à chacune des expositions a été estimé en utilisant comme axe du temps, le temps avant la date index. Les résultats de cette analyse de sensibilité sont présentés dans la Figure 1S contenue dans le matériel supplémentaire de l'article 2.

	Cas (n=1809)		Témoins (n=2503)	
Age à la date index, moy (sd)	60.1	(9.0)	58.2	(9.9)
Départements de résidence, n (%)				
Calvados	207	(11.4)	331	(13.2
Hérault	191	(10.6)	332	(13.3)
Isère	286	(15.8)	361	(14.4)
Loire-Atlantique	250	(13.8)	295	(11.8)
Manche	204	(11.3)	213	(8.5)
Bas-Rhin	223	(12.3)	307	(12.3)
Haut-Rhin	52	(2.9)	86	(3.4)
Somme	209	(11.6)	338	(13.5)
Vendée	102	(5.6)	143	(5.7)
Doubs-Territoire de Belfort	85	(4.7)	97	(3.9)
Statut tabagique, n (%)				
Fumeurs (ex/courant)	1759	(97.2)	1772	(70.8)
Non-fumeurs	50	(2.8)	731	(29.2)
Exposition professionnelle, n (%)				
à l'amiante				
Exposés	1266	(60.0)	1444	(58.1)
Non-exposés	543	(40.0)	1059	(41.9)
Niveau d'éducation, n (%)				
École élémentaire ou moins	600	(33.1)	489	(19.5)
Collège	779	(43.1)	1028	(41.1)
Lycée	177	(9.8)	293	(11.7)
Université ou plus	253	(14.0)	693	(27.7)
Niveau d'éducation binaire, n (%)				
0 *(École élémentaire & Collège)*	1319	(76.2)	1517	(60.6)
1 *(Lycée & Université)*	430	(23.8)	986	(39.4)

Tableau 2.2 – Description des variables démographiques des sujets utilisés pour l'analyse complémentaire

Pour la relation entre le tabac et le cancer du poumon (figure 1S à droite), les estimations ponctuelles de la fonction de poids restent inchangées avec ou sans ajustement sur le niveau d'éducation tel qu'il a été défini précédemment. En effet, nous pouvons observer que les deux courbes se superposent l'une sur l'autre. Pour la relation entre l'exposition professionnelle à l'amiante et le cancer du poumon (figure 1S à gauche), bien que l'ajustement sur le niveau d'éducation ne semble pas changer la forme de la fonction de poids, l'effet de l'intensité en chaque temps semble être sous estimé par rapport à la courbe sans ajustement. En effet, certains auteurs ont déjà pu montrer que l'ajustement sur le niveau d'éducation pouvait sous estimer les relations entre les expositions professionnelles et l'effet de l'intensité (Guida [2012], Brisson et al. [1987]). Pour expliquer ceci, Richiardi et al. [2008] ont pu mettre en avant, notamment en utilisant des graphes orientés acycliques (DAG), que l'ajustement sur le niveau d'étude pouvait gommer l'effet des expositions professionnelles puisque l'on compare des sujets de même catégorie socio-professionnelle et donc avec des expositions professionnelles potentiellement similaires. Avec les données cas-témoins utilisées dans le cadre de cette thèse, la prise en compte du niveau d'éducation ne semble pas modifier la forme des relations étudiées.

2.5.2.2 La modélisation des facteurs de confusion

Dans la relation entre l'exposition professionnelle à l'amiante et le cancer du poumon, nous avons considéré l'histoire tabagique, qui a été représentée par le CSI puisqu'il permet un meilleur ajustement sur ce facteur de risque que le paquet-année ou une variable binaire fumeurs/non-fumeurs (Leffondré et al. [2006], Papadopoulos et al. [2014]). Cependant, dans la relation entre le tabac et le cancer du poumon, l'exposition professionnelle à l'amiante a été prise en compte en introduisant, dans le modèle de régression, un indice cumulé d'exposition classique catégorisé selon les quartiles de la distribution. Afin d'avoir un meilleur ajustement sur ce facteur de risque, une alternative pourrait être le développement d'un indice agrégé pour l'amiante, similaire au CSI, incorporant divers aspects de l'exposition au sein du même indice. En effet, le CSI a été développé pour la relation tabac-cancer du poumon comme on a pu le définir précédemment et ne peut pas être appliqué en l'état pour une autre relation que celle pour lequel il a été développé. Une piste intéressante de recherche est donc l'extension d'un tel indice pour d'autres relations, comme celle entre l'amiante et le cancer du poumon, par exemple. Ce qui a d'ailleurs pu faire l'objet d'un stage de M2 biostatistique, que j'ai pu co-encadrer, dont l'objectif était d'étudier la faisabilité de développer une version du CSI pour la relation entre l'amiante et le mésothéliome pleural/cancer du poumon.

2.5.2.3 Comparaison avec l'approche par fenêtres de temps d'exposition

En analyse complémentaire de l'article 2, nous avons comparé les résultats obtenus avec une méthode moins flexible décrite en introduction de ce chapitre qui est l'approche par fenêtres de temps d'exposition (Finkelstein [1991]). Nous avons utilisé les mêmes données que celle utilisées pour l'article 2. Nous avons considéré 6 périodes de temps de 10 ans chacune. Chacune de ces périodes était associée à la somme des intensités de consommation de tabac, c'est-à-dire la somme du nombre moyen de cigarettes fumées sur cette période de 10 ans. On écrit le modèle de régression de la manière suivante :

$$\begin{aligned}\text{logit}(P(Y_i|x_i, Age_i, Depart_i, ICE_{Abs_i})) =& \beta_0 + \beta_1 \sum_{t=0}^{9} x_i(t) + \beta_2 \sum_{t=10}^{19} x_i(t) + \beta_3 \sum_{t=20}^{29} x_i(t) + \\ & \beta_4 \sum_{t=30}^{39} x_i(t) + \beta_5 \sum_{t=40}^{49} x_i(t) + \\ & \beta_6 \sum_{t=50}^{60} x_i(t) + f(Age_i) + f(Depart_i) + f(ICE_{Abs_i})\end{aligned}$$

Tout comme pour l'approche avec le WCIE, nous avons ajusté sur les facteurs d'appariement (*Age* : âge à la date index en splines et *Depart* : départements de résidence en indicatrices) et l'exposition professionnelle à l'amiante par un indice cumulé d'exposition classique (ICE_{Asb}, catégorisé suivant les quartiles de la population des témoins).
L'interprétation d'un des coefficients d'intérêt ($\beta_1, \ldots, \beta_6$) est la suivante : β_1 est le ln OR associé à l'augmentation d'une unité de l'intensité totale reçue en cigarettes-années sur la fenêtre de temps "0-9" ans avant la date index. Le poids des intensités reçues est constant sur chaque fenêtre de temps d'exposition considérée. D'après la figure 2.4, la forme obtenue avec l'approche par fenêtres de temps d'exposition est cohérente avec nos résultats utilisant l'approche du WCIE (figure 2A de l'article 2). En effet, ce sont les intensités les plus récentes qui ont le plus de poids sur le risque de cancer du poumon.

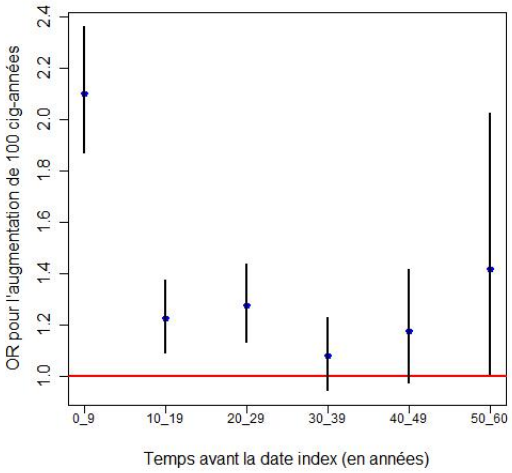

FIGURE 2.4 – Approche par fenêtres de temps d'exposition : OR pour l'augmentation de 100 cigarettes-années pour estimer l'effet de l'intensité du tabagisme sur le risque de cancer du poumon suivant le temps avant la date index. *Ajusté sur l'âge à la date index (splines), départements de résidence, quartiles de l'indice cumulé de l'exposition professionnelle à l'amiante*

Une des principales difficultés de cette approche est la définition du nombre de fenêtres. De plus, la corrélation qui peut exister entre chaque fenêtre de temps entraîne des instabilités dans les estimations. Pour illustrer cela, nous avons également estimé le même modèle mais en découpant l'axe du temps en 12 périodes de 5 ans.

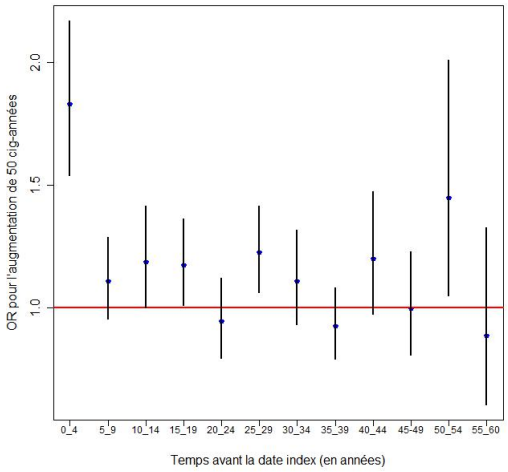

FIGURE 2.5 – Approche par fenêtres de temps d'exposition : OR pour l'augmentation de 50 cigarettes-années pour estimer l'effet de l'intensité du tabagisme sur le risque de cancer du poumon suivant le temps avant la date index. *Ajusté sur l'âge à la date index (splines), départements de résidence, quartiles de l'indice cumulé de l'exposition professionnelle à l'amiante*

On peut observer que la forme des courbes obtenues en figures 2.4 et 2.5 sont similaires. Cependant, on peut constater de fortes instabilités des estimations en figure 2.5 passant du non significatif au significatif sans tendance claire. Les intervalles de confiance sont plus larges que ceux obtenus avec les 6 fenêtres d'exposition. Ce sont les fortes corrélations entre les fenêtres de temps qui engendrent ces estimations biaisées instables. Ceci permet, notamment, de mettre en avant la force de l'approche du WCIE par rapport à celle par fenêtres de temps d'exposition.

2.5.2.4 Test de rapport de vraisemblance

Nous avons réalisé ce test sur les deux expositions pour comparer i) l'apport du WCIE tabac suivant le temps avant la date index vs ICE tabac puis ii) l'apport du WCIE amiante suivant le temps avant la date index vs ICE amiante.

Pour le tabac :

Statistique de test$=(-2\hat{L}_{ICE}) - (-2\hat{L}_{WCIE}) = 98.1 > \chi^2(4) = 9.5$

La valeur calculée de la statistique de test appartient à la région critique (toutes les valeurs supérieures ou égales à 9.5), on rejette donc l'hypothèse nulle. On met en évidence que le WCIE apporte significativement plus d'information quant à un ICE classique. Il y a des poids significativement différents de l'intensité aux différents temps.

Pour l'amiante :

Statistique de test$=(-2\hat{L}_{ICE}) - (-2\hat{L}_{WCIE}) = 7 < \chi^2(4) = 9.5$

La valeur calculée de la statistique de test n'appartient pas à la région critique, on ne rejette pas l'hypothèse nulle. Nous n'avons donc pas pu mettre en évidence que le WCIE apporte significativement plus d'information quant à un ICE classique pour la relation entre l'exposition professionnelle à l'amiante et le cancer du poumon.

En conclusion de ces résultats au test, dans le cadre du cancer du poumon, pour les analyses sur le tabac, il est important de prendre en compte ces poids dans l'intensité cumulée, alors que pour l'amiante, cela ne semble pas nécessaire. Ainsi, cela permettrait de dire que de prendre en compte l'aspect temporel de l'exposition a vraiment de l'importance pour le tabac, alors que pour l'amiante, l'intensité totale cumulée non pondérée suffirait comme information pour évaluer son association avec le cancer du poumon. Néanmoins, le même test a été réalisé pour évaluer l'apport du WCIE dans la relation entre l'exposition professionnelle à l'amiante et le mésothéliome pleural par rapport à l'ICE classique. Le résultat du test a pu mettre en

évidence que le WCIE apportait significativement plus d'information que l'ICE pour l'analyse de la relation dose-réponse (statistique de test= 21 > 9.5). Ainsi, l'aspect temporel de l'exposition professionnelle à l'amiante semblerait important à prendre en compte pour évaluer son association avec le risque de MP. Finalement, on peut se demander si la force d'association entre l'exposition d'intérêt et le cancer pourrait influencer l'apport significatif du WCIE par rapport à un ICE classique. Pour confirmer ceci,, il faudrait envisager l'application du WCIE sur d'autres données ainsi que sur d'autres relations dose-réponse.

2.6 Conclusion générale

Dans le cadre de ce premier travail de thèse, nous avons évalué l'effet de l'intensité d'exposition en chaque temps de l'histoire de l'exposition sur le risque de cancer en utilisant un indice cumulé pondéré par une fonction de poids flexible (WCIE).

Quelle que soit la relation étudiée (Amiante-MP, Amiante-Poumon, Tabac-Poumon), les intensités d'exposition reçues dans les années proches de la date index avaient des poids non négligeables sur le risque de cancer associé. Compte tenu de ces résultats, on peut se poser la question de la pertinence de ne pas considérer les expositions récentes dans le calcul de l'ICE. En effet, un grand nombre d'études épidémiologiques portant sur la relation entre expositions prolongées et cancer ne considèrent pas les années d'exposition récentes dans le calcul de l'ICE en justifiant que ces expositions ne contribuent pas au sur-risque du cancer observé. Ici, nos résultats montrent que cette hypothèse peut s'avérer, dans certains cas, fausse.
Pour l'exposition professionnelle à l'amiante, les intensités d'exposition reçues plus de 40 ans avant la date index sont celles qui contribuent le plus au risque de cancer (MP ou Poumon). De même, ce sont les intensités reçues aux âges les plus jeunes (moins de 25 ans) qui ont le plus de poids. Nos résultats sont en accord avec la littérature puisque ce sont les intensités reçues à un temps distant de la date index qui ont un plus fort impact sur le risque de cancer.
Pour le tabac, les expositions anciennes ne jouent pas un rôle aussi fort que les expositions reçues récemment (dans les 20 ans précédant la date index). De plus, les poids des intensités associées aux jeunes âges, bien que différents de zéro, n'étaient pas plus élevés que ceux des intensités associées à des âges plus avancés.

Ce premier travail de thèse comporte des limites, que l'on a pu soulever à travers les articles et les discussions complémentaires de ce chapitre. Notamment, certaines peuvent être liées aux données et d'autres plutôt à l'approche utilisée. En conclusion de ce chapitre sur l'utilisation d'un indice cumulé pondéré par une fonction de poids flexible, nous allons revenir sur deux limites qui ont des perspectives intéressantes à envisager suite à ce premier travail de thèse. Cependant, d'autres limites peuvent être relevées comme le fait de ne pas avoir pris en compte d'autres facteurs de confusion, ni considéré l'interaction entre le tabac et l'amiante pour le can-

cer du poumon ou encore n'avoir réalisé les analyses seulement avec les sujets qui avaient des histoires d'expositions complètes. Toutes ces limites seront davantage détaillées en conclusion générale de la thèse car elles ne sont pas seulement spécifiques à la méthode utilisée dans ce chapitre.

Limite liée à l'interprétation épidémiologique

Lorsque l'on considère le temps avant la date index comme axe du temps, on estime l'effet de l'intensité sur le risque de cancer en chaque temps avant le diagnostic ou l'interview, ajusté sur toutes les intensités aux autres temps. Si l'on considère la fonction de poids dans son ensemble, cela peut donner une indication sur la notion de latence définie comme le temps entre la première exposition et le diagnostic du cancer. Lorsque l'on considère l'âge comme axe du temps du WCIE, c'est l'effet de l'intensité en chaque âge qui est estimé, ce qui peut être vu comme la susceptibilité liée à l'âge (âge comme modificateur d'effet de l'intensité). Quel que soit l'axe du temps considéré, il est difficile de s'assurer que l'effet observé sur un axe est indépendant de l'autre puisque les deux axes sont fortement corrélés. Pour s'assurer de n'observer que l'effet de l'un ou l'autre sur le risque de cancer, il faudrait pouvoir se positionner dans certains scénarios d'histoires d'exposition définissant un sous groupe d'individus avec des caractéristiques bien définies. Par exemple, pour estimer la fonction de poids en chaque temps avant la date index, il faudrait pouvoir l'estimer en fixant l'âge et le temps depuis la dernière exposition. A contrario pour estimer de façon totalement indépendante l'effet de l'intensité en fonction de l'âge, il faudrait estimer la fonction de poids en fixant l'âge, le temps depuis la première exposition, ...
Illustrons ceci par l'exemple suivant. Considérons un premier sous groupe où les sujets sont des fumeurs courants qui ont un âge à la date index entre 60 et 70 ans et un second sous groupe défini par des ex fumeurs depuis 10 ans à même âge à la date index. Si on s'intéresse à ce qu'il se passe entre 55 et 60 ans dans le premier sous groupe et entre 45 et 50 ans dans le second, on se positionne donc exactement au même moment dans leurs histoires d'exposition respectives. Si l'on obtient des OR similaires associés à une différence de 10 cigarettes-années sur ces deux périodes respectives dans chacun des groupes (groupe 1 : 55-60, groupe 2 : 45-50), cela signifie donc que l'on évalue la récence des intensités d'exposition sur le risque de cancer. Néanmoins, si on constate une différence entre ces deux OR calculés, cela représente l'effet de l'âge à l'exposition puisque la seule différence entre leurs histoires d'exposition serait liée au fait qu'ils n'ont pas eu leurs expositions récentes aux mêmes âges. A partir des données provenant d'ICARE, 1455 sujets ont un âge à la date index entre 60 et 70 ans. Parmi eux, 162 sujets sont fumeurs courants et seulement 47 ont arrêté depuis exactement 10 ans. Cependant, pour assurer la convergence du modèle, il faut avoir assez de sujets pour le scénario envisagé. Le manque de puissance statistique est un point qui justifie la difficulté de réaliser une telle analyse de sensibilité avec les données utilisées dans le cadre de cette thèse. Néanmoins, il serait intéressant d'envisager la mise en place de ces analyses de sensibilité comme perspectives de ce travail sur cet indice cumulé pondéré.

Limite statistique de la modélisation de l'intensité en chaque temps

Comme nous avons pu le préciser dans une précédente discussion en section 2.4.2.3.b, il y a une hypothèse de linéarité de l'intensité reçue en chaque temps avec le WCIE. Pour relâcher cette hypothèse, une alternative pourrait être de modéliser l'intensité en chaque temps par l'utilisation de fonctions flexibles splines. Au delà des difficultés d'interprétation des coefficients, il y aurait également des problèmes de convergence du modèle liés au grand nombre de paramètres à estimer. Une alternative intermédiaire pourrait être de considérer cet effet non linéaire par des fonctions paramétriques potentielles, comme par exemple, avec une fonction logarithmique (Vacek [1997]). L'inconvénient de cette approche réside dans le choix de la fonction non linéaire paramétrique la plus appropriée. Pour pouvoir relâcher l'hypothèse de linéarité des intensités dans le WCIE, des investigations supplémentaires sont donc à envisager en perspectives de l'application d'un tel indice sur d'autres relations dose-réponse.

2.7 Contribution et Valorisation

2.7.1 Contribution

Dans le cadre de ce premier travail de thèse, ma contribution personnelle tient en :

— l'implémentation sous SAS de la vraisemblance du modèle nécessaire à expliciter pour la proc NLP, du bootstrap pour obtenir les ICs de la fonction de poids ;

— l'analyse des données, à savoir : la sélection des sujets, la modélisation des facteurs de confusion, l'estimation du meilleur modèle, l'interprétation des résultats ;

— la rédaction de la première version de l'article 1 puis suivi des modifications suivant les retours des co-auteurs et contribution aux réponses aux relecteurs dans le processus de révision de l'article ;

— la réalisation de l'analyse complémentaire sur laquelle nous avons appuyé nos propos dans la lettre publiée par Boffeta et al. ;

— la rédaction de l'article 2 de la première version à la version soumise, la rédaction des réponses aux relecteurs dans le processus de révision de l'article.

Co-encadrement de Mme Céline Gramond, étudiante en Master 2 Biostatistique (encadrement principal : Karen Leffondré, co-encadrement : Aude Lacourt)

2.7.2 Valorisation scientifique

Articles publiés

— Lacourt A*, **Lévêque E*** (co-first), Guichard E,Gilg Soit Ilg A, Sylvestre M-P, Leffondré K. Dose-time-response association between occupational asbestos exposure and pleural mesothelioma Occup Environ Med 2017 ;74 :691-697.

— **Lévêque E**, Lacourt A, Luce D, et al.Time-dependent effect of intensity of smoking and of occupational exposure to asbestos on the risk of lung cancer : results from the ICARE case–control study. Occup Environ Med Published Online First : 18 May 2018.

Lettre

Lacourt A, **Lévêque E**, Goldberg M, et al. Dose–time response association between occupational asbestos exposure and pleural mesothelioma : authors' response. Occup Environ Med 2018 ;75 :161-162.

Présentation orale

Dynamic longitudinal effects of increasing intensity of smoking and occupational exposure to asbestos on lung cancer : results from the icare case-control study. The 25th International Epidemiology in Occupational Health (EPICOH) Conference, Barcelona, September 5-7, 2016.

Dont le résumé a été publié :
Lévêque E, Lacourt A, Luce D, et al. O15-2 Dynamic longitudinal effects of increasing intensity of smoking and occupational exposure to asbestos on lung cancer : results from the icare case-control study.Occup Environ Med 2016 ;73 :A28. (EPICOH Barcelone 2016).

Présentation affichée

Analyse de l'aspect temporel de la relation entre expositions prolongées et risque de cancer Application à 3 relations : amiante-mésothéliome pleural, amiante-cancer du poumon et tabac-cancer du poumon. Rencontres scientifiques de l'ANSES pour les 10 ans du Programme national de recherche environnement santé travail (PNREST) Paris 14 Novembre 2016.

2.7.3 Travail en collaboration

Un travail collaboratif a été mis en place en parallèle de la thèse avec Aurélie Danjou, pendant son doctorat en épidémiologie au centre Léon Bérard de Lyon supervisé par Béatrice Ferves. Durant sa thèse, elle s'intéressait à l'impact des expositions aux dioxines dans l'air sur le cancer du sein chez les femmes dans la région Rhône Alpes provenant de l'étude de cohorte E3N. Dans le cadre de cette étude, les expositions aux dioxines ont pu être évaluées grâce à l'historique résidentiel des femmes dès leur inclusion dans l'étude et durant le suivi.
L'objectif de ce travail collaboratif était d'évaluer une telle relation par l'application d'un WCIE sur ses données. Une des limites est qu'elle dispose de données d'exposition sur une période de dix ans, difficile de caractériser ceci comme une réelle exposition vie entière. De plus, l'évaluation de cette relation est différente suivant le type/stade du cancer mais ceci n'a pas pu être considéré dans des analyses de sensibilité à cause des effectifs trop faibles dans certaines de ses catégories de type de cancer.

Un article lié à ce travail est actuellement en révision dans *Environmental International* :
A.Danjou, T.Coudon, D.Praud, **E.Lévêque**, E.Faure, P.Salizzoni, M.Le Romancer, G.Severi, F. Mancini, K.Leffondré, L.Dossus and B.Fervers. Long-term Airborne Dioxin Exposure and Breast Cancer Risk in a case-control study nested within the French E3N Prospective Cohort.

3 Chapitre 3 : Identification des profils de trajectoires d'intensité d'exposition et comparaison des risques de cancer associés (JLCMM)

Sommaire

Afin de caractériser l'effet de l'intensité d'exposition au cours du temps sur le risque de cancer, une première méthode (présentée dans le chapitre précédent) a été de considérer un indice cumulé pondéré (WCIE). Celle-ci nous a permis de comparer des rapports de côtes de cancer entre différents profils hypothétiques d'intensité dans une analyse post hoc, à partir des poids estimés des intensités reçues en chaque temps de l'histoire d'exposition (article 1). Cependant, cette méthode ne permettait pas d'identifier des profils d'histoires d'exposition existants dans les données observées.

Dans la littérature, divers auteurs ont proposé des méthodes statistiques pour l'identification de profils de trajectoires à partir de données longitudinales, mais les applications ont davantage été menées pour des données de biomarqueurs (Lin et al. [2000], Boucquemont et al. [2017]), de tests psychométriques (Proust et al. [2006]) ou des données comportementales (Nagin and Tremblay [2005], Muthén [2004], Park et al. [2018], Milanzi et al. [2017]). A ce jour, peu d'études ont utilisé ces modèles avec des données d'expositions environnementales comme le tabac (Weden and Miles [2012], De Genna et al. [2017], Brook et al. [2008], Dutra et al. [2017]) et, à notre connaissance, aucun avec des données d'expositions professionnelles provenant d'études cas-témoins.

Les objectifs de ce travail étaient :

i) d'identifier les profils d'exposition au tabac et à l'amiante chez les cas et les témoins hommes exposés dans l'étude ICARE ;
ii) et de comparer les risques de cancer du poumon associés.

La méthode principale et les principaux résultats sont présentés dans l'article 3 qui a été soumis au journal épidémiologique, *International Journal of Epidemiology*. Dans ce chapitre, la méthode utilisée sera davantage détaillée. De plus, des méthodes et des résultats complémentaires seront présentés à titre de comparaison. Plus précisément, dans une première partie, la méthode statistique envisagée permettant l'identification des profils de trajectoires sera détaillée, puis celles permettant de comparer les risques de cancer associés. Dans un second temps, la méthodologie pour les deux applications sur le cancer du poumon sera décrite. L'article 3 sera ensuite inclus pour présenter les résultats obtenus. Une analyse complémentaire sur la comparaison des différentes approches détaillées en première partie sera présentée. Enfin, nous terminerons avec une discussion complémentaire à l'article, notamment sur l'aspect statistique et la considération de telles données dans ce contexte méthodologique.

3.1 Les méthodes statistiques pour l'identification de trajectoires longitudinales d'intensité d'exposition

3.1.1 Les méthodes de classification

Dans cette section, on s'intéresse à des méthodes de classification permettant de distinguer des groupes d'individus avec une même trajectoire longitudinale d'intensité d'exposition au cours de la vie. Ne sont pas considérées dans cette section, des méthodes permettant d'obtenir un arbre de classification, de faire de la sélection des variables ou encore de faire de la classification supervisée spécifique aux données fonctionnelles ou de grandes dimensions par exemple. En effet, ces méthodes n'ont pas pour objectif de dégager des profils longitudinaux, et surtout la structure des données utilisée dans ces méthodes n'est pas adaptée à celle des données que l'on utilise pour répondre à l'objectif de ce travail.

Ces dernières années, les méthodes de classification ont fait l'objet de nombreux développements à la fois méthodologiques et computationnels. La littérature s'est très largement enrichie sur ce sujet. Elle peut parfois paraître un peu confuse tant les modèles sont nombreux, distincts suivant les domaines d'application (Eshghi et al. [2011]) ou spécifiques suivant une extension méthodologique d'un modèle plus simple (Muthén and Shedden [1999]). Le but de cette section n'est pas de faire une vue d'ensemble de tous les modèles existants mais de bien comprendre la méthode que nous avons identifiée comme exploitable afin de répondre à notre objectif.

Les approches non paramétriques sont classiquement utilisées lorsque l'on est confronté à un problème de classification. De plus, elles sont assez faciles à mettre en oeuvre car elles ne sont pas basées sur un modèle statistique. Ces approches reposent sur l'hypothèse que, plus deux individus sont proches plus ils ont de chance d'appartenir à la même classe de sujets (appelée aussi cluster). Cette classe (non observée) va donc être définie par un groupe de sujets indépendants entre eux mais qui partagent une similarité sur un critère particulier. Ces approches peuvent différer entre elles, à la fois par leur définition d'une classe mais également par la définition de la distance (ou la similarité) entre deux individus. Parmi ces dernières, les deux approches les plus utilisées sont l'approche par K-moyennes (ou nuées dynamiques) (MacQueen et al. [1967]) et la classification hiérarchique (Szekely and Rizzo [2005]). Elles sont moins adaptées lorsque les classes doivent être définies à partir de données présentant des structures plus complexes, comme les mesures répétées d'un même sujet. Or, pour répondre à notre objectif, les classes doivent représenter les classes de sujets (cas et témoins confondus) qui ont des trajectoires d'exposition similaires. La classification doit donc se faire sur l'ensemble des mesures répétées de l'intensité d'exposition, en tenant compte des mesures intra sujets et de leur corrélation. Ces approches ne peuvent donc pas être envisagées pour répondre à notre objectif. Cependant, une extension de l'approche des K-moyennes pour données longitudinales a

récemment été proposée par Genolini and Falissard [2011]. Néanmoins, cette extension, associée au package R kml (Genolini et al. [2015]), requiert que le même nombre de mesures répétées par individu soit disponible, rendant impossible l'utilisation de cette approche pour répondre à notre objectif. En effet, dans notre contexte des expositions prolongées, le nombre de mesures répétées (les intensités d'exposition) dépend de la durée de l'histoire d'exposition de chaque sujet.

Pour les méthodes de classification basées sur des modèles statistiques, chaque cluster est représenté par une distribution paramétrique. L'ensemble des données est décrit par un mélange de ces distributions qui optimise l'ajustement entre les données et le modèle spécifié. Les modèles statistiques paramétriques permettent de considérer les données comme des mesures répétées au cours du temps et d'en estimer une trajectoire moyenne ; c'est la théorie des modèles à effets mixtes (Laird and Ware [1982]). En considérant une extension de cette théorie pour des groupes hétérogènes d'individus à trajectoires moyennes homogènes, ces modèles permettent donc de faire de la classification. Ces modèles reposent sur l'idée générale qu'une variable latente va caractériser un processus non observé permettant de définir des groupes d'individus dont on ne connaît pas la composition a priori dans les données (Muthén and Shedden [1999]). C'est donc ce modèle à variable latente reposant sur un modèle de régression, qui permet d'estimer un profil moyen d'évolution au cours du temps pour chacun des groupes d'individus. Néanmoins, il ne peut être utilisé que sous certaines conditions d'application liées aux données longitudinales puisqu'il repose sur certaines hypothèses. De manière générale, la nature et la distribution des données répétées doivent être considérées pour choisir le modèle longitudinal le plus adapté (par exemple, modèle linéaire mixte ou modèle mixte généralisé).

3.1.2 La méthode envisagée : le modèle à classes latentes

3.1.2.1 Définition

Le modèle linéaire mixte à classes latentes est une extension du modèle linéaire mixte introduit par Laird and Ware [1982] et Verbeke and Molenberghs [2000].

Considérons le vecteur du marqueur longitudinal $Y_i = (Y_{i1}, ..., Y_{in_i})$ correspondant aux n_i mesures répétées du marqueur pour le sujet i ($i \in \{1, \ldots, n\}$). Le modèle linéaire mixte a été défini par Laird and Ware [1982] comme suit :

$$Y_{ij} = X_{Li}(t_{ij})^T \beta + Z_i(t_{ij})^T u_i + w_i(t_{ij}) + \epsilon_{ij} \tag{19}$$

où

— $X_{Li}(t_{ij})$ est un vecteur des variables explicatives observées au temps t_{ij} associé au vecteur des effets fixes β ;

— $Z_i(t_{ij})$ est un vecteur des variables explicatives observées au temps t_{ij} associé au vecteur

des effets aléatoires u_i. Le vecteur des effets aléatoires est supposé suivre une distribution multivariée gaussienne de moyenne nulle avec une matrice de variance-covariance B : $u_i \sim \mathcal{N}(0, B)$;

— $w_i(t_{ij})$ est un processus stochastique caractérisant la corrélation sérielle dans les données (par exemple un mouvement brownien ou un processus auto-régressif) ;

— $\epsilon_{ij} \sim \mathcal{N}(0, \sigma^2)$ est un terme d'erreur de mesure. Les ϵ_{ij_j} sont indépendants entre eux et des u_i.

Ce modèle a été développé pour données homogènes. Il permet de définir une trajectoire moyenne pour les individus en faisant l'hypothèse que la population est homogène en terme d'évolution au cours du temps. La modélisation à effets mixtes est la méthode la plus utilisée pour étudier l'évolution au cours du temps d'un marqueur longitudinal.

Pour prendre en compte des données hétérogènes lorsqu'une ou plusieurs caractéristiques de la population ne sont pas observées, des modèles de mélange ont ensuite été considérés (Böhning [1999]). La population est alors constituée de sous groupes au sein desquels les observations ont la même distribution.

A l'origine, Verbeke and Lesaffre [1996] ont étendu les modèles de mélange pour des données longitudinales pour tenir compte de l'écart à l'hypothèse de normalité des effets aléatoires mais il est depuis utilisé et étendu pour traiter des données hétérogènes. Ces dernières années, Muthen et son équipe (Muthén and Shedden [1999], Muthén [2004], Muthén and Muthén [2005]) ont proposé d'étendre ces modèles pour décrire des sous populations non observées dans les données avec l'idée qu'une population hétérogène est constituée de G sous populations avec des profils d'évolution différents, ces profils ne pouvant être entièrement définis à partir des caractéristiques observées. L'hétérogénéité est prise en compte par la distribution des effets aléatoires qui va suivre G lois normales : $u_i \sim \sum_{g=1}^{G} \pi_g \mathcal{N}(0, B_g)$ où chaque sous-population a une probabilité π_g ($\sum_{g=1}^{G} \pi_g = 1$).

Muthén and Shedden [1999] ont étendu une première version du modèle de Verbeke and Lesaffre [1996] afin de prédire la probabilité d'appartenir à une classe latente en fonction des caractéristiques du sujet (d'où l'émergence d'un modèle pour cette probabilité). On parle donc de modèles à classes latentes. Dans la suite du manuscrit, deux modèles à classes latentes seront distingués, le modèle LCMM ("latent class mixed model") et le modèle LCGM ("latent class growth mixture model") dont on détaillera les spécificités par la suite.

Ces modèles (LCMM & LCGM) font l'hypothèse que la population est hétérogène et est constituée de G groupes non observés appelés classes latentes. Ces G classes latentes sont

caractérisées par leur propre profil moyen de trajectoires. Chaque sujet appartient à une et une seule classe latente. LCMM & LCGM sont composés de deux sous-modèles qui sont estimés de manière simultanée en utilisant le maximum de vraisemblance.

3.1.2.2 LCMM : le modèle mixte à classes latentes

3.1.2.2.a Spécification du modèle

Le premier sous-modèle caractérise la probabilité de l'individu d'appartenir à la classe g ($g \in \{1, \ldots, G\}$) tandis que le second sous-modèle caractérise la distribution individuelle du marqueur conditionnellement à chaque classe d'appartenance g à partir d'un modèle linéaire mixte. Ces deux sous-modèles sont ajustés à partir de covariables explicatives individuelles.

Pour l'individu i, l'appartenance à la classe latente est définie par une variable latente discrète c_i qui est égale à g s'il appartient à la classe g. La probabilité individuelle π_{ig} d'appartenir à chaque classe g est obtenue à partir d'un modèle logistique multinomial (sous-modèle 1). Elle peut dépendre de certaines caractéristiques du sujet (X_{pi}^T) :

$$\pi_{ig} = P(c_i = g|X_{pi}) = \frac{\exp(\zeta_{0g} + X_{pi}^T\zeta_{1g})}{\sum_{l=1}^{G} \exp(\zeta_{0l} + X_{pi}^T\zeta_{1l})} \tag{20}$$

où ζ_{0l} sont les intercepts spécifiques pour chaque classe l ($l \in \{1, \ldots, G\}$) ; ζ_{1l} sont les vecteurs des coefficients spécifiques à chaque classe l associés au vecteur de covariables X_{pi}^T.

Lorsqu'aucune variable d'ajustement n'est introduite dans le modèle, on obtient une probabilité spécifique à chaque classe indépendante des caractéristiques de l'individu.

Le second sous-modèle (sous-modèle 2) estime les profils moyens de trajectoires à travers un modèle mixte spécifique à chaque classe.

Pour Y_i gaussien,

$$Y_{ij_{|c_i=g}} = X_{oi}(t_{ij})^T\beta + X_{si}(t_{ij})^T\upsilon_g + Z_i(t_{ij})^T u_{ig} + w_i(t_{ij}) + \epsilon_{ij} \tag{21}$$

où

— Le vecteur X_{Li} défini dans le modèle linéaire mixte classique (équation 19)) est divisé ici en $X_{oi}(t_{ij})$ vecteur des variables communes aux classes associé au vecteur des effets fixes β, $X_{si}(t_{ij})$ vecteur des variables spécifiques aux classes associé au vecteur des effets fixes υ_g ;

— Le vecteur $Z_i(t_{ij})$ est défini par des variables au temps t_{ij}, et associé aux effets aléatoires u_{ig} ;

— La distribution des effets aléatoires u_{ig} est maintenant spécifique à chaque classe en supposant $u_{ig} \sim \mathcal{N}(0, \omega_g^2 B)$ où ω_g est un coefficient de proportionnalité qui permet une variabilité individuelle spécifique à la classe. $\omega_G = 1$ pour assurer l'identifiabilité du modèle.

Dans chaque classe latente, la trajectoire moyenne en fonction du temps est modélisée à travers 1) une évolution temporelle moyenne et 2) un départ individuel de cette trajectoire moyenne (pour LCMM seulement et non pour LCGM). Pour ajuster au mieux les trajectoires individuelles observées, l'évolution temporelle peut être représentée avec des fonctions polynomiales ou splines. Le départ individuel de cette trajectoire moyenne (variabilité intra-classe) est caractérisé par les effets aléatoires qui peuvent être ajoutés aux paramètres définissant l'évolution temporelle. D'autres covariables peuvent être ajoutées au modèle mixte. La matrice de variance-covariance B peut être diagonale ou non pour prendre en compte la corrélation entre les effets aléatoires.

Cas particulier d'un Y_i non gaussien : modèle mixte à processus latent

Lorsque Y_i a une distribution non gaussienne, on peut utiliser une extension du sous-modèle 2, à travers un **modèle mixte à processus latent** spécifique à la classe (Proust et al. [2006], Proust-Lima et al. [2013]). L'idée de ce modèle est l'introduction d'un processus latent qui sépare le modèle structurel et le modèle de mesure. Il a été principalement défini dans le contexte de modèle mixte multivarié (lorsque l'on souhaite estimer les évolutions d'au moins deux marqueurs simultanément) mais il peut très bien être utilisé lorsque l'on s'intéresse à l'évolution d'un seul marqueur.

Le modèle de mesure est défini pour lier les mesures répétées individuelles au processus latent par une fonction de lien paramétrique H :

$$Y_{ij} = H(\tilde{Y}_{ij}; \eta) = H(\Lambda_i(t_{ij}) + \epsilon_{ij}; \eta) \tag{22}$$

Où

— $\tilde{Y}_{ij}$ est le processus latent (non observé) bruité au temps t_{ij} ;
— Λ est le processus latent sans erreur ;
— ϵ_{ij} sont les erreurs de mesures indépendantes gaussiennes de variance σ_ϵ^2 ;
— η sont les paramètres à estimer de la fonction de lien H.

Suivant le type de mesures répétées (continues, ordinales, ...), différentes fonctions de lien peuvent être utilisées (notamment dans le package R *lcmm* de Proust-Lima et al. [2017] : linéaire, beta cumulée, I-splines, probit cumulé). Il faut choisir une fonction de lien de manière à

ajuster au mieux les données répétées et valider les hypothèses sous-jacentes sur la distribution de ces données.

Donnons un exemple de transformation, où H^{-1} est la fonction de lien inverse et doit être une fonction continue définie monotone croissante :

— Base de I-splines quadratiques (Ramsay et al. [1988]) avec m noeuds :

$$H^{-1}(Y_{ij};\eta) = \eta_0 + \sum_{l=1}^{m+1} \eta_l^2 B_l^I(Y_{ij}) \tag{23}$$

où $B_l^I(Y_{ij})$ est la $l^{ème}$ fonction de base associée au paramètre η_l à estimer. Cette transformation peut être utilisée pour des mesures répétées quantitatives.

Le modèle structurel décrit le processus latent pour chaque classe latente par un modèle mixte, suivant des fonctions du temps (X_{oi}^T) et des variables spécifiques (X_{si}^T), similaire à l'équation (21) :

$$\Lambda_i(t_{ij})|_{c_i=g} = X_{oi}(t_{ij})^T\beta + X_{si}^T\upsilon_g + Z_i(t_{ij})u_{ig} + w_i(t_{ij}) \tag{24}$$

Il est nécessaire de fixer deux contraintes pour l'estimation de ce modèle : β_{01}=0 (intercept de la première classe est contraint à 0) et σ_ϵ^2=1 (la variance des erreurs de mesures est contrainte à 1). En utilisant une intégration numérique, les estimations des profils moyens de trajectoires peuvent être obtenues dans l'échelle initiale des observations.

3.1.2.2.b Estimation

L'un des problèmes dans l'estimation des modèles de mélange tels que les modèles à classes latentes est que l'on ne peut pas estimer simultanément, par méthodes classiques, le nombre de classes latentes G et les paramètres liés aux deux sous-modèles dont le nombre dépend du nombre de classes latentes. Ainsi, dans la littérature, l'estimation se fait donc pour un nombre de classes latentes G fixé, choisi en estimant le modèle pour différents nombres de classes. Le modèle retenu est celui qui ajuste le mieux les données. Différents critères d'ajustement sont étudiés et présentés dans la section 3.1.2.5.

Pour un nombre G fixé de classes latentes, la contribution individuelle à la vraisemblance s'écrit :

$$\mathrm{lv}_i(\theta_G) = \sum_{g=1}^{G} \pi_{ig}\phi_{ig}(Y_i|c_i = g; \theta_G) \tag{25}$$

où

- ϕ_{ig} est une fonction de densité ; par exemple une fonction de densité gaussienne si Y_i est gaussien ;
- θ_G l'ensemble des paramètres des deux sous modèles.

La log vraisemblance totale des n individus ($L(\theta_G) = \sum_{i}^{n} \log \mathrm{lv}_i(\theta_G)$) est directement maximisable par un algorithme d'optimisation itératif. Il est à noter que l'expression de la log vraisemblance peut être différente avec un **modèle mixte à processus latent** car il fait intervenir une transformation des données dont il faut tenir compte dans l'écriture de la vraisemblance du modèle.

Dans la littérature, l'algorithme EM (Dempster et al. [1977]) est souvent utilisé pour estimer ces modèles de mélange (Komárek et al. [2002], Muthén and Muthén [2005]). Cependant, cet algorithme comporte de nombreux inconvénients (Lindstrom and Bates [1988]). En effet, il converge assez lentement, il ne donne pas d'estimations directes de la variance des paramètres et aucun critère de convergence n'est vraiment satisfaisant pour assurer un maximum global. De plus, dans le cadre des modèles de mélanges dont dérive les modèles à classes latentes, l'étape M requiert l'estimation d'un modèle mixte conditionnellement à la classe qui fait intervenir un algorithme itératif de type Newton-Raphston et cela allonge considérablement le temps de calcul.

Ainsi, Proust and Jacqmin-Gadda [2005] ont développé un algorithme modifié de Marquardt (Marquardt [1963]) pour l'estimation de ces modèles de mélange dans leur package R *lcmm* (Proust-Lima et al. [2017]). Celui-ci repose sur la maximisation de la log-vraisemblance de manière itérative. C'est un algorithme de type Newton-Raphston où la diagonale de la matrice Hessienne à l'itération d est "gonflée" pour assurer l'obtention d'une matrice définie positive. Il a été montré plus robuste qu'un simple algorithme de Newton-Raphston (Proust and Jacqmin-Gadda [2005]). Ses performances ont été comparées à celles de la macro SAS HETMIXED (Komárek et al. [2002]) sur deux bases de données distinctes. Sur l'une d'entre elles, basée sur un échantillon de 20 filles préadolescentes dont on veut estimer les évolutions moyennes de leur taille entre 6 et 10 ans (utilisé par Komárek et al. [2002]), le temps computationnel était bien moindre avec le Marquardt modifié (environ 30 secondes vs plusieurs minutes) et la convergence vers le maximum global était plus souvent atteinte (22 sur 32 ensembles différents de valeurs initiales vs 11) (Proust and Jacqmin-Gadda [2005]) .

L'algorithme de Marquardt utilisé dans le package R *lcmm* (Proust-Lima et al. [2017]) :

▷ Définition des éléments de la diagonale de la matrice hessienne H à l'itération d tels que :

$$H_{ii}^{\star} = H_{ii} + \lambda[(1-\eta)|H_{ii}| + \eta tr(H)] \tag{26}$$

où λ et η sont des paramètres permettant d'assurer que la matrice soit définie positive. Les valeurs initiales sont λ=0.01 et η=0.01. Les valeurs des paramètres diminuent si la matrice est définie positive, et augmentent sinon.

▷ Les estimations des paramètres θ sont définies d'une itération à l'autre par la formule suivante :
A l'itération d,

$$\theta^{d+1} = \theta^{d} - \kappa(H^{\star(d)})^{-1} g(\theta^{(d)}) \tag{27}$$

où $g(\theta^{(d)})$ est le gradient courant des paramètres ; κ est initialement égal à 1 puis modifié pour assurer que la log vraisemblance soit meilleure à chaque itération.

▷ La convergence est finalement atteinte lorsque les trois critères de convergence sont satisfaits :

— stabilité des paramètres : $\sum_{r=1}^{n_\theta} (\theta_G(r)^{(d)} - \theta_G(r)^{(d-1)})^2 \leq \epsilon_a$; où $\theta_G(r)^{(d)}$ est le vecteur des paramètres à l'itération d et n_θ la taille du vecteur de paramètres θ_G ;

— stabilité de la log vraisemblance : $|L^{(d)} - L^{(d-1)}| \leq \epsilon_b$ où $L^{(d)}$ est la log vraisemblance à l'itération d ;

— taille des dérivées : $g(\theta^{(d)})^T (H^{(d)})^{-1} g(\theta^{(d)}) \leq \epsilon_c$ où $g(\theta^{(d)})$ est le gradient des paramètres à l'itération d.

Les valeurs par défaut sont $\epsilon_a = \epsilon_b = \epsilon_c = 10^{-4}$ (Proust-Lima et al. [2017]).

Les dérivées premières et secondes qui entrent dans le calcul de cette log vraisemblance sont calculées par la méthode des différences finies. Les écarts-types des paramètres de la matrice B sont calculés par la delta-méthode (Oehlert [1992]). Les écarts-types des autres paramètres sont directement calculés à partir de l'inverse de la matrice Hessienne observée.

Due à la multimodalité de la vraisemblance des modèles de mélange, chaque modèle est estimé à partir de différentes combinaisons de valeurs initiales des paramètres pour assurer la convergence vers le maximum global. Aucune stratégie n'existe pour définir l'ensemble des valeurs initiales, bien que de nombreux articles explicitent ce problème de convergence. Néanmoins, Proust and Jacqmin-Gadda [2005] ont remarqué que les résultats étaient plus sensibles aux valeurs initiales de π_{ig} et μ_g mais moins aux autres paramètres (B, β et σ).

3.1.2.3 LCGM : un cas particulier du LCMM

Comme le modèle LCMM, le modèle LCGM est composé de deux sous-modèles, l'un pour l'appartenance aux classes (défini par l'équation 20) et le second pour l'évaluation de la trajectoire moyenne au sein de chaque classe. Cependant, il peut être considéré comme un cas particulier du modèle LCMM puisqu'il ne considère que des effets fixes (Nagin [2005]). Le fait qu'il n'y ait pas d'effet aléatoire signifie que l'on ne considère pas de déviation individuelle par rapport à la trajectoire moyenne estimée d'une classe. Par conséquent, on ne tient pas compte de la corrélation entre les mesures répétées d'un même individu au sein de la classe. Ceci peut amener à surestimer le nombre de classes latentes puisque l'on fait l'hypothèse que les sujets qui appartiennent à la même classe ont une trajectoire moyenne similaire à celle de la classe sans écart possible à celle-ci (Muthén and Asparouhov [2009]).

Malgré cette limite importante du LCGM, il reste aujourd'hui largement plus utilisé que LCMM pour identifier des sous groupes homogènes de trajectoires au sein d'une population. Cette popularité du LCGM est liée à son implémentation dans SAS par la proc TRAJ développée par Nagin et ses collaborateurs (Jones et al. [2001], Nagin [2005]). En effet, cette procédure est très facile à utiliser, même pour les utilisateurs ne connaissant pas le modèle sous-jacent. Cependant, il faut être vigilant à sa correcte utilisation pour éviter d'obtenir des résultats peu consistants.

Avec la procédure TRAJ, il est possible d'identifier des trajectoires à partir de données répétées de différents types. Pour des mesures répétées quantitatives continues, un modèle gaussien censuré ou non peut être utilisé. Pour des données de comptage, les trajectoires pourront être identifiées avec un modèle de Poisson. Lorsque ces données de comptage ont un nombre fréquent de zéros, un modèle de Poisson avec sur-représentation de zéros (ZIP) peut être estimé. Et enfin, le modèle logistique est également implémenté pour des données binaires. Tous ces modèles linéaires généralisés sont faciles à mettre en œuvre et à estimer, même dans un contexte de classes latentes. En effet, envisager des modèles mixtes généralisés dans les modèles à classes latentes complexifie l'estimation puisque ce sont les effets aléatoires, qui les définissent, qui peuvent amener à une solution non analytique de la vraisemblance du modèle.

3.1.2.4 Classification a posteriori

Il est possible de réaliser une classification a posteriori des sujets suivant les classes latentes estimées par le modèle. Après l'estimation des paramètres du modèle à classes latentes pour un nombre G fixé de classes latentes, on peut calculer la probabilité a posteriori $\hat{\pi}_{ig}$ qu'un individu i appartienne à chacune des g classes latentes ($g = 1, \ldots, G$).

En utilisant le théorème de Bayes, on a

$$\hat{\pi}_{ig} = P(c_i = g|Y_i, \hat{\theta}_G) = \frac{\hat{\pi}_{ig}\phi_{ig}(Y_i, \hat{\theta}_G)}{\sum_{l=1}^{G} \hat{\pi}_{il}\phi_{il}(Y_i, \hat{\theta}_G)} \tag{28}$$

où $\hat{\theta}_G$ est le vecteur des paramètres estimés.

Chaque sujet i est classé a posteriori dans la classe latente pour laquelle il a la plus grande probabilité a posteriori. Dans la littérature, c'est ce que l'on appelle la règle du maximum de la probabilité a posteriori (Nagin [2005]). La variable de classification C_i est égale à :

$$C_i = \arg\max_g \hat{\pi}_{ig} \tag{29}$$

Cette classification peut être utilisée pour décrire et comparer les caractéristiques des sujets a posteriori classés dans chaque classe latente obtenue (Xu and Hedeker [2001], Verbeke and Lesaffre [1996]).

3.1.2.5 Sélection / Adéquation du modèle

Pour déterminer le nombre optimal de classes latentes, on pourrait vouloir tester l'apport d'une nouvelle classe latente dans le modèle à l'aide d'un test de rapport de vraisemblance comme ce qui est habituellement fait pour comparer deux modèles statistiques emboîtés. Cependant, la statistique de test ne suit pas une distribution connue (le nombre de degrés de liberté du χ_2 est incertain (Nagin [2005], Proust-Lima [2006])). De plus, les estimations peuvent ne pas être identifiables pour certaines hypothèses nulles ou alternatives (Stram and Lee [1994], Proust-Lima [2006]). Pour résoudre ce problème, une approche par bootstrap a été proposée pour approximer la distribution asymptotique du test de rapport de vraisemblance. Mais cette approche peut nécessiter des temps de calculs assez longs si des effets aléatoires sont inclus dans le modèle (Schlattmann [2003], Han et al. [2007]).

Ainsi, on utilise des critères d'information tel que le Bayesian Information Criteria (BIC) (Schwarz et al. [1978]) pour évaluer la qualité de l'ajustement du modèle et obtenir le nombre de classes latentes optimal. Le BIC pénalise la déviance du modèle par le nombre de paramètres

multiplié par le logarithme du nombre de sujets. Il est souvent préféré à l'Akaike Information Criteria (AIC) (Akaike [1974]) qui pénalise la déviance du modèle par deux fois le nombre de paramètres et a tendance à surestimer le nombre de classes latentes (Hawkins et al. [2001], Han et al. [2007]). Un autre critère a été proposé par Han et al. [2007] nommé ICL-BIC. Il consiste à ajouter au calcul du BIC, deux fois la valeur de l'entropie estimée (calculée à partir des estimations des probabilités a posteriori). L'avantage d'utiliser ce critère pour le choix du nombre de classes latentes est qu'il permet de prendre en compte la qualité de la classification via l'entropie, ce qui n'est pas le cas avec les deux critères plus classiques (BIC et AIC). Cependant, ce critère reste sous utilisé dans la littérature par rapport au BIC. Nagin [2005] a, notamment, comparé les performances des 3 critères (AIC,BIC, ICL-BIC), et a conclu que le critère ICL-BIC ne permettait pas une meilleure évaluation du nombre de classes latentes par rapport à l'AIC ou au BIC.

Le critère BIC ne doit pas être le seul critère sur lequel la décision du nombre optimal de classes latentes doit être prise. En effet, il se peut que les valeurs du BIC décroissent indéfiniment pour un nombre croissant de classes latentes bien que ces classes supplémentaires n'apportent pas d'information supplémentaire (Nagin [2005]). Pour sélectionner le nombre optimal de classes latentes, il est ainsi également recommandé de se baser sur i) l'adéquation du modèle ; ii) la pertinence clinique des trajectoires identifiées et les effectifs associés aux classes et iii) la capacité discriminante du modèle à partir de la table de classification a posteriori (Muthén [2001]). La table de classification a posteriori représente la moyenne des probabilités a posteriori d'appartenance à chaque classe latente parmi les sujets classés a posteriori dans la classe. Une bonne discrimination est observée lorsque les éléments diagonaux de la table sont tous proches de 1 et les autres sont tous proches de 0 (Muthén [2001], Proust-Lima et al. [2017]). Dans ce cas, on pourra également dire que le modèle à classes latentes a un fort pouvoir discriminant.

3.1.2.6 Domaines d'applications des modèles à classes latentes

Twisk and Hoekstra [2012] ont relevé que les modèles à classes latentes étaient populaires en psychologie et en sciences sociales, et moins fréquemment utilisés en épidémiologie. Les applications concernent notamment le comportement face aux substances illicites, l'évolution des limites fonctionnelles chez les personnes âgées, le mal de dos, les désordres liées à l'anxiété ou le stress ou encore l'obésité.

Le modèle LCGM, en particulier, a d'abord été utilisé dans le domaine de la criminologie et du comportement des adolescents (Nagin [1999], Nagin and Tremblay [2005]). Depuis, beaucoup de travaux dans ces domaines utilisent le modèle LCGM (Côté et al. [2006], Monahan et al. [2009], Kokko et al. [2006]). Les travaux de Muthen (Muthén et al. [2002], Muthén [2004]) ont, quant à eux, largement contribué à l'application des modèles LCMM dans le domaine de

la pyschologie. En épidémiologie, le modèle LCMM a été utilisé pour étudier l'évolution de biomarqueurs comme l'antigène prostatique spécifique (PSA) (Proust-Lima and Taylor [2009], Lin et al. [2000]) ou les réponses aux tests psychométriques réalisés chez les personnes âgées (Proust et al. [2006]).

A notre connaissance, très peu de modèles à classes latentes ont été utilisés pour des données d'expositions environnementales comme le tabac (Dutra et al. [2017], Brook et al. [2008], De Genna et al. [2017], Weden and Miles [2012]) et aucun pour les expositions professionnelles.

3.1.2.7 Avantages et limites des modèles à classes latentes

Le modèle mixte à classes latentes est un modèle qui permet de traiter des données hétérogènes en faisant l'hypothèse qu'il existe un ensemble de sous-populations qui ont chacune leur propre profil d'évolution au cours du temps. Il permet également de faire de la classification a posteriori afin de décrire chacune de ces sous populations selon différentes caractéristiques. De plus, la variabilité des mesures répétées au sein d'un même individu peut être capturée par les effets aléatoires, pour LCMM, afin d'obtenir une meilleure classification.

Cependant, les modèles à classes latentes ont deux principaux inconvénients. Du fait de la multimodalité de la vraisemblance des modèles de mélange, on doit estimer le modèle pour différentes combinaisons de valeurs initiales des paramètres pour assurer la convergence vers un maximum global. Cependant, pour éviter à l'utilisateur de faire lui-même cette recherche de valeurs initiales, le package R *lcmm* propose la fonction gridsearch() qui permet l'estimation d'un modèle à G classes à partir de plusieurs combinaisons de valeurs initiales déterminées à partir de l'estimation du modèle à 1 classe latente. Le deuxième inconvénient qui découle du premier est que le temps de calcul pour estimer ces modèles est assez long (exemple : une dizaine d'heures environ pour l'estimation d'un modèle avec des trajectoires splines, des variances spécifiques à chaque classe et une transformation du marqueur quantitatif, pour 3807 sujets). La procédure SAS TRAJ basée sur le modèle LCGM plus simple (sans effet aléatoire), quant à elle, permet d'obtenir des résultats plus rapidement, mais c'est l'utilisateur qui doit penser à relancer l'estimation du modèle avec différentes combinaisons de valeurs initiales des paramètres.

3.1.2.8 Comparaison des modèles dans la littérature

Ces dernières années, les études comparatives entre les différents modèles de classification se sont multipliées dans la littérature. Beaucoup d'études ont comparé les modèles de classification les plus communément utilisés dans la littérature (Twisk and Hoekstra [2012]), à savoir K-moyennes, LCMM et LCGM. En effet, cela permet à la fois de comparer les comportements d'une approche non paramétrique versus semi-paramétrique (K-moyennes vs modèles à variable

latente), ainsi que deux approches dont l'une est le cas particulier de l'autre (LCGM vs LCMM). Cependant, la principale difficulté pour comparer ces modèles entre eux est de trouver la mesure ou le critère de performance le plus pertinent (Gelbard et al. [2007], Nguyen and Rayward-Smith [2008]). Afin de déterminer la qualité de la classification, l'homogénéité entre les individus d'un même groupe et l'hétérogénéité entre les groupes sont notamment à évaluer. De plus, classiquement, on calcule le taux de concordance entre les clusters identifiés et ceux d'origine lorsque la comparaison des méthodes se fait à partir de données simulées. Des variantes peuvent être considérées comme Gelbard et al. [2007] qui ont calculé un score d'évaluation globale de la classification basé sur les calculs du taux de concordance pour chaque cluster et la moyenne de ces taux sur le nombre total de clusters. Par ailleurs, les performances obtenues (Gelbard et al. [2007]) différaient d'un jeu de données à l'autre pour une même méthode puisque ces méthodes sont assez sensibles aux hypothèses associées et aux données utilisées. L'étude de Eshghi et al. [2011] a également montré que l'on pouvait obtenir des interprétations potentiellement différentes d'une méthode de classification à l'autre (K-moyennes & modèles à classes latentes). A travers ces deux études, on se rend compte de la difficulté d'opter pour la méthode de classification la plus adaptée au contexte dans lequel on veut l'utiliser.

Genolini et ses collaborateurs ont comparé le package R *kml* et la procédure SAS TRAJ afin de montrer l'efficacité de leur extension de l'approche K-moyennes pour données longitudinales. Ils définissent deux critères pour évaluer les performances dont un qui diffère de ceux utilisés classiquement dans les études comparatives citées précédemment (Genolini and Falissard [2011]). Ils se basent sur la distance entre la trajectoire observée et la trajectoire théorique à travers un critère appelé DOT. 4 bases de données ont été simulées, caractérisées chacune par différentes formes de trajectoires. Les résultats étaient assez proches du point de vue de ce critère et du taux de classifiés pour 3 bases de données. Mais ils ont observé une forte différence des résultats entre les deux approches quand les trajectoires avaient été simulées suivant des lois normales. Ils ont pu constater des échecs de convergence pour TRAJ dans ce cas précis.

Twisk et ses collaborateurs ont comparé les résultats en terme de trajectoires et du nombre de classes latentes identifiées suivant 5 méthodes dont K-moyennes, LCGM et LCMM, à partir de deux jeux de données simulés et une base de données provenant d'une étude longitudinale observationnelle sur des adolescents vivants près d'Amsterdam (Twisk and Hoekstra [2012]). Le premier jeu de données simulé était constitué de quatre classes de trajectoires linéaires, le second de deux classes de trajectoires linéaires et deux classes avec trajectoires quadratiques. LCGM performait bien pour détecter les trajectoires linéaires ; avec 91% des sujets qui étaient classés dans leurs classes d'origine et des trajectoires linéaires assez proches de celles simulées. Pour ce scénario, l'approche K-moyennes avait tendance à donner de moins bons résultats en terme de classification et de trajectoires moyennes estimées. Pour le scénario avec les trajectoires quadratiques, aucune des 5 méthodes ne semblaient bien performer, même si LCMM

donnait les résultats les plus cliniquement interprétables. LCGM donnait toujours un nombre plus grand de classes que LCMM, comme attendu. L'approche K-moyennes donnait par contre le bon nombre attendu de classes, même si le critère était basé sur des distances et non sur un critère de sélection tel que le BIC.

Dans ces études comparatives (Gelbard et al. [2007], Nguyen and Rayward-Smith [2008], Eshghi et al. [2011]), les aspects logiciel ont très peu été décrits. Seuls Twisk and Hoekstra [2012] ont précisé avoir réalisé leur étude comparative sous le même logiciel (Mplus) afin de s'assurer que les différences observées n'étaient pas dues aux logiciels. Aucune de ces études comparatives n'a indiqué avoir utilisé le package R *lcmm* (Proust-Lima et al. [2017]) pour estimer le modèle LCMM.

3.2 Les méthodes pour évaluer l'association entre les trajectoires d'intensité d'exposition et la survenue d'un évènement

Muthén and Shedden [1999] puis Lin et al. [2000] ont été les premiers à proposer de modéliser conjointement les profils d'évolution par un modèle de classification et un évènement clinique. D'autres approches, plus simples à mettre en oeuvre, peuvent aussi être considérées pour évaluer l'association entre la survenue d'un évènement clinique et les différents profils d'évolution. Parmi celles-ci, les approches les plus communes sont les approches dites "classify-analyze" (par Lanza and Rhoades [2013]), que l'on nommera "approche en 2 étapes" dans la suite de ce manuscrit.

3.2.1 Approche en 2 étapes

L'approche en 2 étapes est souvent associée à deux méthodes assez communément utilisées dans la littérature. La première est connue sous "maximum probability assigment rule" (Nagin [2005]), que l'on nommera approche "naïve" par la suite. La seconde est appelée "multiple pseudo-class draws approach" (Bandeen-Roche et al. [1997], Wang et al. [2005]), que l'on nommera approche par Pseudo-classes dans la suite du manuscrit.

3.2.1.1 Approche dite "naïve"

L'approche "naïve" a l'avantage d'être simple à mettre en œuvre. Dans un premier temps, on estime les différents profils d'évolution par le modèle de classification et on attribue à chaque individu une classe latente pour laquelle il a une probabilité a posteriori maximum d'appartenir.

Puis l'appartenance aux classes est traitée comme observée, elle est utilisée pour prédire l'évènement D dans un second temps de l'analyse (Clogg [1995]). Cette appartenance aux classes est représentée par une variable qualitative C à G modalités pour une estimation d'un modèle à G classes latentes. Pour estimer l'association avec l'évènement D, un modèle à risques instantanés proportionnels peut être considéré pour des données de cohortes ou d'essais cliniques si le délai d'évènement est connu pour les sujets présentant l'évènement. Un modèle de régression logistique peut être envisagé pour des données cas-témoins (D représente alors le statut cas-témoin). Dans le cadre de cette thèse, nous disposions de données provenant d'une étude cas-témoins, et c'est donc le modèle logistique qui sera utilisé dans la suite de ce manuscrit pour illustrer les différentes approches.

Le modèle logistique pour l'approche "naïve" est le suivant :

$$\text{logit}(P(D_i = 1|C_i, X_{ri})) = \gamma_0 + \gamma_{1l} C_i + X_{ri}^T \gamma_p \tag{30}$$

où

- γ_0 est l'intercept ;
- γ_{1l} l=2,. . . ,G sont les $G-1$ coefficients associés aux $G-1$ modalités de la variable de classification C_i (référence : C_i=1) ;
- X_{ri}^T est le vecteur des p variables explicatives associé au vecteur des coefficients γ_p.

Pour comparer la côte de maladie de la classe l par rapport à celle de la classe de référence, on estime le rapport de côte (OR) suivant :

$\widehat{OR} = \exp(\hat{\gamma}_{1l})$, avec son intervalle de confiance associé $\exp\left(\hat{\gamma}_{1l} \pm 1.96\sqrt{\widehat{Var}(\hat{\gamma}_{1l})}\right)$

où $\hat{\gamma}_{1l}$ est l'estimateur du coefficient de régression γ_{1l} associé et $\widehat{Var}(\hat{\gamma}_{1l})$ l'estimateur de sa variance.

La principale limite de cette approche "naïve" est qu'elle fait l'hypothèse que la classe est assurée, ce qui n'est pas le cas. Pour attribuer une classe à chaque individu, on s'appuie sur les estimations des paramètres du modèle afin de calculer les probabilités individuelles d'appartenir à chacune des classes et on classe le sujet dans la classe à laquelle il a la plus forte probabilité d'appartenir. Il existe donc une incertitude sur l'appartenance aux classes, qui n'est pas correctement prise en compte avec cette approche naïve. Elle peut donc amener à des biais quant à l'estimation de l'effet de C sur D. De plus, la corrélation entre le marqueur longitudinal et la survenue d'évènement n'est pas prise en compte dans cette approche.

3.2.1.2 Approche par régression logistique pondérée

Une alternative à cette approche naïve est de tenir compte de l'incertitude de la classification en pondérant chaque sujet par sa probabilité d'appartenance à la classe. Cette approche est appelée "proportional assignment". Chaque sujet contribue à toutes les classes mais avec des poids différents associés à ses probabilités d'appartenance aux classes. Les sujets contribuent donc G fois dans le modèle et ne sont plus affectés à une seule classe latente. L'estimateur sandwich de la variance (Huber et al. [1967], White [1996]) peut être envisagé afin de tenir compte de la corrélation intra-sujet. Cependant, divers auteurs, comme Bray et al. [2015], ont montré qu'il subsiste toujours un biais dans cette approche (et l'approche naïve) car le processus longitudinal et celui de l'évènement ne sont pas estimés simultanément.

3.2.1.3 Approche par Pseudo-classes

L'approche par Pseudo-classes, introduite par Bandeen-Roche et al. [1997] et Wang et al. [2005], est similaire à la technique d'imputation multiple. Elle considère que la variable de classification (C) est manquante. Une classe latente est affectée de manière aléatoire à chaque sujet, en considérant que la variable de classification suit une loi multinomiale dont les paramètres sont les probabilités a posteriori du sujet d'appartenir à chacune des classes. L'estimation du modèle logistique est réalisée en incluant comme covariable cette variable de pseudo-classification (PC). Cette procédure est répétée M fois. Wang et al. [2005] ont considéré que répéter la procédure 20 fois était suffisant pour tenir compte de l'incertitude liée à l'appartenance aux classes latentes. Les résultats sont ensuite combinés suivant ces M modèles estimés en utilisant les règles provenant de la théorie de l'imputation multiple des données manquantes (Rubin [1987]). Les coefficients de régression finaux estimés sont calculés en effectuant la moyenne des coefficients de régression de chacun des modèles de régressions. Les variances associées sont calculées en tenant compte de la variabilité intra-pseudo-classe et inter-pseudo-classe.

Plus précisément, l'algorithme de la mise en place de l'approche est le suivant :

— M tirages aléatoires réalisés pour affecter une pseudo-classe à tous les individus de l'échantillon en se basant sur la distribution des probabilités a posteriori $(\hat{\pi}_{i1}, .., \hat{\pi}_{iG})$; –> PC_i^m, m=1,. . . , M et i=1,. . . , n

— M modèles de régressions logistiques estimés :

$$\text{logit}(P(D_i = 1|X_{ri}, PC_i^m)) = \gamma_0 + \gamma_l^{(C_m)} PC_i^m + X_{ri}^T \gamma_p \tag{31}$$

où

— la variable PC_i^m est la variable de pseudo-classification à la $m^{è}$ régression logistique associée aux coefficients $\gamma_l^{(C_m)}$, l=2, . . . , G ;

— X_{ri}^T est le vecteur des p variables explicatives associé au vecteur des coefficients γ_p.

L'estimateur de $\hat{\gamma}_g^{(C)}$ est : $\hat{\gamma}_g^{(C)} = \frac{1}{M}\sum_m \hat{\gamma}_g^{(C_m)}$.

La variance asymptotique de $\hat{\gamma}_g^{(C)}$ est obtenue à partir de la théorie d'imputation multiple (Rubin [1987], Schafer [1997]) par :

$$\widehat{Var}(\hat{\gamma}_g^{(C)}) = \hat{V}_W + (1 + \frac{1}{M})\hat{V}_B \tag{32}$$

où

— $\hat{V}_W$ est la variance intra-pseudo-classe : $\hat{V}_W = \frac{1}{M}\sum_m \widehat{Var}(\hat{\gamma}_k^{C_m})$

— $\hat{V}_B$ est la variance inter-pseudo-classe : $\hat{V}_B = \frac{1}{M-1}\sum_m (\hat{\gamma}_k^{(C_m)} - \hat{\gamma}_k^{(C)})^2$

Cette approche reste assez simple à mettre en oeuvre. Elle permet de prendre en compte l'incertitude liée à l'appartenance aux classes dans l'estimation de l'association entre la variable de classification et l'évènement. Cependant, l'étude comparative de Bray et al. [2015] n'a pas montré que l'approche par Pseudo-classes était supérieure à l'approche naïve. Cette dernière montrait moins de biais en terme d'estimations et une moindre atténuation de l'effet. De plus, dans leur étude de simulations, lorsque la mesure d'association entre les classes latentes et l'évènement était considérée forte, les résultats avec l'approche naïve avaient de plus faibles RMSE que ceux obtenus avec l'approche Pseudo-classes. Néanmoins, ils ont relevé qu'il y avait une moins grande variabilité dans les résultats pour l'approche Pseudo-classes. Pourtant les biais engendrés étaient trop importants pour que l'on puisse dire qu'elle performe mieux que l'approche naive. D'autres auteurs (Bray et al. [2015] , Asparouhov and Muthén [2014]) ont pu faire les mêmes constats. Bolck et al. [2004] ont démontré que quelque soit l'approche en 2 étapes considérée, l'association entre l'évènement et la variable de classification était atténuée et que plus le taux d'erreur de classification était grand plus le biais sur les estimations était important. Ce qui peut s'expliquer par la non prise en compte de la corrélation entre le marqueur longitudinal et la survenue d'évènement dans ces approches.

3.2.2 La méthode envisagée : le modèle conjoint à classes latentes (JLCMM)

3.2.2.1 Définition

L'hypothèse d'un modèle conjoint à classes latentes est qu'il existe des sous-populations homogènes en terme d'évènement et d'évolution du marqueur longitudinal. Chacune d'entre elles peut avoir sa propre évolution du marqueur longitudinal et par conséquent son propre risque

d'évènement associé. Pour prendre en compte de manière conjointe les profils d'évolution et un évènement clinique sous forme de variable binaire, Muthén and Shedden [1999] puis Lin et al. [2000] ont donc défini la probabilité de l'évènement D dans chaque classe latente g par un modèle de régression logistique.

$$P(D_i|c_i = g, X_{ri}) = \frac{\exp(\gamma_{0g} + X_{ri}^T \gamma_{lg})}{1 + \exp(\gamma_{0g} + X_{ri}^T \gamma_{lg})} \tag{33}$$

où

— γ_{0g} est l'intercept spécifique à chaque classe ;

— X_{ri} est le vecteur des variables explicatives associé au vecteur de coefficients γ_{lg} potentiellement différents dans chaque classe latente.

Lin et al. [2002b] et McCulloch et al. [2002] ont étendu le modèle en prenant en compte le délai jusqu'à l'évènement. Le risque instantané d'évènement était modélisé par un modèle à risques proportionnels :

$$\lambda\left(t|c_i = g\right) = \lambda_{0g}(t) \exp(X_{ri}^T \gamma_{1g}) \tag{34}$$

où $\lambda_{0g}(t)$ est le risque instantané de base de l'évènement dans la classe latente g.

Le modèle est estimé simultanément avec le modèle multinomial pour l'appartenance aux classes et le modèle mixte pour le marqueur longitudinal (équations 20 et 21 de la section 3.1.2.2.a). Dans le modèle conjoint à classes latentes, la variable c_i représente le seul lien entre l'évolution des mesures répétées et l'évènement.

Ces dernières années, ces modèles conjoints à classes latentes ont principalement été utilisés avec des modèles de survie car ils sont souvent appliqués dans des études de cohortes pour étudier le lien entre un biomarqueur quantitatif répété dans le temps et un évènement (Lin et al. [2000], Muthén and Shedden [1999], Proust et al. [2006]). Ils peuvent également être utilisés dans un but prédictif (Blanche et al. [2015], Proust-Lima et al. [2007]), même si dans ce cas, les modèles conjoints à effets aléatoires partagés sont plus souvent utilisés que les modèles conjoints à classes latentes (Rizopoulos [2012]). Par ailleurs, les modèles conjoints à classes latentes ont récemment été étendus pour des données longitudinales multivariées (Proust-Lima et al. [2007], Proust-Lima et al. [2009]), ainsi que pour l'étude d'évènements récurrents (Han et al. [2007]).

A notre connaissance, aucune étude n'a utilisée un modèle conjoint à classes latentes pour identifier différents profils de trajectoires d'exposition environnementale ou professionnelle et estimer leur association avec le risque de cancer, à partir de données cas-témoins.

3.2.2.2 Estimation

Une modélisation conjointe, qui lie l'évolution des données répétées à l'évènement d'intérêt grâce à la variable de classification c_i, permet une estimation simultanée des deux modèles considérés. Cette variable latente partagée est discrète, on peut donc écrire la densité conjointe comme la somme sur les classes latentes.

Sous l'hypothèse d'indépendance conditionnelle des mesures répétées et de l'occurrence de l'évènement sachant la classe latente à laquelle appartient le sujet i (Lin et al. [2002a]), la distribution conjointe peut s'écrire comme suit :

$$f(Y_i, D_i) = \sum_{g=1}^{G} f(Y_i|_{c_i=g}) f(D_i|_{c_i=g}) f(c_i = g) \tag{35}$$

La vraisemblance s'écrit :

$$L(\theta; Y_i, D_i) = \prod_{i=1}^{n} \sum_{g=1}^{G} \pi_{ig} f(Y_i|_{c_i=g}) f(D_i|_{c_i=g}) \tag{36}$$

où

— $f(Y_i|_{c_i=g})$ est la distribution multivariée gaussienne si Y_i est gaussien ;
— $f(D_i|_{c_i=g})$ est la distribution du modèle logistique ou du modèle à risques proportionnels.

Les effets aléatoires n'interviennent que dans le modèle mixte qui, dans le cas de Y_i gaussien, ont une vraisemblance analytique. Ainsi, même si l'approche du modèle conjoint paraît plus complexe en considérant 3 sous-modèles simultanément, la vraisemblance d'un tel modèle a souvent une forme analytique.

Les estimateurs du maximum de vraisemblance des paramètres sont obtenus par une méthode itérative en utilisant l'algorithme de Marquardt modifié par Proust et al (2005) précédemment détaillé en section 3.1.2.2.b.

Le modèle conjoint à classes latentes est associé aux mêmes difficultés d'estimations que le modèle à classes latentes :

— une estimation à un nombre fixé de classes latentes ;
— une estimation du modèle suivant plusieurs combinaisons de valeurs initiales des paramètres pour assurer la convergence vers un maximum global.

3.2.2.3 Classification a posteriori

Pour un modèle conjoint à classes latentes avec une variable binaire D, les probabilités a posteriori conditionnelles aux données longitudinales et à l'évènement sont définies par :

$$\hat{\pi}_{ig}^{y,D} = P(c_i = g|Y_i, D_i; \hat{\theta}) = \frac{P(c_i = g) f((Y_i, D_i)|c_i = g; \hat{\theta})}{\sum_{l=1}^{G} P(c_i = l) f((Y_i, D_i)|c_i = l; \hat{\theta})} \quad (37)$$

où le dénominateur représente la contribution individuelle à la vraisemblance à l'optimum et $\hat{\theta}$ est le vecteur des paramètres estimés (Proust-Lima et al. [2007]).

3.2.2.4 Sélection / Adéquation du modèle

Le choix du nombre optimal de classes latentes dans un modèle conjoint (JLCMM) s'appuie sur les stratégies détaillées pour le modèle à classes latentes (voir section 3.1.2.5). En effet, plusieurs JLCMM sont donc estimés avec un nombre différent de classes latentes et ensuite sont comparées, les valeurs du BIC, la table de classification a posteriori, et la pertinence clinique des trajectoires moyennes prédites du marqueur longitudinal.

Ce qui change par rapport au LCMM, c'est que l'on doit également tenir compte de l'adéquation du troisième sous-modèle du JLCMM, ici un modèle de régression logistique. Ceci est classiquement fait par 1) le choix pertinent des variables considérées dans le modèle et 2) l'étude de la linéarité de l'effet des variables quantitatives sur le logit.

3.2.2.5 Avantages et limites

Contrairement aux approches en 2 étapes, un tel modèle permet de prendre en compte correctement l'incertitude liée à l'estimation de l'appartenance aux classes latentes (Proust-Lima et al. [2007]) et donc de réduire les variances des effets estimés dans le troisième sous-modèle (logistique ou à risques proportionnels). Grâce à la variable latente discrète, la corrélation entre les deux processus (évolution longitudinale et évènement) est prise en compte. La corrélation des mesures répétées d'un même individu est prise en compte par les effets aléatoires. Ces deux types de corrélation sont bien distincts pour le modèle conjoint à classes latentes. L'interprétation des résultats est facilitée par la représentation, pour chacune des classes latentes, des profils de trajectoires du marqueur longitudinal d'une part, et des courbes de survie ou des OR (95% IC) d'autre part. Enfin, la vraisemblance reste analytique dans la majeure partie des cas et reste facile à maximiser.

Concernant les limites d'une telle approche, elles sont similaires à celles relevées pour les modèles à classes latentes (convergence vers des maxima locaux, estimation pour un nombre fixé de classes latentes). Une forte hypothèse d'indépendance conditionnelle aux classes latentes est une autre limite de ce modèle. C'est ce qui nous permet d'écrire la vraisemblance conjointe (équation 36) puisqu'elle signifie qu'il y a indépendance entre le marqueur longitudinal et l'évènement suivant les classes latentes. Elle peut être vérifiée par la mise en oeuvre d'un score test, avec l'hypothèse nulle d'indépendance entre le marqueur longitudinal et l'évènement sachant les classes latentes versus l'hypothèse alternative que le risque d'évènement dépend des effets aléatoires du modèle mixte. Le calcul de la statistique du score s'appuie sur les dérivées premières de la vraisemblance suivant les paramètres estimés, ce qui a été détaillé par Jacqmin-Gadda et al. [2010].

3.3 Applications au cancer du poumon

Nous avons identifié les trajectoires d'intensité d'exposition prolongée (consommation de tabac et exposition professionnelle à l'amiante) et comparé les risques de cancer du poumon associés entre ces différentes trajectoires. Pour cela, nous avons estimé un modèle conjoint à classes latentes pour chaque exposition avec les données provenant de l'étude cas-témoins ICARE, présentée dans le chapitre 1, avec le package R *lcmm* (Proust-Lima et al. [2017]). Pour les deux relations, nous avons limité la population d'étude aux exposés car nous nous intéressions aux trajectoires d'intensité d'exposition dans cette population. Comme nous avons pu le discuter dans le chapitre 1, seuls les hommes ont été considérés puisque les hommes et les femmes ont des niveaux d'exposition très différents et potentiellement aussi des risques de cancer de poumon différents (Papadopoulos et al. [2014]). Il n'était donc pas envisageable de les regrouper sur un plan conceptuel, ni de répliquer les analyses dans les deux groupes pour une question de temps. Nous avons choisi les hommes parce qu'ils étaient plus nombreux que les femmes (5056 vs 1425, voir diagramme de flux du chapitre 1, section 4) et nous permettent donc de réduire des problèmes de convergence potentiels.

3.3.1 Relation tabac-cancer du poumon

3.3.1.1 Sélection des sujets

Pour cette relation, nous avons sélectionné les hommes ayant fumé en moyenne au moins une cigarette par jour pendant au moins une année. L'estimation du modèle conjoint a ainsi été réalisée sur 1969 cas et 1838 témoins (voir diagramme de flux se trouvant dans le matériel supplémentaire de l'article 3).

3.3.1.2 Spécification du modèle

Les mesures répétées d'intérêt étaient les intensités annuelles de consommation de tabac représentées par le nombre moyen de cigarettes fumées par jour sur une année. L'axe du temps utilisé était le temps depuis la date index (date de diagnostic pour les cas et date d'interview pour les témoins) en années discrètes. Les données manquantes dans un modèle conjoint à classes latentes sont considérées comme manquantes aléatoirement (Missing At Random, MAR). Un tel processus considère que pour les temps où il n'y a pas d'information, la trajectoire garde la même évolution moyenne que celle estimée aux autres temps. Ainsi, cela peut surestimer la valeur de cette trajectoire en ces temps, ne reflétant pas correctement toute l'information sur les histoires d'exposition individuelles. Pour l'éviter, nous avons défini une période d'exposition potentielle sur laquelle le sujet était considéré potentiellement exposé. Ainsi, en dehors de l'histoire d'exposition du sujet (définie entre l'âge à l'initiation et l'âge à l'arrêt) sur cette période, des intensités nulles étaient attribuées.

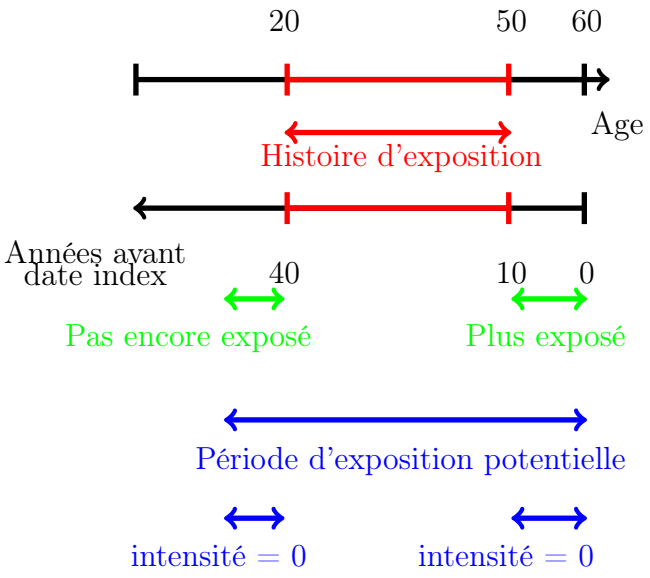

FIGURE 3.1 – Schéma explicatif de la période d'exposition potentielle pour un sujet fictif âgé de 60 ans à la date index, exposé entre 20 et 50 ans

Prenons l'exemple d'un sujet fictif âgé de 60 ans à la date index et exposé de 20 à 50 ans (figure 3.1), il aurait donc des intensités d'exposition associées aux âges entre 20 et 50. Comme nous voulions estimer les profils de trajectoires en fonction du temps avant la date index, nous avons associé ces intensités aux temps 10 à 40 avant la date index. Or, en dehors de ces temps, on ne distingue pas que le sujet ne soit plus exposé dans les 10 années avant la date index, ni qu'il n'ait été exposé qu'à partir de 20 ans. Ainsi, pour en tenir compte et éviter une surestimation de la trajectoire aux temps en dehors de 10 à 40 ans avant la date index, nous avons attribué des intensités nulles sur toute une période d'exposition potentielle en dehors des années de l'histoire d'exposition. La période d'exposition potentielle était considérée entre 12 ans et

l'âge à la date index. Un âge minimum de 12 ans a été considéré afin de ne pas attribuer des intensités nulles à des âges pour lesquels une exposition semblait peu réaliste. Nous avons eu le même raisonnement pour les intensités liées à l'exposition professionnelle à l'amiante.

Sous-modèle 1 :

$$\pi_{ig} = P(c_i = g) = \frac{\exp(\zeta_{0g})}{\sum_{l} \exp(\zeta_{0l})} \tag{38}$$

où

— c_i : variable latente discrète qui est égale à g si le sujet i appartient à la classe latente g (g=1,..., G) ;

— ζ_{0g} : intercept pour la classe latente g. Par soucis d'identifiabilité ζ_{0G}=0 ;

— Aucune variable explicative n'a été considérée prédictive de la probabilité d'appartenance aux classes.

Sous-modèle 2 :

Pour tenir compte de la non normalité des intensités annuelles de consommation de tabac (figure 3.2), nous avons utilisé une transformation des données via un **modèle mixte à processus latent** spécifique à chaque classe qui est lié aux données par une fonction de lien paramétrique (Proust et al. [2006], Proust-Lima et al. [2013]). Nous avons choisi la transformation I-spline implémentée dans le package R *lcmm*. Elle était la plus adaptée aux types de données parmi celles disponibles dans le package.

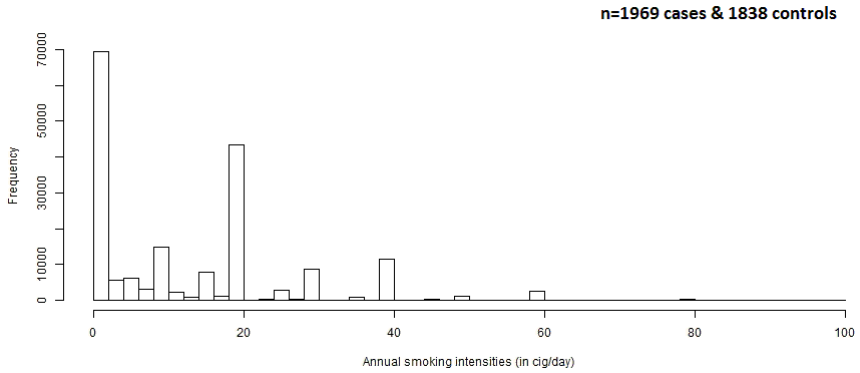

FIGURE 3.2 – Distribution des intensités annuelles de consommation tabagique

Pour définir les fonctions I-splines quadratiques utilisées pour la transformation, nous avons testé entre 3 et 5 noeuds sur le modèle à une classe latente. Nous avons évalué l'impact de la position des noeuds en les fixant a priori, à équidistance, ou en fonction de la distribution des intensités. Après avoir comparé les résultats en terme de BIC et de la forme de la transforma-

tion estimée, nous avons choisi de considérer 3 nœuds ; deux noeuds positionnés aux valeurs extrêmes de la distribution (0 et 100 cigarettes/jour) et un noeud positionné à 20 cigarettes/jour.

Les effets aléatoires :
Pour éviter l'application des effets aléatoires sur les temps en dehors de l'histoire d'exposition du sujet, nous avons défini une variable indicatrice ($Hist_i$) qui valait 1 pour les temps t_{ij} appartenant à l'histoire de l'exposition du sujet i et 0 sinon. Cette indicatrice était associée à l'intercept aléatoire afin qu'il soit défini seulement sur les temps de l'histoire d'exposition du sujet.

La forme des trajectoires :
Au vu des trajectoires individuelles observées dans la population d'étude (figure 3.3) , nous avons décidé de les modéliser de manière flexible avec des fonctions splines naturelles cubiques plutôt que des fonctions polynomiales. En considérant 3 noeuds intérieurs, cela permettait d'obtenir une évolution assez flexible tout en limitant le nombre de paramètres à estimer. Ils ont été positionnés suivant les quartiles de la distribution des temps de mesures de la population d'étude pour éviter les problèmes de convergence du modèle.

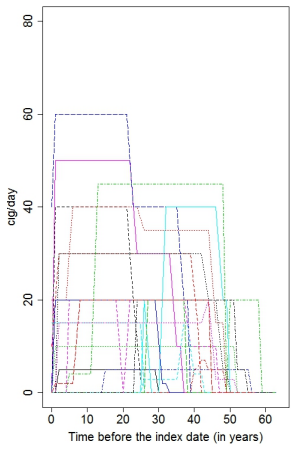

FIGURE 3.3 – Quelques trajectoires individuelles observées

Le sous-modèle 2 considéré s'écrit donc :

$$Y_{ij} = H(\tilde{Y}_{ij}) = H(\Lambda_i(t_{ij}) + \epsilon_{ij}) \tag{39}$$

où

— Y_{ij} : nombre moyen de cigarettes fumées par jour par l'individu i au cours de l'année j ;

— H : fonction de lien paramétrique où H^{-1} est une fonction représentée par une base de I-splines quadratiques avec 3 noeuds placés à 0, 20 et 100 cigarettes/jour ;

— $\tilde{Y}_{ij}$: processus latent bruité au temps t_{ij} ;

— t_{ij} : la jème année avant la date index pour l'individu i ;

— ϵ_{ij} : erreurs de mesures gaussiennes supposées indépendantes de variance σ^2_ϵ.

$\Lambda_i(t_{ij})$ est le processus latent défini par un modèle mixte spécifique à chaque classe latente :

$$\Lambda_i(t_{ij})|_{c_i=g} = \beta_{0g} + u_{0ig}\mathbb{1}_{t_{ij}\in Hist_i} + \sum_l \beta_{lg}B_l(t_{ij}) \tag{40}$$

Où

— $\beta = c(\beta_{0g}, \beta_{lg})$: vecteur des effets fixes spécifiques à chaque classe ;

— $B_l(t)$: base de fonction de splines cubiques du temps avant la date index avec 3 noeuds intérieurs placés aux quartiles (10, 21, et 33 ans) ;

— u_{0ig} : intercept aléatoire spécifique à chaque classe supposant que $u_{0ig} \sim \mathcal{N}(0, \omega^2_g\sigma^2_u)$ où σ^2_u est une variance commune et ω_g un coefficient de proportionnalité, ce qui permet une variabilité spécifique à chaque classe ;

— $\mathbb{1}_{t_{ij}\in Hist_i}$: variable indicatrice qui vaut 1 si t_{ij} se situe durant l'histoire d'exposition du sujet i, 0 sinon.

Sous-modèle 3 :

Les variables d'ajustement :

Les facteurs d'appariement (âge à la date index et départements de résidence) ont été inclus dans le modèle de régression logistique. Pour ajuster sur l'exposition professionnelle à l'amiante, nous avons utilisé l'indice cumulé classique de cette exposition (ICE). Son effet sur le logit de la probabilité n'était pas linéaire. Pour tenir compte de cet effet non linéaire, il a été catégorisé suivant les quartiles de la distribution, afin de limiter le nombre de paramètres à estimer par rapport à une modélisation plus flexible avec des splines, nécessitant plus de paramètres à estimer.

$$\text{logit}(P(D_i = 1|Age, Dpart, CIE_{Am}, c_i = g)) = \gamma_{0g} + f_1(Age_i) + f_2(Dpart_i) + f_3(CIE_{Am_i}) \quad (41)$$

où

— D_i : variable binaire qui correspond au statut cas-témoin (=1 pour cas, 0 pour témoin) ;

— γ_{0g} : intercept spécifique à la classe latente g ;

— $f_1(Age)$: splines naturelles cubiques de l'âge à la date index avec 3 noeuds placés aux 5ème, 50ème et 95ème percentiles de la distribution pour tenir compte de l'effet non linéaire de l'âge sur le logit (43, 60, 73 ans) ;

— $f_2(Dpart)$: combinaison linéaire des 9 variables indicatrices des départements de résidence (département pris en référence : la Somme) ;

— $f_3(CIE_{Am_i})$: combinaison linéaire des indicatrices définissant les quartiles de la distribution de l'indice cumulé de l'exposition professionnelle à l'amiante (0, <=0.00153, <=0.8159, >0.8159 f/mL-années).

Calcul des OR ajustés et des IC à 95% associés :
Pour un modèle JLCMM à 4 classes latentes, si on prend la classe 1 comme classe de référence, on calcule les OR associés, respectivement, entre les classes 2, 3, 4 et la classe 1.

Par exemple,
- l'OR pour le cancer du poumon entre classe 2 et classe 1 : $\exp(\hat{\gamma}_{02} - \hat{\gamma}_{01})$
- l'IC associé : $\exp((\hat{\gamma}_{02} - \hat{\gamma}_{01}) \pm 1.96 \times \sqrt{\hat{V}_{int}})$
avec $\hat{V}_{int} = var(\hat{\gamma}_{02}) + var(\hat{\gamma}_{01}) - 2cov(\hat{\gamma}_{01}, \hat{\gamma}_{02})$

3.3.2 Relation amiante-cancer du poumon

3.3.2.1 Sélection des sujets

Pour cette relation, nous avons sélectionné les hommes exposés professionnellement à l'amiante. Les hommes considérés comme exposés devaient avoir une probabilité d'exposition à l'amiante non nulle sur au moins un emploi occupé (voir chapitre 1 section 1.3.2). Nous avions ainsi une population d'étude de 1417 cas et 1520 témoins. Nous avons rencontré des problèmes de convergence pour l'estimation du modèle conjoint à classes latentes à partir de cet échantillon. Ces problèmes étaient principalement dus à une forte proportion de très faibles intensités d'exposition dont la distribution était par conséquent très éloignée de la distribution gaussienne.

Pour pallier à ce problème, nous avons exclu de cet échantillon les sujets qui ont été très faiblement exposés durant leur carrière professionnelle. Pour ce faire, nous avons donc exclu les hommes exposés qui avaient cumulé, tout au long de leur carrière professionnelle, une dose inférieure à 0.26 f/mL-années. Ce seuil, que nous avons choisi a priori, correspond à une intensité journalière de 0.01 f/mL pendant 26 ans. Le seuil de 0.01 f/mL correspond au seuil de la VLEP actuelle fixée par la législation française (voir Introduction). La durée de 26 ans correspond à la durée totale moyenne d'exposition professionnelle à l'amiante de l'échantillon des 1417 cas et 1520 témoins. Cette sélection a conduit à un échantillon final de 912 cas et 798 témoins (voir diagramme de flux du matériel supplémentaire de l'article 3).

3.3.2.2 Spécification du modèle

Sous-modèle 1 :

$$\pi_{ig} = P(c_i = g) = \frac{\exp(\zeta_{0g})}{\sum_{l=1}^{G} \exp(\zeta_{0l})} \tag{42}$$

où

— c_i : variable latente discrète qui est égale à g si le sujet i appartient à la classe latente g (g=1,. . . , G) ;

— ζ_{0g} : intercept pour la classe latente g. Par soucis d'identifiabilité ζ_{0G}=0.

Sous-modèle 2 :

Malgré l'exclusion des sujets ayant été très faiblement exposés, la distribution des intensités annuelles n'était pas gaussienne (figure 3.4). Nous avons donc utilisé une transformation spline des données comme pour l'application pour le tabac.

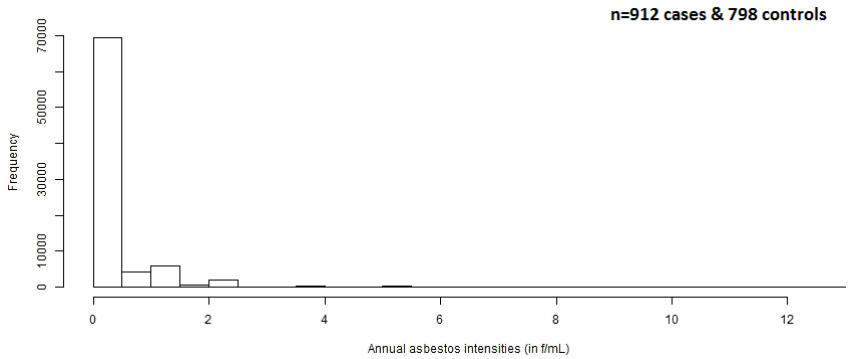

FIGURE 3.4 – Distribution des intensités annuelles d'exposition professionnelle à l'amiante

De la même manière que pour le tabac, nous avons défini une indicatrice $Hist_i$ qui sera associée à l'intercept aléatoire afin qu'il soit seulement appliqué sur les temps correspondant à

l'histoire d'exposition du sujet.

Au vu des trajectoires individuelles observées (figure 3.5), nous avons également utilisé des fonctions splines naturelles cubiques du temps pour modéliser la trajectoire moyenne au sein de chaque classe.

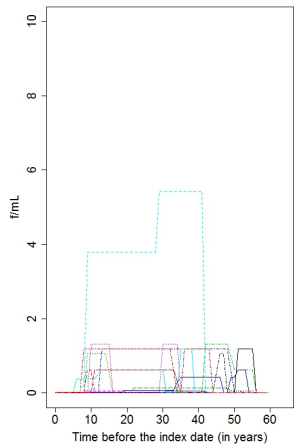

FIGURE 3.5 – Quelques trajectoires individuelles observées

Le sous-modèle 2 considéré s'écrit comme :

$$Y_{ij} = H(\tilde{Y}_{ij}) = H(\Lambda_i(t_{ij}) + \epsilon_{ij}) \tag{43}$$

où

— Y_{ij} est l'intensité annuelle moyenne journalière en éq f/mL pour l'individu i durant l'année j ;

— H : fonction de lien paramétrique où H^{-1} est une fonction représentée par une base de I-splines quadratiques avec 3 noeuds placés à 0, 0.1 et 12.6 éq f/mL ;

— $\tilde{Y}_{ij}$: processus latent bruité au temps t_{ij} ;

— t_{ij} : la jème année avant la date index pour l'individu i ;

— ϵ_{ij} : erreurs de mesures gaussiennes supposées indépendantes de variance σ^2_ϵ.

$\Lambda_i(t_{ij})$ est le processus latent défini par un modèle mixte spécifique à chaque classe latente :

$$\Lambda_i(t_{ij})|_{c_i=g} = \beta_{0g} + u_{0ig}\mathbb{1}_{t_{ij}\in Hist_i} + \sum_l \beta_{lg}B_l(t_{ij}) \tag{44}$$

où

— $\beta = c(\beta_{0g}, \beta_{lg})$: vecteur des effets fixes spécifiques à chaque classe ;

— $B_l(t)$: base de fonction de splines cubiques du temps avant la date index avec 3 noeuds intérieurs placés aux quartiles (12, 24, et 36 ans) ;

— u_{0ig} : intercept aléatoire spécifique à chaque classe supposant que $u_{0ig} \sim \mathcal{N}(0, \omega_g^2\sigma_u^2)$ où σ_u^2 est une variance commune et ω_g un coefficient de proportionnalité, ce qui permet une variabilité spécifique à chaque classe ;

— $\mathbb{1}_{t_{ij}\in Hist_i}$ est la variable indicatrice qui vaut 1 si t_{ij} se situe durant l'histoire d'exposition du sujet i, 0 sinon.

Sous-modèle 3 :

Comme pour le tabac, nous avons ajusté sur les facteurs d'appariement. Pour ajuster sur la consommation de tabac, nous avons utilisé le CSI précédemment détaillé dans le chapitre 2 section 2.3.2.2.b (Leffondré et al. [2006]). Rappelons la formule de cet indice agrégé :

$$CSI = (1 - 0.5^{dur^\star/\tau})(0.5^{tsc^\star/\tau})ln(int + 1) \tag{45}$$

où

— dur est la durée totale de tabagisme en années ;
— tsc est le temps depuis l'arrêt en années ;
— int le nombre moyen de cigarettes fumées par jour vie entière en cig-années ;
— $tsc^\star$=max(tsc-δ,0) ;
— $dur^\star$=max(dur+tsc - δ,0)-$tsc^\star$;
— τ est un paramètre de forme. Il est égal à 26 ans (basé sur les estimations réalisées dans Papadopoulos et al. [2014] sur les données provenant de l'étude ICARE) ;
— δ est un paramètre de décalage. Il est égal à 1 an (basé sur les estimations réalisées dans Papadopoulos et al. [2014] sur les données provenant de l'étude ICARE).

Le sous-modèle 3 s'écrit :

$$\text{logit}(P(D_i = 1|Age, Dpart, CSI, c_i = g)) = \gamma_{0g} + f_1(Age_i) + f_2(Dpart_i) + \gamma_3 CSI_i \quad (46)$$

où

— D_i : variable binaire qui correspond au statut cas-témoin (=1 pour cas, 0 pour témoin) ;

— γ_{0g} : intercept spécifique à la classe latente g ;

— $f_1(Age)$: splines naturelles cubiques de l'âge à la date index avec 3 noeuds placés aux 5ème, 50ème et 95ème percentiles de la distribution pour tenir compte de l'effet non linéaire de l'âge sur le logit (45, 61, 73 ans) ;

— $f_2(Dpart)$: combinaison linéaire des 9 variables indicatrices des départements de résidence (département pris en référence : le Bas-Rhin) ;

— CSI_i : variable représentant le Comprehensive Smoking Index.

3.3.3 Écriture de la vraisemblance conjointe individuelle

Dans la section concernant l'estimation du modèle conjoint, nous avons défini la vraisemblance conjointe pour un modèle linéaire mixte à classes latentes avec un modèle logistique (équation 36). Dans le cadre des deux applications au tabac et à l'amiante, nous avons utilisé un modèle mixte à processus latent spécifique à chaque classe latente, ce qui a modifié l'écriture de la vraisemblance conjointe pour la partie de ce modèle mixte en question.

Pour faciliter la visualisation de la contribution de chacun des 3 sous-modèles à la vraisemblance conjointe individuelle, nous avons utilisé un code couleur dans l'équation suivante :

$$\text{lv}_i(\theta_G) = \sum_{g=1}^{G} \pi_{ig} \phi_{ig}(\tilde{Y}_i|c_i = g; \theta_G) \prod_{j=1}^{n_i} J(H^{-1}(Y_{ij})) P(D_i = 1|c_i = g, X_{ri})^{D_i} P(D_i = 0|c_i = g, X_{ri})^{1-D_i} \quad (47)$$

où

— La partie en orange concerne le modèle d'appartenance aux classes défini par l'équation 20 donnant l'expression de $P(c_i = g|X_{pi})$;

— La partie en rouge concerne le modèle mixte à processus latent spécifique à chaque classe latente g qui fait intervenir la fonction de densité gaussienne sur le processus latent bruité $(\tilde{Y}_{ij})$ au temps t_{ij} et la matrice jacobienne (J) de la fonction de lien inverse H^{-1} (ici la matrice des dérivées des I-splines quadratiques). Les estimations des paramètres du

modèle mixte sont obtenues dans l'échelle originale des mesures répétées Y_{ij} pour garder l'interprétation possible des valeurs numériques associées aux évolutions longitudinales ;

— La partie en bleu concerne le modèle de régression logistique (équation 33) où D_i est la variable binaire représenté par le statut cas-témoin.

3.3.4 Aspect logiciel : appel de la fonction Jointlcmm du package R *lcmm*

Pour l'estimation d'un modèle conjoint avec un modèle logistique, nous avons donc adapté le code de la fonction Jointlcmm du package R *lcmm* (Proust-Lima et al. [2017]). En effet, dans le package R *lcmm*, le modèle conjoint est programmé avec un modèle à risques instantanés proportionnels car il est principalement utilisé pour étudier le délai jusqu'à un évènement clinique et non une variable binaire tel que le statut cas-témoin.
Les principaux changements ont concerné, entre autres, la vraisemblance conjointe avec la vraisemblance d'un modèle de régression logistique. Ainsi que le calcul des probabilités a posteriori conditionnelles aux données longitudinales et à l'évènement.

Voici un exemple de l'appel de la fonction Jointlcmm pour estimer un modèle conjoint avec une partie logistique. L'écriture de ce modèle est celui qui a été défini pour le modèle conjoint concernant les trajectoires d'intensités de consommation de tabac (équations 38 - 41).

```
Jointlcmm(fixed = doseTab_10 ~ ns(t_TSI, knots = c(10, 21, 33),
 Boundary = c(0, 64)),
 mixture = ~ns(t_TSI, knots = c(10,
 21, 33), Boundary = c(0, 64)),
 random = ~-1 + Ind_IntEa,
 subject = "numid",
 ng = 4,
 idiag = 1,
 nwg = TRUE,
 survival = kt ~
  ns(AgeIndexDate, knots = c(43, 60, 73), Boundary = c(23,76))
  + depthab2 + ICE_Cat2 + ICE_Cat3 + ICE_Cat4,
 hazard = "Logistic",
 link = "3-manual-splines",
 intnodes = c(2),
 data = BaseDupl_Tab,
```

```
  maxiter = 50,
  B=Val_Init)
```

où les variables sont :

— doseTab_10 : le nombre moyen de cigarettes fumées par jour sur l'année ;

— t_TSI : le temps avant la date index ;

— Ind_IntEa : l'intercept aléatoire ;

— numid : le nom de la variable correspondant aux identifiants des sujets ;

— kt : le statut cas-témoin ;

— AgeIndexDate : l'âge à la date index ;

— depathab2 : les départements de résidence (en type "factor") ;

— ICE_Cat2, ICE_Cat3, ICE_Cat4 : les indicatrices correspondantes aux 3 catégories de l'indice cumulé d'exposition classique de l'exposition professionnelle à l'amiante.

et les arguments (en bleu, ceux qui ont été modifiés pour remplacer le modèle de survie par un modèle de régression logistique) sont :

— **fixed** : à gauche du $\sim$ la variable des mesures répétées et à droite la formule des effets fixes (communs et spécifiques aux classes latentes).

— **mixture** : les effets fixes qui sont spécifiques à chaque classe.
Dans le cadre de nos deux applications, les effets fixes sont tous spécifiques. Ils sont définis par une base de splines naturelles cubiques (fonction R *ns*) pour estimer une trajectoire du temps flexible plutôt que polynomial ;

— **random** : la formule définissant les variables associées aux effets aléatoires.
Ici, on a -1 pour dire qu'on n'estime pas un intercept "classique" mais on va utiliser notre variable (Ind_IntEa) qui été définie plus haut par l'indicatrice $\mathbb{1}_{t_{ij} \in Hist_i}$ pour chaque individu i en chaque temps t_{ij} ;

— **subject** : le nom de la variable définissant l'identifiant du sujet dans la base de données ;

— **ng** : le nombre de classes latentes ;

— **idiag** : une variable binaire qui définit la matrice de variance-covariance des effets aléatoires. La variable vaut 1 si on considère que les effets sont indépendants (matrice diagonale) sinon 0 quand les effets sont considérés corrélés (matrice non-structurée) ;

— **nwg** : une variable logique déterminant si la matrice de variance-covariance des effets aléatoires est spécifique à chaque classe (TRUE) ou commune sur les classes latentes (FALSE). Les coefficients de proportionnalité ω_g seront estimés quand nwg=TRUE. Pour assurer l'identifiabilité du modèle, ω_G=1 ;

— **survival** : la formule intégrant à gauche du $\sim$ la variable binaire (statut cas-témoin) et à droite les variables d'ajustement considérées pour la régression logistique ;

— **hazard** : une variable définissant les familles des fonctions de risque des modèles de survie (Weibull, Splines, ...). Ici, on a ajouté la valeur "Logistic" qui permet de préciser que le sous-modèle pour l'évènement n'est plus un modèle de survie mais un modèle logistique ;

— **link** : une variable qui définit la famille de la fonction de lien paramétrique à considérer ("linear","beta","thresholds","splines").
Pour nos applications, nous avons utilisé les splines où il est nécessaire de préciser le nombre de noeuds et leur position (manual, equi, quant) comme suit "3-manual splines" ;

— **intnodes** : le vecteur des positions des noeuds intérieurs. Cet argument est seulement requis quand link="splines" ;

— **data** : le nom de la base de données ;

— **maxiter** : le nombre maximum d'itérations pour l'algorithme itératif de Marquardt ;

— **B** : le vecteur des valeurs initiales de tous les paramètres des trois sous-modèles. Il doit être écrit en considérant l'ordre dans lequel est attendu les différents paramètres des sous-modèles. L'aide sous R est très bien expliquée pour le créer correctement.

3.4 Article 3 soumis à *International Journal of Epidemiology* (IJE)

Association between lung cancer and lifetime profiles of intensity of exposure to occupational asbestos and smoking: Results from the ICARE case-control study.

Emilie Lévêque*, Aude Lacourt, Viviane Philipps, Danièle Luce, Pascal Guénel, Isabelle Stücker, Cécile Proust-Lima and Karen Leffondré

Key messages : (up to 5 succinct points)

- Our results provide an illustration of the importance of the timing of exposure intensity to smoking and asbestos on the risk of lung cancer
- Recent smoking intensities seem to play a substantial role, as opposed to asbestos for which distant intensities have stronger contribution to the risk
- Joint latent class mixed model constitutes an interesting tool to identify different patterns of lifetime exposure trajectory and associated risk of disease onset

Abstract

Objective: To identify lifetime trajectories of intensity of smoking and occupational exposure to asbestos in the ICARE case-control study and to compare their risk of lung cancer in males.

Methods: Incident lung cancer cases were recruited in 2001-2007 in 10 French territorial *départements. C*ontrols were selected via incidence density sampling within the general population of the same *département* as the case with matching on age and sex. Smoking and occupational history were reported during face-to-face interviews. Exposure to asbestos was assessed using a Job Exposure Matrix. Longitudinal profiles of intensity of smoking and asbestos exposure were identified and compared using a joint latent class mixed model in ever smokers (1938/1837 cases/controls) and ever occupationally exposed to asbestos (912/798 cases/controls).

Results: Four classes of trajectories were identified for smoking and asbestos: a class with low to moderate constant intensity over lifetime (reference class, 52.4% of subjects for smoking, 44.4% for asbestos), a class with recent high intensity (22.5%, 14.9%), a class with long term high intensity (12.9%, 23.5%), and a class with distant high intensity (12.2%, 17.3%). Strongest risk of lung cancer were found in classes characterized by recent high intensity for smoking and distant high intensity for asbestos. Classes with distant high smoking intensity or recent high asbestos intensity did not have a stronger risk of lung cancer compared to the reference class, despite higher cumulative doses of exposure.

Conclusions: Our results illustrate the importance of the timing of exposure in dose-response relationship between smoking or asbestos and lung cancer.

INTRODUCTION

Occupational and environmental exposures often extend over lifetime. Their intensity generally varies over time and one of the main methodological challenges in epidemiologic cancer studies is to capture this temporal variation and to assess its relationship with the risk of cancer(1) (2) (3, 4). While the association between lung cancer and smoking or occupational exposure to asbestos have been extensively investigated in the literature, few studies have specifically investigated the dynamic aspects of exposure intensity over lifetime. Some previous studies have attempted to study the impact on lung cancer of intensity of smoking and occupational exposure to asbestos at each year of the exposure history. For example, we found that recent intensity of smoking had a stronger contribution than distant intensity of smoking on the risk of lung cancer and the opposite for asbestos(5) using a weighted cumulative index of exposure(6) . While such an approach allows the comparison of the risk of lung cancer associated with different hypothetical profiles of exposure over lifetime(7), it does not allow the identification of different types of longitudinal profiles of exposure intensities in the actual data.

Several statistical methods such as Latent Class Mixed Model (LCMM) have been developed to identify trajectories of individual longitudinal quantitative indicators(8, 9). These models have been largely used to describe heterogeneous evolution of a quantitative biomarker over time (10) (11), psychometric tests(12) or delinquency behavior(13) but rarely for identifying longitudinal profiles of environmental or occupational exposures. They have been used to identify trajectories of smoking (14) (15, 16) but never to explore their association with the risk of lung cancer. Yet, such an approach could give new insight on the dose-response relationship between smoking or occupational exposure to asbestos and the risk of lung cancer.

The objective of the present study was to identify profiles of trajectories of the intensity of occupational exposure to asbestos and smoking over lifetime and quantify their association with the risk of lung cancer, using data from a multicentric population-based case-control study.

Material & Methods

Study design

The ICARE study is a large French case-control study of respiratory cancers(17). Briefly, all histologically confirmed primary malignant lung or upper aerodigestive tract incident cancer cases, aged 18-75 years and living in 10 French *départements* were recruited from French cancer registries in 2001-2007. Controls were selected within the general population using a random digit dialing procedure through incidence density sampling and were frequency-matched to cases by sex, age and *départements*. Additional stratification has been made on socioeconomic status within each sex-age-*départements* group to represent the general population(17). The present analysis was restricted to males only.

Data collection

Subjects were face-to-face interviewed by trained interviewers with a detailed standardized questionnaire to collect information on sociodemographic characteristics, lifetime occupational history (all jobs held for at least 1 month) and lifetime smoking history.

For each smoking episode, start and end years, number of cigarettes smoked per day were reported. For each job, information related to start and end years, industrial activity further coded using the the French Nomenclature Activities (NAF) and occupation further coded using the International Standard Classification of Occupations 1968 (ISCO) were recorded. The present analysis was restricted to males who had complete history on smoking and occupational history.

Occupational asbestos exposure assessment

Occupational exposure to asbestos was assessed by a job exposure matrix (JEM) which accounted for asbestos exposure levels changes in France over calendar time(7, 18). For each job defined as a combination of an ISCO and NAF code, the JEM assigned 1) a probability of exposure defined as the proportion of exposed workers for that job (from 0 to 0.85) ; 2) a frequency of exposure defined as

the proportion of exposed working time on a typical 8h working day for that job (from 0.025 to 0.85) and 3) an intensity of exposure defined as the equivalent annual average daily concentration of asbestos fibers in the air at workplace for that job (from 0.0005 to 20 equivalent fibers per mL, Table 1S in Supplementary Material). By linking the JEM with individual lifetime occupational history, annual levels of exposure were defined as the product of intensity, probability, and frequency of exposure and expressed in equivalent fibres per mL. The mean level of exposure over the year was prorated by the duration of each job occupied within that year. In the present analysis, we referred to this level of exposure in a given year, as the "annual average daily intensity of occupational exposure to asbestos". The cumulative index of exposure (CIE) to asbestos over lifetime was the sum of these annual intensities over job history.

Statistical analysis

We used two distinct joint latent class mixed models (JLCMM) both aiming at identifying profiles of trajectories of intensity over lifetime and quantify the associated risk of lung cancer at the index date. The first model identified the trajectories of intensity of smoking (in cig/day) over lifetime. To this aim, we restricted the analysis to ever smokers only, i.e cases and controls who had at least one non-null intensity over lifetime. The second model identified the trajectories of intensity of occupational exposure to asbestos (in equivalent f/mL) over lifetime. To this aim, we restricted the analysis to cases and controls who had a CIE to asbestos over lifetime higher than 0.26 f/mL-years. This cut-off of 0.26 f/mL-years corresponded to the limit value of occupational exposure on 8h working day fixed by the French law (0.01 f/mL) multiplied by the mean duration exposure of cases and controls (26 years). We excluded subjects below this cut-off in order to solve convergence issues of the JLCMM. In each model, the time axis was the time before the index date (diagnosis for case, interview for control).

The JLCMM was made of three sub-models (equations of the three sub-models are provided in the supplementary material but are not necessary to understand what follows), one for the latent class membership, one for the trajectory of repeated measures over time and one for the risk of event(19). In the present analysis, the repeated measures were the number of cigarettes smoked per day on average in each year, and the annual average daily intensity of occupational exposure to asbestos in each year. The event was the case/control status at the index date.

JLCMM assumes that the population is constituted of G non-observed subgroups of subjects, called latent classes, with different trajectory profiles and risks of event. Each subject belongs to only one latent class. Sub-model 1 was thus a multinomial logistic regression model that estimated for each subject his probability to belong to each latent class included only class-specific intercepts.

Sub-model 2 was a class-specific mixed model that estimated the mean trajectory of intensities over lifetime in each class. To account for non-normal distribution of both smoking and occupational asbestos annual intensities (Figure 1S in Supplementary Material), we used a mixed model which transforms the annual intensities using splines (12) and simultaneously models the transformed intensities using a linear mixed model(20). More specifically, we transformed the data using a I-Splines function with 3 knots: at 0, 20, 100 cig/day for smoking and at 0, 0.1, 12.6 f/mL for asbestos. This fitted the best our data according to Akaike Information Criteria (AIC). To account for the discrete changes of intensities of both smoking and asbestos exposure intensity over time (Figure 2S in Supplementary Material) and allow for a flexible modelling of individual trajectories we used splines functions of time. More specifically, we used natural cubic splines with 3 inner knots, placed at quartiles of time before the index date: 10, 21, 33 years for smoking and 12, 24, 36 years for asbestos. Individual departures from the mean trajectory in each class were modelled via a random effect on the intercept with a class-specific variance. For each subject, the random effect applied on his entire time window of exposure but not outside when the dose of exposure equaled zero.

Sub-model 3 was a class specific logistic regression model that estimated the probability of lung cancer in each class conditionally on matching factors (age and *départements*) and potential confounders (smoking for asbestos, and asbestos for smoking). More specifically, association between profiles of asbestos and lung cancer risk was adjusted for smoking via the Comprehensive Smoking Index (CSI), which is a single aggregate measure of smoking history accounting for average intensity over lifetime, total duration of smoking and time since cessation at the index date(21). Association between profiles of smoking and lung cancer risk was adjusted for occupational exposure to asbestos using the quartiles of CIE.

JLCMM were estimated using an extension of the lcmm R package(19) to handle binary endpoints. To select the optimal number of latent classes, JLCMM with different number of classes (one to six) were estimated and compared in terms of i) quality of adjustment (Bayesian Information Criteria (BIC), AIC, plot of residuals), ii) the relevance of identified trajectories and iii) the discriminatory performance of the model based on the posterior classification table (see Supplementary Materials). For each model (with one to six latent classes), a grid of 20 initial values were tested to prevent any convergence toward a local maximum. The authors followed the GRoLTS-Checklist(22) .

Results

A total of 1969 male cases and 1838 male controls were ever smokers and thus included in the analysis which aimed at identifying smoking trajectories (Figure 3S in supplementary material). For occupational exposure to asbestos, 912 male cases and 798 male controls had a CIE to asbestos higher than 0.26 f/mL-years (Figure 3S in Supplementary Material), they had similar socio demographics characteristics than all the subjects ever exposed to asbestos (Table S4 in Supplementary Material).

[Table 1 here]

For occupational exposure to asbestos, cases and controls had similar total duration of exposure, time since first exposure, age at first exposure or time since last exposure (Table 1). The average annual daily intensity of occupational exposure was however higher in cases than in controls (median of 0.24 versus 0.15 f/mL). For smoking, cases smoked for a longer duration than controls (median of 39 versus 26 years), stopped smoking more recently (median of 3 versus 16 years), and had a stronger average intensity over lifetime (median of 20 versus 15 cig/day) (Table 1).

[Table 2 here]

For smoking, the best model had four latent classes (Table 3S in Supplementary Material). Class 1 (52.4% of subjects) had a constant mean trajectory of smoking intensity over lifetime, at about 8 cig/day (reference Class): “constant moderate intensity” (Figure 1). Class 2 (22.5% of subjects) had in mean high intensities of smoking reaching 15 cig/day within the 20 years before index date: “recent high intensity”. Class 3 (12.9% of subjects) had rather very high intensities from 40 to 10 years before the index date at around 25 cig/day: “long-term very high intensity”. Class 4 (12.2% of subjects) had a very high-intensity episode at in mean about 22 cig/day occurring around 40 years before index date: “distant very high intensity”.

Class 2 with recent high intensity of smoking had the strongest risk of lung cancer (OR=4.89, 95% CI: 4.01; 5.96, compared to the reference Class, Table 2). Despite subjects a posteriori classified in Class 2 were younger than in all other classes (median age of 53 years at the index date), they had a high total number of cigarette-years (median of 805), a long duration of smoking (median of 37 years), and thus the lowest proportion of ex-smokers (51.5%) and the shortest time since smoking cessation (median of 1 year). Class 3 with long term very high intensity had a significant increased risk of lung cancer compared to the reference Class (OR=2.60, 95% CI: 2.02; 3.31) but did not have a stronger risk of lung cancer compared to class 2, despite Class 3 had the highest median total number of cigarette-years over lifetime (907) and the longest duration of smoking (median of 46 years). This is likely because the high intensities in Class 3 were mainly accumulated on average more than 20 years

before the index date compared to Class 2 (Figure 1), and subjects stopped smoking for a longer duration (median of 6 years versus 1 year). Despite Class 4 had a higher median total number of cigarette-years compared to the reference Class (512 vs 350), it had a similar risk of lung cancer (OR=0.91, 95% CI: 0.71; 1.17), likely because it had a higher proportion of ex-smokers (99.1% versus 94.1%) who stopped smoking for a much longer duration (median of 25 vs 10 years).

For occupational exposure to asbestos, the best model included also four latent classes (Table 4S in Supplementary Material). Class 1 (44.4% of subjects) had a constant mean trajectory of exposure intensity at a level on average at 0.04 f/mL each year of occupational history (reference Class): "constant low intensity" (Figure 2). Class 2 (14.9% of subjects) had a high-intensity episode at 0.4 f/mL from 13 to 19 years before the index date: "recent moderate intensity". Class 3 (23.5% of subjects) had a high-intensity episode which reached a maximum at 1f/mL around 20 years before the index date: "20 year very high intensity". Class 4 (17.3% of subjects) had a high-intensity episode at around 0.6 f/mL from 42 to 29 years before index date: "distant high intensity".

[Table 3 here]

As expected, Class 4 with the most distant high-intensity episode of asbestos was made of older subjects at the index date (median of 65 years, Table 3). It had the strongest risk of lung cancer compared to the reference Class, without reaching statistical significance (OR=1.26, 95% CI: 0.9; 1.79). Despite the strongest cumulative exposure in Class 3 (13.5 f/mL-years versus 12.2 f/mL-years in Class 4), Class 3 did not have the strongest risk of lung cancer (OR=1.17, 95% CI: 0.85; 1.79 compared to the reference class), likely because the high-intensity episode occurred more recently than in Class 4 (around 25 years before the index date versus 35 years, Figure 2). Despite a much higher median cumulative dose of exposure in Class 2 than in the reference Class (5.9 f/mL-years versus 1.4 f/mL-years), Class 2 had a similar risk of lung cancer (OR=1.03, 95% CI: 0.71; 1.50), likely because doses were accumulated much more recently (median time since first exposure of 31 years versus 44 years).

Discussion

For smoking, we identified four latent classes with different patterns of intensity over lifetime. The class with heavy current smokers had by far the strongest risk of lung cancer (Class 2). The Class of heavy ex-smokers (Class 4) who stopped for the longest duration did not have a stronger risk of lung cancer than the reference class of constant moderate intensity, although they had accumulated a much higher number of cigarette-years over lifetime. For lifetime occupational exposure to asbestos, we also identified four latent classes with distinct patterns of intensity over time. Only the class with the most distant high intensity (Class 4) tended to have a marginally significant higher risk of lung cancer. These results illustrate the importance of the timing of exposure with an apparent stronger weight of recent intensity for smoking and distant intensity for asbestos.

All these results for both exposures are consistent with our previous study(5). In this study, we estimated the relative weight of intensity of smoking and occupational exposure to asbestos in each year of exposure history. We used the same case-control data but a totally different statistical method which did not allow us to identify the different classes of exposure trajectories and compare their associated risk of lung cancer. For smoking, the important role of recent exposures has been reported in other studies (23, 24) . The decrease in the risk of lung cancer after smoking cessation is also consistent with the literature(25, 26). For occupational exposure to asbestos, previous studies have also suggested that the distant intensities have a stronger contribution to the risk of lung cancer than recent intensities(27) (28) . However, it would be of interest to confirm our results using the same analytical approach on other case-controls studies on lung cancer. It should be noted that the identified classes of exposure in our case-control study are not expected to represent the classes of exposure in the general French population, or in other general population, just because of over representation of cases, as in all case-control studies. However, our results on the association between the classes of trajectories and the risk of lung cancer should be reproducible in other populations.

Our study has limitations. First, we used the approximation of an annual average daily intensity of exposure for both smoking and occupational asbestos from data retrospectively reported, even if it was reported in a standardized questionnaire face-to-face administered by trained interviewers. For each period of smoking consumption, the mean number of cigarettes smoked per day over that period has been reported by subjects. It may be reasonable to consider that intensity was approximately constant within each of those reported periods. It has indeed been shown that self-reported smoking histories tend to be reliable(29) (30). For asbestos, the complete occupational histories of subjects have also been reported by subjects and several studies have shown that self-reported occupational histories tend also to be valid(31). To infer occupational asbestos exposure from reported job histories, we applied a job-exposure matrix which is a method known to produce non-differential misclassification(32). However, we acknowledge that the average intensity daily derived from this matrix is a crude estimation of the actual intensities. Further studies are needed to investigate the impact of measurement errors on the results of JCLMM. Further methodological developments are also needed to improve the modelling of the skewed distributions of exposures in the JLCMM. Indeed, many subjects may be exposed to low doses of exposure during their entire lifetime, while few are exposed to very high doses. The intensity distributions may also include a pike at zero due to non-exposed periods. In the present study, we used a spline transformation of the intensity of exposure in each year to account for such non Gaussian distributions(19). However, this was not sufficient for asbestos, thus we had to exclude subjects who were exposed to very low doses of asbestos. We believe that this exclusion did not bias our estimations since excluded subjects had similar socio demographics characteristics as included subjects (Table 4S in Supplementary Material). The discriminatory power of both models is extremely high with mean posterior probabilities above 95% in each class.

In conclusion, we believe that our study provides a new illustration of the importance of the timing of exposure in the dose-response relationship between smoking, asbestos and the risk of lung cancer, as well as how to investigate trajectories of exposures over lifetime. Nevertheless, future

methodological developments are still needed to accurately handle all trajectories of environmental and occupational exposures in cancer etiology.

References

1. Thomas D. Statistical Methods in Environmental Epidemiology Oxkord; 2009.
2. Vacek PM. Assessing the effect of intensity when exposure varies over time. Statistics in medicine. 1997;16(5):505-13.
3. Richardson DB. Latency models for analyses of protracted exposures. Epidemiology. 2009;20(3):395-9.
4. Vermeulen R, Chadeau-Hyam M. Dynamic aspects of exposure history-do they matter? Epidemiology. 2012;23(6):900-1.
5. Lévêque E, Lacourt A, Luce D, Sylvestre M-P, Guénel P, Stücker I, et al. Time-dependent effect of intensity of smoking and of occupational exposure to asbestos on the risk of lung cancer: results from the ICARE case–control study. Occupational and environmental medicine. 2018;75(8):586.
6. Hauptmann M, Wellmann J, Lubin JH, Rosenberg PS, Kreienbrock L. Analysis of exposure-time-response relationships using a spline weight function. Biometrics. 2000;56(4):1105-8.
7. Lacourt A, Lévêque E, Guichard E, Gilg Soit Ilg A, Sylvestre M-P, Leffondré K. Dose-time-response association between occupational asbestos exposure and pleural mesothelioma. Occupational and environmental medicine. 2017;74(9):691.
8. Berlin KS, Parra GR, Williams NA. An Introduction to Latent Variable Mixture Modeling (Part 2): Longitudinal Latent Class Growth Analysis and Growth Mixture Models. Journal of Pediatric Psychology. 2014;39(2):188-203.
9. Proust-Lima C, Séne M, Taylor JMG, Jacqmin-Gadda H. Joint latent class models for longitudinal and time-to-event data: A review. Statistical methods in medical research. 2014;23(1):74-90.
10. Lin H, Turnbull BW, McCulloch CE, Slate EH. Latent Class Models for Joint Analysis of Longitudinal Biomarker and Event Process Data. Journal of the American Statistical Association. 2002;97(457):53-65.
11. Boucquemont J, Loubère L, Metzger M, Combe C, Stengel B, Leffondre K. Identifying subgroups of renal function trajectories. Nephrology Dialysis Transplantation. 2017;32(suppl_2):ii185-ii93.
12. Proust C, Jacqmin-Gadda H, Taylor JM, Ganiayre J, Commenges D. A nonlinear model with latent process for cognitive evolution using multivariate longitudinal data. Biometrics. 2006;62(4):1014-24.
13. Nagin Daniel S, Tremblay Richard E. DEVELOPMENTAL TRAJECTORY GROUPS: FACT OR A USEFUL STATISTICAL FICTION?*. Criminology. 2005;43(4):873-904.
14. Weden MM, Miles JNV. Intergenerational Relationships Between the Smoking Patterns of a Population-Representative Sample of US Mothers and the Smoking Trajectories of Their Children. American Journal of Public Health. 2012;102(4):723-31.
15. De Genna NM, Goldschmidt L, Day NL, Cornelius MD. Maternal Trajectories of Cigarette Use as a Function of Maternal Age and Race. Addictive behaviors. 2017;65:33-9.
16. Brook DW, Brook JS, Zhang C, Whiteman M, Cohen P, Finch SJ. Developmental Trajectories of Cigarette Smoking from Adolescence to the Early Thirties: Personality and Behavioral Risk Factors. Nicotine & tobacco research : official journal of the Society for Research on Nicotine and Tobacco. 2008;10(8):1283-91.
17. Luce D, Stucker I, Group IS. Investigation of occupational and environmental causes of respiratory cancers (ICARE): a multicenter, population-based case-control study in France. BMC public health. 2011;11:928.
18. Févotte J, Dananché B, Delabre L, Ducamp S, Garras L, Houot M, et al. Matgéné: A Program to Develop Job-Exposure Matrices in the General Population in France. The Annals of occupational hygiene. 2011;55(8):865-78.

19. Proust-Lima C, Philipps V, Liquet B. Estimation of Extended Mixed Models Using Latent Classes and Latent Processes: The R Package lcmm. Journal of Statistical Software; Vol 1, Issue 2 (2017). 2017.
20. Proust-Lima C, Amieva H, Jacqmin-Gadda H. Analysis of multivariate mixed longitudinal data: A flexible latent process approach. British Journal of Mathematical and Statistical Psychology. 2012;66(3):470-87.
21. Leffondre K, Abrahamowicz M, Xiao Y, Siemiatycki J. Modelling smoking history using a comprehensive smoking index: application to lung cancer. Statistics in medicine. 2006;25(24):4132-46.
22. van de Schoot R, Sijbrandij M, Winter SD, Depaoli S, Vermunt JK. The GRoLTS-Checklist: Guidelines for Reporting on Latent Trajectory Studies. Structural Equation Modeling: A Multidisciplinary Journal. 2017;24(3):451-67.
23. Hazelton WD, Clements MS, Moolgavkar SH. Multistage Carcinogenesis and Lung Cancer Mortality in Three Cohorts. Cancer Epidemiology Biomarkers & Prevention. 2005;14(5):1171.
24. Hauptmann M, Pohlabeln H, Lubin JH, Jockel KH, Ahrens W, Bruske-Hohlfeld I, et al. The exposure-time-response relationship between occupational asbestos exposure and lung cancer in two German case-control studies. American journal of industrial medicine. 2002;41(2):89-97.
25. Peto R, Darby S, Deo H, Silcocks P, Whitley E, Doll R. Smoking, smoking cessation, and lung cancer in the UK since 1950: combination of national statistics with two case-control studies. BMJ. 2000;321(7257):323.
26. Vlaanderen J, Portengen L, Schuz J, Olsson A, Pesch B, Kendzia B, et al. Effect modification of the association of cumulative exposure and cancer risk by intensity of exposure and time since exposure cessation: a flexible method applied to cigarette smoking and lung cancer in the SYNERGY Study. American journal of epidemiology. 2014;179(3):290-8.
27. Hillerdal G, Henderson DW. Asbestos, asbestosis, pleural plaques and lung cancer. Scandinavian Journal of Work, Environment & Health. 1997;23(2):93-103.
28. Selikoff IJ HE, Seidman H. Latency of asbestos disease among insulation workers in the United States and Canada. . Cancer causes & control : CCC. 1980(46):2736-40.
29. Soulakova JN, Hartman AM, Liu B, Willis GB, Augustine S. Reliability of Adult Self-Reported Smoking History: Data from the Tobacco Use Supplement to the Current Population Survey 2002–2003 Cohort. Nicotine & Tobacco Research. 2012;14(8):952-60.
30. Huerta M, Chodick G, Balicer RD, Davidovitch N, Grotto I. Reliability of self-reported smoking history and age at initial tobacco use. Preventive Medicine. 2005;41(2):646-50.
31. Teschke K, Olshan A, Daniels J, De Roos AJ, Parks C, Schulz M, et al. Occupational exposure assessment in case–control studies: opportunities for improvement. Occupational and environmental medicine. 2002;59(9):575-94.
32. Bouyer J, Dardenne J, xe, mon D. Performance of odds ratios obtained with a job-exposure matrix and individual exposure assessment with special reference to misclassification errors. Scandinavian Journal of Work, Environment & Health. 1995;21(4):265-71.

Table 1 Characteristics of ever-smokers and ever occupational exposed to asbestos from ICARE, case control study, 2001-2007

	Smoking		Occupational exposure to asbestos	
	Ever smoking cases (n=1969)	Ever smoking controls (n=1838)	Ever exposed[+] cases (n=912)	Ever exposed[+] controls (n=798)
Age at index date (in years) Median (5^{th} – 95^{th} percentile)	61 (46-74)	59 (41-73)	60.0 (45-73)	59.0 (41-73)
Area of residence (*département*) (n, %)				
Calvados	234 (11.9)	260 (14.1)	115 (12.6)	109 (13.7)
Doubs - Territoire de Belfort	102 (5.2)	66 (3.6)	39 (4.3)	36 (4.5)
Hérault	220 (11.2)	243 (13.2)	99 (10.9)	94 (11.8)
Isère	330 (16.8)	258 (14.0)	149 (16.3)	107 (13.4)
Loire Atlantique	250 (12.7)	205 (11.2)	129 (14.1)	102 (12.8)
Manche	219 (11.1)	149 (8.1)	123 (13.5)	65 (8.1)
Bas-Rhin	240 (12.2)	240 (13.1)	111 (12.2)	115 (14.4)
Haut-Rhin	50 (2.5)	59 (3.2)	26 (2.9)	29 (3.6)
Somme	222 (11.3)	266 (14.5)	76 (8.3)	90 (11.3)
Vendée	102 (5.2)	92 (5.0)	45 (4.9)	51 (6.4)
Total duration of exposure (in years) Median (5^{th} – 95^{th} percentile)	39 (18-54)	26 (4-48)	34 (6-47)	32 (5-46)
Age at first exposure (in years) Median (5^{th} – 95^{th} percentile)	17 (13-22)	17 (13-24)	16 (14-27)	16 (14-25)
Exposed status (n, %)				
Current exposed	363 (18.4)	203 (11.0)	158 (17.3)	150 (18.8)
Ex-exposed	1606 (81.6)	1635 (89.0)	754 (82.7)	648 (81.2)
Time since last exposure[#] (in years) Median (5^{th} – 95^{th} percentile)	3 (1-29)	16 (1-40)	12 (1-43)	11 (1-43)
Time since first exposure (in years) Median (5^{th} – 95^{th} percentile)	44 (28-57)	41 (23-55)	43 (25-57)	42 (20-57)
Cumulative index of exposure (cig/years or f/mL-years)				

Median (5^{th} – 95^{th} percentile)	772.5 (195.0-1752.4)	330.2 (9.9-1047)	5.1 (0.3-51.1)	3.3 (0.3-42.5)
Average intensity over lifetime (cig/day or f/mL)				
Median (5^{th} – 95^{th} percentile)	20.0 (7.7 -40)	14.9 (1.4-33)	0.24 (0.01-1.68)	0.15 (0.01-1.28)

SD, standard deviation; Med, median; IQR, interquartile range,[#] in ex exposed

[t]Ever exposed to occupational exposure who had a cumulative index of exposure over lifetime higher than 0.26 f/mL-years

Table 2 Association between trajectories of smoking intensity and lung cancer, ICARE case-control study, 2001-2007, France.

Trajectory of smoking intensity	Cases/Controls a posteriori classified in the class	Age at index date (years)	Cigarettes-years	Total duration of smoking (years)	Average intensity over lifetime (cig/day)	Ex-smokers	Time since smoking cessation in ex-smokers (years)	Age at initiation (years)	
	n	median (5^{th}-95^{th} percentile)	median (5^{th}-95^{th} percentile)	median (5^{th}-95^{th} percentile)	median (5^{th}-95^{th} percentile)	n (%)	median (5^{th}-95^{th} percentile)	median (5^{th}-95^{th} percentile)	OR[±] (95%CI)
Class 1 « constant moderate intensity »	799/1195	58 (42-72)	350 (10-1047)	27 (4-51)	15 (1-31)	1879 (94.1%)	10 (1-40)	18 (14-25)	1.00
Class 2 « recent high intensity »	614/247	53 (41-66)	805 (280-1810)	37 (24-49)	21 (10-45)	442 (51.5%)	1 (1-8)	16 (12-22)	4.89 (4.01 ; 5.96)
Class 3 « long-term very high intensity »	343/147	68 (61-74)	907 (391-1942)	46 (33-55)	20 (10-40)	462 (94.3%)	6 (1-16)	16 (12-21)	2.60 (2.02 ; 3.31)
Class 4 « distant very high intensity »	213/249	70 (62-75)	512 (215-1200)	28 (16-45)	19 (9-40)	459 (99.1%)	25 (13-36)	17 (13-21)	0.91 (0.71 ; 1.17)

OR, odds ratio; CI, confidence interval

[±] Adjusted for age at the index date (in years, spline), area of residence (*département*) and occupational exposure to asbestos (quartiles of the standard cumulative index of occupational exposure to asbestos).

Table 3 Association between trajectories of occupational exposure to asbestos intensity and lung cancer, ICARE case-control study, 2001-2007, France.

Trajectory of occupational exposure intensity to asbestos	Cases/ Controls n	Age at index date (years) median (5^{th}-95^{th} percentile)	Cumulative index of exposure (f/mL-years) median (5^{th}-95^{th} percentile)	Total duration of exposure (years) median (5^{th}-95^{th} percentile)	Average intensity over lifetime (f/mL) median (5^{th}-95^{th} percentile)	Time since last exposure (years) median (5^{th}-95^{th} percentile)	Time since first exposure (years) median (5^{th}-95^{th} percentile)	$OR^{\pm}$ (95%CI)
Class 1 « constant low intensity »	379/381	61 (41-74)	1.4 (0.3-15.9)	23 (3-46)	0.08 (0.01-1.19)	14 (0-47)	44 (20-58)	1.00
Class 2 « recent moderate intensity »	132/122	50 (41-64)	5.9 (0.4-41.3)	31 (19-45)	0.2 (0.01-1.29)	0 (0-9)	31 (20-47)	1.03 (0.71 ; 1.50)
Class 3 « 20 year very high intensity »	224/177	60 (50-70)	13.5 (0.9-73.4)	39 (26-47)	0.36 (0.02-1.76)	5 (0-14)	42 (33-52)	1.17 (0.85 ; 1.60)
Class 4 « distant high intensity »	177/118	65 (49-74)	12.2 (1.1-59.6)	36 (10-47)	0.4 (0.04-1.80)	15 (2-28)	48 (34-59)	1.26 (0.9 ; 1.79)

OR, odds ratio; CI, confidence interval

$^{\pm}$ Adjusted for age at the index date (in years), area of residence (*département*) and smoking (Comprehensive Smoking Index at the index date).

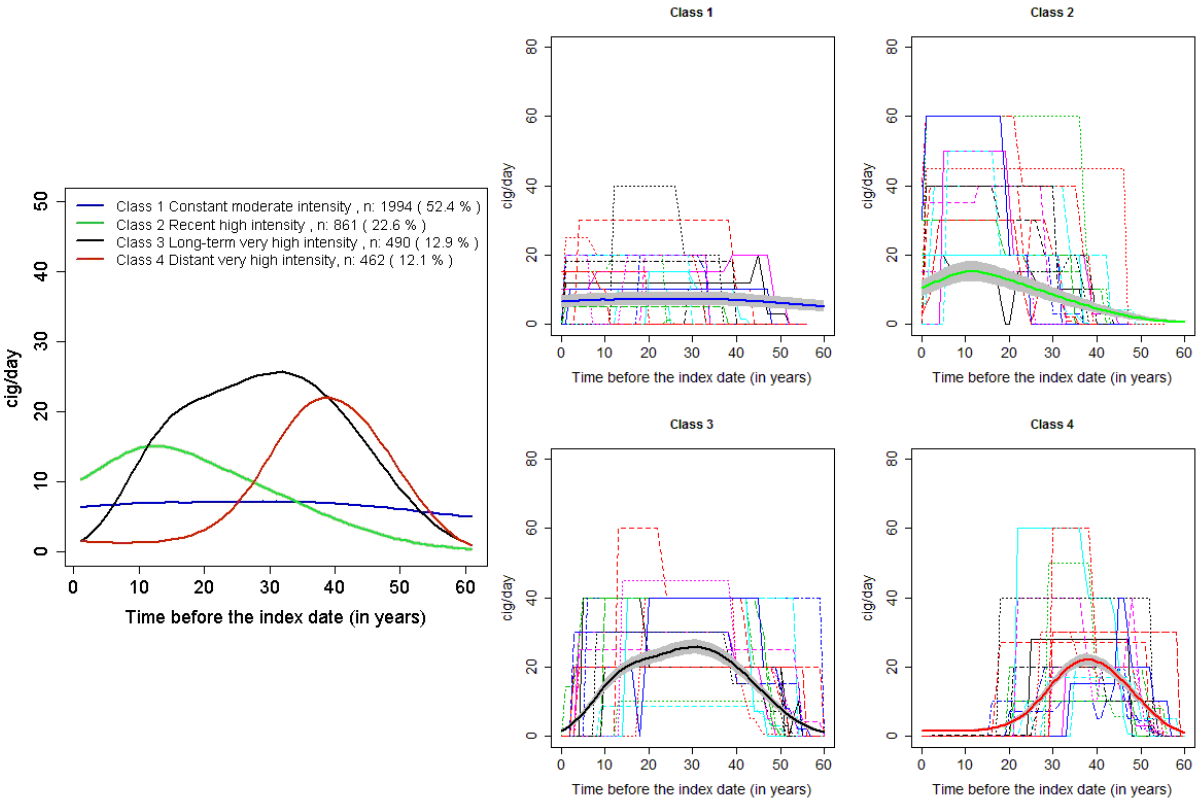

Figure 1 *Lifetime trajectories of smoking intensities. The left panel shows the estimated mean trajectory of smoking intensity in the four latent classes. The right panel shows for each class, 20 randomly selected observed individual trajectories of subjects who had a high probability (close to 1) to belong to the class, with the bold line representing the estimated mean trajectory in the Class, with its 95% CI.*

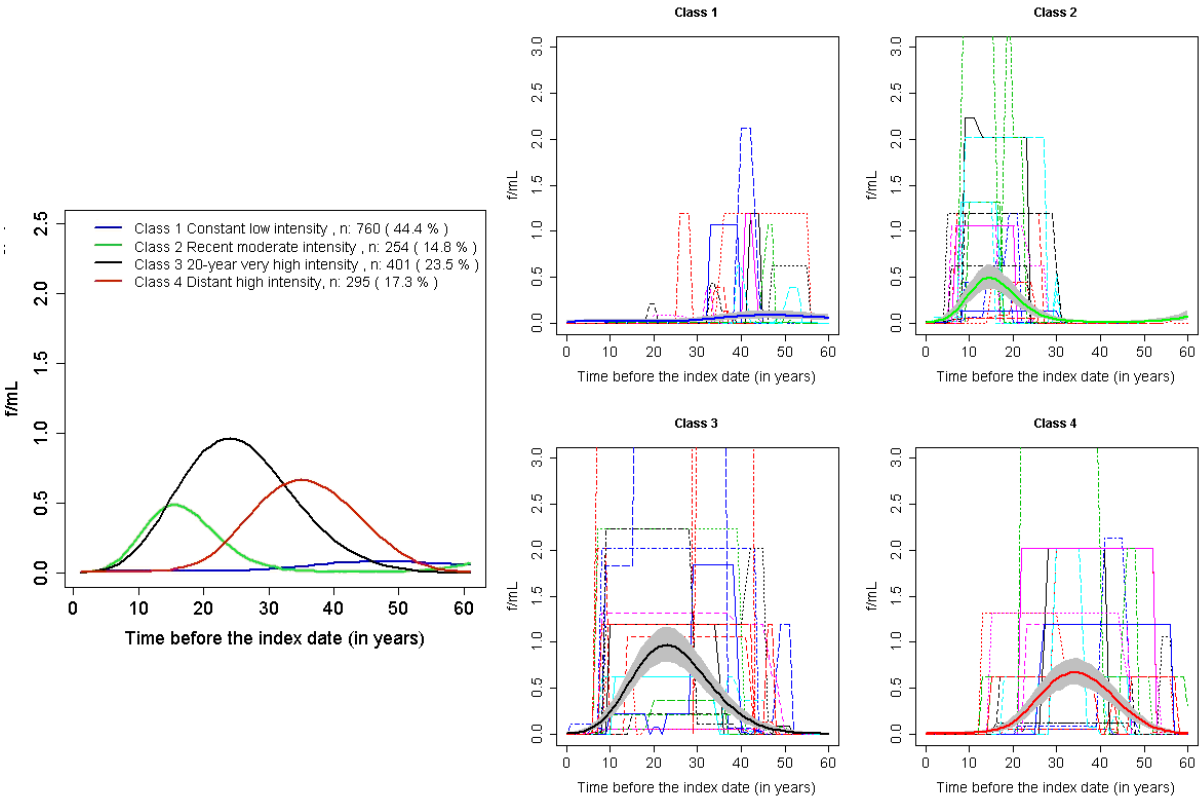

Figure 2 *Lifetime trajectories of intensities of occupation exposure to asbestos. The left panel shows the estimated mean trajectory of asbestos intensity in the four latent classes. The right panel shows for each class, 20 randomly selected observed individual trajectories of subjects who had a high probability (close to 1) to belong to the class, with the bold line representing the estimated mean trajectory in the Class, with its 95% CI.*

Supplementary Material

Equations of the joint latent class mixed model for a longitudinal marker and a binary outcome

In the following part, we give the equations of the three sub-models used in the two joint latent class models (JLCMM) for smoking and occupational exposure to asbestos.

For smoking exposure:

Sub-model 1: multinomial logistic regression for latent class membership

The probability that a subject i belongs to latent class g was given by:

$$\pi_{ig} = P(c_i = g) = \frac{e^{\gamma_{0g}}}{\sum_{l=1}^{G} e^{\gamma_{0l}}}$$

Where c_i denotes the discrete random variable which equals g if the subject i belongs to latent class g (g=1,…,G).

- γ_{0g} is the intercept for class g. For identifiability γ_{0G}=0.

Sub-model 2: class-specific mixed model

Let $Y_i=(Y_{i1}, .., Y_{ij},.. Y_{ini})$ be the vector of repeated measures where Y_{ij} is the annual intensity of exposure at j^{th} year before the index date of subject i and n_i is the number of years of exposure for subject i.

The observed annual smoking intensity of subject i in the j^{th} year before diagnosis/interview is modelled according to time using a latent process mixed model which simultaneously normalizes the intensities using a link function H and models the change over time of the transformed intensities using a linear mixed model:

$$H(Y_{ij})|ci = g = \left(b_{0g} + u_{0ig}\mathbf{1}_{t_{ij}\in Hist_i}\right) + \sum_{l} \mathrm{b}_{\mathrm{lg}}\, B_l(t_{ij}) + \varepsilon_{ij}$$

Where

- H is a I-splines function to account for non-normality of annual intensities with 3 manual knots at 0, 20 and 100 cig/day (20)
- t_{ij} is the j^{th} year before the index date for individual i
- ε_{ij} are assumed to be independent Gaussian measurement errors with variance σ^2_{ε}
- $b = c(b_{0g}, b_{lg})$ is the vector of class-specific fixed effects
- $B_l(t)$ are the splines basis function of time before index date with 3 inner knots placed at quartiles (10, 21 and 33)
- u_{0ig} the intercept individual class-specific random effect with $\mathbf{1}_{t_{ij}\in Hist_i}$ an indicator which equals to one if time t_{ij} is during the individual history exposure ($Hist_i$) defined by the

different times between the first exposure and the last exposure of individual i. Assuming $u_{0ig} \sim N(0, w_g^2 \sigma^2_u)$ where σ^2_u was an unspecified common variance and w_g a proportional coefficient allowing for a class-specific between subjects variability

Sub-model 3: logistic regression model for the probability of lung cancer

The case-control status D is modelled using a logistic regression model :

$$logit\big(P(D_i = 1|Age, AreaResidence, CIE\_Asbestos, c_i = g)\big) = \delta_{0g} + f_1(Age) + f_2(AreaResidence) + f_3(CIE\_Asbestos)$$

Where

- D is the binary outcome which corresponds to the case-control status (= 1 for case, 0 for control)
- δ_{0g} is the class-specific intercept
- $f_1(Age)$ is natural cubic splines of Age at the index date with 3 knots placed at 5th, 50th, 95th percentiles of age distribution
- $f_2(AreaResidence)$ is a linear combination of 9 indicators variables for the French *départements* of residence
- $f_3(CIE\_Asbestos)$ is a linear combination of the indicators of the quartiles of Cumulative Index of Exposure (CIE) of occupational exposure to asbestos (0, <=0.00153, <=0.8159, >0.8159 f/mL-years)

The adjusted Odd Ratio (OR) for lung cancer between Class 2 and Class 1 equals exp (δ_{02}- δ_{01}).

For occupational exposure to asbestos:

Sub-model 1: multinomial logistic regression for latent class membership

The probability that a subject i belongs to latent class g was given by:

$$\pi_{ig} = P(c_i = g) = \frac{e^{\gamma_{0g}}}{\sum_{l=1}^{G} e^{\gamma_{0l}}}$$

Where c_i denotes the discrete random variable which equals g if the subject i belongs to latent class g (g=1,…,G).

- γ_{0g} is the intercept for class g. For identifiability γ_{0G}=0.

Sub-model 2: class-specific mixed model

Let $Y_i=(Y_{i1}, .., Y_{ij},.. Y_{ini})$ be the vector of repeated measures where Y_{ij} is the annual intensity of exposure at jth year before the index date of subject i and n_i is the number of years of exposure for subject i.

The observed annual occupational asbestos intensity of subject i in the j^{th} year before diagnosis/interview is modelled according to time using a latent process mixed model which simultaneously normalizes the intensities using a link function H and models the change over time of the transformed intensities using a linear mixed model:

$$H(Y_{ij})|ci = g = \left(b_{0g} + u_{0ig}\mathbf{1}_{t_{ij}\in Hist_i}\right) + \sum_l \mathrm{b_{lg}}\, B_l(t_{ij}) + \varepsilon_{ij}$$

Where

- H is a I-splines function to account for non-normality of annual intensities with 3 manual knots at 0,0.1,12.6 f/mL
- t_{ij} is the j^{th} year before the index date for individual i
- ε_{ij} are assumed to be independent Gaussian measurement errors with variance σ^2_ε
- $b = c(b_{0g}, b_{lg})$ is the vector of class-specific fixed effects
- $B_l(t)$ are the splines basis function of time before index date date with 3 inner knots placed at quartiles (12, 24 and 36)
- u_{0ig} the intercept individual class-specific random effect with $\mathbf{1}_{t_{ij}\in Hist_i}$ an indicator which equals to one if time t_{ij} is during the individual history exposure ($Hist_i$) defined by the different times between the first exposure and the last exposure of individual i. Assuming $u_{0ig} \sim N(0, w_g^2\ \sigma^2_u)$ where σ^2_u was an unspecified common variance and w_g a proportional coefficient allowing for a class-specific between subjects variability

Sub-model 3: logistic regression model for the probability of lung cancer

The case-control status D is modelled using a logistic regression model:

$$logit\big(P(D_i = 1|Age, AreaResidence, CSI\ , c_i = g)\big) = \delta_{0g} + f_1(Age) + f_2(AreaResidence) + \delta_3 CSI_i$$

Where

- D_i is the binary outcome which corresponds to the case-control status (= 1 for case, 0 for control)
- δ_{0g} is the class-specific intercept
- $f_1(Age)$ is natural cubic splines of Age at the index date with 3 knots placed at 5^{th}, 50^{th}, 95^{th} percentiles of age distribution
- $f_2(AreaResidence)$ is a linear combination of 9 indicators variables for the French *départements* of residence
- CSI_i is the Comprehensive Smoking Index (CSI) which is a single aggregate measure for each subject of his lifetime smoking history derived from his reported average number of cigarettes smoked per day over lifetime, total duration of smoking, and time since smoking cessation at the index date(21)

The adjusted Odd Ratio (OR) for lung cancer between Class 2 and Class 1 equaled exp (δ_{02}- δ_{01}).

Table 1S Numerical values of probability, frequency, and intensity of asbestos exposure used in the Job Exposure Matrix (JEM) to derive individual average annual daily intensity of exposure.

Asbestos exposure characteristics	Definition	Numerical values used to calculate annual doses		
Probability of exposure (% of workers exposed)				
Non exposed	0	0		
Possible	> 0 - 5	0.025		
Probable	5 - 30	0.175		
Likely	30 - 70	0.5		
Definite	≥ 70	0.85		
Frequency of exposure (% of work time)				
Sporadic	> 0-5	0.025		
Occasional	5-30	0.175		
Frequent	30-70	0.5		
Continuous	≥ 70	0.85		
Intensity of exposure (equivalent fibres/ml)*		Passive exposure	Indirect exposure	Direct exposure
Very low	> 0 - 0.01	0.0005	0.0025	0.005
Low	0.01 - 0.1	0.005	0.025	0.05
Medium	0.1 - 1	0.05	0.25	0.5
High	1 - 10	0.5	2.5	5
Very high	≥ 10	2	10	15

* Intensity of exposure was defined as a combination of the intensity of exposure due to specific task and work environment contamination. Since the asbestos JEM was based on expert judgment, intensity of exposure was expressed in equivalent fibres/ml. Three types of exposure were defined: Passive exposure (workers were exposed according to diffuse contamination of buildings); indirect exposure (workers were exposed by other workers using asbestos materials); direct exposure (workers used directly asbestos materials).

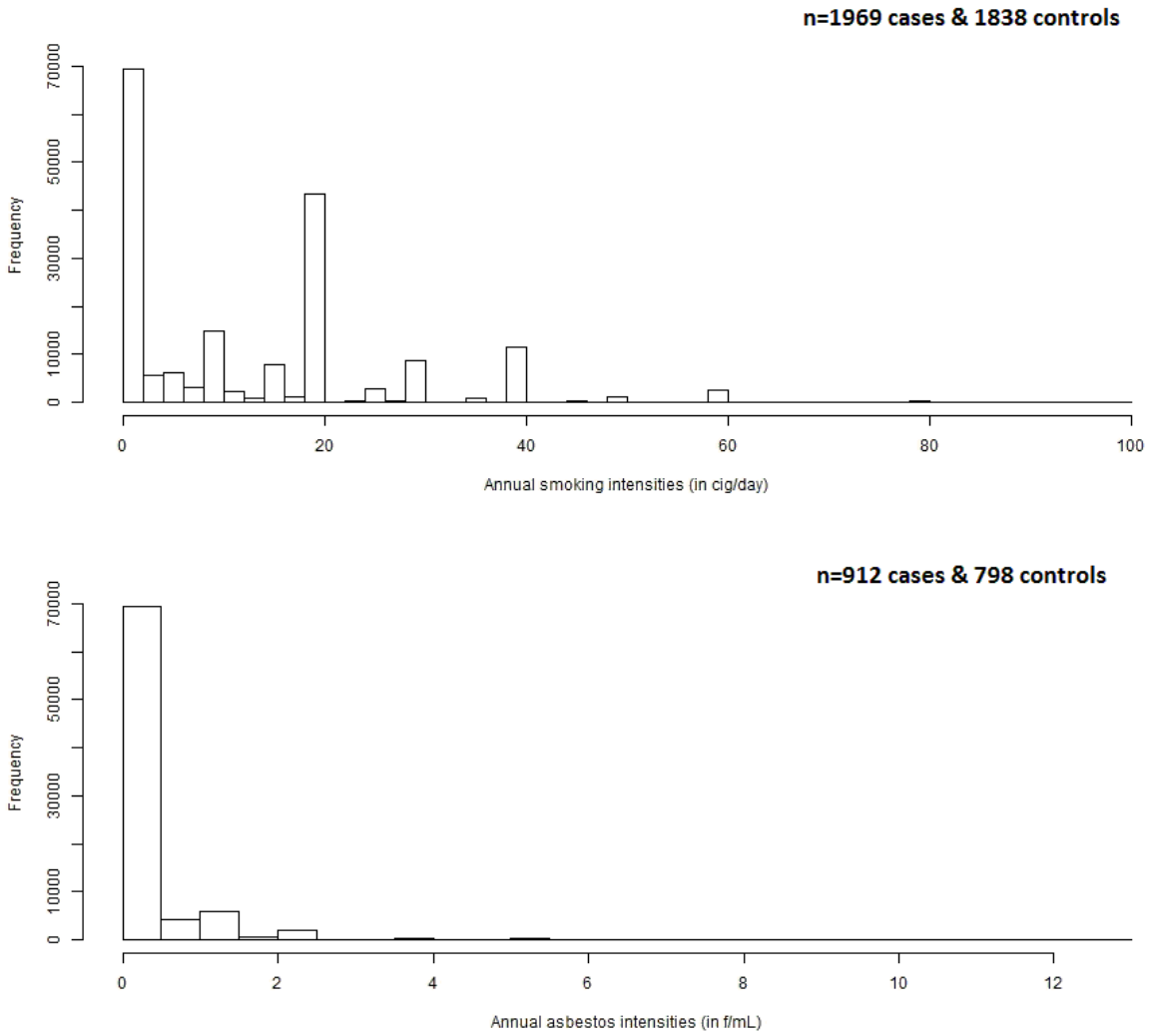

Figure 1S Distribution of the annual average daily intensities of exposure. *At the top, distribution of repeated measures for the annual average number of cigarettes smoked per day. At the bottom, distribution of the repeated measures for the annual average intensities of occupational exposure to asbestos (in f/mL).* ICARE case-control study, 2001-2007, France.

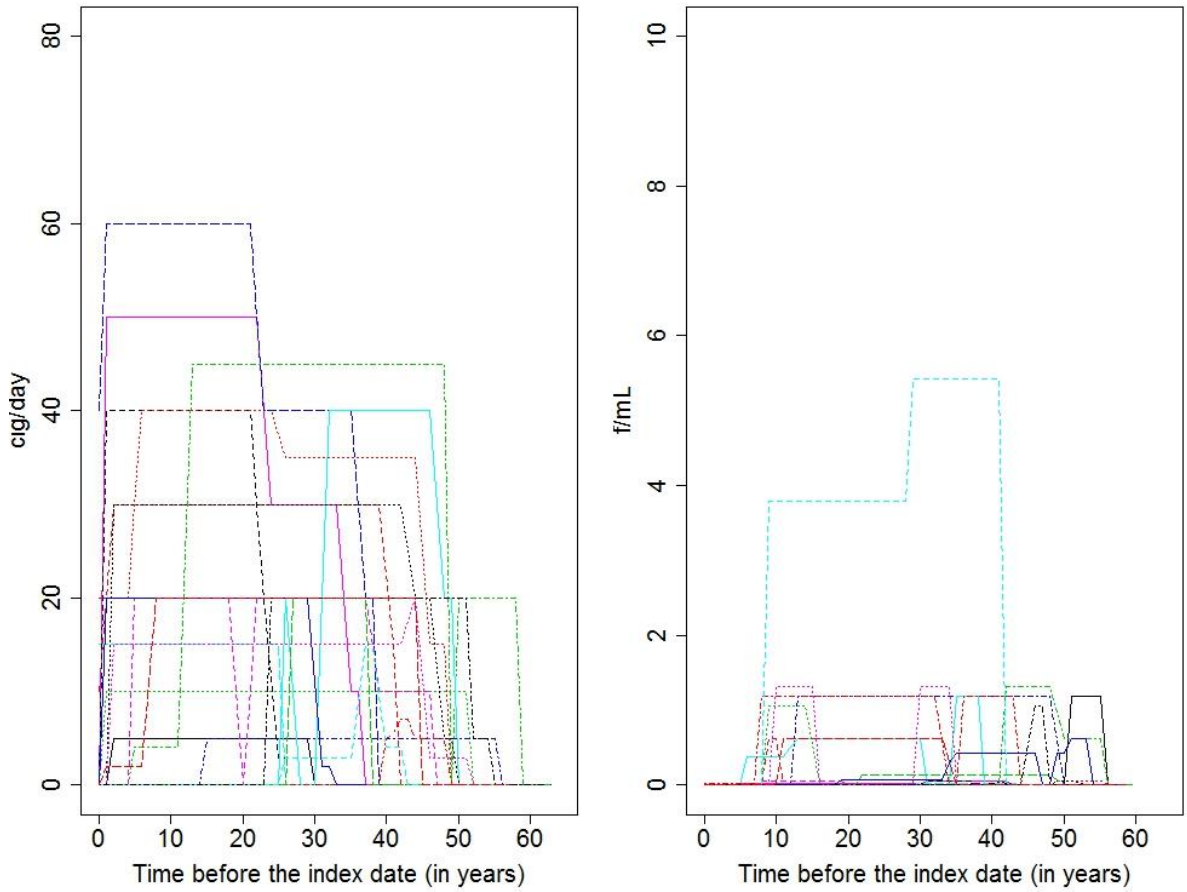

Figure 2S 20 random observed individual trajectories; left panel for smoking intensities and right panel for intensities of occupational exposure to asbestos. ICARE case-control study, 2001-2007, France.

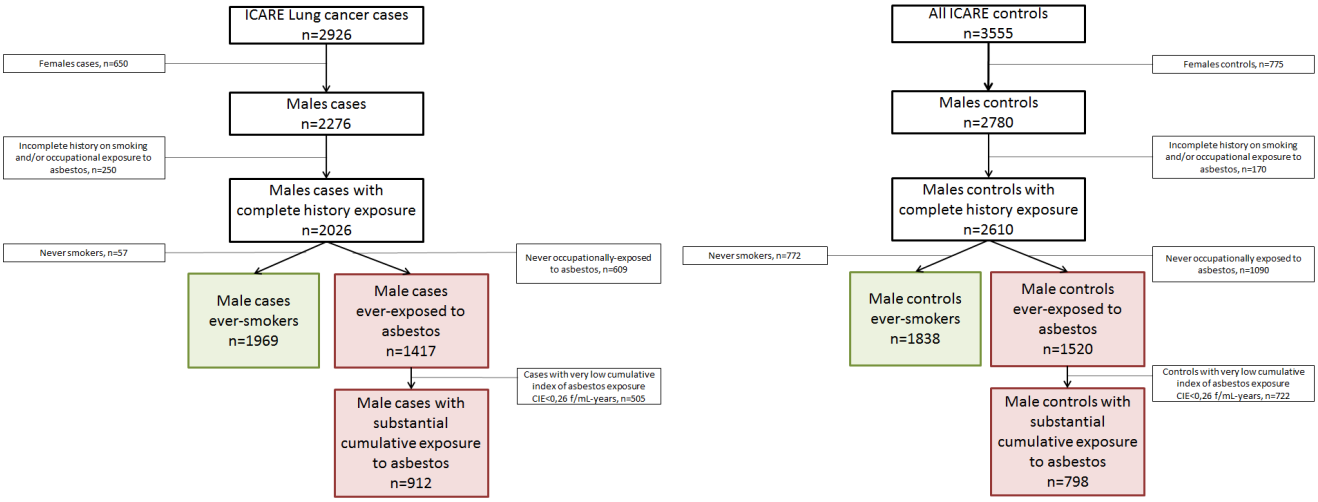

Figure 3S. Selection of subjects for the joint lcmm analysis in the ICARE case-control study, 2001-2007, France.

Table 2S Socio-demographics characteristics for the excluded and included subjects exposed occupationally to asbestos. ICARE case-control study, 2001-2007, France.

	Ever exposed asbestos cases (n=1417)	Ever exposed asbestos controls (n=1520)	Ever exposed asbestos included cases (n=912)	Ever exposed asbestos included controls (n=798)
Age at index date (in years)				
Mean (SD)	60.0 (9.0)	58.3 (10.0)	60.0 (8.6)	58.5 (9.9)
Area of residence (département) (n, %)				
Calvados	169 (11.9)	215 (14.1)	115 (12.6)	109 (13.7)
Doubs - Territoire de Belfort	72 (5.1)	76 (5.0)	39 (4.3)	36 (4.5)
Hérault	145 (10.2)	168 (11.1)	99 (10.9)	94 (11.8)
Isère	234 (16.5)	210 (13.8)	149 (16.3)	107 (13.4)
Loire Atlantique	182 (12.8)	181 (11.9)	129 (14.1)	102 (12.8)
Manche	159 (11.2)	123 (8.1)	123 (13.5)	65 (8.1)
Bas-Rhin	180 (12.7)	195 (12.8)	111 (12.2)	115 (14.4)
Haut-Rhin	40 (2.8)	50 (3.3)	26 (2.9)	29 (3.6)
Somme	151 (10.7)	199 (13.1)	76 (8.3)	90 (11.3)
Vendée	85 (6.0)	103 (6.8)	45 (4.9)	51 (6.4)
Education level (n, %)				
Elementary school or less	475 (33.5)	346 (22.8)	303 (33.2)	187 (23.4)
Middle school	608 (42.9)	724 (47.6)	434 (47.6)	430 (53.9)
High school	71 (5.0)	141 (9.3)	26 (2.9)	65 (8.1)
University	112 (7.9)	233 (15.3)	50 (5.5)	73 (9.1)
Other	19 (1.3)	15 (1.0)	11 (1.2)	10 (1.3)
NA	132 (9.3)	61 (4.0)	88 (9.6)	33 (4.1)

Discrimination capacity of the two JLCMM

From the estimated JLCMM, we derived the estimated posterior probability for each subject to belong to each latent class given his repeated observations and his case-control status. Each subject was thus a posteriori classified in the class he had the highest probability to belong. We further derived the posterior classification table where for each latent class we calculated the mean posterior probability to belong to each latent class among subjects a posteriori classified in the given class. The classification has a good discriminatory capacity if diagonal terms are close to 1 and all others close to 0.

In the posterior classification Table 2S for smoking given below, 481 subjects a posteriori classified in the second class have for example a mean of 94,6% to be classified in second class whereas only 1% and 1,2% to be classified in first and fifth class respectively.

Table 3S Posterior classification table for the four identified latent classes of smoking intensities. ICARE case-control study, 2001-2007,France.

	N*	Mean of the posterior probabilities of belonging to each class			
		1	2	3	4
Class 1	1994	0.9839	0.0122	0.0029	0.0011
Class 2	861	0.0285	0.9706	0.0008	0.0000
Class 3	490	0.0093	0.0017	0.9855	0.0035
Class 4	462	0.0020	0.0000	0.0039	0.9941

*Number of subjects a posteriori classified in the class

Table 4S Posterior classification table for the four identified latent classes of occupational asbestos intensities. ICARE case-control study, 2001-2007,France.

	N*	Mean of the posterior probabilities of belonging to each class			
		1	2	3	4
Class 1	760	0.9785	0.0065	0.0061	0.0089
Class 2	254	0.0156	0.9665	0.0080	0.0050
Class 3	401	0.0068	0.0102	0.9727	0.0102
Class 4	295	0.0206	0.0000	0.0174	0.9714

3.5 Analyses complémentaires sur les données

3.5.1 Comparaison entre proc SAS TRAJ et package R *lcmm*

Pour comparer les résultats entre les deux logiciels, nous avons utilisé les données de consommation de tabac. Nous avons considéré un modèle longitudinal avec une simple évolution linéaire. Sous le package R *lcmm*, nous avons donc estimé le modèle sans transformation des données ni effet aléatoire afin d'estimer exactement le même modèle à classes latentes dans les deux logiciels considérés. Nous donnons les résultats obtenus pour l'estimation des modèles à 4 classes latentes. En figure 3.6, on retrouve les trajectoires moyennes prédites avec les deux modèles, puis les tableaux 3.1 & 3.2 qui représentent les tables de classification a posteriori.

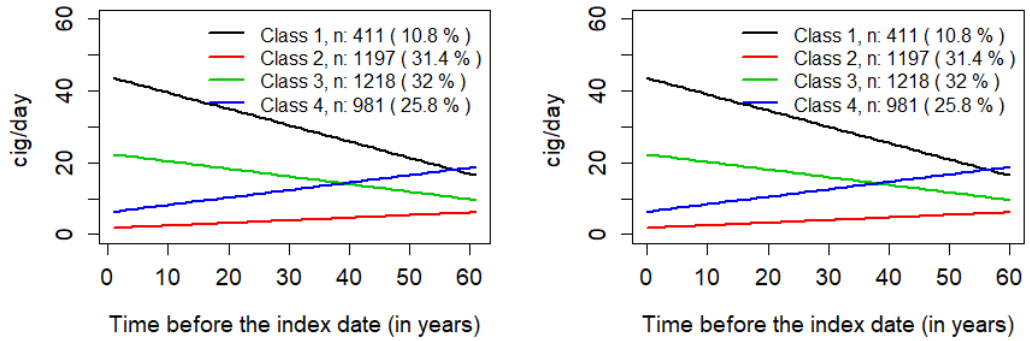

FIGURE 3.6 – Trajectoires moyennes prédites d'intensité de consommation de tabac au cours du temps avant la date index. A gauche, avec le modèle LCMM à 4 classes latentes. A droite, avec le modèle LCGM à 4 classes latentes.

	N	Moyenne des probabilités a posteriori d'appartenance à chaque classe			
		1	2	3	4
Class 1	411	0.9846	0.0000	0.0145	0.0001
Class 2	1197	0.0000	0.9745	0.0003	0.025
Class 3	1218	0.0043	0.0007	0.9713	0.0238
Class 4	981	0.0000	0.0388	0.0241	0.9371

Tableau 3.1 – Table de classification a posteriori pour LCGM à 4 classes latentes

	N	Moyenne des probabilités a posteriori d'appartenance à chaque classe			
		1	2	3	4
Class 1	411	0.9846	0.0000	0.0145	0.0009
Class 2	1197	0.0000	0.9745	0.0003	0.0252
Class 3	1218	0.0043	0.0007	0.9713	0.0238
Class 4	981	0.0000	0.0388	0.0241	0.9371

Tableau 3.2 – Table de classification a posteriori pour LCMM à 4 classes latentes

Comme attendu, nous retrouvons bien les mêmes résultats puisque l'on estime exactement le même modèle à classes latentes. Les différences, s'il y en avaient eu, auraient été expliquées par les logiciels utilisés pour estimer ces modèles.

Nous avons voulu ajouter un effet aléatoire au modèle estimé sous le package R *lcmm*, pour évaluer l'apport de ce dernier par rapport au LCGM estimé sous SAS, qui ne comporte pas d'effet aléatoire. Cependant, sans aucune transformation des données, le modèle LCMM ne converge pas. Si l'on avait transformé les données pour obtenir la convergence du modèle, nous n'aurions pas pu comparer les résultats avec ceux obtenus sous TRAJ en réalisant un test de rapport de vraisemblance car les modèles n'auraient plus été emboîtés.

3.5.2 Comparaison entre le modèle conjoint et les approches en 2 étapes

Pour obtenir les résultats suivant les approches en 2 étapes, il est donc nécessaire d'estimer, dans un premier temps, un modèle mixte à classes latentes. Nous avons estimé deux LCMM similaires à ceux décrits pour les JLCMM estimés, respectivement, pour les relations tabac-cancer du poumon (équations 38 & 40) et amiante-cancer du poumon (équations 42 & 44). Les trajectoires moyennes prédites obtenues sont montrées dans les figures 3.7 et 3.8, auxquelles sont ajoutées les trajectoires moyennes prédites obtenues avec les JLCMM estimés associés aux mêmes relations (provenant de l'article 3).

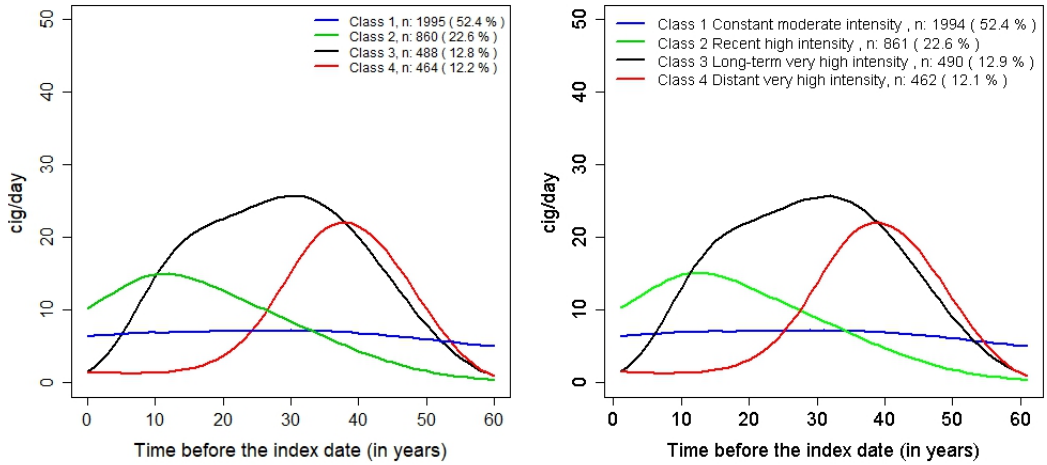

FIGURE 3.7 – Trajectoires moyennes prédites d'intensité de consommation de tabac au cours du temps avant la date index. A gauche, avec le modèle LCMM à 4 classes latentes. A droite, avec le modèle JLCMM à 4 classes latentes.

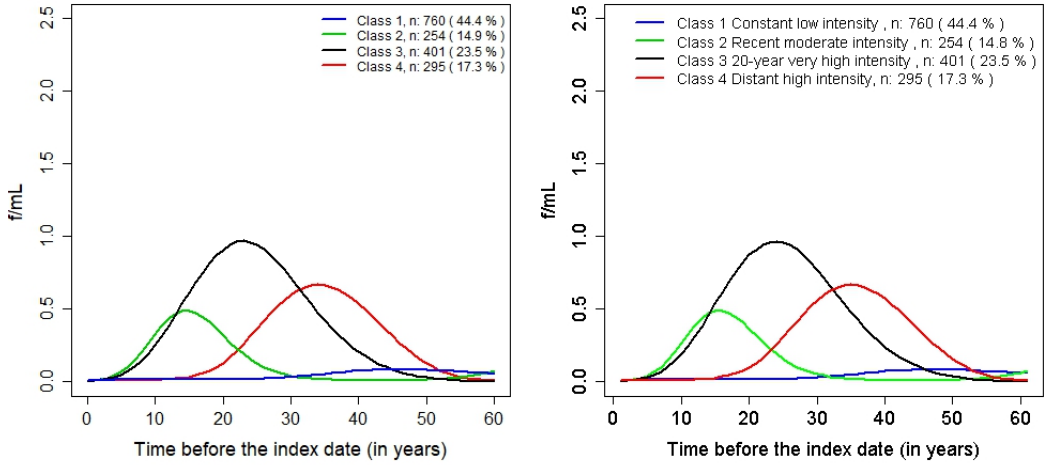

FIGURE 3.8 – Trajectoires moyennes prédites d'intensité d'exposition professionnelle à l'amiante au cours du temps avant la date index. A gauche, avec le modèle LCMM à 4 classes latentes. A droite, avec le modèle JLCMM à 4 classes latentes.

On peut voir que les trajectoires moyennes prédites obtenues entre LCMM et JLCMM sont exactement les mêmes. En terme de classification a posteriori entre LCMM et JLCMM, pour le tabac, on peut constater une légère différence d'un ou deux sujets classés a posteriori pour chacune des 4 classes. Pour l'amiante, ce sont exactement les mêmes effectifs entre les classes. Les comparaisons en terme d'OR entre les approches en 2 étapes et le modèle conjoint vont donc se faire par rapport aux mêmes trajectoires et mêmes classifications a posteriori entre les

approches.

Nous avons ensuite estimé les OR associés à la variable de classification par les approches en 2 étapes (Naïve, Logistique pondérée et Pseudo-classes) (Tableaux 3.3 et 3.4). Les deux dernières colonnes de ces tableaux sont les résultats retrouvés dans l'article 3 concernant le modèle JLCMM estimé respectivement pour les deux applications.

Pour le tabac, les ORs obtenus avec les 4 approches sont très similaires. On constate une tendance à l'atténuation de l'association avec les approches en 2 étapes par rapport au modèle conjoint sauf pour la classe 4. Ceci est cohérent avec les études de comparaison précédemment introduites en section 3.2.1.3. Néanmoins, les ICs associés sont tous de même étendues avec les 4 approches. Pour l'amiante, on constate une atténuation de l'association entre deux approches en 2 étapes et le modèle conjoint, elle est cependant moins marquée que pour le tabac. Cependant, les ORs obtenus avec l'approche naïve sont légèrement plus élevés par rapport à ceux avec le modèle conjoint.

Les résultats sont donc très similaires entre les 4 approches. Cependant, rappelons que le modèle conjoint est le modèle le plus juste puisqu'il prend mieux en compte l'incertitude liée à l'appartenance aux classes latentes par rapport aux approches en 2 étapes. Pour les deux applications et quelle que soit l'approche (en 2 étapes ou conjoint), le pouvoir discriminant était fort puisque la diagonale des tables de classification a posteriori contenait des moyennes de probabilités a posteriori supérieures à 95% (pour LCMM et JLCMM). Cela peut expliquer que l'étendue des intervalles de confiance était similaire entre les approches en 2 étapes et le modèle conjoint. Dans la littérature, il a été reporté une différence entre les approches en 2 étapes et le conjoint mais qui serait essentiellement attribuable à la non prise en compte de la sortie d'étude informative par les approches en 2 étapes pour les données prospectives (utilisant donc un modèle de Cox). Compte tenu du schéma d'étude cas-témoins, nos analyses ne sont pas impactées par cela et nos résultats semblent donc indiqués une faible variabilité dans les estimations des ORs et des intervalles de confiance associés entre les approches. Cela reste, néanmoins, à conforter avec l'application de ces 4 approches à d'autres données cas-témoins.

Classe	N	Approche Naïve	Logistique pondérée	Pseudo-classes	N	Modèle conjoint (JLCMM)
1	1995	1-	1-	1-	1994	1-
2	860	4.38 (3.63 ;5.28)	4.29 (3.59 ;5.14)	4.28 (3.53 ;5.20)	861	4.89 (4.02 ;5.96)
3	488	2.54 (1.99 ;3.25)	2.48 (1.95 ;3.15)	2.49 (1.94 ;3.19)	490	2.60 (2.03 ;3.33)
4	464	0.90 (0.71 ;1.16)	0.89 (0.70 ;1.13)	0.89 (0.70 ;1.14)	462	0.91 (0.71 ;1.17)

Tableau 3.3 – Association entre les classes de trajectoires de consommation de tabac et le cancer du poumon, selon 4 approches, ICARE 2001-2007, France.

Classe	N	Approche Naïve	Logistique pondérée	Pseudo-classes	N	Modèle conjoint (JLCMM)
1	760	1-	1-	1-	760	1-
2	254	1.04 (0.72 ;1.49)	1.03 (0.72 ;1.48)	1.03 (0.71 ;1.50)	254	1.03 (0.71 ;1.50)
3	401	1.19 (0.87 ;1.62)	1.16 (0.85 ;1.58)	1.16 (0.85 ;1.59)	401	1.17 (0.85 ;1.60)
4	295	1.28 (0.92 ;1.79)	1.25 (0.91 ;1.73)	1.23 (0.87 ;1.73)	295	1.26 (0.90 ;1.79)

Tableau 3.4 – Association entre les classes de trajectoires d'exposition professionnelle à l'amiante et le cancer du poumon, selon 4 approches, ICARE 2001-2007, France.

3.6 Discussion

3.6.1 Synthèse des résultats de l'article

Pour le tabac, nous avons identifié 4 profils distincts de trajectoires d'intensité d'exposition : une trajectoire moyenne constante de 8 cigarettes fumées en moyenne par jour (cig/jour) tout au long des années avant la date index (Classe 1) ; une trajectoire moyenne avec des intensités d'exposition plus élevées (plus de 10 cig/jour) dans les 20 années précédant la date index (Classe 2) ; une trajectoire moyenne avec des intensités de plus 20 cig/jour de 40 à 10 ans avant la date index (Classe 3) et une trajectoire avec des intensités d'exposition plus élevées (plus de 15 cig/jour) au delà de 30 ans avant la date index avec un arrêt d'exposition dans les 30 années avant cette date (Classe 4). L'OR le plus élevé était entre la Classe 2 et la Classe 1 (OR=4.89, 95% IC :4.01 :5.96), ce qui peut s'expliquer par des intensités très élevées proche de la date index, regroupant ainsi seulement 51.5% d'ex-fumeurs contrairement aux autres classes qui en ont plus de 90% et un temps médian d'arrêt de consommation d'un an. Bien que la Classe 3 ait l'intensité cumulée la plus élevée (médiane : 907 cigarettes-années) associée à la durée totale la plus élevée (médiane : 46 ans) comparativement à la Classe 1, l'OR associé (OR=2.60, 95% IC : 2.02 ;3.31) était plus faible que celui pour la Classe 2 mais restait élevé. La Classe 4 avec les intensités les plus élevées distantes de la date index n'était pas associée à un sur-risque (OR=0.91, 95% IC :0.71 :1.17) par rapport à la Classe 1 avec la trajectoire moyenne constante. Cette Classe 4 était constituée d'ex fumeurs de longue date (25 ans depuis l'arrêt en médiane). Nous retrouvons donc des résultats cohérents avec ceux du WCIE (chapitre 2) puisque les intensités reçues récemment contribueraient fortement au risque de cancer du poumon. Ces résultats confirment aussi la diminution importante du risque de cancer du poumon après l'arrêt de consommation de tabac.

Pour l'amiante, nous avons identifié 4 profils distincts de trajectoires d'intensité d'exposition : une trajectoire moyenne constante d'intensité très faible (Classe 1) ; une trajectoire moyenne avec des intensités non nulles (moins de 0.5 f/mL) dans les 30 ans avant la date index (Classe 2) ; une trajectoire moyenne avec des intensités très élevées (près de 1 f/mL) autour de 20 ans avant la date index (Classe 3) et une trajectoire moyenne avec des intensités plus élevées (plus de 0.5 f/mL) entre 30 et 50 avant la date index avec un arrêt d'exposition dans les 20 ans précédant la date index (Classe 4). Bien que la Classe 2 ait des sujets avec en médiane un nombre de f/mL-années plus élevé (5.9 f/mL-années) que dans la Classe 1 (1.4 f/mL-années), elle n'était pas associée à une augmentation du risque de cancer du poumon (OR=1.03, 95% IC :0.71 ;1.50). C'est une classe qui est composée principalement de sujets encore exposés à la date index vu que le temps médian depuis la dernière exposition était nul. Même si le temps médian depuis la dernière exposition est similaire entre les Classes 4 et 1, on pouvait constater une augmentation du risque (OR= 1.26, 95 % IC :0.90 ;1.79). Les sujets de cette Classe 4

sont en médiane les plus âgés (65 ans) par rapport aux autres classes avec un temps médian depuis la première exposition plus long (48 ans). Bien que la Classe 3 soit caractérisée par la dose cumulée la plus élevée (13.5 f/mL-années) et la durée totale la plus longue (médiane : 39 ans), elle n'était pas associée au risque de cancer du poumon le plus élevé parmi les 3 classes (OR=1.17, 95% IC :0.85 ;1.60). Tous ces résultats obtenus pour l'amiante sont concordants avec ceux obtenus avec le WCIE. En effet, les intensités reçues longtemps avant la date index auraient la plus forte contribution sur le risque de cancer du poumon.

3.6.2 Perspectives autour de l'utilisation d'un modèle conjoint à classes latentes pour l'identification de trajectoires d'expositions environnementales/professionnelles dans les études cas-témoins

A notre connaissance, c'est la première étude appliquant un modèle conjoint à classes latentes sur des données d'expositions environnementales et professionnelles provenant d'une étude cas-témoins. Même si les résultats issus de cette étude sont très encourageants, il y a néanmoins un certain nombre de limites qui doivent être discutées et qui doivent orienter vers des développements complémentaires.

3.6.2.1 L'axe du temps

Nous disposions de données rétrospectives liées à la consommation tabagique et à l'histoire d'exposition professionnelle à l'amiante des sujets. Pour identifier des profils longitudinaux de ces expositions, il est donc primordial de fixer un axe du temps. Pour coller au schéma de l'étude cas-témoins, c'est une axe du temps rétrospectif qui a été considéré. Les trajectoires ont donc été estimées pour chaque année avant la date index. Identifier les profils de trajectoires d'intensités selon une échelle prospective telle que l'âge courant, comme nous l'avons fait pour le WCIE, ne semble pas trivial dans le contexte du modèle conjoint. En effet, dans ce dernier cas, il serait plus naturel de considérer un modèle à risques instantanés proportionnels prenant l'âge comme axe du temps comme modèle d'évènement. Dans le contexte des études cas-témoins, ce modèle nécessite en particulier de redresser les ensembles à risque pour tenir compte du sur-échantillonnage des cas, propre aux études cas-témoins (Leffondré et al. [2010], Gauvin et al. [2013]). Ce modèle appelé modèle de Cox pondéré nécessite d'attribuer un poids différent aux cas et témoins afin de refléter leur probabilité d'inclusion dans l'étude. Ces poids sont définis par des probabilités conditionnellement à l'âge de développer le cancer dans la population source. Lorsque l'on ne dispose pas de telles informations sur la population source, ces poids peuvent être estimés à partir de statistiques nationales de santé sur le cancer étudié. Ce modèle conjoint avec un modèle de Cox pondéré demanderait donc une implémentation supplémentaire par rapport au modèle conjoint avec un modèle de Cox classique déjà implémenté dans le package R *lcmm*. De plus, les performances d'un tel modèle conjoint devront être étudiées par simulation.

3.6.2.2 Modélisation de l'évolution longitudinale

Nous avons choisi de modéliser l'évolution moyenne des profils longitudinaux par des fonctions du temps flexibles représentées par des splines du fait des trajectoires individuelles en escalier observées dans les données. En considérant l'indicatrice $Hist_i$ appliquée à l'intercept aléatoire, on estime des écarts individuels à l'évolution moyenne seulement durant l'histoire d'exposition du sujet ce qui permet d'être plus précis qu'avec un intercept aléatoire classique. Cependant, les évolutions moyennes, quant à elles, étaient estimées à partir de splines définies entre le temps depuis la date index et 60 ans avant cette date. Pour estimer plus précisément cette évolution moyenne, une solution intéressante serait de l'estimer à partir de splines qui seraient définies seulement entre les temps de mesure caractérisant le début et la fin de l'histoire d'exposition du sujet. Cependant, la principale difficulté serait liée à la position des noeuds des splines puisque les noeuds devront alors être associés aux quantiles de la distribution des temps de mesure propre à chaque sujet et non plus suivant la distribution marginale de tous les temps de mesures des sujets comme pour ce travail. Pour éviter de rencontrer cette difficulté, il faudrait envisager d'utiliser un autre type de splines pour l'évolution moyenne des profils telles que des splines permettant de considérer un noeud en chaque temps de mesure qui la définie. Cependant, cela augmenterait considérablement le temps computationnel d'estimation du modèle. Néanmoins, c'est un aspect méthodologique qui pourrait être envisagé en perspectives de ce travail réalisé.

3.6.2.3 Ajustement sur l'âge

Utiliser comme axe du temps, le temps avant la date index permet de situer l'histoire d'exposition du sujet par rapport à sa date index. Par exemple, en t=2, on observe pour tous les sujets leur intensité d'exposition reçue deux ans avant leur diagnostic/interview. Cependant, ces sujets peuvent avoir des âges complètement différents en ce t=2. Un ajustement supplémentaire sur l'âge dans le modèle mixte pourrait être à envisager pour permettre de caler les trajectoires sur le même âge. Nous avons fait le choix de ne pas ajuster sur l'âge dans le sous-modèle mixte puisque l'âge était considéré dans le sous-modèle logistique, du fait de l'appariement des cas et témoins sur l'âge. Un ajustement sur l'âge aurait également pu être considéré dans le sous-modèle multinomial d'appartenance aux classes, pour tenir compte du fait, par exemple, que les sujets jeunes à la date index ont une probabilité potentiellement plus faible que les sujets plus âgés d'appartenir à une classe d'ex-fumeurs de long terme. L'impact de l'ajustement supplémentaire sur l'âge dans le sous-modèle mixte et/ou le sous-modèle multinomial, en plus du sous-modèle logistique mérite d'être étudier dans des études ultérieures.

3.6.2.4 Généralisation des résultats

Aucune validation externe sur les trajectoires estimées n'a été réalisée dans ce travail. Il n'y avait aucun but à la généralisation de telles trajectoires à la population générale française.

En effet, par définition de toute étude cas-témoin, nous avons une sur-représentation des cas ce qui rend impossible une telle généralisation. Ce travail avait un but étiologique afin d'enrichir les connaissances de l'association entre les expositions prolongées et le risque de cancer du poumon. Néanmoins, il serait intéressant de pouvoir appliquer ces modèles sur d'autres études cas-témoins du cancer du poumon afin de pouvoir comparer avec nos résultats obtenus. Même si les trajectoires moyennes estimées seraient probablement différentes en terme de valeurs des intensités d'exposition, elles devraient amener à des conclusions similaires quant à l'impact des expositions récentes et anciennes sur le risque de cancer du poumon.

3.6.2.5 Les mesures répétées

Si l'on considère que la trajectoire individuelle observée est définie entre le temps à la première exposition et celui à la dernière exposition, le modèle mixte considère qu'au-delà de ces temps propres de l'histoire d'exposition du sujet, la trajectoire a en moyenne la même évolution estimée qu'aux temps observés de son exposition (processus MAR). Pour relâcher cette hypothèse, nous avons considéré des intensités nulles supplémentaires sur la période d'exposition potentielle. Pour les deux applications, tabac et amiante, nous avons ainsi considéré qu'un sujet pouvait avoir été potentiellement exposé à partir de l'âge de 12 ans jusqu'à l'âge à la date index. Ainsi, par exemple, pour un sujet qui avait déclaré avoir commencé à fumer à 20 ans, les intensités entre 12 et 20 ans étaient considérées comme nulles et non manquantes. Si le sujet avait arrêté de fumer 5 ans avant la date index, les intensités entre la date index (t=0) et cinq ans avant (t=5) étaient nulles. Ce choix de la période d'exposition potentielle pour compléter les trajectoires individuelles observées peut être discutable, et son impact sur les résultats mériterait là aussi d'être davantage étudié.

3.6.2.6 Distribution des mesures répétées

Nous avons dû faire face à une distribution des mesures répétées assez éloignée de la distribution gaussienne supposée pour l'application de modèle conjoint linéaire mixte à classes latentes. Pour corriger ceci, nous avons utilisé une transformation des données répétées via l'estimation d'un modèle mixte à processus latent. Comme aucune des transformations classiques (linéaire, logarithme, Box-Cox (Sakia [1992])) n'a donné des résultats satisfaisants pour rendre la distribution proche d'une distribution gaussienne, nous avons utilisé des Isplines quadratiques. Cependant, cette transformation n'a pas été suffisante pour l'amiante puisque nous avons dû exclure les sujets très faiblement exposés afin d'atteindre la convergence du modèle.
Afin de mieux prendre en compte cette particularité des données d'expositions prolongées, un développement méthodologique paraît donc nécessaire et c'est ce qui fait l'objet du chapitre suivant.

3.7 Contribution et Valorisation

3.7.1 Contribution

Dans le cadre de ce second travail de thèse, ma contribution personnelle tient en :

— l'implémentation sous R permettant l'estimation du modèle JLCMM et des approches en deux étapes. En sachant que pour l'estimation du JLCMM, j'ai modifié la fonction *jointlcmm* du package R *lcmm* ;

— l'analyse des données, à savoir : la sélection des sujets, l'évaluation des transformations, la spécification du modèle, l'interprétation des résultats ;

— la rédaction de l'article 3 ;

Encadrement de M. Amin El Gareh, étudiant en Master 2 MIGS (co-encadrement : Karen Leffondré, Aude Lacourt) sur la faisabilité de l'application d'un modèle conjoint à classes latentes sur des données d'expositions professionnelles à l'amiante.

3.7.2 Valorisation Scientifique

Article soumis

Lévêque E, Lacourt A, Phillips V, Luce D, Guénel P, Stücker I, Proust-lima C and Leffondré K. Association between lung cancer and lifetime profiles of intensity of exposure to occupational asbestos and smoking : Results from the ICARE case-control study. *International Journal of Epidemiology.*

Présentation affichée

Identification of lifetime profiles of smoking intensities and association with lung cancer risks : Results from the ICARE case-control study. The 51st Annual Meeting of Society for Epidemiological Research (SER), Baltimore, June 19-22, 2018.

Bourse obtenue dans le cadre de congrès

SPC Travel Scholarship for The 51st Annual Meeting of Society for Epidemiological Research (SER), Baltimore, June 19-22, 2018 : 400$

4 Chapitre 4 : Développement du modèle mixte ZIP à classes latentes (ZIP-LCMM) (travail en cours et en collaboration avec Cécile Proust-Lima)

Sommaire

Dans le chapitre 3, afin d'identifier des profils de trajectoires d'intensité de tabagisme et d'exposition professionnelle à l'amiante, et d'estimer leur association avec le risque de cancer du poumon, nous avons estimé un modèle conjoint mixte à classes latentes sur les données de l'étude cas-témoins ICARE. Pour estimer ce modèle, nous avons dû faire face à des distributions non gaussiennes des mesures répétées des intensités d'exposition. Nous avons donc considéré une transformation des intensités d'exposition, en utilisant un modèle à processus latent mixte à classes latentes. Un développement méthodologique pour mieux tenir compte de la particularité des distributions des mesures répétées a été envisagé dans le cadre d'un dernier travail de thèse présenté dans ce dernier chapitre. Nous détaillerons précisément le modèle envisagé et son estimation. Nous aborderons ensuite les résultats obtenus pour la consommation de tabac. Cependant, ce travail est encore en cours, et les résultats obtenus restent donc préliminaires. Les perspectives seront détaillées en conclusion de ce chapitre.

4.1 Justification

La particularité de la distribution des mesures répétées de l'intensité d'exposition vient d'un nombre important de zéros (voir histogrammes des figures 3.2 & 3.4 au chapitre 3). En effet, comme nous avons pu le détailler dans le chapitre précédent, nous avons dû considérer des zéros supplémentaires en dehors de l'histoire d'exposition de chaque sujet en attribuant des intensités nulles sur une période d'exposition potentielle définie de 12 ans jusqu'à l'âge de la date index. La conséquence d'un tel ajout est l'augmentation non négligeable du nombre de zéros, qui représente une proportion de 33.8% des intensités annuelles pour le tabac et de 37.8% des intensités annuelles pour l'amiante.

Nous avons donc considéré une distribution permettant de gérer ces zéros. Ces zéros pouvaient s'expliquer par deux processus puisque nous avions

1) des zéros correspondant aux arrêts au minimum d'un an durant l'histoire d'exposition du sujet, considérés comme des "vrais" zéros ;
2) des zéros définis comme "structurels" qui ont été ajoutés pour tenir compte du fait que le sujet n'ait pas été exposé sur la période d'exposition potentielle en dehors de son histoire d'exposition

Dans le processus 1, à un temps j, le sujet est considéré "actif" mais n'a pas été exposé en ce temps qui est durant son histoire d'exposition. Dans le processus 2, le sujet est considéré "inactif" en ce temps j car le temps j se trouve en dehors de son histoire d'exposition observé, mais dans la période d'exposition potentielle s'étendant de 12 ans jusqu'à l'âge de la date index. Il est important d'appuyer sur ce point car c'est ce qui explique notre choix d'utiliser un modèle à sur-représentation de zéros ("Zero-Inflated") au lieu des modèles appelés "hurdle" (Mullahy [1986]) ou "two-part" (Manning [1981]).

En effet, ces derniers modèles permettent également de gérer des zéros, mais ils font l'hypothèse que ces zéros ne s'expliquent que par un seul processus. Ils se traduisent donc par une modélisation des deux parties bien distinctes :
1) la probabilité que la mesure du marqueur longitudinal soit nulle, et
2) la modélisation des mesures répétées non nulles par une distribution tronquée pour les valeurs positives du marqueur longitudinal.

Un modèle à sur-représentation de zéros (ZI) est, au contraire, un mélange de distribution où il y a une distribution pour les zéros structurels et une seconde distribution pour les mesures répétées qui peuvent aussi contenir des zéros (les "vrais" zéros) (Lambert [1992], Nagin and Tremblay [2005]).

Étant donné les distributions d'intensité observées à la fois pour le tabac et l'exposition professionnelle à l'amiante (figures 3.2 & 3.4, chapitre 3), il nous a semblé approprié de considérer une distribution de Poisson pour les mesures répétées hors zéros structurels. En effet, le nombre moyen de cigarettes fumées par jour sur l'année pouvait être considéré comme des données de comptage. Les intensités d'exposition professionnelle à l'amiante en éq f/mL issues de l'application de la MEE pouvaient également être considérées comme plutôt discrètes puisque l'intensité moyenne annuelle dérive d'un produit de paramètres d'exposition représentés par des variables semi-quantitatives, comme détaillé dans le chapitre 1. Il est donc à noter que cette distribution semble adaptée pour des intensités d'exposition que l'on peut considérer comme discrètes et non pour des intensités qui seraient plutôt considérées comme continues (par exemple, les expositions aux polluants atmosphériques dont la concentration est quantifiée directement par une station de mesure). Ainsi, dans la suite de ce chapitre, nous avons considéré une distribution de Poisson à "sur-représentation" de zéros, nommée ZIP (Zero-Inflated Poisson), dans un modèle mixte à classes latentes.

Le modèle envisagé, que l'on nomme ZIP-LCMM (Zero-Inflated Poisson Latent Class Mixed Model) par la suite, n'est pas implémenté dans le package R *lcmm* ni dans d'autres packages à notre connaissance. Dans le logiciel SAS, la procédure TRAJ permet de faire du ZIP mais ne permet pas de considérer des effets aléatoires (Roeder et al. [1999], Nagin [2005]). Le logiciel Mplus permet l'estimation d'un tel modèle (Muthén and Asparouhov [2009]) mais nous allons voir que certaines hypothèses étaient restrictives. De plus, nous voulions utiliser le même algorithme d'optimisation que celui utilisé pour le modèle conjoint JLCMM du chapitre 3, ce qui n'était pas possible avec Mplus qui utilise un algorithme de type EM (Dempster et al. [1977]).

Les objectifs de ce travail étaient donc de développer et d'implémenter le modèle ZIP-LCMM dans le logiciel R puis d'appliquer ce modèle pour identifier des profils de trajectoires d'intensité de consommation de tabac vie entière. Leur association avec le risque de cancer du poumon a été estimée par une approche en 2 étapes (présentée au chapitre 3), car le développement d'un modèle conjoint ZIP-LCMM fera l'objet d'un travail futur.

4.2 Le modèle mixte ZIP à classes latentes (ZIP-LCMM)

4.2.1 Définition

Comme le modèle mixte à classes latentes, le modèle ZIP-LCMM est composé de deux sous-modèles estimés simultanément. Le premier sous-modèle caractérise la probabilité de l'individu i d'appartenir à la classe latente g ($g \in \{1, \ldots, G\}$). La probabilité individuelle π_{ig} d'appartenir à la classe latente g est obtenue à partir d'un modèle logistique multinomial :

$$\pi_{ig} = P(c_i = g|X_{pi}) = \frac{\exp(\zeta_{0g} + X_{pi}^T \zeta_{1g})}{\sum_{l=1}^{G} \exp(\zeta_{0l} + X_{pi}^T \zeta_{1l})} \tag{48}$$

où ζ_{0l} sont les intercepts spécifiques pour chaque classe l ($l \in \{1, \ldots, G\}$) ; ζ_{1l} sont les vecteurs des coefficients spécifiques à chaque classe l associés au vecteur de covariables X_{pi}^T. Par soucis d'identifiabilité, ζ_{0G}=0 et ζ_{1G}=0.

Le second sous-modèle estime les profils moyens de trajectoires à travers un modèle mixte ZIP spécifique à chaque classe latente. Conditionnellement à la classe latente g, les mesures répétées d'un même individu i suivent une distribution ZIP (Roeder et al. [1999]). Une telle distribution est composée de deux processus qui sont estimés en même temps. Le premier concerne la partie des zéros structurels qui est représentée par une variable binaire α d'être un zéro structurel ou non et modélisée par un modèle logit. Le second processus concerne les données, autres que les zéros structurels, qui sont représentées par une distribution de Poisson associée à son paramètre λ défini par un modèle mixte.

Pour Y_{ij}, l'intensité d'exposition moyenne au cours de l'année j ($j \in \{1, \ldots, n_i\}$) pour le sujet $i \in \{1, \ldots, n\}$,

$$Y_{ij}|_{c_i=g} \sim ZIP(p_{ijg}, \lambda_{ijg})$$

où

— $p_{ijg} = P(\alpha_{ij} = 1|c_i = g)$ représente la probabilité spécifique à la classe g d'être un zéro structurel au temps j. Elle est représentée par α_{ij} qui est l'indicateur binaire qui vaut 1 si Y_{ij} est un zéro structurel (Muthén and Asparouhov [2009]).

Ceci est modélisé par :

$$p_{ijg} = P(\alpha_{ij} = 1 | c_i = g) = \frac{\exp(\varrho_{0g} + M_{ij}^T \varrho_{lg})}{1 + \exp(\varrho_{0g} + M_{ij}^T \varrho_{lg})} \tag{49}$$

- ▷ ϱ_{0g} sont les intercepts spécifiques à chaque classe ;
- ▷ M_{ij}^T est le vecteur des covariables associé au vecteur des coefficients spécifiques à chaque classe ζ_{lg}

— $\lambda_{ijg} = E(Y_{ij}|_{c_i=g})$ définie par

$$\ln(\lambda_{ijg}) = \beta_{0g} + X_{ij}^T \beta_{lg} + Z_{ij}^T b_i \tag{50}$$

- ▷ X_{ij}^T est le vecteur des covariables associé au vecteur des coefficients des effets fixes β_{lg} spécifiques à chaque classe ;
- ▷ Z_{ij}^T est le vecteur des covariables associé au vecteur des effets aléatoires b_i.
  On suppose que $b_i|_{c_i=g} \sim \mathcal{N}(0, \sigma_{0g}^2 Br)$ avec σ_{0g}^2 les coefficients de proportionnalité associés à la matrice de variance-covariance Br. Par soucis d'identifiabilité, $\sigma_{0G}^2 = 1$.

Ce modèle proposé est différent de celui introduit par Muthén and Asparouhov [2009] et, par conséquent, de celui implémenté dans le logiciel MPlus. Ici, nous avons voulu que la probabilité d'être un zéro structurel puisse être différente d'une classe latente à l'autre. En effet, il est important que la probabilité d'être en dehors ou non de l'histoire d'exposition du sujet en chaque temps puisse varier d'une classe à l'autre, puisque l'on a vu dans les chapitres 2 et 3 que les périodes d'exposition jouent un rôle important dans le risque de cancer.

4.2.2 Estimation par maximum de vraisemblance

Le modèle ZIP-LCMM peut être estimé par maximum de vraisemblance. Pour rappel, dans un tel modèle, la vraisemblance individuelle du modèle, représentée par la densité des mesures répétées $f(Y_i)$ de l'individu i, s'écrit conditionnellement aux effets aléatoires et aux classes latentes. Pour la calculer, l'ordre selon lequel on conditionne n'a pas d'influence sur les résultats obtenus. Cependant, l'ordre est important du point de vue de l'implémentation. En effet, il va jouer sur la dimension des matrices manipulées et donc influencer la rapidité du temps de calcul pour obtenir la convergence du modèle.

Ici, il y a deux choix possibles pour le conditionnement :

1- Soit $f(Y_i)$ est conditionnée sur les classes puis sur les effets aléatoires :

$$f(Y_i) = \sum_{g=1}^{G} f(Y_i|_{c_i=g}) f(c_i = g) \tag{51}$$

$$= \sum_{g=1}^{G} \int f(Y_i|_{c_i=g,b_i}) f(b_i|_{c_i=g}) f(c_i = g) db_i \tag{52}$$

$$= \sum_{g=1}^{G} \int f(Y_i|_{c_i=g,b_i}) \overbrace{f(b_i|_{c_i=g}) f(c_i = g)}^{f(b_i,c_i=g)} db_i \tag{53}$$

2- Soit $f(Y_i)$ est conditionnée sur les effets aléatoires puis sur les classes latentes :

$$f(Y_i) = \int f(Y_i|_{b_i}) f(b_i) db_i \tag{54}$$

$$= \int \{\sum_{g=1}^{G} f(Y_i|_{b_i,c_i=g}) f(c_i = g|_{b_i})\} f(b) db_i \tag{55}$$

$$= \int \sum_{g=1}^{G} f(Y_i|_{b_i,c_i=g}) \overbrace{f(c_i = g|_{b_i}) f(b_i)}^{f(b_i,c_i=g)} db_i \tag{56}$$

$$= \int \sum_{g=1}^{G} f(Y_i|_{b_i,c_i=g}) f(b_i|_{c_i=g}) f(c_i = g) db_i \tag{57}$$

On voit que nous avons équivalence dans l'écriture de la vraisemblance quelque soit l'ordre du conditionnement puisque la densité jointe $f(b_i, c_i = g)$ peut s'écrire de deux manières équivalentes. Cependant, dans le cas de Q effets aléatoires, avec le premier conditionnement, on a GxQ intégrales à calculer pour un même individu alors qu'avec le second on a seulement Q intégrales.

Dans la suite de ce travail, nous avons donc considéré le second conditionnement, ce qui nous a permis d'écrire la contribution individuelle comme suit, avec $f(c_i = g) = \pi_{ig}$:

$$f(Y_i) = \int \sum_{g=1}^{G} \pi_{ig} f(b_i|_{c_i=g}) f(Y_i|_{b_i,c_i=g}) db_i \tag{58}$$

$$= \int \sum_{g=1}^{G} \pi_{ig} f(b_i|_{c_i=g}) \prod_{j=1}^{n_i} f(Y_{ij}|_{b_i,c_i=g}) db_i \tag{59}$$

La densité de la distribution ZIP est une combinaison linéaire de la densité associée au modèle logit et de celle associée au modèle mixte ZIP spécifique à chaque classe latente (Roeder et al. [1999]). Elle s'écrit comme suit :

$$f(Y_{ij}|_{b_i,c_i=g}) = P(\alpha_{ij}=0|_{c_i=g}) f(Y_{ij}|_{\alpha_{ij}=0,b_i,c_i=g}) + P(\alpha_{ij}=1|_{c_i=g}) f(Y_{ij}|_{\alpha_{ij}=1,b_i,c_i=g}) \quad (60)$$

$$= (1-p_{ijg}) \frac{\exp(-\lambda_{ijg})\lambda_{ijg}^{Y_{ij}}}{Y_{ij}!} + p_{ijg}\mathbb{1}_{Y_{ij}=0} \quad (61)$$

Ainsi, on peut écrire la contribution individuelle de la manière suivante :

$$\mathrm{lv}_i = f(Y_i) = \int \sum_{g=1}^{G} \pi_{ig} f(b_i|_{c_i=g}) \prod_{j=1}^{n_i} \{(1-p_{ijg}) \frac{\exp(-\lambda_{ijg})\lambda_{ijg}^{Y_{ij}}}{Y_{ij}!} + p_{ijg}\mathbb{1}_{Y_{ij}=0}\} dbi \quad (62)$$

Le calcul de la log-vraisemblance totale ($\sum_{i=1}^{n} \log \mathrm{lv}_i$) requiert donc le calcul d'une intégrale sur les effets aléatoires. La dimension de cette intégrale est égale au nombre d'effets aléatoires considérés dans le modèle. Ici, la vraisemblance individuelle (équation 62) n'a pas de solution analytique. Ceci implique l'utilisation d'une approche numérique pour approximer cette intégrale pour chaque sujet dans la log-vraisemblance totale.

4.2.3 Intégration numérique sur les effets aléatoires

4.2.3.1 Généralités sur les approximations existantes

Il existe trois méthodes pour l'intégration numérique : la quadrature gaussienne, l'approximation de Laplace et les simulations de Monte-Carlo. Avec l'approximation de Laplace, l'intégrant est approché par un développement de Taylor autour du mode pour avoir un calcul analytique de l'intégrale (Wolfinger [1993]). Dans la suite de ce chapitre, le mode est défini comme le maximum de la fonction de densité de l'effet aléatoire. Le temps de calcul de cette approximation est acceptable mais la précision est moins bonne qu'avec une approximation par quadrature gaussienne. L'approximation par Monte Carlo consiste à approcher l'intégrale par une somme de l'intégrant calculé en des points simulés suivant la loi a priori des effets aléatoires. Le nombre de points nécessite d'être très grand pour avoir une bonne précision. Or, si le nombre de points simulés est trop grand, le temps de calcul devient considérable.

L'approximation par quadrature gaussienne est la plus couramment utilisée quand le nombre d'effets aléatoires n'est pas trop grand car elle est un bon compromis entre précision et rapidité. L'intégrale est approchée par une somme de l'intégrant calculé en des points prédéfinis (appelés points de quadratures) et pondéré par des poids dont les valeurs ont été calculées en fonction

du type de l'intégrale. Si l'effet aléatoire est considéré gaussien alors les abscisses des points de quadrature et les poids associés sont calculés en utilisant la règle de Gauss-Hermite (Abramowitz and Stegun [1972]). Le nombre fixé des points de quadrature détermine la précision de l'approximation. En effet, ces points de quadrature doivent être positionnés de façon à être répartis correctement par rapport à la densité de l'effet aléatoire et plus précisément autour du point de masse 0 de l'intégrant puisqu'on fait l'hypothèse que l'effet aléatoire b_i suit une loi gaussienne de moyenne nulle. De plus, ces points doivent également être répartis autour de ce point de masse 0 avec la même dispersion que l'intégrant. Lesaffre and Spiessens [2001] ont très bien illustré ce problème avec une figure qui se trouve ci-après. Sur la figure 4.1 a), on constate que, dû à la position des points de quadratures calculés de manière prédéfinie ((Abramowitz and Stegun [1972])), le mode de l'intégrant n'est pas correctement situé (décalé sur la droite) par rapport à la densité associée. De plus, on voit que les points sont étendus entre -6 et 6 alors que la densité se disperse plutôt entre -1 et 1. Pour évaluer précisément l'intégrale, il faut donc considérer de nombreux points de quadrature ce qui va augmenter le temps de calcul. Lesaffre and Spiessens [2001] ont donc proposé une solution à ce problème en introduisant la quadrature gaussienne adaptative qui va permettre une approximation plus précise. En effet, le but est de centrer et mettre à l'échelle ("rescaler") les points de quadrature autour de la prédiction de b_i à chaque itération. Ainsi, les nouveaux points de quadrature sont répartis autour du mode de la densité avec la même dispersion, comme on peut le constater sur la figure 4.1 b).

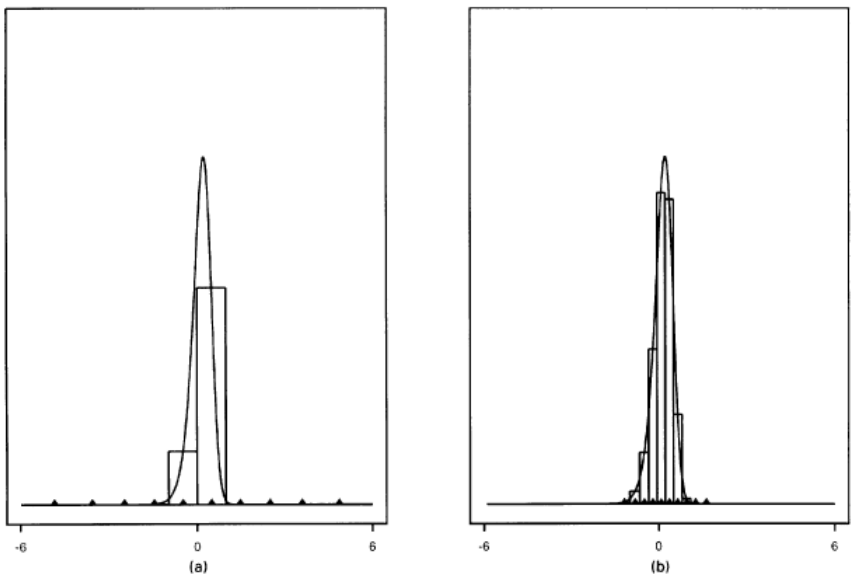

FIGURE 4.1 – Comparaison de la position de 10 points de quadrature obtenu par a) une quadrature gaussienne oridinaire et b) une quadrature gaussienne adaptative pour le même intégrant. Les triangles sur l'axe des abscisses représentent la position des points de quadrature. *Figure tirée de Lesaffre and Spiessens [2001]*

L'avantage de cette approche est que l'intégrale peut être approximée avec une meilleure précision qu'avec l'approche non adaptative, avec moins de points de quadrature puisque les points seront mieux positionnés. Cependant, le problème avec cette quadrature gaussienne adaptative

est qu'elle requiert la localisation du mode, $\hat{b}_i$, et le calcul de la matrice hessienne, $\hat{H}_i$, pour chaque individu, impliquant ainsi d'autres calculs d'intégrales. De plus, ceci doit être fait à chaque itération de l'algorithme d'optimisation, cela augmente donc considérablement le temps de calcul.

Rizopoulos [2012] a proposé une solution pour diminuer ce temps de calcul en introduisant la quadrature gaussienne pseudo-adaptative pour l'estimation de son modèle conjoint à effets aléatoires partagés. Au lieu de recentrer et remettre à l'échelle l'intégrale autour de la prédiction $\hat{b}_i$ pour chaque sujet à chaque itération, l'intégrale est recentrée et remise à l'échelle une seule fois au début de l'optimisation à partir d'informations provenant d'un modèle statistique plus simple, dans son cas, le modèle mixte pour les données répétées. De manière pratique, les points de quadrature vont donc être redéfinis par rapport aux estimations des $\hat{b}_i$ et $\hat{H}_i$ provenant du modèle linéaire mixte. L'efficacité de cette adaptation a été évaluée par simulations en la comparant à l'approche non adaptative (Rizopoulos [2012]). Les résultats étaient très proches entre l'approche pseudo-adaptative avec 3 points de quadrature et l'approche non adaptative avec 15 points de quadrature. De plus, le temps moyen pour estimer le modèle conjoint était plus faible avec l'approche pseudo-adaptative par rapport à la non-adaptative (de 251.3 secondes à 13.5 secondes).

La qualité de l'approximation de l'intégrale par l'approche pseudo-adaptative réside dans les estimations des prédictions individuelles $\hat{b}_i$ et de leur matrice de variance-covariance associée $\hat{H}_i$ qui doivent permettre d'obtenir des points de quadrature correctement localisés et dispersés. Une solution pour améliorer la précision de l'approximation peut être de mettre en place une stratégie d'estimation similaire à celle proposée par Ferrer et al. [2016] nommée "multi-step pseudo-adaptative gauss hermite rule". L'idée est de ré-estimer une n-ème fois le même modèle en adaptant les points de quadrature à partir des estimations obtenues à la $(n-1)$-ème estimation du modèle. Ceci permet d'éviter de choisir un trop grand nombre de points de quadrature tout en ayant des estimations correctes pour le modèle. Dès n=2 étapes, Ferrer et al. [2016] ont montré de bons résultats dans les modèles conjoints à effets aléatoires partagés.

Au vu de ses avantages, nous avons choisi d'utiliser l'approche quadrature gaussienne pseudo-adaptative en deux étapes (2-step pseudo-adaptative gauss hermite rule) pour approximer l'intégrale de chaque sujet dans la log-vraisemblance de notre modèle ZIP-LCMM. Cependant, cette approche a été initialement développée pour l'estimation d'un modèle conjoint à effets aléatoires partagés, la stratégie d'estimation a ainsi été adaptée au contexte du modèle à classes latentes et est présentée en section 4.2.5.

4.2.3.2 Détail de l'approximation choisie

En utilisant une approximation de quadrature gaussienne non adaptative, on pourrait réécrire la contribution individuelle du modèle ZIP-LCMM comme :

$$\begin{aligned} f(Y_i) &= \int_{\mathbb{R}^q} \sum_{g=1}^{G} \pi_{ig} f(b_i|_{c_i=g}) \prod_{j=1}^{n_i} f(Y_{ij}|_{b_i,c_i=g}) dbi \\ &\approx \sum_{k_1..k_q} \sum_{g=1}^{G} \pi_{ig} w_k f(b_k) \{ \prod_{j=1}^{n_i} f(Y_{ij}|_{b_i=b_k,c_i=g}) \} \end{aligned}$$

où

— $\sum_{k_1..k_q} = \sum_{k_1=1}^{K} .. \sum_{k_q=1}^{K}$ est le produit cartésien avec K points de quadrature et q effets aléatoires ;

— w_k sont les poids associés à b_k qui sont les abscisses des points de quadratures (définis par Abramowitz and Stegun [1972]) ;

— π_{ig} est la probabilité d'appartenance à la classe g pour l'individu i définie par l'équation 48 ;

— $f(b_k)$ est la densité gaussienne correspondante au point b_k ;

— $f(Y_{ij}|_{b_i=b_k,c_i=g})$ est la densité donnée à l'équation 60 au point b_k

En utilisant l'approche de quadrature gaussienne pseudo-adaptative basée sur Rizopoulos [2012], nous allons définir les nouveaux points de quadrature :

$$r_{ik} = \tilde{b}_i + \sqrt{2}\tilde{B}_i^{-1} b_k \tag{63}$$

où

— b_k sont les abscisses des points de quadratures (définis par Abramowitz and Stegun [1972]) ;

— $\tilde{b}_i$, $\tilde{H}_i$ sont les estimations individuelles des effets aléatoires et de la matrice de variance-covariance associée obtenues à partir d'un modèle statistique plus simple ;

— $\tilde{B}_i$ est le facteur de Cholesky de la matrice de variance-covariance $\tilde{H}_i$

Les poids associés aux nouveaux points de quadrature se réécrivent comme suit (Rizopoulos [2012]) :

$$w_k^{'} = w_k 2^{q/2} exp(||b_k||^2)|\tilde{B}_i|^{-1} \tag{64}$$

où

— w_k sont les poids associés aux points de quadrature b_k calculés avec la règle de Gauss-Hermite ;

— q est le nombre d'effets aléatoires ;

— $|\tilde{B}_i|^{-1}$ est le déterminant du facteur de Cholesky $\tilde{B}_i$

Avec (63) et (64), la contribution individuelle approximée peut s'écrire comme suit :

$$f(Y_i) = \sum_{k_1 \ldots k_q} \sum_{g=1}^{G} \pi_{ig} w_k^{'} f(r_{ik}) \{ \prod_{j=1}^{n_i} f(Y_{ij}|_{b_i = r_{ik}, c_i = g}) \}$$

De manière plus détaillée, on peut donc réécrire la vraisemblance individuelle du modèle ZIP-LCMM :

$$f(Y_i) = \sum_{k_1 \ldots k_q} w_k 2^{q/2} \exp(||b_k||^2)|\tilde{B}_i|^{-1} \sum_{g=1}^{G} \pi_{ig} f(r_{ik}) \{ \prod_{j=1}^{n_i} f(Y_{ij}|_{b_i = r_{ik}, c_i = g}) \} \tag{65}$$

4.2.4 Algorithme d'optimisation

Comme pour l'estimation du modèle conjoint mixte à classes latentes (chapitre 3), nous avons utilisé l'algorithme itératif d'optimisation de Marquardt. Ce dernier a été détaillé dans le chapitre 3. Pour l'implémentation sous le logiciel R, nous avons utilisé une version adaptée par Viviane Philipps du package R *MarqLevAlg*, initialement développé par Mélanie Prague et Daniel Commenges (Prague et al. [2012]). La fonction R adaptée que nous avons utilisée, nommée *MarqLevAlgParallel*, permet une meilleure estimation des écart-types des paramètres du modèle grâce à l'ajout d'un critère de convergence lié à l'inversibilité de la matrice Hessienne. Grâce à la parallélisation, elle est également plus efficace en temps de calcul.

4.2.5 Stratégies d'estimation

4.2.5.1 Règle de quadrature gaussienne pseudo adaptative en 2 étapes

Pour estimer le modèle conjoint à effets aléatoires partagés avec la règle de quadrature gaussienne pseudo-adaptative, Rizopoulos [2012] a utilisé l'information provenant de l'estimation du modèle mixte pour définir ses points de quadrature. Nous avons adapté ceci au contexte du modèle à classes latentes. Ainsi, pour l'estimation du modèle à G classes latentes, on a utilisé l'information provenant du modèle à $G-1$ classes latentes. Cependant, pour être plus précis, nous avons estimé le modèle avec la règle de quadrature gaussienne pseudo-adaptative en deux étapes. Ce qui se traduit par :

— 1ère étape : calcul des points de quadrature (r_{ik}) à partir des estimations des effets aléatoires marginaux aux classes $\tilde{b}_{iM}$ et de la matrice de variance-covariance associée $\tilde{H}_{iM}$ obtenus avec le modèle à $G-1$ classes (équations 63 & 64) ;

— 2ème étape : nouveau calcul des points de quadrature à partir des estimations des effets aléatoires spécifiques aux classes $\tilde{b}_{ig}$ et de la matrice de variance-covariance associée $\tilde{H}_{ig}$ obtenues avec le modèle à G classes estimé à la première étape. Dans cette deuxième étape, les points de quadrature et leurs poids associés deviennent donc spécifiques à chaque classe g. On redéfinit donc r_{ikg} et $w^{'}_{kg}$ comme suit :

$$r_{ikg} = \tilde{b}_{ig} + \sqrt{2}\tilde{B}_{ig}^{-1} b_k \tag{66}$$

$$w^{'}_{kg} = w_k 2^{q/2} \exp(||b_k||^2)|\tilde{B}_{ig}|^{-1} \tag{67}$$

4.2.5.2 Estimation des effets aléatoires et de leur matrice de variance-covariance associée

Effets aléatoires prédits spécifiques à la classe ($\tilde{b}_{ig}$)

Pour pouvoir déterminer les points de quadrature, nous avons besoin d'estimer les effets aléatoires individuels. De manière générale, ils sont estimés par l'espérance de leur distribution a posteriori $E(b_i|Y_i = y_i, \hat{\theta})$ où $\hat{\theta}$ est le vecteur des paramètres estimés. Pour le modèle linéaire mixte, la distribution a posteriori a une forme analytique utilisant les propriétés de la loi normale multivariée. Ainsi est obtenu l'estimateur bayésien empirique appelé BLUP (*Best Linear Unbiased Predictor*). Cependant, pour un modèle mixte ZIP, il n'y a pas de forme analytique de l'espérance de la distribution $f(b_i|_{Y_i})$.

Pour pouvoir la calculer, on utilise la relation générale suivante :

$$f_{b_i|_{Y_i}}(b_i|_{Y_i}) = \frac{f_{Y_i|_{b_i}}(Y_i|_{b_i}) f_{b_i}(b_i)}{f_{Y_i}(Y_i)} \approx f_{Y_i|_{b_i}}(Y_i|_{b_i}) f_{b_i}(b_i) \tag{68}$$

Ainsi, l'espérance $E(b_i|Y_i = y_i, \hat{\theta})$ est donc approchée par le mode de cette fonction (équation 68) au point des estimations des paramètres $\hat{\theta}$.

Pour la deuxième étape de l'estimation du modèle ZIP-LCMM, nous obtenons ainsi les effets aléatoires prédits spécifiques à la classe $\tilde{b}_{ig}$ et leur matrice de variance-covariance associée $\tilde{H}_{ig}$ en maximisant l'équation 68 propre pour chaque classe g par l'algorithme Marquardt.

Effets aléatoires prédits marginaux ($\tilde{b}_{iM}$)

Pour la première étape de l'estimation du modèle ZIP-LCMM, nous utilisons les effets aléatoires marginaux aux classes. De manière intuitive, cet effet aléatoire aurait pu être calculé de manière séquentielle en utilisant les $\tilde{b}_{ig}$ obtenus avec la maximisation ci-dessus puis en les pondérant par les probabilités d'appartenance aux classes $\hat{\pi}_{ig}$. Le principal problème réside dans le calcul de matrice de variance-covariance associée. En effet, la variance de cette somme pondérée des effets aléatoires prédits spécifiques à chaque classe n'est égale à la somme pondérée des variances seulement sous l'hypothèse d'indépendance entre ces variances. Cette hypothèse semble forte et pourrait impacter nos résultats. Nous avons donc préféré calculer $\tilde{b}_{iM}$ et $\tilde{H}_{iM}$ en une seule étape en maximisant la fonction qui pouvait approcher cette distribution.

Cette fonction peut être obtenue via le calcul suivant :

$$\begin{aligned}
f(b_i|_{Y_i}) =& \frac{f(b_i, Y_i)}{f(Y_i)} && (69)\\
=& \frac{1}{f(Y_i)} \sum_{g=1}^{G} f(b_i, Y_i|_{c_i=g}) f(c_i = g) && (70)\\
=& \frac{1}{f(Y_i)} \sum_{g=1}^{G} f(b_i|_{c_i=g}) f(Y_i|_{b_i,c_i=g}) f(c_i = g) && (71)\\
\approx& \sum_{g=1}^{G} f(b_i|_{c_i=g}) f(Y_i|_{b_i,c_i=g}) f(c_i = g) && (72)
\end{aligned}$$

Comme pour les effets aléatoires prédits spécifiques à la classe, nous obtenons les effets aléatoires prédits marginaux et leur matrice de variance covariance associée en maximisant l'équation 71 au point des estimations des paramètres $\hat{\theta}$ par l'algorithme Marquardt.

Par ailleurs, de manière similaire au modèle LCMM, le modèle ZIP-LCMM doit être estimé à partir de différentes combinaisons de valeurs initiales des paramètres pour éviter la convergence vers un maximum local due à la multimodalité de la vraisemblance.

4.2.6 Classification a posteriori & Sélection du modèle

Une fois l'estimation réalisée, nous pouvons réaliser une classification a posteriori des sujets. La probabilité a posteriori $\hat{\pi}_{ig}$ que l'individu i appartienne à la classe latente g peut être calculée de manière similaire à LCMM.

$$\hat{\pi}_{ig} = P(c_i = g|Y_i, \hat{\theta}_G) = \frac{\hat{\pi}_{ig}\phi_{ig}(Y_i, \hat{\theta}_G)}{\sum_{l=1}^{G} \hat{\pi}_{il}\phi_{il}(Y_i, \hat{\theta}_G)} \tag{73}$$

où

— $\hat{\theta}_G$ est le vecteur des paramètres estimés ;

— $\phi_{ig}(Y_i, \hat{\theta}_G)$ est la densité de la distribution ZIP associée à la classe g calculée suivant les estimations des paramètres, précédemment détaillée (équation 60)

Pour déterminer le nombre optimal de classes latentes, nous avons utilisé les mêmes stratégies que celles présentées dans le chapitre précédent. A savoir, la comparaison des valeurs du BIC de modèles ayant un nombre différent de classes latentes, la pertinence des classes identifiées suivant les trajectoires moyennes prédites et les effectifs associés, la capacité discriminante à partir de la table de classification a posteriori.

4.2.7 Implémentation sous R

Le second objectif de ce travail était d'implémenter le ZIP-LCMM sous le logiciel R. Pour cela, nous avons pu utiliser des fonctions R déjà programmées comme *MarqLevAlgParallel* pour l'algorithme d'optimisation, *gaussher* pour le calcul de points de quadrature suivant la règle de Gauss-Hermite, *dmvnorm* pour la densité multivariée gaussienne. Par ailleurs, nous avons dû implémenter la définition des points de quadrature adaptée pour les classes latentes, la densité de la distribution ZIP, la log vraisemblance du modèle ZIP-LCMM, les calculs des prédictions individuelles et des trajectoires moyennes prédites ainsi que tout ce qui était lié à la classification a posteriori.

4.3 Application à la relation tabac - cancer du poumon

Nous avons identifié les trajectoires d'intensité de consommation de tabac par l'estimation du modèle ZIP-LCMM. Ensuite, nous avons estimé leur association avec le risque de cancer du poumon en utilisant une des approches en 2 étapes présentées au chapitre 3.

4.3.1 Spécification du modèle ZIP-LCMM

Nous allons préciser l'écriture des 3 sous-modèles considérés pour notre application. Dans le sous-modèle concernant la probabilité d'appartenance aux classes, nous avons choisi de ne considérer aucune covariable. Pour le sous-modèle représentant les zéros stucturels, on a choisi que la probabilité d'être un de ces zéros dépende d'un intercept et du temps auquel est associé ce zéro. Ainsi, nous avons pu évaluer l'évolution de la probabilité d'être un zéro structurel selon le temps avant la date index en fonction des différentes classes latentes. Et enfin, pour le sous-modèle mixte ZIP, nous avons, pour l'instant et pour simplifier, considéré une évolution linéaire en tenant compte de la corrélation entre les mesures répétées d'un même sujet au sein de la classe grâce à un intercept aléatoire. Une évolution spline des intensités sera considérée dans un travail futur.

Modèle multinomial logistique pour l'appartenance aux classes latentes :

$$\pi_{ig} = P(c_i = g) = \frac{\exp(\zeta_{0g})}{\sum_{l=1}^{G} \exp(\zeta_{0l})}$$

où ζ_{0g} est l'intercept spécifique à la classe latente g (avec ζ_{0G} =0) et c_i est la variable latente discrète égale à g si l'individu i appartient à la classe g.

Modèle pour la probabilité d'être un zéro structurel (α_{ij}) :

$$p_{ijg} = P(\alpha_{ij} = 1) = \frac{\exp(\varrho_{0g} + \varrho_{1g} t_{ij})}{1 + \exp(\varrho_{0g} + \varrho_{1g} t_{ij})}$$

où

— γ_{0g} est l'intercept spécifique à la classe g ;

— t_{ij} est la $j^{ème}$ année avant la date index associé au coefficient fixe spécifique à la classe γ_{1g}.

Modèle mixte ZIP spécifique à la classe latente g :

$$\ln(\lambda_{ijg}) = \beta_{0g} + \beta_{1g} t_{ij} + b_{0i}$$

où

— β_{0g} est l'intercept spécifique à la classe g ;

— t_{ij} est la $j^{ème}$ année avant la date index associé au coefficient fixe spécifique à la classe β_{1g} ;

— b_{0i} : intercept aléatoire avec $b_{0i}|_{c_i=g} \sim \mathcal{N}(0, w_g^2\sigma^2)$ où w_g est le coefficient de proportionnalité propre à g (avec $w_G = 1$) et σ^2 la variance commune. L'intercept aléatoire est défini sur toutes les années de l'axe du temps, et pas seulement sur l'histoire d'exposition du sujet comme pour le modèle JLCMM du chapitre 3.

Après estimation du modèle spécifié, nous avons calculé les trajectoires moyennes prédites associées aux classes latentes en utilisant la formule suivante :

$$\begin{aligned}
E(Y_{ij}|_{b_{0i},c_i=g}; \hat{\theta}_G) =& P(\alpha_{ij} = 1|_{c_i=g}; \hat{\theta}_G) E(Y_{ij}|_{\alpha_{ij}=1,b_{0i},c_i=g}; \hat{\theta}_G) + \\
& (1 - P(\alpha_{ij} = 1|_{c_i=g}; \hat{\theta}_G) E(Y_{ij}|_{\alpha_{ij}=0,b_{0i},c_i=g}; \hat{\theta}_G) \\
=& 0 + (1 - p_{ij}) E(\lambda_{ijg}; \hat{\theta}_G) \\
=& (1 - p_{ij}) \int_{b_{0i}} \exp(\beta_{0g} + \beta_{1g} t_{ij}) \exp(b_{0i}) f(b_{0i}) db_{0i} \\
=& (1 - p_{ij}) \{\exp(\beta_{0g} + \beta_{1g} t_{ij}) \int_{b_{0i}} \exp(b_{0i}) f(b_{0i}) db_{0i}\}
\end{aligned}$$

Nous avons utilisé la fonction R *integrate* pour approximer l'intégrale sur l'intercept aléatoire. Cette trajectoire était donc calculée à partir des estimations des paramètres et pour un vecteur de temps de mesure allant de 0 à 63.

4.3.2 Résultats

Nous avons estimé les modèles de 1 à 4 classes latentes en utilisant 6 points de quadrature pour l'approximation de l'intégrale sur l'intercept aléatoire. La table ci-après résume les résultats selon le BIC et les effectifs des sujets classés a posteriori dans chaque classe.

G	Nombre de paramètres	BIC	Effectifs dans chaque classe
1	5	1167301	Classe 1 : 3807 (100%)
2	11	985940	Classe 1 : 1391 (36.5%) Classe 2 : 2416 (63.5%)
3	17	966214	Classe 1 : 1265 (33.2%) Classe 2 : 1610 (42.3%) Classe 3 : 932 (24.5%)
4	23	949301	Classe 1 : 891 (23.4%) Classe 2 : 1212 (31.8%) Classe 3 : 806 (21.2%) Classe 4 : 898 (23.6%)

Tableau 4.1 – Sélection du modèle ZIP-LCMM

Le modèle ZIP-LCMM avec 4 classes latentes est celui qui avait le BIC le plus faible avec des effectifs suffisants dans chaque classe (tableau 4.1). Ce modèle avait également une bonne capacité discriminante comme le montre la table de classification a posteriori (tableau 4.2). Les probabilités moyennes des sujets a posteriori classés dans chaque classe étaient en effet supérieures à 0.965 pour chaque classe (cf les éléments diagonaux). Nous n'avons pas essayé d'estimer le modèle au-delà de 4 classes latentes, afin de pouvoir, à terme de ce travail, comparer les résultats avec ceux obtenus par le JLCMM présentés dans l'article 3.

	N	Moyenne des probabilités a posteriori d'appartenance à chaque classe			
		1	2	3	4
Classe 1	891	0.9773	0.0005	0.0015	0.0207
Classe 2	1212	<0.0001	0.9742	0.0188	0.0070
Classe 3	806	0.0018	0.0245	0.9654	0.0083
Classe 4	898	0.0115	0.0123	0.0067	0.9696

Tableau 4.2 – Table de classification a posteriori pour le modèle ZIP-LCMM à 4 classes latentes

Les trajectoires moyennes prédites de la consommation de tabac au sein de chaque classe sont montrées en figure 4.2 (gauche). Les probabilités prédites d'être un zéro structurel sont montrées en figure 4.2 (droite).

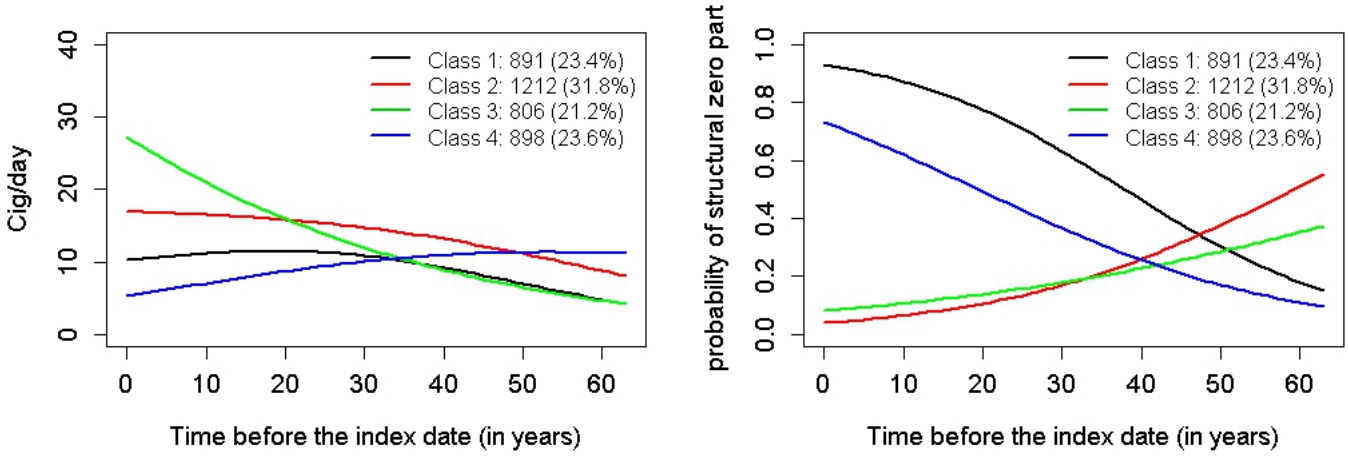

FIGURE 4.2 – Modèle ZIP-LCMM à 4 classes latentes. A gauche, les trajectoires moyennes prédites d'intensité de consommation de tabac au cours du temps avant la date index. A droite, les probabilités prédites d'être un zéro structurel en chaque temps avant la date index.

La classe 2 (trajectoire rouge) est la plus représentée, avec 32 % de sujets classés a posteriori dans cette classe. Les intensités de consommation de tabac passent en moyenne de 10 cigarettes par jour 60 ans avant la date index à presque 20 cigarettes par jour à la date index. La classe 1 (trajectoire noire) a un comportement similaire mais on peut constater que les intensités dans les 20 dernières années avant la date index ont tendance à stagner voire diminuer légèrement. La classe 3 (trajectoire verte) est caractérisée par des intensités moyennes de consommation de tabac les plus élevées (plus de 20 cigarettes par jour) dans les 15 dernières années avant la date index. La classe 4 (trajectoire bleue) a des intensités de consommation de tabac qui diminuent de 60 ans avant la date index jusqu'à la date index, avec moins de 10 cigarettes par jour en moyenne dans les 30 ans avant la date index. Ces trajectoires moyennes prédites sont cohérentes avec les probabilités prédites (Figure 4.2, droite). Les probabilités plus élevées pour les années proches de la date index sont pour les classes 1 et 4. Pour les classes 2 et 3, les probabilités prédites sont proches de zéro pour les années proches de la date index.
Les caractéristiques des sujets classés a posteriori dans chaque classe sont montrées dans le tableau 4.3 et sont cohérentes avec les résultats de la Figure 4.2. En effet, les sujets classés a posteriori dans les classes 2 et 3 ont un temps d'arrêt médian de tabac de 1 an seulement et 70 % d'entre eux sont des ex-fumeurs. Les classes 1 et 4 ont, quant à elles, 99% d'ex-fumeurs qui ont arrêté de fumer depuis, respectivement, 28 et 15 ans en médiane. Cependant, on peut trouver surprenant que la classe 1 soit composée de sujets ayant arrêté de fumer, en médiane, 28 ans alors que la trajectoire moyenne prédite a des intensités récentes plus élevées que la

classe 4 qui a un temps d'arrêt médian de 15 ans. Cela s'explique par le fait que les fumeurs courants dans la classe 4 fument en moyenne moins que les fumeurs courants de la classe 1 aux mêmes temps (14 cigarettes par jour versus 26 cigarettes par jour 1 an avant la date index).

L'association entre les classes de trajectoires identifiées et le risque de cancer du poumon a été estimée dans une deuxième étape, par un modèle logistique ajusté sur l'âge à la date index, les départements de résidence, et l'indice cumulé d'exposition professionnelle à l'amiante. Cette étape a été réalisée avec deux méthodes décrites dans le chapitre 3, l'approche naïve et l'approche par Pseudo-classes. Les résultats étaient similaires, et seuls ceux de l'approche par Pseudo-classes sont montrés dans la dernière colonne du tableau 4.3. La classe de référence était la classe 1 (trajectoire noire), avec la plus faible dose cumulée médiane (180 cigarettes-années) et le plus long temps depuis l'arrêt du tabagisme médian (28 ans). Avec des sujets plus jeunes mais qui avaient, en médiane, l'intensité cumulée la plus forte (819 cigarettes-années) et un court temps médian depuis l'arrêt (1 an), la classe 3 avait le risque de cancer du poumon le plus élevé (OR=11.3, 95% IC : 8.9-14.3) par rapport à la classe 1. La classe 4 était plus à risque que la classe 1 (OR=2.4, 95% IC : 2.0-3.1) car les sujets classés a posteriori avaient en médiane davantage fumé tout au long de leur vie (465 versus 180 cigarettes-années) et un temps médian depuis l'arrêt plus court parmi les ex-fumeurs (15 ans versus 28 ans).

	Cas/Témoins	Age à la date index (années)	Cig-années	Durée totale d'exposition (années)	Intensité moyenne vie entière (cig/jour)	Ex-fumeurs	Temps depuis l'arrêt (années)	Age à l'initiation (années)	OR
	n	Médiane ($5^{è} - 95^{è}$ percentile)	Médiane ($5^{è} - 95^{è}$ percentile)	Médiane ($5^{è} - 95^{è}$ percentile)	Médiane ($5^{è} - 95^{è}$ percentile)	n (%)	Médiane ($5^{è} - 95^{è}$ percentile)	Médiane ($5^{è} - 95^{è}$ percentile)	(95% IC)
Classe 1	198/693	61 (44-73)	180 (4-924)	15 (3-35)	15 (0.5-35)	879 (98.6%)	28 (8-45)	18 (13-25)	1-
Classe 2	816/396	59 (43-73)	714 (169-1586)	41 (25-55)	18.8 (4.8-37.2)	872 (71.9%)	1 (1-6)	17 (13-23)	8.2 [6.6 ;10.1]
Classe 3	566/240	55 (41-71)	819 (243-1763)	38 (23-52)	21.7 (8.5-43.4)	606 (75.2%)	1 (1-10)	16 (12-22)	11.3 [8.9 ;14.4]
Classe 4	389/509	65 (44-74)	465 (80-1326)	30 (15-48)	17.7 (3.6-37.2)	885 (98.5%)	15 (3-30)	17 (13-22)	2.4 [2.0 ; 3.1]

Tableau 4.3 – Association entre les trajectoires d'intensité de consommation de tabac et cancer du poumon. OR ajusté sur l'âge à la date index, le département de résidence et l'indice cumulé d'exposition professionnelle à l'amiante.

4.3.3 Comparaison avec les résultats du JLCMM du chapitre 3

Les résultats du modèle JLCMM dans le chapitre 3 indiquaient que la classe avec la trajectoire des intensités très élevées mais les plus distantes était moins à risque par rapport à la classe de référence avec une dose cumulée plus faible mais des intensités récentes plus élevées. Les sujets des deux classes avaient, en médiane, la même durée d'exposition, donc la différence de risque venait principalement du moment auquel intervenait les intensités élevées. Ici, malgré les zéros structurels, les périodes effectives d'exposition sont beaucoup moins visibles sur les trajectoires prédites de chaque classe. En effet, le temps depuis l'arrêt important dans la classe 1 (28 ans en médiane) est mal reflété par la courbe de la trajectoire prédite. Ceci s'explique sans doute en grande partie par le fait d'avoir imposé une trajectoire linéaire et au fait que nous avons imposé un intercept aléatoire sur toutes les années de l'axe et pas seulement sur l'histoire d'exposition. Ces résultats sont donc très préliminaires et à confirmer en faisant évoluer le modèle mixte ZIP considéré.

L'avantage du modèle ZIP-LCMM par rapport au modèle LCMM est qu'il permet de gérer la large proportion de zéros dans les données. De plus, il permet d'estimer les probabilités prédites de ces zéros en chaque temps pour chaque classe que l'on peut analyser en association avec les trajectoires moyennes prédites. Cependant, dans sa forme actuelle, il peut être considéré comme trop restrictif avec un intercept aléatoire appliqué sur toutes les années, y compris celles en dehors de l'histoire d'exposition, et une évolution linéaire. Il faudra donc notamment considérer 1) un intercept aléatoire qui sera évalué sur l'histoire d'exposition du sujet et non sur toute l'étendue de l'axe du temps considéré et 2) une évolution flexible afin de mieux ajuster les trajectoires individuelles.

4.4 Perspectives

Les perspectives à ce travail en cours sont donc les suivantes :

— Considérer un modèle mixte ZIP spécifique à chaque classe latente avec 1) une évolution flexible en utilisant des splines plutôt que l'évolution linéaire considérée dans ces résultats préliminaires ; 2) un intercept aléatoire évalué sur l'histoire d'exposition observé des sujets ; 3) d'autres effets aléatoires sur l'évolution temporelle ;

— Implémenter un modèle conjoint avec un modèle logistique (J-ZIP-LCMM) pour prendre en compte la corrélation entre les données longitudinales et l'occurrence du cancer. On pourra également comparer les résultats avec ceux obtenus avec l'approche en 2 étapes réalisée pour le moment. Cependant, nous avons vu, dans le chapitre précédent, que les OR estimés et leurs intervalles de confiance associés étaient assez proches entre le modèle conjoint et les approches en 2 étapes dans le cadre de nos applications, pouvant

s'expliquer par une capacité discriminante assez élevée. Ici, on peut également constater une telle capacité discriminante au vu de la table de classification a posteriori (table 4.2). Cela, laisse donc à penser que nous obtiendrions des OR (et 95% IC) assez proches entre le J-ZIP-LCMM et l'approche en 2 étapes ;

— Réaliser des simulations afin d'évaluer les performances du modèle et leur impact sur la classification ainsi que la sensibilité des résultats aux nombres de points de quadrature utilisés pour l'approximation de l'intégrale sur les effets aléatoires. Cependant, en utilisant une méthode pseudo-adaptative et une estimation du modèle en deux étapes, nous pouvons penser que les biais des estimations engendrés par cette approximation ne sont pas plus importants que les biais que l'on peut avoir avec une estimation d'un modèle LCMM ;

— Calculer les intervalles de confiance des trajectoires moyennes prédites en utilisant une méthode de Monte-Carlo

4.5 Contribution & Valorisation

4.5.1 Contribution

Dans le cadre de ce dernier travail de thèse, ma contribution personnelle tient en :
— l'intégralité de l'implémentation sous R concernant les points de quadrature adaptée pour les classes latentes, la densité de la distribution ZIP, la log vraisemblance du modèle ZIP-LCMM, les prédictions individuelles, les trajectoires moyennes prédites ainsi que tout ce qui était lié à la classification a posteriori ;
— l'analyse des données à savoir la spécificité du modèle et l'interprétation des résultats.

4.5.2 Valorisation scientifique

Présentation orale

A Zero-Inflated Poisson latent class mixed model for the identification of longitudinal profiles of environmental and occupational exposure intensities in epidemiological studies. The 38th Annual Conference of the International Society for Clinical Biostatistics (ISCB), Melbourne, Australia, August 26-30, 2018.

Article en préparation

Lévêque E, Lacourt A, Leffondré K and Proust-lima C. Zero-Inflated Poisson latent class mixed model for the identification of longitudinal profiles of environmental and occupational exposure intensities in epidemiological studies.

Conclusion générale de la thèse

Cette thèse s'est concentrée sur la problématique de la modélisation de l'intensité des expositions prolongées dans le cadre des études épidémiologiques cas-témoins. Plus particulièrement, elle a permis d'éprouver deux types de méthodes statistiques avancées sur des relations déjà bien documentées, à savoir l'association entre l'intensité du tabagisme et de l'exposition professionnelle à l'amiante, et le risque de mésothéliome pleural et de cancer du poumon. Ces deux méthodes ont, plus spécifiquement, permis de prendre en compte la variation de l'intensité au cours de la vie, d'estimer son effet (le WCIE), et d'identifier les différents profils de trajectoires d'intensité d'exposition vie entière puis de comparer leur risque de cancer associé (le JLCMM). Tandis que la première méthode (WCIE) existait déjà et avait déjà été utilisée dans le contexte des expositions environnementales et professionnelles en étiologie du cancer (Hauptmann et al. [2000b], Hauptmann et al. [2001]), la seconde a très peu été utilisée dans ce contexte et a nécessité des développements nouveaux (le ZIP-LCMM).

Dans l'ensemble, les résultats épidémiologiques de cette thèse apportent de nouveaux éclairages sur l'effet de l'intensité d'exposition au cours de la vie tout en confirmant certains éléments comme le rôle majeur des intensités d'exposition à l'amiante reçues quelques dizaines d'années avant la date du diagnostic pour le MP (Day and Brown [1980]) et l'effet cancérigène du tabac sur le cancer du poumon agissant à la fois au stade précode du processus mais aussi au stade plus avancé du processus (Doll [1978], Brown and Chu [1987]). Notamment, nos résultats montrent la contribution importante des intensités récentes de tabagisme sur le risque de cancer du poumon. Ces résultats tendent à remettre en question la non prise en compte des intensités d'exposition les plus récentes dans le calcul du nombre de paquets-années. De plus, ces résultats remettent aussi en question l'utilisation d'un simple indice cumulé d'exposition attribuant le même poids à l'intensité tout au long de la vie. En plus de fournir des résultats épidémiologiques importants sur les relations étudiées, les deux articles sur le WCIE ont permis de clarifier l'interprétation des résultats obtenus avec cet indice cumulé pondéré, qui reste peu utilisé dans la littérature.

Ce travail comporte plusieurs limites importantes. Premièrement, nous n'avons pas pris en compte la potentielle interaction entre le tabac et l'amiante pour le cancer du poumon. La nature de l'interaction entre ces deux expositions reste encore controversée dans la littérature (Nielsen et al. [2014], El Zoghbi et al. [2017]). Il serait intéressant d'étudier cette interaction avec le WCIE, afin d'étudier comment le poids de l'intensité d'une exposition en chaque temps est potentiellement modifié par l'intensité de l'autre exposition. L'utilisation d'un produit de tenseur splines pourrait être une solution potentielle. En effet, Berhane et al. [2008] ont déve-

loppé une approche en utilisant cet outil mathématique afin de visualiser l'effet de la latence du cancer du poumon pour différents niveaux d'exposition au radon à partir de données d'une cohorte minière d'uranium. Cependant, cela nécessiterait un développement méthodologique important pour adapter leur approche dans le but d'utiliser cet outil plutôt pour étudier une interaction entre deux expositions avec le WCIE à partir de données d'études cas-témoins. De plus, cela aboutirait à des résultats complexes, à la fois en terme d'interprétation et en terme de visualisation (surface en 2D).

L'ajustement sur l'histoire d'exposition professionnelle à l'amiante avec l'indice cumulé d'exposition en catégories pour estimer la relation tabac-cancer du poumon peut paraitre trop approximatif. Une autre perspective que celle évoquée dans la discussion du chapitre 2, serait d'utiliser aussi un WCIE pour représenter l'histoire d'exposition à l'amiante, donc intégrer plusieurs WCIE au sein d'un même modèle de régression. Il serait intéressant d'étudier la faisabilité d'une telle approche d'un point de vue numérique.

Par ailleurs, parmi les autres limites du travail, concernant le cancer du poumon, nous avons seulement pris en compte, comme facteurs de risque, le tabagisme actif et l'exposition professionnelle à l'amiante. Cependant, d'autres facteurs de risque ont été mis en évidence pour causer le cancer du poumon. En effet, à ce jour, le CIRC a classé 13 facteurs en cancérogènes certains (groupe 1) et 4 facteurs en cancérogènes probables (groupe 2A). Il est facilement concevable qu'il soit impossible de tenir compte de tous ces facteurs à la fois au sein du même modèle. Nous avons donc fait le choix de considérer les deux plus importants facteurs de risque dans le développement du cancer du poumon, qui sont également ceux qui ont été largement étudiés dans la littérature.

Une limite supplémentaire importante est la non prise en compte des erreurs de mesures sur les intensités d'exposition. Nous avons utilisé une approximation de l'intensité annuelle moyenne journalière pour le tabac et l'exposition professionnelle à l'amiante provenant de données rétrospectives reportées.

L'histoire tabagique était, en effet, auto-reportée via des questionnaires. Nous ne pouvons donc pas exclure un biais de mémorisation, évoqué en chapitre 2, même si la validité et la fiabilité du mode de recueil de données ont été démontrées pour l'histoire du tabagisme (Huerta et al. [2005], Soulakova et al. [2012]).

L'intensité de l'exposition professionnelle à l'amiante a été estimée par le biais d'une matrice emploi exposition appliquée sur les emplois et activités retracés à travers le calendrier professionnel. Le calendrier professionnel était également auto-reporté via des questionnaires. Un biais de mémorisation ne peut donc pas être exclu, bien que la validité et la fiabilité du mode de recueil de ces données aient également été démontrées dans la littérature (Bouyer et al. [1995], Teschke et al. [2002]). Comme discuté dans le chapitre 2, il a été montré que les matrices emplois expositions peuvent engendrer des erreurs de classification, plutôt considérées comme non différentielles (Kauppinen et al. [1992]). Il serait donc important d'évaluer leur impact sur les

résultats des méthodes utilisées dans cette thèse. Pour le WCIE, Hauptmann et al. [2000b] a montré que la méthode semblait robuste aux erreurs de classification non différentielles, comme on a déjà pu le relever dans le chapitre 2. Néanmoins, un développement intéressant, pour prendre en compte ces erreurs de classification, serait d'adapter les méthodes considérées à l'approche bayésienne, en intégrant dans le modèle, de l'information concernant les erreurs de mesures. Ces approches pourraient permettre, par exemple, d'intégrer en prior l'information fournie par la MEE ou encore d'ajouter un terme d'erreur aléatoire directement à l'intensité d'exposition. Deux articles (Espino-Hernandez et al. [2011], Zhang et al. [2013]) ont proposé de considérer ces erreurs de mesures observées pour les données d'expositions provenant d'études cas-témoins en utilisant une approche bayésienne sur des modèles de régression assez simples. Il serait intéressant d'évaluer ces méthodes sur les modèles considérés au sein de cette thèse.
Par ailleurs, les erreurs de classification engendrées par la MEE peuvent également avoir un impact sur la définition d'un sujet exposé. Dans le cadre de cette thèse, nous avons défini un sujet exposé par une probabilité non nulle d'être exposé à l'amiante durant un emploi. Cependant, il ne peut pas être exclu que les erreurs de classification soient potentiellement plus importantes pour les sujets ayant une probabilité d'exposition relativement faible. Afin de voir l'impact de celles-ci, on pourrait envisager de réaliser des analyses de sensibilité en faisant varier la valeur seuil de la probabilité à partir de laquelle on pourrait considérer un sujet comme exposé.

Dans les données provenant de l'étude cas-témoins ICARE, des sujets étaient très faiblement exposés à l'amiante. Ces très faibles valeurs de l'intensité annuelle moyenne journalière (en équivalent f/mL) proviennent donc du calcul du niveau d'exposition annuel (équation 2 du chapitre 1) réalisé à partir des paramètres d'exposition définis dans la MEE (probabilité, fréquence, intensité). Les intensités liées à l'exposition d'ambiance (passive et indirecte) ont été prises en compte, or elles sont associées à de très faibles valeurs numériques. Dans le cadre du second travail de thèse concernant le JLCMM, ces très faibles valeurs de l'intensité annuelle moyenne journalière ont engendré des problèmes de convergence du modèle. Ainsi, une limite importante du JLCMM considéré dans le chapitre 3 était l'exclusion des sujets très faiblement exposés à l'amiante avec une dose cumulée inférieure à 0.26 f/mL-années, pour résoudre ces problèmes de convergence. Le choix du seuil peut être discutable, même s'il repose sur la VLEP qui est une valeur limite d'exposition professionnelle à l'amiante définie par la législation française. Les développements du JLCMM proposés dans le chapitre 4 pourraient permettre de pallier à l'exclusion de ces sujets, mais ces développements restent encore à valider par simulation.

Enfin, une autre limite importante est que nous avons seulement considéré les sujets qui avaient des histoires d'exposition complètes. Il serait intéressant de pouvoir intégrer les sujets avec des données manquantes pour éviter le problème potentiel de biais de sélection (même si nous avons montré que les sujets exclus ne différaient pas des sujets inclus pour ces données, dans le matériel supplémentaire de l'article 2). Pour cela, il faudrait développer des méthodes

permettant de gérer les données manquantes qui n'augmentent pas considérablement le temps computationnel nécessaire à la convergence des modèles utilisés.

Les approches statistiques considérées dans cette thèse ont été appliquées sur des relations dose-réponse bien documentées. Nous pensons que ce travail peut être vu comme une validation de ces méthodes, pas au sens propre du terme mais plutôt comme une validation auprès de la communauté des épidémiologistes œuvrant dans le domaine de l'étiologie du cancer. Il est, en effet, souvent légitimement attendu que les méthodes statistiques soient d'abord éprouvées sur des relations connues, avant d'être utilisées pour l'analyse de relations davantage controversées. Puisque ce travail de thèse permet à la fois de confirmer des connaissances et donner de nouveaux éclairages sur les relations étudiées, nous espérons qu'il suscitera des développements futurs pour une meilleure analyse des expositions prolongées en étiologie du cancer pour d'autres relations.

Bibliographie

International Standard Industrial Classification of all Economic Activities (Revision 2).New york : United States., 1975. 39

Nomenclature d'activités et de produits françaises NAF-CPF.Paris : Insee, 1999. 39

Nomenclature des professions et catégories socioprofessionnelles PCS. seconde édition.Paris : Insee, 1994. 39

International standard classification of occupations.Geneva. CITP : International Labour Office, 1968. 39

Abramowitz M. and Stegun I. A. *Handbook of Mathematical Functions, 10th printing with corrections.* Dover, 1972. 186, 188

Agudo A., González C. A., Bleda M. J., Ramírez J., Hernández S., López F., Calleja A., Panadès R., Turuguet D., Escolar A., et al. Occupation and risk of malignant pleural mesothelioma : a case–control study in Spain. *American Journal of Industrial Medicine*, 37(2) : 159–168, 2000. 26

Akaike H. A new look at the statistical model identification. *IEEE transactions on automatic control*, 19(6) : 716–723, 1974. 117

Alberg A., Ford J., and Samet J. Epidemiology of lung cancer. *Chest*, 132(3) : 29S–55S, 2007. 28

Armitage P. and Doll R. Stochastic models for carcinogenesis. In *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, Volume 4 : Contributions to Biology and Problems of Medicine*, pages 19–38, Berkeley, Calif., 1961. University of California Press. 30

Asparouhov T. and Muthén B. Auxiliary variables in mixture modeling : Three-step approaches using m plus. *Structural Equation Modeling : A Multidisciplinary Journal*, 21(3) : 329–341, 2014. 123

Atkinson K. E. *An introduction to numerical analysis, 2nd Edition*, volume 528. John Wiley, 1991. 54

Bandeen-Roche K., Miglioretti D. L., Zeger S. L., and Rathouz P. J. Latent variable regression for multiple discrete outcomes. *Journal of the American Statistical Association*, 92(440) : 1375–1386, 1997. 120, 122

Baris I., Artvinli M., Saracci R., Simonato L., Pooley F., Skidmore J., and Wagner C. Epidemiological and environmental evidence of the health effects of exposure to erionite fibres : A four-year study in the cappadocian region of turkey. *International Journal of Cancer*, 39(1) : 10–17, 1987. 23

Baris Y. I. and Grandjean P. Prospective Study of Mesothelioma Mortality in Turkish Villages With Exposure to Fibrous zeolite. *JNCI : Journal of the National Cancer Institute*, 98(6) : 414–417, 2006. 23

Beck F., Guignard R., Richard J. B., Wilquin J.-L., and Peretti-Watel P. Augmentation récente du tabagisme en France : principaux résultats du Baromètre santé, France, 2010. *Bulletin Epidemiologique Hebdomadaire*, pages 20–21, 2011. 29

Belot A., Grosclaude P., Bossard N., Jougla E., Benhamou E., Delafosse P., Guizard A.-V., Molinié F., Danzon A., Bara S., et al. Cancer incidence and mortality in france over the period 1980–2005. *Revue d'épidémiologie et de santé publique*, 56(3) : 159–175, 2008. 22

Benhamou S. and Benhamou E. To the editor, letter. *Epidemiology*, 5(5) : 560, 1994. 30

Berhane K., Hauptmann M., and Langholz B. Using tensor product splines in modeling exposure–time–response relationships : Application to the Colorado Plateau Uranium Miners cohort. *Statistics in Medicine*, 27(26) : 5484–5496, 2008. 201

Bianchi C., Giarelli L., Grandi G., Brollo A., Ramani L., and Zuch C. Latency periods in asbestos-related mesothelioma of the pleura. *European Journal of Cancer Prevention*, 6(2) : 162–166, 1997. 25

Bianchi C. and Bianchi T. Global mesothelioma epidemic : Trend and features. *Indian Journal of Occupational and Environmental Medicine*, 18(2) : 82–88, 2014. 22

Binder-Foucard F., Belot A., Delafosse P., Remontet L., Woronoff A., Bossard N., et al. Estimation nationale de l'incidence et de la mortalité par cancer en France entre 1980 et 2012, 2013. 28

Binder-Foucard F., Bossard N., Delafosse P., Belot A., Woronoff A., and Remontet L. e. a. Cancer incidence and mortality in France over the 1980-2012 period : solid tumors. *Revue Epidemiologie de Santé Publique*, 62(2) : 95–108, 2014. 28

Blanche P., Proust-Lima C., Loubère L., Berr C., Dartigues J.-F., and Jacqmin-Gadda H. Quantifying and comparing dynamic predictive accuracy of joint models for longitudinal marker and time-to-event in presence of censoring and competing risks. *Biometrics*, 71(1) : 102–113, 2015. 124

Böhning D. *Computer-assisted analysis of mixtures and applications : meta-analysis, disease mapping and others*, volume 81. CRC press, 1999. 109

Bolck A., Croon M., and Hagenaars J. Estimating latent structure models with categorical variables : One-step versus three-step estimators. *Political Analysis*, 12(1) : 3–27, 2004. 123

Boucquemont J., Loubère L., Metzger M., Combe C., Stengel B., Leffondre K., and Group N. S. Identifying subgroups of renal function trajectories. *Nephrology Dialysis Transplantation*, 32 (suppl_2) : ii185–ii193, 2017. 106

Bouyer J., Dardenne J., and Hémon D. Performance of odds ratios obtained with a job-exposure matrix and individual exposure assessment with special reference to misclassification errors. *Scandinavian Journal of Work, Environment & Health*, pages 265–271, 1995. 202

Bray B. C., Lanza S. T., and Tan X. Eliminating bias in classify-analyze approaches for latent class analysis. *Structural equation modeling : a multidisciplinary journal*, 22(1) : 1–11, 2015. 122, 123

Brisson C., Loomis D., and Pearce N. Is social class standardisation appropriate in occupational studies ? *Journal of Epidemiology & Community Health*, 41(4) : 290–294, 1987. 95

Brook D. W., Brook J. S., Zhang C., Whiteman M., Cohen P., and Finch S. J. Developmental trajectories of cigarette smoking from adolescence to the early thirties : personality and behavioral risk factors. *Nicotine & Tobacco Research*, 10(8) : 1283–1291, 2008. 106, 118

Brown C. C. and Chu K. C. Use of multistage models to infer stage affected by carcinogenic exposure : example of lung cancer and cigarette smoking. *Journal of Chronic Diseases*, 40 : 171S–179S, 1987. 201

Bruno C., Tumino R., Fazzo L., Cascone G., Cernigliaro A., De Santis M., Giurdanella M. C., Nicita C., Rollo P. C., Scondotto S., et al. Incidence of pleural mesothelioma in a community exposed to fibres with fluoro-edenitic composition in Biancavilla (Sicily, Italy). *Annali dell'Istituto superiore di sanita*, 50 : 111–118, 2014. 23

CDC. State-specific trends in lung cancer incidence and smoking–United States, 1999-2008. *MMWR. Morbidity and mortality weekly report*, 60(36) : 1243, 2011. 28

Chadeau-Hyam M., Tubert-Bitter P., Guihenneuc-Jouyaux C., Campanella G., Richardson S., Vermeulen R., De Iorio M., Galea S., and Vineis P. Dynamics of the risk of smoking-induced lung cancer : a compartmental hidden Markov model for longitudinal analysis. *Epidemiology*, 25(1) : 28–34, 2014. 31

Chen M., Tse L. A., Au R. K., Ignatius T., Wang X.-r., Lao X.-q., and Au J. S.-k. Mesothelioma and lung cancer mortality : a historical cohort study among asbestosis workers in Hong Kong. *Lung Cancer*, 76(2) : 165–170, 2012. 32

Clogg C. C. Latent class models. In *Handbook of statistical modeling for the social and behavioral sciences*, pages 311–359. Springer, 1995. 121

Conti S., Minelli G., Manno V., Iavarone I., Comba P., Scondotto S., and Cernigliaro A. Health impact of the exposure to fibres with fluoro-edenitic composition on the residents in Biancavilla (Sicily, Italy) : mortality and hospitalization from current data. *Annali dell'Istituto superiore di sanita*, 50 : 127–132, 2014. 23

Côté S., Vaillancourt T., LeBlanc J. C., Nagin D. S., and Tremblay R. E. The development of physical aggression from toddlerhood to pre-adolescence : A nation wide longitudinal study of Canadian children. *Journal of Abnormal Child Psychology*, 34(1) : 68–82, 2006. 117

Day N. E. and Brown C. C. Multistage models and primary prevention of cancer. *Journal of the National Cancer Institute*, 64(4) : 977–989, 1980. 201

De Angelis R., Sant M., Coleman M. P., Francisci S., Baili P., Pierannunzio D., Trama A., Visser O., Brenner H., Ardanaz E., et al. Cancer survival in Europe 1999–2007 by country and age : results of EUROCARE-5—a population-based study. *The Lancet Oncology*, 15(1) : 23–34, 2014. 23

De Genna N. M., Goldschmidt L., Day N. L., and Cornelius M. D. Maternal trajectories of cigarette use as a function of maternal age and race. *Addictive behaviors*, 65 : 33–39, 2017. 106, 118

de l'Afsset A. Les fibres courtes et les fibres fines d'amiante. 2009. 24

Delva F., Margery J., Brochard P., Laurent F., Petitprez K., Pairon J., André M., Bessette D., Certin J., Chouaid C., et al. Recommandations de bonne pratique : surveillance médico-professionnelle des travailleurs exposés ou ayant été exposés à des agents cancérogènes pulmonaires. *Archives des Maladies Professionnelles et de l'Environnement*, 77(3) : 579–620, 2016. 28

Dempster A. P., Laird N. M., and Rubin D. B. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B (methodological)*, pages 1–38, 1977. 113, 181

Dennis J. E., Gay D. M., and Welsch R. E. Algorithm 573 : Nl2sol—an adaptive nonlinear least-squares algorithm [e4]. *ACM Transactions on Mathematical Software (TOMS)*, 7(3) : 369–383, 1981. 63, 64

Doll R. Mortality from Lung Cancer in Asbestos Workers. *British Journal of Industrial Medicine*, 12(2) : 81–86, 1955. 32

Doll R. The age distribution of cancer : Implications for models of carcinogenesis. *Journal of the Royal Statistical Society. Series A (General)*, 134(2) : 133–166, 1971. 30

Doll R. and Peto R. Cigarette smoking and bronchial carcinoma : dose and time relationships among regular smokers and lifelong non-smokers. *Journal of Epidemiology and Community Health*, 32(4) : 303–313, 1978. 30

Doll R. An epidemiological perspective of the biology of cancer. *Cancer Research*, 38(11 Part 1) : 3573–3583, 1978. 201

Doll R. and Hill A. B. Smoking and carcinoma of the lung. *British Medical Journal*, 2(4682) : 739, 1950. 30

Doll R. and Peto J. *Effects on health of exposure to asbestos*. Health & Safety Commission, 1985. 25

Dutra L. M., Glantz S. A., Lisha N. E., and Song A. V. Beyond experimentation : Five trajectories of cigarette smoking in a longitudinal sample of youth. *PloS one*, 12(2) : e0171808, 2017. 106, 118

Efron B. and Tibshirani R. J. *An introduction to the bootstrap*. CRC press, 1994. 57

El Zoghbi M., Salameh P., Stücker I., Brochard P., Delva F., and Lacourt A. Absence of multiplicative interactions between occupational lung carcinogens and tobacco smoking : a systematic review involving asbestos, crystalline silica and diesel engine exhaust emissions. *BMC public health*, 17(1) : 156, 2017. 201

Eshghi A., Haughton D., Legrand P., Skaletsky M., and Woolford S. Identifying groups : A comparison of methodologies. *Journal of Data Science*, 9(2) : 271–291, 2011. 107, 119, 120

Espino-Hernandez G., Gustafson P., and Burstyn I. Bayesian adjustment for measurement error in continuous exposures in an individually matched case-control study. *BMC medical research methodology*, 11(1) : 67, 2011. 203

Ferrante D., Mirabelli D., Tunesi S., Terracini B., and Magnani C. Pleural mesothelioma and occupational and non-occupational asbestos exposure : a case-control study with quantitative risk assessment. *Occupational and Environmental Medicine*, 73(3) : 147–153, 2016. 25

Ferrer L., Rondeau V., Dignam J., Pickles T., Jacqmin-Gadda H., and Proust-Lima C. Joint modelling of longitudinal and multi-state processes : application to clinical progressions in prostate cancer. *Statistics in Medicine*, 35(22) : 3933–3948, 2016. 187

Finkelstein M. M. Use of "time windows" to investigate lung cancer latency intervals at an ontario steel plant. *American Journal of Industrial Medicine*, 19(2) : 229–235, 1991. 50, 96

Galateau-Sallé F., Ilg A. G. S., Le Stang N., Brochard P., Pairon J.-C., Astoul P., Frenay C., Blaizot G., Ducamp S., Rousvoal T., et al. Mésothéliome : les dispositifs en place en France «le réseau mésothéliome» 1998–2013. In *Annales de pathologie*, volume 34, pages 51–63. Elsevier, 2014. 22

Gauvin H., Lacourt A., and Leffondré K. On the proportional hazards model for occupational and environmental case-control analyses. *BMC Medical Research Methodology*, 13(1) : 18, 2013. 174

Gay D. M. Algorithm 611 : Subroutines for unconstrained minimization using a model/trust-region approach. *ACM Transactions on Mathematical Software (TOMS)*, 9(4) : 503–524, 1983. 63, 64

Gelbard R., Goldman O., and Spiegler I. Investigating diversity of clustering methods : An empirical comparison. *Data & Knowledge Engineering*, 63(1) : 155–166, 2007. 119, 120

Genolini C. and Falissard B. Kml : A package to cluster longitudinal data. *Computer Methods and Programs in Biomedicine*, 104(3) : e112–e121, 2011. 108, 119

Genolini C., Alacoque X., Sentenac M., Arnaud C., et al. kml and kml3d : R packages to cluster longitudinal data. *Journal of Statistical Software*, 65(4) : 1–34, 2015. 108

Gilg Soit Ilg A., Goldberg M., Rolland P., Ducamps S., et al. Programme national de surveillance du mésothéliome–principaux résultats 1998–2006. 2009. 22

Goldberg M. and Luce D. The health impact of nonoccupational exposure to asbestos : what do we know ? *European Journal of Cancer Prevention*, 18(6) : 489–503, 2009. 24

Goldberg M., Imbernon E., Rolland P., Soit Ilg A. G., Savès M., de Quillacq A., Frenay C., Chamming's S., Arveux P., Boutin C., Launoy G., Pairon J. C., Astoul P., Galateau-Sallé F., and Brochard P. The French National Mesothelioma Surveillance Program. *Occupational and Environmental Medicine*, 63(6) : 390–395, 2006. 23, 36

Goldberg M., Banaei A., Goldberg S., Auvert B., Luce D., and Guéguen A. Past occupational exposure to asbestos among men in France. *Scandinavian journal of work, environment & health*, pages 52–61, 2000. 36, 38, 74

Greenland S. Problems in the average-risk interpretation of categorical dose-response analyses. *Epidemiology*, pages 563–565, 1995. 26

Guida F. *Rôle de l'exposition professionnelle aux laines minérales dans les cancer broncho-pulmonaires : analyse de l'étude cas-témoins ICARE*. PhD thesis, Paris 11, 2012. 95

Guignard R., Beck F., Richard J. a., Lermenier A., Wilquin J. a., and Nguyen-Thanh V. La consommation de tabac en France en 2014 : caractéristiques et évolutions récentes. *Évolutions, INPES*, (31) : 6, 2015. 29

Gustavsson P., Nyberg F., Pershagen G., Scheele P., Jakobsson R., and Plato N. Low-Dose Exposure to Asbestos and Lung Cancer : Dose-Response Relations and Interaction with Smoking in a Population-based Case-Referent Study in Stockholm, Sweden. *American Journal of Epidemiology*, 155(11) : 1016–1022, 2002. 32

Han J., Slate E. H., and Peña E. A. Parametric latent class joint model for a longitudinal biomarker and recurrent events. *Statistics in Medicine*, 26(29) : 5285–5302, 2007. 116, 117, 124

Hauptmann M., Berhane K., Langholz B., and Lubin J. Using splines to analyse latency in the Colorado Plateau uranium miners cohort. *Journal of Epidemiology and Biostatistics*, 6(6) : 417–424, 2001. 201

Hauptmann M., Pohlabeln H., Lubin J. H., Jöckel K., Ahrens W., Brüske-Hohlfeld I., and Wichmann H. The exposure-time-response relationship between occupational asbestos exposure and lung cancer in two German case-control studies. *American Journal of Industrial Medicine*, 41(2) : 89–97, 2002. 33

Hauptmann M., Lubin J. H., Rosenberg P., Wellmann J., and Kreienbrock L. The use of sliding time windows for the exploratory analysis of temporal effects of smoking histories on lung cancer risk. *Statistics in Medicine*, 19(16) : 2185–2194, 2000a. 31

Hauptmann M., Wellmann J., Lubin J. H., Rosenberg P. S., and Kreienbrock L. Analysis of exposure-time-response relationships using a spline weight function. *Biometrics*, 56(4) : 1105–1108, 2000b. 31, 50, 53, 54, 57, 81, 201, 203

Hawkins D. S., Allen D. M., and Stromberg A. J. Determining the number of components in mixtures of linear models. *Computational Statistics & Data Analysis*, 38(1) : 15–48, 2001. 117

Hegmann K., Fraser A. M., Keaney R. P., Moser D. S. S. E. N., Sedlars M., Higham-Gren L., and Lyon J. L. The effect of age at smoking initiation on lung cancer risk. *Epidemiology*, 4 (5) : 444–448, 1993. 30

Henderson D. W., Rödelsperger K., Woitowitz H.-J., and Leigh J. After Helsinki : a multidisciplinary review of the relationship between asbestos exposure and lung cancer, with emphasis on studies published during 1997–2004. *Pathology*, 36(6) : 517–550, 2004. 32

Hill C. and Doyon F. The frequency of cancer in France : mortality trends since 1950 and summary of the report on the causes of cancer. *Bull Cancer*, 95(1) : 5–10, 2008. 29

Huber P. J. et al. The behavior of maximum likelihood estimates under nonstandard conditions. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, volume 1, pages 221–233. University of California Press, 1967. 122

Huerta M., Chodick G., Balicer R. D., Davidovitch N., and Grotto I. Reliability of self-reported smoking history and age at initial tobacco use. *Preventive Medicine*, 41(2) : 646–650, 2005. 202

IARC. *IARC monographs on the evaluation of carcinogenic risk of chemicals to man. Asbestos. Volume 14.* World Health Organization, 1977. 24

IARC. *IARC Monographs on the Evaluation of Carcinogenic Risks to Humans : Tobacco Smoking Volume 37.* World Health Organization, 1986. 29

IARC. *IARC monographs on the evaluation of carcinogenic risk of chemicals to man. Tobacco smoke and involuntary smoking. Volume 83.* World Health Organization, 2004. 29

IARC. *IARC. World cancer report 2008.* World Health Organization, 2008. 29

Inserm. Effets sur la santé des principaux types d'exposition à l'amiante. Technical report, Les éditions Inserm, 1997. 32

Iwatsubo Y., Pairon J., Boutin C., Menard O., Massin N., Caillaud D., Orlowski E., Galateau-Salle F., Bignon J., and Brochard P. Pleural mesothelioma : dose-response relation at low levels of asbestos exposure in a French population-based case-control study. *American Journal of Epidemiology*, 148(2) : 133–142, 1998. 26, 36, 37

Jacqmin-Gadda H., Proust-Lima C., Taylor J. M., and Commenges D. Score test for conditional independence between longitudinal outcome and time to event given the classes in the joint latent class model. *Biometrics*, 66(1) : 11–19, 2010. 127

Janssen-Heijnen M. L. and Coebergh J.-W. W. The changing epidemiology of lung cancer in Europe. *Lung cancer*, 41(3) : 245–258, 2003. 28

Jones B. L., Nagin D. S., and Roeder K. A sas procedure based on mixture models for estimating developmental trajectories. *Sociological methods & research*, 29(3) : 374–393, 2001. 115

Kang D., Myung M., Kim Y., and Kim J. Systematic review of the effects of asbestos exposure on the risk of cancer between children and adults. *Annals of Occupational and Environmental Medicine*, 25(1) : 10, 2013. 32

Kauppinen T. P., Mutanen P. O., and Seitsamo J. T. Magnitude of misclassification bias when using a job-exposure matrix. *Scandinavian Journal of Work, Environment & Health*, pages 105–112, 1992. 202

Kokko K., Tremblay R. E., Lacourse E., Nagin D. S., and Vitaro F. Trajectories of prosocial behavior and physical aggression in middle childhood : Links to adolescent school dropout and physical violence. *Journal of research on adolescence*, 16(3) : 403–428, 2006. 117

Komárek A., Verbeke G., and Molenberghs G. A SAS-macro for linear mixed models with finite normal mixtures as random-effects distribution. *URL https ://ibiostat. be/online-resources/online-resources/longitudinal# Mixturelin*, 2002. 113

Lacourt A., Gramond C., Rolland P., Ducamp S., Audignon S., Astoul P., Ilg A. G. S., Rinaldo M., Raherison C., Galateau-Salle F., et al. Occupational and non-occupational attributable risk of asbestos exposure for malignant pleural mesothelioma. *Thorax*, 69(6) : 532–539, 2014. 24, 25, 26

Lacourt A. *Mésothéliome : étiologie professionnelle à partir d'enquêtes cas-témoins françaises*. PhD thesis, Bordeaux 2, 2010. 36, 38, 73, 74

Lacourt A., Leffondré K., Gramond C., Ducamp S., Rolland P., Ilg A. G. S., Houot M., Imbernon E., Févotte J., Goldberg M., et al. Temporal patterns of occupational asbestos exposure and risk of pleural mesothelioma. *European Respiratory Journal*, 39(6) : 1304–1312, 2012. 26, 73, 82

Laird N. M. and Ware J. H. Random-effects models for longitudinal data. *Biometrics*, pages 963–974, 1982. 108

Lambert D. Zero-Inflated Poisson Regression, with an Application to Defects in Manufacturing. *Technometrics*, 34(1) : 1–14, 1992. 181

Langholz B., Thomas D., Xiang A., and Stram D. Latency analysis in epidemiologic studies of occupational exposures : Application to the Colorado Plateau Uranium miners cohort. *American Journal of Industrial Medicine*, 35(3) : 246–256, 1999. 20, 25, 50, 53

Lanphear B. P. and Buncher C. R. Latent period for malignant mesothelioma of occupational origin. *JOM*, 34(7) : 718–721, 1992. 25

Lanza S. T. and Rhoades B. L. Latent class analysis : An alternative perspective on subgroup analysis in prevention and treatment. *Prevention Science*, 14(2) : 157–168, 2013. 120

Le Stang N., Belot A., Gilg Soit Ilg A., Rolland P., Astoul P., Bara S., Brochard P., Danzon A., Delafosse P., Grosclaude P., et al. Evolution of pleural cancers and malignant pleural mesothelioma incidence in france between 1980 and 2005. *International Journal of Cancer*, 126 (1) : 232–238, 2010. 22

Leffondré K., Abrahamowicz M., Siemiatycki J., and Rachet B. Modeling smoking history : a comparison of different approaches. *American Journal of Epidemiology*, 156(9) : 813–823, 2002. 20

Leffondré K., Abrahamowicz M., Xiao Y., and Siemiatycki J. Modelling smoking history using a comprehensive smoking index : application to lung cancer. *Statistics in Medicine*, 25(24) : 4132–4146, 2006. 21, 31, 62, 95, 135

Leffondré K., Wynant W., Cao Z., Abrahamowicz M., Heinze G., and Siemiatycki J. A weighted cox model for modelling time-dependent exposures in the analysis of case–control studies. *Statistics in Medicine*, 29(7-8) : 839–850, 2010. 174

Lesaffre E. and Spiessens B. On the effect of the number of quadrature points in a logistic random effects model : an example. *Journal of the Royal Statistical Society : Series C (Applied Statistics)*, 50(3) : 325–335, 2001. 186

Lin H., McCulloch C. E., Turnbull B. W., Slate E. H., and Clark L. C. A latent class mixed model for analysing biomarker trajectories with irregularly scheduled observations. *Statistics in Medicine*, 19(10) : 1303–1318, 2000. 106, 118, 120, 124

Lin H., McCulloch C. E., and Mayne S. T. Maximum likelihood estimation in the joint analysis of time-to-event and multiple longitudinal variables. *Statistics in Medicine*, 21(16) : 2369–2382, 2002a. 125

Lin H., Turnbull B. W., McCulloch C. E., and Slate E. H. Latent class models for joint analysis of longitudinal biomarker and event process data : application to longitudinal prostate-specific antigen readings and prostate cancer. *Journal of the American Statistical Association*, 97 (457) : 53–65, 2002b. 124

Lindstrom M. J. and Bates D. M. Newton—Raphson and EM algorithms for linear mixed-effects models for repeated-measures data. *Journal of the American Statistical Association*, 83(404) : 1014–1022, 1988. 113

Loomis D., Dement J. M., Wolf S. H., and Richardson D. B. Lung cancer mortality and fiber exposures among North Carolina asbestos textile workers. *Occupational and Environmental Medicine*, 2009. 33

Lopez A. D., Collishaw N. E., and Piha T. A descriptive model of the cigarette epidemic in developed countries. *Tobacco Control*, 3(3) : 242, 1994. 29

Lubin J. H. and Caporaso N. E. Cigarette smoking and lung cancer : Modeling total exposure and intensity. *Cancer Epidemiology and Prevention Biomarkers*, 15(3) : 517–523, 2006. 20, 31

Luce D. and Stücker I. Investigation of occupational and environmental causes of respiratory cancers (ICARE) : a multicenter, population-based case-control study in france. *BMC Public Health*, 11(1) : 928, 2011. 29, 39, 40, 41

Luce D., Bugel I., Goldberg P., Goldberg M., Salomon C., Billon-Galland M.-A., Nicolau J., Quénel P., Fevotte J., and Brochard P. Environmental exposure to tremolite and respiratory cancer in new Caledonia : a case-control study. *American journal of epidemiology*, 151(3) : 259–265, 2000. 32

MacQueen J. et al. Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, volume 1, pages 281–297. Oakland, CA, USA, 1967. 107

Magnani C., Dalmasso P., Biggeri A., Ivaldi C., Mirabelli D., and Terracini B. Increased risk of malignant mesothelioma of the pleura after residential or domestic exposure to asbestos : a case-control study in Casale Monferrato, Italy. *Environmental Health Perspectives*, 109(9) : 915–919, 2001. 24

Magnani C., Ferrante D., Barone-Adesi F., Bertolotti M., Todesco A., Mirabelli D., and Terracini B. Cancer risk after cessation of asbestos exposure : a cohort study of italian asbestos cement workers. *Occupational and Environmental Medicine*, 65(3) : 164–170, 2008. 32

Manning W. G. A two-part model of the demand for medical care : Preliminary results from the Health Insurance Study. *Health, Economics, and Health Economics*, pages 103–123, 1981. 180

Marinaccio A., Binazzi A., Cauzillo G., Cavone D., De Zotti R., Ferrante P., Gennaro V., Gorini G., Menegozzo M., Mensi C., et al. Analysis of latency time and its determinants in asbestos related malignant mesothelioma cases of the Italian register. *European Journal of Cancer*, 43(18) : 2722–2728, 2007. 25

Marquardt D. W. An algorithm for least-squares estimation of nonlinear parameters. *Journal of the society for Industrial and Applied Mathematics*, 11(2) : 431–441, 1963. 113

McCulloch C., Lin H., Slate E., and Turnbull B. Discovering subpopulation structure with latent class mixed models. *Statistics in medicine*, 21(3) : 417–429, 2002. 124

McDonald A. D. and McDonald J. C. Malignant mesothelioma in north America. *Cancer*, 46 (7) : 1650–1656, 1980. 25

Milanzi E. B., Brunekreef B., Koppelman G. H., Wijga A. H., Rossem L., Vonk J. M., Smit H. A., and Gehring U. Lifetime secondhand smoke exposure and childhood and adolescent asthma : findings from the piama cohort. *Environmental Health*, 16(1) : 14, 2017. 106

Monahan K. C., Steinberg L., Cauffman E., and Mulvey E. P. Trajectories of antisocial behavior and psychosocial maturity from adolescence to young adulthood. *Developmental psychology*, 45(6) : 1654, 2009. 117

Mullahy J. Specification and testing of some modified count data models. *Journal of Econometrics*, 33(3) : 341–365, 1986. 180

Muthén B. Latent variable analysis. *The Sage handbook of quantitative methodology for the social sciences*, 345 : 368, 2004. 106, 109, 117

Muthén B. and Shedden K. Finite mixture modeling with mixture outcomes using the em algorithm. *Biometrics*, 55(2) : 463–469, 1999. 107, 108, 109, 120, 124

Muthén B., Brown C. H., Masyn K., Jo B., Khoo S.-T., Yang C.-C., Wang C.-P., Kellam S. G., Carlin J. B., and Liao J. General growth mixture modeling for randomized preventive interventions. *Biostatistics*, 3(4) : 459–475, 2002. 117

Muthén B. O. Latent variable mixture modeling. In *New developments and techniques in structural equation modeling*, pages 21–54. Psychology Press, 2001. 117

Muthén L. K. and Muthén B. O. *Mplus : Statistical analysis with latent variables : User's guide*. Muthén & Muthén Los Angeles, 2005. 109, 113

Muthén B. and Asparouhov T. Growth mixture modeling : Analysis with non-gaussian random effects. In *Fitzmaurice, G and Davidian, M and Verbeke, G and Molenberghs, G (eds.) Longitudinal Data Analysis*, pages 143–165. Boca Raton : Chapman & Hall/CRC Press, 2009. 115, 181, 182, 183

Nagin D. S. Analyzing developmental trajectories : a semiparametric, group-based approach. *Psychological methods*, 4(2) : 139, 1999. 117

Nagin D. S. *Group-Based Modeling of Development*. Harvard University Press, 2005. 115, 116, 117, 120, 181

Nagin D. S. and Tremblay R. E. What has been learned from group-based trajectory modeling ? examples from physical aggression and other problem behaviors. *The Annals of the American Academy of Political and Social Science*, 602(1) : 82–117, 2005. 106, 117, 181

Nelder J. A. and Mead R. A simplex method for function minimization. *The computer journal*, 7(4) : 308–313, 1965. 64

Nguyen Q. H. and Rayward-Smith V. J. Internal quality measures for clustering in metric spaces. *International Journal of Business Intelligence and Data Mining*, 3(1) : 4–29, 2008. 119, 120

Nielsen L., Bælum J., Rasmussen J., Dahl S., Olsen K., Albin M., Hansen N., and Sherson D. Occupational asbestos exposure and lung cancer—a systematic review of the literature. *Archives of Environmental & Occupational Health*, 69(4) : 191–206, 2014. PMID : 24410115. 32, 82, 201

Oehlert G. W. A note on the delta method. *The American Statistician*, 46(1) : 27–29, 1992. 114

Papadopoulos A., Guida F., Leffondré K., Cénée S., Cyr D., Schmaus A., Radoi L., Paget-Bailly S., Carton M., Menvielle G., et al. Heavy smoking and lung cancer : are women at higher risk ? result of the ICARE study. *British Journal of Cancer*, 110(5) : 1385, 2014. 46, 62, 95, 127, 135

Park E., McCoy T. P., Erausquin J. T., and Bartlett R. Trajectories of risk behaviors across adolescence and young adulthood : the role of race and ethnicity. *Addictive behaviors*, 76 : 1–7, 2018. 106

Pesch B., Kendzia B., Gustavsson P., Jöckel K.-H., Johnen G., Pohlabeln H., Olsson A., Ahrens W., Gross I. M., Brüske I., et al. Cigarette smoking and lung cancer—relative risk estimates for the major histological types from a pooled analysis of case–control studies. *International journal of cancer*, 131(5) : 1210–1219, 2012. 30

Peto J., Seidman H., and Selikoff I. Mesothelioma mortality in asbestos workers : implications for models of carcinogenesis and risk assessment. *British Journal of Cancer*, 45(1) : 124, 1982. 25

Peto J., Doll R., Hermon C., Binns W., Clayton R., and Goffe T. Relationship of mortality to measures of environmental asbestos pollution in an asbestos textile factory. *The Annals of Occupational Hygiene*, 29(3) : 305–355, 1985. 25

Peto R., Darby H. S. D., Silcocks P., Whitley E., and Doll R. Smoking, smoking cessation, and lung cancer in the UK since 1950 : combination of national statistics with two case-control studies. *British Medical Journal*, 321(7257) : 323–329, 2000. 20, 30

Peto R., Lopez A., Boreham J., Thun M., and Heath C. J. Mortality from tobacco in developed countries : indirect estimation from national vital statistics. *Lancet*, 339(8804) : 1268–78, 1992. 29

Pira E., Pelucchi C., Piolatto P., Negri E., Discalzi G., and La Vecchia C. First and subsequent asbestos exposures in relation to mesothelioma and lung cancer mortality. *British Journal of cancer*, 97(9) : 1300, 2007. 32

Pira E., Romano1 C., Violante F., Farioli A., Spatari G., La Vecchia C., and Boffetta P. Updated mortality study of a cohort of asbestos textile workers. *Cancer Medicine*, 5(9) : 2623–2628, 2016. 26

Prague M., Diakite A., and Commenges D. Package 'marqlevalg' - algorithme de levenberg-marquardt en r : Une alternative à 'optimx' pour des problèmes de minimisation, 1ères rencontres r, jul 2012, bordeaux, france. <hal-00717566>. 2012. 189

Prazakova S., Thomas P. S., Sandrini A., and Yates D. H. Asbestos and the lung in the 21st century : an update. *The Clinical Respiratory Journal*, 8(1) : 1–10, 2014. 32

Proust C. and Jacqmin-Gadda H. Estimation of linear mixed models with a mixture of distribution for the random effects. *Computer methods and programs in biomedicine*, 78(2) : 165–173, 2005. 113, 115

Proust C., Jacqmin-Gadda H., Taylor J. M., Ganiayre J., and Commenges D. A nonlinear model with latent process for cognitive evolution using multivariate longitudinal data. *Biometrics*, 62(4) : 1014–1024, 2006. 106, 111, 118, 124, 129

Proust-Lima C. *Modèles mixtes à structure latente pour données longitudinales multivariées hétérogènes : application à l'étude du vieillissement cognitif et de la démence*. PhD thesis, Université Victor Segalen-Bordeaux II, 2006. 116

Proust-Lima C. and Taylor J. M. Development and validation of a dynamic prognostic tool for prostate cancer recurrence using repeated measures of posttreatment psa : a joint modeling approach. *Biostatistics*, 10(3) : 535–549, 2009. 118

Proust-Lima C., Letenneur L., and Jacqmin-Gadda H. A nonlinear latent class model for joint analysis of multivariate longitudinal data and a binary outcome. *Statistics in Medicine*, 26 (10) : 2229–2245, 2007. 124, 126

Proust-Lima C., Joly P., Dartigues J.-F., and Jacqmin-Gadda H. Joint modelling of multivariate longitudinal outcomes and a time-to-event : a nonlinear latent class approach. *Computational statistics & data analysis*, 53(4) : 1142–1154, 2009. 124

Proust-Lima C., Amieva H., and Jacqmin-Gadda H. Analysis of multivariate mixed longitudinal data : a flexible latent process approach. *British Journal of Mathematical and Statistical Psychology*, 66(3) : 470–487, 2013. 111, 129

Proust-Lima C., Philipps V., and Liquet B. Estimation of extended mixed models using latent classes and latent processes : The r package lcmm. *Journal of Statistical Software, Articles*, 78(2) : 1–56, 2017. 111, 113, 114, 117, 120, 127, 137

Rachet B., Siemiatycki J., Abrahamowicz M., and Leffondre K. A flexible modeling approach to estimating the component effects of smoking behavior on lung cancer. *Journal of Clinical Epidemiology*, 57(10) : 1076 – 1085, 2004. 20, 30, 31

Rake C., Gilham C., Hatch J., Darnton A., Hodgson J., and Peto J. Occupational, domestic and environmental mesothelioma risks in the British population : a case–control study. *British Journal of Cancer*, 100(7) : 1175, 2009. 26

Ramsay J. O. et al. Monotone regression splines in action. *Statistical science*, 3(4) : 425–441, 1988. 112

Reid A., De Klerk N., Magnani C., Ferrante D., Berry G., Musk A., and Merler E. Mesothelioma risk after 40 years since first exposure to asbestos : a pooled analysis. *Thorax*, 69 : 843–850, 2014. 26

Richardson D. B. and Ashmore J. P. Investigating time patterns of variation in radiation cancer associations. *Occupational and Environmental Medicine*, 62(8) : 551–558, 2005. 53

Richardson D. B., Cole S. R., Chu H., and Langholz B. Lagging exposure information in cumulative exposure-response analyses. *American Journal of Epidemiology*, 174(12) : 1416–1422, 2011. 20, 26, 50

Richiardi L., Barone-Adesi F., Merletti F., and Pearce N. Using directed acyclic graphs to consider adjustment for socioeconomic status in occupational cancer studies. *Journal of Epidemiology & Community Health*, 62(7) : e14–e14, 2008. 93, 95

Rizopoulos D. *Joint models for longitudinal and time-to-event data : With applications in R.* Chapman and Hall/CRC, 2012. 124, 187, 188, 189, 190

Rödelsperger K., Jöckel K.-H., Pohlabeln H., Römer W., and Woitowitz H.-J. Asbestos and man-made vitreous fibers as risk factors for diffuse malignant mesothelioma : Results from a German hospital-based case-control study. *American Journal of Industrial Medicine*, 39(3) : 262–275, 2001. 26

Roeder K., Lynch K. G., and Nagin D. S. Modeling uncertainty in latent class membership : A case study in criminology. *Journal of the American Statistical Association*, 94(447) : 766–776, 1999. 181, 182, 185

Rolland P., Gramond C., Lacourt A., Astoul P., Chamming's S., Ducamp S., Frenay C., Galateau-Salle F., Ilg A. G. S., Imbernon E., et al. Occupations and industries in france at high risk for pleural mesothelioma : A population-based case–control study (1998–2002). *American journal of industrial medicine*, 53(12) : 1207–1219, 2010. 76

Roos F. Symposium «amiante et risques professionnels : études épidémiologiques récentes. *Documents pour le médecin du travail*, 97 : 103–7, 2004. 22

Rubin D. B. *Multiple Imputation for Nonresponse in Surveys (Wiley Series in Probability and Statistics).* Wiley, 1987. 93, 122, 123

Rusch V. A proposed new international tnm staging system for malignant pleural mesothelioma. *CHEST*, 108(4) : 1122 – 1128, 1995. 22

Rushton L., Bagga S., Bevan R., Brown T. P., Cherrie J. W., Holmes P., Fortunato L., Slack R., Van Tongeren M., Young C., and Hutchings S. J. Occupation and cancer in Britain. *British Journal of Cancer*, 102 : 1428–1437, 2010. 25

Sakia R. The box-cox transformation technique : a review. *The statistician*, pages 169–178, 1992. 176

Salvan A., Stayner L., Steenland K., and Smith R. Selecting an exposure lag period. *Epidemiology*, pages 387–390, 1995. 26

Schafer J. L. *Analysis of incomplete multivariate data.* Chapman and Hall/CRC, 1997. 123

Scherpereel A. and Astoul P. Mésothéliome pleural malin. *EMC-Pneumologie*, pages 1–15, 2007. [Article 6-002-H-10]. 22

Schlattmann P. Estimating the number of components in a finite mixture model : the special case of homogeneity. *Computational Statistics & Data Analysis*, 41(3-4) : 441–451, 2003. 116

Schwarz G. et al. Estimating the dimension of a model. *The Annals of Statistics*, 6(2) : 461–464, 1978. 116

Seidman H., Selikoff I. J., and Gelb S. K. Mortality experience of amosite asbestos factory workers : Dose-response relationships 5 to 40 years after onset of short-term work exposure. *American Journal of Industrial Medicine*, 10(5-6) : 479–514, 1986. 25

Selikoff I. J. and Seidman H. Asbestos-associated deaths among insulation workers in the United States and Canada, 1967–1987. *Annals of the New York Academy of Sciences*, 643 (1) : 1–14, 1991. 25

Siesling S., Van der Zwan J., Izarzugaza I., Jaal J., Treasure T., Foschi R., Ricardi U., Groen H., Tavilla A., Ardanaz E., and Group. R. W. Rare thoracic cancers, including peritoneum mesothelioma. *European Journal of Cancer*, 48(7) : 949 – 960, 2012. 23

Soulakova J. N., Hartman A. M., Liu B., Willis G. B., and Augustine S. Reliability of adult self-reported smoking history : data from the tobacco use supplement to the current population survey 2002–2003 cohort. *Nicotine & Tobacco Research*, 14(8) : 952–960, 2012. 202

Spirtas R., Heineman E. F., Bernstein L., Beebe G. W., Keehn R. J., Stark A., Harlow B. L., and Benichou J. Malignant mesothelioma : attributable risk of asbestos exposure. *Occupational and Environmental Medicine*, 51(12) : 804–811, 1994. 25, 26

Stram D. O. and Lee J. W. Variance components testing in the longitudinal mixed effects model. *Biometrics*, pages 1171–1177, 1994. 116

Sylvestre M.-P. and Abrahamowicz M. Flexible modeling of the cumulative effects of time-dependent exposures on the hazard. *Statistics in Medicine*, 28(27) : 3437–3453, 2009. 50, 53, 54, 57, 63

Szekely G. J. and Rizzo M. L. Hierarchical clustering via joint between-within distances : Extending ward's minimum variance method. *Journal of classification*, 22(2) : 151–183, 2005. 107

Teschke K., Olshan A., Daniels J., De Roos A., Parks C., Schulz M., and Vaughan T. Occupational exposure assessment in case–control studies : opportunities for improvement. *Occupational and Environmental Medicine*, 59(9) : 575–594, 2002. 81, 202

Thomas D. C. Models for exposure-time-response relationships with applications to cancer epidemiology. *Annual Review of Public Health*, 9(1) : 451–482, 1988. 20

Thurston S. W., Liu G., Miller D. P., and Christiani D. C. Modeling lung cancer risk in case-control studies using a new dose metric of smoking. *Cancer Epidemiology and Prevention Biomarkers*, 14(10) : 2296–2302, 2005. 20, 30, 31

Twisk J. and Hoekstra T. Classifying developmental trajectories over time should be done with great caution : a comparison between methods. *Journal of Clinical Epidemiology*, 65(10) : 1078–1087, 2012. 117, 118, 119, 120

Vacek P. M. Assessing the effect of intensity when exposure varies over time. *Statistics in Medicine*, 16(5) : 505–513, 1997. 53, 101

van der Bij S., Koffijberg H., Lenters V., Portengen L., Moons K. G. M., Heederik D., and Vermeulen R. C. H. Lung cancer risk at low cumulative asbestos exposure : meta-regression of the exposure–response relationship. *Cancer Causes & Control*, 24(1) : 1–12, 2013. 32

Verbeke G. and Lesaffre E. A linear mixed-effects model with heterogeneity in the random-effects population. *Journal of the American Statistical Association*, 91(433) : 217–221, 1996. 109, 116

Verbeke G. and Molenberghs G. *Linear mixed models for longitudinal data*. Springer, 2000. 108

Vermeulen R. and Chadeau-Hyam M. Commentary : Dynamic aspects of exposure history—do they matter ? *Epidemiology*, 23(6) : 900–901, 2012. 20

Vineis P., Kogevinas M., Simonato L., Brennan P., and Boffetta P. Levelling-off of the risk of lung and bladder cancer in heavy smokers : an analysis based on multicentric case-control studies and a metabolic interpretation. *Mutation Research*, 463(1) : 103 – 110, 2000. 30, 82

Vlaanderen J., Portengen L., Schüz J., Olsson A., Pesch B., Kendzia B., Stücker I., Guida F., Brüske I., Wichmann H.-E., et al. Effect modification of the association of cumulative exposure and cancer risk by intensity of exposure and time since exposure cessation : a flexible method applied to cigarette smoking and lung cancer in the synergy study. *American Journal of Epidemiology*, 179(3) : 290–298, 2013. 31

Wagner J. C., Sleggs C., and Marchand P. Diffuse pleural mesothelioma and asbestos exposure in the North Western Cape Province. *Occupational and Environmental Medicine*, 17(4) : 260–271, 1960. 25

Wang C.-P., Hendricks Brown C., and Bandeen-Roche K. Residual diagnostics for growth mixture models : Examining the impact of a preventive intervention on multiple trajectories

of aggressive behavior. *Journal of the American Statistical Association*, 100(471) : 1054–1076, 2005. 120, 122

Weden M. M. and Miles J. N. Intergenerational relationships between the smoking patterns of a population-representative sample of us mothers and the smoking trajectories of their children. *American Journal of Public Health*, 102(4) : 723–731, 2012. 106, 118

Weiss W. Cigarette smoking and lung cancer trends. *CHEST*, 111(5) : 1414 – 1416, 1997. 30

White H. *Estimation, inference and specification analysis*. Number 22. Cambridge university press, 1996. 122

Wolfinger R. Laplace's approximation for nonlinear mixed models. *Biometrika*, 80(4) : 791–795, 1993. 185

Wynder E. and Graham E. Tobacco smoking as a possible etiologic factor in bronchiogenic carcinoma ; a study of 684 proved cases. *J Am Med Assoc*, 143(4) : 329–36, 1950. 30

Xu W. and Hedeker D. A random-effects mixture model for classifying treatment response in longitudinal clinical trials. *Journal of Biopharmaceutical Statistics*, 11(4) : 253–273, 2001. 116

Zhang J., Cole S. R., Richardson D. B., and Chu H. A bayesian approach to strengthen inference for case-control studies with multiple error-prone exposure assessments. *Statistics in medicine*, 32(25) : 4426–4437, 2013. 203

Après l'année de publication se trouve le numéro de la page où la référence a été citée.

Annexe A :
Activités complémentaires réalisées pendant la thèse

2016 : Mission complémentaire en conseil statistique auprès de la société Capionis (Bordeaux) avec M. Sébastien Marque à hauteur de 32 jours

2016-2017 : Poste de moniteur à hauteur de 64h éq. TD
Travaux Dirigés dispensés en :
M1 Santé Publique : UE EPI101, UE STA102
M2 Epidémiologie : Initiation logiciel R, UE EPI202 cas témoins
DFGSM2 : logiciel de bureautique

2017-2018 : Poste de moniteur à hauteur de 64h éq. TD
Travaux Dirigés dispensés en :
M1 Santé Publique : UE EPI101, UE STA102
M2 Epidémiologie : Initiation logiciel R, UE EPI202 cas témoins
DFGSM2 : logiciels de bureautique
Ecole Orthophonie de l'Université de Bordeaux : soutien statistique auprès des étudiantes de 3ème année

Annexe B :
CV

Emilie LEVEQUE

190 Cours de la Somme
Bordeaux, France 33800
(+33) 6 74 54 08 05
emilie.leveque@u-bordeaux.fr

Formation

2008 - 2013 **INSA Rouen Normandie - spécialisation Génie Mathématique**.
Diplôme d'ingénieur de l'Institut National des Sciences Appliquées en Mathématiques et Informatique

Aout - Déc 2012 **Semestre à Lappeenranta University of Technology (LUT) en Finlande**.

2008 **Baccalauréat Scientifique**, *Lycée Georges Brassens*, Neufchatel-en-Bray, Mention Bien.

Expériences

2016-2017 & 2017-2018 **Activité complémentaire : Enseignement**, *Poste de moniteur de 64h éq TD* .
- M1 Santé Publique / M2 Epidémiologie au sein de l'ISPED à Bordeaux
- DFGSM2 (logiciels bureautique) de l'Université de Bordeaux
- Formation Orthophonie (soutien statistique) de l'Université de Bordeaux

2016 **Activité complémentaire : Conseil statistique**, *auprès de la société Capionis (Bordeaux)* .

Janv 2016 - Déc 2018 **Doctorat en Santé Publique option Biostatistique. Équipe Biostatistique du centre INSERM U1219 Bordeaux Population Heatlh**, *Modèles statistiques pour l'étude de l'impact des expositions professionnelles ou environnementales prolongées sur le risque de cancer. Applications au mésothéliome pleural et cancer du poumon*.
Application d'un indice cumulé pondéré d'exposition avec une fonction de poids flexible dépendante du temps
Application d'un modèle conjoint linéaire mixte à classes latentes avec un modèle logistique
Développement d'un modèle mixte Poisson à classes latentes

Nov 2013 - Déc 2015 **Ingénieure Biostatisticienne. Équipe Biostatistique du centre INSERM U897 à Bordeaux(33)**.
Projet sur l'aspect temporel de la relation entre expositions professionnelles ou environnementales et la survenue du cancer

Avril - Oct 2013 **Stage ingénieur. Équipe Biostatistique du centre INSERM U897 à Bordeaux(33)**, *La pollution influence-t-elle les sorties SMUR pour pathologies cardio-vasculaires ?* .
Analyse de données rétrospectives regroupant des données sanitaires et environnementales. Utilisation d'une méthode case-crossover

Juin - Sept 2011 **Stage technicien. Équipe de Recherche Associée "Accidentologie, Trajectographie, Risques Routiers" au CETE Normandie Centre à Grand-Quevilly(76)**, *Influence des véhicules croiseurs sur la trajectoire des véhicules* .
Analyse et traitement de données microscopiques du trafic relevées grâce à une station bord de voie

Compétences informatiques

Bonne maîtrise de R et Latex.
Langages : C, C++, Java, Fortran, UML.
Web : PHP, HTML, SQL.
Logiciels : SAS, Matlab, Eclipse, Maple.
Bureautique : Word, Excel, Power Point, Open Office

Compétences biostatistiques et épidémiologiques

Analyse de données cas-témoins
Analyse de séries chronologiques et données cas-croisés
Modélisation d'expositions prolongées (professionnelles à partir de matrice emploi-exposition et environnementales) en étiologie du cancer
Modélisation des variables quantitatives (splines, polynomes fractionnaires)
Utilisation de Directed Acyclic Graphs (DAG)
Utilisation de modèles linéaires généralisés (linéaire, logistique, Poisson) et de modèles de survie

Utilisation de modèles linéaires mixtes conjoints à classes latentes
Adaptation d'un modèle mixte de Poisson à classes latentes

Langues

Anglais **Courant**
Espagnol **Intermediaire**

Publications

Articles scientifiques publiés

Lévêque E, Lacourt A, Luce D, et al Time-dependent effect of intensity of smoking and of occupational exposure to asbestos on the risk of lung cancer : results from the ICARE caseâĂŞcontrol study.Occup Environ Med 2018 ;75 :586-592.

Lacourt A*, **Lévêque E*** (co-first), Guichard E,Gilg Soit Ilg A, Sylvestre M-P, Leffondré K. Dose-time-response association between occupational asbestos exposure and pleural mesothelioma Occup Environ Med 2017 ;74 :691-697.

Pradeau C, Rondeau V, **Lévêque E**, Guernion PY, Tentillier E, Thicoipé M, Brochard P. Air pollution and activation of mobile medical team for out-of-hospital cardiac arrest. American Journal of Emergency Medicine (2015) 367-372.

Lettre

Lacourt A, **Lévêque E**, Goldberg M, et al. Dose-time response association between occupational asbestos exposure and pleural mesothelioma : authors' response. Occup Environ Med 2018 ;75 :161-162.

Résumés publiés

Lévêque E, Lacourt A, Luce D, et al. O15-2 Dynamic longitudinal effects of increasing intensity of smoking and occupational exposure to asbestos on lung cancer : results from the icare case-control study.Occup Environ Med 2016 ;73 :A28. (EPICOH Barcelone 2016).

Lacourt A, **Lévêque E**, Leffondré K. 0291 Dose-time-response association between occupational asbestos exposure and mesothelioma Occup Environ Med 2014 ;71 :A101. (EPICOH Chicago 2014).

Article soumis

Lévêque E, Lacourt A, Phillips V, Luce D, Guénel P, Stücker I, Proust-lima C and Leffondré K. Association between lung cancer and lifetime profiles of intensity of exposure to occupational asbestos and smoking : Results from the ICARE case-control study. *International Journal of Epidemiology*. .

Article en préparation

Lévêque E, Lacourt A, Leffondré K and Proust-lima C. Zero-Inflated Poisson latent class mixed model for the identification of longitudinal profiles of environmental and occupational exposure intensities in epidemiological studies.

Congrès

Présentations orales :

- Dynamic longitudinal effects of increasing intensity of smoking and occupational exposure to asbestos on lung cancer : results from the icare case-control study. The 25th International Epidemiology in Occupational Health (EPICOH) Conference, Barcelona, September 5-7, 2016.

- A Zero-Inflated Poisson latent class mixed model for the identification of longitudinal profiles of environmental and occupational exposure intensities in epidemiological studies. The 38th Annual Conference of the International Society for Clinical Biostatistics (ISCB), Melbourne, Australia, August 26-30, 2018.

Prśentations affichées :

- Identification of lifetime profiles of smoking intensities and association with lung cancer risks : Results from the ICARE case-control study. The 51st Annual Meeting of Society for Epidemiological Research (SER), Baltimore, June 19-22, 2018.

- Analyse de l'aspect temporel de la relation entre expositions prolongées et risque de cancer Application à 3 relations : amiante-mésothéliome pleural, amiante-cancer du poumon et tabac-cancer du poumon. Rencontres scientifiques de l'ANSES pour les 10 ans du Programme national de recherche environnement santé travail (PNREST) Paris 14 Novembre 2016.

- Dose-time-response association between occupational asbestos exposure and mesothelioma. The 24th International Epidemiology in Occupational Health (EPICOH) Conference, Chicago, June 24-27, 2014.

Bourses

- Subvention doctorale en "Sciences Humaines et Sociales, Epidémiologie, Santé Publique" par l'Institut National du Cancer (INCa) (2016-2018)

- SPC Travel Scholarship pour le congrès Society for Epidemiologic Research (SER) à Baltimore Juin 2018 : 400$

Modélisation statistique de l'intensité des expositions prolongées en étiologie du cancer : application au tabac, à l'amiante, au cancer du poumon, et au mésothéliome pleural.

Résumé : L'association entre le tabac et le cancer du poumon ou entre l'exposition professionnelle à l'amiante et le mésothéliome pleural ont largement été étudiées. Cependant, comme pour de nombreuses autres relations expositions prolongées-cancer, le rôle de l'intensité d'exposition tout au long de la vie a été peu étudié. La prise en compte de la variation de l'intensité au cours de la vie et de son effet dépendant du temps dans les analyses statistiques des données cas-témoins pose en effet quelques défis méthodologiques.

Les objectifs de la thèse étaient 1) d'étudier l'effet dépendant du temps de l'intensité d'exposition au cours de la vie sur le risque de cancer et 2) d'identifier les profils de trajectoires d'intensité d'exposition sur la vie entière et comparer les risques de cancer associés. Pour répondre à ces deux objectifs, nous avons utilisé un indice cumulé d'exposition pondéré flexible déjà existant et nous avons développé un nouveau modèle conjoint à classes latentes, pour analyser les données de deux études cas-témoins françaises sur le mésothéliome pleural et le cancer du poumon.

Les résultats montrent la contribution importante de l'intensité de la consommation de tabac récente pour le cancer du poumon et des expositions professionnelles anciennes à l'amiante pour les deux cancers. Ils confirment l'importance de considérer l'aspect temporel des expositions pour évaluer l'association avec le risque de cancer et illustrent l'intérêt des approches statistiques considérées.

Mots-clés : Expositions prolongées, intensité d'exposition, tabac, amiante, cancer du poumon, mésothéliome pleural, relation dose-réponse, effet dépendant du temps, modélisation flexible, modèle mixte à classes latentes, modèle conjoint.

Statistical modelling of the intensity of protracted exposures in etiology of cancer : application to smoking, asbestos, lung cancer and pleural mesothelioma.

Summary : The association between smoking and lung cancer or between occupational exposure to asbestos and pleural mesothelioma have been extensively investigated. Nevertheless, as for many protracted exposures-cancer relationships, the role of exposure intensity over lifetime has been rarely addressed. Accounting for individual variation of intensity over lifetime and investigating time-dependent effect in the statistical analysis of case-control data indeed raise several methodological issues.

The thesis objectives were 1) to study the time-dependent effect of exposure intensity over lifetime on the risk of cancer and 2) to identify lifetime profiles of exposure intensities and to compare their associated risks of cancer. To address these objectives, we used an existing flexible weighted cumulative index of exposure and we developed a new joint latent class mixed model, to analyze the data from two French case-control studies on lung cancer and pleural mesothelioma.

The results show the important contribution of recent smoking intensity for lung cancer and distant intensity of exposure to asbestos for both cancers. They confirm the importance of the timing of exposure in the association with the risk of cancer and illustrate the relevance of the proposed statistical approaches.

Key words : Protracted exposures, exposure intensity, smoking, asbestos, lung cancer, pleural mesothelioma, dose-time-response relationship, flexible modeling, latent class mixed model, joint model.

Discipline : Santé publique – option : Biostatistiques
Laboratoire : Unité INSERM U1219, Bordeaux Population Health Center - Université de Bordeaux 146 rue Léo-Saignat CS 61292 33076 BORDEAUX CEDEX