Approche hiérarchique bayésienne pour l'estimation du risque de cancers radio-induits en situation d'expositions professionnelles multiples et incertaines. Application aux travailleurs du cycle du combustible nucléaire

Marion Belloni

Thèse de doctorat

NNT : 2021UPASR004

Approche hiérarchique bayésienne pour l'estimation du risque de cancers radio-induits en situation d'expositions professionnelles multiples et incertaines : Application aux travailleurs du cycle du combustible nucléaire

Bayesian hierarchical approach to estimate the risk of radiation-induced cancers in the situation of multiple and uncertain occupational exposures: Application to workers in the nuclear fuel cycle

Thèse de doctorat de l'université Paris-Saclay

École doctorale n° 570, Santé Publique (EDSP)
Spécialité de doctorat: Biostatistique
Unité de recherche : Laboratoire d'épidémiologie des rayonnements ionisants (LEPID)
Référent : Faculté de médecine

Thèse présentée et soutenue à Paris-Saclay, le 10/03/2021, par

Marion BELLONI

Composition du Jury

Bruno FALISSARD Professeur des Universités Praticien Hospitalier, Université Paris Saclay	Président
Marc COLONNA Directeur de recherche, Registre des cancers de l'Isère et CHU de Grenoble	Rapporteur
Karen LEFFONDRE Professeur des Universités, Université de Bordeaux ISPED	Rapporteur & Examinatrice
Nicolas BOUSQUET Professeur associé, Université de La Sorbonne	Examinateur

Direction de la thèse

Chantal GUIHENNEUC Professeur des Universités, Université de Paris	Directrice de thèse
Sophie ANCELET Chargée de Recherche, Institut de Radioprotection et de Sûreté Nucléaire	Co-encadrante

Remerciements

Je tiens d'abord à remercier très sincèrement Sophie Ancelet et Chantal Guihenneuc. Elles ont joué un rôle primordial dans l'aboutissement de ce travail par leurs qualités professionnelles et humaines, par leurs conseils, leur encadrement et leur soutien. Merci beaucoup Sophie pour ton enthousiasme à transmettre, tes encouragements et ta gentillesse tout au long de cette thèse. Chantal, je te remercie pour le temps que tu m'as apporté et pour ton optimisme face aux difficultés rencontrées.

Je tiens également à remercier Klervi Leuraud, responsable du LEPID, pour ses encouragements et sa bienveillance. Je la remercie ainsi que Sylvaine Caer-Lorho, Estelle Rage et Dominique Laurier pour leur expertise sur la cohorte des mineurs d'uranium français.

Je souhaite remercier Marc Colonna et Karen Leffondré d'avoir accepté de prendre le temps d'être rapporteurs de cette thèse. Je remercie également Nicolas Bousquet et Bruno Falissard d'avoir accepté d'en être les examinateurs.

Je voudrais remercier tous les membres du LEPID côtoyés quelques mois ou tout au long de cette thèse pour leur accueil chaleureux et même gourmand. Je ne pouvais guère espérer rejoindre une équipe aussi accueillante ! De pauses café en rallyes pédestres, de séminaires en fêtes de la musique, leur présence a égayé tous ces moments et bien d'autres encore. Je tiens à remercier en particulier Ségolène Bouet, Sabrina Cossais, Kossi Abalo et Lucie Fournier avec qui j'ai partagé de

nombreux moments aussi réconfortants qu'indispensables. La bonne humeur de tous, à différentes étapes de la thèse, a participé à l'aboutissement de ce projet. Le confinement m'a éloignée de tous ces collègues et j'espère pouvoir les retrouver à l'occasion de la soutenance.

À ma famille et mes amis, vous qui arrêterez sûrement la lecture de ce manuscrit à ces quelques lignes, j'adresse un immense merci. Merci à tous ces amis dont la présence à Paris ou beaucoup plus loin, pour deux heures, deux jours ou deux semaines, m'a apporté confiance, énergie et bonne humeur ! Merci aussi à tous les membres de ma famille qui, mois après mois, m'ont écoutée parler de mineurs d'uranium, de radon et d'algorithmes : votre patience est d'or et je ne l'oublierai pas ! Un très grand merci à ma mère : son soutien pendant ces années de thèse et toutes les autres années m'a été indispensable. Et finalement, un immense merci à toi Alexis, merci pour tout !

Valorisation scientifique de la thèse

Publications

Belloni, M., Guihenneuc, C., Rage, E., and Ancelet, S. (2020). A Bayesian hierarchical approach to account for left-censored and missing radiation doses prone to classical measurement error when analyzing lung cancer mortality due to γ-ray exposure in the French cohort of uranium miners. Radiation and Environmental Biophysics, 59(3), 423-437. https ://doi.org/10.1007/s00411-020-00859-6

Belloni, M., Laurent, O., Guihenneuc, C. and Ancelet, S. (2020). Bayesian Profile Regression to Deal With Multiple Highly Correlated Exposures and a Censored Survival Outcome. First Application in Ionizing Radiation Epidemiology. Frontiers Public Health 8 :557006. doi : 10.3389/fpubh.2020.557006

Samson, E., Leuraud, K., Rage, E., Caër-Lorho, S., Ancelet, S., Cléro, E., Bouet, S., Hoffmann, S., Fournier, L., **Belloni, M.**, Jovanovic, I., Bah, T., Davesne, E., Blanchardon, E., Challeton - de Vathaire, C., Laurier, D., Laurent, O. (2018) Bilan de la surveillance épidémiologique des travailleurs du cycle électronucléaire en France. Radioprotection 53(3) :175–184. doi : 10.1051/radiopro/2018026.

Communications orales

Belloni, M., Guihenneuc, C. et Ancelet, S. *Prise en compte d'erreurs de mesure d'exposition combinées à un processus de censure avec une approche hiérarchique bayésienne. Application en épidémiologie des rayonnements ionisants.* 50ème journées de la Société Française de la Statistique (SFdS), Saclay, Mai 2018.

Ancelet, S., Hoffmann, S., **Belloni, M.** et Guihenneuc, C. *Approche hiérarchique bayésienne pour la prise en compte d'erreurs de mesure d'exposition complexes dans les études de cohorte. Application en épidémiologie des rayonnements ionisants.* Journées de Statistique de Rennes (INSA), Rennes, Avril 2019.

Belloni, M., Ancelet, S. et Guihenneuc, C. *Régression bayésienne sur profils d'exposition : application en épidémiologie des rayonnements ionisants.* 51ème journées de la Société Française de la Statistique (SFdS), Nancy, Juin 2019.

Belloni, M., Ancelet, S. et Guihenneuc, C. *Régression bayésienne sur profils d'exposition : application en épidémiologie des rayonnements ionisants.* Séminaire de statistiques du laboratoire MAP5, Université de Paris, Novembre 2019.

Belloni, M., Guihenneuc, C. et Ancelet, S. *Régression bayésienne sur profils d'exposition : application en épidémiologie des rayonnements ionisants.* Appli-BUGS, Visioconférence, Juin 2020.

Belloni, M., Laurent, O., Guihenneuc, C. et Ancelet, S. (2020). *Régression bayésienne sur profils d'exposition : application en épidémiologie des rayonnements ionisants.* 11ème congrès de la Société Francophone Santé Environnement (SFSE), Visioconférence, Novembre 2020.

Ancelet, S., **Belloni, M.**, Laurent O. et Guihenneuc, C. (2020) *Accounting for highly correlated environmental exposures and a censored disease outcome with a Bayesian profile regression mixture model. First application in ionizing radiation epidemiology.* TIES 2020 Virtual conference of the International Environmetrics

Society - Frontiers in Statistics, Epidemiology and the Environment, Visioconférence, Décembre 2020.

Communication affichées

Belloni, M., Guihenneuc, C. et Ancelet, S. *A Bayesian hierarchical approach to account for exposure measurement errors combined with a censoring process when analyzing lung cancer mortality in the French cohort of uranium miners.* International Society for Bayesian Analysis (ISBA), Édimbourg, Royaume-Uni, Juin 2018.

Belloni, M., Guihenneuc, C. et Ancelet, S. *Estimation of lung cancer risk associated to multiple correlated sources of ionizing radiation in the post-55 French cohort of uranium miners.* European Radiation Protection Week (ERPW), Stockholm, Suède, Octobre 2019.

Table des matières

Table des figures

Liste des tableaux

Liste des abréviations

ADN : Acide DésoxyriboNucléique
Bq : Becquerel
CEA : Commissariat à l'Énergie Atomique
CepiDC : Centre d'épidémiologie sur les causes médicales de décès
CIM : Classification Internationale des Maladies
CIPR : Commission Internationale de Protection Radiologique
CIRC : Centre International de Recherche sur le Cancer
COGEMA : COmpagnie GÉnérale des MAtières nucléaires
DIC : Deviance Information Criterion
EHR : Excess Hazard Ratio / Excès de Risque Instantané
ERR : Excès de Risque Relatif
EWAS : Environment-Wide Association Study
GWAS : Genome-Wide Association Studies
Gy : Gray
IC : Intervalle de Confiance ou de Crédibilité (selon le contexte inférentiel)
INSEE : Institut National de la Statistique et des Études Économiques
IPSN : Institut de Protection et de Sûreté Nucléaire
IRSN : Institut de Radioprotection et de Sûreté Nucléaire
$kBq.m^{-3}.h$: kilos Becquerels par mètre cube heure
LCA : Analyse de classe latente (Latent Class Analysis en anglais)

LD : Limite de Détection

LEPID : Laboratoire d'ÉPIDémiologie des rayonnements ionisants

MAR : Missing At Random

MCAR : Missing Completely At Random

MCMC : Monte-Carlo par Chaînes de Markov

MeV : Mégaélectron-Volt

mGy : MilliGray

MNAR : Missing Not At Random

mSv : MilliSievert

OMS : Organisation Mondiale de la Santé

PCR : Régression sur composantes principales

PRM : Profile Regression Mixture model

PUMA : Pooled Uranium Miners Analysis

RI : Rayonnement Ionisant

RNIPP : Répertoire National d'Identification des Personnes Physiques

RPRM : Restricted Profile Regression Mixture model

RR : Risque Relatif

RWMH : Random Walk Metropolis-Hastings

SI : Système International d'unité

SIDI : Système Intégré de Dosimétrie Individuelle

SIMEX : SIMulation EXtrapolation

SMR : Ratio de mortalité standardisé (Standardized Mortality Ratio)

Sv : Sievert

TLD : Dosimètre Thermo Luminescent

UNSCEAR : United Nations Scientific Committee on the Effects of Atomic Radiations - Comité scientifique des Nations Unies pour l'étude des effets des rayonnements ionisants

VIF : Coefficients d'inflation de la Variance (Variance Inflation Factor)

WAIC : Widely Applicable / Watanabe-Akaike Information Criterion

WL : Working Level

WLM : Working Level Month

CHAPITRE 1

Introduction

Depuis les dix dernières années, l'exposome humain est apparu comme un nouveau paradigme de recherche prometteur en épidémiologie (Buck Louis et al. 2013 ; Rappaport, Barupal et al. 2014 ; Vrijheid 2014). Proposé à l'origine par le Dr Christopher Wild en 2005 (Wild 2005), il englobe la totalité des expositions environnementales (i.e., non génétiques) de l'Homme au cours de sa vie - de la conception à la mort. Ce concept, qui plaide pour une prise en compte globale et simultanée de toutes les expositions environnementales de l'Homme (Wild 2012 ; Rappaport et Smith 2010), est le complément clé du génome en matière de compréhension de la santé humaine. Son objectif initial est d'appréhender comment des situations complexes d'exposition environnementale peuvent conduire au développement de pathologies. Son objectif finalisé est de mieux comprendre l'étiologie de pathologies chroniques et multifactorielles afin d'aboutir à de meilleures stratégies de prévention en santé publique.

De toute évidence, les cancers font partie de ces pathologies pour lesquelles le concept d'exposome est essentiel. En effet, ils résultent de l'influence combinée de nombreux facteurs de risque génétiques, environnementaux (i.e., physiques, biologiques, chimiques) et comportementaux auxquels l'Homme est susceptible d'être exposé simultanément et qui peuvent interagir les uns avec les autres (Eiband

et al. 1989 ; Krejs 2010 ; R. S. Lin et Kessler 1981 ; Zabaleta 2012 ; Steliga et Dresler 2011). Ce sont des situations qualifiées de «multi-expositions» voire de «co-expositions» en cas d'expositions simultanées. Les expositions concernées sont plus ou moins fortement influencées par les modes de vie (alimentation, addictions, mobilités...), eux-mêmes fortement conditionnés par le contexte socio-économique. Dans les études épidémiologiques, il est donc important de raisonner dans un cadre d'exposition multifactorielle lorsque des risques de cancer sont estimés ou prédits au niveau individuel ou populationnel. Pourtant, historiquement, les études épidémiologiques se sont surtout concentrées sur la caractérisation de l'effet d'un seul facteur de risque, généralement considéré comme facteur d'intérêt principal pour l'étude (Li et al. 2020 ; Dominici et al. 2010). D'autres facteurs de risque ont été pris en compte avec des modèles de régression standard, mais le plus souvent en raison de leur rôle redouté de facteurs de confusion potentiels. L'effet du facteur d'intérêt principal a ainsi été estimé indépendamment de l'influence potentielle de ces autres facteurs de risque (Greenland, Pearl et Robins 1999 ; Bell, J. Kim et Dominici 2006). Aussi, si certains travaux de recherche ont pu documenter des exemples de synergies ou d'antagonismes suite à des expositions conjointes à différents agents environnementaux, les effets sanitaires d'expositions en mélanges demeurent très mal caractérisés. Seules quelques études ont visé à estimer l'interaction entre l'exposition à un facteur de risque environnemental et d'autres facteurs de risque (ex : le tabagisme et l'amiante ou le radon) (Klebe et al. 2019 ; K. Leuraud et al. 2011), et, plus rarement encore, les effets conjoints de l'exposition à plusieurs facteurs de risque, par exemple environnementaux (ex : les particules ambiantes et l'ozone) (Wu et al. 2015 ; H. Lin et al. 2019).

Dans le domaine spécifique de l'épidémiologie des rayonnements ionisants (RIs), l'estimation des risques de cancers radio-induits et de leur incertitude est un objectif clé depuis des décennies. L'enjeu, pour le système de radioprotection actuel, est d'utiliser cette connaissance, issue des études épidémiologiques, pour fixer des seuils limites d'exposition (United Nations Scientific Committee on the Effects of Atomic Radiation 2017) visant à réduire les risques à un niveau aussi bas que raisonnablement possible, compte-tenu des contraintes socio-économiques. Le suivi épidémiologique des cancers chez les survivants des bombardements atomiques de Hiroshima et Nagasaki a notamment permis de mettre

en évidence une augmentation de la fréquence de certains cancers, en particulier des leucémies, des cancers du sein, du côlon ou du poumon, dans ces populations exposées à des doses modérées à fortes. Si de nombreuses connaissances ont été accumulées sur les effets cancer des expositions aux RIs, des interrogations persistent, en particulier sur les effets des doses faibles et modérées. De même, la manière dont des expositions simultanées à de multiples facteurs de risque radiologiques peuvent affecter les risques de cancers n'a pas encore été étudiée de manière approfondie (National Research Council 2006). Ainsi, l'élaboration de normes de radioprotection reste principalement basée sur un cadre d'exposition mono-factoriel. Dans le cas particulier de l'estimation de risques de cancers radio-induits en situation d'expositions chroniques à faibles doses de RIs, il est ainsi légitime de se demander dans quelle mesure l'estimation d'un risque faible, dans un cadre simplifié d'exposition radiologique supposée unique, reste faible en situation de co-expositions à de multiples sources radiologiques. Plus généralement, dans ce contexte, les enjeux en épidémiologie des RIs sont de : 1/ mieux estimer les effets sanitaires faisant suite à des expositions conjointes prolongées à faibles doses de RIs et à d'autres facteurs de risque ; 2/ mieux appréhender les potentielles interactions entre expositions radiologiques et autres facteurs de risque ; 3/ mettre en perspective ou mieux positionner les effets sanitaires de l'exposition aux RIs par rapport à ceux associés à d'autres facteurs de risque.

Au-delà des difficultés posées par la collecte de données d'expositions multiples de qualité pour les études épidémiologiques (Vrijheid 2014 ; Slama et Vrijheid 2015 ; Lenters, Portengen, Smit et al. 2015), deux grands défis statistiques se posent lors de l'estimation de risques sanitaires en situation de co-expositions environnementales (Billionnet et al. 2012). Tout d'abord, un problème de multicolinéarité survient lorsque plusieurs facteurs de risque d'intérêt sont fortement corrélés. En épidémiologie des RIs, cela peut notamment être le cas, par exemple, lorsqu'un travailleur du cycle du combustible nucléaire est simultanément exposé à plusieurs facteurs de risque radiologiques, chimiques et/ou biologiques, au cours de son activité professionnelle. Dans ce contexte, il est bien connu que l'application de modèles standard de régression multiple - dans lesquels au moins deux prédicteurs sont fortement corrélés - peut conduire à des estimations instables et des coefficients de risque avec une variance élevée. Une telle approche peut donc conduire

à des conclusions trompeuses et à de mauvaises interprétations concernant l'effet de chacun des prédicteurs colinéaires sur la variable réponse d'intérêt (Farrar et Glauber 1967 ; Mela et Kopalle 2002 ; Tu YK 2004). Bien qu'elles ne soient pas encore largement utilisées dans la pratique (Vatcheva et al. 2016), différentes approches statistiques ont été développées pour pallier un problème de multicolinéarité et étudier le potentiel effet combiné de facteurs de risque environnementaux fortement corrélés, sur un évènement d'intérêt (Patel et Butte 2010 ; Rage et al. 2015 ; Lenters, Portengen, Rignell-Hydbom et al. 2016 ; Bottolo et S. Richardson 2010 ; Jain et al. 2018 ; Wold, Ruhe et al. 1984 ; Marshall 2001 ; Patterson, Dayton et Graubard 2002 ; Molitor et al. 2010). Aucune d'entre elles n'a néanmoins été utilisée jusqu'alors en épidémiologie des RIs dans le but d'estimer des risques sanitaires radio-induits. Cela est notamment dû au fait que ces approches sont soit non utilisables tel quel via des packages informatiques publiés, soit difficiles à implémenter, soit non pertinentes pour estimer un risque et son incertitude. En épidémiologie, le deuxième grand défi statistique posé par l'estimation de risques sanitaires en situation de co-expositions environnementales concerne le fait que les données d'exposition collectées soient inévitablement entachées d'incertitude, ce qui est bien sûr déjà le cas en situation d'exposition unique (Carroll 2005 ; Thomas, D. Stram et Dwyer 1993). Cette incertitude, associée à un manque de connaissances concernant la «vraie» valeur des expositions d'intérêt, a différentes origines - appelées sources d'incertitude par la suite - parmi lesquelles : a) les erreurs de mesure sur les expositions qui surviennent lorsque les expositions d'intérêt ne peuvent être mesurées avec précision et que seules des mesures de substitution imparfaites sont disponibles ; b) l'existence d'une limite de détection i.e., d'une plus petite valeur d'activité détectable (avec une incertitude acceptable) pour chaque appareil utilisé pour mesurer les expositions. Ces limites donnent lieu à des données d'exposition potentiellement censurées à gauche ; c) l'existence de données d'exposition manquantes. Si non ou mal prises en compte, ces diverses sources d'incertitude, qui sont omniprésentes dans les études observationnelles, peuvent contribuer à remettre en question la validité de l'inférence statistique dans les études épidémiologiques (Thomas, D. Stram et Dwyer 1993 ; H.-M. Kim, Yasui et Burstyn 2006 ; Physick et al. 2007). Elles peuvent mener à des estimateurs de risque biaisés, une déformation des relations dose-réponse et

à une sur- ou sous-estimation de l'incertitude associée aux estimations de risque (Carroll 2005 ; Carroll et al. 2006 ; Lubin, Colt et al. 2004 ; Xue, M. Y. Kim et Shore 2006). En épidémiologie des RIs, cela peut conduire à ne pas détecter certaines associations déjà faibles entre expositions chroniques à faibles doses de RIs et risques de cancers ou, peut-être même pire, à détecter par erreur des associations importantes là où il n'y en a pas. Malgré leurs conséquences potentiellement délétères et l'existence de différentes approches statistiques proposées dans la littérature (Keogh et al. 2020 ; Shaw et al. 2020 ; Hoffmann, Estelle Rage et al. 2017 ; Andrew Gelman et al. 2013), les différentes sources d'incertitude pré-citées sont généralement non ou mal prises en compte en épidémiologie des RIs (Rage et al. 2015 ; Kreuzer, Dufey et al. 2013 ; Klervi Leuraud, David B Richardson et al. 2015 ; Zablotska et al. 2018 ; Ron 1998 ; Laurent et al. 2016 ; Yoder et al. 2018). Le manque de flexibilité de certaines approches statistiques existantes pour tenir compte de sources d'incertitude multiples et hétérogènes et l'absence de logiciel suffisamment simple d'utilisation en sont les raisons principales.

Le cas d'étude considéré dans cette thèse a porté sur l'analyse de l'association entre expositions radiologiques à faibles doses et mortalité par cancer du poumon dans la cohorte française des mineurs d'uranium. En effet, dans le cadre de leur activité professionnelle, les mineurs d'uranium sont soumis, de manière chronique, à un ensemble d'expositions radiologiques à faibles doses (Vacquier, Estelle Rage et al. 2011 ; Estelle Rage, Vacquier et al. 2012) dont le radon et ses descendants à vie courte, les rayonnements gamma et des poussières d'uranium en suspension dans l'atmosphère. Toutes ces expositions sont associées au même phénomène de désintégration radioactive de l'uranium-238 qui est omniprésent dans les mines d'uranium (Vacquier, Estelle Rage et al. 2011). Elles sont ainsi fortement corrélées. À ce stade, un effet additif ou synergique de la co-exposition à ces différentes sources radiologiques, sur les risques de cancer du poumon, ne peut donc être exclu. Jusqu'à présent, la plupart des études épidémiologiques portant sur la cohorte française des mineurs d'uranium se sont concentrées sur l'étude de l'association entre la mortalité par cancer du poumon et une exposition chronique et à faibles doses au radon (et ses descendants à vie courte), comme si ce dernier - reconnu comme cancérigène pulmonaire chez l'homme depuis 1988 (IARC et al. 1988) et comme deuxième cause de cancer du poumon après le tabagisme (J. M. Samet

1989 ; Birchall et J. Marsh 2005) - avait un effet nécessairement indépendant des autres sources radiologiques auxquelles sont exposés les mineurs. Seuls quelques travaux ont été menés afin d'estimer l'impact du radon (et ses descendants à vie courte), des poussières d'uranium et des rayonnements gamma sur le risque de décès par cancer du poumon dans la cohorte française des mineurs d'uranium (Rage et al. 2015 ; Vacquier, Estelle Rage et al. 2011 ; Estelle Rage, Vacquier et al. 2012). Ces travaux n'ont permis ni d'estimer l'effet combiné de ces trois expositions radiologiques ni d'appréhender les potentielles interactions entre expositions au regard du risque d'intérêt. Par ailleurs, ces travaux ne tenaient pas compte des incertitudes de mesure associées aux expositions radiologiques. Hoffmann *et al.* (Hoffmann, Estelle Rage et al. 2017) ont proposé une approche hiérarchique bayésienne pour tenir compte des erreurs de mesure sur les expositions au radon (et ses descendants à vie courte) dans l'estimation du risque de décès par cancer du poumon, dans la cohorte française des mineurs d'uranium. Ces travaux ont notamment permis d'illustrer la très grande flexibilité de l'approche hiérarchique bayésienne pour traiter des structures complexes d'erreurs de mesure, telles que celles rencontrées dans les études de cohortes professionnelles en épidémiologie des RIs. Une potentielle limite de ces travaux au regard de l'estimation du risque d'intérêt est d'avoir considéré toutes les expositions nulles ou manquantes de la cohorte française des mineurs d'uranium comme la marque d'une absence d'exposition. Par ailleurs, les co-expositions radiologiques aux rayonnements gamma et aux poussières d'uranium ont été ignorées.

L'objectif principal de ce travail était de promouvoir l'utilisation de l'approche hiérarchique bayésienne pour l'estimation d'un risque de cancer radio-induit à faibles doses (et de son incertitude) en situation d'expositions radiologiques multiples, fortement corrélées et entachées de sources d'incertitude multiples et hétérogènes. L'enjeu est ainsi de contribuer à améliorer la connaissance des effets des expositions à de faibles doses de RIs. Dans le cadre de cette thèse, les problèmes de multicolinéarité et de prise en compte des incertitudes d'exposition dans la cohorte française des mineurs d'uranium ont été traités séparément, afin d'ouvrir la voie à la mise en oeuvre de modèles plus complexes intégrant simultanément ces deux dimensions. Dans un premier temps, l'objectif était de proposer différents modèles hiérarchiques pour tenir compte des incertitudes sur l'exposition aux rayonnements

gamma dont les erreurs de mesure sont différentes de celles associées à l'exposition au radon (et ses descendants à vie courte), traitées dans Hoffmann *et al.* (ibid.), et aux poussières d'uranium (très similaires à celles du radon). Il s'agissait également de conduire l'inférence bayésienne de ces modèles afin d'obtenir une estimation corrigée du risque de décès par cancer du poumon associé à une exposition chronique aux rayonnements gamma dans cette cohorte. Dans un deuxième temps, l'objectif était d'apporter une première réponse, souple et élégante, au problème des co-expositions radiologiques lors de l'estimation du risque de décès par cancer du poumon dans la cohorte française des mineurs d'uranium. Plus spécifiquement, il s'agissait d'étendre la classe des modèles de mélange par régression bayésienne sur profils d'exposition (Molitor et al. 2010 ; Silvia Liverani et al. 2015) aux modèles de survie classiquement utilisés en épidémiologie des RIs. L'approche par modélisation hiérarchique, basée sur une combinaison de sous-modèles reposant sur des hypothèses d'indépendances conditionnelles, fournit un cadre de travail flexible pour la description de sources d'incertitude multiples et hétérogènes. Le choix du paradigme bayésien permet quant à lui de mener l'inférence de ces modèles complexes dans un cadre générique, naturel et cohérent vis à vis de l'apprentissage de structures reposant sur des hypothèses d'indépendances conditionnelles. Il permet également de pouvoir injecter des connaissances *a priori* sur certains paramètres inconnus des modèles considérés via la littérature ou des dires d'experts. Cela peut faciliter leur estimation en situation où les données disponibles sont faiblement informatives, ce qui est le cas lorsqu'on s'intéresse au risque de décès par cancer du poumon dans la cohorte française des mineurs d'uranium.

Le chapitre 2 du manuscrit présente des éléments de contexte en épidémiologie des RIs ainsi que le cas d'étude qui a servi de fil conducteur pour tous les développements méthodologiques réalisés dans le cadre de cette thèse. Le chapitre 3 du manuscrit illustre - à partir des données de la cohorte française des mineurs d'uranium - les trois difficultés posées par l'analyse statistique de données d'expositions radiologiques qui ont fait l'objet de travaux dans cette thèse. Il s'agit : a) des erreurs de mesure sur les expositions aux RIs ; b) des données d'exposition manquantes ou censurées à gauche ; c) de la multicolinéarité. Le chapitre 4 fournit une courte présentation des principaux outils mathématiques utilisés dans ce travail à savoir la modélisation hiérarchique et la statistique bayésienne. Le chapitre

5 porte sur la prise en compte des incertitudes sur les expositions aux rayonnements gamma dans l'estimation du risque de décès par cancer du poumon dans la cohorte française des mineurs d'uranium. Le chapitre 6 porte sur la prise en compte des co-expositions au radon, aux poussières d'uranium et aux rayonnements gamma dans l'estimation du risque de décès par cancer du poumon chez les mineurs d'uranium français. Le chapitre 7 présente les principaux résultats obtenus parmi lesquels figurent les résultats d'une importante étude par simulations ainsi que les estimations de risque de décès par cancer du poumon après prise en compte des incertitudes sur les expositions aux rayonnements gamma d'une part et des co-expositions radiologiques d'autre part. Enfin, le chapitre 8 fait la synthèse et permet de discuter des avantages et limites de ce travail. Il présente également des perspectives possibles.

CHAPITRE 2

Expositions aux rayonnements ionisants dans les cohortes de mineurs d'uranium

Ce chapitre a pour objectifs d'apporter quelques éléments de contexte, nécessaires à la compréhension de ces travaux, concernant l'exposition humaine aux RIs puis de décrire le cas d'étude en épidémiologie qui a servi de fil conducteur pour tous les développements méthodologiques réalisés dans le cadre de cette thèse.

2.1 Les rayonnements ionisants : généralités

2.1.1 Qu'est-ce qu'un rayonnement ionisant ?

Un atome est composé d'un noyau autour duquel gravitent un ou plusieurs électrons (Gooch 2007). Le noyau contient deux types de particules subatomiques (i.e., de taille inférieure à celle d'un atome) : les protons chargés positivement et les neutrons qui n'ont pas de charge électrique (Jevremovic 2005). Les électrons, quant à eux, sont chargés négativement. Dans la nature, la plupart des noyaux d'atomes sont stables mais certains atomes ont des noyaux instables. Cela peut être dû à un excès de protons, de neutrons, ou encore à un excès des deux. Pour

acquérir une meilleure stabilité, ces noyaux instables appelés radionucléides expulsent une quantité d'énergie à un moment donné - appelé désintégration - sous forme de rayonnements et/ou de particules : ce phénomène est appelé radioactivité. Un rayonnement est ainsi qualifié de ionisant lorsqu'il est capable de transférer aux atomes qu'il croise une énergie suffisante pour leur arracher un électron (*La radioprotection en milieu hospitalier* p. d.), comme illustré dans la Figure 2.1. Cette altération peut déstabiliser les molécules constituant les cellules du vivant, induisant des perturbations biochimiques. Les conséquences biologiques peuvent être importantes, particulièrement lors de lésions cytoplasmiques et membranaires. Il peut s'agir d'effets déterministes, aussi appelés «réactions tissulaires» (e.g., brûlure cutanée, perte de cheveux, syndrome hématopoïétique), qui sont liés à un processus de mort cellulaire et peuvent survenir de façon précoce. Il peut également s'agir d'effets stochastiques, survenant de manière aléatoire et tardive (quelques années à plusieurs décennies), comme le développement de tumeurs lorsque la molécule d'Acide désoxyribonucléique (ADN) est touchée.

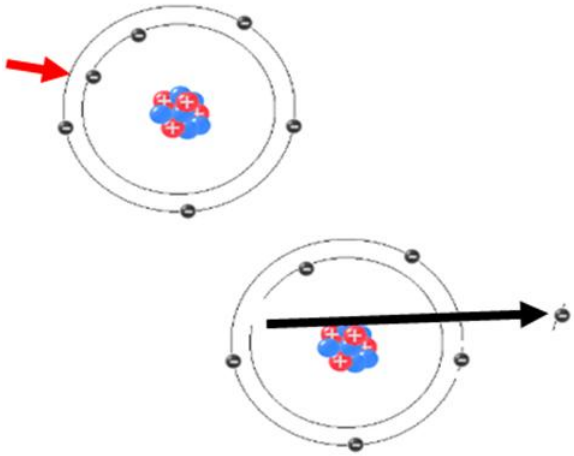

FIGURE 2.1 – Un rayonnement ionisant arrachant un électron à un atome. La flèche rouge représente le rayonnement ionisant et la flèche noire représente l'électron arraché.

2.1.2 Les principaux types de rayonnements ionisants

Cinq principaux types de RIs peuvent être émis lors d'une désintégration radioactive. Ceux-ci se distinguent par la quantité d'énergie transférée et, comme illustré dans la Figure 2.2, par leur pouvoir de pénétration respectif dans la matière traversée (ibid.) :

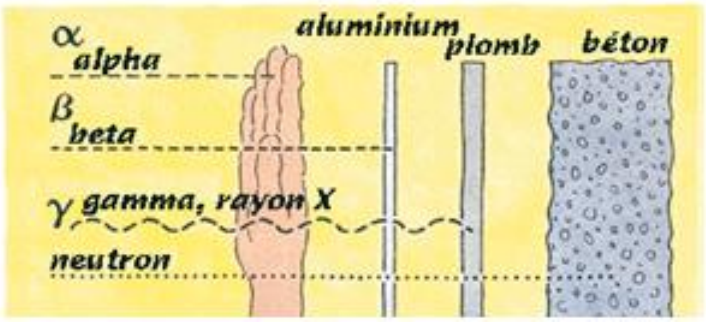

FIGURE 2.2 – Pouvoir de pénétration de différents types de rayonnements ionisants

- Les rayonnements alpha - notés rayonnements α par la suite - sont de nature particulaire. Ils sont constitués de noyaux d'hélium, eux-mêmes composés de deux protons et deux neutrons, et sont émis par les noyaux d'atomes trop chargés en protons et en neutrons (noyaux lourds). Ils sont libérés, par exemple, lors de la désintégration de l'uranium 238 présent à l'état naturel dans la croûte terrestre, donnant du thorium 234. Ces rayonnements ont une faible capacité de pénétration dans la matière (quelques dizaines de microns) : une feuille de papier suffit à les arrêter. Néanmoins, comme ces particules se diffusent très peu dans la matière, toute leur énergie est déposée de façon localisée : elles sont donc hautement ionisantes et seront particulièrement nocives pour la santé lors d'une contamination interne au cours de laquelle la source de rayonnement α est en contact permanent avec l'organisme ;

- Les rayonnements bêta - notés rayonnements β par la suite - sont également de nature particulaire. Ils sont soit : a/ émis lors de la transformation d'un neutron en proton et alors constitués d'un électron et d'un antineutrino (appelés dans ce cas, rayonnement $\beta-$) soit b/ émis lors de la transformation

d'un proton en neutron et alors constitués d'un positon et d'un neutrino (appelés dans ce cas, rayonnement $\beta+$). À titre d'exemple, le césium 137 émet une particule $\beta-$ quand il se désintègre en barium 137. Ces particules sont moins ionisantes que les particules α mais ont une plus grande capacité de pénétration dans la matière (jusqu'à un centimètre) : elles peuvent traverser une main mais être en revanche arrêtées par une feuille d'aluminium ;

- Les rayonnements gamma - notés rayonnements γ par la suite - sont des ondes électromagnétiques portées par des photons qui sont des particules électriquement neutres. Ils sont souvent précédés par l'émission de particules α ou β qui laissent le noyau dans un état excité. C'est par exemple le cas du nickel 60 qui provient de la désintégration du cobalt 60 via l'émission d'un rayonnement $\beta-$ et de deux photons. L'émission d'un rayonnement γ permet alors au noyau de retourner vers un état plus stable. Le ré-agencement des nucléons dans les couches nucléaires émet également des rayonnements γ. Les rayonnements γ sont plus faiblement ionisants que les rayonnements α et β. En revanche, ils ont une très forte capacité de pénétration qui dépend de l'énergie du rayonnement. Ils peuvent parcourir quelques kilomètres dans l'air et peuvent être arrêtés par une forte épaisseur de plomb ;

- Les rayonnements X sont des ondes électromagnétiques, essentiellement produites par des moyens artificiels tel que la radiographie. Ils sont de même nature que les rayonnements γ mais se distinguent par leur origine : les rayonnements X sont émis par le réarrangement des électrons autour du noyau alors que les rayonnements γ sont émis par le noyau. Comme les rayonnements γ, les rayonnements X sont moins ionisants que les rayonnements α et β. Ils sont en revanche plus pénétrants dans la matière ;

- Les neutrons sont de nature particulaire. Comme ils ne sont pas chargés électriquement, les neutrons ne sont pas directement ionisants. Cependant, du fait de leur forte masse, ils peuvent interagir avec les noyaux des atomes, les rendant instables et provoquant différentes réactions nucléaires (e.g., fission), elles-mêmes sources de RIs. Un neutron peut, par exemple, heurter un proton et lui transmettre toute son énergie, provoquant ainsi une forte ionisation.

Les rayonnements neutroniques ont un important pouvoir pénétrant dans la matière : ils peuvent parcourir quelques kilomètres dans l'air et être bloqués par un mètre d'eau ou un mur de béton dopés en absorbants neutroniques.

2.1.3 De l'exposition humaine à la dose

L'ensemble de la population humaine est exposé tout au long de sa vie aux RIs. Comme illustré dans la Figure 2.3, cette exposition provient de multiples sources - d'origine naturelle ou artificielle - que ce soit dans le cadre de l'environnement résidentiel, de différentes activités (professionnelles ou autres, comme lors de voyages en avion) ou pour des raisons médicales. L'exposition aux rayonnements d'origine naturelle provient principalement de quatre sources : les rayonnements cosmiques, les rayonnements terrestres dit telluriques, l'ingestion de radionucléides présents en petite quantité dans les aliments et l'eau potable et l'inhalation de radon, un gaz radioactif qui se forme lors de la désintégration de l'uranium naturellement présent dans la croûte terrestre. L'exposition aux rayonnements cosmiques augmente avec l'altitude, l'exposition aux rayonnements telluriques varie en fonction de la géologie et l'exposition au radon dans l'habitat dépend de la géologie, de la construction des bâtiments et du mode de vie des ménages (aération des pièces...). En outre, des sources artificielles d'exposition se sont développées au cours du siècle dernier. Elles résultent principalement de l'utilisation des RIs à visée thérapeutique et/ou diagnostique dans le secteur médical et, dans une plus faible proportion par rapport au secteur médical, de l'énergie nucléaire (ex : retombées des accidents et des anciens essais nucléaires atmosphériques), de la recherche et de l'industrie (ex : rejets des installations nucléaires).

Il existe deux modalités d'exposition aux RIs pour l'Homme : exposition externe et contamination interne. Lors d'une exposition externe, la source radioactive se trouve à l'extérieur du corps (ex : irradiation par des rayonnements X lors d'une radiographie). L'exposition peut alors concerner le corps entier ou seulement une partie de celui-ci (ex : un tissu, un organe) selon le champ d'irradiation. L'exposition cesse dès que l'organisme n'est plus situé dans le champ de la source de RIs. Lors d'une contamination interne, la source radioactive se trouve à l'intérieur du corps suite à l'incorporation de radionucléides par inhalation (ex : radon), in-

gestion ou voie percutanée (ex : après une blessure) ou encore, suite à l'injection de radionucléides dans le cadre de procédures de médecine nucléaire (ex : iode). Les radionucléides se répartissent alors dans l'organisme. La contamination interne cesse lorsque le ou les radionucléides incorporés sont éliminés de l'organisme, grâce à la combinaison d'une décroissance de leur activité au cours du temps et de leur élimination naturelle par excrétion.

Le becquerel (Bq) est l'unité de mesure traditionnelle du Système International d'unités (SI) permettant de quantifier le niveau d'activité d'une certaine quantité de matière radioactive. Ainsi, pour une source radioactive donnée, une activité de 1 Bq signifie qu'une désintégration par seconde s'y produit (Domenech 2017). Cette unité de mesure n'est pas adaptée pour refléter les effets sanitaires potentiels d'une exposition aux RIs. En effet, les dommages biologiques causés par les RIs dépendent d'un certain nombre de facteurs comme : a/ la quantité d'énergie déposée en un point (ex : cellule, tissu, organe, organisme entier) et le débit de dose (ex : dose délivrée de façon aigue ou chronique) associés aux conditions d'exposition ; b/ le type de RIs impliqués ; c/ la radiosensibilité des cellules/tissus irradiés (« Dosimetry and Biological effects of Ionizing Radiation » 2004). Afin de tenir compte de ces différents facteurs, la Commission Internationale de Protection Radiologique (CIPR) a défini différentes quantités dosimétriques (Protection 2007). Nous évoquerons ici trois d'entre elles :

- La dose absorbée à un organe/tissu mesure l'énergie moyenne déposée par unité de masse de cet organe/tissu, par un RI. Elle s'exprime en gray (Gy) : 1 Gy correspond au dépôt d'un joule d'énergie par kilogramme de matière (ibid.). Il s'agit d'une quantité physique qui peut être mesurée. Elle est utilisée pour tous les types de RIs et pour toutes les situations d'irradiation ;

- La dose équivalente à un organe/tissu prend en compte la l'effet relatif des différents types de RIs pour la matière vivante. Elle s'exprime en sievert (Sv) et est égale à la dose absorbée moyenne pondérée par un facteur de pondration radiologique W_R fixé par la CIPR pour chaque type de RI (ibid.). Par exemple, pour les photons (*e.g.*, rayonnements X et γ), le W_R est égal à 1 ; pour les rayonnements α, il est égal à 20 ; pour les neutrons, le W_R est compris entre 5 et 20 selon l'énergie des neutrons considérés. La dose

équivalente est une grandeur de gestion de risque qui permet de quantifier le dommage biologique à un tissu ou à un organe. Elle ne peut pas être mesurée ;

- La dose efficace est généralement utilisée lorsque l'exposition à plusieurs organes ou au corps entier est considérée. Elle permet de prendre en compte les différences de radiosensibilité des tissus irradiés. Également exprimée en Sv, elle est obtenue en multipliant la dose équivalente à chaque organe/tissu par un facteur de pondération tissulaire W_T spécifique à l'organe/tissu, puis en sommant ces produits sur l'ensemble des organes. Les W_T sont déterminés par la CIPR (ibid.). La dose efficace est définie pour les besoins de la radioprotection.

Le Comité scientifique des Nations Unies pour l'étude des effets des RIs (UNSCEAR) a proposé une classification des niveaux de doses (United Nations Scientific Committee on the Effects of Atomic Radiation 2012). Il définit ainsi comme «faibles» doses les doses inférieures à 100 milliGray (mGy). Cette définition s'applique à des rayonnements à faible transfert d'énergie linéique [1], pour des doses au corps entier ou pour des organes ou tissus spécifiques. Le terme de faible débit de dose est utilisé pour des débits de dose inférieurs à 0.1 mGy par minute, en moyenne, sur une heure.

En France, selon le bilan IRSN 2015 illustré avec la Figure 2.3, la dose efficace annuelle moyenne reçue par la population est de 4.5 mSv dont 64% est d'origine naturelle. L'exposition environnementale de la population française provient principalement de l'inhalation de radon présent dans l'air inspiré (32% de la dose totale), mais également de l'exposition externe aux rayonnements telluriques (14%) et cosmiques (7%) et enfin de l'ingestion de radionucléides présents en petite quantité dans l'eau et les aliments (12%). L'exposition médicale due aux examens à visée diagnostique ou thérapeutique représente un peu plus d'un tiers de la dose moyenne annuelle reçue (35%). Enfin, les sources industrielles constituent une partie mineure de l'exposition de la population française.

1. Quantité qui décrit l'énergie transférée par une particule ionisante traversant la matière, par unité de distance. Il varie selon la nature et l'énergie du rayonnement

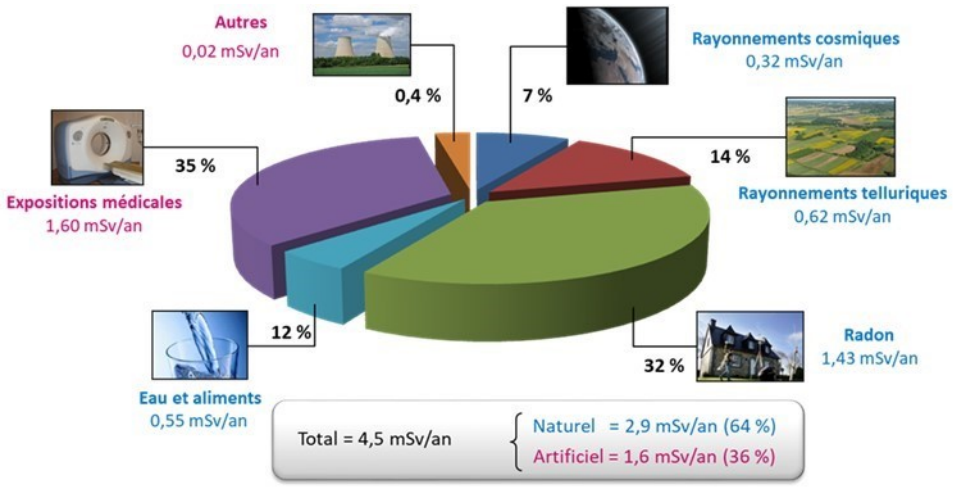

FIGURE 2.3 – Exposition moyenne de la population française aux rayonnements ionisants (Source IRSN 2015)

2.2 Expositions radiologiques dans les cohortes de mineurs d'uranium

Des taux de mortalité élevés ont été observés dès le 15ème siècle au sein des populations de travailleurs des mines de fond d'Europe centrale et ce, bien avant la découverte de la radioactivité (IONISING RADIATION (AGIR) 2009). Ce n'est qu'en 1879 que la cause des maladies pulmonaires dont étaient décédés ces travailleurs a été identifiée : il s'agissait du cancer du poumon (A. C. GEORGE 2008 ; ROBINSON 2015 ; G. SACCOMANNO et al. 1964). Après cette découverte, il a fallu 70 ans pour reconnaître que cette mortalité élevée associée au cancer du poumon pouvait être causée par l'inhalation de produits de filiation à vie courte du radon présents dans les endroits clos tels que les mines de fond (IONISING RADIATION (AGIR) 2009) et presque 40 ans de plus avant que le Centre international de recherche sur le cancer (CIRC) ne reconnaisse le radon comme cancérigène pulmonaire certain chez l'homme (IARC et al. 1988). Si les concentrations en radon étaient élevées dans les premières années d'exploitation des mines, la mise en place de normes de radioprotection et notamment de la ventilation forcée a entraîné une forte diminution de ces concentrations au début des années 1950 pour certains sites miniers (R. S. ALLODJI, Klervi LEURAUD et al. 2012 ; ROGEL et al. 2002 ;

Geno Saccomanno et al. 1988) et à la fin des années 1960 (Navaranjan et al. 2016) pour d'autres. Ainsi, sur les dernières années d'exploitation, l'exposition annuelle au radon des mineurs travaillant dans les mines de fond est globalement comparable aux niveaux d'exposition mesurés dans la plupart des habitations humaines (M. Tirmarche et al. 2012). Parallèlement, comme nous le verrons plus en détails dans ce chapitre, les mineurs sont co-exposés à de nombreux radionucléides et à plusieurs types de RIs dans le cadre de leur activité professionnelle. Une surveillance dosimétrique ayant été instaurée relativement tôt dans les mines, les mineurs d'uranium constituent une population pertinente pour l'étude des effets sanitaires à long terme d'une exposition chronique et à faibles doses aux RIs. Plusieurs cohortes de mineurs de fond ont ainsi été mises en place à l'échelle internationale puis suivies, dont neuf cohortes de mineurs d'uranium : la cohorte française dont il sera question dans ce travail de thèse, la cohorte tchèque, la cohorte allemande (aussi appelée cohorte de la Wismut), la cohorte de Radium Hill (Australie), trois cohortes canadiennes (cohorte de l'Ontario, de Beaverlodge et de Port Radium), deux cohortes américaines (cohorte du plateau du Colorado et du Nouveau-Mexique). Alors que l'ensemble de ces cohortes a mis en évidence une relation entre le risque de décès par cancer du poumon et l'exposition cumulée au radon, la question des effets liés aux RIs autres que le radon se pose.

2.2.1 Un contexte de co-expositions associé à la désintégration radioactive de l'uranium-238

L'activité radiologique au sein d'une mine provient de la désintégration radioactive naturelle de l'uranium-238 présent dans les sols, les roches et l'eau (Emsley 2011). L'uranium-238 (U_{92}^{238}) est l'isotope [2] de l'uranium le plus courant dans la nature : il représente plus de 99% de l'uranium naturel et sa demie-vie [3] est de 4 468 milliards d'années. La chaîne de désintégration radioactive (cf. section 2.1.1) de l'uranium-238 est l'une des trois chaînes de désintégration se produisant naturellement sur Terre (Griffin 2011). La Figure 2.4 représente les nombreux

2. Plusieurs atomes d'un même élément chimique qui ont le même nombre de protons mais un nombre différent de neutrons sont appelés isotopes

3. La demie-vie d'un radio-isotope est le temps moyen nécessaire à la moitié des noyaux de ce radio-isotope pour se désintégrer

produits de filiation (encore appelés descendants) de l'uranium-238, leur demi-vie ainsi que le type le plus probable de RIs émis lors de leur désintégration respective. L'uranium-238, le thorium-234, le protactinium-234, l'uranium-234, le thorium-230 et le radium-226 sont tous à l'état solide dans les mines. Ils se désintègrent principalement en émettant des rayonnements α ou β qui, compte-tenu de leur faible capacité de pénétration dans la matière (cf. section 2.1.2), ne constituent pas une exposition externe dangereuse pour les mineurs. En revanche, ces radio-isotopes peuvent être inhalés et se déposer dans les poumons lorsqu'ils sont présents dans la poussière en suspension, ce qui est notamment le cas dans les chantiers fortement empoussiérés où la teneur en minerai est élevée. Par ailleurs, ils peuvent provoquer une exposition externe aux rayonnements γ - ce type de RIs suivant généralement l'émission de rayonnements α et β. Contrairement aux autres produits de filiation de l'uranium-238, le radon-222 ($^{222}_{86}$Rn) est un gaz - incolore, insipide et inodore - qui peut être inhalé par les mineurs. Cependant, en tant que gaz noble, il a une très faible réactivité chimique (Lecomte 2012) si bien que la majeure partie du gaz inhalée est rapidement exhalée (James W Marsh et al. 2010). En revanche, comme le montre la Figure 2.4, le radon-222 se désintègre en une série de radio-isotopes à vie courte comme le polonium-218 (^{218}Po) et le plomb-214 (^{214}Po) qui peuvent former des grappes en se fixant aux particules d'aérosol présentes dans l'atmosphère (J. Marsh et al. 2012). Lorsqu'ils sont inhalés, ils peuvent ainsi se déposer dans les poumons où ils peuvent se désintégrer en émettant des rayonnements α, β et γ. En particulier, le polonium-218 (^{218}Po) et le plomb-214 (^{214}Po) peuvent entraîner de fortes doses de rayonnements α au niveau du poumon (Jelle 2012). Par conséquent, ce n'est donc pas tant l'exposition au radon-222 qui est dangereuse pour les mineurs mais plutôt l'exposition aux produits de filiation du radon, également appelés descendants du radon. Dans un souci de simplicité et parce qu'il est courant de procéder ainsi dans la littérature, nous utiliserons simplement le terme «radon» pour faire référence au radon et ses descendants à vie courte dans la suite de ce manuscrit.

Dans le cadre de leur activité professionnelle, les mineurs de fond sont donc simultanément exposés à plusieurs radionucléides descendants de l'uranium-238 et plusieurs types de RIs (α, β et γ). Parmi ces différentes co-expositions radiologiques, issues du même mécanisme de désintégration radioactive de l'uranium-238,

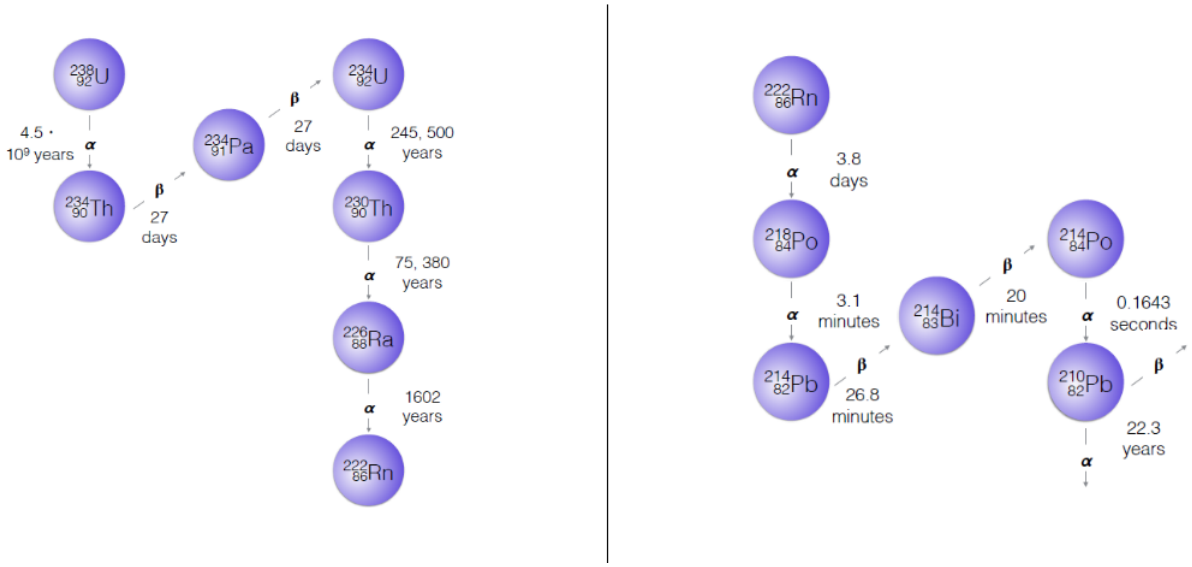

FIGURE 2.4 – Chaîne de désintégration radioactive de l'uranium-238. Dans un premier temps (figure gauche), l'uranium-238 se désintègre en thorium-234 ($^{234}_{90}$Th), protactinium-234 ($^{234}_{91}$Pa), uranium-234 ($^{234}_{92}$U), thorium-230 ($^{230}_{90}$Th), radium-226 ($^{226}_{88}$Ra) et radon-222 ($^{222}_{86}$Rn). Puis (figure droite), le radon-222 se désintègre en polonium-218 ($^{218}_{84}$Po), plomb-214 ($^{214}_{82}$Pb), bismuth-214 ($^{214}_{83}$Bi), polonium-214 ($^{214}_{84}$Po), plomb-210 ($^{210}_{82}$Pb) puis en radionucléides à vie longue non représentés ici.

trois sont potentiellement cancérigènes pour l'Homme :

- le radon pénétrant dans l'organisme par inhalation (contamination interne) ;
- les poussières fines de minerai, en suspension dans l'atmosphère, qui sont chargées des éléments de la chaîne de désintégration de l'uranium-238 émetteurs α à vie longue : $^{238}_{92}$U, $^{234}_{92}$U, $^{230}_{90}$Th, $^{226}_{88}$Ra, $^{210}_{82}$Pb. Ces poussières appelées «poussières d'uranium» ou parfois même «poussières» par la suite pénètrent dans l'organisme par inhalation (contamination interne) ;
- les rayonnements γ (exposition externe).

En raison de l'omniprésence de l'uranium dans les mines, tous les mineurs de fond sont chroniquement exposés aux descendants de l'uranium, et ce, quel que soit le type de mine. Néanmoins, la concentration en uranium dans le minerai extrait des mines d'uranium est généralement plus élevée que dans les autres types de mines.

2.2.2 Estimation des expositions radiologiques dans les mines

Dans la plupart des cohortes de mineurs d'uranium, les expositions aux RIs ont principalement été évaluées à partir de trois techniques : une estimation rétrospective des expositions pour les premières années d'exploitation des mines, des stratégies d'évaluation groupée des expositions, basée sur des mesures ambiantes puis, pour les années les plus récentes, des mesures individuelles d'exposition avec des dosimètres personnels portés par les mineurs d'uranium.

2.2.2.1 Estimation rétrospective des expositions

Lorsque l'exploitation de l'uranium a débuté, au lendemain de la seconde guerre mondiale (Waggitt 2008), les risques sanitaires et environnementaux potentiellement associés à l'exposition aux RIs étaient méconnus (R. S. Allodji, Klervi Leuraud et al. 2012 ; Lane et al. 2010). Il n'y avait donc pas d'évaluation systématique de l'exposition aux RIs dans les mines d'uranium (R. S. Allodji, Klervi Leuraud et al. 2012 ; Lane et al. 2010). Pour le besoin des études épidémiologiques, les valeurs d'exposition reçues au cours de ces premières années d'exploitation de l'uranium ont généralement été estimées rétrospectivement soit par

dires d'experts (R. S. Allodji, Klervi Leuraud et al. 2012 ; Michaela Kreuzer et al. 2010 ; Vacquier, Agnès Rogel et al. 2009) soit en utilisant des données d'exposition collectées sur des années plus récentes (Navaranjan et al. 2016 ; Ladislav Tomasek et al. 2008). Dans la cohorte française des mineurs d'uranium, par exemple, un groupe d'experts des conditions d'exposition dans les mines a estimé les niveaux d'exposition mensuels au radon pour les années antérieures à 1956 et ce, pour différents sites miniers. Ils se sont basés sur les caractéristiques du minerai, le type de ventilation des sites et quelques mesures disponibles de concentration en radon (Tirmarche, Raphalen et al. 1993 ; Tirmarche, Brenot et al. 1985). Les estimations de l'exposition mensuelle ont ensuite été multipliées par le nombre de mois pendant lesquels un mineur a travaillé dans chaque site minier afin d'obtenir des estimations individuelles de l'exposition annuelle (Tirmarche, Brenot et al. 1985). Les niveaux d'exposition aux rayonnements γ et aux poussières d'uranium n'ont pas été estimés pour les années antérieures à 1956 dans la cohorte française des mineurs d'uranium. En revanche, les niveaux d'exposition aux rayonnements γ ont été estimés rétrospectivement pour les plus anciennes années d'exploitation de l'uranium dans les cohortes tchèque et allemande de mineurs d'uranium (Estelle Rage, David B Richardson et al. 2020).

2.2.2.2 Stratégies d'évaluation groupée, basée sur des mesures ambiantes

Dans les années 1950, des méthodes indirectes de mesures de l'exposition au radon ont été mises en place dans les mines d'uranium de plusieurs pays, à des fins de radioprotection (Geno Saccomanno et al. 1988 ; Lane et al. 2010). Elles ont été introduites en 1949 dans la cohorte tchèque (Ladislav Tomasek et al. 2008), en 1955 dans la cohorte de la Wismut (Michaela Kreuzer et al. 2010), en 1956 dans la cohorte française (R. S. Allodji, Klervi Leuraud et al. 2012) et en 1958 dans la cohorte de l'Ontario (Navaranjan et al. 2016). Ces méthodes indirectes étaient basées sur une stratégie d'évaluation groupée des expositions à partir de mesures ambiantes. Dans la cohorte française des mineurs d'uranium, par exemple, un certain nombre de mesures ambiantes hebdomadaires de la concentration en radon étaient réalisées en différents endroits dans les mines. Ces mesures étaient

effectuées avec des flacons à scintillation (R. S. Allodji, Klervi Leuraud et al. 2012 ; Navaranjan et al. 2016). Dans la cohorte de la Wismut, des mesures ambiantes ont été utilisées pour définir des matrices emplois-expositions [4] (Michaela Kreuzer et al. 2010). Dans les autres cohortes, des mesures ambiantes ont été utilisées pour estimer des expositions individuelles au radon en multipliant le temps passé par un mineur dans une certaine zone de la mine par l'exposition estimée dans cette zone. Des estimations individuelles annuelles de l'exposition cumulée au radon pouvaient ensuite être obtenues en additionnant ces termes pour toutes les zones dans lesquelles un mineur avait travaillé pendant l'année (Lane et al. 2010). Le nombre de mesures effectuées a généralement augmenté au fil des années, ce qui a permis d'obtenir des estimations de plus en plus précises de l'exposition au radon dans les mines d'uranium (R. S. Allodji, Klervi Leuraud et al. 2012 ; Navaranjan et al. 2016 ; Ladislav Tomasek et al. 2008).

2.2.2.3 Dosimétrie individuelle

Un programme de développement de dosimètres personnels a été lancé par le CEA, en France, en 1974 (Zeetwoog 1981). Les dosimètres sont des instruments de mesure destinés à estimer l'exposition reçue par un individu exposé à des RIs, par exemple, dans le cadre de son activité professionnelle. Traditionnellement portés à la ceinture par chaque mineur d'uranium, ils permettent d'obtenir des estimations précises de son exposition individuelle au radon, aux poussières d'uranium et aux rayonnements γ (Marušiaková, Gregor et Tomášek 2011). Les dosimètres personnels ont été introduits dans les mines d'uranium françaises en 1983 et dans la dernière mine en exploitation en République tchèque en 2000 (ibid.).

4. Matrices dans lesquelles sont répertoriés les niveaux d'exposition moyens estimés à différents agents chimiques et physiques en fonction du type d'emploi.

2.3 La sous-cohorte post-55 des mineurs d'uranium français

2.3.1 Exploitation de l'uranium en France

Comme dans beaucoup de pays, l'exploitation de l'uranium en France a débuté au lendemain de la seconde guerre mondiale. En 1946, la prospection de l'uranium a été lancée par le Commissariat à l'Énergie Atomique (CEA). À partir de 1976, la COmpagnie GÉnérale des Matières Nucléaires (Cogema) a été créée par le CEA pour prendre en charge l'exploitation de l'uranium (Mabile 1985). La Cogema est devenue AREVA NC en 2006 puis Orano depuis 2018. Nous parlerons donc par la suite des mines du groupe CEA-Cogema. Durant l'exploitation des mines, il y a eu deux pics d'embauche : le premier lors de l'extension de l'activité minière du début des années 1950 et le second lors de la relance de l'industrie de l'uranium consécutive au choc pétrolier de 1973. L'exploitation des mines d'uranium du groupe a vécu son apogée au cours des années 1980 pour décliner ensuite. La dernière mine d'uranium française a cessé son activité en mai 2001. Actuellement, l'uranium utilisé en France provient des mines exploitées par les filiales d'Orano installées principalement au Canada, au Niger et au Kazakhstan.

Comme illustré dans la Figure 2.5, les mines d'uranium françaises exploitées par le groupe CEA-Cogema étaient localisées dans quatre districts miniers (ibid.) :

- la division de La Crouzille, exploitée de 1949 à 1995 et localisée dans le département de la Haute-Vienne ;
- la division du Forez et du Morvan, exploitée de 1953 à 1978 et localisée près de Vichy ;
- la division de Vendée et de Bretagne, exploitée de 1954 à 1991 ;
- la division de l'Hérault, exploitée de 1978 à 1997 et localisée près de Lodève, au sud du Massif Central.

Les mines de Jouac, fermées depuis 2001, étaient initialement exploitées par une société minière indépendante, sont devenues Cogema en 1993. Les mineurs

de Jouac ont donc été intégrées plus tardivement dans la cohorte française des mineurs d'uranium.

FIGURE 2.5 – Localisations des divisions minières exploitées par le groupe CEA-Cogema en France

Le type de gisement conditionnait son mode d'exploitation. Il s'agissait de gisements de type sédimentaire dans la division minière de l'Hérault et de gisements de type granitique dans les autres divisions minières citées plus haut. Les premiers gisements exploités étaient de type granitique alors que les gisements de type sédimentaire ont plutôt été exploités dans les années les plus récentes. Par ailleurs, selon la profondeur du gisement, l'extraction du minerai d'uranium pouvait s'effectuer dans une mine à ciel ouvert ou une mine souterraine. Grâce à la ventilation naturelle, le travail dans les mines françaises à ciel ouvert induisait des niveaux d'exposition plus faibles au radon que dans les mines souterraines, milieu plus confiné où le radon s'accumulait plus facilement.

De nombreux métiers bénéficiant du statut de «mineur» ont été exercés sur les sites miniers français. Au cours des années 1970, la généralisation de la mécanisation dans le processus d'extraction de l'uranium (camions, chargeuses...) a amélioré les conditions de travail des mineurs et ouvert des emplois demandant de nouvelles compétences. Ainsi, plusieurs types d'emplois peuvent être retrouvés :

foreur, mineur de fond, chauffeur, électricien, mécanicien, géologue, topographe, etc. Ces différents métiers étaient sujets à des niveaux d'expositions différents dépendant notamment de l'activité réalisée et du temps passé au niveau du gisement. Un électricien était notamment moins exposé qu'un mineur de fond.

2.3.2 Historique, statuts vitaux et causes de décès

La cohorte française des mineurs d'uranium est une cohorte prospective qui a été mise en place au début des années 1980 dans le cadre d'une collaboration entre la médecine du travail de la Cogema et l'IPSN [5] (devenu IRSN en 2002). Elle inclut 5 086 mineurs d'uranium (uniquement de sexe masculin) ayant travaillé pendant au moins un an dans le groupe CEA-Cogema entre 1946 et 1990.

Le suivi des mineurs d'uranium français a débuté en 1946. La date de point des analyses réalisées dans cette thèse a été fixée au 31 décembre 2007. La date d'entrée dans la cohorte est définie comme la date d'embauche du mineur plus un an. La date de sortie d'étude est définie comme la date minimum entre la date de décès, la date du 85$^{\text{ème}}$ anniversaire, la date de dernières nouvelles (si mineur perdu de vue) et le 31 décembre 2007. Un âge limite de 85 ans a été fixé car la proportion de causes de décès non identifiables est significativement plus élevée après 85 ans du fait de la présence fréquente de multipathologies (Bouvier-Colle, Vallin et Hatton 1990). Les statuts vitaux ont été recueillis grâce au Répertoire National d'Identification des Personnes Physiques (RNIPP) de l'Institut National de la Statistique et des Études Économiques (INSEE). À partir de ces données, quatre modalités ont été définies : vivants, âgés de plus de 85 ans (les mineurs d'uranium encore vivants et âgés de plus de 85 ans et les sujets décédés au-delà de 85 ans), décédés et perdus de vue (Table 2.1). Pour la période de 1946 à 1967, les causes de décès ont été fournies par le Service de Santé au Travail de la Cogema car l'information sur les causes médicales de décès n'est disponible en France pour l'ensemble de la population qu'à partir de 1968. Pour la période de 1968 à 1990, elles ont été obtenues à partir des données nationales de mortalité de l'Institut National de la Santé et de la Recherche Médicale (INSERM) via le Centre d'épidémiologie sur les causes médicales de décès (CepiDC), les informations fournies par le Service

5. Institut de Protection et de Sûreté Nucléaire

de Santé au Travail de la Cogema peuvent compléter les données recueillies par l'INSERM. Depuis 1990, elles proviennent exclusivement des données nationales de mortalité de l'INSERM. Les causes de décès sont codées selon la version de la Classification Internationale des Maladies (CIM) en vigueur à la date du décès et définie par l'Organisation Mondiale de la Santé (OMS).

L'ensemble des informations administratives (date de début et de fin d'embauche, poste occupé...) ont été collectées dans les fichiers du personnel CEA-Cogema. Des informations supplémentaires concernant la localisation de la mine et le type d'emploi ont également été collectées individuellement pour chaque année de travail. En particulier, cinq catégories de type d'emploi ont été définies : a) les foreurs avant mécanisation ; b) les foreurs après mécanisation ; c) les autres postes dans les mines souterraines avant mécanisation ; d) les autres postes dans les mines souterraines après mécanisation ; e) les autres postes dans les mines à ciel ouvert.

Dans ce travail, nous nous intéresserons au risque de décès par cancer du poumon associé, dans un premier temps, à l'exposition aux rayonnements γ puis à l'ensemble des expositions radiologiques auxquelles ont été exposés les mineurs d'uranium français. Les expositions aux rayonnements γ n'ayant été enregistrées systématiquement qu'à partir de 1956 et les expositions aux poussières d'uranium n'ayant été enregistrées qu'à partir 1959 et estimées rétrospectivement entre 1956 et 1958 dans les mines françaises, seule la sous-cohorte des mineurs d'uranium français embauchés après le 31 décembre 1955 a été utilisée dans les analyses statistiques de cette thèse. Cette sous-cohorte de mineurs d'uranium français - nommée «sous-cohorte post-55» par la suite - inclut 3 377 mineurs dont 94 sont décédés par cancer du poumon à la date de point. Quelques caractéristiques de la sous-cohorte post-55 ainsi que de la cohorte complète des mineurs d'uranium français sont données dans la Table 2.1. Dans la cohorte complète, 2 924 (57.5%) mineurs étaient encore en vie à la fin du suivi, 1 935 (38.0%) étaient décédés, 187 (3.7%) avaient atteint l'âge de 85 ans. Comme on pouvait s'y attendre compte-tenu de l'entrée plus récente des mineurs de la sous-cohorte post-55, les proportions observées de mineurs décédés (25.8%) et de mineurs ayant atteint l'âge de 85 ans (2.2%) dans la sous-cohorte post-55 sont plus faibles que celles de la cohorte complète. En contrepartie, la proportion observée de mineurs vivants au 31 décembre 2007

(71.4%) est plus élevée dans la sous-cohorte post-55 que dans la cohorte complète. Les proportions de mineurs perdus de vue sont similaires et très faibles dans les deux populations (0.8% dans la cohorte complète et 0.6% dans la sous-cohorte), témoignant d'une très bonne qualité de suivi de ces travailleurs. De même, l'âge moyen à l'entrée dans l'étude et la durée de travail dans les mines sont similaires dans les deux populations : ils sont respectivement et approximativement de 28 ans et 17 ans. Bien que le suivi des mineurs d'uranium soit légèrement plus court dans la sous-cohorte post-55 (32.8 ans) que dans la cohorte complète (35.4 ans), il reste suffisamment long pour réaliser des analyses d'effets sanitaires à long terme d'une exposition chronique aux RIs.

	Sous-cohorte post-55	Cohorte totale
Mineurs, n	3 377	5 086
Âge à l'entrée dans l'étude, moyenne [min-max]	28.3 [16.9-57.7]	28.8 [16.0-68.0]
Durée de travail en années, moyenne [min-max]	16.7 [1.0-40.9]	17.0 [1.0-43.0]
Durée de suivi en années, moyenne [min-max]	32.8 [0.1-51.0]	35.4 [0.1-61.0]
Statut vital, n (%)		
Vivant	2 412 (71.4)	2 924 (57.5)
Âgé $\geq$ 85 ans	74 (2.2)	187 (3.7)
Décès par cancer du poumon	94 (2.8)	211 (4.2)
Décès d'une autre cause	777 (23.0)	1 724 (33.9)
Perdu de vue	20 (0.6)	40 (0.8)

TABLE 2.1 – Caractéristiques principales de la sous-cohorte post-55 et de la cohorte complète des mineurs d'uranium français au 31 décembre 2007.

2.3.3 Co-expositions radiologiques dans la sous-cohorte post-55

Dans la sous-cohorte post-55 des mineurs d'uranium français, l'exposition au radon, aux rayonnements γ et aux poussières d'uranium a été estimée pour chaque mineur et pour chaque année de suivi. La Figure 2.6 présente les techniques utilisées pour estimer ces expositions (soit pour la période de suivi 1956-2007). Entre 1956 et 1982, les niveaux individuels et annuels d'exposition au radon ont été estimés à partir : a) de mesures ambiantes hebdomadaires de concentration en gaz

radon réalisées en différents endroits dans les mines avec des flacons à scintillation (R. S. Allodji, Klervi Leuraud et al. 2012) et b) d'informations complémentaires concernant l'activité de chaque mineur (ex : type d'emploi, localisation de la mine d'emploi, temps passé chaque année dans différentes zones d'une même mine). Puis, à partir de 1983, ils ont été mesurés à l'aide d'un dosimètre personnel appelé «système intégré de dosimétrie individuelle» (SIDI). Les niveaux d'exposition au radon sont exprimé en Working Level Months (WLMs). Le WLM est une unité de mesure historique permettant de quantifier l'exposition au radon et ses descendants à vie courte dans les cohortes de mineurs d'uranium. Pour calculer l'exposition cumulée au radon en WLM, la concentration d'énergie α mesurée dans un litre d'air et exprimée en Working Level (WL) est multipliée par le temps pendant lequel le mineur a travaillé dans cet environnement. 1 WLM équivaut à une exposition de 1 WL (i.e., $1.3 \cdot 10^5$ Mégaélectron-Volt (MeV) d'énergie α potentielle par litre d'air) sur un mois de travail (défini comme 170 heures). L'avantage est que la concentration peut être mesurée directement. Aucune autre hypothèse n'est nécessaire pour la distribution de la dose dans le corps.

Les doses équivalentes associées à l'exposition aux rayonnements γ ont quant à elles été estimées individuellement à partir de deux dosimètres personnels différents selon la période calendaire : des films badges individuels (type PS1 du CEA) sur la période 1956-1985 puis des dosimètres thermo luminescents (TLD) intégrés au système SIDI à partir de 1986. Les doses sont exprimées en sieverts (Sv).

Enfin, les niveaux d'exposition individuels et annuels aux poussières d'uranium ont été estimés rétrospectivement pour la période 1956-1958 (Estelle Rage, Vacquier et al. 2012) puis mesurés à partir de 1959. Entre 1959 et 1982, ils ont été estimés, comme pour le radon, à partir : a) de mesures ambiantes de l'activité α volumique des poussières en suspension réalisées mensuellement en différents endroits dans les mines à partir de différents appareils (ex : tubes d'extraction équipés de filtres à papier puis de filtres membraneux, photomultiplicateurs, scintillateurs au sulfure de zinc) (Bernhard, Kraemer et Zettwoog 1991) et b) d'informations complémentaires concernant l'activité de chaque mineur (i.e., type d'emploi, localisation de la mine d'emploi, temps passé chaque année dans différentes zones d'une même mine). À partir de 1983, ils ont été mesurés, comme pour le radon, par dosimétrie individuelle intégrée au système SIDI. Les niveaux

d'exposition sont mesurés en kilos Becquerels par mètre cube heure (kBq.m^{-3}.h).

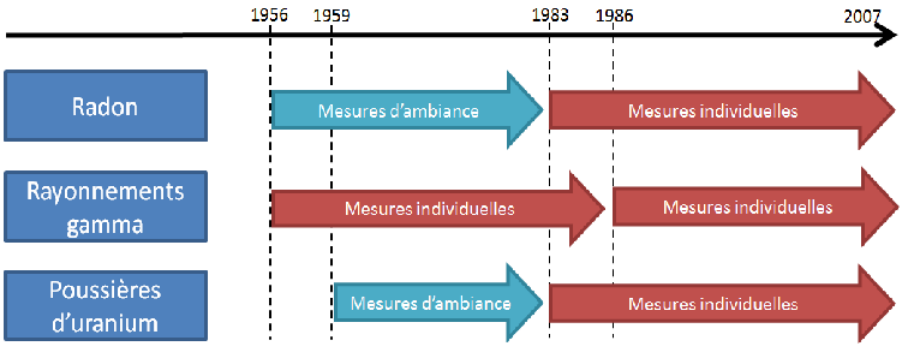

FIGURE 2.6 – Évolution des modalités d'enregistrement des expositions aux rayonnements ionisants dans les mines d'uranium françaises entre 1956 et 2007

La Table 2.2 indique quelques caractéristiques chiffrées relatives aux expositions au radon, aux poussières d'uranium et aux rayonnements γ dans la sous-cohorte post-55 des mineurs d'uranium français. Sur les 3 377 mineurs d'uranium concernés : a) 86.2% ont été exposés au radon pendant une durée d'exposition moyenne de 12.9 années et pour une exposition cumulée[6] moyenne[7] de 17.8 WLMs ; b) 95.9% ont été exposés aux rayonnements γ pendant 13.2 ans en moyenne pour une dose équivalente cumulée moyenne de 54.9 mSv ; c) 81.3% ont été exposés aux poussières d'uranium pendant 12.9 ans en moyenne pour une exposition cumulée moyenne de 1.64 kBq.m^{-3}.h. La Figure 2.7 présente les histogrammes des expositions radiologiques cumulées observées dans la sous-cohorte post-55.

2.4 Principaux résultats concernant l'association entre rayonnements ionisants et cancer du poumon chez les mineurs d'uranium

L'existence d'un excès de décès par cancer du poumon dans les populations de mineurs d'uranium est connue depuis 1879 (ROBINSON 2015 ; G. SACCOMANNO

6. sur l'ensemble du suivi
7. sur l'ensemble des 3 377 mineurs d'uranium de la sous-cohorte post-55

	Cohorte post-55
Exposition au radon	
Mineurs exposés, n (%)	2 910 (86.2)
Durée d'exposition en années, moyenne [min-max]	12.9 [1.0-35.0]
Exposition cumulée (en WLM), moyenne [min-max]	17.8 [0.01-128.4]
Exposition aux rayonnements γ	
Mineurs exposés, n (%)	3 240 (95.9)
Durée d'exposition en années, moyenne [min-max]	13.2 [1.0-36.0]
Equivalent de dose cumulée (en mSv), moyenne [min-max]	54.9 [0.20-470.1]
Exposition aux poussières d'uranium	
Mineurs exposés, n (%)	2 746 (81.3)
Durée d'exposition en années, moyenne [min-max]	12.9 [1.0-35.0]
Exposition cumulée (en kBq.m^{-3}.h), moyenne [min-max]	1.64 [0.01-10.4]

TABLE 2.2 – Quelques caractéristiques chiffrées des co-expositions radiologiques dans la sous-cohorte post-55 des mineurs d'uranium français au 31 décembre 2007.

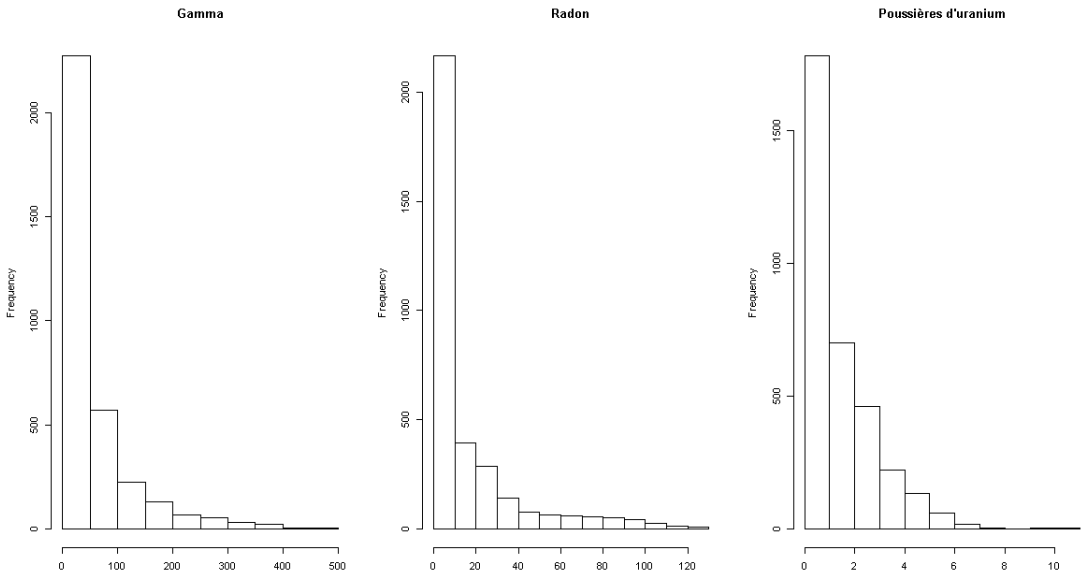

FIGURE 2.7 – Histogrammes des niveaux d'exposition cumulée aux rayonnements γ (en mSv), au radon (en WLM) et aux poussières d'uranium (en kBq.m^{-3}.h) dans la sous-cohorte post-55

et al. 1964). L'impact de l'exposition chronique des mineurs au radon sur le risque de décès par cancer du poumon est suspecté à partir de 1924-1932 (A. C. George 2008). Les premières études épidémiologiques se mettent ainsi en place dans les années 1960 afin d'estimer l'association potentielle entre une exposition chronique au radon et un risque augmenté de décès par cancer du poumon chez les mineurs. Classiquement, cette estimation a été et est toujours réalisée avec des modèles de régression de Poisson stratifiée (Preston et D. O. Stram 2017). Dans ce contexte, les personnes-années [8] sont stratifiées en groupes homogènes selon certaines variables comme l'âge, la période calendaire et surtout le niveau d'exposition cumulée. Le nombre de décès au sein de chaque strate est modélisé avec une distribution de Poisson de paramètre d'intensité spécifique à la strate : $\lambda = \lambda_0(1 + ERR(X_{cum}))$ où λ_0 désigne le taux de mortalité de base et $ERR(X_{cum})$ l'excès de risque relatif [9] (ERR) de décès par cancer du poumon associé à une exposition radiologique cumulée d'intérêt X_{cum}. Celui-ci est souvent modélisé comme une fonction linéaire de X_{cum} : $ERR(X_{cum}) = \beta \cdot X_{cum}$ avec β le paramètre inconnu permettant de quantifier l'augmentation de l'ERR par unité d'exposition cumulée (Council et al. 2006 ; Lubin, Boice Jr, Edling, Hornung, G. Howe et al. 1995).

Jusqu'à présent, les études de mortalité par cancer du poumon dans les cohortes de mineurs d'uranium se sont majoritairement concentrées sur l'exposition au radon. Les ERRs estimés pour les plus importantes cohortes de mineurs d'uranium (Estelle Rage, David B Richardson et al. 2020) ainsi que les nombres de décès par cancer du poumon et ratios de mortalité standardisés sont indiqués dans la Table 2.3. Les ERR estimés sont tous statistiquement supérieurs à zéro (leurs intervalles de confiance (IC) respectifs ne contiennent pas 0). L'exposition chronique au radon est donc associée à une augmentation statistiquement significative du risque de décès par cancer du poumon dans toutes ces cohortes. Néanmoins, on peut remarquer que les ERRs estimés sont variables d'une cohorte à l'autre. Un ERR pour 100 WLM minimum de 0.19 est estimé dans la cohorte de la Wis-

8. Unité de mesure des personnes-temps en épidémiologie. Cela correspond à la durée de suivi d'une personne non-malade pendant un an aussi bien qu'à la durée de suivi de deux personnes non malades pendant 6 mois. En particulier, si une personne n'est pas malade pendant une année de suivi, elle aura été susceptible de produire un nouveau cas de maladie pendant 1 an.

9. Dans d'autres domaines de l'épidémiologie, il est courant de modéliser le risque relatif (RR) par une régression de Poisson. Le lien entre le RR et le ERR est RR=1+ERR. Les paramètres d'un modèle ERR sont contraints par la relation ERR>-1.

Étude	Nombre mineurs ; Nombre décès par cancer du poumon	SMR IC 95%	ERR/100 WL IC 95%
Eldorado, Beaverlodge Lane et al. 2010	9 498 ; 279	1.28 [1.13 ; 1.43]	0.96 [0.56 ; 1.5
Eldorado, Port Radium Lane et al. 2010	3 047 ; 230	1.63 [1.42 ; 1.84]	0.37 [0.23 ; 0.5
Ontario Navaranjan et al. 2016	28 546 ; 1 230	1.34 [1.27 ; 1.42]	0.64 [0.43 ; 0.8
Czech Ladislav Tomasek 2012	9 978 ; 1 141	3.47 [3.27 ; 3.67]	0.97 [0.74 ; 1.2
France Rage et al. 2015	5 086 ; 211	1.34 [1.16 ; 1.53]	0.71 [0.31 ; 1.3
Wismut Walsh et al. 2010	58 987 ; 3 942	1.95 [1.90 ; 2.01]	0.19 [0.16 ; 0.2
Colorado Plateau, White Schubauer-Berigan, Daniels et Pinkerton 2009	3 255** ; 549	4.96 [4.55 ; 5.39]	NE
Colorado Plateau, American Indian Schubauer-Berigan, Daniels et Pinkerton 2009	767** ; 63	3.18 [2.45 ; 4.07]	NE
Nouveau-Mexique J. M. Samet et al. 1991	3 469 ; 68	4.00 [3.1 ; 5.1]	1.8 [0.7 ; 5.4]

* IC 90%
** Nombre de mineurs d'uranium vivants en janvier 1960
NE : Non estimé

Table 2.3 – Nombres de mineurs d'uranium, nombres de décès par cancer du poumon, ratios de mortalité standardisés (SMRs), excès de risque relatifs (ERRs) pour 100 WLM et intervalles de confiance à 95% (IC 95%) associés dans les plus importantes cohortes internationales de mineurs d'uranium

mut et un ERR pour 100 WLM maximum de 1.8 est estimé dans la cohorte des mineurs d'uranium de Nouveau-Mexique. Ces différences peuvent être dues à des caractéristiques différentes des données disponibles (ex : effectifs, âges, facteurs de risques...), à des distributions différentes des expositions au radon, à la prise en compte de facteurs modifiant la forme de la relation dose-réponse (classiquement appelés facteurs modifiants) et de facteurs de confusion différents ou encore à des méthodes différentes d'estimation de l'exposition au radon donnant lieu à des erreurs de mesure différentes.

Bien que l'ERR de décès par cancer du poumon soit souvent modélisé comme une fonction linéaire de l'exposition cumulée au radon, un certain nombre de variables peuvent néanmoins modifier cette association (Hornung, Deddens et Roscoe 1995). Les plus importantes sont : le temps écoulé depuis l'exposition (M. Tirmarche et al. 2012 ; Hunter et al. 2013), l'âge atteint (M. Tirmarche et al. 2012), l'âge à la première exposition (Ladislav Tomasek et al. 2008 ; Kreuzer, Sobotzki et al. 2017), le tabagisme (Lubin, Boice Jr, Edling, Hornung, G. R. Howe et al. 1995) et le débit de dose (Vacquier, Agnès Rogel et al. 2009). En ce qui concerne le temps écoulé depuis l'exposition, il a été constaté, dans la plupart des cohortes de mineurs d'uranium, que les expositions récentes (c'est-à-dire reçues au cours des 15 dernières années) étaient associées à des ERRs plus élevés de décès par cancer du poumon que les expositions anciennes (Ladislav Tomasek 2012). Certaines études de cohortes pour lesquelles le statut tabagique des mineurs d'uranium était disponible ont également observé une association significative entre l'exposition au radon et la mortalité par cancer du poumon, tant pour les fumeurs que pour les non-fumeurs, avec un coefficient de risque estimé qui tendait à être plus élevé pour les non-fumeurs que pour les fumeurs (Kreuzer, Sobotzki et al. 2017 ; Tomasek 2002). Certaines études ont également suggéré que le tabagisme ne serait pas un facteur de confusion dans l'estimation du risque de décès par cancer du poumon dans les cohortes de mineurs d'uranium (David B Richardson et al. 2014 ; Keil, David B Richardson et Troester 2015). En outre, des études cas-témoins nichées dans la cohorte française de mineurs d'uranium (Klervi Leuraud, Billon et al. 2007) et dans deux autres cohortes européennes de mineurs d'uranium (Klervi Leuraud, Schnelzer et al. 2011) ont montré que, si le statut tabagique était pris en compte, le lien entre exposition

au radon et risque de décès par cancer du poumon persistait et cet ajustement ne modifiait pas de manière substantielle le coefficient de risque estimé, associé à l'exposition au radon. Enfin, un effet positif du débit de dose selon lequel le risque estimé pour un travailleur est plus élevé dans le cas d'une exposition cumulée reçue sur une période longue que dans le cas d'une exposition cumulée reçue sur une période plus courte a été observé pour le radon dans la plupart des cohortes de mineurs d'uranium (Lubin, Boice Jr, Edling, Hornung, G. Howe et al. 1995). Il est néanmoins important de noter que cet effet positif disparaît si les niveaux d'exposition au radon sont faibles (M. Tirmarche et al. 2012 ; Michaela Kreuzer et al. 2010 ; Ladislav Tomasek et al. 2008 ; Vacquier, Agnès Rogel et al. 2009 ; Ladislav Tomasek 2012 ; G. R. Howe et Stager 1996). Comme les niveaux d'exposition au radon ont généralement diminué au fil des années dans les cohortes de mineurs d'uranium, il a été suggéré à plusieurs reprises qu'un effet positif du débit de dose pourrait en fait être dû, au moins en partie, à l'existence d'une erreur de mesure substantielle sur les valeurs d'exposition au radon (Lubin, Boice Jr, Edling, Hornung, G. Howe et al. 1995 ; Ladislav Tomasek 2012 ; Ladislav Tomášek et al. 1994 ; H. Morrison et al. 1998). Conformément à cette hypothèse, Stram *et al.* (1999) ont d'ailleurs constaté que l'effet positif du débit de dose observé dans la cohorte des mineurs d'uranium du plateau du Colorado était affaibli après correction des erreurs de mesure d'exposition (D. O. Stram et al. 1999).

À notre connaissance, le risque de décès par cancer du poumon associé à une exposition chronique aux rayonnements γ ou aux poussières d'uranium n'a été estimé que dans la sous-cohorte post-55 des mineurs d'uranium français. L'ERR estimé à partir de modèles de régression de Poisson stratifiée était de 0.74 (IC 95% [0.23 ; 1.73]) pour une dose équivalente de 100 mSv pour l'exposition aux rayonnements γ et de 32.18 (IC 95% [9.16 ; 72.56]) pour 100 kBq.m^{-3}.h pour les poussières d'uranium (Rage et al. 2015). Dans le cadre d'une étude en cours baptisée PUMA (pour *Pooled Uranium Miners Analysis*) et qui porte sur une cohorte internationale conjointe de 120 000 mineurs d'uranium (inclus dans les cohortes de la Table 2.3), il est notamment prévu d'estimer le risque de décès par cancer du poumon associé aux rayonnements γ (Estelle Rage, David B Richardson et al. 2020). Enfin, à notre connaissance, aucune étude n'a été réalisée jusqu'à présent pour

estimer le risque de décès par cancer du poumon associé à un effet combiné du radon, des rayonnements γ et des poussières d'uranium et appréhender les potentielles interactions entre ces expositions radiologiques chez les mineurs d'uranium. Dans Vacquier *et al.* (2011) (Vacquier, Estelle Rage et al. 2011) et Rage *et al.* (2015) (Rage et al. 2015), l'impact respectif de chaque source d'exposition a été estimé séparément dans la sous-cohorte post-55 des mineurs d'uranium français, à partir de modèles de régression de Poisson stratifiée en ERR. Dans Rage *et al.* (2012) (Estelle Rage, Vacquier et al. 2012), le risque de décès par cancer du poumon dans la sous-cohorte post-55 a été estimé en partant d'une estimation des doses absorbées au poumon incluant la contribution des rayonnements α, β et γ intervenant dans le processus de désintégration radioactive de l'uranium-238 (cf. section 2.2.1).

CHAPITRE 3

Difficultés statistiques posées par les données de co-expositions radiologiques en épidémiologie : cas de la sous-cohorte post-55

Lorsque l'on cherche à estimer un risque sanitaire à partir de données d'exposition radiologiques, on peut être confronté à de nombreuses difficultés statistiques. Si ces difficultés sont ignorées ou traitées de façon inappropriée, elles peuvent affecter la qualité des estimations de risques radio-induits et ainsi, remettre en question la validité de l'inférence statistique dans les études épidémiologiques (Thomas, D. Stram et Dwyer 1993 ; H.-M. Kim, Yasui et Burstyn 2006 ; Physick et al. 2007). Dans ce chapitre sont présentées les trois difficultés statistiques qui ont fait l'objet de travaux dans cette thèse et qui seront traitées dans les chapitres suivants de ce manuscrit. Il s'agit de la prise en compte, dans l'estimation de risques sanitaires radio-induits : a) des erreurs de mesure sur les expositions aux RIs ; b) de données d'exposition manquantes ou censurées à gauche du fait de la présence d'une limite de détection sur les appareils de mesure (ex : dosimètres personnels) ; c) de la multicolinéarité dans le cas d'expositions radiologiques corrélées. Toutes ces difficultés seront directement illustrées dans cette section à partir des données d'expositions radiologiques de la sous-cohorte post-55 des mineurs d'uranium

français décrite dans le chapitre 2.

3.1 Erreurs de mesure sur les expositions

3.1.1 Qu'est-ce qu'une erreur de mesure d'exposition ?

L'objectif principal de la plupart des études en épidémiologie des RIs est d'estimer l'association entre une variable réponse (typiquement l'occurrence ou le décès d'une pathologie) et une ou plusieurs covariables d'expositions radiologiques (voire d'autres facteurs de risque d'intérêt). Comme l'épidémiologie est par nature une science observationnelle, les conditions d'exposition des individus d'une étude ne peuvent cependant pas être contrôlées expérimentalement, si bien que les covariables d'expositions radiologiques ne sont souvent pas connues parfaitement (Broadbent 2001). La détermination des vraies valeurs d'expositions d'un individu peut en effet être difficile, coûteuse voire même impossible. Il est ainsi fréquent de passer par des questionnaires ou, comme décrit dans le chapitre 2, par des stratégies d'évaluation groupée de l'exposition, basée sur des mesures ambiantes, afin de dériver des mesures de substitution imparfaites des vraies expositions. L'association entre ces mesures de substitution imparfaites des expositions et la variable réponse d'intérêt est ensuite estimée, menant à un problème dit d'erreurs de mesure. En effet, c'est le lien entre des mesures de substitution imparfaites des expositions radiologiques et un évènement d'intérêt qui est estimé, alors que c'est bien le lien entre les vraies expositions et l'évènement d'intérêt que nous souhaiterions estimer. Une erreur de mesure d'exposition fait donc référence à l'écart entre une mesure de substitution (aussi appelée « exposition observée» par la suite) et une valeur d'exposition réelle (et donc inconnue).

3.1.2 Une structure complexe dans les cohortes professionnelles

Dans les études de cohortes professionnelles telles que les cohortes de mineurs d'uranium, on s'intéresse classiquement à l'association entre une pathologie et des expositions chroniques cumulées aux RIs. Comme cela est le cas dans la sous-

cohorte post-55 des mineurs d'uranium français décrite dans le chapitre 2, il s'agit généralement de traiter un ensemble de données d'expositions longitudinales dont les modalités d'enregistrement ont évolué au cours du temps. Dans ce contexte, les erreurs de mesure sur les expositions radiologiques ont souvent une structure complexe. Leur type et leur magnitude peuvent en effet changer au cours du temps, en fonction des techniques utilisées pour mesurer ou estimer ces expositions (Hoffmann, Estelle Rage et al. 2017 ; R. S. Allodji, Thiébaut et al. 2012). L'utilisation d'appareils permettant d'effectuer des mesures ambiantes en différents endroits dans les mines (ex : période 1956-1983 pour le radon dans la sous-cohorte post-55) ou de dosimètres personnels (ex : à partir de 1983 pour le radon dans la sous-cohorte post-55) donne généralement lieu à des erreurs de mesure de type «classique». Celles-ci correspondent au fait que si plusieurs appareils de mesure de même type sont utilisés pour mesurer une même exposition (i.e., en un instant et en un point donné), les valeurs enregistrées sont généralement légèrement différentes d'un appareil à l'autre. Cela est principalement dû à des différences de précision des appareils de mesure. Les progrès techniques accomplis en matière de dosimétrie personnalisée ont permis de bénéficier d'appareils de plus en plus précis au fil des années et donc de réduire la magnitude de ces erreurs de mesure. Dans le cas d'une stratégie d'évaluation groupée de l'exposition radiologique, une même exposition (ex : une mesure ambiante) est attribuée à tous les mineurs partageant le même type d'emploi et travaillant dans la même zone au sein d'une mine. Une telle stratégie est à l'origine d'erreurs de mesure de type «Berkson». En effet, elle sous-estime la variabilité des expositions individuelles. Une description plus formelle des erreurs de type classique et Berkson sera donnée dans le chapitre 5. Nous verrons alors que les erreurs de mesure classique et Berkson ne se traitent pas de manière identique et que leur impact potentiel sur les estimations de risque peut être bien différent.

Au-delà de l'erreur Berkson inhérente à toute stratégie d'évaluation groupée de l'exposition radiologique, le manque de précision des mesures ambiantes (ou de toute autre estimation groupée de l'exposition) peut venir ajouter une erreur de type classique qui est partagée entre plusieurs travailleurs appartenant à un même groupe c'est-à-dire travaillant, par exemple, dans une même zone d'une mine. Par ailleurs, les pratiques de travail individuelles des mineurs d'uranium n'étant pas

prises en compte dans le cas d'une stratégie d'évaluation groupée de l'exposition, cela peut donner lieu à des erreurs qui sont partagées sur plusieurs années de suivi d'un même travailleur (Hoffmann, Laurier et al. 2018). Ainsi, par exemple, si un mineur avait l'habitude de prendre ses pauses dans des endroits peu ventilés au sein des mines (afin d'échapper, par exemple, au vacarme des systèmes de ventilation), il pouvait être exposé à des niveaux de RIs bien plus élevés que ceux enregistrés par les capteurs d'ambiance situés dans les zones de travail, qui elles étaient toutes ventilées.

Enfin, l'incertitude sur le temps passé par un mineur dans différentes zones de la mine vient également accroître les erreurs de mesure associées à l'utilisation d'une stratégie d'évaluation groupée de l'exposition, lors de l'estimation des expositions annuelles.

3.1.3 Exemple des expositions aux rayonnements γ dans la sous-cohorte post-55

Pour rappel, les équivalents de dose associés à l'exposition aux rayonnements γ ont été estimés individuellement dans la sous-cohorte post-55 à partir de deux dosimètres personnels différents selon la période calendaire : des films badges individuels sur la période 1956-1985 puis des dosimètres TLD intégrés au système SIDI à partir de 1986. D'après la section 3.1.2, les erreurs de mesure associées sont donc de type classique avec une magnitude d'erreur qui a changé au cours du temps, en fonction du dosimètre utilisé.

Les travaux de thèse de Rodrigue Allodji (S. R. Allodji 2011) ont permis d'identifier et de caractériser plus finement les principales composantes de cette erreur classique selon la période calendaire. Il s'agit tout d'abord des erreurs de laboratoire, des erreurs radiologiques et des erreurs liées à l'environnement (E.S. Gilbert, J. J. Fix et W. V. Baumgartner 1995). Les erreurs de laboratoire désignent les erreurs introduites lors de l'étalonnage en laboratoire des films badges utilisés pendant la période 1956-1985, de leur traitement chimique et de la lecture des densités optiques. Les erreurs radiologiques se rapportent à l'imprécision de ces films badges pour mesurer tous les niveaux d'énergie γ dans toutes les directions. Les erreurs d'environnement correspondent à la surestimation des expositions γ

qui peut se produire lorsque l'étalonnage des films badges est réalisé à l'air libre plutôt que sur des fantômes. La perte de lecture des dosimètres TLD utilisés après 1983 ainsi que les inévitables erreurs humaines lors de la transcription des données (ex : localisation, temps de travail, valeurs d'exposition,...) et de la tenue des dossiers sont d'autres composantes possibles de l'erreur de mesure d'exposition aux rayonnements γ.

Une estimation de la magnitude relative de ces différentes composantes d'erreur de mesure a également été réalisée dans la thèse de Rodrigue Allodji pour les périodes 1956-1985 et 1986-2007 (S. R. Allodji 2011). Elle est résumée dans la Table 3.1. Ces estimations reposent sur la littérature et des dires d'experts. L'estimation de la magnitude de l'erreur de mesure totale associée à l'utilisation de films badges et de dosimètres TLD est basée sur l'hypothèse simple selon laquelle toutes les composantes d'erreur identifiées sont indépendantes.

Composantes	1956-1985	1986-2007
Erreurs de laboratoire	0.2	0.0
Erreurs radiologiques	0.1	0.0
Erreurs de l'environnement	0.1	0.0
Perte de lecture de l'instrument de mesure	0.0	0.16
Tenue des dossiers et transcription des données	0.015	0.01
Erreur de mesure totale	0.245	0.16

Table 3.1 – Ecarts-types estimés des principales composantes d'erreur (supposées indépendantes) et de l'erreur de mesure totale associée à l'utilisation de films-badges (période 1956-1985) et de dosimètres TLD (période 1986-2007) pour mesurer les expositions aux rayonnements γ dans la sous-cohorte post-55 des mineurs d'uranium français.

3.1.4 Impact général des erreurs de mesure en épidémiologie des RIs

Malgré leur omniprésence dans les études épidémiologiques, les erreurs de mesure sur les expositions sont encore rarement prises en compte dans l'estimation de coefficients de risque sanitaire (Jurek et al. 2006). Cela est d'ailleurs une critique récurrente et une source de discussion dans la plupart des études sur les

cohortes de mineurs d'uranium (Estelle Rage, Vacquier et al. 2012 ; Ladislav Tomasek et al. 2008). En effet, il est bien connu que, lorsqu'elles ne sont pas ou mal prises en compte dans les analyses, de telles erreurs de mesure peuvent mener à : a) une estimation biaisée de leur association avec un risque d'intérêt ; b) une mauvaise quantification de l'incertitude associée à cette estimation ; c) une déformation de la relation exposition-risque d'intérêt et, plus généralement, à une perte de puissance statistique (Carroll et al. 2006). En épidémiologie des RIs, cela peut notamment conduire à ne pas détecter certaines associations déjà faibles entre expositions chroniques à faibles doses de RI et risques de cancers ou, peut-être même pire, à détecter par erreur des associations importantes là où il n'y en a pas.

3.2 Données d'exposition manquantes et censurées à gauche

3.2.1 Origine des données manquantes et censurées à gauche

Dans les études menées en épidémiologie des RIs, les données d'exposition radiologiques peuvent être censurées. C'est en particulier le cas lorsque le niveau d'exposition est déterminé à l'aide d'un appareil de mesure qui ne peut pas détecter et mesurer des niveaux d'exposition en dessous d'un certain seuil fixé, appelé limite de détection (LD). Ainsi, une exposition radiologique peut être censurée à gauche de manière déterministe : la valeur associée est soit mesurée, soit seulement connue pour être inférieure à la LD de l'appareil de mesure utilisé. La LD correspond à la plus petite valeur d'activité détectable - avec une incertitude acceptable - d'un appareil de mesure. Elle dépend donc principalement du type d'appareil utilisé (Fournier et al. 2017). Les films badges et les dosimètres TLD successivement utilisés pour mesurer les équivalents de dose de rayonnements γ dans la sous-cohorte post-55 des mineurs d'uranium français en sont un exemple : ils ne peuvent pas mesurer des équivalents de dose inférieurs à 2.2 mSv et 0.55 mSv respectivement. On peut remarquer que la valeur de la LD a diminué au cours du temps. Cela est principalement dû aux progrès technologiques réalisés pour la

mesure des expositions radiologiques. Enfin, il est important de noter que des niveaux d'exposition radiologiques censurés à gauche peuvent également être sujets aux erreurs de mesure. En effet, une exposition légèrement supérieure à la LD peut, à cause des erreurs de mesure, être enregistrée comme inférieure à la LD et donc être non mesurée. Au contraire, une exposition réellement inférieure à la LD peut, à cause des erreurs de mesure, être détectée supérieure à la LD et donc être mesurée.

Dans les études de cohortes professionnelles, les expositions radiologiques peuvent également être manquantes et ce, pour diverses raisons. Tout d'abord, certaines données d'exposition peuvent être perdues (conséquence d'une erreur humaine) ou non mesurées du fait d'un dosimètre personnel défectueux ou non porté par un travailleur. Elles peuvent également être manquantes pour les travailleurs expatriés. Cela est par exemple le cas des mineurs d'uranium travaillant occasionnellement dans une mine étrangère : ceux-ci ont alors été exposés aux RIs mais leur dosimétrie individuelle n'a pas été systématiquement transmise en France. Ainsi, certains travailleurs peuvent avoir des valeurs d'expositions manquantes sur toute leur période d'expatriation. Il est également possible que l'exposition d'un travailleur soit manquante pour certaines années, si celui-ci n'a pas été exposé pendant un certain temps en raison d'un congé maladie ou du service militaire par exemple. Enfin, comme les valeurs d'expositions radiologiques disponibles pour les années les plus récentes sont généralement le fruit d'un cumul d'expositions enregistrées par chaque dosimètre personnel, celles-ci peuvent être sous-estimées dès lors qu'un travailleur n'aurait pas porté son dosimètre pendant quelques jours. Malheureusement, cela n'est pas réellement détectable dans les données disponibles car l'occurrence de ces données d'exposition manquantes n'y est évidemment pas renseignée.

À notre connaissance, dans la sous-cohorte post-55, il n'y a aucune raison de supposer que les valeurs de dose de rayonnements γ manquantes sont manquantes non aléatoirement (MNAR, Missing Not At Random) : cela signifierait que la raison pour laquelle ces données sont manquantes dépend de la valeur de la donnée elle-même. Si ces données ont été perdues, il semble raisonnable de supposer que les données sont manquantes complètement aléatoirement (MCAR, Missing Completely At Random). D'autres raisons pouvant entrer en jeu dans l'occurrence de

données d'exposition manquantes (ex : congé maladie, expatriation), l'hypothèse la plus simple est de supposer que ces doses manquantes sont manquantes aléatoirement (MAR, Missing At Random), ce qui signifie que le processus de réponse (ex : décès par cancer du poumon) n'est pas impacté par ces données manquantes. Il est plutôt impacté par des informations observées annexes telles que la période calendaire, la localisation de la mine, le type de poste...

3.2.2 Des données censurées parfois mal identifiées dans les bases de données : une approche possible

La base de données relative à la sous-cohorte post-55 des mineurs d'uranium français contient un certain nombre de valeurs d'expositions radiologiques nulles ou manquantes. Si certaines de ces valeurs représentent bien une absence d'exposition, d'autres peuvent raisonnablement être supposées strictement positives, même si potentiellement proches de zéro. D'une certaine façon, cela peut être vu comme une nouvelle composante d'erreur de mesure sur les expositions, d'origine humaine, créée lors de la transcription des informations dans la base de données.

Une première étape de correction de cette composante d'erreur de mesure a été proposée dans cette thèse et appliquée au cas des équivalents de dose (simplement appelés «doses» par la suite) de rayonnements γ nuls ou manquants de la sous-cohorte post-55. Elle a consisté à tirer partie du contexte de co-expositions radiologiques des mineurs d'uranium pour faire des hypothèses permettant de classer chaque dose nulle ou manquante comme 1/ un vrai zéro ; 2/ une dose censurée à gauche. Cela correspond à une valeur de dose strictement positive si faible que non détectable par l'appareil de mesure utilisé (i.e., inférieure à sa LD) : elle a ainsi été enregistrée comme faux zéro ou donnée manquante dans la base de données ; 3/ une dose strictement positive sans restriction c'est-à-dire que sa valeur peut être supérieure à la LD de l'appareil de mesure utilisé.

La Table 3.2 résume les hypothèses faites concernant les valeurs de dose de rayonnements γ nulles ou manquantes dans la sous-cohorte post-55. Il indique également la fréquence relative (en pourcentage) des valeurs de dose concernées par chaque catégorie. Ainsi, dans la sous-cohorte post-55, on peut observer que 7.60% des valeurs de doses de rayonnements γ ont été enregistrées comme zéros,

Dose γ enregistrée	Mesures des co-expositions	Poste du mineur	Hypothèse sur la valeur de dose γ	Fréquence relative des valeurs de dose γ (en %)
0	Radon=0 et Poussières=0	Normal	= 0	2.75%
	Radon>0 ou Poussières>0	Normal	< LD**	4.85%
NA*	Radon=0 et Poussières=0 Radon=NA* et Poussières=0 Radon=0 et Poussières=NA*	Normal	=0	0.50%
	Radon>0 ou Poussières>0	Normal	<LD**§	0.45%
	Radon=NA* et Poussières=NA*	Normal	<LD**	1.41%
		Expatrié	>0	0.30%
		Autre***	=0	0.28%

TABLE 3.2 – Résumé des hypothèses sur les valeurs de dose nulles et manquantes associées à une exposition aux rayonnements γ dans la sous-cohorte post-55 dans mineurs d'uranium français.
* Valeur de dose manquante
** Limite de détection du dosimètre
*** Poste d'un mineur différent de «normal» et «expatrié». «Service militaire», «congé maladie» et «licencié» sont des exemples non-exhaustifs de dénominations utilisées pour la catégorie «Poste» dans la base de données

et 2.94% des valeurs de doses ont été enregistrées comme données manquantes. Afin de faire des hypothèses plausibles concernant les doses enregistrées comme nulles, les historiques de co-expositions radiologiques des mineurs d'uranium de la sous-cohorte post-55 ont été utilisés pour différencier les vrais zéros (correspondant à une absence réelle d'exposition aux rayonnements γ) des faux zéros (correspond à une fausse absence d'exposition i.e., à des valeurs de dose strictement positives mais inférieures à la LD de l'appareil de mesure utilisé). Ainsi, lorsque les expositions au radon et aux poussières d'uranium mesurées à l'année t étaient nulles dans la base de données, la valeur de dose γ correspondante a été supposée être un vrai zéro. En revanche, lorsque l'exposition au radon ou aux poussières d'uranium était enregistrée comme strictement positive à l'année t, la valeur de dose γ correspondante a été supposée être un faux zéro. En effet, une exposition strictement positive à au moins une source de RI à l'année t signifie que le mineur travaillait effectivement dans la mine cette année là et donc qu'il a bien été exposé aux rayonnements γ. Au final, parmi les doses γ enregistrées comme nulles, la catégorie la plus fréquente concerne les faux zéros (4.85%), c'est-à-dire les doses enregistrées comme nulles alors qu'il semble plus raisonnable de supposer qu'elles soient censurées à gauche et même sujettes à de possibles erreurs de mesure.

Afin de faire des hypothèses plausibles concernant les doses de rayonnements γ enregistrées comme manquantes, les postes de travail occupés par les mineurs d'uranium de la sous-cohorte post-55 ont été utilisés en complément des historiques de co-expositions radiologiques. Si les données d'expositions au radon ou aux poussières d'uranium étaient disponibles à l'année t, elles ont été utilisées afin de faire les hypothèses suivantes :

- Lorsque les expositions au radon et aux poussières d'uranium étaient enregistrées comme nulles à l'année t, la valeur de dose γ manquante a été supposée égale à zéro à l'année t ;
- Lorsque l'exposition au radon (respectivement aux poussières d'uranium) était manquante à l'année t mais que l'exposition aux poussières d'uranium (respectivement au radon) était nulle à l'année t, la valeur de dose γ manquante a été supposée nulle à l'année t ;
- Lorsque l'exposition au radon ou aux poussières d'uranium était strictement

positive à l'année t, la valeur de dose γ manquante a été supposée strictement positive mais inférieure à la LD de l'appareil de mesure utilisé à l'année t. En effet, dans ce cas, les valeurs d'exposition au radon et aux poussières d'uranium étaient principalement inférieures au premier quartile des distributions respectives de chaque exposition, ce qui suggère que les valeurs de dose γ correspondantes étaient faibles. Des hypothèses plus élaborées, prenant notamment en compte la corrélation entre les trois co-expositions radiologiques auraient pu être posées mais, sachant que cette catégorie ne représente que 0.45% des valeurs de dose γ, l'hypothèse la plus simple a été privilégiée.

Si les co-expositions aux trois sources radiologiques étaient simultanément manquantes à l'année t pour un mineur i mais qu'au moins une valeur de dose γ de ce même mineur i était enregistrée comme strictement positive avant l'année t et une autre après l'année t, alors le poste de travail indiqué dans le dossier administratif du mineur i a été utilisé pour distinguer trois situations :

- Si le mineur i était enregistré en poste régulier à l'année t, sa valeur de dose γ manquante a été supposée strictement positive mais inférieure à la LD de l'appareil de mesure utilisé. En effet, il semble peu probable qu'un mineur occupant un poste régulier à l'année t et ayant été exposé aux rayonnements γ avant et après cette même année t n'ait pas été exposé (c'est-à-dire que sa dose γ soit réellement nulle) à l'année t. De plus, comme les expositions aux trois sources radiologiques, qui ont été mesurées avec différents appareils de mesure, étaient manquantes, il est fort probable que toutes ces expositions étaient en fait inférieures aux LDs respectives de ces appareils ;

- Si le mineur i était enregistré comme «expatrié» à l'année t, il a été supposé exposé aux trois sources radiologiques comme s'il avait travaillé dans une mine d'uranium française à l'année t. En effet, les conditions d'exposition des mineurs d'uranium à l'étranger n'étaient pas disponibles. Cette hypothèse peut sembler un peu forte, mais elle ne concerne que 0.30% des valeurs de dose γ, qui sont alors supposées strictement positives mais pas nécessairement inférieures à la LD. Dans cette situation, les valeurs de dose γ manquantes sont supposées suivre la même distribution, à l'année t, que les valeurs de dose

γ strictement positives observées dans la sous-cohorte post-55 des mineurs d'uranium français ;

- Si à l'année t, la base de données indique que le mineur i ne travaillait pas dans une mine d'uranium, sa valeur de dose γ manquante a été supposée nulle à l'année t. Ce mineur pouvait par exemple être au service militaire, en congé maladie ou licencié à l'année t.

Enfin, si les expositions aux trois sources radiologiques étaient manquantes à l'année t et qu'aucune valeur de dose γ était strictement positive avant et après l'année t alors la valeur de dose γ manquante a été supposée nulle à l'année t.

Classiquement, en épidémiologie des RIs, les données manquantes et censurées à gauche sont remplacées soit par zéro, soit par la LD, soit par la moitié de la LD de l'appareil de mesure utilisé (Ron 1998 ; Laurent et al. 2016 ; Yoder et al. 2018 ; ES Gilbert, J. Fix et W. Baumgartner 1996). Même si la LD est faible, en particulier pour les périodes d'exposition récentes, cette approche naïve sous-estime la variabilité des expositions et sous-estime ou sur-estime l'exposition cumulée reçue par un mineur sur une période déterminée. L'exposition cumulée étant utilisée dans l'estimation de relations exposition-risque, la sous-estimation de sa variabilité peut causer un biais dans l'estimation de risques radio-induits tels que le cancer du poumon (Lubin, Colt et al. 2004 ; Xue, M. Y. Kim et Shore 2006).

3.3 Multi-colinéarité des expositions

3.3.1 Des coefficients de Pearson élevés

Comme expliqué en détails dans le chapitre 2, les mineurs de fond sont simultanément exposés au radon, aux poussières d'uranium et aux rayonnements γ. Ceux-ci sont tous le fruit du même mécanisme de désintégration radioactive de l'uranium-238 présent dans les mines. Ainsi, dans la sous-cohorte post-55 des mineurs d'uranium français, les niveaux d'exposition associés à ces trois sources radiologiques sont très corrélés entre eux. Les coefficients de corrélation de Pearson estimés pour chaque paire d'exposition sont ainsi de 0.90 entre l'exposition

cumulée au radon et la dose cumulée de rayonnements γ, de 0.82 entre l'exposition cumulée aux poussières d'uranium et la dose cumulée de rayonnements γ et de 0.78 entre l'exposition cumulée au radon et aux poussières d'uranium. Même si les coefficients estimés sont très élevés, il est cependant important de noter que ceux-ci peuvent être sous-estimés du fait de la présence d'erreurs de mesure sur les valeurs de dose γ et d'expositions au radon et aux poussières d'uranium. La Figure 3.1 montre les nuages de points relatifs aux trois paires d'expositions cumulées observées : radon/rayonnements γ, rayonnements γ/poussières d'uranium, radon/poussières d'uranium. Cela vient confirmer visuellement que les valeurs d'exposition aux trois sources radiologiques considérées sont corrélées linéairement, impliquant un potentiel problème de multicolinéarité lors de l'estimation d'un risque sanitaire radio-induit.

3.3.2 Des coefficients d'inflation de la variance élevés

Un coefficient d'inflation de la variance (appelé VIF par la suite pour «Variance Inflation Factor») permet de quantifier l'augmentation de la variance d'un coefficient de régression estimé, dans un contexte de multicolinéarité entre plusieurs variables. Le VIF relatif à une covariable k est défini comme suit :

$$VIF_k = \frac{1}{1 - R_k^2}$$

avec R_k^2 le R^2 de la régression linéaire de la covariable k en fonction de toutes les autres covariables. Le R^2, également appelé coefficient de détermination, mesure la qualité de la prédiction d'une régression linéaire. C'est le rapport entre la variance expliquée par la régression et la variance totale.

Les VIFs estimés, relatifs aux trois co-expositions radiologiques des mineurs d'uranium - sont les suivants dans la sous-cohorte post-55 :

$$VIF_{\text{radon}} = 1.80 \tag{3.1}$$

$$VIF_{\text{gamma}} = 2.01 \tag{3.2}$$

$$VIF_{\text{poussières}} = 1.25 \tag{3.3}$$

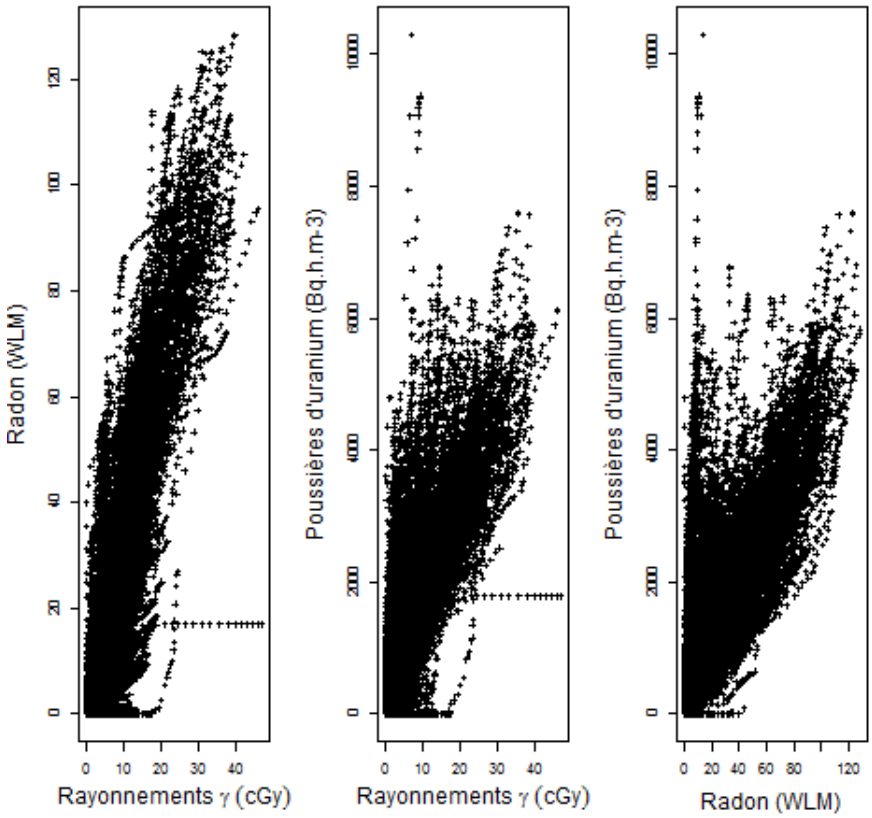

FIGURE 3.1 – Nuage de points des expositions cumulées observées au radon et aux rayonnements γ (à gauche), aux rayonnements γ et aux poussières d'uranium (au milieu), au radon et aux poussières d'uranium (à droite)

Les valeurs de VIF estimées sont relativement éloignées de la valeur seuil de 1 indiquant un potentiel manque de robustesse des estimations de risques sanitaires radio-induits dans le cas de l'application d'un modèle de régression multiple impliquant simultanément les trois expositions radiologiques. Elles restent néanmoins inférieures à la limite critique de 2.5 au delà de laquelle la robustesse du modèle n'est plus du tout garantie même si potentiellement sous-estimées par l'existence d'erreurs de mesure sur les expositions.

3.3.3 Des estimations de risque instables

Afin de mettre en évidence un potentiel manque de robustesse des estimations de risques sanitaires radio-induits dans le cas de l'application d'un modèle de régression multiple, des modèles de survie classiques en épidémiologie des RIs (décrits plus en détails dans le chapitre 5 de ce manuscrit) ont été mis en oeuvre sur les données de la sous-cohorte post-55 des mineurs d'uranium français. Le taux de risque instantané $h_i(t)$ de décès par cancer du poumon d'un mineur i au temps t est alors directement fonction des expositions, sans tenir compte d'un éventuel problème de multicolinéarité. Une première approche consiste à considérer chaque source radiologique séparément dans un modèle de régression univariée et une deuxième à inclure simultanément les trois dans un modèle de régression multiple.

Dans un cadre univarié, $h_i(t)$ a été modélisé comme suit : $h_i(t) = h_0(t) \cdot (1 + \beta \cdot X_i(t))$ où h_0 désigne le taux de base de décès par cancer du poumon et β l'excès de risque instantané de décès par cancer du poumon associé à une exposition radiologique cumulée d'intérêt X. Celle-ci peut-être plus précisément notée X^R, X^G ou X^P selon qu'elle désigne une exposition au radon, aux rayonnements γ ou aux poussières d'uranium respectivement. Ainsi, $X_i(t)$ désigne l'exposition cumulée à une source radiologique du mineur i au temps t (lagguée de 5 ans afin de tenir compte d'un délai de latence entre l'exposition et l'expression d'un risque de décès par cancer du poumon radio-induit). Le modèle de survie univarié a été estimé dans un cadre bayésien. Les médianes *a posteriori* de β et les intervalles de crédibilité à 95% associés sont de 2.7 [1.1 ; 5.2] pour le radon, 0.78 [0.28 ; 1.67] pour les rayonnements γ et $3.34{\cdot}10^{-2}$ $[1.07{\cdot}10^{-2}$; $7.00{\cdot}10^{-2}]$ pour les poussières d'uranium. Comme zéro est exclu de chaque intervalle de crédibilité estimé, on peut conclure

que l'excès de risque de décès par cancer du poumon est associé de manière positive et statistiquement significative à chaque exposition radiologique.

Dans un cadre multivarié i.e., considérant simultanément les trois expositions radiologiques, $h_i(t)$ a été modélisé comme suit : $h_i(t) = h_0(t) \cdot (1 + \beta_R \cdot X_i^R(t) + \beta_G \cdot X_i^G(t) + \beta_P \cdot X_i^P(t))$. Les médianes *a posteriori* des coefficients de risque β_R, β_G et β_P ainsi que les intervalles de crédibilité à 95% associés sont désormais de 2.7 [-0.2 ; 5.8] pour le radon, 0.00 [-0.39 ; 1.17] pour les rayonnements γ et -$0.15{\cdot}10^{-2}$ [-$1.66{\cdot}10^{-2}$; $3.81{\cdot}10^{-2}$] pour les poussières d'uranium. On peut remarquer que zéro est désormais inclus dans chaque intervalle de crédibilité estimé. Aucune des expositions radiologiques n'apparaît donc associée de manière statistiquement significative au risque de décès par cancer du poumon.

Cette analyse vient clairement mettre en évidence un manque de robustesse des estimations de risque de décès par cancer du poumon dans le cadre multivarié. Ce problème de multicolinéarité des expositions radiologiques chez les mineurs d'uranium peut conduire à des conclusions trompeuses et à de mauvaises interprétations concernant l'effet de chaque exposition sur la variable réponse d'intérêt. En effet, dans un cadre univarié où l'effet de chaque source radiologique est évalué séparément, les valeurs de risques estimées sont difficiles à interpréter puisqu'elles pourraient être le reflet d'un effet confondant des deux autres sources radiologiques (non considérées dans le modèle) qui sont à la fois corrélées avec le risque de décès par cancer du poumon et l'exposition d'intérêt principal. D'un autre coté, l'application d'un modèle de régression multiple conduit à des estimations d'excès de risque instables avec une variance élevée. Des méthodes statistiques plus sophistiquées doivent donc être utilisées pour traiter ce problème de multicolinéarité.

CHAPITRE 4

Approche hiérarchique bayésienne

Ce chapitre fournit une courte présentation des principaux outils mathématiques utilisés dans ce travail de thèse à savoir la modélisation hiérarchique et la statistique bayésienne.

4.1 Qu'est-ce qu'un modèle hiérarchique ?

Les ingrédients d'un modèle hiérarchique

La méthode la plus adaptée pour construire un modèle probabiliste complexe est l'approche hiérarchique. Comme un modèle probabiliste élémentaire, un modèle hiérarchique est composé de grandeurs observables et de paramètres inconnus. Dans ce chapitre, nous désignerons les variables aléatoires observables par la lettre majuscule Y. Il peut s'agir, par exemple, d'un vecteur $Y = (Y_1, Y_2, \ldots, Y_n)$ d'âges au décès ou au diagnostic par une pathologie donnée ou de temps à la première récidive d'une pathologie dans une cohorte de n individus. Par défaut, dans ce chapitre, nous désignerons les paramètres inconnus (de dimension D) par la lettre grecque $\theta = (\theta_1, ..., \theta_D)$. Ces derniers décrivent un état récapitulatif des propriétés d'intérêt du phénomène aléatoire étudié. Il peut s'agir, par exemple, de la ma-

gnitude d'une erreur de mesure d'exposition, d'un coefficient de risque de décès par une pathologie donnée. À noter que, dans ce manuscrit, nous utiliserons la notation entre crochets [], introduite en 1990 par A. Gelfand et A.F.M Smith et justifiée dans Parent et Bernier (Parent et Bernier 2007), pour désigner des lois de probabilité comme, par exemple, $[Y|\theta]$ pour désigner la loi de probabilité conditionnelle du vecteur aléatoire Y sachant un vecteur de paramètres θ.

Les modèles hiérarchiques - aussi appelés modèles multiniveaux ou modèles mixtes - ont pour spécificité d'inclure des variables aléatoires non observables dites latentes (ou cachées). Dans ce chapitre (uniquement), elles seront désignées par la lettre majuscule X. Ces variables sont généralement associées à un mécanisme aléatoire non observable. Il peut s'agir par exemple du vecteur $X = (X_1, \ldots, X_n)$ des « vraies» valeurs d'expositions radiologiques cumulées (c'est-à-dire sur l'ensemble du suivi) d'une cohorte de n individus ou des labels de classe inconnus permettant de regrouper des individus possédant des caractéristiques d'expositions similaires à un polluant environnemental donné.

Une démarche de construction brique par brique

La démarche de modélisation hiérarchique est souvent utilisée pour décrire un phénomène aléatoire complexe car elle permet de distinguer les différents niveaux de sources d'information et d'incertitude relatifs à ce phénomène. Plus précisément, elle consiste à décomposer un phénomène aléatoire complexe en processus aléatoires plus simples qui peuvent être modélisés avec des structures probabilistes «standard». En pratique, cette décomposition s'appuie sur une unique formule probabiliste : la formule des probabilités composées. Par exemple, celle-ci permet d'écrire la distribution jointe de grandeurs observables Y et de variables aléatoires latentes X, notées $[Y, X|\theta]$, sous la forme :

$$[Y, X|\theta] = [Y|X, \theta][X|\theta]$$

Le modèle $[Y|X, \theta]$, appelé sous-modèle d'observations, décrit l'occurrence des variables observables Y conditionnellement aux variables latentes X et à certains paramètres θ. Le modèle $[X|\theta]$, dit sous-modèle ou couche latente, décrit l'occurrence des variables latentes X conditionnellement aux paramètres θ. Comme un

jeu de LEGO, la démarche de construction hiérarchique consiste ainsi à « emboîter » judicieusement différents sous-modèles par conditionnements probabilistes successifs afin d'élaborer des structures fonctionnelles plus complexes (JORDAN et al. 2004).

Dans ce travail de thèse, par exemple, tous les modèles hiérarchiques bayésiens développés (qui seront détaillés dans les chapitres 4 et 5) combinent trois sous-modèles conditionnellement indépendants. Suivant la terminologie introduite par Clayton (1992) (CLAYTON et al. 1992) puis reprise par Richardson (2002) (Sylvia RICHARDSON et al. 2002) dans un papier traitant de la prise en compte des erreurs de mesure dans les études épidémiologiques, le sous-modèle d'observations sera systématiquement appelé « sous-modèle de maladie » par la suite. Il décrit la probabilité d'occurrence d'une variable-réponse (par exemple, âge au décès par cancer du poumon) en fonction d'une exposition ou de plusieurs co-expositions aux RIs. Un « sous-modèle d'exposition » décrira la loi des « vraies» expositions radiologiques latentes ou la loi des expositions radiologiques observées au sein d'un groupe d'individus ayant des caractéristiques d'exposition similaires. Enfin, un troisième sous-modèle - qui sera différent selon l'objectif de l'étude - sera inclus dans les structures proposées. Concernant la prise en compte d'expositions radiologiques incertaines (chapitre 4), ce troisième sous-modèle sera appelé « sous-modèle de mesure » : il décrira la relation entre les « vraies » expositions radiologiques et les expositions mesurées/estimées qui sont entachées d'incertitudes de mesure. Concernant la prise en compte d'expositions multiples (chapitre 5), ce troisième sous-modèle sera appelé « sous-modèle d'attribution » : il associe chaque individu à un groupe partageant des caractéristiques d'exposition similaires. Ainsi, les différents sous-modèles (définissant différents niveaux de hiérarchie) permettent de combiner différentes sources d'incertitude (par exemple : erreur de mesure d'expositions, mécanisme de censure, appartenance d'un individu à un groupe d'individus partageant des caractéristiques d'exposition communes) dans un même modèle conjoint. Un des grands intérêts de l'apprentissage statistique d'un modèle hiérarchique est de pouvoir obtenir des estimations conjointes de tous les paramètres inconnus et variables latentes des différents sous-modèles.

Les graphes acycliques orientés pour la représentation graphique de

modèles hiérarchiques

Dans une structure hiérarchique, les variables observables, les variables latentes et les paramètres inconnus entretiennent des relations de dépendances conditionnelles toujours orientées dans le même sens. Les paramètres sont toujours des grandeurs conditionnantes. Certains conditionnent l'occurrence des variables latentes, d'autres celle des observations. Les observations dépendent conditionnellement des variables latentes et de certains paramètres. Ainsi, les variables latentes jouent un double rôle : celui de variables conditionnantes et de variables conditionnées. La modélisation graphique permet de visualiser ces relations de dépendances conditionnelles. En particulier, les graphes acycliques orientés (ou DAG pour Direct Acyclic Graph) sont traditionnellement utilisés dans le cas de modèles hiérarchiques (Parent et Bernier 2007). Ils sont constitués de noeuds désignant soit une variable observable, soit une variable latente soit un paramètre. Les relations de dépendance sont représentées par des flèches orientées qui partent des grandeurs conditionnantes (les enfants) et pointent vers les grandeurs conditionnées (les parents). Les noeuds initiaux qui n'ont pas de parents sont les paramètres : ils ne dépendent d'aucune autre variable aléatoire. Les noeuds terminaux sans enfants, encore appelés feuilles, sont les observables. Enfin, tous les noeuds qui ne sont ni des paramètres ni des observables désignent des variables aléatoires latentes. Les DAG sont représentés avec les conventions suivantes : a) les noeuds aléatoires (quantités inconnues) sont représentés sous la forme de cercles tandis que les variables observables ou paramètres fixes apparaissent sous la forme de rectangles ; b) les flèches en trait plein indiquent des liens probabilistes et les doubles flèches indiquent des liens déterministes orientés entre deux noeuds.

Comme décrit dans la section suivante, l'approche statistique bayésienne relie l'information contenue dans les données observées (à travers la vraisemblance globale) et l'information disponible sur les quantités inconnues du modèle (à travers une loi de probabilité dite *a priori*) pour dériver une loi de probabilité dite *a posteriori*. Dans le cas de modèles hiérarchiques, cette loi *a posteriori* est complexe c'est-à-dire multidimensionnelle et souvent non disponible analytiquement. Néanmoins, l'approche statistique bayésienne offre une panoplie d'algorithmes stochastiques génériques (qui seront décrits dans la partie 4.3) permettant l'inférence

de tels modèles hiérarchiques via une approximation stochastique de la loi *a posteriori*.

4.2 Généralités sur l'approche statistique bayésienne

En statistique fréquentiste, tout paramètre d'intérêt θ est supposé fixe et inconnu. En statistique bayésienne, il est fixe et inconnu mais l'incertitude sur ce paramètre, appelée incertitude par ignorance (ou épistémique), est décrite à l'aide d'une loi de probabilité ; en ce sens, on dit que le paramètre d'intérêt est considéré comme une variable aléatoire.

La principale caractéristique de l'approche statistique bayésienne est la possibilité d'introduire de la connaissance *a priori* sur tout paramètre inconnu θ. Celle-ci peut provenir de la littérature ou de dires d'expert du domaine. Cette connaissance *a priori* doit nécessairement être représentée sous forme de distribution de probabilité appelée loi *a priori* du paramètre et notée $[\theta]$ par la suite.

L'idée de la démarche statistique bayésienne est de mettre à jour, compte-tenu de l'information apportée par des données observées y [1], la loi *a priori* $[\theta]$ en une loi *a posteriori* $[\theta|y]$ désignant la loi de probabilité conditionnelle de θ sachant y. Cette mise à jour s'opère grâce à une célèbre formule probabiliste - la formule de Bayes - donnée par :

$$[\theta|y] = \frac{[y|\theta][\theta]}{[y]} \tag{4.1}$$

dans laquelle $[y|\theta]$ désigne la vraisemblance du modèle considéré (i.e. la probabilité d'observer la réalisation y de Y sachant θ) et $[y]$ la vraisemblance marginale (ou évidence). À part dans les cas simples dits de conjugaison (cas dans lesquels la loi *a priori* et la loi *a posteriori* appartiennent à la même famille de lois de probabilité) le calcul explicite de la loi *a posteriori* est rarement accessible. Ceci est notamment dû à la difficulté du calcul de la constante de normalisation $[y]$. Même si la loi *a posteriori* est accessible, les calculs dérivés de la loi *a posteriori* ne le sont souvent pas car ils résultent d'intégrales non explicites. Par exemple, l'espérance *a posteriori*, la variance *a posteriori* et les quantiles de la loi *a posteriori* $[\theta|y]$ font ap-

1. y désigne une réalisation de Y

pel à du calcul intégral complexe. Il est donc utile de disposer d'outils permettant d'approcher l'ensemble de ces intégrales. Dans la littérature sur l'approximation d'intégrales, deux grandes familles existent : ce sont les approximations dites numériques (comme la méthode des trapèzes) et les approximations dites stochastiques. Les deux inconvénients majeurs des approximations numériques sont la lenteur dans le cas multidimensionnel (fréquent dans le cas de l'utilisation d'un modèle hiérarchique) et l'utilisation de la même précision dans les régions à plus ou moins fortes probabilités. Pour ces raisons, les approximations stochastiques sont utilisées dans les approches bayésiennes. Une des grandes familles de telles approximations est l'approximation de Monte Carlo par chaînes de Markov (MCMC) qui est une généralisation des méthodes de Monte Carlo dans le cas d'une dépendance de type markovien. Les algorithmes utilisés sont appelés les algorithmes MCMC. Ils sont décrits ci-après.

4.3 Généralités sur les algorithmes MCMC

4.3.1 Qu'est-ce qu'un algorithme MCMC ?

En statistique bayésienne, les algorithmes MCMC permettent de générer un échantillon markovien selon la loi *a posteriori* $[\theta|y]$ d'un modèle donné. À chaque itération t de l'algorithme, une nouvelle valeur $\theta^{(t)}$ est affectée au paramètre θ en fonction de la valeur du paramètre à l'itération précédente $\theta^{(t-1)}$. Cette chaîne de Markov converge vers une loi stationnaire : la loi *a posteriori*, cible de l'inférence bayésienne. L'échantillon de valeurs ainsi obtenu peut être utilisé pour estimer de nombreux résumés statistiques de la loi *a posteriori* (i.e., moyenne, variance, médiane, percentiles permettant de dériver des intervalles de crédibilité, ...).

Pendant cette thèse, deux principaux échantillonneurs MCMC ont été utilisés pour mener l'inférence bayésienne des modèles hiérarchiques proposés : l'échantillonneur de Gibbs (Casella et E. I. George 1992) et celui de Metropolis Hastings (Chib et Greenberg 1995). Dans ces deux cas, les chaînes de Markov doivent être initialisées pour chacun des paramètres, c'est-à-dire nécessitent le choix arbitraire de valeurs de départ $\theta^{(o)} = (\theta_1^{(o)}, \ldots, \theta_D^{(o)})$. Une description générale de ces deux échantillonneurs est donnée dans les deux parties suivantes.

4.3.2 Échantillonneur de Gibbs

L'échantillonneur de Gibbs nécessite la connaissance des distributions conditionnelles complètes de chaque paramètre à savoir $[\theta_i|\theta_{-i}, y]$, avec θ_{-i} le vecteur des paramètres du modèle privé de la composante θ_i, soit $\theta_{-i} = (\theta_1, \ldots, \theta_{i-1}, \theta_{i+1}, \ldots, \theta_D)$. À chaque itération, et successivement pour chaque composante, on tire une nouvelle valeur θ_i dans la distribution conditionnelle complète en remplaçant les autres paramètres par la dernière valeur obtenue. L'itération $t+1$ de l'algorithme est la suivante :

1. $[\theta_1^{t+1}] \sim [\theta_1|\theta_2^t, \ldots, \theta_D^t, y]$
2. $[\theta_2^{t+1}] \sim [\theta_2|\theta_1^{t+1}, \theta_3^t \ldots, \theta_D^t, y]$
3. ...
4. $[\theta_D^{t+1}] \sim [\theta_D|\theta_1^{t+1}, \ldots, \theta_{D-1}^{t+1}, y]$

Ainsi, après K itérations, une chaîne de Markov multidimensionnelle $\theta^1, \cdots, \theta^K$ est obtenue. Celle-ci converge nécessairement vers la loi stationnaire égale à la loi cible recherchée : la loi *a posteriori* $[\theta|y]$. L'avantage principale à l'utilisation d'un échantillonneur de Gibbs est qu'à chaque itération une nouvelle valeur est systématiquement affectée à chaque composante θ_i du paramètre θ (c'est-à-dire pas de rejet possible d'une nouvelle valeur candidate). En revanche, l'utilisation d'un échantillonneur de Gibbs nécessite la connaissance des lois conditionnelles complètes ainsi que la disponibilité d'un outil performant et efficace pour simuler des valeurs selon ces lois.

4.3.3 Échantillonneur de Metropolis-Hastings adaptatif

L'échantillonneur de Metropolis Hastings requiert le choix d'une loi instrumentale notée g. Certaines conditions doivent être respectées pour ce choix afin que l'algorithme converge bien vers la loi stationnaire $[\theta|y]$. Ces conditions portent notamment sur le support de la loi g. De plus, la loi g doit être entièrement connue et simulable efficacement. En effet, à chaque itération, une valeur dite «candidate» sera simulée (θ^{cand}) selon la loi g puis acceptée ou non comme nouvelle valeur de

θ. Si elle est refusée, alors la valeur affectée à θ lors de cette itération sera la valeur obtenue à l'itération précédente. La loi g peut dépendre de la valeur de θ de l'itération précédente. La probabilité d'acceptation α de θ^{cand} est donc cruciale et vaut :

$$\alpha = \min\left\{1; \frac{[\theta^{cand}|y]}{[\theta^{curr}|y]} \frac{g(\theta^{curr}|\theta^{cand})}{g(\theta^{cand}|\theta^{curr})}\right\} = \min\left\{1; \frac{[y|\theta^{cand}]\,[\theta^{cand}]}{[y|\theta^{curr}]\,[\theta^{curr}]} \frac{g(\theta^{curr}|\theta^{cand})}{g(\theta^{cand}|\theta^{curr})}\right\} \tag{4.2}$$

où θ^{curr} est le vecteur des valeurs des paramètres à l'itération précédente.

Contrairement à l'échantillonneur de Gibbs, où chaque valeur tirée devient la nouvelle valeur du paramètre, la valeur candidate est évaluée par rapport à la valeur courante du paramètre dans le cas d'un échantillonneur de Metropolis-Hastings . L'examen de la probabilité d'acceptation α permet de comprendre le fondement de cette approche. Pour alléger son expression, imaginons que la loi instrumentale g soit symétrique (et donc que $g(x|y) = g(y|x)$), la probabilité d'acceptation devient dans ce cas : $\alpha = \min\left\{1; \frac{[\theta^{cand}|y]}{[\theta^{curr}|y]}\right\}$. Dans le cas où la nouvelle valeur candidate θ^{cand} est plus probable sous la loi *a posteriori* que la valeur θ^{curr} simulée à l'itération précédente, alors elle est automatiquement acceptée car $\alpha = 1$. Par contre, quand elle est moins probable, elle n'est pas automatiquement rejetée mais acceptée avec une probabilité de $\frac{[\theta^{cand}|y]}{[\theta^{curr}|y]}$. Cette propriété permet à l'algorithme de visiter des régions de la loi *a posteriori* à faible probabilité.

A l'itération t, $\theta^t = \theta^{curr}$. L'itération $t+1$ de l'algorithme est la suivante :

1. $\theta^{cand} \sim g(\theta|\theta^{curr})$
2. $\theta^{t+1} = \begin{cases} \theta^{cand} & \text{avec probabilité } \alpha \\ \theta^{curr} & \text{sinon} \end{cases}$

Le choix de la loi instrumentale g est essentiel dans cet algorithme. Deux grandes classes existent pour ces lois instrumentales : les lois instrumentales indépendantes ($g(\theta|\theta^{curr}) = g(\theta)$) et les marches aléatoires ($g(\theta|\theta^{curr}) = g(\theta - \theta^{curr})$).

Dans ce travail de thèse, les marches aléatoires ont été systématiquement privilégiées pour le choix des lois instrumentales. Par ailleurs, lorsque l'échantillonneur de Metropolis-Hastings a été utilisé pour la mise à jour d'un (ou de plusieurs)

paramètres, l'algorithme a systématiquement commencé par une phase adaptative, c'est-à-dire que pendant les premières itérations (qui n'ont bien sûr pas été conservées dans l'échantillon *a posteriori*), les paramètres de la loi instrumentale $g(.|\theta^{curr})$ ont été calibrés afin d'atteindre les taux d'acceptation recommandés dans la littérature (Andrew Gelman et al. 2014) à savoir un taux d'acceptation de la valeur candidate de 40% pour un paramètre seul et de 20% pour un vecteur de paramètres. Par exemple, si la loi instrumentale est la loi normale centrée en θ^{curr} de variance σ^2, le paramètre de variance variera pour atteindre le taux d'acceptation cible de 40%. Augmenter la variance σ^2 permet d'obtenir des valeurs candidates plus éloignées de la valeur centrale θ^{curr} (et donc plus souvent rejetées) alors que diminuer cette variance permet de rester localement autour de cette valeur centrale. Les valeurs fixées pour les taux d'acceptation d'un algorithme Metropolis-Hastings permettent aux chaînes de Markov de «visiter» efficacement l'ensemble des valeurs possibles de la loi *a posteriori*. En effet, en cas de taux d'acceptation trop élevés, les chaînes de Markov ne visitent que très lentement l'ensemble du support des valeurs possibles de θ et restent souvent bloquées dans une partie du support à forte probabilité. En revanche, en cas de taux d'acceptation trop faibles, les valeurs proposées tombent trop souvent dans des zones peu probables de la loi *a posteriori* et sont ainsi souvent rejetées : les chaînes de Markov «bougent» alors très peu.

4.4 Convergence d'un algorithme MCMC

En statistique bayésienne, les chaînes de Markov $\theta = (\theta^1, \cdots, \theta^k, \cdots, \theta^K)$ générées à partir d'un algorithme MCMC (par exemple basé sur un échantillonneur de Gibbs ou de Metropolis-Hastings) convergent théoriquement vers la loi stationnaire cible - la loi *a posteriori* - quand le nombre d'itérations tend vers l'infini. Partant d'un point de départ arbitraire θ^0, combien d'itérations initiales (appelé «temps de chauffe» ou «burnin») semblent suffisantes pour espérer que la chaîne ait atteint son régime stationnaire ? Une fois le régime stationnaire supposé atteint, combien d'itérations supplémentaires doivent être réalisées pour obtenir une approximation stochastique «correcte» de la loi *a posteriori* ? Les réponses à ces deux questions font référence aux diagnostics de convergence d'un algorithme MCMC.

Dans cette thèse, plusieurs chaînes de Markov partant de positions initiales

différentes ont été générées pour inférer chaque modèle hiérarchique proposé. Un examen visuel des chaînes a été systématiquement réalisé afin de vérifier graphiquement que celles-ci se «mélangeaient» correctement et oscillaient bien autour des mêmes valeurs. Ceci a notamment permis d'évaluer les temps de chauffe nécessaires pour espérer avoir atteint l'état stationnaire et de s'assurer de l'absence de problèmes de convergence majeurs. Il s'agissait notamment de s'assurer que quelque soit leur position initiale, les chaînes convergeaient bien vers la même distribution stationnaire. Nous verrons dans la suite de ce travail que ce point peut être crucial pour détecter notamment des convergences dites locales.

Le diagnostic de convergence de Gelman & Rubin (Gelman et Rubin 1992) fondé sur le ratio entre la variabilité intra-chaîne et la variabilité inter-chaînes a été estimé et comparé à 1. En effet, un ratio typiquement inférieur à 1.05 indique qu'un temps de chauffe suffisant a été considéré et que les chaînes de Markov «se recouvrent» convenablement.

Des problèmes dits de «mélangeance» doivent également être détectés. Ce phénomène correspond à des lenteurs de convergence. Un outil possible est l'estimation des auto-corrélations intra-chaînes pour chaque paramètre estimé, en fonction des itérations. Ce critère a été systématiquement examiné dans ce travail, pour chaque chaîne de Markov et pour chaque modèle. Une forte autocorrélation indique un mélange très lent de la chaîne de Markov concernée. Afin de pallier ce problème, une solution consiste à sous-échantillonner la chaîne en ne conservant que les itérations espacées d'un pas p (appelé "thinning") fixé par l'utilisateur (une seule itération est conservée sur p itérations).

L'estimation de résumés statistiques de la loi *a posteriori* s'effectue à partir des échantillons *a posteriori* générés par algorithmes MCMC et de méthodes de Monte-Carlo. Ainsi, par exemple, l'espérance *a posteriori* de toute fonction $h(\theta)$ du paramètre θ sera estimée par $\hat{\mathbb{E}}(h(\theta)|y) = \frac{1}{M}\sum_{m=1}^{M} h(\theta^{(m)})$ avec M le nombre total d'itérations conservées. Il est possible de quantifier l'erreur de Monte-Carlo, c'est-à-dire l'erreur commise en estimant la moyenne *a posteriori* d'un paramètre à l'aide de la moyenne empirique d'un échantillon simulé de la distribution *a posteriori*. Cette évaluation de la différence entre la moyenne empirique et la vraie espérance *a posteriori* est alors comparée à l'estimation de l'écart-type *a posteriori* du paramètre. Une règle pratique couramment utilisée est de vérifier que

cette erreur est inférieure à 5% de l'écart-type *a posteriori*.

4.5 Quelques critères de sélection de modèles hiérarchiques

4.5.1 Le critère DIC

Le «Deviance Information Criterion» (DIC) (Spiegelhalter et al. 2002) est un critère de type vraisemblance pénalisée qui peut être vu comme une version bayésienne du critère AIC [2]. Il est défini par :

$$DIC = D_m(y) + p_D \tag{4.3}$$

$D_m(y) = E_{[\theta|y]}(D(y,\theta))$ désigne l'espérance *a posteriori* de la déviance. Elle mesure la qualité d'ajustement du modèle de paramètres θ aux données y. Elle peut être estimée par $\frac{1}{K}\sum_{k=1}^{K} D(y,\theta^k)$, avec K la taille de l'échantillon *a posteriori* de θ et $D(y,\theta^k) = -2\cdot\log([y|\theta^k])$ la déviance en θ^k. p_D est le paramètre de pénalisation. Il peut être interprété comme le nombre effectif de paramètres ou la complexité du modèle. $p_D = D_m(y) - D(y,\mathbb{E}(\theta|y))$ où $D(y,\mathbb{E}(\theta|y))$ est la déviance en l'espérance *a posteriori* de θ . Cette déviance peut être estimée par $D(y,\frac{1}{K}\sum_{k=1}^{K}\theta^k)$. A noter que le DIC peut donc être facilement estimé à partir d'un échantillon *a posteriori* issu d'un algorithme MCMC.

Plus le DIC estimé est faible, meilleures sont les capacités d'ajustement aux données du modèle associé.

4.5.2 Le critère WAIC

Le critère de Watanabe-Akaike, également appelé «Widely Applicable Information Criterion» (WAIC) (Watanabe 2010), a été récemment proposé comme un critère de sélection de modèle alternatif prometteur au critère DIC et se voulant "plus bayésien" (Andrew Gelman et al. 2013). Ce critère quantifie les performances prédictives d'un modèle tout en tenant compte du nombre effectif de

2. Akaike Information Criterion

paramètres à ajuster. Ainsi, le WAIC est notamment basé sur la distribution prédictive *a posteriori* «point par point» [3] et à échelle logarithmique, notée $log[y_i|\theta]$ pour l'observation i. Comme le DIC, le WAIC peut être facilement estimé à partir d'un échantillon *a posteriori* issu d'un algorithme MCMC.

Soit K le nombre d'itérations d'un échantillon *a posteriori* généré avec un algorithme MCMC. La capacité prédictive du modèle considéré peut être estimée par :

$$lppd = \sum_{i=1}^{n} log\left(\frac{1}{K}\sum_{k=1}^{K}[y_i|\theta^k]\right) \tag{4.4}$$

Le terme de pénalité p_{WAIC} désignant le nombre «effectif» de paramètres à ajuster peut être estimé de deux façons. La première méthode proposée dans (Andrew Gelman et al. 2013) pour estimer p_{WAIC} est basée sur une différence similaire à celle utilisée dans le calcul de la pénalité du DIC :

$$p_{WAIC,1} = 2 \cdot \sum_{i=1}^{n}\left(log\left(\frac{1}{K}\sum_{k=1}^{K}[y_i|\theta^k]\right) - \frac{1}{K}\sum_{k=1}^{K} log[y_i|\theta^k]\right) \tag{4.5}$$

La deuxième méthode proposée dans (ibid.) pour estimer p_{WAIC} est la suivante :

$$p_{WAIC,2} = \sum_{i=1}^{n} V_{k=1}^{K}(log[y_i|\theta^k]) \tag{4.6}$$

où $V_{k=1}^{K}(log[y_i|\theta^k])$ désigne la variance estimée de la densité prédictive *a posteriori*.

Finalement, on a :

$$WAIC = -2 \cdot lppd + 2 \cdot p_{WAIC} \tag{4.7}$$

Pour des raisons de simplicité calculatoire, les WAIC indiqués dans ce manuscrit ont été estimés avec la pénalité $p_{WAIC,1}$.

Plus la valeur du WAIC est faibles meilleures sont les capacités prédictives du modèle concerné.

3. c'est-à-dire variable réponse par variable réponse

CHAPITRE 5

Prise en compte d'expositions radiologiques incertaines dans l'estimation d'un risque radio-induit

Ce chapitre porte sur la prise en compte des incertitudes de mesure sur les doses de rayonnements γ dans l'estimation du risque de décès par cancer du poumon dans la sous-cohorte post-55 des mineurs d'uranium français. Après un état de l'art sur la prise en compte des erreurs de mesure d'exposition dans les études épidémiologiques, il présente le modèle hiérarchique proposé dans le cadre de cette thèse pour prendre en compte simultanément trois sources d'incertitude sur ces expositions : les erreurs de mesure sur les doses, la limite de détection des dosimètres personnels utilisés et les valeurs de dose manquantes.

5.1 Erreurs de mesure d'exposition : nature et état de l'art

Dans ce travail, on parlera d'« erreur de mesure d'exposition » pour un individu i lorsqu'il existe un écart - noté U_i par la suite - entre la valeur réelle d'une exposition d'intérêt de cet individu - notée X_i par la suite - et la valeur de substi-

tution mesurée - notée Z_i par la suite - et ce, quelles que soient les raisons de cet écart. À noter que, dans ce manuscrit, le terme d'«erreur de mesure» sera aussi strictement réservé au cas où une valeur de substitution Z_i strictement positive a été mesurée afin de le distinguer des données censurées et manquantes.

L'impact des erreurs de mesure des covariables d'exposition sur les estimations des risques sanitaires en épidémiologie dépend bien sûr de la taille des erreurs de mesure, mais également du type des erreurs de mesure et du modèle de risque utilisé. On va aborder dans cette partie uniquement les erreurs de mesure qui affectent une variable explicative continue. Les erreurs de mesure sur des variables discrètes et sur les variables réponses ne seront pas traitées dans cette thèse.

5.1.1 Différents types d'erreurs de mesure

Afin de pouvoir modéliser correctement les erreurs de mesure, il faut au préalable les caractériser. Les paragraphes suivants vont décrire les différentes caractéristiques des erreurs de mesure. Ce travail se concentre sur des expositions de type environnemental, variables strictement positives par nature.

Erreurs de mesure additives et multiplicatives

Les modèles d'erreurs de mesures peuvent être essentiellement classés en deux catégories correspondant à des hypothèses additives ou multiplicatives. En effet, une première hypothèse consiste à supposer que l'erreur de mesure U sur une variable est la différence entre la vraie exposition inconnue X et l'exposition mesurée Z. L'erreur de mesure est alors additive. Une seconde hypothèse considère que l'erreur de mesure peut être vue comme le ratio entre X et Z, ainsi l'erreur est multiplicative. Des études empiriques suggèrent que le modèle d'erreurs de mesure multiplicatives est plus adapté dans les études épidémiologiques (Armstrong 1998 ; Lyles et Kupper 1997). Il offre de plus l'avantage de respecter la positivité de la vraie exposition X, ce qui n'est pas le cas pour un modèle fondé sur une différence. Ainsi, nous supposerons les erreurs de mesure multiplicatives dans la suite de ce travail.

Erreurs de mesure systématiques et aléatoires

Parmi les erreurs de mesure aléatoires, on distingue deux types d'erreurs de

mesure : les erreurs systématiques aléatoires et les erreurs purement aléatoires. On parle d'erreurs systématiques lorsque la valeur mesurée est systématiquement sur-évaluée (ou sous-évaluée). Ce type d'erreurs fait référence à l'existence d'un biais non nul qui existerait pour tous les individus, dû par exemple, à la mauvaise calibration d'un appareil de mesure. On parle d'erreur de mesure purement aléatoire lorsque le biais est supposé nul et donc la valeur d'exposition peut, d'une mesure à l'autre, être sur-évaluée puis sous-évaluée et inversement.

Erreurs de mesure différentielles et non-différentielles

L'erreur de mesure est dite non-différentielle lorsqu'il y a indépendance entre la variable réponse Y et l'erreur de mesure U. Cette hypothèse peut être vue comme une indépendance entre la variable réponse Y et l'exposition mesurée Z conditionnellement à la vraie exposition X. Dans les études de cohorte, les variables explicatives étant mesurées avant la variable réponse, il est raisonnable de considérer les erreurs de mesure comme non-différentielles. Au contraire, on parle d'erreur de mesure différentielle lorsqu'il n'y a pas indépendance entre l'erreur de mesure et la variable réponse. Ce type d'erreur de mesure est plus fréquent dans les études cas-témoin puisque quu'à la différence des études de cohorte, la variable réponse est obtenue avant l'évaluation de la variable d'exposition. On peut imaginer, par exemple, que les cas sont plus à même de reconstruire leurs expositions car plus concernés par la maladie et donc par la recherche de causes.

Erreurs de mesure Berkson et classique

Dans la littérature, on distingue essentiellement deux modèles d'erreur de mesure qui sont les erreurs dites classiques et les erreurs dites de Berkson. Selon les cas, on considère la distribution de l'exposition mesurée Z conditionnellement à la vraie exposition X (classique) ou inversement, la distribution de X sachant Z (Berkson).

Dans le cas d'erreurs de mesure de nature classique, l'exposition mesurée au temps t, $Z(t)$, peut être modélisée comme fonction de $X(t)$ et de $U(t)$ par :

$$Z(t) = X(t) \cdot U(t) \tag{5.1}$$

Les erreurs de mesure de nature classique surviennent à cause de l'appareil de

mesure utilisé pour mesurer l'exposition. Ce modèle d'erreurs de mesure permet de refléter qu'une même exposition mesurée par plusieurs appareils de mesure ne donnera pas exactement la même valeur d'exposition, mais plusieurs valeurs centrées autour de la vraie valeur. La variance de ces différentes mesures dépend de la précision de l'appareil de mesure. En cas d'erreurs de mesure de nature classique, on suppose que la variance des expositions mesurées est plus importante que la variance des vraies expositions.

Dans le cas d'erreurs de mesure Berkson, on modélise au contraire la vraie exposition au temps t, $X(t)$, comme fonction de $Z(t)$ et $U(t)$. Ce modèle permet de transcrire que lorsque l'on affecte une même valeur d'exposition à tous les individus, en fonction des conditions d'exposition, les vraies valeurs d'expositions oscillent autour de la valeur mesurée. Dans ce cas :

$$X(t) = Z(t) \cdot U(t) \tag{5.2}$$

Dans le cas d'erreurs de mesure Berkson, on suppose souvent que la valeur mesurée est connue avec précision. Or, cette exposition est elle même soumise à des erreurs de mesure de nature classique dues à l'appareil de mesure. Il est donc possible de combiner les deux types d'erreurs de mesure. En cas d'erreurs de mesure Berkson, on suppose que la variance des vraies expositions est plus importante que la variance des expositions mesurées.

Erreurs de mesure partagées et non-partagées

L'erreur de mesure peut être partagée par plusieurs individus ou par plusieurs mesures d'un même individu. L'erreur de mesure est partagée par plusieurs individus lorsque la composante d'erreur de mesure est la même pour plusieurs individus. Ainsi, l'erreur de mesure $U_i(t)$ ne dépend plus de l'individu et donc $U_i(t) = U(t)$. Ce type d'erreur de mesure est typique de la reconstruction d'expositions. Dans cette situation, une valeur d'exposition est attribuée à tous les mineurs d'une mine, donc l'erreur de mesure sur cette valeur est la même pour tous les mineurs à qui elle a été attribuée. Les erreurs de mesure sont partagées par plusieurs mesures d'un individu lorsque la composante d'erreurs de mesure est identique pour plusieurs pas de temps et ainsi, $U_i(t) = U_i$. Cela arrive fréquemment lorsque l'exposition est enregistrée par capteur d'ambiance au cours du temps. L'exposition reçue par

un individu dépend de plusieurs facteurs dont le poste du mineur, mais aussi de son comportement, de sa technique individuelle de travail... que l'on peut supposer constant au cours du temps. Ainsi, l'erreur de mesure est partagée par plusieurs mesures d'un même individu.

Nature des erreurs de mesure dans la sous-cohorte post-55

Dans la cohorte française des mineurs d'uranium, la dose de rayonnements γ est mesurée individuellement depuis 1956, et à partir de 1983 pour les expositions au radon et aux poussières d'uranium. Les erreurs de mesure sur ces expositions peuvent être modélisées comme des erreurs de mesure de nature classique car résultant de l'appareil de mesure. Pour les mesures d'exposition au radon et aux poussières d'uranium entre 1956 et 1982, les erreurs de mesure sont de type Berkson puisque causées par des capteur d'ambiance mesurant les niveaux d'exposition de tous les mineurs. Les erreurs de mesure sont, dans la sous-cohorte post-55 des mineurs d'uranium, supposées non-différentielles et aléatoires, hypothèses raisonnables dans le contexte d'une cohorte prospective. Les erreurs de mesure de nature classique, associées à l'appareil de mesure, sont supposées non-partagées. Au contraire, les erreurs de mesure de type Berkson, associées à des capteurs d'ambiance, sont partagées par plusieurs temps pour un même individu. En fonction de l'appareil de mesure utilisé, et donc de sa précision, la variance d'erreur de mesure varie.

5.1.2 Impacts des erreurs de mesure sur l'inférence statistique

Malgré la diversité des types d'erreurs de mesure rencontrés en épidémiologie, les études qui évaluent les effets de l'erreur de mesure sur l'inférence statistique ainsi que les méthodes de correction des erreurs de mesure portent principalement sur des erreurs homoscédastiques, non différentielles et non partagées (Zhang et al. 2017). De plus, la plupart des hypothèses émises concernant les impacts potentiels des erreurs de mesure sur l'inférence statistique est basé sur des résultats analytiques disponibles pour le modèle de régression linéaire simple incluant une seule covariable sujette à erreur de mesure additive. Cependant, en épidémiologie

des RIs, la pertinence de ce modèle est extrêmement limitée car il est rarement possible de modéliser les variables réponses d'intérêt comme des variables continues non censurées. Les variables réponse sont plus communément des variables binaires (ex : présence ou absence de maladie), des temps d'occurrence d'événements (ex : âge au décès d'une pathologie, âge au diagnostic, temps jusqu'à la guérison ou la récidive) ou des variables de comptage (ex : nombre de cas d'une pathologie). Par conséquent, l'impact des erreurs de mesure dans la régression logistique, la régression de Poisson et les modèles de survie sont d'un intérêt indéniable en épidémiologie des RIs. En effet, il est bien connu que les conséquences exactes des erreurs de mesure d'exposition sur l'inférence statistique dépendent de leur type (ex : classique / Berkson, partagée / non partagée, différentielle / non différentielle), de leur variance mais aussi du type de modèle de maladie considéré (ex : régression de Poisson, régression logistique, modèles de survie) (Hoffmann, Laurier et al. 2018 ; Flegal, Keyl et Nieto 1991 ; Heid et al. 2002 ; Stefanski et Carroll 1985 ; Reeves et al. 1998 ; D. Richardson et Loomis 2004).

Régression linéaire

Les conséquences des erreurs de mesure dans les modèles de régression sont plus étudiées, et donc plus connues. Dans le cadre d'une régression linéaire simple :

$$Y_i = \beta_0 + \beta_1 \cdot X_i + \epsilon_i \tag{5.3}$$

où Y_i est la variable réponse continue de l'individu i, β_0 et β_1 sont les coefficients de régression inconnus et ϵ_i les termes d'erreur tels que $\mathbb{E}(\epsilon_i) = 0$. En supposant les erreurs de mesure classiques, non-différentielles et additives, le modèle est $Z_i = X_i + U_i$ avec U_i indépendant de Z_i et de X_i. À cause des erreurs de mesure, on observe uniquement Z_i comme substitut de X_i. Ainsi, on ajuste classiquement le modèle suivant :

$$Y_i = \beta_0^* + \beta_1^* \cdot Z_i + \epsilon_i^* \tag{5.4}$$

Dans ce modèle, l'estimateur des moindres carrés de β_1^* est

$$\begin{aligned}\hat{\beta_1^*} &= \frac{Cov(Z_i, Y_i)}{Var(Z_i)} \\ &= \frac{Cov(X_i, Y_i) + Cov(U_i, Y_i)}{Var(X_i) + Var(U_i)} \\ &= \frac{Cov(X_i, Y_i)}{\sigma_x^2 + \sigma_u^2}\end{aligned} \tag{5.5}$$

car l'erreur est non-différentielle. L'estimateur sans biais de β_1 étant $\hat{\beta_1} = \frac{Cov(X_i,Y_i)}{\sigma_x^2}$, l'estimateur de β_1* est atténué par le coefficient $\frac{\sigma_x^2}{\sigma_x^2+\sigma_u^2}$. Ainsi, plus la variance d'erreur de mesure est importante, plus l'atténuation sera également importante.

On suppose maintenant l'erreur de mesure toujours additive et non-différentielle, mais cette fois de type Berkson au lieu de nature classique. Comme $X_i = Z_i + U_i$, en partant de $Y_i = \beta_0 + \beta_1 \cdot X_i + \epsilon_i$ on obtient :

$$\begin{aligned}Y_i &= \beta_0 + \beta_1 \cdot Z_i + \beta_1 \cdot U_i + \epsilon_i \\ &= \beta_0 + \beta_1 \cdot Z_i + \epsilon_i^*\end{aligned} \tag{5.6}$$

Le nouveau terme d'erreur $\epsilon_i^* = \beta_1 \cdot U_i + \epsilon_i$, est bien centré en 0. En effet,

$$\begin{aligned}\mathbb{E}(\epsilon_i^*) &= \mathbb{E}(\beta_1 \cdot U_i + \epsilon_i) \\ &= \beta_1 \cdot \mathbb{E}(U_i) + \mathbb{E}(\epsilon_i) \\ &= \beta_1 \cdot 0 + 0 \\ &= 0\end{aligned} \tag{5.7}$$

Ainsi l'estimateur obtenu par la méthode des moindres carrés ordinaires est un estimateur sans biais. En revanche, la variance de l'estimateur est plus importante :

$$\begin{aligned}Var(\epsilon_i^*) &= Var(\beta_1 \cdot U_i + \epsilon_i) \\ &= \beta_1^2 \cdot Var(U_i) + Var(\epsilon_i) \\ &= \beta_1^2 \cdot \sigma_u^2 + \sigma_\epsilon^2 \\ &> \sigma_\epsilon^2\end{aligned} \tag{5.8}$$

Régression logistique

Le modèle de régression logistique se présente comme suit :

$$P(Y_i = 1|X_i) = \frac{exp(\beta_0 + \beta_1 X_i)}{1 + exp(\beta_0 + \beta_1 X_i)} \tag{5.9}$$

avec Y_i la variable réponse binaire de l'individu i. Dans ce cadre, les erreurs de mesure sur les covariables, de nature classique et de Berkson, créent un biais sur l'estimateur de risque (Reeves et al. 1998 ; Hoffmann 2017). Le biais dépend de la variance d'erreurs de mesure (Heid et al. 2002).

Modèles de survie

Dans les études épidémiologiques, la variable réponse est souvent le temps jusqu'à un évènement qui est une variable censurée à droite puisqu'on n'observe pas le temps jusqu'à cet évènement pour tous les individus. Les modèles de survie sont alors utilisés dans le contexte de telles données où parmi eux, les modèles à hasards proportionnels sont les plus populaires car ils s'affranchissent de l'estimation du risque de base. De manière générale, pour un individu i, on note T_i le délai jusqu'à l'événement d'intérêt (âge au décès par cancer du poumon) et C_i le délai jusqu'à la censure (par exemple le délai jusqu'aux dernières nouvelles), alors le délai observé Y_i est le minimum entre les deux délais précédents, soit $Y_i = \min(T_i, C_i)$. L'indicatrice de non censure, δ_i, vaut 1 si $Y_i = T_i$ et vaut 0 si $Y_i = C_i$. On définit alors le risque instantané de décès $h_i(t)$ par

$$h_i(t) = \lim_{\Delta_t \to 0} \frac{P(t < T_i \leq t + \Delta_t | T_i > t)}{\Delta_t} \tag{5.10}$$

Dans le cadre de ce travail, le risque instantané d'un évènement à un instant t est modélisé comme le produit d'un risque de base $h_0(t)$ (en général celui des individus non exposés) supposé identique pour tous les individus et d'un coefficient $\rho(\beta, X_i(t))$ représentant le ratio de risque instantané entre un individu non-exposé et un individu ayant reçu une exposition X_i au temps t soit

$$h_i(t; \beta) = h_0(t) \cdot \rho(X_i(t); \beta) \tag{5.11}$$

où β est le coefficient de risque de l'évènement étudié.

Différentes formes de la fonction ρ existent dans la littérature, la plus courante étant celle correspondant au modèle de Cox, le terme $\rho(\beta; X_i(t))$ est $exp(\beta \cdot X_i(t))$. En épidémiologie du rayonnement ionisant étudiant l'association entre la mortalité

par cancer solide et l'exposition aux radiations, la forme de ρ privilégiée est celle appelée EHR (Excess Hazard Ratio) définie par

$$\rho(X_i(t);\beta) = 1 + EHR_i(t;\beta) = 1 + \beta \cdot X_i(t) \tag{5.12}$$

Comme dans le cas de la régression logistique, les erreurs de mesure de nature classique et de Berkson conduisent à un biais dans les estimations de risque β (H.-M. Kim, Yasui et Burstyn 2006).

Régression de Poisson

La régression de Poisson est un modèle souvent utilisé en épidémiologie des rayonnements ionisants (Preston et D. O. Stram 2017). Comme décrit dans la section 2.4, les individus sont regroupés, notamment sur le niveau d'exposition qui est alors discrétisé. Ainsi, à cause des erreurs de mesure, des individus peuvent être mal classifiés. En cas de réelle association entre l'évènement et l'exposition, l'exposition des cas a tendance à être sous-estimée, et celle des témoins à être sur-estimée. L'estimation du risque associé à l'exposition est ainsi sous-estimée et donc biaisée.

Atténuation de la relation dose-réponse

Une atténuation de la relation dose-réponse est souvent observée pour les niveaux d'exposition plus élevés (Stayner et al. 2003). Il est suggéré que cela soit la conséquence d'erreurs de mesure sur les expositions plus importantes. En effet, les expositions les plus importantes correspondent aux plus anciennes années pour lesquelles les niveaux d'exposition ont été reconstruit *a posteriori* à partir de dires d'experts, ce qui conduit à des erreurs de mesure plus importantes. Hoffmann *et al.* (Hoffmann, Laurier et al. 2018) a montré, par des simulations, que les erreurs de mesure partagées par un individu au cours du temps causent un biais et une atténuation de la relation dose-réponse plus important que les erreurs de mesure non-partagées ou partagées par plusieurs individus.

5.1.3 Méthodes de correction des erreurs de mesure

Il existe plusieurs méthodes pour corriger les effets des erreurs de mesure. On distingue les approches fonctionnelles plutôt fréquentistes telles que la régression

calibration (Stefanski et Carroll 1985) et la méthode SIMEX (SIMulation EXtrapolation) (Cook et Stefanski 1994) et les approches structurelles plutôt bayésiennes (Carroll et al. 2006). La différence principale entre les approches fonctionnelles et les approches structurelles, est que les vraies expositions inconnues X sont supposées fixes, tandis qu'elles sont supposées suivre une distribution, à spécifier, dans le cas des approches structurelles.

5.1.3.1 Régression calibration

La régression calibration est la méthode la plus fréquemment utilisée (Guolo 2008). L'idée de la régression calibration est de remplacer les expositions mesurées par l'espérance de la vraie valeur d'exposition (Stefanski et Carroll 1985). Le risque radio-induit est estimé avec l'espérance de la vraie valeur d'exposition comme covariable au lieu de l'exposition mesurée. L'espérance des vraies valeurs de $X_i(t)$ sont modélisées comme fonction des expositions mesurées $Z_i(t)$ et des covariables $V_1, ..., V_p$. Par exemple, $\mathbb{E}(X_i(t)|Z_i(t), V_{i1}, ..., V_{ip}) = a_0 + a_1 \cdot Z_i(t) + \sum_{j=1}^{k} b_j \cdot V_{ij}$. Les paramètres inconnus a_0, a_1 et $b_j \forall j \in [1, ..., p]$ sont estimés grâce à un jeu de données de validation dans lequel la vraie exposition et l'exposition mesurée sont disponibles. L'inconvénient de la régression calibration est que cette méthode nécessite un échantillon de validation. De plus, cette méthode consiste en deux étapes disjointes. Ainsi, l'incertitude sur la vraie valeur d'exposition X de la première étape n'est pas prise en compte dans l'estimation du risque en épidémiologie.

5.1.3.2 Méthode SIMEX

L'idée de la méthode SIMEX est que l'effet des erreurs de mesure sur un estimateur peut être déterminé grâce à des simulations (Carroll et al. 2006 ; Stefanski et Carroll 1985). Des données d'exposition sont simulées avec plusieurs valeurs d'erreurs de mesure. On estime alors naïvement le risque associé à ces expositions. Ensuite, on estime la relation (classiquement linéaire ou quadratique) entre le biais de l'estimateur et la magnitude de l'erreur de mesure. Cette fonction est utilisée pour estimer le risque d'intérêt sans erreurs de mesure. La qualité de l'estimateur de risque dépend de la qualité de la fonction d'extrapolation (Misumi et al. 2018).

5.1.3.3 Approches structurelles basées sur la vraisemblance

Dans les deux approches présentées précédemment, le risque est estimé sans prendre en compte l'incertitude sur l'espérance des vraies expositions (dans le cas de la régression-calibration) ou celle de la fonction d'extrapolation (dans le cas SIMEX), d'où l'intérêt de prendre en compte les erreurs de mesure et d'estimer le risque d'intérêt conjointement en une seule étape.

Il faut alors spécifier un sous-modèle de maladie qui relie la vraie exposition $X_i(t)$ et la réponse. Un deuxième sous-modèle est nécessaire, le sous-modèle de mesure qui relie la vraie exposition à l'exposition mesurée. Dans le cas d'erreurs de mesure Berkson, le sous-modèle de mesure décrit la distribution de la vraie exposition $X_i(t)$ conditionnellement à l'exposition mesurée $Z_i(t)$. Au contraire, en cas d'erreurs de mesure de nature classique, le sous modèle de mesure décrit la distribution de l'exposition mesurée $Z_i(t)$ conditionnellement à la vraie exposition $X_i(t)$. Dans ce dernier cas, un troisième sous-modèle qui décrit la distribution de la vraie exposition est nécessaire, via une famille de distribution de probabilité connue. Les sous-modèles sont supposés être conditionnellement indépendants, en partie justifié par l'hypothèse d'erreurs de mesure non-différentielles. Ces approches peuvent se faire dans un cadre fréquentiste où la vraisemblance sera maximisée ou dans un cadre bayésien avec la spécification des lois *a priori* sur les paramètres, la vraisemblance sera utilisée pour obtenir la loi *a posteriori* des paramètres. Concernant ces deux approches, la complexité du modèle complet rend souvent impossible des calculs explicites. Dans le cadre fréquentiste, la maximisation de la vraisemblance nécessaire à l'estimation des paramètres est obtenue via des algorithmes de type EM ou SEM (Stochastic Expectation-Maximisation) et dans le cadre bayésien, les lois *a posteriori* sont approchées via des algorithme MCMC (Monte-Carlo par chaînes de Markov).

5.2 Présentation des modèles hiérarchiques bayésiens proposés

On propose différents modèles hiérarchiques pour prendre en compte les expositions censurées à gauche et manquantes sujettes aux erreurs de mesure dans

l'estimation du risque de décès par cancer du poumon associé à la dose de rayonnements γ dans la sous-cohorte post-55 des mineurs d'uranium français.

Les modèles proposés sont largement inspirés du modèle proposé pour estimer le risque de décès par cancer du poumon associé à l'exposition au radon dans la cohorte française des mineurs d'uranium avec prise en compte des erreurs de mesure (Hoffmann, Estelle Rage et al. 2017). Ces modèles sont composés de trois sous-modèles :

- le *sous-modèle de maladie* : il définit la relation entre la vraie dose professionnelle cumulée aux rayonnements γ et le risque de décès par cancer du poumon ;
- le *sous-modèle de mesure* : il définit la relation entre la vraie dose annuelle aux rayonnements γ et la dose annuelle aux rayonnements γ mesurée ;
- le *sous-modèle d'exposition* : il décrit la distribution de probabilité de la vraie dose de rayonnements γ dans la sous-cohorte post-55 des mineurs d'uranium français.

Comme suggéré par Hoffmann *et al* (ibid.), on suppose que l'erreur de mesure sur la dose de rayonnements γ est non-différentielle dans la sous-cohorte post-55 des mineurs d'uranium français. Le sous-modèle de maladie, le sous-modèle de mesure et le sous-modèle d'exposition sont également supposés être indépendants conditionnellement.

5.2.1 Sous-modèle de maladie

Soit T_i l'âge au décès par cancer du poumon du mineur i, $i \in \{1, 2, \ldots, n\}$ où n est le nombre total de mineurs d'uranium. Soit C_i l'âge censuré à droite définit comme le plus petit âge du mineur i entre l'âge au décès d'une autre cause que le cancer du poumon, l'âge au 31 décembre 2007, l'âge en jours correspondant au $85^{\text{ème}}$ anniversaire du mineur i et âge auquel le mineur a été perdu de vue. Pour chaque mineur i, l'évènement observé est la variable continue positive $Y_i = \min(T_i, C_i)$ et la variable binaire δ_i, où $\delta_i = 1$ si $T_i \leq C_i$ (c'est-à-dire que le mineur i est décédé par cancer du poumon à l'âge $Y_i = T_i$), et $\delta_i = 0$ si $T_i > C_i$ (c'est-à-dire

que le mineur i aurait été décédé après l'âge C_i). Soit $X_i^{cum}(t-5)$ la vraie dose cumulée de rayonnements γ inconnue du mineur i au temps t, lagguée de 5 ans. En effet, on suppose une période de latence de 5 ans entre une dose de rayonnements γ et son potentiel impact sur la mortalité par cancer du poumon (Rage et al. 2015 ; Langholz et al. 1999).

On modélise la relation entre la dose cumulée de rayonnements γ et l'âge au décès par cancer du poumon par un modèle de survie en EHR :

$$h_i(t;\beta) = h_0(t)(1 + EHR_i(t;\beta)) \tag{5.13}$$

où $EHR_i(t;\beta)$ est l'excès de risque instantané (EHR) de décès par cancer du poumon potentiellement associé à la dose de rayonnements γ reçue par le mineur i au temps t. Comme supposé classiquement lorsque l'on modélise l'association entre un cancer solide et la dose de rayonnements γ (Rage et al. 2015 ; Vacquier, Estelle Rage et al. 2011), on a supposé une structure linéaire en fonction de la dose cumulée, sans facteur de confusion :

$$EHR_i(t;\beta) = \beta \cdot X_i^{cum}(t-5) \tag{5.14}$$

Ainsi, β est le coefficient de risque inconnu d'intérêt, sujet à la contrainte suivante : $\beta \cdot X_i^{cum}(t-5) > -1, \forall t \forall i$. Cela permet d'assurer la positivité de $h_i(t;\beta)$. Finalement, $h_0(t)$ correspond au risque instantané de base, c'est-à-dire au risque instantané de décès par cancer du poumon au temps t pour un mineur non-exposé aux rayonnements γ. On suppose le risque de base $h_0(t)$ constant par morceaux :

$$h_0(t) = \lambda_j, \forall t \in (s_{(j-1)}, s_j] \tag{5.15}$$

avec $s_0 = 0$, $s_1 = 40$, $s_2 = 55$, $s_3 = 70$ et $s_4 = 85$ ans. Ainsi, on considère quatre intervalles de temps pour lesquels la valeur du taux de base λ_j est supposé constant. Pour la suite, $\boldsymbol{\lambda} = (\lambda_1, \lambda_2, \lambda_3, \lambda_4)$ désigne le vecteur des paramètres inconnus du taux de base. D'autres hypothèses sur $h_0(t)$ peuvent être faites notamment par des modèles paramétriques (de type exponentiel), changement faisable dans ce contexte.

Soit $h_i(t)$ le risque instantané du mineur i au temps t, $f_i(t)$ la densité de T_i et

$F_i(t)$ sa fonction de répartition, ainsi :

$$h_i(t) = \lim_{\Delta t \to 0} \frac{P(t < T_i \leq t + \Delta t | T_i > t)}{\Delta t} = \frac{f_i(t)}{1 - F_i(t)} = \frac{f_i(t)}{S_i(t)} \tag{5.16}$$

où $S_i(t) = 1 - F_i(t) = \int_t^\infty f_i(u)du$ est la fonction de survie du mineur i au temps t. Ainsi, $\Delta t \cdot h_i(t)$ est une approximation de la probabilité que le mineur i décède par cancer du poumon pendant l'intervalle de temps $[t, t + \Delta t]$ sachant que ce mineur est toujours vivant au temps t (Ibrahim, Chen et Sinha 2014). Sachant que $h_i(t) = \frac{f_i(t)}{S_i(t)}$, alors $f_i(t) = h_i(t) \cdot S_i(t)$, et que $f_i(t) = \frac{d}{dt}F_i(t) = -\frac{d}{dt}(1 - F_i(t)) = -\frac{d}{dt}S_i(t)$, on obtient :

$$h_i(t) = \frac{-\frac{d}{dt}S_i(t)}{S_i(t)} = -\frac{d}{dt}\log S_i(t) \tag{5.17}$$

Ainsi, on obtient l'expression de la fonction de survie comme suit :

$$S_i(t) = \exp\left(-\int_0^t h_i(u)du\right) \tag{5.18}$$

Dans le cas où la dose cumulée X_i^{cum} serait constante (i.e. $X_i^{cum}(t-5) = X_i^{cum}, \forall t$), la probabilité $S_i(y_i, X_i^{cum})$ que le mineur i survive jusqu'au temps y_i, sachant sa dose cumulée X_i^{cum}, est :

$$\begin{aligned} S_i(y_i, X_i^{cum}) &= \exp\left(-\int_0^{y_i} h_i(t)dt\right) \\ &= \exp\left(-\int_0^{y_i} h_0(t)\rho(X_i^{cum}, \beta)dt\right) \\ &= \exp\left(-\rho(X_i^{cum}, \beta)\sum_{j=1}^{J}\delta_{ij}\left[\lambda_j(y_i - s_{j-1}) + \sum_{g=1}^{j-1}\lambda_g(s_h - s_{g-1})\right]\right) \end{aligned} \tag{5.19}$$

avec $s_0 = 0$ et $\delta_{ij} = 1$ si $Y_i \in]s_{j-1}; s_j]$, c'est-à-dire si le mineur i est décédé par cancer du poumon ou censuré dans le $j^{\text{ème}}$ intervalle, et 0 sinon.

Dans notre cas, la variable explicative c'est-à-dire la dose cumulée dépend du temps. Ainsi, il est classique de «découper» le passé de chaque individu en plusieurs intervalles d'âge (correspondant à plusieurs pseudo-individus) dans lesquels la dose

cumulée reste constante. On obtient alors une partition $r_i^0 < r_i^1 < \cdots < r_i^{M_i} = y_i - 5$ en M_i intervalles avec r_i^0 l'âge du mineur à l'embauche et $r_i^{M_i} = y_i - 5$. Nous noterons $w_{i,m}$ la dose cumulée constante sur l'intervalle $]r_i^{m-1}, r_i^m]$.

Ainsi, la probabilité que le mineur i survive jusqu'au temps y_i est :

$$\begin{aligned} S(y_i, X_i^{cum}(t-5)) &= \exp\left(-\int_0^{y_i} h_0(t)\rho(X_i^{cum}(t-5), \beta)dt\right) \\ &= \prod_{m=1}^{M_i} \frac{S_m(r_i^m, w_{i,m})}{S_m(r_i^{m-1}, w_{i,m})} \end{aligned} \tag{5.20}$$

On en déduit que la vraisemblance du modèle de maladie est :

$$\begin{aligned} L(\mathbf{y}, \delta^Y | X^{cum}, \beta, \lambda) &= \prod_{i=1}^{n} f(y_i)^{\delta_i^Y} S(y_i, X_i^{cum}(t-5))^{1-\delta_i^Y} \\ &= \prod_{i=1}^{n} \left[S(y_i, X_i^{cum}(t-5)) \cdot h_i(y_i)\right]^{\delta_i^Y} S(y_i, X_i^{cum}(t-5))^{1-\delta_i^Y} \\ &= \prod_{i=1}^{n} h_i(y_i)^{\delta_i^Y} S(y_i, X_i^{cum}(t-5)) \\ &= \prod_{i=1}^{n} [h_0(y_i)\rho(X_i^{cum}(y_i-5), \beta)]^{\delta_i^Y} \cdot \prod_{m=1}^{M_i} \frac{S_m(r_i^m, w_{i,m})}{S_m(r_i^{m-1}, w_{i,m})} \end{aligned} \tag{5.21}$$

5.2.2 Sous-modèle de mesure

Ce sous-modèle décrit la relation entre la vraie dose de rayonnements γ, la dose de rayonnements γ enregistrée et la dose de rayonnements γ présumée (voir partie 3.2).

Toutes les doses de rayonnements γ supposées nulles dans la Table 3.2 sont considérées comme des vrais zéros par la suite. Ainsi, ces doses ne sont pas considérées sujettes aux erreurs de mesure, et donc ne sont pas inclues dans le sous-modèle de mesure présenté ci-après. De plus, les doses présumées strictement positives (censurées à gauche ou non) dans la Table 3.2 sont traitées comme variables

latentes dans le modèle et leurs erreurs de mesure modélisées de la même manière que les doses de rayonnements γ mesurées strictement positives.

Soit $X_i(t)$ la vraie dose de rayonnements γ reçue par le mineur i au temps t. $Z_i(t)$ représente soit la dose mesurée strictement positive aux rayonnements γ du mineur i au temps t, soit la dose de rayonnements γ du mineur i au temps t qui aurait été mesurée si la dose n'était pas manquante et si les dosimètres n'avait pas de LD au temps t (noté $LD(t)$). $Z_i(t)$ est supposé sujet aux erreurs de mesure. En outre, $X_i(t)$ et $Z_i(t)$ sont traités comme des variables latentes, sauf lorsque $Z_i(t)$ correspond à une dose de rayonnements γ enregistrée comme strictement positive. Si $Z_i(t)$ n'est pas observé, $Z_i(t)$ est supposé suivre une distribution de probabilité, et est supposé inférieur à $LD(t)$ lorsque la dose est présumée censurée à gauche. À noter que dans ce cas, le processus de censure est déterministe puisque la valeur de $LD(t)$ est fixée pour chaque temps t : $LD(t) = 2.2$ mSv pour t entre 1956 et 1985, et $LD(t) = 0.55$ mSv pour t entre 1986 et 2007 (S. R. Allodji 2011).

Le sous-modèle de mesure est défini par le couple $(W_i(t), \delta_i^W(t))$ où $W_i(t) = \max(Z_i(t), LD(t))$ et $\delta_i^W(t)$ indique si $Z_i(t)$ est supérieur à LD ou non (c'est-à-dire $\delta_i^W(t) = 1$ si $Z_i(t) \geq LD(t)$ et $\delta_i^W(t) = 0$ si $Z_i(t) < LD(t)$). Si le poste du mineur d'uranium a été enregistré comme «normal» avec une dose présumée inférieure à LD (voir la Table 3.2), alors on suppose $W_i(t) = LD(t)$ et $\delta_i^W(t) = 0$. Si le poste du mineur d'uranium a été enregistré comme «expatrié» (voir la Table 3.2), alors $W_i(t) = \max(Z_i(t), LD(t))$ et $\delta_i^W(t) = 0$ ou 1.

Sachant que les doses de rayonnements γ sont basées sur des dosimètres individuels pendant toute la durée de l'exposition observée dans la sous-cohorte post-55 des mineurs d'uranium français (1956-2007), une erreur de mesure de type classique a été naturellement supposée pour refléter le manque de précision de l'appareil de mesure. Des erreurs de mesure partagées se produisent lorsque la différence entre la vraie dose et la dose mesurée dépend d'une composante d'erreur qui affecte simultanément plusieurs mineurs ou plusieurs temps d'un même mineur d'uranium. Dans notre contexte de dosimètres individuels, il n'y a aucune raison de supposer une erreur de mesure partagée dans le sous-modèle d'exposition, pour toute la période d'exposition.

Conformément avec la littérature suggérant le modèle d'erreurs de mesure multiplicatif comme plus réaliste en épidémiologie professionnelle en général (Armstrong

1998), on suppose une structure multiplicative et lognormale d'erreurs de mesure. Sachant que la dose de rayonnements γ a été mesurée à l'aide à des films badge de 1956 à 1985 et ensuite avec des TLDs de 1986 à 2007, la variance d'erreur de mesure a changé au cours du temps justifiant une approche hétéroscédastique. Ainsi, nous modélisons l'erreur de mesure de nature classique comme suit :

$$Z_i(t) = X_i(t) \cdot U_i(t) \tag{5.22}$$

Les termes d'erreur de mesure $U_i(t)$ sont supposé indépendant et suivre une distribution lognormale. L'espérance du logarithme de $U_i(t)$ est $-\sigma^2_{U,q(t)}/2$ et son écart-type est $\sigma^2_{U,q(t)}$:

$$U_i(t) \sim LogN(-\sigma^2_{U,q(t)}/2, \sigma^2_{U,q(t)}) \tag{5.23}$$

où $q(t)$ vaut 1 pour la période d'exposition entre 1956 et 1985, et vaut 2 pour la période d'exposition entre 1986 et 2007. Cette forte hypothèse sur la distribution log-normale implique que $\mathbb{E}(U_i(t)) = 1, \forall t, \forall i$ et ainsi $\mathbb{E}(Z_i(t)|X_i(t)) = X_i(t)$. $Z_i(t)$ est un estimateur sans biais de la vraie dose de rayonnements γ $X_i(t)$. $Z_i(t)$ étant sujet aux erreurs de mesure de nature classique, il est possible que la vraie dose de rayonnements γ $X_i(t)$ soit supérieure à $LD(t)$ même si $Z_i(t) < LD(t)$.

Il n'y a pas de données de validation qui permettraient d'estimer la vraie valeur de la variance de l'erreur de mesure. L'écart-type du logarithme du terme d'erreur de mesure $U_i(t)$ a été fixé grâce aux travaux de Allodji *et al.* (S. R. Allodji 2011) pour chaque période d'intérêt : $\sigma_{U,1} = 0.245$ entre 1956 et 1985 et $\sigma_{U,2} = 0.16$ entre 1986 et 2007. Pour estimer ces valeurs, un inventaire des potentielles causes des erreurs de mesure a été réalisé dans la sous-cohorte post-55 des mineurs d'uranium français : la précision de l'appareil de mesure à cause des champs médicaux, radiologiques et environnementaux, la perte de lecture et des problèmes dans la transmission des données. Ensuite, leur impact respectif sur l'erreur de mesure globale a été estimé à partir des études de Brady *et al.* (Brady 1985) et Gilbert *et al.* (ES Gilbert, J. Fix et W. Baumgartner 1996).

5.2.3 Sous-modèle d'exposition

La correction des erreurs de mesure de nature classique nécessite la spécification de la distribution de probabilité de la variable latente $X_i(t)$ représentant la vraie dose de rayonnements γ reçue par le mineur i au temps t. Les doses professionnelles sont souvent supposées suivre une distribution log-normale (Steenland et al. 2015). Comme suggéré par Allodji *et al.* (S. R. Allodji 2011), on suppose que les vraies doses de rayonnements γ suivent une distribution log-normale :

$$\log(X_i(t)) \sim N(\mu_{x,i}(t), \sigma^2_{x,i}(t)) \tag{5.24}$$

où $\mu_{x,i}(t)$ et $\sigma_{x,i}(t)$ sont l'espérance et l'écart-type du logarithme de la vraie dose $X_i(t)$ du mineur i au temps t. Pour des raisons de parcimonie, différentes hypothèses de modélisation sur la structure de $\mu_{x,i}(t)$ et $\sigma_{x,i}(t)$ ont été faites afin d'éviter d'avoir autant de paramètres inconnus que d'années d'exposition pour tous les mineurs d'uranium exposés dans la sous-cohorte post-55 des mineurs d'uranium. $\sigma_{x,i}(t)$ est supposé constant par morceaux, soit $\sigma_{x,i}(t) = \sigma_{x,p(t)} \forall i$ avec $p(t)$ qui vaut entre 1 et 5 correspondant à cinq périodes d'environ 10 ans. Les valeurs estimées de $\sigma_{x,1}, \sigma_{x,2}, \sigma_{x,3}, \sigma_{x,4}, \sigma_{x,5}$ étant similaires, on a finalement opté pour supposer $\sigma_{x,i}(t)$ constant au cours du temps : $\sigma_{x,i}(t) = \sigma_x, \forall i \forall t$.

Pour permettre à l'espérance de la dose à l'échelle logarithme $\mu_{x,i}(t)$ de varier au cours du temps, trois sous-modèles ont été proposés. Le premier sous-modèle d'exposition, noté $\mathcal{M}_1$, suppose $\mu_{x,i}(t)$ constant par morceaux au cours du temps pour tous les mineurs :

$$\mu_{x,i}(t) = \mu_{x,p(t)} \forall i \tag{5.25}$$

où $p(t)$ qui peut prendre comme valeur un entier entre 1 et 5, correspondant à cinq périodes d'exposition d'environ 10 ans, qui prennent en compte le changement de dosimètre en 1986 : 1956-1965, 1966-1975, 1976-1985, 1986-1995 et 1996-2007. Ainsi, $\boldsymbol{\mu_x} = (\mu_{x,1}, \mu_{x,2}, \mu_{x,3}, \mu_{x,4}, \mu_{x,5})$ est le vecteur des espérances de la vraie dose de rayonnements γ à l'échelle logarithme pour les cinq périodes d'exposition. Le vecteur des six paramètres inconnus du sous-modèle d'exposition $\mathcal{M}_1$ est donc $(\mu_{x,1}, \mu_{x,2}, \mu_{x,3}, \mu_{x,4}, \mu_{x,5}, \sigma_x)$.

Le second sous-modèle d'exposition, noté $\mathcal{M}_2$, est un modèle hiérarchique qui représente l'incertitude sur le paramètre $\mu_{x,i}(t)$ par une moyenne définie comme une fonction linéaire du temps de pente a et d'ordonnée à l'origine b inconnues. Cette fonction linéaire du temps est supposée identique pour tous les mineurs d'uranium français, ainsi :

$$\mu_{x,i}(t) \sim N(a \cdot f(t) + b, \sigma_\mu^2), \text{ où } f(t) = t - 1956 \tag{5.26}$$

Le paramètre b peut être interprété comme la valeur de l'espérance de la vraie dose de rayonnements γ à l'échelle logarithme en 1956 dans la sous-cohorte post-55 des mineurs d'uranium français. Pour des raisons de parcimonie, l'écart-type σ_μ de l'espérance du logarithme de la vraie dose de rayonnements γ est supposé constant au cours du temps. Le vecteur des quatre paramètres inconnus du sous-modèle d'exposition $\mathcal{M}_2$ est le suivant : $(a, b, \sigma_x, \sigma_\mu)$.

Finalement, on suppose que la tendance temporelle des vraies doses de rayonnements γ introduite dans $\mathcal{M}_2$ diffère entre les mines de jour et les mines de fond. Le troisième sous-modèle d'exposition, noté $\mathcal{M}_3$, est également un modèle hiérarchique qui prend en compte cette hypothèse, ainsi :

$$\mu_{x,i}(t) = \mu_{x,m_i(t)}(t) \tag{5.27}$$

avec

$$\mu_{x,J}(t) \sim N(a_J \cdot f(t) + b_J, \sigma_\mu^2) \tag{5.28}$$

$$\mu_{x,F}(t) \sim N(a_F \cdot f(t) + b_F, \sigma_\mu^2) \tag{5.29}$$

où F et J désignent respectivement les mines de fond et les mines de jour et $m_i(t) = \{F, J\}$ selon le type de mine dans laquelle le mineur i a travaillé au temps t. Ici, le sous-modèle dépend du mineur à travers la mine dans laquelle il a travaillé. La pente et l'ordonnée à l'origine de l'évolution linéaire dépendent tous les deux du type de mine. Le type de mine est manquant pour 11% des doses strictement positives. Dans ce cas, on suppose une distribution de Bernouilli pour la variable

$m_i(t)$ représentant le type de mine dans laquelle le mineur i a travaillé au temps t. Son paramètre est p_F, c'est-à-dire la probabilité inconnue pour un mineur de travailler dans une mine de fond. Le vecteur des sept paramètres inconnus du sous-modèle d'exposition $\mathcal{M}_3$ est le suivant : $(a_J, a_F, b_J, b_F, \sigma_x, \sigma_\mu, p_F)$.

5.3 Inférence bayésienne

5.3.1 Choix des distributions *a priori*

La Figure 5.1 montre le diagramme acyclique dirigé du modèle hiérarchique complet, soit la combinaison du sous-modèle de maladie, du sous-modèle de mesure et du sous-modèle d'exposition $\mathcal{M}_3$. Pour inférer le modèle hiérarchique dans le cadre bayésien, des distributions *a priori* ont été choisies pour tous les paramètres inconnus. Ces distributions *a priori* sont résumées dans la Table 5.1.

Comme précédemment, une distribution *a priori* normale centrée avec une large variance (10^6) a été choisie pour le coefficient de risque β. Cette distribution n'a pas été tronquée, mais la positivité du risque instantané $h_i(t)$ a été garantie durant l'inférence. Comme suggéré par Hoffmann *et al.* (Hoffmann, Estelle Rage et al. 2017), la distribution *a priori* du taux de mortalité par cancer du poumon de base λ_1 est une distribution *a priori* informative se basant sur les données de mortalité des hommes en France entre 1968 et 2005. Dans la sous-cohorte post-55 des mineurs d'uranium français, il n'y a qu'un seul cas de décès par cancer du poumon avant 40 ans, ainsi l'information contenue dans les données pour l'estimation du paramètre λ_1 est faible, d'où l'intérêt d'inclure de l'information *a priori*. Pour les autres paramètres du taux de mortalité par cancer du poumon de base, soit λ_2, λ_3 et λ_4, les distributions *a priori* choisies sont des lois uniformes entre 0 et 100.

Finalement, les distributions *a priori* des paramètres des trois sous-modèles d'exposition sont inspirées des distributions *a priori* non-informatives de Jeffrey (Jeffrey et al. 1992), qui sont invariantes par reparamétrisation. Le but est de refléter notre absence de connaissance sur la distribution des doses de rayonnements γ à l'échelle logarithme dans la sous-cohorte post-55 des mineurs d'uranium français, son évolution au cours du temps et d'autant plus son évolution au cours du temps en fonction du type de mine. Les distributions de tous les paramètres

des sous-modèles d'exposition sont données dans la Table 5.1.

	Paramètre	Famille	
Sous-modèle de maladie			
	β	Normal	$N(0; 10^6)$
	λ_1	Gamma	$G(23.7; 4.9 \cdot 10^8)$
	λ_2	Uniforme	$U[0; 100]$
	λ_3	Uniforme	$U[0; 100]$
	λ_4	Uniforme	$U[0; 100]$
Sous-modèle d'exposition $\mathcal{M}_1$			
	$\mu_{x,1}, \mu_{x,2}, \mu_{x,3}, \mu_{x,4}, \mu_{x,5}$		$\propto 1$
	τ_x		$\propto 1/\tau_x$
Sous-modèle d'exposition $\mathcal{M}_2$			
	a		$\propto 1$
	b		$\propto 1$
	τ_x		$\propto 1/\tau_x$
	τ_μ		$\propto 1/\tau_\mu$
Sous-modèle d'exposition $\mathcal{M}_3$			
	a_J, a_F		$\propto 1$
	b_J, b_F		$\propto 1$
	p_F	Gamma	$G(1/2; 1/2)$
	τ_x		$\propto 1/\tau_x$
	τ_μ		$\propto 1/\tau_\mu$

TABLE 5.1 – Distributions *a priori* de tous les paramètres du modèle hiérarchique prenant en compte les incertitudes sur la dose de rayonnements γ

5.3.2 Détails de l'algorithme MCMC implémenté

Un algorithme MCMC a été codé en Python pour échantillonner dans la loi *a posteriori* jointe des paramètres inconnus et des variables latentes des modèles hiérarchiques combinant le sous-modèle de maladie, le sous-modèle de mesure et un des trois sous-modèles d'exposition. Les distributions *a posteriori* n'ayant pas toutes une écriture analytique, un algorithme de Metropolis-within-Gibbs adaptatif (ROBERTS et ROSENTHAL 2009) a été implémenté pour inférer le modèle hiérarchique dans le cadre bayésien. Une étape d'ajustement de la variance de la loi de proposition est réalisée pour améliorer l'efficacité et la convergence de l'algorithme, en visant un taux d'acceptation de 40% pour les paramètres seuls, et

20% pour les vecteurs de paramètre (Roberts et Rosenthal 2009). La Table 5.2 indique le type d'échantillonneur MCMC utilisé pour mettre à jour chaque quantité inconnue inférée et donne des précisions supplémentaires concernant le type de mise à jour réalisée (par bloc/individuel, par calcul vectoriel/utilisation d'une boucle for). Deux types d'échantillonneurs MCMC différents ont été utilisés pour les mises à jour : un échantillonneur de Gibbs dans les cas de conjugaison donnant accès à une loi conditionnelle complète explicite ou un échantillonneur de Métropolis-Hasting avec marche aléatoire Gaussienne (loi de proposition) dans les autres cas. Les paramètres et les variables latentes ont été mis à jour soit individuellement (ex : paramètres définissant le risque de base de décès par cancer du poumon) soit par bloc (ex : vecteur des doses vraies d'un même mineur d'uranium).

Trois chaînes avec des valeurs initiales différentes ont été lancées afin de vérifier l'absence de problème de convergence. Après 100 étapes de 100 itérations chacune comme phase adaptative, puis 10 000 itérations pour le temps de chauffe, on échantillonne finalement 55 000 itérations pour chaque modèle. Pour réduire la corrélation intra-chaînes, on fait un pas de 30, c'est-à-dire qu'on sauvegarde 1 itération toutes les 30 itérations. L'échantillon *a posteriori* contient donc 5 500 valeurs pour chaque paramètre et variable latente.

La mise à jour des vecteurs de paramètres grâce au calcul vectoriel, c'est-à-dire sans boucle «for», a permis de gagner un temps de calcul considérable. Cela a été possible grâce aux packages numpy et pandas disponibles sous Python. Un effort particulier a ainsi été réalisé afin d'utiliser autant que possible le calcul vectoriel pour la mise à jour des plus longs vecteurs de variables latentes. Il s'agissait principalement des vecteurs regroupant l'ensemble des dose $X_i(t)$ ou $Z_i(t)$ relatives à un même mineur i : leur longueur correspondait donc au nombre (variable) d'années d'exposition du mineur i. Concernant les paramètres mis à jour via un échantillonneur de Metropolis-Hastings, les ratios d'acceptation ont été calculés soit en utilisant des fonctions pré-codées sous Python (i.e., calculant directement les densités des distributions considérées), soit en codant directement ces ratios dont l'écriture mathématique peut parfois se simplifier selon les choix de modélisation effectués. Les deux méthodes ont été chronométrées et comparées afin de sélectionner la plus rapide.

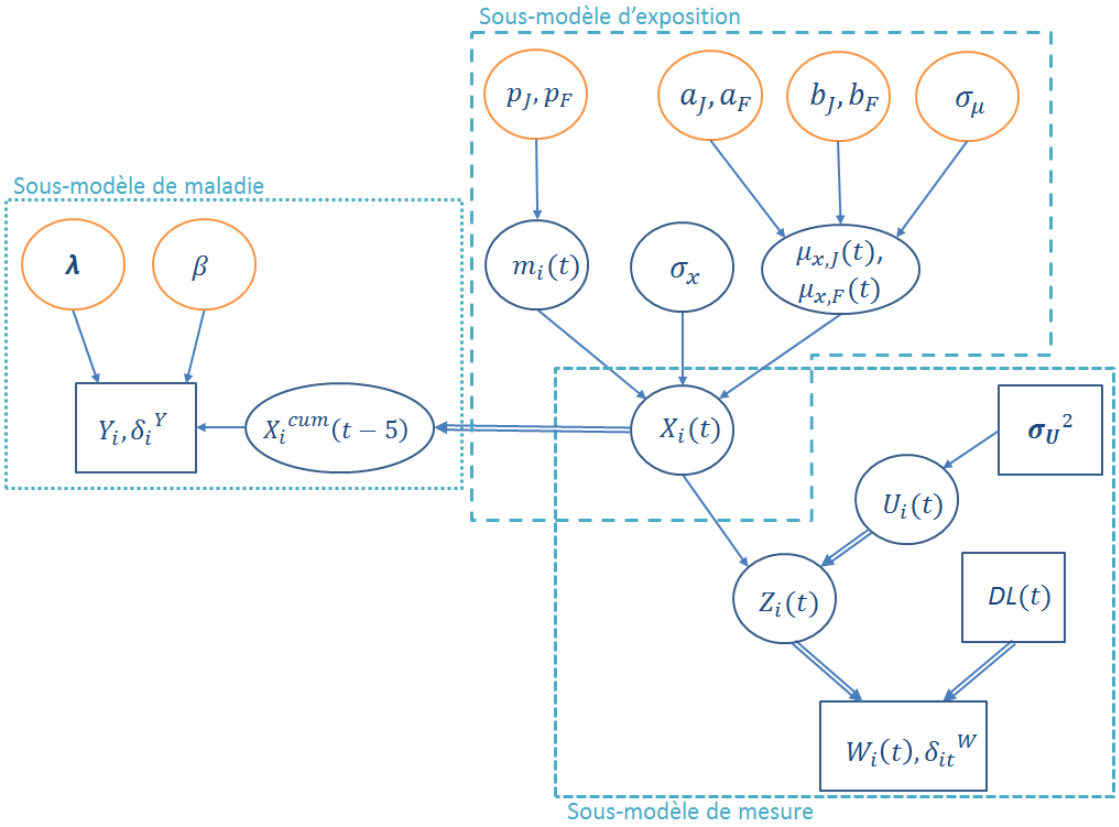

FIGURE 5.1 – Graphique acyclique dirigé pour le modèle hiérarchique complet combinant le sous-modèle de maladie, du sous-modèle de mesure et du sous-modèle d'exposition $\mathcal{M}_3$. Les cercles indiquent les quantités inconnues et les rectangles les variables observées ou les paramètres fixes. Les flèches simples indiquent des liens probabilistes orientés entre deux quantités et les doubles flèches indiquent des liens déterministes orientés entre deux quantités.

Paramètre	Échantillonneur	Par paramètre/ Par mineur	Par boucle/ calcul vectoriel
$\lambda_1, \lambda_2, \lambda_3, \lambda_4$	RWMH	Par paramètre	Boucle
β	RWMH	.	.
$X_i(t)$	RWMH	Par mineur	Vectoriel
$Z_i(t)$	RWMH	Par mineur	Vectoriel
$m_i(t)$	RWMH	Par paramètre	Vectoriel
σ_x	Gibbs	.	.
$\mu_{x,1}, \mu_{x,2}, \mu_{x,3}, \mu_{x,4}, \mu_{x,5}$	RWMH	Par paramètre	Boucle
$\mu_x(t)$	RWMH	Par paramètre	Boucle
a	Gibbs	.	.
b	Gibbs	.	.
$\mu_{x,J}(t), \mu_{x,F}(t)$	RWMH	Par paramètre	Boucle
a_J	Gibbs	.	.
a_F	Gibbs	.	.
b_J	Gibbs	.	.
b_F	Gibbs	.	.
σ_μ	Gibbs	.	.
p_F	Gibbs	.	.

TABLE 5.2 – Nature de l'échantillonneur MCMC et du mouvement réalisé (par bloc=par mineur, individuel=par paramètre) pour la mise de jour de chaque paramètre et de chaque vecteur latent, défini dans les 3 modèles hiérarchiques proposés. RWMH désigne un échantillonneur de type Metropolis-Hastings avec marche aléatoire, Gibbs désigne un échantillonneur de Gibbs. La troisième colonne du tableau précise, pour les vecteurs de paramètres, si le mouvement est accepté pour chaque paramètre individuellement ou pour l'ensemble des paramètres d'un même mineur d'uranium. La dernière colonne indique comment cela a été codé en python c'est-à-dire soit à l'aide d'une boucle parcourant tous les éléments du vecteur, soit par calcul vectoriel.

CHAPITRE 6

Prise en compte d'expositions radiologiques multiples et fortement corrélées dans l'estimation d'un risque radio-induit

La notion d'exposome regroupe différentes situations notamment de par le nombre de variables considérées dans cette notion d'exposome. Une première situation consiste à considérer un très grand nombre de variables corrélées servant à définir l'exposome. C'est typiquement le cas dans le domaine de la génétique où les associations entre de nombreux marqueurs génétiques et une pathologie d'intérêt sont recherchées. Le génome entier est parfois même considérée dans les approches dites «Genome-Wide Association Studies» (GWAS). Typiquement, dans de telles situations à grand nombre de variables, les techniques de réduction de dimension sont privilégiées. Elles permettent de sélectionner un plus petit nombre de variables représentatives de la globalité. Une seconde situation concerne un nombre plus petit de variables fortement corrélées composant l'exposome. C'est typiquement le cas dans ce travail de thèse. En effet, les différentes expositions considérées correspondent à différentes sources d'exposition aux RIs mais leur nombre reste très raisonnable. Les méthodes recherchées dans ce contexte sont donc bien différentes.

6.1 État de l'art

Bien qu'elles ne soient pas encore largement utilisées dans la pratique (Vatcheva et al. 2016), plusieurs méthodes statistiques ont été proposées pour traiter la multicolinéarité et étudier le potentiel effet combiné de facteurs de risque environnementaux fortement corrélés sur l'évènement d'intérêt. De nombreuses études antérieures s'appuyaient sur une méthode dite «environment-wide association study» (EWAS) où l'association entre chaque facteur d'exposition et l'évènement d'intérêt était estimée séparément (Patel et Butte 2010 ; Rage et al. 2015). Même si elle est potentiellement utile pour découvrir les facteurs de risque prioritaires, cette approche est principalement envisagée dans une phase de recherche exploratoire et conduit à des recherches limitées sur une association santé-exposome. D'autres approches proposées dans ce contexte spécifique reposent principalement sur : a) la sélection de variables dans un contexte de régression en utilisant, par exemple, le critère elastic-net (Lenters, Portengen, Rignell-Hydbom et al. 2016) ou la «Graphical Unit Evolutionary Stochastic Search» (GUESS), la recherche stochastique évolutive en unité graphique (Bottolo et S. Richardson 2010) ; b) la réduction de dimensions basée sur les données en utilisant la régression sur les composantes principales (Massy 1965) ou la régression partielle sur les moindres carrés (Jain et al. 2018 ; Wold, Ruhe et al. 1984) ; c) des algorithmes d'apprentissage automatique (ou machine learning) comme les k-means et le partitionnement récursif utilisant des forêts aléatoires (Marshall 2001) et d) des approches de regroupement de données corrélées multiples (Forgy 1965) comme l'analyse de classe latente (LCA) (Patterson, Dayton et Graubard 2002) et les modèles de mélange bayésiens dits de régression sur profil (PRM, Profile Regression Mixture) (Molitor et al. 2010).

Les méthodes de sélection de variables sont des outils très intéressants pour identifier un petit sous-ensemble de facteurs de risque environnementaux qui ont le plus d'influence sur l'évènement de santé qui nous intéresse. Elles sont particulièrement adaptées lorsque l'on considère un très grand nombre de facteurs de risque. Toutefois, lorsque seules quelques covariables d'exposition fortement corrélées sont disponibles, l'idée n'est pas d'en omettre certaines dans l'étude, mais plutôt d'estimer une relation exposition-risque en utilisant toutes les covariables

disponibles et des méthodes statistiques appropriées pour traiter les questions de multicolinéarité. Elles peuvent également être limitées dans leur capacité à différencier efficacement les vrais prédicteurs des covariables corrélées lorsque ces dernières sont très fortement corrélées (Agier et al. 2016). La réduction de dimensions reposant sur les données vise à construire des variables latentes non corrélées en tant que combinaisons linéaires des covariables d'exposition d'origine et ensuite, à inclure ces nouvelles variables non corrélées résumant les données dans un modèle de régression multiple (Wold, Esbensen et Geladi 1987). Un inconvénient majeur est que ces variables sont construites sans prendre en compte l'évènement d'intérêt dans la construction des variables latentes utilisées par la régression en composante principale (PCR). Même si l'approche sPLS (Jain et al. 2018) corrige ce problème en construisant des variables latentes non corrélées comme des combinaisons linéaires des covariables d'origine et de la variable réponse, un autre inconvénient des approches de réduction de dimensions basées sur les données concerne les incertitudes liées à cette construction. En effet, étant donné que les risques de maladie sont estimés dans une deuxième étape disjointe, la perte d'informations sur l'incertitude associée à cette construction peut conduire à une interprétation trompeuse des estimations de risque. Enfin, les algorithmes d'apprentissage automatique sont des approches à la fois pertinentes et efficaces pour traiter un grand nombre de facteurs de risque mais ne permettent pas d'obtenir des estimations de risques sanitaires.

Dans ce travail, nous nous sommes concentrés sur le problème spécifique de l'estimation de l'effet combiné sur la santé - en termes d'excès de risque - de quelques covariables d'exposition environnementale, mais fortement corrélées, à partir d'un résultat de survie censuré. Nous avons choisi les modèles PRM. Ce sont des modèles de mélange infini qui relient la survenue d'une maladie à un ensemble de covariables corrélées par l'appartenance à un groupe. Ils sont basés sur un processus de mélange de Dirichlet comme sous-modèle d'attribution. En saisissant l'hétérogénéité entre les covariables, les modèles PRM permettent à la fois d'identifier des schémas spécifiques de valeurs de covariables - appelés profils de covariables - qui sont représentatifs d'une sous-population (c'est-à-dire d'un groupe) et de les associer à la survenue de la maladie par le biais d'un modèle de régression. Ensuite, l'inférence de ce modèle probabiliste permet à la fois de regrouper simultanément

des individus présentant des risques similaires et des caractéristiques d'exposition proches et d'estimer le risque associé pour chaque groupe. Par rapport à la LCA et l'algorithme k-means, l'un des principales avantages des modèles PRM est que la survenue de la maladie influence l'appartenance à un groupe. Ainsi, la survenue de la maladie peut guider l'inférence vers les structures de regroupement des individus les plus pertinentes et n'est pas seulement utilisée pour le post-traitements. Une autre motivation des modèles PRM est que le nombre de groupes est inconnu et estimé grâce aux données. De plus, l'adaptation des modèles PRM au paradigme bayésien offre des avantages supplémentaires. Tout d'abord, il permet de traiter les nombreuses variables latentes incluses dans ces modèles complexes et d'obtenir des réponses probabilistes à la question étudiée. Deuxièmement, toute l'incertitude, y compris l'incertitude associée au regroupement des individus, se reflète dans des intervalles de crédibilité sur les paramètres de risque. Troisièmement, elle offre la possibilité d'inclure des informations externes sur les paramètres sous la forme de distributions *a priori*, ce qui est particulièrement utile lorsque certaines quantités inconnues d'intérêt ne sont pas ou seulement mal renseignées par les données. Enfin, elle permet de prédire le risque de maladie d'un individu multi-exposé tout en conservant l'incertitude des paramètres estimés. Ces modèles ont déjà été utilisés dans divers domaines, notamment la génétique (M. Papathomas et al. 2012), l'épidémiologie environnementale (Molitor et al. 2010 ; Michail Papathomas et al. 2011 ; Coker et al. 2018 ; Pirani et al. 2015 ; S. Liverani, Lavigne et Blangiardo 2016) et l'épidémiologie professionnelle (D. Hastie et al. 2013) mais jamais en épidémiologie des rayonnements ionisants. À noter qu'un package R appelé PReMiuM (Silvia Liverani et al. 2015) existe pour inférer des modèles PRM dans le cadre bayésien pour une variable d'intérêt modélisée par une distribution gaussienne, binaire, ordinale, catégorielle, de Poisson et de Weibull pour une variable censurée.

Nous avons étendu les modèles PRM pour traiter une variable de survie censurée suivant un modèle d'excès de risque instantané. Ce type de modèles de survie est couramment utilisé pour estimer les risques de cancer en épidémiologie des rayonnements ionisants (Hoffmann, Estelle Rage et al. 2017) mais n'est pas mis en oeuvre dans le package PReMiuM. L'inférence bayésienne du modèle PRM proposé est réalisée avec un algorithme adaptatif de Metropolis-Within-Gibbs im-

plémenté en Python et comprenant trois mouvements de changement de label de groupe. Nous avons appliqué notre modèle PRM au problème spécifique de l'estimation du risque de décès par cancer du poumon chez les mineurs d'uranium français multi-exposés. En effet, dans le cadre de leur travail, les mineurs d'uranium sont simultanément exposés au radon, aux rayonnements γ externes et à la poussière d'uranium (ainsi qu'à d'autres agents chimiques et physiques). Il est intéressant de noter que ces trois sources d'exposition aux rayonnements ionisants sont fortement corrélées entre elles dans la cohorte des mineurs d'uranium français. En fait, elles sont associées au même phénomène de désintégration de l'uranium, qui est omniprésent dans les mines d'uranium (Vacquier, Estelle Rage et al. 2011). De plus, à ce stade, un effet additif ou synergique de la co-exposition à ces différents composants radiologiques sur les risques de cancer du poumon ne peut être exclu. Jusqu'à présent, la plupart des études épidémiologiques sur la cohorte française des mineurs d'uranium se sont concentrées sur l'étude de l'association entre une exposition chronique et à faible dose au radon et la mortalité par cancer du poumon, comme si le radon - qui est considéré comme la deuxième cause de cancer du poumon après le tabagisme (IARC et al. 1988) - avait un effet isolé. Comme présenté dans la section 2.4, une approche EWAS a été réalisée, dans laquelle l'association entre chaque source unique de rayonnement ionisant et le risque de décès par cancer du poumon a été estimée séparément, en utilisant un modèle de régression de Poisson. Cela a montré que chaque source de rayonnement ionisant était associée de manière significative à un risque plus élevé de décès par cancer du poumon dans la cohorte française des mineurs d'uranium (Rage et al. 2015). Nous proposons de traiter la question de la multicolinéarité dans cette étude de cas, en utilisant le modèle PRM bayésien proposé. À notre connaissance, il s'agit de la première application des modèles PRM bayésiens pour traiter la forte corrélation entre les co-expositions en épidémiologie des rayonnements ionisants.

6.2 Extension des modèles de régression bayésienne sur profils d'exposition au contexte de données de survie

Pour traiter de la multicolinéarité dans le contexte spécifique de l'estimation de l'effet combiné de quelques covariables d'exposition fortement corrélées, nous avons opté pour un modèle PRM bayésien. Il s'agit d'un modèle hiérarchique dans lequel trois sous-modèles doivent être spécifiés et reliés, par hypothèse d'indépendance conditionnelle :

- le *sous-modèle de maladie* : il décrit l'association entre la survenue d'une maladie (par exemple, l'âge au décès par cancer du poumon d'un mineur) et un profil d'exposition ;
- le *sous-modèle d'exposition* : il définit la distribution de probabilité des différentes covariables d'intérêt dans chaque groupe, afin de caractériser des profils d'exposition spécifiques (sous-modèle d'exposition) ;
- le *sous-modèle d'attribution* : il décrit l'affectation aléatoire d'un individu à un profil (ou groupe) donné.

6.2.1 Sous-modèle de maladie

Le sous-modèle de maladie utilisé classiquement en épidémiologie des rayonnements ionisants est un modèle en excès de risque instantané (EHR). Le taux de risque instantané de décès par cancer du poumon d'un mineur d'uranium i au moment t, noté $h_i(t)$ est défini par :

$$h_i(t) = h_0(t) \cdot (1 + \beta_{C_i}) \tag{6.1}$$

Le risque de base $h_0(t)$ est le risque instantané de décès par cancer du poumon au moment t du profil des non exposés (le groupe de référence des mineurs non exposés aux rayonnements ionisants), C_i est le label du groupe auquel appartient le mineur i et β_c est l'excès de risque instantané de décès par cancer du poumon du groupe c. Ainsi, deux mineurs appartenant au même groupe c ont le même risque

de décès par cancer du poumon. À noter que pour pour assurer la positivité de $h_i(t)$, β_c est soumis à la contrainte $\beta_c > -1, \forall c$.

Selon Hoffmann *et al.* (Hoffmann, Estelle Rage et al. 2017), on suppose $h_0(t)$ constant par morceaux sur quatre intervalles d'âge pour lesquels les valeurs du risque instantané de base sont supposées être constantes. Ces quatre intervalles correspondent à une partition de l'axe des âges : avant 40 ans, entre 40 et 55, entre 55 et 70 et enfin après 70 ans. Les quatre constantes correspondantes du risque de base sont désignées par $\lambda_1, \lambda_2, \lambda_3$ et λ_4.

6.2.2 Sous-modèle d'exposition

Le sous-modèle d'exposition définit des groupes basés sur des niveaux de covariables et sur un risque de décès par cancer du poumon similaires. On définit donc la distribution de probabilité des covariables conditionnellement à un groupe. Les différentes covariables considérées pour les groupes comprennent les expositions cumulées aux rayonnements ionisants et d'autres caractéristiques des mineurs d'uranium français. Les détails de ces covariables sont les suivants :

- L'exposition professionnelle cumulée au radon X_i^R, aux rayonnements γ X_i^G et aux poussières d'uranium X_i^P pendant toute la période de suivi du mineur i ;

- Le type d'emploi J_i le plus occupé par le mineur i. Cette variable catégorielle comporte cinq modalités : 1) foreur avant la mécanisation, 2) foreur après la mécanisation, 3) autres travaux souterrains avant la mécanisation, 4) autres travaux souterrains après la mécanisation et 5) travaux de surface ;

- L'âge à la première exposition A_i du mineur i. La sensibilité aux rayonnements ionisants peut être fonction de l'âge au moment de l'exposition (Crosfill, Lindop et Rotblat 1959) ;

- La localisation de la mine M_i. Nous avons distingué la mine de l'Hérault et les autres en fonction du type de gisement ;

- La durée de l'exposition T_i du mineur i. On considère quatre catégories avec un nombre similaire de mineurs d'uranium : les mineurs qui ont été exposés

5 ans et moins, 6 à 12 ans, 13 à 18 ans et enfin ceux qui ont été exposés pendant au moins 19 ans.

La distribution de probabilité de chaque covariable dépend de paramètres qui sont fonction du groupe c. Nous avons supposé des distributions lognormales $\text{LogN}(\mu_c^X, \sigma_c^X)$ pour les variables positives et continues et des distributions multinomiales $\text{Multinomial}(p_c^X)$ pour les variables catégorielles. Les différentes distributions sont les suivantes :

$$\begin{cases} X_i^R|C_i = c, \mu_c^R, \sigma_c^R & \sim & LogN(\mu_c^R, \sigma_c^R) \\ X_i^G|C_i = c, \mu_c^G, \sigma_c^G & \sim & LogN(\mu_c^G, \sigma_c^G) \\ X_i^P|C_i = c, \mu_c^P, \sigma_c^P & \sim & LogN(\mu_c^P, \sigma_c^P) \\ A_i|C_i = c, \mu_c^A, \sigma_c^A & \sim & LogN(\mu_c^A, \sigma_c^A) \\ J_i|C_i = c, p_c^J & \sim & Multinomial(p_c^J) \\ M_i|C_i = c, p_c^M & \sim & Multinomial(p_c^M) \\ T_i|C_i = c, p_c^T & \sim & Multinomial(p_c^T) \end{cases} \tag{6.2}$$

6.2.3 Sous-modèle d'attribution

Le sous-modèle d'attribution associe le mineur i à un groupe C_i en fonction de la probabilité ϕ_c d'appartenir au groupe c. Soit C_{max} le nombre maximum de groupes, $\boldsymbol{\phi} = (\phi_1, \phi_2, ..., \phi_{C_{max}})$ définit le vecteur des probabilités d'affectation à chaque groupe. Le vecteur de paramètres $\boldsymbol{\phi}$ suit un processus de Dirichlet. En raison du processus de Dirichlet, le nombre de groupes non-vides n'est pas fixé arbitrairement mais estimé, seul le nombre maximum de groupes C_{max} est donné. La construction de ces poids de mélange $\boldsymbol{\phi} = (\phi_1, \phi_2, ..., \phi_{C_{max}})$, appelée «stick-breaking», est la suivante :

$$V_c \sim Beta(1, \alpha), c \in \{1, ..., C_{max} - 1\} \tag{6.3}$$

$$\phi_c = V_c \cdot (1 - \sum_{k=1}^{c-1} \phi_k), c \in \{1, ..., C_{max} - 1\} \tag{6.4}$$

$$\phi_{C_{max}} = 1 - \sum_{k=1}^{C_{max}-1} \phi_k \tag{6.5}$$

Le nombre de groupes non-vides est guidé par α. Une petite valeur de α réduit la probabilité d'avoir un grand nombre de groupes non-vides, et respectivement. Cette construction «stick-breaking» est une approximation finie du modèle de regroupement infini. La valeur de C_{max} doit être choisie assez grande pour donner une bonne approximation mais suffisamment petite pour éviter des calculs inutiles. C_{max} doit être réglé de manière à ce que la probabilité $\phi_{C_{max}}$ soit faible (Ohlssen, Sharples et Spiegelhalter 2007). Le choix de C_{max} est fortement influencé par la valeur de α, et pour une valeur de α jusqu'à 10, la probabilité $\phi_{C_{max}}$ est négligeable avec C_{max} égal à 50 (Ishwaran et Zarepour 2000). Quelques lignes directrices et une description plus détaillée sont données dans Molitor *et al* (Molitor et al. 2010).

6.3 Inférence bayésienne par algorithme MCMC

6.3.1 Choix des distributions *a priori*

Les distributions *a priori* choisies sont peu informative, sauf pour les paramètres impliqués dans le taux de base de mortalité par cancer du poumon, dans le «stick-breaking» et les moyennes d'exposition pour lesquelles des informations externes étaient disponibles.

Ainsi, des distributions normales centrées avec une grande variance ont été considérées pour les paramètres de risque β_c et pour les moyennes d'âge à la première exposition μ_c^A (à l'échelle logarithmique) dans chaque groupe $c, c = 1, ..., C_{max}$. De larges distributions uniformes ont été choisies pour les paramètres d'écart-type géométrique des distributions lognormales σ_c^R, σ_c^G, σ_c^P et σ_c^A. Les distributions *a priori* de Dirichlet avec des paramètres égaux à $1/2$ ont été supposées pour les paramètres des distributions multinomiales, à savoir p_c^J, p_c^M et p_c^T.

Concernant la moyenne des niveaux d'exposition aux rayonnements γ μ_c^G, au radon μ_c^R et aux poussières d'uranium μ_c^P (à l'échelle logarithmique), des informations sont disponibles auprès de la cohorte allemande des mineurs d'uranium (M. Kreuzer et al. 2017). Des distributions *a priori* normales ont été supposées pour μ_c^G, μ_c^R et μ_c^P avec des moyennes et des variances basées sur les niveaux

d'exposition de cette cohorte.

Comme les paramètres impliqués dans le risque de base sont mal renseignés par les données, en particulier pour les jeunes mineurs, des données externes sur la mortalité par cancer du poumon chez les hommes en France entre 1968 et 2005 ont été utilisées pour préciser les distributions gamma *a priori* informatives sur les paramètres λ_1, λ_2, λ_3 et λ_4 définissant le risque de base de décès par cancer du poumon chez les mineurs d'uranium français (supposé constant par morceaux).

Enfin, comme recommandé par Molitor (Molitor et al. 2010), nous avons utilisé une distribution *a priori* uniforme sur l'intervalle $[0.3\,;\,10]$ pour le paramètre α qui influence le nombre de groupes non-vides *a posteriori*. Tous les détails des distributions *a priori* sont donnés dans la Table 6.1.

Sous-modèle de maladie	Paramètre	Famille	
	β_c	Normale	N(0, 10^6)
	λ_1	Gamma	G$(23.7, 4.9 \cdot 10^8)$
	λ_2	Gamma	G$(35.5, 2.6 \cdot 10^7)$
	λ_3	Gamma	G$(88.1, 1.6 \cdot 10^7)$
	λ_4	Gamma	G$(29.7, 3.2 \cdot 10^6)$
Sous-modèle d'exposition			
	μ_c^G	Normale	N(0.10, 2.25)
	μ_c^R	Normale	N(-2.3, 8.08)
	μ_c^P	Normale	N(1.01, 11.79)
	μ_c^A	Normale	N(0, 10^6)
	$\sigma_c^G, \sigma_c^R, \sigma_c^P, \sigma_c^A$	Uniforme	U[0, 100]
	p_c^J, p_c^M, p_c^T	Dirichlet	D(0.5,..., 0.5)
Sous-modèle d'attribution			
	α	Uniforme	U[0.3,10]

Table 6.1 – Distributions de probabilités *a priori* des paramètres inconnus d'un modèle bayésien PRM pour le sous-modèle de maladie, le sous-modèle d'exposition et le sous-modèle d'attribution.

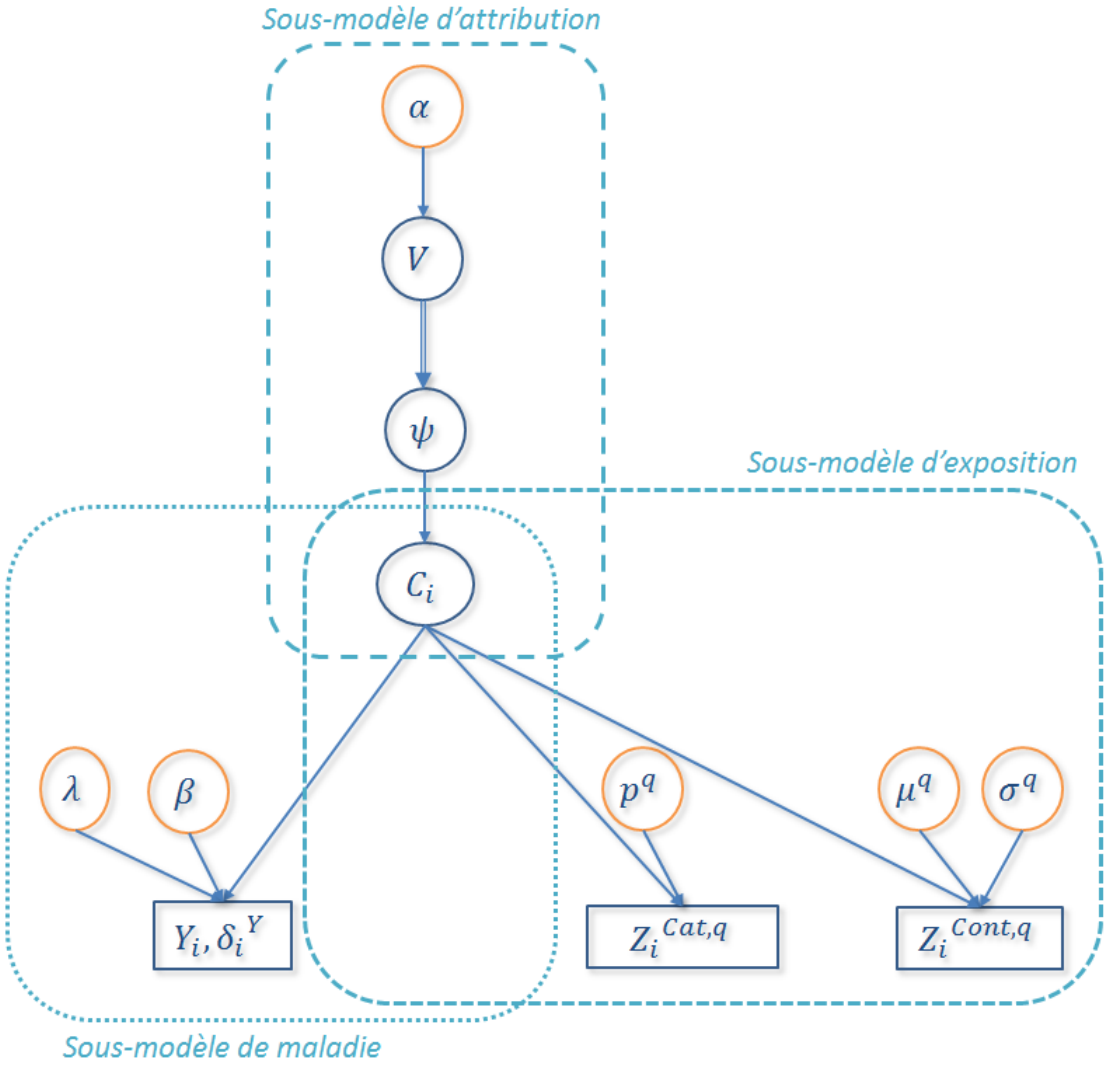

FIGURE 6.1 – Diagramme acyclique dirigé associé au modèle bayésien complet PRM. Les cercles indiquent les quantités inconnues et les rectangles les variables observées. Les flèches simples indiquent des liens probabilistes orientés entre deux quantités et les doubles flèches indiquent des liens déterministes orientés entre deux quantités. $Z_i^{Cat,q}$ indique la valeur observée de toute covariable catégorielle q du mineur d'uranium i et $Z_i^{Cont,q}$ indique la valeur observée de toute covariable continue q du mineur d'uranium i.

6.3.2 Détails de l'algorithme MCMC implémenté

La Figure 6.1 montre le graphique acyclique dirigé pour le modèle hiérarchique complet combinant le sous-modèle de maladie, le sous-modèle d'exposition et le sous-modèle d'attribution. Le package R «PReMiuM» existe déjà pour mettre en œuvre la régression du profil bayésien (Silvia Liverani et al. 2015) pour une variable réponse modélisée par une distribution de Bernoulli, binomiale, de Poisson, Normale, catégorielle ainsi que le modèle de survie de Weibull. Malheureusement, le modèle de survie en EHR n'est pas une option possible dans ce package. Ainsi, un algorithme MCMC a été implémenté en Python pour échantillonner à partir de la distribution *a posteriori* jointe de tous les paramètres et les variables latentes inconnus.

Nous avons utilisé un algorithme de Metropolis-within-Gibbs (Roberts et Rosenthal 2009) pour effectuer l'inférence bayésienne, car les distributions conditionnelles complètes n'étaient pas toujours calculable analytiquement. Une phase adaptative d'étapes de Metropolis-Hastings, qui est nécessaire pour améliorer la convergence et l'efficacité de l'algorithme, met à jour la variance de chaque distribution de proposition pour viser un taux d'acceptation de 40% pour les paramètres uniques et de 20% pour les vecteurs de paramètre (ibid.). La Table 6.2 indique le type d'échantillonneur MCMC utilisé pour mettre à jour chaque quantité inconnue inférée et donne des précisions supplémentaires concernant le type de mise à jour réalisée (par bloc/individuel, par calcul vectoriel/utilisation d'une boucle for). Deux types d'échantillonneurs MCMC différents ont été utilisés pour les mises à jour : un échantillonneur de Gibbs dans les cas de conjugaison donnant accès à une loi conditionnelle complète explicite ou un échantillonneur de Métropolis-Hasting avec marche aléatoire Gaussienne (loi de proposition) dans les autres cas. Les paramètres et les variables latentes ont été mis à jour soit individuellement soit par bloc (ex : vecteur des V_c).

Nous avons effectué 100 étapes de 100 itérations pour la phase adaptative, puis 10 000 itérations n'ont pas été conservées car correspondant à la phase de chauffe et enfin 150 000 itérations supplémentaires ont été effectuées. Pour diminuer les autocorrélations intra-chaîne, nous avons réduit l'échantillon *a posteriori* en ne stockant qu'une itération toutes les 20 itérations. Cela a par ailleurs permis de réduire

les temps de sauvegarde de l'ensemble des valeurs échantillonnées. L'échantillon *a posteriori* de chaque quantité inconnue contient donc 7500 valeurs.

Paramètre	Échantillonneur	Par paramètre/ Par vecteur	Par boucle/ Calcul vectoriel
$\beta_c, \forall c$	RWMH	Par paramètre	Vectoriel
$\lambda_1, \lambda_2, \lambda_3, \lambda_4$	RWMH	Par paramètre	Boucle
α	RWMH	.	.
$C_i, \forall i$	Gibbs	Par paramètre	Vectoriel
$V_c, \forall c$	RWMH	Par vecteur	Vectoriel
$\mu_c^G, \forall c$	RWMH	Par paramètre	Vectoriel
$\sigma_c^G, \forall c$	RWMH	Par paramètre	Vectoriel
$\mu_c^R, \forall c$	RWMH	Par paramètre	Vectoriel
$\sigma_c^R, \forall c$	RWMH	Par paramètre	Vectoriel
$\mu_c^P, \forall c$	RWMH	Par paramètre	Vectoriel
$\sigma_c^P, \forall c$	RWMH	Par paramètre	Vectoriel
$\mu_c^A, \forall c$	RWMH	Par paramètre	Vectoriel
$\sigma_c^A, \forall c$	RWMH	Par paramètre	Vectoriel
$p_c^J, \forall c$	Gibbs	Par paramètre	Vectoriel
$p_c^T, \forall c$	Gibbs	Par paramètre	Vectoriel
$p_c^M, \forall c$	Gibbs	Par paramètre	Vectoriel

TABLE 6.2 – Nature de l'échantillonneur MCMC et du mouvement réalisé (par bloc=par mineur, individuel=par paramètre) pour la mise de jour de chaque paramètre et de chaque vecteur latent, défini dans le modèle PRM. RWMH désigne un échantillonneur de type Metropolis-Hastings avec marche aléatoire, Gibbs désigne un échantillonneur de Gibbs. La troisième colonne du tableau précise, pour les vecteurs de paramètres, si le mouvement est accepté pour chaque paramètre individuellement ou pour l'ensemble des paramètres d'un même mineur d'uranium. La dernière colonne indique comment cela a été codé en python c'est-à-dire soit à l'aide d'une boucle parcourant tous les éléments du vecteur, soit par calcul vectoriel.

La mise à jour des vecteurs de paramètres grâce au calcul vectoriel, c'est-à-dire sans boucle «for», a permis de gagner un temps de calcul considérable. Cela a été possible grâce aux packages numpy et pandas disponibles sous Python. Les paramètres λ_1, λ_2, λ_3 et λ_4 sont les seuls à être mis à jour via une boucle. En effet, pour ces paramètres du taux de mortalité de base, il n'était pas possible de passer par un calcul vectoriel puisque la probabilité d'accepter la valeur candidate de λ_2 dépendait de la nouvelle valeur de λ_1. Concernant les paramètres mis à

jour via un échantillonneur de Metropolis-Hastings, les ratios d'acceptation ont été calculés soit en utilisant des fonctions pré-codées sous Python (i.e., calculant directement les densités des distributions considérées), soit en codant directement ces ratios dont l'écriture mathématique peut parfois se simplifier selon les choix de modélisation effectués. Les deux méthodes ont été chronométrées et comparées afin de sélectionner la plus rapide.

Une attention particulière a été portée sur la convergence vers les modes locaux. En outre, comme le suggère Liverani *et al.* (Silvia LIVERANI et al. 2015), nous avons introduit trois mouvements de changement de label de groupe afin d'essayer d'éviter la convergence de l'algorithme vers un mode local et de réduire l'autocorrélation intra-chaîne (PAPASPILIOPOULOS et ROBERTS 2008 ; D. I. HASTIE, Silvia LIVERANI et Sylvia RICHARDSON 2015). L'utilisation de ces trois mouvements de changement de label de groupe est justifiée par la difficulté d'identifier les labels des groupes, ce qui conduit à plusieurs modes des distribution *a posteriori* des ϕ_c. Pour explorer l'espace des distributions *a posteriori* multimodales, Papaspiliopoulos et Roberts (PAPASPILIOPOULOS et ROBERTS 2008) introduisent deux mouvements de changement de labels de groupe qui permettent notamment de faire des changements au début de l'algorithme MCMC. Pour améliorer l'efficacité des changements de labels de groupes, Hastie *et al.* (D. I. HASTIE, Silvia LIVERANI et Sylvia RICHARDSON 2015) suggère d'ajouter un troisième mouvement de changement de label de groupe. L'idée de base des changements de label est d'inverser deux labels j et k selon une probabilité $\min(1,r_{jk})$. Les détails sur r_{jk} sont donnés dans la Table 6.3. Les principales caractéristiques de ces trois changements sont les suivantes. Le premier changement a une probabilité élevée d'accepter l'échange de j et k lorsque les poids ϕ_j et ϕ_k sont proches. D'autre part, deux groupes ayant un nombre similaire de mineurs d'uranium sont rarement échangés. Les deux autres changements proposent uniquement l'échange de labels de groupes voisins, à savoir j et $j+1$. Lorsque le changement de label est accepté selon le deuxième ou le troisième mouvement, les composantes V correspondantes impliquées dans la construction «stick-breaking» sont simultanément modifiées (et par conséquent les poids ϕ).

Le deuxième mouvement de changement de label de groupe à une probabilité d'acceptation élevée pour des groupes voisins comprenant un nombre différent de

mineurs.

Pour le troisième changement de labels de groupes, les composantes V correspondantes sont modifiées de sorte que les poids correspondants ϕ_j et ϕ_{j+1} sont proches de leurs espérances sachant les nouveaux labels. Les détails concernant r et V sont donnés dans la Table 6.3. Pour les trois procédures de changements de label, les excès de risque β correspondants et les autres paramètres spécifiques aux groupes sont simplement échangés lorsque le déplacement est accepté.

	Mouvement 1	Mouvement 2	Mouvement 3
r_{jk}	$(\frac{\phi_j}{\phi_k})^{n_k-n_j}$	$\begin{cases} \frac{(1-V_{j+1})^{n_j}}{(1-V_j)^{n_{j+1}}} & k=j+1 \\ 0 & \text{sinon} \end{cases}$	$\begin{cases} (\frac{\phi^+}{\phi_{c+1}R_1+\phi_c R_2})^{n_j+n_{j+1}} R_1^{n_{j+1}} R_2^{n_j} & k=j+1 \\ 0 & \text{sinon} \end{cases}$
V'_l	V_l	$\begin{cases} V_{j+1} & l=j \\ V_j & l=j+1 \\ V_l & \text{sinon} \end{cases}$	$\begin{cases} \frac{\phi'_j}{\prod_{k<j}(1-V_k)} & l=j \\ \frac{\phi'_{j+1}}{(1-V'_j)\prod_{k<j}(1-V_k)} & l=j+1 \\ V_l & \text{sinon} \end{cases}$

TABLE 6.3 – Changement de label de groupe. L'échange entre les labels j et k est acceptée avec la probabilité $\min(1, r_{jk})$. Si le changement d'étiquette est accepté, V' est la nouvelle valeur de la composante bêta V. n_j est le nombre de mineurs dans le groupe j. $\phi^+ = \phi_j + \phi_k$, $\phi' = \phi_{j+1}\frac{\mathbb{E}(\phi_j|C',\alpha)}{\mathbb{E}(\phi_{j+1}|C,\alpha)} + \phi_j\frac{\mathbb{E}(\phi_{j+1}|C',\alpha)}{\mathbb{E}(\phi_j|C,\alpha)}$, $R_1 = \frac{1+\alpha+n_{j+1}+\sum_{l>j+1}n_l}{\alpha+n_{j+1}+\sum_{l>j+1}n_l}$ et $R_2 = \frac{\alpha+n_j+\sum_{l>j+1}n_l}{1+\alpha+n_j+\sum_{l>j+1}n_l}$

6.4 Traitements *a posteriori*

Comme décrit dans Molitor *et al.* (MOLITOR et al. 2010) et dans Liverani *et al.* (Silvia LIVERANI et al. 2015), le post-traitement est réalisé après l'exécution de l'algorithme MCMC. L'objectif est de déterminer une seule partition optimale des mineurs d'uranium qui résume l'échantillon de la chaîne MCMC. Il existe différentes techniques pour obtenir cette partition optimale. Nous choisissons d'utiliser une approche de post-traitement basée sur la matrice de similarité *a posteriori*. Si K est le nombre d'itérations, K matrices binaires carrées S_k de dimension $n \times n$ sont calculées à chaque itération k où $S_k(i,j) = 1$ si les mineurs i et j appartiennent

au même groupe à l'itération k de l'échantillon MCMC, et 0 sinon. La moyenne S de ces K matrices $(S_1, ..., S_K)$ contient donc la probabilité que deux mineurs appartiennent au même groupe dans l'échantillon MCMC. La meilleure partition estimée, appelée $\boldsymbol{C}^{best}$, est celle obtenue à l'itération k qui minimise la distance des moindres carrées entre la matrice S_k et la matrice S. $\boldsymbol{C}^{best}$ est un vecteur tel que C_i^{best} est le label du groupe du mineur i dans cette partition optimale.

Les distributions *a posteriori* des paramètres sont obtenues conditionnellement à la meilleure partition $\boldsymbol{C}^{best}$. Si θ_c désigne un paramètre dépendant du groupe c, un échantillon de la distribution *a posteriori* du paramètre θ_c conditionnellement à la partition $\boldsymbol{C}^{best}$ est $\{\bar{\theta}_{c,k}, k = 1, ..., K\}$ tel que

$$\bar{\theta}_{c,k} = \frac{1}{n_c} \sum_{i:C_i^{best}=c} \theta_{C_i^k,k} \tag{6.6}$$

avec n_c le nombre de mineurs d'uranium dans le groupe c et C_i^k le groupe du mineur i à l'itération k. Cette procédure de post-traitement s'applique à tous les paramètres impliqués dans les trois sous-modèles pour tous les labels de groupe, à savoir $(\beta, \mu^R, \sigma^R, \mu^G, \sigma^G, \mu^P, \sigma^P, \mu^A, \sigma^A, p^J, p^M, p^T)$ ainsi que les poids ϕ des groupes.

CHAPITRE 7

Résultats

7.1 Impact des incertitudes de mesure sur la dose de rayonnements γ dans l'estimation du risque de décès par cancer du poumon dans la sous-cohorte post-55 des mineurs d'uranium français

Cette partie présente les résultats obtenus après l'inférence du modèle décrit dans la section 5.2. Après avoir comparé les résultats des différents sous-modèles dans la section 7.1.1, c'est le sous-modèle $\mathcal{M}_3$ qui est sélectionné. Les résultats complets de l'inférence avec ce sous-modèle d'exposition sont présentés dans la section 7.1.2. Les valeurs de seuils de détection et de variances d'erreurs de mesure ayant été fixées dans le modèle inféré, différentes valeurs ont été testées afin de vérifier la sensibilité du modèle à ces valeurs dans les sections 7.1.3 et 7.1.4. Finalement, l'impact des différentes sources d'incertitudes a été mesuré et présenté dans la section 7.1.5.

7.1.1 Sensibilité au sous-modèle d'exposition

Comme décrit précédemment, trois sous-modèles d'exposition différents ont été présentés pour décrire l'incertitude des variables latentes des vraies doses de rayonnements γ dans la sous-cohorte post-55 des mineurs d'uranium français. Les médianes et les intervalles de crédibilité à 95% (IC à 95%) *a posteriori* de l'excès de risque instantané (EHR) pour 100 mSv ainsi que les valeurs des critères DIC et WAIC pour les différents sous-modèles d'exposition sont indiquées dans la Table 7.1.

Sous-modèle d'exposition	EHR par 100 mSv	WAIC	DIC
$\mathcal{M}_1$	0.82 [0.29 ; 1.71]	2462.0	2461.8
$\mathcal{M}_2$	0.81 [0.30 ; 1.75]	2461.9	2461.7
$\mathcal{M}_3$	0.81 [0.28 ; 1.75]	2461.7	2461.5

TABLE 7.1 – Médianes et intervalles de crédibilité (IC) à 95% *a posteriori* de l'excès de risque (EHR) (pour 100 mSv) de décès par cancer du poumon dans la sous-cohorte post-55 des mineurs d'uranium français, en supposant trois sous-modèles d'exposition différents. Watanabe-Akaike Information Criterion (WAIC) et Deviance Information Criterion (DIC) pour les trois sous-modèles d'exposition

Il y a peu de différences entre les estimations de risque, les intervalles de crédibilité à 95% et les critères DIC et WAIC pour les trois sous-modèles d'exposition. Cela montre que les estimations du risque sont robustes au choix du sous-modèle d'exposition. Étant donné que le sous-modèle $\mathcal{M}_3$ fournit plus d'informations que les autres sous-modèles sur l'évolution au cours du temps des vraies doses de rayonnements γ log-transformées en fonction du type de mine et le plus petit écart-type géométrique estimé σ_x des variables latentes des vraies doses de rayonnements γ, l'accent sera mis sur ce sous-modèle par la suite.

7.1.2 Résultats avec le sous-modèle d'exposition $\mathcal{M}_3$

La Table 7.2 donne les médianes et les intervalles de crédibilité à 95% *a posteriori* de l'EHR (pour 100 mSv) de la mortalité par cancer du poumon, du risque instantané de base $\boldsymbol{\lambda}$ et des paramètres du modèle hiérarchique complet basé sur le sous-modèle d'exposition $\mathcal{M}_3$ (voir Figure 5.1).

Paramètre	Médiane	IC 95%
EHR par 100 mSv	0.81	[0.28 ; 1.75]
λ_1 (10^{-6})	0.05	[0.03 ; 0.07]
λ_2 (10^{-6})	0.78	[0.45 ; 1.27]
λ_3 (10^{-6})	4.24	[2.72 ; 6.22]
λ_4 (10^{-6})	7.33	[4.18 ; 11.88]
a		
Mine de jour (J)	-0.025	[-0.034 ; -0.016]
Mine de fond (F)	-0.047	[-0.054 ; -0.039]
b		
Mine de jour (J)	-1.59	[-1.81 ; -1.37]
Mine de fond (F)	-0.05	[-0.26 ; 0.16]
p		
Mine de fond (F)	0.67	[0.66 ; 0.68]
σ_μ	0.33	[0.27 ; 0.45]
σ_x	0.93	[0.90 ; 0.96]

TABLE 7.2 – Médianes et intervalles crédibles (IC) à 95% *a posteriori* des paramètres du modèle hiérarchique complet combinant le sous-modèle de maladie, le sous-modèle de mesure et le sous-modèle d'exposition $\mathcal{M}_3$. EHR - excès de risque instantané.

Le coefficient de risque de décès par cancer du poumon corrigé associé aux doses cumulées de rayonnements γ dans la sous-cohorte post-55 des mineurs d'uranium français a été estimé à 0.81 par 100 mSv (IC à 95% : [0.28 ; 1.75]). Par conséquent, après avoir pris en compte l'incertitude sur les doses de rayonnements γ, il existe toujours une association positive statistiquement significative entre les doses de rayonnements γ et le risque de décès par cancer du poumon dans la sous-cohorte post-55 des mineurs d'uranium français. Les paramètres de pente a_J et a_F définissant la tendance temporelle des doses de rayonnements γ log-transformées sont significativement négatifs pour les deux types de mines. Cela confirme que la valeur de l'espérance de la vraie dose log-transformée de rayonnements γ a diminué avec le temps dans les mines d'uranium françaises. Cette tendance à la baisse est plus forte dans les mines souterraines que dans les mines à ciel ouvert, comme le montre la Figure 7.1. Comme prévu, le paramètre d'ordonnée à l'origine b_F, qui indique l'espérance de la vraie dose de rayonnements γ log-transformée en 1956, est plus élevé pour les mines souterraines que le paramètre d'ordonnée à l'origine b_J pour les mines à ciel ouvert. Il est intéressant de noter que la Figure 7.1 montre également qu'à partir de 1996, il n'y a plus de différence significative entre les moyennes des vraies doses annuelles de rayonnements γ dans les mines souterraines et dans les mines à ciel ouvert. Cela ne signifie pas que les niveaux de dose de rayonnements γ étaient les mêmes dans les mines souterraines et à ciel ouvert, mais plutôt que le nombre de mineurs exposés était trop faible à partir de 1996 pour mettre en évidence une différence significative, si elle existe. En effet, les mines d'uranium françaises ont fermé peu après cette date.

7.1.3 Sensibilité aux valeurs de limite de détection

Étant donné que les valeurs fixées pour la LD des dosimètres peuvent être incertaines, la robustesse de l'EHR estimé à ces valeurs a été testée. Pour cela, le modèle hiérarchique bayésien incluant le sous-modèle d'exposition $\mathcal{M}_3$ a été ajusté après avoir considéré deux scénarios alternatifs. Premièrement, les LD définies précédemment ont été augmentées de 50% (3.3 mSv avant 1986 et 0.825 mSv après 1986). Pour ce faire, toute valeur de dose entre la LD initiale et la LD augmentée a été supposée censurée à gauche. En d'autres termes, $Z_i(t)$ - qui a été

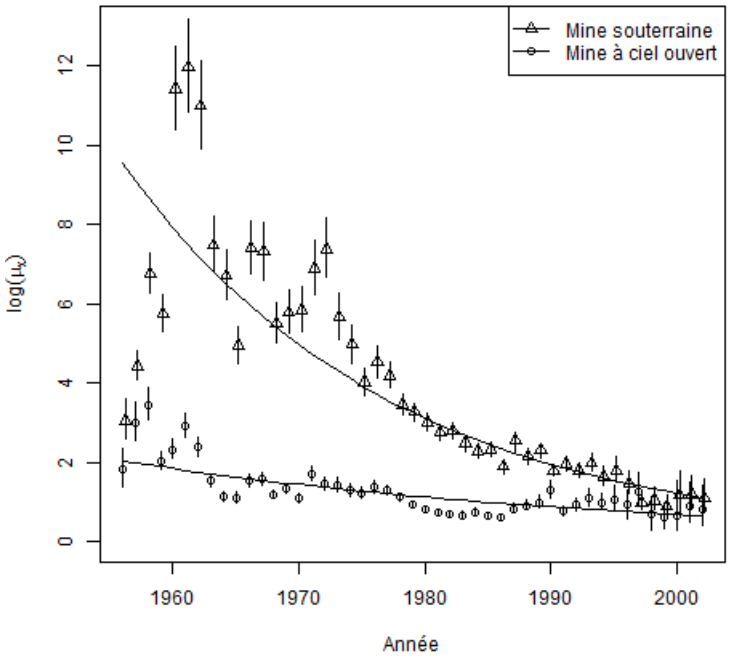

FIGURE 7.1 – Doses de rayonnements γ réelles log-transformés attendues dans les mines souterraines (c'est-à-dire $log(\mu_{x,F}(t))$) et les mines à ciel ouvert (c'est-à-dire $log(\mu_{x,J}(t))$) au fil du temps, dans la sous-cohorte post-55 des mineurs d'uranium français. Les cercles et les triangles représentent les médianes postérieures de $log(\mu_{x,J}(t))$ et de $log(\mu_{x,F}(t))$, respectivement, et les segments indiquent les intervalles crédibles à 95%. Les lignes pleines montrent $log(a_J \cdot f(t) + b_J)$ et $log(a_F \cdot f(t) + b_F)$ qui sont estimés à partir des médianes postérieures de a_J, a_F, b_J et b_F.

observé avec la LD initiale - a été supposé inconnu mais plus petit que la $LD(t)$, et a également été supposé suivre la distribution lognormale donnée par l'équation 5.23. La médiane *a posteriori* de l'EHR a ensuite été estimée à 0.80 pour 100 mSv avec un intervalle de crédibilité à 95% de [0.28 ; 1.73]. Finalement, les LD définies précédemment ont été augmentées de 75% (3.85 mSv avant 1986 et 0.9625 mSv après 1986). La médiane *a posteriori* de l'EHR a alors été estimée à 0.83 pour 100 mSv avec un intervalle de crédibilité à 95% de [0.29 ; 1.75]. Par conséquent, on observe que l'augmentation de la valeur de la LD ne modifie pas sensiblement les estimations du risque.

7.1.4 Sensibilité aux valeurs de variance d'erreurs de mesure

Afin de tester la robustesse des résultats par rapport aux écarts-types géométriques fixés pour les composantes des erreurs de mesure lognormale (voir la section 5.2.2 sur le sous-modèle de mesure), le modèle hiérarchique bayésien complet avec le sous-modèle d'exposition $\mathcal{M}_3$, a également été ajusté après avoir augmenté les écarts-types correspondants de 50% et 100%. Dans l'hypothèse d'une augmentation de 50%, l'EHR pour 100 mSv reste positif et statistiquement significatif avec une médiane *a posteriori* égale à 0.86 pour 100 mSv et un intervalle de crédibilité à 95% de [0.33 ; 1.84]. Dans l'hypothèse d'une augmentation de 100%, la médiane *a posteriori* estimée de l'EHR était de 0.90 pour 100 mSv avec un intervalle de crédibilité à 95% de [0.34 ; 1.89]. Comme prévu, les résultats sont sensibles à l'ampleur de l'erreur de mesure. Ils montrent également que, dans les cas où ce paramètre est sous-estimé (resp. surestimé), le risque de décès par cancer du poumon peut également être sous-estimé (resp. surestimé), de même que son incertitude d'estimation.

7.1.5 Impact des différentes sources d'incertitude sur l'estimation du risque

Enfin, le sous-modèle de maladie a été ajusté sans tenir compte de l'erreur de mesure de l'exposition et après avoir remplacé toutes les valeurs de dose de rayonnements γ censurées à gauche et manquantes par zéro. L'EHR non corrigé de

décès par cancer du poumon dû à l'exposition professionnelle aux rayonnements γ a ensuite été estimé à 0.78 par 100 mSv avec un intervalle de crédibilité à 95% [0.28 ; 1.64]. Ce résultat est similaire à l'excès de risque relatif non corrigé estimé à partir d'une régression de Poisson qui était de 0.74 pour 100 mSv avec un intervalle de crédibilité à 95% de [0.23 ; 1.73] (Rage et al. 2015). Étant donné que l'estimation corrigée de l'EHR est de 0.81 pour 100 mSv avec un intervalle de crédibilité à 95% [0.28 ; 1.75], aucune modification substantielle de l'EHR estimé pour 100 mSv n'est observée en tenant compte de l'incertitude de la dose. L'intervalle de crédibilité à 95% associé est plus large lorsqu'on tient compte de l'incertitude de la dose, mais il n'inclut pas zéro, ce qui signifie que l'association positive entre la mortalité par cancer du poumon et la dose de rayonnements γ reste statistiquement significative.

Afin de distinguer l'impact des erreurs de mesure classiques et des valeurs de dose censurées à gauche ou manquantes sur l'estimation du risque, un modèle hiérarchique a été ajusté en tenant compte de l'erreur de mesure avec le sous-modèle d'exposition $\mathcal{M}_3$ mais après avoir remplacé toutes les données d'exposition censurées à gauche et manquantes par zéro. En d'autres termes, les erreurs de mesure classiques ont été prises en compte, mais l'existence de données d'exposition censurées à gauche ou manquantes a été négligée. L'EHR du décès par cancer du poumon dû à une exposition professionnelle aux rayonnements γ a alors été estimé à 0.80 par 100 mSv (IC à 95% : [0.29 ; 1.70]). Ainsi, si l'on tient compte uniquement de l'erreur de mesure classique, on n'observe là encore aucune modification substantielle de l'EHR estimé pour 100 mSv. Enfin, un modèle hiérarchique a été ajusté en tenant compte de l'existence de valeurs de dose censurées à gauche et manquantes, mais en négligeant l'existence des erreurs de mesure classiques. L'EHR du décès par cancer du poumon dû à une exposition professionnelle aux rayonnements γ a ensuite été estimé à 0.80 pour 100 mSv (IC à 95% : [0.28 ; 1.75]). Là encore, si l'on tient compte uniquement des données d'exposition censurées à gauche et manquantes, aucune modification substantielle de l'EHR pour 100 mSv n'a été observée. Néanmoins, l'intervalle de crédibilité à 95% associé est légèrement plus large lorsque cette source d'incertitude est prise en compte par le modèle, par rapport à l'intervalle de crédibilité à 95% obtenu en ne tenant compte que de l'erreur de mesure classique. Dans tous les cas, l'intervalle de crédibilité à 95% de l'EHR pour 100 mSv n'incluait pas 0 et, par conséquent, l'association positive

entre la mortalité par cancer du poumon et la dose de rayonnements γ est restée statistiquement significative. La Figure 7.2 présente les densités *a posteriori* du coefficient de risque (c'est-à-dire β) pour ces différents modèles.

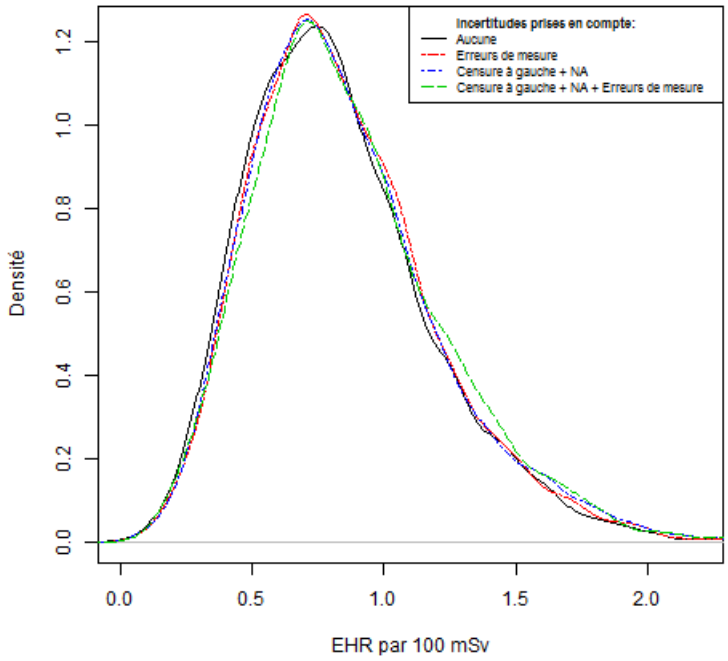

FIGURE 7.2 – Densités *a posteriori* de l'EHR pour 100 mSv de décès par cancer du poumon dû à une exposition professionnelle aux rayonnements γ dans la sous-cohorte post-55 des mineurs d'uranium français, en fonction des sources d'incertitude de l'exposition qui ont été prises en compte

7.2 Prise en compte de la multi-exposition dans l'estimation du risque de décès par cancer du poumon radio-induit dans la sous-cohorte post-55 des mineurs d'uranium

Les résultats du modèle de régression bayésienne sur profils d'exposition sont présentés dans cette partie avec deux cas. Le premier cas correspond au modèle complet PRM tel qu'il a été présenté dans la section 6.2 où les différents sous-modèles permettent de relier le risque de décès par cancer du poumon à des groupes ou profils de mineurs d'uranium français. La construction d'un groupe sous ce modèle est fondée non seulement sur des caractéristiques d'exposition similaires mais également sur le partage d'un même risque sanitaire. Dans ce premier cas, le nombre de groupes est inconnu et donc également estimé comme l'ensemble des autres paramètres du modèle. Comme décrit dans la section 7.2.1.1, des problèmes de convergence vers des modes locaux nous ont contraints à envisager le second cas où le nombre de groupes est cette fois-ci fixé. Après une discussion sur le choix du nombre de groupes, les résultats sous ce modèle restreint sont présentés dans la section 7.2.1.2. Enfin, une étude des performances de ce modèle restreint est présentée dans la section 7.2.2 avec un regard particulier sur l'impact d'une mauvaise spécification de ce nombre.

7.2.1 Application du modèle de régression bayésienne sur profils d'exposition

7.2.1.1 Modèle avec nombre de groupes de mineurs d'uranium inconnu

Le modèle PRM tel que défini dans la section 6.2 est mis en oeuvre sur la sous-cohorte post-55 des mineurs d'uranium français.

Comme déjà mentionné dans Liverani *et al.* (Silvia Liverani et al. 2015), le paramètre α de l'équation 6.3 est directement lié au nombre de groupes non-vides. Dans le modèle PRM, ce nombre est également estimé (seul le nombre maximum de groupes C_{max} est fixé) et une attention particulière doit être accordée à la ques-

tion de la convergence locale même si des mouvements de changement de label de groupe ont été ajoutés à l'algorithme MCMC. Pour évaluer la convergence vers un mode local, les échantillonneurs MCMC ont été exécutés à partir de différentes valeurs initiales de α. Les valeurs initiales sont choisies entre 0.5 et 9.5 avec un pas de 1, ce qui couvre le support de la distribution *a priori* de α. Pour une valeur initiale donnée, le nombre de groupes non-vides converge systématiquement vers une valeur unique sans déplacement au cours de l'échantillonneur, alors qu'il n'y a pas de problème de convergence pour les autres paramètres. Les résultats sont présentés dans la Figure 7.3 où le nombre de groupes non-vides prend quatre valeurs possibles de 5 à 8 (y compris la groupe des mineurs d'uranium non exposés) selon les différentes valeurs initiales de α. Un problème de convergence locale est également suspecté malgré les trois procédures de changement de labels de groupe. Une raison possible est la faible proportion de cas de décès par cancer du poumon. Cette proportion est en effet proche de 3%, ce qui donne un signal faible pour déduire le risque entre les groupes et le cancer du poumon. Par conséquent, un modèle RPRM de mélange de régression sur profil restreint est inféré, dans lequel le nombre K de groupes non-vides est fixé. Le sous-modèle d'attribution défini section 6.2 est simplifié, les poids ϕ ont maintenant un nombre fixe K de composantes. Nous avons lancé l'algorithme MCMC avec deux ensembles de valeurs initiales pour s'assurer de la convergence de ce modèle.

Une solution pour choisir K est de choisir une valeur parmi les quatre valeurs suggérées par la Figure 7.3. Les critères d'information de déviation (DIC) (Spiegelhalter et al. 2002) ainsi que le Watanabe-Akaike, également appelé Widely Applicable, Information Criterion (WAIC) (Watanabe 2010) sont présentés dans la Table 7.3 pour un nombre de groupes K entre 5 et 8. Ces deux critères sont concordants en faveur de 8 groupes non-vides. Comme la déviance pénalisée est connue pour sélectionner les modèles les plus complexes, nous préférons présenter les résultats avec K égal à 8 groupes non-vides mais aussi comparer avec les trois autres modèles RPRM correspondant à 5, 6 et 7 groupes non-vides (résultats donnés en annexe A). À noter que lorsque le nombre de groupes non-vides est fixé, aucun problème de convergence n'a été trouvé pour tous les autres paramètres.

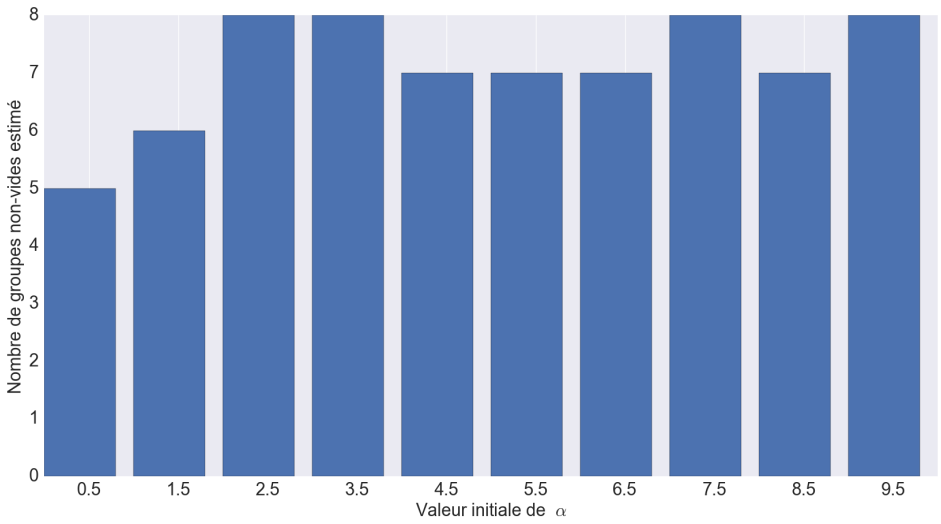

FIGURE 7.3 – Nombre de groupes non-vides estimé selon la valeur initiale de α

Nombre de groupes non-vides K	DIC	WAIC
5	146345	110872
6	136714	108773
7	118602	107004
8	104566	105704

TABLE 7.3 – DIC et WAIC du modèle bayesien PRM selon K, le nombre de groupes non-vides fixé

7.2.1.2 Modèle avec nombre fixé de groupes de mineurs d'uranium

Les résultats pour le modèle à 8 groupes sont résumés sur la Figure 7.4 et la Figure 7.5 tandis que les résultats pour les modèles avec 5 à 7 groupes peuvent être trouvés en annexe A.

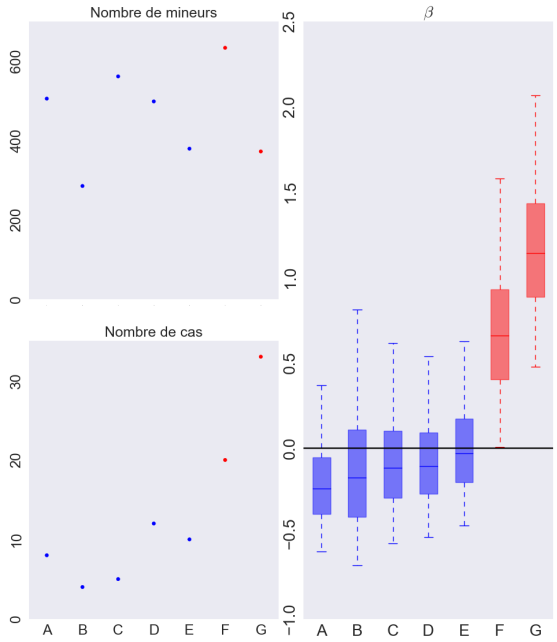

FIGURE 7.4 – Nombre de mineurs d'uranium français (en haut à gauche), nombre de décès par cancer du poumon (en bas à gauche) et excès de risque instantané de décès par cancer du poumon β (à droite) dans chaque groupe, en ajustant un modèle bayésien RPRM supposant 8 groupes non-vides dans la cohorte française des mineurs d'uranium. Le groupe des mineurs non exposés n'est pas affichée. Les boîtes représentent les trois quartiles (1^{er} quartile, médiane et $3^{\text{ème}}$ quartile) de la distribution *a posteriori* de β et les moustaches des boxplots montrent l'intervalle de crédibilité à 95% de la distribution *a posteriori* pour chaque groupe.

À gauche de la Figure 7.4, le nombre de mineurs (en haut) et le nombre de cas (en bas) par groupe sont représentés, sauf pour le groupe des mineurs d'uranium non exposés. Les sept groupes résultants sont désignées par les lettres A à G. L'ordre de représentation des groupes suit l'ordre croissant du risque estimé associé à chaque groupe. Ainsi, le groupe A correspond à la médiane *a posteriori* estimée la plus faible de β et le groupe G à la médiane la plus élevée. Le nombre

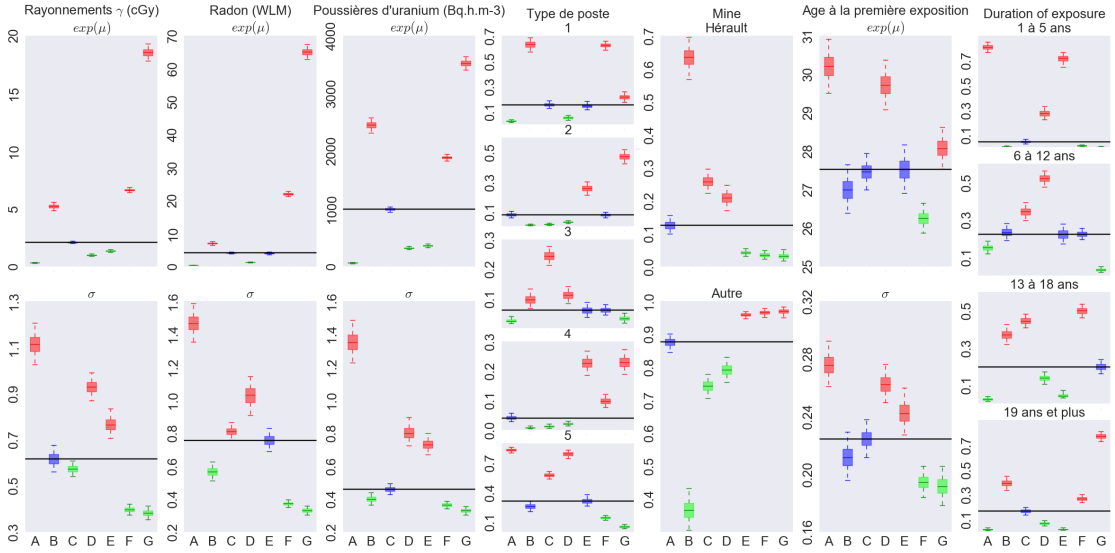

FIGURE 7.5 – Caractérisation des profils d'exposition associés à chaque groupe, dans le cadre d'un modèle bayésien RPRM supposant 8 groupes non-vides. Le groupe des mineurs d'uranium non exposés n'est pas affiché.
Types de poste :
1) foreur avant la mécanisation ;
2) foreur après la mécanisation ;
3) autres travaux souterrains avant la mécanisation ;
4) autres travaux souterrains après la mécanisation ;
5) travaux de surface.

de mineurs varie de 285 à 633 et le nombre de cas de 4 à 30 par groupe. À droite de la Figure 7.4, on trouve les résultats sur l'excès de risque de décès par cancer du poumon de chaque groupe (β_A à β_G). Les boîtes correspondent aux quartiles *a posteriori* de β et les moustaches s'étendent aux quantiles *a posteriori* à 2.5% et à 97.5% illustrant un intervalle de crédibilité de β à 95%. Les couleurs indiquent si l'intervalle de crédibilité *a posteriori* à 95% de β est supérieur à zéro (rouge) ou inclut zéro (bleu). Un groupe est appelé «groupe à haut risque significatif» (ou respectivement «groupe à faible risque significatif») si les moustaches sont supérieures à 0 (respectivement inférieures à 0). Deux groupes à haut risque significatif sont ici identifiés, à savoir les groupes F et G. L'excès de risque *a posteriori* médian du groupe G est estimé à 1.14 et à 0.66 pour le groupe F. Notez qu'un excès de risque de 1.14 signifie que les mineurs appartenant à ce groupe ont un risque multiplié par 2.14 par rapport aux mineurs d'uranium non exposés.

La caractérisation de chaque groupe en terme de covariables est illustrée sur la Figure 7.5. Chaque colonne correspond à une covariable, les labels des groupes sont précisées sur l'axe horizontal. Pour les covariables continues, c'est-à-dire les expositions cumulées et l'âge à la première exposition, les résultats sur les médianes (e^{μ}) se trouvent en haut tandis que les résultats sur l'écart type (à l'échelle logarithmique) se trouvent en bas. Pour les covariables catégorielles, soit le type de poste, la mine et la durée d'exposition, la distribution *a posteriori* de la probabilité de chaque catégorie est indiquée. Les boîtes et les moustaches sont définies comme précédemment. Les deux couleurs différentes, le vert et le rouge, correspondent à un intervalle de crédibilité *a posteriori* à 95% respectivement en-dessous ou au-dessus de la médiane globale sur toutes les médianes *a posteriori* des groupes alors que la couleur bleue ne montre aucune valeur particulière de la covariable pour ce groupe.

Le groupe G présentant le plus grand risque de décès par cancer du poumon correspond aux mineurs d'uranium les plus exposés, car les intervalles de crédibilité de la moyenne d'exposition cumulée au radon, aux rayonnements γ et aux poussières d'uranium sont élevés. Ils travaillaient principalement avant la mécanisation ou comme foreur après la mécanisation, pas dans la mine de l'Hérault, assez âgés lorsqu'ils ont commencé à travailler par rapport aux autres groupes et était exposés pendant une longue période (plus de 19 ans). Ce groupe correspond aux

conditions de travail les plus difficiles. Ce groupe à haut risque se retrouve pour les modèles à 5, 6 ou 7 groupes non-vides (voir l'annexe A). Son identification systématique est rassurante en termes de validité du modèle car elle est conforme aux hypothèses standards dans le domaine.

Le groupe F associé au deuxième risque le plus élevé de décès par cancer du poumon est caractérisé par des mineurs d'uranium qui étaient également très exposés mais moins que dans le groupe G, qui travaillaient comme foreur après la mécanisation ou un autre travail souterrain avant la mécanisation, qui ne travaillaient pas dans la mine de l'Hérault, qui étaient jeunes lorsqu'ils ont commencé à travailler par rapport aux autres groupes et qui étaient exposés depuis plus de 13 ans. Les conditions de travail de ce deuxième groupe peuvent également être considérées comme difficiles mais moins que celles du groupe G, notamment en ce qui concerne le forage avant ou après la mécanisation et la durée d'exposition un peu plus courte. D'autre part, ce deuxième groupe met en évidence le profil de risque des mineurs qui ont commencé à travailler tôt par rapport aux autres groupes.

Les résultats concernant ce deuxième groupe diffèrent légèrement en fonction du nombre de groupes non-vides fixé (voir annexe A). En effet, ce groupe est associé à un excès de risque positif qui est significatif pour le RPRM avec $K = 7$, presque significatif avec $K = 6$ mais non significatif avec $K = 5$. Les médianes *a posteriori* de β_F et β_G ainsi que les caractéristiques de ces deux groupes sont très similaires avec $K = 6$ et $K = 7$ à celles déjà trouvées avec $K = 8$. En ce qui concerne le modèle RPRM avec 5 groupes non-vides, les résultats sur le groupe G sont similaires tandis que la médiane *a posteriori* de l'excès de risque β_F et les caractéristiques du groupe F sont différents. En effet, ce deuxième groupe F ne contient pas exactement le même nombre de mineurs d'uranium selon les valeurs de K. Près de 630 mineurs appartiennent au deuxième groupe pour tout nombre fixe K, sauf pour $K = 5$ où il y a environ 250 mineurs de plus (groupe F dans les Figures A.1 et A.2). En comparant les 630 mineurs à ces 250 mineurs supplémentaires, on constate que les mineurs communs ont reçu une exposition cumulée au radon plus élevée et qu'ils travaillaient tous dans d'autres mines que celle de l'Hérault. Les 250 mineurs qui diffèrent avec K=5, ont une exposition cumulée au radon plus faible et un peu plus de la moitié d'entre eux travaillaient

dans la mine de l'Hérault. Enfin, il n'y a que 2 cas de décès par cancer du poumon parmi ces 250 mineurs. Le risque associé aux 630 mineurs communs est également plus élevé que celui associé aux 880 mineurs appartenant au second groupe avec la partition en 5 groupes non-vides. Par conséquent, ce deuxième groupe F est à nouveau significatif ou presque significatif avec des partitions en 6 ou 7 groupes mais pas avec la partition en 5 groupes non-vides. La médiane *a posteriori* de β_F estimée est proche de la même valeur pour 6 et 7 groupes non-vides que pour 8 groupes mais est environ de 0.3 pour le modèle à 5 groupes non-vides. Malgré ces différences, ce deuxième groupe à haut risque existe pour tous les modèles avec des caractéristiques très proches, en particulier avec des expositions cumulées moins importantes au radon, aux rayonnements γ et à la poussière d'uranium mais avec un jeune âge au début du travail.

Nous n'observons pas systématiquement un risque croissant correspondant à des niveaux d'exposition croissants. C'est particulièrement le cas lorsqu'on se concentre sur le groupe B (Figure 7.5). Ce groupe est associé au deuxième risque le plus faible alors que les mineurs de ce groupe sont très exposés. Les principales différences par rapport aux autres groupes sont la proportion importante de mineurs d'uranium travaillant dans la mine de l'Hérault et la période de travail après la mécanisation. La modélisation de l'association entre les profils et la mortalité permet d'obtenir une interprétation plus fine de l'effet des niveaux d'exposition que les études incluant des associations directes avec les expositions n'auraient pu le faire.

7.2.1.3 Analyse du statut tabagique dans les différents groupes

Dans ce travail, on ne tient pas compte de la consommation de tabac des mineurs alors qu'elle est connue comme étant la principale cause de cancer du poumon. En fait, le statut tabagique n'est disponible que pour 4.2% des mineurs de la sous-cohorte post-55 des mineurs d'uranium français. Ce manque crucial d'informations rend très peu fiable l'ajustement sur le statut tabagique lors de l'estimation du risque de décès par cancer du poumon dû à des expositions radiologiques multiples.

Une analyse par validation croisée a néanmoins été réalisée afin d'estimer la

capacité d'un modèle de régression logistique à prédire correctement le statut tabagique d'un mineur (variable binaire : fumeur / non-fumeur) à partir des informations disponibles suivantes au sein de la sous-cohorte post-55 : type de poste, année de naissance du mineur (par décennie), mine (localisation), mine de jour/mine de fond, évènement (décès par cancer du poumon), niveaux d'expositions radiologiques. Seuls 65% des statuts tabagiques ont été correctement retrouvés, ce qui est faible. En particulier, la probabilité de prédire un statut fumeur sachant que le mineur est non-fumeur est de 67%, ce qui est très élevé. Cette analyse confirme qu'à l'heure actuelle, il est très peu fiable d'imputer près de 96% des statuts tabagiques étant donné qu'aucun prédicteur potentiel du statut de fumeur n'est disponible dans la sous-cohorte post-55.

En fait, si la consommation de tabac est principalement associée à l'excès de risque de décès par cancer du poumon estimé dans la sous-cohorte post-55, une proportion plus élevée de fumeurs devrait être a minima observée dans les groupes présentant un excès de risque élevé par rapport à ceux présentant un excès de risque faible (et réciproquement). Compte tenu des données de statut tabagique disponibles, cela ne semble pas être le cas. Les rapports entre le nombre de fumeurs et le nombre de non-fumeurs pour les groupes A, B ,C, D, E, F et G (définis dans la Figure 7.4) sont respectivement de 12/3, 7/0, 14/5, 17/4, 5/5, 16/8 et 34/12, où les groupes F et G sont les groupes avec les excès de risque de décès par cancer du poumon les plus élevés. De plus, les proportions de fumeurs associées pour les groupes A, B ,C, D, E, F et G sont respectivement de 0.8, 1.0, 0.74, 0.81, 0.50, 0.67 et 0.74. Bien entendu, ces ratios estimés doivent être interprétés avec prudence étant donné le peu de données disponibles (c'est-à-dire seulement 142 mineurs avec statut tabagique connu parmi les 3 377 de la sous-cohorte post-55).

7.2.2 Performances du modèle de régression bayésienne sur profils d'exposition à nombre de groupes fixé : Étude de simulation

Dans ce travail, les résultats ont tous été présentés sous des modèles à nombre de groupes fixé dû au problème de convergence. Ce choix du nombre de groupes pourrait avoir un impact sur la qualité des différentes estimations. Cette partie a

pour objectif d'étudier la qualité des estimations dans le cas d'un bon nombre de groupes fixé et de quantifier l'impact d'une mauvaise spécification du nombre de groupes.

7.2.2.1 Présentation du protocole de simulations

Afin d'être proche du cas d'étude réel présenté dans les parties précédentes, les valeurs des paramètres utilisées pour simuler les jeux de données sont inspirées des valeurs de la sous-cohorte post-55 des mineurs d'uranium et des résultats obtenus sous le modèle RPRM à 8 groupes non-vides (dont un groupe de mineurs non exposés). Rappelons que la sous-cohorte post-55 est composée de 3 377 mineurs d'uranium dont 144 n'ayant reçu aucune exposition. Le modèle RPRM à 8 groupes non-vides a identifié deux groupes à haut risque (groupes G et F) où les intervalles de crédibilité à 95% des excès de risque β_G et β_F excluaient la valeur 0 par valeurs supérieures. Les variables caractérisant ces deux groupes étaient, essentiellement, les expositions aux trois sources de rayonnements ionisants et l'âge à la première exposition. Pour les cinq autres groupes exposés (groupes A à E), les intervalles de crédibilité à 95% des excès de risque incluaient la valeur 0.

Afin de réduire le temps algorithmique nécessaire à l'inférence des paramètres et ainsi de permettre la mise en place de plusieurs scénarios, nous avons choisi de réduire le nombre de paramètres tout en gardant les principales caractéristiques des données réelles. Tout d'abord, le nombre de groupes a été réduit à 4 (au lieu de 8), les trois groupes associés aux exposés sont notés A, B et C. Le groupe A est systématiquement associé à un excès de risque nul ($\beta_A = 0$). Il permettra d'étudier la capacité du modèle à ne pas détecter un groupe à risque à tort. Les deux autres groupes d'exposés (B et C) sont associés à des excès de risque strictement positifs avec $\beta_B < \beta_C$. Ainsi, le groupe C sera qualifié de groupe à risque élevé et le groupe B de groupe à risque moyen. Enfin, le nombre de variables caractérisant les groupes a également été réduit à 4 correspondant aux trois sources de rayonnements ionisants et à l'âge à la première exposition.

Trois scénarios de simulations $\mathcal{S}_1$, $\mathcal{S}_2$ et $\mathcal{S}_3$ de 3 377 mineurs sont mis en place avec 4 groupes (144 mineurs non-exposés et 3 233 mineurs exposés répartis équitablement dans les groupes A, B et C) et 4 variables d'exposition. Pour chaque

scénario, les paramètres à fixer sont les deux excès de risque β_B et β_C ainsi que les moyennes et écart-type des quatre variables d'exposition (les trois expositions radiologiques et l'âge à la première exposition) pour chacun des trois groupes des exposés. Les modèles de simulations correspondent aux sous-modèles de maladie et d'exposition présentés dans la section 6.2 à savoir le modèle de survie EHR et les distributions lognormales des expositions continues. Les vecteurs des moyennes (en échelle log) du radon, des rayonnements γ, des poussières d'uranium et de l'âge à la première exposition sont notées, respectivement, μ^R, μ^G, μ^P et μ^A. De la même manière, les écarts-types sont σ^R, σ^G, σ^P et σ^A. L'ensemble de ces paramètres se trouve dans la Table 7.4.

Le premier scénario de simulation $\mathcal{S}_1$ correspond aux excès de risque $\beta_B = 5$ et $\beta_C = 10$. Les moyennes et écarts-types des quatre variables d'exposition correspondent aux valeurs médianes obtenues sur les données réelles sous le modèle RPRM à 8 groupes non-vides. Les valeurs associées au groupe C dit à risque élevé sont celles du groupe identifié à plus haut risque (groupe G sur données réelles) et celles associées au groupe B dit à risque moyen sont celles du second groupe identifié à haut risque (groupe F sur données réelles). Concernant le groupe A de risque nul, les caractéristiques (moyenne et écart-type) des variables d'exposition ont été fixées aux valeurs médianes d'un groupe identifié comme à risque nul (groupe C sur données réelles).

Le deuxième scénario de simulation $\mathcal{S}_2$ correspond à la situation où les excès de risque sont identiques mais où les variables d'exposition radiologique sont moins fluctuantes. On peut imaginer alors que ces variables d'exposition radiologique sont plus discriminantes et ainsi, permettent une meilleure identification des différents groupes. Ainsi, les valeurs des écarts-types σ^R, σ^G et σ^P ont été réduits de 70%.

Finalement, le troisième scénario $\mathcal{S}_3$ correspond à des valeurs d'excès de risque plus faibles. Ainsi, les excès de risque β_B et β_C sont fixés à 2.5 et 5 respectivement. Les moyennes et écarts-types des variables d'exposition sont identiques aux valeurs du scénario $\mathcal{S}_1$. Ce dernier scénario a pour objectif d'appréhender la capacité du modèle à capturer des excès de risque plus faibles et ayant des écarts entre eux plus petits. Ainsi, les groupes sont dans ce sens moins «éloignés» que dans le premier scénario et donc, peut-être, plus difficiles à identifier lors de l'estimation.

Pour chacun des trois scénarios de simulation, 100 jeux de données D_s, avec

$s = 1, ..., 100$, ont été simulés.

	Groupe	β	μ^R	μ^G	μ^P	μ^A	σ^R	σ^G	σ^P	σ^A
	A	0	1.41	0.74	6.9	3.31	0.81	0.57	0.46	0.22
$\mathcal{S}_1$	B	5	3.09	1.89	7.54	3.27	0.37	0.4	0.36	0.19
	C	10	4.18	2.92	8.17	3.33	0.33	0.38	0.33	0.19
	A	0	1.41	0.74	6.9	3.31	0.24	0.17	0.14	0.22
$\mathcal{S}_2$	B	5	3.09	1.89	7.54	3.27	0.11	0.12	0.11	0.19
	C	10	4.18	2.92	8.17	3.33	0.1	0.11	0.1	0.19
	A	0	1.41	0.74	6.9	3.31	0.81	0.57	0.46	0.22
$\mathcal{S}_3$	B	2.5	3.09	1.89	7.54	3.27	0.37	0.4	0.36	0.19
	C	5	4.18	2.92	8.17	3.33	0.33	0.38	0.33	0.19

TABLE 7.4 – Vraies valeurs utilisées pour simuler les 100 jeux de données pour chaque scénario de simulation envisagé

7.2.2.2 Estimations et indicateurs de performance

Pour chaque jeu de données, les estimations de l'ensemble des paramètres du modèle RPRM sont obtenues de la même manière que sur les données réelles (section 7.2.1.2) en fixant le nombre K de groupes non-vides. De manière générale, la médiane *a posteriori* a été choisie comme estimation ponctuelle des paramètres. Deux situations sont envisagées lors des estimations : soit une situation de bonne spécification (en fixant le nombre de groupes non-vides $K = 4$) soit en situation de mauvaise spécification (en fixant le nombre de groupes non-vides $K = 3$ ou $K = 5$). Les résultats de ces deux situations sont présentés dans des parties différentes.

Afin de comparer la qualité des estimateurs obtenus sous les différents scénarios de simulation et selon le nombre K de groupes fixé pour l'estimation, plusieurs indicateurs de performances ont été utilisés. Ces indicateurs visent à renseigner essentiellement deux notions :

- le biais ;
- la bonne et mauvaise classification des individus.

Dans la situation de bonne spécification du nombre de groupes non-vides lors de l'estimation ($K = 4$), le nombre de coefficients β estimés est identique au

vrai nombre de coefficients utilisés pour simuler les jeux de données. Il est donc possible de les comparer directement à l'échelle du groupe. Notamment, les estimations des excès de risque de chaque groupe seront comparées aux vraies valeurs, des intervalles de variabilités (IV) correspondant aux quantiles à 5% et 95% des 100 estimations seront donnés ainsi que les taux de recouvrement fondés sur les intervalles de crédibilité à 95%.

Par contre, en cas de mauvaise spécification de ce nombre de groupes dans le modèle d'estimation ($K \neq 4$), le nombre de coefficients β estimés n'est pas le même que celui des vrais coefficients β. Afin de permettre le calcul des biais et des proportions de bon ou mauvais classement, un changement d'échelle est alors proposé à savoir de descendre au niveau des individus et non des groupes. En effet, pour un jeu de données s, comme chaque individu appartient à un groupe lors de la simulation, chaque individu i a donc une «vraie» valeur d'excès de risque $\beta_{i,s}$ égale à la valeur de l'excès de risque du groupe auquel il appartient, $\beta_{i,s} \in \{\beta_A, \beta_B, \beta_C\}$.

Ainsi, pour chaque jeu de données s, l'estimation $\hat{\beta}_{i,s}$ correspond à la médiane *a posteriori* de l'excès de risque du groupe où l'individu i a été classé lors de l'estimation, $\hat{\beta}_{i,s} \in \left\{\hat{\beta}_{A,s}, \hat{\beta}_{B,s}, \hat{\beta}_{C,s}\right\}$. Trois biais (absolus) sont calculés comme suit :

$$Biais^{abs}_{A,s} = \frac{\sum_{i \in I_{A,s}} \hat{\beta}_{i,s}}{n_{A,s}} \tag{7.1}$$

$$Biais^{abs}_{B,s} = \frac{\sum_{i \in I_{B,s}} (\hat{\beta}_{i,s} - \beta_B)}{n_{B,s}} \tag{7.2}$$

$$Biais^{abs}_{C,s} = \frac{\sum_{i \in I_{C,s}} (\hat{\beta}_{i,s} - \beta_C)}{n_{C,s}} \tag{7.3}$$

avec $I_{A,s}$ (respectivement $I_{B,s}$ et $I_{C,s}$) l'ensemble des $n_{A,s}$ (respectivement $n_{B,s}$ et $n_{C,s}$) individus appartenant au groupe A (respectivement aux groupes B et C) lors de la simulation du jeu de données s.

La valeur de β_A étant nulle, les biais relatifs ne peuvent donc être calculés que pour les groupes B et C.

$$Biais^{rel}_{B,s} = \frac{Biais^{abs}_{B,s}}{\beta_B} \tag{7.4}$$

$$Biais^{rel}_{C,s} = \frac{Biais^{abs}_{C,s}}{\beta_C} \tag{7.5}$$

Des statistiques résumées des 100 valeurs de chaque biais ainsi obtenues sur les 100 jeux de données seront présentées.

Les proportions d'individus bien classés (BC) et mal classés (MC) sont également évaluées à l'échelle individuelle. Pour chaque jeu de donnée s, la proportion $\pi_{BC,s}$ d'individus globalement bien classés (que ce soit classé à raison comme non à risque ou classé à raison comme à risque), la proportion $\pi_{BC-HR,s}$ d'individus correctement identifiés comme à risque le plus élevé, la proportion d'individus faussement classés comme à risque («faux positifs» : $\pi_{MC-FP,s}$) et la proportion d'individus faussement classés comme non à risque («faux négatifs» : $\pi_{MC-FN,s}$) sont calculées.

$$\pi_{BC,s} = \frac{\text{Card}(I_{A,s} \cap \hat{I}_{NR,s}) + \text{Card}((I_{B,s} \cup I_{C,s}) \cap \hat{I}_{R,s})}{n_{A,s} + n_{B,s} + n_{C,s}} \tag{7.6}$$

$$\pi_{BC-HR,s} = \frac{\text{Card}(I_{C,s} \cap \hat{I}_{HR,s})}{n_{C,s}} \tag{7.7}$$

$$\pi_{MC-FP,s} = \frac{\text{Card}(I_{A,s} \cap \hat{I}_{R,s})}{n_{A,s}} \tag{7.8}$$

$$\pi_{MC-FN,s} = \frac{\text{Card}((I_{B,s} \cup I_{C,s}) \cap \hat{I}_{NR,s})}{n_{B,s} + n_{C,s}} \tag{7.9}$$

avec, Card(M) le cardinal de l'ensemble M et, pour le jeu de données s, $\hat{I}_{HR,s}$ la liste des individus classés uniquement dans le groupe estimé à *plus* haut risque ; $\hat{I}_{R,s}$ la liste des individus classés dans un groupe estimé à haut risque et $\hat{I}_{NR,s}$ la liste des individus classés dans un groupe «non significatif» (où la valeur 0 est contenue dans l'intervalle de crédibilité à 95% de l'excès de risque).

7.2.2.3 Résultats avec le vrai nombre de groupes fixé

La Table 7.5 donne les résultats des estimations des excès de risque obtenus sur les 100 jeux de données simulés dans le cas de bonne spécification du nombre de groupes non-vides soit $K = 4$ groupes lors de l'estimation. Dans cette table, la médiane des 100 estimations de β, les intervalles de variabilités IV (quantiles à

5% et à 95% des 100 β médians estimés) et le taux de couverture fondés sur les intervalles de crédibilité à 95% sont présentés.

Comme attendu dans ce cas de bonne spécification du nombre de groupes, les médianes des estimations des excès de risque sont proches des vraies valeurs, ceci pour tous les scénarios et tous les groupes. Les intervalles de variabilité de ces estimations montrent que dans 90% des cas, les estimations restent raisonnablement proches des vraies valeurs.

En comparant les performances des scénarios, le scénario $\mathcal{S}_2$ donne des différences entre β et $\hat{\beta}$ légèrement plus faibles, et ceci de manière plus nette concernant β_A. Les meilleures performances du scénario $\mathcal{S}_2$ semblent logiques car ce scénario correspond aux variables d'exposition radiologique plus précises et donc plus discriminantes entre les groupes. Les taux de couverture de ces paramètres sont également proches du niveau nominal de 95% même si systématiquement un peu plus élevés.

	A			B			C		
	β_A	Med($\hat{\beta_A}$) IV($\hat{\beta_A}$)	couv	β_B	Med($\hat{\beta_B}$) IV($\hat{\beta_B}$)	couv	β_C	Med($\hat{\beta_C}$) IV($\hat{\beta_C}$)	couv
$\mathcal{S}_1$	0	0.10 [-0.30 ; 0.59]	0.97	5	4.96 [3.92 ; 5.99]	0.98	10	9.92 [8.65 ; 11.23]	0.99
$\mathcal{S}_2$	0	-0.00 [-0.37 ; 0.51]	0.97	5	5.01 [3.96 ; 6.30]	0.97	10	9.87 [8.11 ; 11.56]	0.97
$\mathcal{S}_3$	0	0.10 [-0.39 ; 0.50]	0.98	2.5	2.43 [1.82 ; 3.32]	0.99	5	4.95 [3.87 ; 6.30]	0.96

TABLE 7.5 – Médianes (Med) et intervalles de variabilité (IV) des β médians estimés et taux de couverture (couv) des β estimés dans les 100 jeux de données pour chacun des 3 scénarios de simulation. Modèle estimé à 4 groupes non-vides. IV(β) : quantiles 5% et 95% des β médians estimés dans les 100 jeux de données

Les indicateurs de performances à l'échelle individuelle décrits dans la section 7.2.2.2 pour les trois scénarios sont résumés dans la Table 7.6. Une représentation graphique des biais relatifs (pour β_B et β_C) se trouve dans la Figure 7.6. En comparant les biais relatifs $Biais_B^{rel}$ et $Biais_C^{rel}$ pour les différents scénarios, les valeurs sont toutes proches de 0, même si légèrement plus élevées dans le scénario de simulation $\mathcal{S}_3$. La Figure 7.6 montre une légère plus grande variabilité des biais

relatifs également pour le scénario $\mathcal{S}_3$. En effet, ce scénario correspond à des valeurs d'excès de risques plus faibles que dans les autres scénarios et donc plus difficiles à capturer. La valeur du biais absolu pour le groupe à risque nul $Biais_A^{abs}$ est, comme attendu, plus faible dans le scénario $\mathcal{S}_2$. Les scénarios $\mathcal{S}_1$ et $\mathcal{S}_3$ donnent des biais médians de $Biais_A^{abs}$ strictement positifs. Il est à noter que 90% des valeurs des biais obtenues (IV) restent relativement peu éloignées de 0 avec une légère moins bonne performance pour le scénario $\mathcal{S}_1$. Une explication pourrait provenir du fait que les vraies valeurs des autres excès de risque (β_B et β_C) sont plus faibles dans le scénario $\mathcal{S}_3$ que dans $\mathcal{S}_1$. Ainsi, un individu mal classé augmentera le biais de façon moins importante dans le scénario $\mathcal{S}_3$ que dans le scénario $\mathcal{S}_1$.

Les différentes proportions calculées à l'échelle individuelle se trouvent également dans la Table 7.6 . Les proportions d'individus bien classés (π_{BC} et π_{BC-HR}) et mal classés (π_{MC-FP} et π_{MC-FN}) sont similaires pour tous les scénarios. Ces proportions montrent de très bons résultats dans tous les cas avec une proportion de faux positifs plus élevée que celle des faux négatifs mais qui reste de l'ordre de 3%.

L'ensemble de ces indicateurs de performance sont globalement similaires pour les trois scénarios de simulation et montrent de bons résultats. Ceci permet de valider l'inférence du modèle lorsque le nombre de groupe non-vides est correctement spécifié.

	$\mathcal{S}_1$	$\mathcal{S}_2$	$\mathcal{S}_3$
$Biais_A^{abs}$	0.23 [-0.11 ; 0.70]	0.00 [-0.37 ; 0.51]	0.17 [-0.31 ; 0.56]
$Biais_B^{rel}$	-0.01 [-0.21 ; 0.18]	0.00 [-0.21 ; 0.26]	-0.03 [-0.28 ; 0.32]
$Biais_C^{rel}$	-0.02 [-0.14 ; 0.12]	-0.01 [-0.19 ; 0.16]	-0.02 [-0.23 ; 0.25]
π_{BC}	0.98 [0.98 ; 0.99]	1.00 [1.00 ; 1.00]	0.98 [0.98 ; 0.99]
π_{BC-HR}	0.98 [0.98 ; 0.99]	1.00 [1.00 ; 1.00]	0.99 [0.98 ; 0.99]
π_{MC-FP}	0.03 [0.02 ; 0.04]	0.00 [0.00 ; 0.00]	0.03 [0.02 ; 0.04]
π_{MC-FN} (10^{-3})	0.47 [0.46 ; 0.47]	0.46 [0.46 ; 0.47]	0.46 [0.45 ; 0.47]

TABLE 7.6 – Indicateurs de performances médians et leurs intervalles de variabilité à 90% pour les différents scénarios de simulation obtenus à l'échelle individuelle. Modèle estimé avec 4 groupes non-vides.

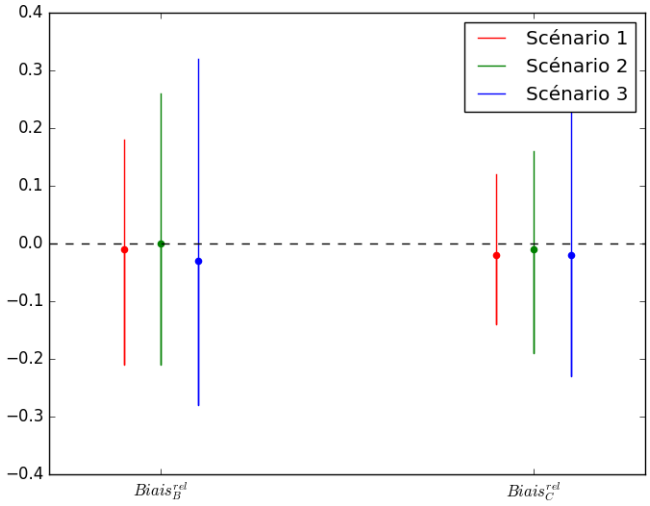

FIGURE 7.6 – Biais relatifs médians (point) et leurs intervalles de variabilité à 90% (barre verticale) pour les scénarios $\mathcal{S}_1$ (rouge), $\mathcal{S}_2$ (vert) et $\mathcal{S}_3$ (bleu) des deux groupes B (à gauche) et C (à droite) obtenus à l'échelle individuelle. Modèle estimé avec 4 groupes non-vides.

7.2.2.4 Sensibilité au nombre de groupes non-vides fixé

Cette partie a pour objectif d'étudier l'impact d'une mauvaise spécification du nombre de groupes non-vides où soit un nombre trop faible de $K = 3$ groupes, soit un nombre trop élevé de $K = 5$ groupes a été fixé lors des estimations. Dans ces deux cas, seuls les indicateurs de performance à l'échelle individuelle sont calculables. Le détail des valeurs de ces différents indicateurs de performance et leurs intervalles de variabilité pour les trois scénarios de simulation pour les modèles à 3, 4 et 5 groupes non-vides sont disponibles dans la Table B.1 en annexe.

Pour des raisons de clarté, nous détaillerons ici les résultats obtenus uniquement dans le cadre du scénario de simulation $\mathcal{S}_1$. La Table 7.7 montre les résultats en terme de biais absolu ou relatif et en terme de proportions de bien ou mal classés. La Figure 7.7 représente les biais relatifs en fonction du nombre de groupes non-vides fixé comprenant le cas de la bonne spécification de ce nombre et la Figure 7.8 illustre les résultats obtenus en terme de proportions d'individus bien et mal classés.

	$K = 3$	$K = 4$	$K = 5$
$Biais_A^{abs}$	1.29 [0.71 ; 1.89]	0.23 [-0.11 ; 0.70]	0.34 [-0.04 ; 0.87]
$Biais_B^{rel}$	0.09 [-0.20 ; 0.24]	-0.01 [-0.21 ; 0.18]	-0.01 [-0.21 ; 0.19]
$Biais_C^{rel}$	-0.20 [-0.30 ; -0.05]	-0.02 [-0.14 ; 0.12]	-0.01 [-0.13 ; 0.12]
π_{BC}	0.66 [0.65 ; 0.68]	0.98 [0.98 ; 0.99]	0.98 [0.83 ; 0.99]
π_{BC-HR}	1.00 [1.00 ; 1.00]	0.98 [0.98 ; 0.99]	0.98 [0.51 ; 0.99]
π_{MC-FP}	1.00 [1.00 ; 1.00]	0.03 [0.02 ; 0.04]	0.03 [0.01 ; 0.49]
π_{MC-FN} (10^{-3})	0.48 [0.48 ; 0.48]	0.47 [0.46 ; 0.47]	1.39 [0.46 ; 2.82]

TABLE 7.7 – Indicateurs de performances médians et leurs intervalles de variabilité à 90% en fonction du nombre K de groupes non-vides fixé lors de l'estimation. Résultats obtenus à l'échelle individuelle pour le scénario $\mathcal{S}_1$.

De manière générale, les modèles avec 4 et 5 groupes non-vides ont des performances similaires en terme de biais et toujours meilleures que dans le cas avec 3 groupes non-vides. Concernant le biais absolu médian du groupe A, il est légèrement plus faible dans le modèle à 4 groupes non-vides par rapport au modèle à 5 groupes non-vides, mais cette différence reste négligeable. Par contre, il est environ 5 fois plus élevé dans le modèle à 3 groupes. Concernant les biais relatifs médians des groupes B et C, les performances des modèles à 4 ou 5 groupes sont très proches entre elles avec des valeurs quasi nulles, les intervalles de variabilité de ces biais relatifs sont également très similaires. Là également, le modèle à 3 groupes montrent des biais relatifs nettement moins bons (environ 10 fois plus élevés), leurs valeurs négatives mettant en évidence une sous estimation de l'excès de risque du groupe C. Les intervalles de variabilité montrent que 90% des biais relatifs avec $K = 3$ varient de 5% à 30% alors que pour les autres valeurs de K ($K = 4$ ou $K = 5$), ces biais sont dans 90% des cas tous inférieurs à 14%.

La Table 7.7 et la Figure 7.8 permettent de comparer les proportions d'individus bien et mal classés en fonction du nombre K. Les proportions médianes d'individus globalement bien classés (π_{BC}) sont élevées et proches de 1 pour les modèles à 4 ou 5 groupes mais cette proportion est beaucoup plus faible valant environ 66% pour le modèle d'estimation à 3 groupes. Les proportions médianes d'individus bien classés dans un groupe à plus haut risque (π_{BC-HC}) sont par contre toutes élevées, cette proportion valant 1 quand $K = 3$. Mais ce dernier résultat est faussement un bon résultat, juste une conséquence d'un défaut de

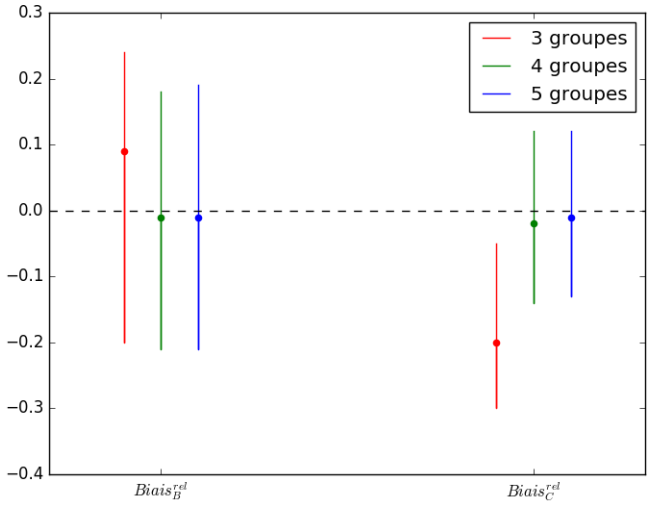

FIGURE 7.7 – Biais relatifs médians (point) et leurs intervalles de variabilité à 90% (barre verticale) pour le modèle estimé à 3 (rouge), 4 (vert) et 5 (bleu) groupes non-vides en fonction des deux vrais groupes B (à gauche) et C (à droite). Résultats obtenus à l'échelle individuelle pour le scénario $\mathcal{S}_1$.

mauvais classement. En effet, ce modèle à $K = 3$ groupes classe la plupart des individus comme à risque, la proportion d'individus faussement classés comme à risque π_{MC-FP} est ainsi proche de 1 alors que cette proportion est raisonnablement proche de zéro pour les deux autres valeurs de K. On peut donc en déduire que dans ce cadre de simulation $\mathcal{S}_1$, le fait de fixer un nombre de groupes non-vides K trop faible engendre une nette sur-estimation du nombre d'individus considérés à risque, augmentant artificiellement la valeur de π_{BC-HC}. Par contre, la proportion d'individus π_{MC-FN} faussement identifiés à risque nul est proche de zéro pour tous les modèles. Finalement, les proportions π_{BC}, π_{BC-HC} et π_{MC-FP} sont légèrement meilleures pour le modèle à 4 groupes par rapport au modèle à 5 groupes, mais les différences restent faibles. L'impact du choix d'un nombre K trop grand est ici négligeable sur les médianes. Par contre, les intervalles de variabilités à 90% montrent une plus grande instabilité pour le modèle à 5 groupes en comparaison à celui à 4 groupes comme par exemple, la borne supérieure de l'intervalle associé à la proportion de faux positifs étant proche de 50%. La largeur de ces intervalles est néanmoins à prendre avec précaution car les quantiles à 5% et 95% sont estimés

sur 100 jeux de données, cette instabilité étant peut-être simplement le reflet de l'instabilité des estimateurs des quantiles eux-mêmes. Une étude sur un nombre plus conséquent de jeux de données doit être envisagée.

Dans la limite des nombres de groupes fixés considérés, l'ensemble de ces résultats amène à deux considérations importantes. La première considération concerne l'impact d'un nombre de groupes fixé trop élevé lors des estimations. Que ce soit en termes de biais ou de proportion d'individus bien et mal classés, les résultats obtenus sont très proches des résultats avec le bon nombre de groupes fixé. L'impact peut donc être considéré ici comme faible. La sensibilité au nombre de groupes est ici négligeable, ceci constitue un atout indéniable.

La seconde considération concerne le cas où le nombre de groupes non-vides est inférieur au vrai nombre. Les conséquences sont cette fois-ci loin d'être négligeables. Notamment, la proportion d'individus considérés à risque à tort est très élevée et les biais sont plus élevés.

L'ensemble de ces remarques est généralement confirmé dans le cas des autres scénarios (voir Table B.1 en annexe) avec des valeurs médianes proches entre les modèles avec $K = 4$ et $K = 5$ que ce soit en terme de biais ou de proportions de bon ou mauvais classement mais beaucoup plus éloignées pour $K = 3$.

Cette étude de simulation a permis de mettre en évidence les bons résultats du modèle RPRM dans le cas où le nombre de groupes non-vides est soit fixé à la vraie valeur soit sur-évalué. Cette étude permet également de conclure à fortement privilégier une sur-évaluation du nombre de groupes car une sous-évaluation de ce nombre engendre des taux de faux positifs trop importants.

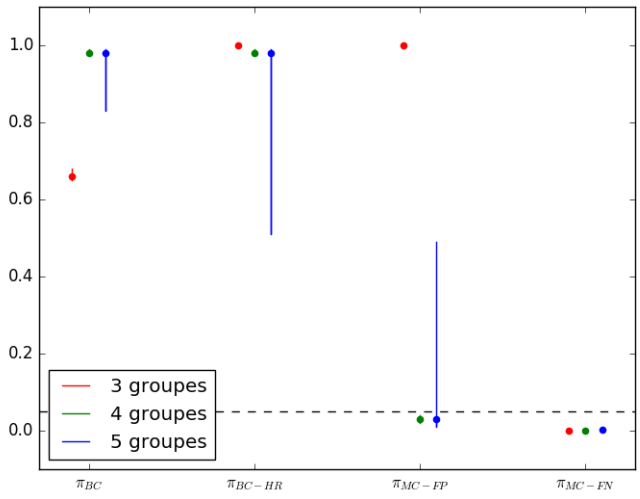

FIGURE 7.8 – Proportions médianes (points) et intervalles de variabilité à 90% d'individus bien classés (π_{BC}), d'individus biens classés à plus haut risque (π_{BC-HR}), d'individus mal classés en faux positifs (π_{MC-FP}) et en faux négatifs (π_{MC-FN}) pour le modèle estimé à 3 (rouge), 4 (vert) et 5 groupes (bleu). Résultats obtenus à l'échelle individuelle pour le scénario $\mathcal{S}_1$.

CHAPITRE 8

Discussion

8.1 Synthèse

Ce travail de thèse se place dans le contexte général de l'estimation de risques de cancers radio-induits, et de leur incertitude, en épidémiologie des RIs. Il contribue à améliorer la connaissance des effets sanitaires faisant suite à une ou plusieurs expositions conjointes prolongées à de faibles doses et débits de dose de RIs. À travers le développement et la mise en œuvre de modèles hiérarchiques bayésiens, ce travail permet de promouvoir cette approche comme une réponse flexible et élégante au regard de deux défis statistiques d'actualité, inhérents à l'estimation de risques sanitaires en situation de co-expositions environnementales : 1) la prise en compte d'erreurs de mesure sur les expositions radiologiques, de données d'exposition manquantes ou censurées à gauche du fait de l'existence d'une limite de détection sur les dosimètres personnels utilisés pour mesurer ces expositions ; 2) la prise en compte d'expositions radiologiques multiples fortement corrélées (communément appelé problème de multicolinéarité). Le cas d'étude traité portait sur l'analyse de l'association entre expositions radiologiques chroniques à faibles doses et mortalité par cancer du poumon dans la sous-cohorte post-55 des mineurs d'ura-

nium français. Cette cohorte professionnelle constitue notamment une population de référence - caractérisée par un suivi long des individus, peu de perdus de vue et des données de bonne qualité - pour l'estimation des risques sanitaires associés à une exposition chronique à faibles doses au radon. Parmi les nombreux radionucléides descendants de l'uranium-238 naturellement présent dans les mines et les différents types de RIs auxquels sont exposés les mineurs d'uranium dans le cadre de leur activité professionnelle, trois co-expositions radiologiques ont été considérées dans cette thèse : le radon, les poussières fines de minerai chargées des radionucléides émetteurs α à vie longue, les rayonnements γ. Ces dernières étaient disponibles dans la sous-cohorte post-55 des mineurs d'uranium français et potentiellement cancérigènes pour l'Homme. Les deux défis statistiques précédemment cités ont été traités séparément dans cette thèse. Une synthèse propre à chacun des travaux menés pour y répondre sont présentés dans la suite de cette section.

Prise en compte d'expositions radiologiques incertaines dans l'estimation d'un risque radio-induit

Hoffmann *et al.* (Hoffmann, Laurier et al. 2018) ont montré qu'une modélisation adéquate des incertitudes d'exposition est essentielle lorsqu'on souhaite étudier leur impact respectif possible sur les estimations de risques sanitaires. Ces incertitudes peuvent être multiples et potentiellement complexes, ce qui est notamment le cas dans les cohortes professionnelles : erreurs de mesure hétéroscédastiques en raison de la diversité des techniques utilisées pour mesurer ces expositions selon la période calendaire, sources d'incertitude intervenant sur des variables d'exposition longitudinales, combinaison d'erreurs de mesure de nature Berkson et classique, existence d'erreurs de mesure partagées. Une question se pose alors : Comment modéliser, de manière appropriée, de telles sources d'incertitude multiples et complexes ? La réponse proposée dans ce travail est d'utiliser une approche par modélisation probabiliste hiérarchique. Cette approche permet la combinaison de sous-modèles probabilistes dans lesquels différentes sources d'incertitude, incluant par exemple erreurs de mesure hétéroscédastiques et mécanisme de censure déterministe, sont prises en compte à différents niveaux de la hiérarchie. Au prix d'hypothèses d'indépendance conditionnelle, un modèle hiérarchique permet de combiner plusieurs sous-modèles probabilistes simples dans un cadre

unique et flexible. Cela permet de décrire avec souplesse des phénomènes aléatoires complexes, tout en bénéficiant d'une interprétation claire des quantités inconnues impliquées dans les différents sous-modèles.

Dans le cadre de l'estimation du risque de décès par cancer du poumon dans la sous-cohorte post-55 des mineurs d'uranium français, trois modèles hiérarchiques - basés sur différents sous-modèles d'exposition - ont ainsi été développés, ajustés sous le paradigme bayésien puis comparés afin de tenir compte explicitement d'équivalents de dose incertains, associés à l'exposition aux rayonnements γ. La flexibilité de la modélisation hiérarchique a été mise à profit pour décrire simultanément plusieurs sources d'incertitude d'exposition : des erreurs de mesure hétéroscédastiques de nature classique, un processus de censure à gauche déterministe dû à la limite de détection des dosimètres et des équivalents de dose manquants. À noter que les modèles proposés ont originalement permis de coupler un processus de censure avec l'existence possible d'erreurs de mesure sur les valeurs censurées latentes. Les approches fonctionnelles classiques pour la correction d'erreurs de mesure telles que SIMEX (Cook et Stefanski 1994) ou la régression-calibration (Stefanski et Carroll 1985) reposent sur plusieurs étapes consécutives : estimation de l'exposition réelle dans un premier temps puis estimation des coefficients de risque inconnus dans un second temps. De même, en cas de données d'exposition manquantes ou censurées, l'approche classique utilisée en épidémiologie des RIs consiste à estimer ces données d'exposition non observées puis, dans un deuxième temps, à estimer les coefficients de risque inconnus en supposant les valeurs imputées comme «vraies» et en négligeant les incertitudes d'estimation associées. Au contraire, l'ajustement de modèles hiérarchiques sous le paradigme bayésien a permis une estimation simultanée - dans un cadre inférentiel cohérent et valide - des coefficients de risque inconnus et de leur incertitude, des vrais équivalents de dose de rayonnements γ ainsi que de ceux qui auraient pu être estimés par les dosimètres si non indiqués comme faux-zéros ou comme données manquantes dans la base de données des mineurs d'uranium français. Ainsi, l'incertitude d'estimation associée à chaque quantité inconnue - qui peut être importante et/ou distribuée de manière non gaussienne autour de la «vraie» valeur - est directement prise en compte lors de l'estimation d'autres quantités inconnues qui en dépendent. Dans ce travail, par exemple, la mise en oeuvre d'un algorithme MCMC a permis d'échantillonner dans

la distribution *a posteriori* jointe des vraies doses de rayonnements γ et donc, *a fortiori*, des vraies doses cumulées entrant en jeu dans les modèles dose-réponse. Cela a permis de prendre directement en compte l'incertitude d'estimation associée aux vraies doses cumulées de rayonnements γ dans l'estimation de l'excès de risque instantané de décès par cancer du poumon. Il est également important de noter que les modèles hiérarchiques bayésiens proposés ont permis de tirer parti des informations disponibles dans les équivalents de dose de rayonnements γ nuls (i.e., vrais zéros) ou estimés par les dosimètres à une période donnée et/ou pour un type de mine donné pour apprendre sur les équivalents de dose manquants ou censurés à gauche pendant la même période et/ou dans le même type de mine. Enfin, un autre avantage à l'utilisation de la statistique bayésienne pour inférer les modèles hiérarchiques proposés était la possibilité d'attribuer une distribution *a priori* informative aux paramètres inconnus qui n'étaient que peu renseignés par les données, en vue d'améliorer leur estimation. Dans la présente étude, cela a été particulièrement utile pour l'estimation du risque instantané de base (i.e., en l'absence d'expositions radiologiques) de décès par cancer du poumon pour les mineurs d'uranium français de moins de 40 ans : cela correspondait au paramètre λ_1 dans le chapitre 5. En effet, seul un mineur est décédé avant 40 ans dans la sous-cohorte post-55 des mineurs d'uranium français. Peu d'informations étaient donc disponibles dans cette sous-cohorte pour estimer λ_1. Les données de mortalité par cancer du poumon dans la population française masculine générale (*Base de données des taux de mortalité de référence. IRSN/PRP-HOM/SRBE/LEPID/2013-07* 2013) ont ainsi été utilisées pour assigner une loi Gamma informative sur λ_1.

Compte-tenu des hypothèses de modélisation faites dans les 3 structures hiérarchiques proposées, aucun impact notable des différentes sources d'incertitude considérées sur les équivalents de dose de rayonnements γ n'a été mis en évidence sur le risque de décès par cancer du poumon, dans la sous-cohorte post-55 des mineurs d'uranium français. Récemment, Hoffmann *et al.* (2018) ont montré, par une étude de simulations, que l'impact d'une erreur de mesure de nature classique, non partagée, multiplicative et de distribution log-normale avec variance géométrique égale à 0.1 est négligeable sur l'estimation du risque dans le cas d'un modèle de survie en excès de risque instantané (EHR) (Hoffmann, Laurier et al. 2018). Dans ce travail, suivant le tableau 3.1 du chapitre 3 (issu des travaux de thèse de

Rodrigue Allodji (2011) (S. R. Allodji 2011)), la variance géométrique de l'erreur de mesure a été supposée inférieure à 0.06 pour toute la période d'exposition aux rayonnements γ. Cela explique, au moins en partie, pourquoi la prise en compte de l'erreur de mesure sur les équivalents de dose de rayonnements γ n'a pas eu d'impact notable sur l'estimation du risque de décès par cancer du poumon dans la sous-cohorte post-55. En outre, la faible proportion (environ 7%) d'équivalents de dose censurés à gauche ou manquants dans la sous-cohorte post-55 des mineurs d'uranium français explique, en moins en partie, pourquoi remplacer simplement par zéro tous ces équivalents de dose n'a finalement pas d'impact significatif sur les estimations de risques. Le coefficient de risque corrigé de décès par cancer du poumon associé à l'exposition au rayonnements γ a été estimé à 0.81 par 100 mSv avec un intervalle de crédibilité à 95% égal à [0.28 ; 1.75]. Cela confirme la robustesse de l'estimation du risque obtenue à partir d'un modèle de survie en EHR qui ne tient pas compte des erreurs de mesure et dans lequel toutes les valeurs d'équivalents de dose de rayonnements γ censurées à gauche et manquantes sont remplacées par zéro (EHR de 0.78 par 100 mSv avec un intervalle de crédibilité à 95% égal à [0.28 ; 1.64]). Après avoir pris en compte différentes sources d'incertitude d'exposition, une association positive statistiquement significative reste mise en évidence entre l'exposition aux rayonnements γ et le risque de décès par cancer du poumon, même si l'intervalle de crédibilité à 95% de l'EHR est plus large que celui obtenu sans prise en compte des erreurs de mesure. Il a été montré que les résultats obtenus sont également robustes aux différents choix de modélisation considérés pour le sous-modèle d'exposition dans lequel est décrite l'évolution temporelle des «vrais» équivalents de dose de rayonnements γ attendus. En outre, une analyse de sensibilité a montré que les résultats obtenus sont robustes à une augmentation de 75% de la limite de détection des dosimètres. Enfin, il est intéressant de noter que le modèle hiérarchique bayésien basé sur le sous-modèle d'exposition $\mathcal{M}_3$ a permis de mettre en évidence une diminution dans le temps statistiquement significative de l'espérance des vrais équivalents de doses de rayonnements γ (à échelle logarithmique) dans les mines souterraines et à ciel ouvert. Il a également montré que cette tendance à la baisse était plus forte dans les mines souterraines que dans les mines à ciel ouvert. Cette tendance à la baisse a également été observée dans la cohorte des mineurs d'uranium allemands (Kreuzer, Dufey et al. 2013). Elle

pourrait s'expliquer d'un part par des opérations minières intensives qui auraient pu contribuer à réduire la concentration en uranium des mines et donc, *a fortiori*, le niveau d'exposition aux rayonnements γ, et d'autre part, par des changements dans les pratiques de radioprotection avec notamment une réduction de la durée d'exposition des mineurs d'uranium.

Prise en compte d'expositions radiologiques multiples et fortement corrélées dans l'estimation d'un risque radio-induit

La deuxième partie de ce travail de thèse apporte une première réponse, souple et élégante, au problème spécifique de l'estimation d'un risque sanitaire radio-induit en présence de plusieurs covariables d'expositions radiologiques fortement corrélées et d'une variable réponse de type survie fortement censurée. Du point de vue méthodologique, une extension de la classe des modèles de mélange par régression bayésienne sur profils d'exposition (modèle PRM pour «Profile Regression Mixture») au contexte des modèles de survie en EHR (classiquement utilisés en épidémiologie des RIs) a été proposée. Il s'agit de la première utilisation de cette classe de modèles en épidémiologie des RIs pour pallier au problème de multicolinéarité classiquement rencontré en situation de multi-expositions radiologiques. Présenté sous forme hiérarchique, le modèle PRM bayésien proposé se compose notamment d'un sous-modèle de maladie en EHR et d'un processus de Dirichlet tronqué comme sous-modèle d'attribution. L'inférence d'un modèle PRM permet : 1) d'identifier et de caractériser des groupes d'individus ayant un profil d'expositions similaire (à différentes sources radiologiques par exemple) et un risque similaire vis-à-vis d'une pathologie d'intérêt ; 2) d'estimer le risque d'intérêt associé à chaque groupe identifié, ainsi que l'incertitude d'estimation associée. Une importante étude par simulations a également été réalisée afin d'étudier les performances du modèle proposé du point de vue de l'estimation des coefficients de risque et de la qualité des classifications d'individus obtenues dans le cas où le nombre de groupes d'individus est fixé.

Le modèle PRM bayésien proposé a été appliqué aux données de la sous-cohorte post-55 des mineurs d'uranium français afin d'estimer l'excès de risque instantané de décès par cancer du poumon associé à l'exposition cumulée au radon, aux poussières d'uranium et aux rayonnements γ ainsi qu'à toute autre exposition

professionnelle par le biais de variables de substitution comme le type de poste et la localisation des mines. Un algorithme MCMC adaptatif de type Metropolis-Within-Gibbs a été implémenté en Python pour échantillonner dans la distribution *a posteriori* jointe de tous les paramètres inconnus et de toutes les variables latentes. Les premiers résultats obtenus ont montré que la loi *a posteriori* recherchée est multimodale. Une attention particulière a ainsi été portée sur la convergence vers des modes locaux de l'algorithme implémenté. Comme suggéré par Liverani *et al* (Silvia Liverani et al. 2015), trois mouvements de changement des labels de groupe ont été introduits afin d'essayer d'éviter une convergence de l'algorithme MCMC implémenté vers un mode local et de réduire les auto-corrélations intra-chaînes. Malheureusement, malgré ces précautions, l'algorithme MCMC implémenté ne visite pas efficacement l'ensemble des points du support de la loi *a posteriori*. Selon les positions initiales choisies, les chaînes de Markov vont se bloquer dans un mode local différent. Une première approche pragmatique, en trois étapes, a été définie pour résoudre ce problème : a/ le nombre de groupes non-vides de mineurs d'uranium (i.e., incluant le groupe des mineurs non exposés) a été estimé en ajustant le modèle PRM complet en fonction de différentes valeurs initiales fixées du paramètre-clé α : des partitions en 5, 6, 7 ou 8 groupes non-vides ont ainsi été identifiées ; b/ quatre modèles PRM bayésiens restreints (RPRM pour «Restricted Profile Regression Model») dans lesquels le nombre de groupes non-vides a été fixé à 5, 6, 7 et 8 respectivement ont été ajustés à la sous-cohorte post-55 ; c/ deux critères bayésiens de sélection de modèles (i.e., DIC et WAIC) ont été calculés et ont conduit à sélectionner la partition en 8 groupes non vides comme étant la « meilleure » selon les deux critères retenus et sachant les données disponibles.

Le modèle RPRM bayésien avec 8 groupes non-vides a permis d'identifier des groupes de mineurs très intéressants. Deux d'entre eux ont été associés à un EHR strictement positif et statistiquement significatif de décès par cancer du poumon. Le premier groupe (EHR=1.4, IC 95%=[0.60 ; 2.60]) correspondait aux mineurs les plus exposés au radon, aux rayonnements γ et aux poussières d'uranium et ce depuis plus de 19 ans (principalement avant la mécanisation ou comme foreur après la mécanisation, hors de la mine située dans l'Hérault). Le second groupe (EHR=1.2, IC 95%=[0.17 ; 2.80]) correspondait aux mineurs très jeunes lors de

leur première exposition et qui ont été fortement exposés au radon, aux rayonnements γ et aux poussières d'uranium (mais dans une moindre mesure que le groupe précédent) et ce pendant plus de 13 ans (principalement comme foreur après la mécanisation ou autre travail souterrain avant la mécanisation). De manière intéressante, les résultats obtenus montrent également que les mineurs ayant travaillé après la mécanisation et principalement dans la mine de l'Hérault - la seule mine d'uranium avec un sol sédimentaire - ont un risque faible [1] et non statistiquement significatif alors que ces mineurs ont été fortement exposés aux trois sources radiologiques (cf. groupe B dans la Figure 7.5 du chapitre 7). Le modèle RPRM bayésien à 8 groupes non-vides a ainsi permis de fournir une interprétation originale, riche et fine de l'association potentielle entre le risque de décès par cancer du poumon et les profils spécifiques d'exposition aux RIs des mineurs d'uranium français. Il permet d'estimer l'effet combiné de plusieurs sources d'expositions radiologiques tout en modulant cet effet par d'autres facteurs de risque potentiels tels que l'âge à la première exposition et la durée d'exposition.

L'analyse par simulations réalisée dans cette thèse a permis de valider l'algorithme d'inférence MCMC implémenté pour estimer les paramètres d'un modèle RPRM bayésien lorsque le nombre de groupes est correctement spécifié. En effet, dans ce cas, les estimations des excès de risque sont proches des «vraies» valeurs fixées lors de la simulation de données et ce, pour tous les scénarios de simulation testés et tous les groupes d'individus. Par ailleurs, les proportions d'individus bien classés sont proches de 1. Lorsque le nombre de groupes fixé est supérieur au «vrai» nombre de groupes simulés, de très bons résultats sont obtenus pour tous les scénarios d'exposition considérés avec un impact négligeable sur les biais d'estimation des excès de risque et les proportions d'individus bien et mal classés. En revanche, lorsque le nombre de groupes fixé est inférieur au «vrai» nombre de groupes simulés, les simulations réalisées mettent en évidence une forte sous-estimation de l'excès de risque pour les groupes à très haut risque qui s'accompagne d'une très forte proportion d'individus classés à risque à tort.

Définir et surveiller l'exposome humain est une tâche très difficile, étant donné la grande variété des facteurs environnementaux, des paramètres biologiques et des interactions gène-environnement (Wild 2005 ; Rappaport et Smith 2010 ;

1. Le deuxième risque le plus faible sur les 8 groupes identifiés

Slama et Vrijheid 2015). Wild a suggéré que la mesure de l'exposition dans l'une des grandes catégories d'exposition suivantes - interne (par exemple, hormones, microflore), externe spécifique (par exemple, substances toxiques) et externe générale (par exemple, sociale, psychologique) - peut permettre de refléter certains aspects de l'exposome global (Wild 2012). En outre, il peut être avantageux pour le développement de méthodes statistiques de restreindre le champ d'application de l'exposome à une classe particulière d'expositions et/ou à des étapes spécifiques de la vie afin de les améliorer et de les valider pour les appliquer ultérieurement au concept plus large d'exposome, dans une évaluation des risques ou un cadre réglementaire (Bennett et al. 2020). Ce fut le cas dans ce travail qui s'est concentré sur l'exposition professionnelle à plusieurs types de RIs des mineurs d'uranium français, en ne considérant qu'un petit nombre (c'est-à-dire 7) de covariables d'exposition. Il montre que les modèles PRM bayésiens sont prometteurs pour la recherche sur l'exposome en épidémiologie des RIs.

8.2 Limites

Une limite commune aux différents travaux menés dans cette thèse concerne la consommation de tabac des mineurs d'uranium qui, malheureusement, n'a pas pu être prise en compte dans les analyses bien que celle-ci soit reconnue comme étant la principale cause de décès par cancer du poumon. En effet, pour rappel, le statut tabagique n'est disponible que pour 4.2% des mineurs de la sous-cohorte post-55. Ce manque crucial d'informations rend très dangereux l'ajustement sur le statut tabagique lors de l'estimation du risque de décès par cancer du poumon dû à l'exposition aux rayonnements γ seule ou à l'ensemble des sources radiologiques disponibles au sein de cette sous-cohorte. Par ailleurs, une analyse par validation croisée, menée dans le cadre de cette thèse, a confirmé qu'aucun prédicteur potentiel du statut tabagique n'était disponible dans la sous-cohorte post-55 (cf. section 7.2.1.3 du chapitre 7). Une imputation fiable des 96% de statuts tabagiques manquants n'était donc pas envisageable. Néanmoins, des analyses antérieures sur l'impact du tabagisme dans les études de cohortes professionnelles de mineurs d'uranium ont suggéré que le tabagisme n'était pas un facteur de confusion dans ces études (David B Richardson et al. 2014 ; Keil, David B Richardson et Troester

2015). Cela n'a rien de vraiment surprenant puisqu'il n'y a en fait aucune raison de penser que le statut tabagique soit fortement associé aux niveaux d'expositions radiologiques dûs à l'activité professionnelle des mineurs d'uranium. En outre, des analyses sur des sous-cohortes pour lesquelles les antécédents tabagiques des travailleurs étaient disponibles ont observé une association positive significative entre l'exposition au radon et la mortalité par cancer du poumon, tant pour les fumeurs que pour les non-fumeurs, avec un coefficient de risque estimé qui avait d'ailleurs tendance à être plus élevé chez les non-fumeurs que chez les fumeurs (Kreuzer, Sobotzki et al. 2017 ; Tomasek 2002 ; Tirmarche, Harrison et al. 2012). Enfin, des études cas-témoins nichées dans la cohorte française de mineurs d'uranium (Klervi Leuraud, Billon et al. 2007) et dans deux autres cohortes européennes de mineurs d'uranium (Klervi Leuraud, Schnelzer et al. 2011) ont montré que, si l'on ajuste les modèles exposition-risque sur le statut tabagique, l'effet de l'exposition au radon sur le risque de décès par cancer du poumon persiste avec un coefficient de risque estimé robuste à cet ajustement. Si la proportion de statut tabagique manquant était raisonnable (environ 30 %), il serait néanmoins intéressant que les modèles PRM bayésiens puissent traiter ces covariables manquantes - tout en tenant compte de leur incertitude - pour identifier les profils d'exposition.

La suite de cette section présente les limites principales spécifiques aux différents travaux menés dans le cadre de la thèse.

Prise en compte d'expositions radiologiques incertaines dans l'estimation d'un risque radio-induit

Lors de la prise en compte des incertitudes sur l'exposition aux rayonnements γ dans l'estimation de l'excès de risque de décès par cancer du poumon dans la sous-cohorte post-55 des mineurs d'uranium français, la variance de l'erreur de mesure classique a été fixée en fonction de la période calendaire, selon les travaux de Rodrigue Allodji (S. R. Allodji 2011) (cf. Table 3.1 du chapitre 3). Néanmoins, une analyse de sensibilité, menée dans le cadre de cette thèse, montre que l'estimation de cet excès de risque est sensible à ce paramètre de variance d'erreur. Cela montre qu'une évaluation minutieuse de ce paramètre crucial doit être faite lors de la prise en compte d'erreurs de mesure de nature classique afin de garantir la validité de l'estimation des coefficients de risque. Ici, par exemple,

l'estimation corrigée du risque de décès par cancer du poumon (0.81 pour 100 mSv avec un IC à 95% égal à [0.28 ; 1.75]) doit être considérée avec prudence, en supposant que l'écart-type géométrique de l'erreur de mesure (multiplicative et lognormale) soit de 0.245 entre 1956 et 1985 et de 0.16 entre 1986 et 2007 (ibid.).

Comme montré dans la section 3.3 du chapitre 3, les valeurs d'exposition au radon et les équivalents de dose de rayonnements γ sont fortement corrélés dans la sous-cohorte post-55 des mineurs d'uranium français. Étant donné que le radon n'a pas été inclus dans les modèles hiérarchiques développés dans le chapitre 5 [2] alors qu'il est reconnu comme carcinogène pulmonaire chez l'Homme depuis 1988 (IARC et al. 1988), l'association positive statistiquement significative constatée entre la dose de rayonnements γ et le risque de décès par cancer du poumon chez les mineurs d'uranium français doit donc être interprétée avec une extrême prudence. En effet, les résultats obtenus pourraient refléter soit une association spécifique réelle entre la dose de rayonnements γ et la mortalité par cancer du poumon, soit, plus probablement, une association fallacieuse due au radon qui agirait comme facteur de confusion dans cette relation.

Prise en compte d'expositions radiologiques multiples et fortement corrélées dans l'estimation d'un risque radio-induit

Comme énoncé précédemment, l'ajustement aux données de la sous-cohorte post-55 des mineurs d'uranium français du modèle PRM bayésien développé dans le chapitre 6 a conduit à une distribution *a posteriori* estimée multimodale. Cela pourrait être dû à un manque de signal dans la base de données disponibles qui impliquerait que plusieurs classifications distinctes de mineurs (en groupes caractérisés par des profils d'expositions proches et des risques proches de décès par cancer du poumon) conduiraient à des probabilités *a posteriori* élevées similaires. En d'autres termes, aucune partition ne se détacherait nettement comme «optimale». Dans ce contexte spécifique et étant donné le grand nombre de paramètres inconnus et de variables latentes inclus dans les modèles de mélange de type PRM, l'utilisation d'un algorithme MCMC de type Metropolis-Within-Gibbs adaptatif tel que celui proposé dans le chapitre 6 semble ne pas être la meilleure méthode d'inférence bayésienne. En effet, l'utilisation d'un échantillonneur de Gibbs ou d'un

2. En raison de problèmes de multicolinéarité non traités dans cette partie de la thèse

échantillonneur de Métropolis-Hastings avec marche aléatoire (dont la variance a cependant été calibrée de manière à bénéficier d'un taux d'acceptation correct) a mené à une exploration lente et seulement locale de la distribution *a posteriori* cible du modèle PRM développé, lorsqu'ajusté aux données de la sous-cohorte post-55. Ce problème a été notamment illustré par Gelman *et al.* (Andrew Gelman et al. 2013) dans le cadre plus général de modèles complexes de grande dimension. Betancourt et Girolami (Betancourt et Girolami 2015) ont par ailleurs montré que, dans le cas de modèles de mélange ou de modèles hiérarchiques, la situation pouvait notamment s'aggraver lorsque le nombre de groupes ou que le nombre de niveaux augmente respectivement. Comme détaillée précédemment, une première approche, se voulant essentiellement pragmatique, a consisté à ajuster plusieurs modèles PRM restreints dans lesquels le nombre inconnu de groupes a été préalablement fixé à différentes valeurs. Néanmoins, parvenir à inférer le modèle PRM complet en visitant efficacement l'ensemble du support de la distribution *a posteriori* multimodale cible serait bien plus optimal. Cela permettrait d'éviter de devoir faire le choix du nombre de groupes étant donné que celui-ci peut avoir un impact important sur les estimations en cas de mauvaise spécification (cf. section 7.2.2 du chapitre 7). Cela permettrait également d'intégrer dans les estimations une prise en compte de l'incertitude (probablement importante) sur le nombre inconnu de groupes.

Le modèle PRM bayésien proposé dans le chapitre 6 présentent deux principales limites. Tout d'abord, il ne tient pas compte de la dimension temporelle des covariables d'expositions radiologiques. En effet, pour chaque source radiologique, seule l'exposition cumulée totale estimée sur l'ensemble de la carrière de chaque mineur a été considérée dans le modèle PRM comme valeur «résumée» de l'exposition sur toute la période calendaire de suivi. Or, comme expliqué dans la section 2.3.3 du chapitre 2, l'exposition au radon, aux rayonnements γ et aux poussières d'uranium est disponible annuellement pour chaque mineur dans la sous-cohorte post-55. Par ailleurs, le modèle PRM bayésien développé ne prend pas en compte les différentes sources d'incertitude sur l'exposition au radon, aux rayonnements γ et aux poussières d'uranium lors de l'identification des groupes de mineurs et de l'estimation des risques de décès par cancer du poumon qui leur sont associés. Bien que ce travail de thèse ait montré qu'une prise en compte des incertitudes

sur les doses de rayonnements γ ne semble pas nécessaire, les travaux de thèse de Sabine Hoffmann (Hoffmann 2017) ont en revanche montré que la structure des erreurs de mesure sur les expositions au radon est bien plus complexe que celle des rayonnements γ (avec notamment l'existence d'erreurs de mesure Berkson partagées) et que, lorsqu'elle n'est pas ou mal prise en compte, peut conduire à des estimations de risque biaisées et à une déformation de la relation exposition-risque. La structure des erreurs de mesure sur les expositions aux poussières d'uranium étant relativement similaire à celle des expositions au radon, une prise en compte de ces erreurs semblerait également importante à intégrer dans le modèle PRM afin d'étudier la robustesse des classifications obtenues et des estimations d'excès de risque de décès par cancer du poumon. En effet, la non-prise en compte des erreurs de mesure sur les expositions au radon et aux poussières d'uranium pourrait avoir un impact non négligeable sur l'identification des groupes et sur l'estimation des risques de décès par cancer du poumon qui leur sont associés.

8.3 Perspectives

À court terme, une perspective intéressante, purement algorithmique, sera d'améliorer l'efficacité de l'algorithme d'inférence bayésienne proposé dans le chapitre 6, afin de pouvoir ajuster le modèle PRM bayésien complet. Plusieurs pistes pourraient être envisagées comme : a) l'implémentation d'échantillonneurs plus sophistiqués pour la mise à jour des labels de classe d'un modèle de mélange comme un échantillonneur de Langevin (Celeux, Hurn et Robert 2000) ou un échantillonneur par tranches (Slice sampling en anglais) (Walker 2007) ; b) l'implémentation de dynamiques Hamiltoniennes (Neal 1994) pour la mise à jour de certains paramètres inconnus. Bien que difficiles à calibrer, les algorithmes MCMC basés sur de telles dynamiques (appelés algorithmes HMC pour « Hamiltonian Monte-Carlo ») peuvent être plus efficaces que les algorithmes MCMC standard de type Metropolis-Within-Gibbs pour traiter les problèmes de multimodalité de la loi *a posteriori* (Betancourt et Girolami 2015) et ainsi, inférer des modèles PRM bayésiens.

À court terme, l'analyse par simulations présentée dans le chapitre 7 sera également complétée afin de simuler un nombre plus élevé de groupes de mineurs (par

exemple 8 groupes non vides au lieu de 4). L'objectif sera de confirmer ou non les résultats obtenus pour 4 groupes simulés. Pour chaque jeu de données simulé, il sera également intéressant de réaliser la même procédure de sélection de modèles que celle proposée dans la section 7.2.1.1 du chapitre 7 en comparant les capacités d'ajustement aux données de modèles RPRM bayésiens à nombre fixé de groupes de mineurs d'uranium (avec un calcul de DIC et/ou WAIC).

Une perspective envisagée pour pallier aux limites précédemment évoquées est d'étendre la classe des modèles PRM bayésiens pour prendre en compte la dynamique temporelle de données de co-expositions longitudinales (telles que généralement rencontrées dans les études de cohorte en épidémiologie des RIs). Chaque individu pourrait être affecté à un groupe unique qui dépendrait de toute sa trajectoire d'exposition. Alternativement, le label du groupe de chaque individu pourrait changer au fil du temps en fonction de la dynamique temporelle de ses expositions. De par leur structure hiérarchique, les modèles PRM bayésiens pourraient également être étendus pour tenir compte des erreurs de mesure et autres sources d'incertitude sur les co-expositions environnementales considérées. En effet, celles-ci constituent, avec la multicolinéarité, l'un des problèmes les plus importants lors de l'évaluation d'associations exposition-santé (Billionnet et al. 2012). En particulier, dans le cas de l'estimation du risque de décès par cancer du poumon dans la sous-cohorte post-55 des mineurs d'uranium français, les sous-modèles de mesure et d'exposition des modèles hiérarchiques décrits dans le chapitre 5 pourraient être inclus dans un modèle PRM basé sur les co-expositions radiologiques annuelles. Ces sous-modèles pourraient également être étendus afin de tenir compte de valeurs d'expositions strictement positives, manquantes ou censurées à gauche sujettes à des erreurs de mesure plus complexes que celles des rayonnements γ et pour lesquelles un impact substantiel sur les estimations de risque est attendu (Hoffmann, Laurier et al. 2018). Cela serait par exemple le cas des expositions au radon ou aux poussières d'uranium pour lesquelles des erreurs de mesure partagées sur plusieurs années de suivi d'un même individu sont suspectées dans la sous-cohorte post-55. Enfin, sous le paradigme bayésien, la possibilité d'inclure des informations *a priori* sur certains paramètres incertains comme ceux du sous-modèle d'exposition pourrait contribuer à réduire l'incertitude sur les estimations des coefficients de risque.

Les modèles PRM bayésiens permettent de pallier au problème de multicolinéarité fréquemment rencontré lors de l'estimation d'effets sanitaires en situation d'expositions environnementales multiples. Ils permettent d'estimer l'effet combiné de plusieurs sources d'expositions tout en modulant cet effet par d'autres facteurs de risque potentiels. En revanche, ils ne permettent pas d'interpréter aisément les potentiels effets de synergie ou d'antagonisme entre des expositions conjointes à différents agents environnementaux et autres facteurs de risque. D'autres hypothèses de modélisation comme celles proposées par exemple dans les «Bayesian Kernel Machine Regression» (Bobb et al. 2015) pourraient être considérées pour répondre à cet enjeu spécifique.

Le modèle PRM bayésien proposé dans le chapitre 6 a été appliqué à la cohorte post-55 des mineurs d'uranium français, sur un nombre limité de variables d'expositions radiologiques (c'est-à-dire 7). La question se pose donc de savoir dans quelle(s) mesure(s) une telle approche serait applicable à la prise en compte d'un nombre plus élevé de variables d'expositions voire de nature différente (e.g., chimique). En théorie, un plus grand nombre de covariables, y compris de facteurs de risque environnementaux et génétiques, pourraient être inclus dans les modèles PRM bayésiens afin d'étudier, par exemple, des interactions gène-environnement. Néanmoins, en pratique, les performances des modèles PRM restent encore à évaluer dans ce contexte plus difficile.

Enfin, pour mieux mettre en évidence l'impact potentiel sur les estimations de risque du remplacement par zéro de valeurs d'expositions radiologiques manquantes ou censurées à gauche, il serait très intéressant d'appliquer les modèles hiérarchiques bayésiens proposés dans le chapitre 5 sur d'autres cohortes (par exemple de travailleurs du cycle du combustible nuéclaire) affectées par une très forte proportion de valeurs d'expositions radiologiques manquantes ou censurées à gauche.

8.4 Conclusion

Pour conclure, l'approche par modélisation (probabiliste) hiérarchique, combinée à un apprentissage statistique bayésien, offre des avantages indéniables pour l'estimation de risques sanitaires en situation d'expositions multiples, fortement

corrélées et entachées de sources d'incertitude multiples et hétérogènes. Elle peut donc contribuer, comme illustré dans cette thèse, à améliorer la connaissance des effets sanitaires faisant suite à des expositions conjointes prolongées à faibles doses de RIs (et à d'autres facteurs de risque). Cela est essentiel en termes d'implications en santé publique et en radioprotection. La modélisation hiérarchique fournit un cadre souple et élégant pour la prise en compte simultanée de sources d'incertitude multiples et complexes. Ainsi, en épidémiologie des RIs, une prise en compte conjointe d'expositions radiologiques fortement corrélées et des erreurs de mesure potentielles sur ces expositions, dans l'estimation de risques radio-induits, est envisageable par combinaison de différents sous-modèles associés à différents niveaux de hiérarchie. La prise en compte simultanée (c'est-à-dire dans un même modèle) de ces deux sources de variabilité pourrait avoir un impact non négligeable sur l'estimation des risques radio-induits et sur l'interprétation de l'effet de ces expositions sur les risques estimés. Quant à la statistique bayésienne, elle fournit non seulement une panoplie d'outils génériques - comme les algorithmes MCMC - pour l'apprentissage statistique de modèles complexes tels que les modèles hiérarchiques mais également la possibilité d'introduire des informations *a priori* sur des paramètres pour lesquels l'information disponible dans les données étudiées est relativement faible. Malgré sa grande flexibilité, l'utilisation d'une approche hiérarchique bayésienne ne peut néanmoins se passer ni du choix d'hypothèses de modélisation parcimonieuses ni d'une étape d'analyse de sensibilité et de validation de modèle vis-à-vis des hypothèses de modélisation fixées. Enfin, il est important de noter que les méthodes proposées dans ce travail de thèse sont généralisables à d'autres contextes que la cohorte française des mineurs d'uranium comme à d'autres pathologies que le cancer du poumon, d'autres professions et d'autres variables d'exposition.

Bibliographie

[Agi+16] L. Agier et al. « A systematic comparison of linear regression-based statistical methods to assess exposome-health associations ». In : *Environmental Health Perspectives* 124.12 (2016), p. 1848-56.

[All+12a] Rodrigue S Allodji, Klervi Leuraud et al. « Assessment of uncertainty associated with measuring exposure to radon and decay products in the French uranium miners cohort ». In : *Journal of Radiological Protection* 32.1 (2012), p. 85.

[All+12b] Rodrigue S Allodji, Anne CM Thiébaut et al. « The performance of functional methods for correcting non-Gaussian measurement error within Poisson regression : corrected excess risk of lung cancer mortality in relation to radon exposure among French uranium miners ». In : *Statistics in medicine* 31.30 (2012), p. 4428-4443.

[All11] Setcheou Rodrigue Allodji. « Prise en compte des erreurs de mesure dans l'analyse du risque associe a l'exposition aux rayonnements ionisants dans une cohorte professionnelle : application à la cohorte française des mineurs d'uranium ». Thèse de doct. 2011.

[Arm98] Ben G Armstrong. « Effect of measurement error on epidemiological studies of environmental and occupational exposures. » In : *Occupational and environmental medicine* 55.10 (1998), p. 651-656.

[13] *Base de données des taux de mortalité de référence. IRSN/PRP-HOM/SRBE/LEPID/2013-07.* Institut de Radioprotection et de Sûreté Nucléaire, 2013.

[BKD06] M.L. Bell, J.Y. Kim et F. Dominici. « Potential confounding of particulate matter on the short-term association between ozone and mortality in multisite time-series studies ». In : *Environ Health Perspect* 115.11 (2006), p. 1591-5.

[Ben+20] B. Bennett et al. « Characterizing the neurodevelopmental pesticide exposome in a children's agricultural cohort ». In : *Int J Environ Res Public Health* 17.5 (2020), p. 1479.

[BKZ91] S. Bernhard, G. Kraemer et P. Zettwoog. « La radioprotection dans les mines et usines de minerai d'uranium françaises ». In : *Radioprotection* 26.2 (1991), p. 329-349.

[BG15] Michael Betancourt et Mark Girolami. « Hamiltonian Monte Carlo for hierarchical models ». In : *Current trends in Bayesian methodology with applications* 79 (2015), p. 30.

[Bil+12] C. Billionnet et al. « Estimating the health effects of exposure to multi-pollutant mixture ». In : *Ann Epidemiol* 22.2 (2012), p. 126-41.

[BM05] A Birchall et JW Marsh. « Radon dosimetry and its implication for risk ». In : *International Congress Series*. T. 1276. Elsevier. 2005, p. 81-84.

[Bob+15] J.F. Bobb et al. « Bayesian kernel machine regression for estimating the health effects of multi-pollutant mixtures ». In : *Biostatistics* 16.3 (2015), p. 493-508.

[BR10] L. Bottolo et S. Richardson. « Evolutionary stochastic search for Bayesian model exploration ». In : *Bayesian Analysis* 5 (2010), p. 583-618.

[BVH90] Marie-Hélène Bouvier-Colle, Jacques Vallin et Françoise Hatton. *Mortalité et causes de décès en France*. Doin, 1990.

[Bra85] WJ Brady. *Radiac instruments and film badges used at atmospheric nuclear tests*. Defense Nuclear Agency, 1985.

[Bro01] A. Broadbent. « Conceptual and methodological issues in epidemiology : an overview. » In : *Preventive Medicine* 53.4-5 (2001), p. 215-216.

[Buc+13] G.M. Buck Louis et al. « The exposome -Exciting opportunities for discoveries in reproductive and perinatal epidemiology ». In : *Paediatr Perinat Epidemiol* 27.3 (2013), p. 229-236.

[Car05] Raymond J Carroll. « Measurement Error in Epidemiologic Studies ». In : *Encyclopedia of Biostatistics* 5 (2005).

[Car+06] Raymond J Carroll et al. *Measurement error in nonlinear models : a modern perspective*. CRC press, 2006.

[CG92] George Casella et Edward I George. « Explaining the Gibbs sampler ». In : *The American Statistician* 46.3 (1992), p. 167-174.

[CHR00] G. Celeux, M. Hurn et C.P. Robert. « Computational and inferential difficulties with mixture posterior distributions. » In : *J. Am. Stat. Assoc.* 95 (2000), p. 957-970.

[CG95] Siddhartha Chib et Edward Greenberg. « Understanding the metropolis-hastings algorithm ». In : *The american statistician* 49.4 (1995), p. 327-335.

[Cla+92] DG Clayton et al. « Models for the analysis of cohort and case-control studies with inaccurately measured exposures ». In : *Statistical models for longitudinal studies of health* (1992), p. 301-331.

[Cok+18] E. Coker et al. « Multi-pollutant Modeling Through Examination of Susceptible Subpopulations Using Profile Regression ». In : *Current Environmental Health Reports* 5 (2018), p. 59-69.

[CS94] John R Cook et Leonard A Stefanski. « Simulation-extrapolation estimation in parametric measurement error models ». In : *Journal of the American Statistical association* 89.428 (1994), p. 1314-1328.

[Cou+06] National Research Council et al. *Health risks from exposure to low levels of ionizing radiation : BEIR VII phase 2*. T. 7. National Academies Press, 2006.

[CLR59] ML Crosfill, Patricia J Lindop et J Rotblat. « Variation of sensitivity to ionizing radiation with age ». In : *Nature* 183.4677 (1959), p. 1729.

[Dom17] Haydee Domenech. « Radiation Safety ». In : *Managment and Programs. Suiza : Springer* (2017).

[Dom+10] F. Dominici et al. « Protecting Human Health from Air Pollution : Shifting from a Single-Pollutant to a Multi-pollutant Approach ». In : *Epidemiology* 21.2 (2010), p. 187-194.

[04] « Dosimetry and Biological effects of Ionizing Radiation ». In : *Handbook of Nuclear Chemistry*. Springer US, 2004, p. 1647-1684.

[Eib+89] Jason D Eiband et al. « Prognostic factors in squamous cell carcinoma of the larynx ». In : *The American journal of surgery* 158.4 (1989), p. 314-317.

[Ems11] John Emsley. *Nature's building blocks : an AZ guide to the elements*. Oxford University Press, 2011.

[FG67] Donald E Farrar et Robert R Glauber. « Multicollinearity in regression analysis : the problem revisited ». In : *The Review of Economic and Statistics* (1967), p. 92-107.

[FKN91] Katherine M Flegal, Penelope M Keyl et F Javier Nieto. « Differential misclassification arising from nondifferential errors in exposure measurement ». In : *American Journal of Epidemiology* 134.10 (1991), p. 1233-1246.

[For65] E. Forgy. « Cluster analysis of multivariate data : efficiency vs interpretability of classifications ». In : *Biometrics* 21 (1965), p. 768-769.

[Fou+17] Lucie Fournier et al. *Reconstruction of individual radiation doses in a cohort of french nuclear workers : considering doses under the recording threshold*. 2017.

[GR92] A Gelman et D Rubin. « Inference from Iterative Simulation using Multiple Sequences ». In : *Statistical Science* 7.4 (1992), p. 457-511.

[Gel+13] Andrew Gelman et al. *Bayesian data analysis*. Chapman et Hall/CRC, 2013.

[Gel+14] Andrew Gelman et al. *Bayesian data analysis*. T. 2. CRC press Boca Raton, FL, 2014.

[Geo08] Andreas C George. « World history of radon research and measurement from the early 1900's to today ». In : *AIP Conference Proceedings*. T. 1034. 1. American Institute of Physics. 2008, p. 20-33.

[GFB95] E.S. Gilbert, J. J. Fix et W. V. Baumgartner. « An Approach to Evaluating Bias and Uncertainty in Estimates of External Dose Obtained from Personal Dosimeters. » In : *Health Physics* 70 (1995), p. 336-345.

[GFB96] ES Gilbert, JJ Fix et WV Baumgartner. « An approach to evaluating bias and uncertainty in estimates of external dose obtained from personal dosimeters. » In : *Health physics* 70.3 (1996), p. 336-345.

[Goo07] JW Gooch. *Encyclopedic dictionary of poylmers*. Springer Science + Business Media, 2007.

[GPR99] S. Greenland, J. Pearl et J.M. Robins. « Causal diagrams for epidemiologic research ». In : *Epidemiology* 10 (1999), p. 37-48.

[Gri11] HC Griffin. « Natural radioactive decay chains ». In : *honc* (2011), p. 667.

[Guo08] Annamaria Guolo. « Robust techniques for measurement error correction : a review ». In : *Statistical Methods in Medical Research* 17.6 (2008), p. 555-580.

[HLR15] David I Hastie, Silvia Liverani et Sylvia Richardson. « Sampling from Dirichlet process mixture models with unknown concentration parameter : mixing issues in large data implementations ». In : *Statistics and computing* 25.5 (2015), p. 1023-1037.

[Has+13] DI. Hastie et al. « A Semi-parametric Approach to Estimate Risk Functions Associated with Multi-dimensional Exposure Profiles : Application to Smoking and Lung Cancer. » In : *BMC Medical Research Methodology* 13.1 (2013), p. 129.

[Hei+02] IM Heid et al. « On the potential of measurement error to induce differential bias on odds ratio estimates : an example from radon epidemiology ». In : *Statistics in medicine* 21.21 (2002), p. 3261-3278.

[Hof17] Sabine Hoffmann. « Approche hiérarchique bayésienne pour la prise en compte d'erreurs de mesure d'exposition chronique et à faible doses aux rayonnements ionisants dans l'estimation du risque de cancers radio-induits : Application à une cohorte de mineurs d'uranium ». Thèse de doct. Université Paris-Saclay (ComUE), 2017.

[Hof+18] Sabine Hoffmann, Dominique Laurier et al. « Shared and unshared exposure measurement error in occupational cohort studies and their effects on statistical inference in proportional hazards models ». In : *PloS one* 13.2 (2018), e0190792.

[Hof+17] Sabine Hoffmann, Estelle Rage et al. « Accounting for Berkson and classical measurement error in radon exposure using a Bayesian structural approach in the analysis of lung cancer mortality in the French cohort of uranium miners ». In : *Radiation research* 187.2 (2017), p. 196-209.

[HDR95] Richard W Hornung, James Deddens et Robert Roscoe. « Modifiers of exposure-response estimates for lung cancer among miners exposed to radon progeny. » In : *Environmental health perspectives* 103.suppl 2 (1995), p. 49-53.

[HS96] Geoffrey R Howe et Ron H Stager. « Risk of lung cancer mortality after exposure to radon decay products in the Beaverlodge cohort based on revised exposure estimates ». In : *Radiation research* 146.1 (1996), p. 37-42.

[Hun+13] Nezahat HUNTER et al. « Joint analysis of three European nested case-control studies of lung cancer among radon exposed miners : exposure restricted to below 300 WLM ». In : *Health physics* 104.3 (2013), p. 282-292.

[IAR+88] IARC et al. *Man-made mineral fibres and radon*. Lyon : IARC, 1988.

[ICS14] Joseph G IBRAHIM, Ming-Hui CHEN et Debajyoti SINHA. « B ayesian Survival Analysis ». In : *Wiley StatsRef : Statistics Reference Online* (2014).

[Ion09] Advisory Group on IONISING RADIATION (AGIR). *Radon and Public Health*. Health Protection Authority, 2009.

[IZ00] Hemant ISHWARAN et Mahmoud ZAREPOUR. « Markov chain Monte Carlo in approximate Dirichlet and beta two-parameter process hierarchical models ». In : *Biometrika* 87.2 (2000), p. 371-390.

[Jai+18] P. JAIN et al. « A multivariate approach to investigate the combined biological effects of multiple exposures ». In : *J Epidemiol Community Health* 72 (2018), p. 564-571.

[Jef+92] Richard JEFFREY et al. *Probability and the Art of Judgment*. Cambridge University Press, 1992.

[Jel12] Bjørn Petter JELLE. « Development of a model for radon concentration in indoor air ». In : *Science of the total environment* 416 (2012), p. 343-350.

[Jev05] T JEVREMOVIC. *Nuclear principles in engineering*. Springer US, 2005.

[Jor+04] Michael I JORDAN et al. « Graphical models ». In : *Statistical science* 19.1 (2004), p. 140-155.

[Jur+06] AM. JUREK et al. « Exposure-measurement error is frequently ignored when interpreting epidemiologic study results ». In : *European Journal of Epidemiology* 21 (2006), p. 871-876.

[KRT15] Alexander P Keil, David B Richardson et Melissa A Troester. « Healthy worker survivor bias in the Colorado Plateau uranium miners cohort ». In : *American journal of epidemiology* 181.10 (2015), p. 762-770.

[Keo+20] R.H. Keogh et al. « STRATOS guidance document on measurement error and misclassification of variables in observational epidemiology : Part 1-Basic theory and simple methods of adjustment. » In : *Statistics in Medicine* 39.16 (2020), p. 2197-2231.

[KYB06] Hyang-Mi Kim, Yutaka Yasui et Igor Burstyn. « Attenuation in risk estimates in logistic and Cox proportional-hazards models due to group-based exposure assessment strategy ». In : *The Annals of occupational hygiene* 50.6 (2006), p. 623-635.

[Kle+19] S. Klebe et al. « Asbestos, Smoking and Lung Cancer : An Update ». In : *Int J Environ Res Public Health* 17.1 (2019), p. 258.

[Kre10] Guenter J Krejs. « Gastric cancer : epidemiology and risk factors ». In : *Digestive diseases* 28.4-5 (2010), p. 600-603.

[Kre+13] M Kreuzer, F Dufey et al. « External gamma radiation and mortality from cardiovascular diseases in the German WISMUT uranium miners cohort study, 1946–2008 ». In : *Radiation and environmental biophysics* 52.1 (2013), p. 37-46.

[Kre+17a] M Kreuzer, C Sobotzki et al. « Factors Modifying the Radon-Related Lung Cancer Risk at Low Exposures and Exposure Rates among German Uranium Miners ». In : *Radiation Research* 189.2 (2017), p. 165-176.

[Kre+17b] M. Kreuzer et al. « Leukaemia mortality and low-dose ionising radiation in the WISMUT uranium miner cohort (1946–2013) ». In : *Occupational and environmental medicine* 74.4 (2017), p. 252-258.

[Kre+10] Michaela Kreuzer et al. « Cohort profile : the German uranium miners cohort study (WISMUT cohort), 1946–2003 ». In : *International journal of epidemiology* 39.4 (2010), p. 980-987.

[] *La radioprotection en milieu hospitalier*. https://www.utc.fr/tsibh/public/2tsibh/09/projet/groupe4/index.html. Accessed : 2020-06-30.

[Lan+10] Rachel SD Lane et al. « Mortality (1950–1999) and cancer incidence (1969–1999) in the cohort of Eldorado uranium workers ». In : *Radiation research* 174.6a (2010), p. 773-785.

[Lan+99] Bryan Langholz et al. « Latency analysis in epidemiologic studies of occupational exposures : application to the Colorado Plateau uranium miners cohort ». In : *American journal of industrial medicine* 35.3 (1999), p. 246-256.

[Lau+16] Olivier Laurent et al. « Concerted Uranium Research in Europe (CURE) : toward a collaborative project integrating dosimetry, epidemiology and radiobiology to study the effects of occupational uranium exposure ». In : *Journal of Radiological Protection* 36.2 (2016), p. 319.

[Lec12] JF Lecomte. « Radon and the system of radiological protection ». In : *Annals of the ICRP* 41.3-4 (2012), p. 389-396.

[Len+16] V. Lenters, L. Portengen, A. Rignell-Hydbom et al. « Prenatal Phthalate, Perfluoroalkyl Acid, and Organochlorine Exposures and Term Birth Weight in Three Birth Cohorts : Multi-Pollutant Models Based on Elastic Net Regression. » In : *Environmental health perspectives* 124.3 (2016), p. 365-372.

[Len+15] V. Lenters, L. Portengen, L.A. Smit et al. « Phthalates, perfluoroalkyl acids, metals and organochlorines and reproductive function : a multipollutant assessment in Greenlandic, Polish and Ukrainian men. » In : *Occup Environ Med* 72.6 (2015), p. 385-93.

[Leu+11a] K. Leuraud et al. « Radon, smoking and lung cancer risk : results of a joint analysis of three European case-control studies among uranium miners. » In : *Radiat Res* 176.3 (2011), p. 375-387.

[Leu+07] Klervi Leuraud, Solenne Billon et al. « Lung cancer risk associated to exposure to radon and smoking in a case-control study of French uranium miners ». In : *Health Physics* 92.4 (2007), p. 371-378.

[Leu+15] Klervi Leuraud, David B Richardson et al. « Ionising radiation and risk of death from leukaemia and lymphoma in radiation-monitored workers (INWORKS) : an international cohort study ». In : *The Lancet Haematology* 2.7 (2015), e276-e281.

[Leu+11b] Klervi Leuraud, Maria Schnelzer et al. « Radon, smoking and lung cancer risk : results of a joint analysis of three European case-control studies among uranium miners ». In : *Radiation research* 176.3 (2011), p. 375-387.

[Li+20] N. Li et al. « Associations between long-term exposure to air pollution and blood pressure and effect modifications by behavioral factors ». In : *Environmental Research* 182.109-109 (2020).

[Lin+19] H. Lin et al. « Ambient PM2.5 and O3 and their combined effects on prevalence of presbyopia among the elderly : A cross-sectional study in six low- and middle-income countries ». In : *Sci Total Environ.* 655 (2019), p. 168-173.

[LK81] Ruey S Lin et Irving I Kessler. « A multifactorial model for pancreatic cancer in man : Epidemiologic evidence ». In : *Jama* 245.2 (1981), p. 147-152.

[LLB16] S. Liverani, A. Lavigne et M. Blangiardo. « Modelling Collinear and Spatially Correlated Data. » In : *Spatial and Spatio-temporal Epidemiology* 18 (2016), p. 63-73.

[Liv+15] Silvia Liverani et al. « PReMiuM : An R package for profile regression mixture models using Dirichlet processes ». In : *Journal of statistical software* 64.7 (2015), p. 1.

[Lub+95a] Jay H Lubin, John D Boice Jr, Christer Edling, Richard W Hornung, Geoffrey Howe et al. « Radon-exposed underground miners and inverse dose-rate (protraction enhancement) effects ». In : *Health physics* 69.4 (1995), p. 494-500.

[Lub+95b] Jay H Lubin, John D Boice Jr, Christer Edling, Richard W Hornung, Geoffrey R Howe et al. « Lung cancer in radon-exposed miners and estimation of risk from indoor exposure ». In : *JNCI : Journal of the National Cancer Institute* 87.11 (1995), p. 817-827.

[Lub+04] Jay H Lubin, Joanne S Colt et al. « Epidemiologic evaluation of measurement data in the presence of detection limits ». In : *Environmental health perspectives* 112.17 (2004), p. 1691-1696.

[LK97] Robert H Lyles et Lawrence L Kupper. « A detailed evaluation of adjustment methods for multiplicative measurement error in linear regression with applications in occupational epidemiology ». In : *Biometrics* (1997), p. 1008-1025.

[Mab85] J. Mabile. *Développement de l'industrie minière de l'uranium en France et dans l'Union Française*. Commissariat à l'énergie atomique, 1985.

[Mar+10] James W Marsh et al. « Dose conversion factors for radon : recent developments ». In : *Health Physics* 99.4 (2010), p. 511-516.

[Mar+12] JW Marsh et al. « Dosimetric calculations for uranium miners for epidemiological studies ». In : *Radiation protection dosimetry* 149.4 (2012), p. 371-383.

[Mar01] RJ. Marshall. « The use of classification and regression trees in clinical epidemiology. » In : *J Clin Epidemiol* 54.6 (2001), p. 603-609.

[MGT11] M Marušiaková, Z Gregor et L Tomášek. « A review of exposures to radon, long-lived radionuclides and external gamma at the Czech uranium mine ». In : *Radiation protection dosimetry* 145.2-3 (2011), p. 248-251.

[Mas65] William F Massy. « Principal components regression in exploratory statistical research ». In : *Journal of the American Statistical Association* 60.309 (1965), p. 234-256.

[MK02] C.F. Mela et P.K. Kopalle. « The impact of collinearity on analysis : the asymmetric effect of negative and positive correlations ». In : *Applied Economics* 34 (2002), p. 667-677.

[Mis+18] Munechika Misumi et al. « Simulation extrapolation for bias correction with exposure uncertainty in radiation risk analysis utilizing grouped data ». In : *Journal of the Royal Statistical Society Series C* 67.1 (2018), p. 275-289.

[Mol+10] John Molitor et al. « Bayesian profile regression with an application to the National Survey of Children's Health ». In : *Biostatistics* 11.3 (2010), p. 484-498.

[Mor+98] HI Morrison et al. « Radon-progeny exposure and lung cancer risk in a cohort of Newfoundland fluorspar miners ». In : *Radiation research* 150.1 (1998), p. 58-65.

[Nat06] NRC. National Research Council. *Heath Risks from Exposure to Low Levels of Ionizing Radiation. BEIR VII Phase 2.* Washington, DC : The National Academies Press, 2006.

[Nav+16] Garthika Navaranjan et al. « Cancer incidence and mortality from exposure to radon progeny among Ontario uranium miners ». In : *Occupational and environmental medicine* 73.12 (2016), p. 838-845.

[Nea94] Radford M Neal. « An improved acceptance procedure for the hybrid Monte Carlo algorithm ». In : *Journal of Computational Physics* 111.1 (1994), p. 194-203.

[OSS07] David I Ohlssen, Linda D Sharples et David J Spiegelhalter. « Flexible random-effects models using Bayesian semi-parametric models : applications to institutional comparisons ». In : *Statistics in medicine* 26.9 (2007), p. 2088-2112.

[PR08] Omiros Papaspiliopoulos et Gareth O Roberts. « Retrospective Markov chain Monte Carlo methods for Dirichlet process hierarchical models ». In : *Biometrika* 95.1 (2008), p. 169-186.

[Pap+12] M. Papathomas et al. « Exploring Data from Genetic Association Studies using Bayesian Variable Selection and the Dirichlet Process : Application to Searching for gene × gene patterns. » In : *Genetic Epidemiology* 36.6 (2012), p. 663-674.

[Pap+11] Michail Papathomas et al. « Examining the Joint Effect of Multiple Risk Factors Using Exposure Risk Profiles : Lung Cancer in Nonsmokers ». In : *Environ Health Perspect* 119.1 (2011), p. 84-91.

[PB07] E. Parent et J. Bernier. *Le raisonnement bay´esien. Modélisation et inférence*. Springer, 2007.

[PB10] J. Patel C.J. Bhattacharya et A.J. Butte. « An Environment-Wide Association Study (EWAS) on type 2 diabetes mellitus. » In : *Plos One* 5.5 (2010), e10746.

[PDG02] B.H. Patterson, C.M. Dayton et B.I. Graubard. « Latent class analysis of complex sample survey data : application to dietary data ». In : *Journal of the American Statistical Association* 97 (2002), p. 721-728.

[Phy+07] William L Physick et al. « An approach for estimating exposure to ambient concentrations ». In : *Journal of exposure science & environmental epidemiology* 17.1 (2007), p. 76-83.

[Pir+15] M. Pirani et al. « Analysing the Health Effects of Simultaneous Exposure to Physical and Chemical Properties of Airborne Particles. » In : *Environment International* 79 (2015), p. 56-64.

[PS17] Dale L Preston et Daniel O Stram. « The growth of biostatistics and estimation of cancer risk estimates : past, current, and future challenges ». In : *Radiation Protection Dosimetry* 173.1-3 (2017), p. 32-35.

[Pro07] Radiological Protection. « ICRP publication 103 ». In : *Ann ICRP* 37.2.4 (2007), p. 2.

[Rag+15] E Rage et al. « Mortality analyses in the updated French cohort of uranium miners (1946–2007) ». In : *International archives of occupational and environmental health* 88.6 (2015), p. 717-730.

[Rag+20] Estelle Rage, David B Richardson et al. « PUMA–pooled uranium miners analysis : cohort profile ». In : *Occupational and Environmental Medicine* 77.3 (2020), p. 194-200.

[Rag+12] Estelle RAGE, Blandine VACQUIER et al. « Risk of lung cancer mortality in relation to lung doses among French uranium miners : follow-up 1956–1999 ». In : *Radiation research* 177.3 (2012), p. 288-297.

[Rap+14] S.M. RAPPAPORT, D.K. BARUPAL et al. « The blood exposome and its role in discovering causes of disease ». In : *Environ Health Perspect* 122 (2014), p. 769-774.

[RS10] S.M. RAPPAPORT et M.T. SMITH. « Environment and Disease Risks ». In : *Science* 330 (2010), p. 460-1.

[Ree+98] GK REEVES et al. « Some aspects of measurement error in explanatory variables for continuous and binary regression models ». In : *Statistics in Medicine* 17.19 (1998), p. 2157-2177.

[Ric+14] David B RICHARDSON et al. « Assessment and indirect adjustment for confounding by smoking in cohort studies using relative hazards models ». In : *American journal of epidemiology* 180.9 (2014), p. 933-940.

[RL04] DB RICHARDSON et D LOOMIS. « The impact of exposure categorisation for grouped analyses of cohort data ». In : *Occupational and Environmental Medicine* 61.11 (2004), p. 930-935.

[Ric+02] Sylvia RICHARDSON et al. « Mixture models in measurement error problems, with reference to epidemiological studies ». In : *Journal of the Royal Statistical Society : Series A (Statistics in Society)* 165.3 (2002), p. 549-566.

[RR09] Gareth O ROBERTS et Jeffrey S ROSENTHAL. « Examples of adaptive MCMC ». In : *Journal of Computational and Graphical Statistics* 18.2 (2009), p. 349-367.

[Rob15] R.F. ROBINSON. *Mining and selling radium and uranium*. Springer International Publishing, 2015.

[Rog+02] A ROGEL et al. « Lung cancer risk in the French cohort of uranium miners ». In : *Journal of Radiological Protection* 22.3A (2002), A101.

[Ron98] Elaine Ron. « Ionizing radiation and cancer risk : evidence from epidemiology ». In : *Radiation research* 150.5s (1998), S30-S41.

[Sac+64] G. Saccomanno et al. « Lung cancer of uranium miners on the colorado plateau. » In : *Health Physics* 10 (1964), p. 1195-1201.

[Sac+88] Geno Saccomanno et al. « Relationship of radioactive radon daughters and cigarette smoking in the genesis of lung cancer in uranium miners ». In : *Cancer* 62.7 (1988), p. 1402-1408.

[Sam89] Jonathan M Samet. « Radon and lung cancer ». In : *JNCI : Journal of the National Cancer Institute* 81.10 (1989), p. 745-758.

[Sam+91] Jonathan M Samet et al. « Lung cancer mortality and exposure to radon progeny in a cohort of New Mexico underground uranium miners ». In : *Health physics* 61.6 (1991), p. 745-752.

[SDP09] Mary K Schubauer-Berigan, Robert D Daniels et Lynne E Pinkerton. « Radon exposure and mortality among white and American Indian uranium miners : an update of the Colorado Plateau cohort ». In : *American journal of epidemiology* 169.6 (2009), p. 718-730.

[Sha+20] P.A. Shaw et al. « STRATOS guidance document on measurement error and misclassification of variables in observational epidemiology : Part 2- More complex methods of adjustment and advanced topics. » In : *Statistics in Medicine* 39.16 (2020), p. 2232-2263.

[SV15] R. Slama et M. Vrijheid. « Some challenges of studies aiming to relate the Exposome to human health ». In : *Occup Environ Med* 72.6 (2015), p. 383-384.

[Spi+02] David J Spiegelhalter et al. « Bayesian measures of model complexity and fit ». In : *Journal of the royal statistical society : Series b (statistical methodology)* 64.4 (2002), p. 583-639.

[Sta+03] Leslie Stayner et al. « Attenuation of exposure-response curves in occupational cohort studies at high exposure levels ». In : *Scandinavian journal of work, environment & health* (2003), p. 317-324.

[Ste+15] Kyle Steenland et al. « Attenuation of exposure-response rate ratios at higher exposures : A simulation study focusing on frailty and measurement error ». In : *Epidemiology* 26.3 (2015), p. 395-401.

[SC85] Leonard A Stefanski et Raymond J Carroll. « Covariate measurement error in logistic regression ». In : *The Annals of Statistics* (1985), p. 1335-1351.

[SD11] Matthew A Steliga et Carolyn M Dresler. « Epidemiology of lung cancer : smoking, secondhand smoke, and genetics ». In : *Surgical Oncology Clinics* 20.4 (2011), p. 605-618.

[Str+99] Daniel O Stram et al. « Correcting for exposure measurement error in a reanalysis of lung cancer mortality for the Colorado Plateau Uranium Miners cohort ». In : *Health physics* 77.3 (1999), p. 265-275.

[TSD93] Duncan Thomas, Daniel Stram et James Dwyer. « Exposure measurement error : influence on exposure-disease relationships and methods of correction ». In : *Annual review of public health* 14.1 (1993), p. 69-93.

[Tir+85] M Tirmarche, J Brenot et al. « The present state of an epidemiological study of uranium miners in France ». In : *Proceedings of the International Conference, Occupational Radiation Safety in Mining*. T. 1. Citeseer. 1985, p. 344-349.

[Tir+12a] M Tirmarche, J Harrison et al. « Risk of lung cancer from radon exposure : contribution of recently published studies of uranium miners ». In : *Annals of the ICRP* 41.3-4 (2012), p. 368-377.

[Tir+93] M Tirmarche, A Raphalen et al. « Mortality of a cohort of French uranium miners exposed to relatively low radon concentrations ». In : *British journal of cancer* 67.5 (1993), p. 1090-1097.

[Tir+12b] M. Tirmarche et al. « Risk of lung cancer from radon exposure : contribution of recently published studies of uranium miners. » In : *Annals of the ICRP* 41.3-4 (2012), p. 368-377.

[Tom02] L Tomasek. « Czech miner studies of lung cancer risk from radon ». In : *Journal of Radiological Protection* 22.3A (2002), A107.

[Tom12] Ladislav Tomasek. « Lung cancer mortality among Czech uranium miners—60 years since exposure ». In : *Journal of radiological protection* 32.3 (2012), p. 301.

[Tom+08] Ladislav Tomasek et al. « Lung cancer in French and Czech uranium miners : radon-associated risk at low exposure rates and modifying effects of time since exposure and age at exposure ». In : *Radiation research* 169.2 (2008), p. 125-137.

[Tom+94] Ladislav Tomášek et al. « Patterns of lung cancer mortality among uranium miners in West Bohemia with varying rates of exposure to radon and its progeny ». In : *Radiation research* 137.2 (1994), p. 251-261.

[Tu 04] Gilthorpe M Tu YK Clerehugh V. « Collinearity in linear regression is a serious problem in oral health research ». In : *Eur J Oral Sci* 112 (2004), p. 389-397.

[Uni12] UNSCEAR. United Nations Scientific Committee on the Effects of Atomic Radiation. *Sources, effects and risks of ionizing radiation. UNSCEAR 2012 Report with scientific Annexes A and B*. New York, NY : United Nations, 2012.

[Uni17] UNSCEAR. United Nations Scientific Committee on the Effects of Atomic Radiation. *Sources, effects and risks of ionizing radiation. UNSCEAR 2017 Report with scientific Annexes A and B*. New York, NY : United Nations, 2017.

[Vac+11] Blandine Vacquier, Estelle Rage et al. « The influence of multiple types of occupational exposure to radon, gamma rays and long-lived radionuclides on mortality risk in the French “post-55” sub-cohort of uranium miners : 1956–1999 ». In : *Radiation research* 176.6 (2011), p. 796-806.

[Vac+09] Blandine Vacquier, Agnès Rogel et al. « Radon-associated lung cancer risk among French uranium miners : modifying factors of the exposure–risk relationship ». In : *Radiation and environmental biophysics* 48.1 (2009), p. 1-9.

[Vat+16] K.P. VATCHEVA et al. « Multicollinearity in regression analyses conducted in epidemiologic studies ». In : *Epidemiology* 6.2 (2016), p. 227.

[Vri14] M. VRIJHEID. « The exposome : a new paradigm to study the impact of environment on health ». In : *Thorax (BMJ journals)* 69 (2014), p. 876-878.

[Wag08] Peter WAGGITT. « Uranium mining legacies remediation and renaissance development : an international overview ». In : *Uranium, Mining and Hydrogeology*. Springer, 2008, p. 11-18.

[Wal07] S.G. WALKER. « Sampling the Dirichlet Mixture Model with Slices ». In : *Communications in Statistics - Simulation and Computation* 36.1 (2007), p. 45-54.

[Wal+10] Linda WALSH et al. « The influence of radon exposures on lung cancer mortality in German uranium miners, 1946–2003 ». In : *Radiation research* 173.1 (2010), p. 79-90.

[Wat10] Sumio WATANABE. « Asymptotic equivalence of Bayes cross validation and widely applicable information criterion in singular learning theory ». In : *Journal of Machine Learning Research* 11.Dec (2010), p. 3571-3594.

[Wil05] Christopher Paul WILD. « Complementing the genome with an "exposome" : the outstanding challenge of environmental exposure measurement in molecular epidemiology ». In : *Cancer Epidemiol Biomark Prev* 14 (2005), p. 1847-50.

[Wil12] Christopher Paul WILD. « The exposome : from concept to utility. » In : *Int J Epidemiol* 41 (2012), p. 24-32.

[WEG87] Svante WOLD, Kim ESBENSEN et Paul GELADI. « Principal component analysis ». In : *Chemometrics and intelligent laboratory systems* 2.1-3 (1987), p. 37-52.

[Wol+84] Svante WOLD, Arnold RUHE et al. « The collinearity problem in linear regression. The partial least squares (PLS) approach to generalized inverses ». In : *SIAM Journal on Scientific and Statistical Computing* 5.3 (1984), p. 735-743.

[Wu+15] M. Wu et al. « Modification Effects of Ozone on the Relationship Between PM10 and Daily Mortality in Three Cities, China ». In : *Sci Total Environ* 44.5 (2015), p. 788-812.

[XKS06] Xiaonan Xue, Mimi Y Kim et Roy E Shore. « Estimation of health risks associated with occupational radiation exposure : addressing measurement error and minimum detectable exposure level ». In : *Health physics* 91.6 (2006), p. 582-591.

[Yod+18] Robert C Yoder et al. « Dosimetry for the study of medical radiation workers with a focus on the mean absorbed dose to the lung, brain and other organs ». In : *International journal of radiation biology* (2018), p. 1-12.

[Zab12] Jovanny Zabaleta. « Multifactorial etiology of gastric cancer ». In : *Cancer Epigenetics*. Springer, 2012, p. 411-435.

[Zab+18] Lydia B Zablotska et al. « Analysis of mortality in a pooled cohort of Canadian and German uranium processing workers with no mining experience ». In : *International archives of occupational and environmental health* 91.1 (2018), p. 91-103.

[Zee81] PR Zeetwoog. « State-of-the-art of the α individual dosimetry in France ». In : *Radiation hazards in mining : control, measurement, and medical aspects*. 1981.

[Zha+17] Z. Zhang et al. « Correction of confidence intervals in excess relative risk models using Monte Carlo dosimetry systems with shared errors. » In : *PLoS One.* 12.4 (2017).

Annexes

A Résultats pour les modèles de mélange RPRM de régression de profil bayésien en supposant 5, 6 et 7 groupes non-vides

Les figures A.1, A.3 et A.5 montrent le nombre de mineurs d'uranium français (en haut à gauche), le nombre de décès par cancer du poumon (en bas à gauche) et l'excès de risque instantané de décès par cancer du poumon (β) dans chaque groupe, en ajustant un modèle bayésien RPRM supposant respectivement 5, 6 et 7 groupes non-vides. Les boîtes représentent les trois quartiles (1^{er} quartile, médiane et $3^{\text{ème}}$ quartile) de la distribution *a posteriori* de β et les moustaches des boxplots montrent l'intervalle de crédibilité à 95% *a posteriori* de β pour chaque groupe. La ligne horizontale noire est à 0, ce qui indique une absence d'excès de risque instantané. Les boxplots rouges montrent les groupes avec un excès de risque instantané de décès par cancer du poumon significatif (c'est-à-dire que l'intervalle de crédibilité à 95% associé est supérieur à 0), et les boxplots bleus montrent les groupes sans excès de risque instantané significatif (c'est-à-dire que l'intervalle de crédibilité à 95% associé contient 0).

Les figures A.2, A.4 et A.6 permettent de caractériser les profils d'exposition

associés à chaque groupe, en ajustant un modèle RPRM bayésien supposant respectivement 5, 6 et 7 groupes non-vides. Chaque colonne est associée à une variable d'exposition d'intérêt, lors de l'estimation du risque de décès par cancer du poumon lié aux rayonnements ionisants dans la cohorte française des mineurs d'uranium. De gauche à droite : exposition aux rayonnements γ (en milliSieverts), exposition au radon (en WLM), exposition à la poussière d'uranium (en becquerel par heure par mètre cube), type de poste, type de mine (c'est-à-dire mine sédimentaire située dans l'Hérault par rapport aux autres mines de granit situées en France métropolitaine), âge à la première exposition (en années) et durée d'exposition (en années). Les types de poste sont les suivants : 1) foreur avant mécanisation, 2) foreur après mécanisation, 3) autres travaux souterrains avant mécanisation, 4) autres travaux souterrains après mécanisation et 5) travaux de surface. Les boîtes représentent les trois quartiles (1^{er} quartile, médiane et $3^{\text{ème}}$ quartile) de la distribution *a posteriori* des paramètres définissant la distribution de probabilité suivie par chaque variable dans chaque groupe et les moustaches des boxplots représentent les intervalles de crédibilité à 95% *a posteriori* associés. La ligne horizontale noire sur chaque sous-graphique d'un paramètre donné θ représente $\tilde{\theta}$ qui est la médiane des médianes *a posteriori* de tous les groupes. Les couleurs des boxplots permettent de comparer les valeurs des paramètres dans les différents groupes. En fait, lorsque l'intervalle de crédibilité à 95% d'un paramètre donné dans un groupe donné contient la valeur $\tilde{\theta}$, la boîte est bleue. Lorsque l'intervalle crédible est supérieur à $\tilde{\theta}$ alors le boxplot est rouge, et s'il est inférieur à $\tilde{\theta}$ alors le boxplot est vert. Les groupes sont classées par ordre croissant de la médiane *a posteriori* de chaque coefficient de risque $\beta_c, c \in \{1, C_{max}\}$ de décès par cancer du poumon dans la cohorte française des mineurs d'uranium.

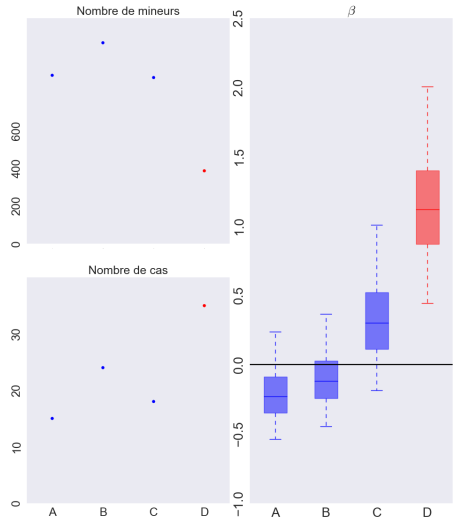

FIGURE A.1 – Nombre de mineurs d'uranium français (en haut à gauche), nombre de décès par cancer du poumon (en bas à gauche) et excès de risque instantané de décès par cancer du poumon (β) dans chaque groupe (à droite), en ajustant un modèle bayésien RPRM supposant 5 groupes non-vides de la cohorte française des mineurs d'uranium. Le groupe comprenant les mineurs non exposés n'est pas affichée.

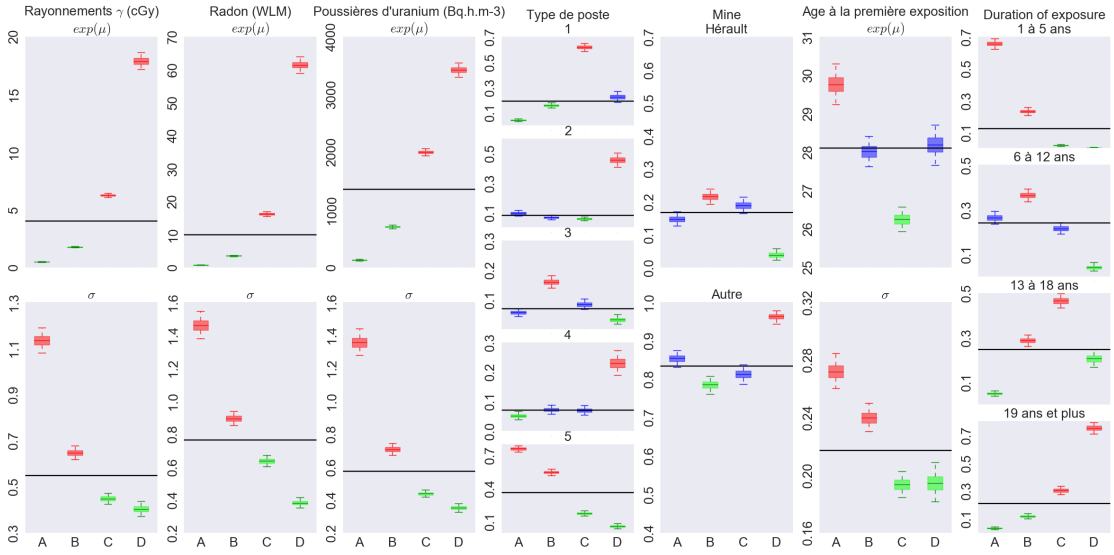

FIGURE A.2 – Caractérisation des profils d'exposition associés à chaque groupe, dans le cadre d'un modèle RPRM bayésien supposant 5 groupes non-vides, le groupe comprenant les mineurs non exposés n'étant pas affiché.

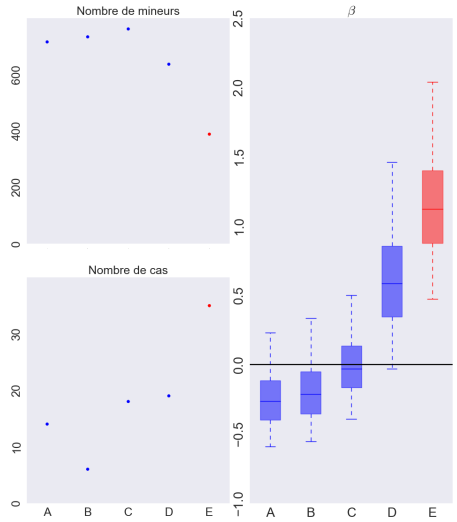

FIGURE A.3 – Nombre de mineurs d'uranium français (en haut à gauche), nombre de décès par cancer du poumon (en bas à gauche) et excès de risque instantané de décès par cancer du poumon (β) dans chaque groupe (à droite), en ajustant un modèle bayésien RPRM supposant 6 groupes non-vides de la cohorte française des mineurs d'uranium. Le groupe comprenant les mineurs non exposés n'est pas affichée.

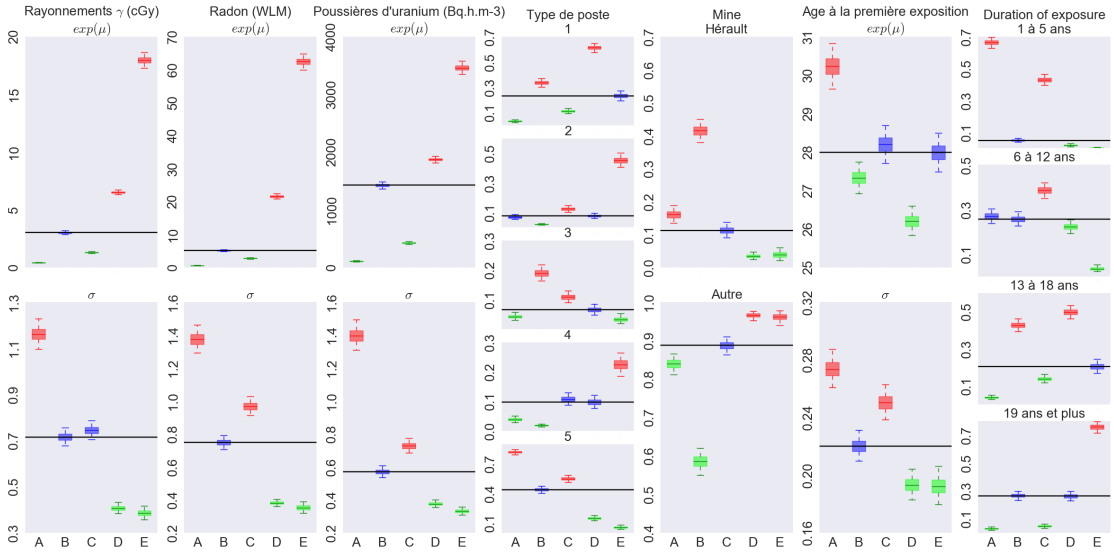

FIGURE A.4 – Caractérisation des profils d'exposition associés à chaque groupe, dans le cadre d'un modèle RPRM bayésien supposant 6 groupes non-vides, le groupe comprenant les mineurs non exposés n'étant pas affiché.

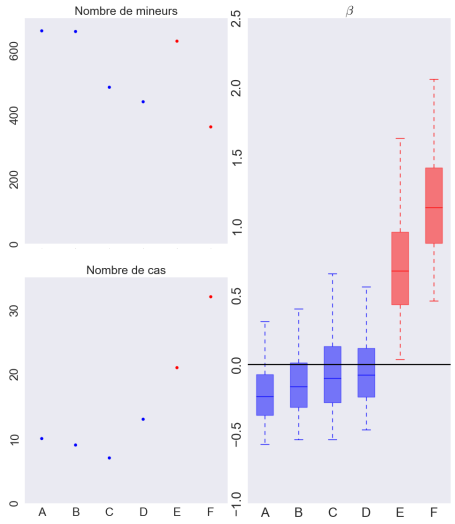

FIGURE A.5 – Nombre de mineurs d'uranium français (en haut à gauche), nombre de décès par cancer du poumon (en bas à gauche) et excès de risque instantané de décès par cancer du poumon (β) dans chaque groupe (à droite), en ajustant un modèle bayésien RPRM supposant 7 groupes non-vides de la cohorte française des mineurs d'uranium. Le groupe comprenant les mineurs non exposés n'est pas affichée.

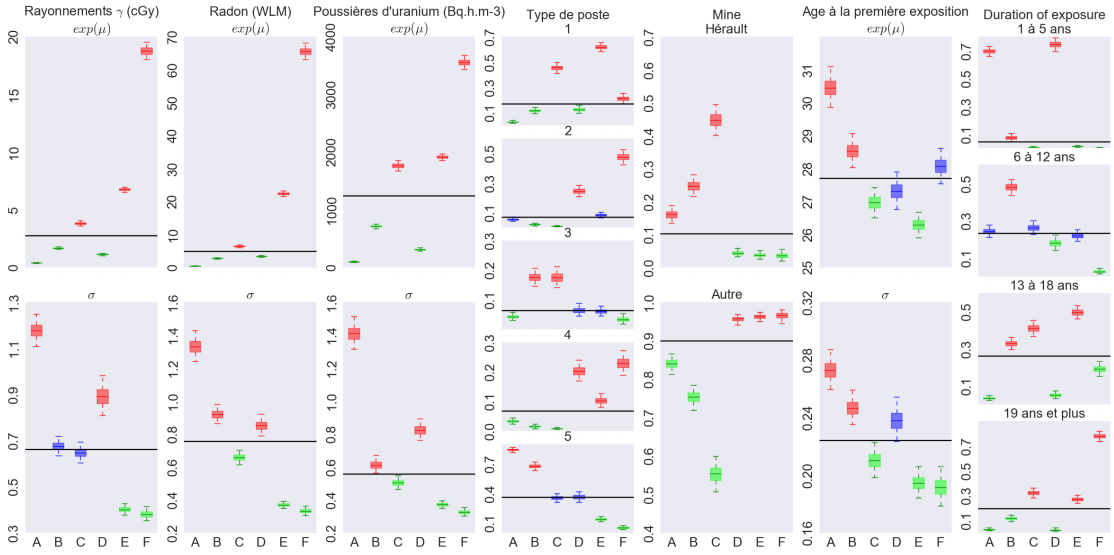

FIGURE A.6 – Caractérisation des profils d'exposition associés à chaque groupe, dans le cadre d'un modèle RPRM bayésien supposant 7 groupes non-vides, le groupe comprenant les mineurs non exposés n'étant pas affiché.

B Indicateurs de performance pour les trois scénarios de simulations, pour un nombre de groupes non-vides fixé à 3, 4 et 5

	$Biais_A^{abs}$	$Biais_B^{rel}$	$Biais_C^{rel}$	π_{BC}	π_{BC-HR}	π_{MC-FP}	π_{MC-FN}
3 groupes non-vides							
$\mathcal{S}_1$	1.29	0.09	-0.2	0.66	1.00	1.00	0.00048
	[0.71 ; 1.89]	[-0.20 ; 0.24]	[-0.30 ; -0.05]	[0.65 ; 0.68]	[1.00 ; 1.00]	[1.00 ; 1.00]	[0.00048 ; 0.00048]
$\mathcal{S}_2$	0.34	0.31	-0.22	1.00	1.00	0.00	0.00092
	[-0.31 ; 2.99]	[-0.63 ; 0.64]	[-0.35 ; 0.11]	[0.65 ; 1.00]	[1.00 ; 1.00]	[0.00 ; 1.00]	[0.00046 ; 0.00094]
$\mathcal{S}_3$	0.75	0.05	-0.19	0.67	1.00	1.00	0.00092
	[0.12 ; 1.18]	[-0.26 ; 0.32]	[-0.34 ; 0.05]	[0.65 ; 0.89]	[1.00 ; 1.00]	[0.00 ; 1.00]	[0.00046 ; 0.00094]
4 groupes non-vides							
$\mathcal{S}_1$	0.23	-0.01	-0.02	0.98	0.98	0.03	0.00047
	[-0.11 ; 0.70]	[-0.21 ; 0.18]	[-0.14 ; 0.12]	[0.98 ; 0.99]	[0.98 ; 0.99]	[0.02 ; 0.04]	[0.00046 ; 0.00047]
$\mathcal{S}_2$	0.00	0.00	-0.01	1.00	1.00	0.00	0.00046
	[-0.37 ; 0.51]	[-0.21 ; 0.26]	[-0.19 ; 0.16]	[1.00 ; 1.00]	[1.00 ; 1.00]	[0.00 ; 0.00]	[0.00046 ; 0.00047]
$\mathcal{S}_3$	0.17	-0.03	-0.02	0.98	0.99	0.03	0.00046
	[-0.31 ; 0.56]	[-0.28 ; 0.32]	[-0.23 ; 0.25]	[0.98 ; 0.99]	[0.98 ; 0.99]	[0.02 ; 0.04]	[0.00045 ; 0.00047]
5 groupes non-vides							
$\mathcal{S}_1$	0.34	-0.01	-0.01	0.98	0.98	0.03	0.00139
	[-0.04 ; 0.87]	[-0.21 ; 0.19]	[-0.13 ; 0.12]	[0.83 ; 0.99]	[0.51 ; 0.99]	[0.01 ; 0.49]	[0.00046 ; 0.00282]
$\mathcal{S}_2$	0.03	0.01	-0.02	1.00	1.00	0.00	0.00141
	[-0.31 ; 0.62]	[-0.20 ; 0.41]	[-0.27 ; 0.17]	[0.83 ; 1.00]	[0.50 ; 1.00]	[0.00 ; 0.51]	[0.00046 ; 0.00277]
$\mathcal{S}_3$	0.22	-0.01	-0.01	0.98	0.98	0.03	0.0014
	[-0.25 ; 0.67]	[-0.28 ; 0.36]	[-0.22 ; 0.26]	[0.84 ; 0.98]	[0.52 ; 0.99]	[0.01 ; 0.48]	[0.00046 ; 0.00322]

TABLE B.1 – Indicateurs de performances médians et leurs intervalles de variabilité à 90% pour les différents scénarios de simulation et pour les modèles estimés à 3, 4 et 5 groupes

C Articles issus de la thèse

C.1 A Bayesian hierarchical approach to account for left-censored and missing radiation doses prone to classical measurement error when analyzing lung cancer mortality due to γ-ray exposure in the French cohort of uranium miners

Radiation and Environmental Biophysics (2020) 59:423–437
https://doi.org/10.1007/s00411-020-00859-6

ORIGINAL ARTICLE

A Bayesian hierarchical approach to account for left-censored and missing radiation doses prone to classical measurement error when analyzing lung cancer mortality due to γ-ray exposure in the French cohort of uranium miners

M. Belloni[1] · C. Guihenneuc[2] · E. Rage[1] · S. Ancelet[1]

Received: 13 September 2019 / Accepted: 13 June 2020 / Published online: 22 June 2020

Abstract
Epidemiological data on cohorts of occupationally exposed uranium miners are currently used to assess health risks associated with chronic exposure to low doses of ionizing radiation. Nevertheless, exposure uncertainty is ubiquitous and questions the validity of statistical inference in these cohorts. This paper highlights the flexibility and relevance of the Bayesian hierarchical approach to account for both missing and left-censored (i.e. only known to be lower than a fixed detection limit) radiation doses that are prone to measurement error, when estimating radiation-related risks. Up to the authors' knowledge, this is the first time these three sources of uncertainty are dealt with simultaneously in radiation epidemiology. To illustrate the issue, this paper focuses on the specific problem of accounting for these three sources of uncertainty when estimating the association between occupational exposure to low levels of γ-radiation and lung cancer mortality in the post-55 subcohort of French uranium miners. The impact of these three sources of dose uncertainty is of marginal importance when estimating the risk of death by lung cancer among French uranium miners. The corrected excess hazard ratio (EHR) is 0.81 per 100 mSv (95% credible interval: [0.28; 1.75]). Interestingly, even if the 95% credible interval of the corrected EHR is wider than the uncorrected one, a statistically significant positive association remains between γ-ray exposure and the risk of death by lung cancer, after accounting for dose uncertainty. Sensitivity analyses show that the results obtained are robust to different assumptions. Because of its flexible and modular nature, the Bayesian hierarchical models proposed in this work could be easily extended to account for high proportions of missing and left-censored dose values or exposure data, prone to more complex patterns of measurement error.

Keywords Measurement errors · Censored data · Bayesian inference · Survival models · Lung cancer · Gamma radiation

Introduction

Epidemiological data on cohorts of occupationally exposed uranium miners are currently used to assess health risks associated with chronic exposure to low levels of ionizing radiation (IR) like internal exposure to radon and its decay products (Clement et al. 2010; Vacquier et al. 2011; Rage et al. 2015) or external exposure to γ-rays (Vacquier et al. 2011; Kreuzer et al. 2013; Rage et al. 2015). These cohorts generally provide valuable information about the individual exposure history of workers to one or several radiological sources but also about other potential risk factors that might modify the exposure-risk relationships of interest (Kreuzer et al. 2017). Similarly, other cohorts of nuclear workers have been set up to study the health effects of external IR exposure (Leuraud et al. 2015).

In any occupational cohort study, exposure uncertainty, which refers to a lack of knowledge about the "true" value of occupational exposures, is inevitably present. This uncertainty arises from different reasons, called sources of uncertainty, including, for example: exposure measurement errors, detection limit of measurement devices, and missing dose values. Exposure measurement error arises whenever an

✉ M. Belloni
marion.belloni@irsn.fr

[1] PSE-SANTE/SESANE/LEPID, Institut de Radioprotection et de Sûreté Nucléaire, Fontenay-aux-Roses, France

[2] UR 7537, Faculté de Pharmacie de Paris, Université de Paris, Paris, France

exposure cannot be measured accurately and only an imperfect surrogate exposure measurement is available. This latter source of uncertainty refers to the discrepancy between a surrogate exposure value (i.e., observed exposure value) and a true (and unknown) exposure value, regardless of the reasons for this discrepancy. Epidemiological studies on occupational cohorts are mainly characterized by complex patterns of exposure measurement error (Allodji et al. 2012; Hoffmann et al. 2017). First, the type and magnitude of error can change over time and space depending on the methods of exposure assessment. Moreover, methods of group-level exposure estimation (e.g., job-exposure matrix) may give rise to errors which are shared between workers belonging to the same group or shared within workers, i.e. errors that affect all exposure values received for several years by the same worker in the same way (Hoffmann et al. 2018). Additionally, in cohorts of uranium miners, exposure data may be censored. This is especially the case when exposure measurements are based on measuring devices, like radiation dosimeters, that cannot measure a radiation exposure which is lower than a fixed threshold, called detection limit (DL). In this context, exposure data are said to be deterministically left-censored: they can be either measured or only known to be lower than a given DL. The DL, which denotes the smallest exposure value that can be detected by a measuring device, mainly depends on the type of the device (Fournier 2017). In cohorts of uranium miners or nuclear workers where exposure measurements were based on radiation dosimeters, the DL usually decreased over the calendar periods due to an improved technology of the measuring devices. It is important to note that left-censored exposure data may also be prone to measurement error. Indeed, a true exposure value slightly higher (resp. lower) than the DL can be wrongly reported as lower (resp. higher) than the DL due to measurement error. Finally, in cohorts of uranium miners or nuclear workers, exposure data can be missing. First, they can have been lost. Moreover, they can be missing because uranium miners were expatriated, which means that the miners worked in a mine abroad and, consequently, were potentially exposed to IR. In the case of French expatriated miners, dosimetry data were not always retrieved in France and miners have missing exposure data for the time of their expatriation. Additionally, exposure data can be missing if the miner was not exposed and so, did not wear his dosimeter. This could be the case when he was on sick leave for example. Finally, if the available exposure data is the cumulative exposure over a year, it is highly probable that a miner did not wear his dosimeter for a few days. The associated missing values can result in an underestimation of his cumulative exposure.

Exposure uncertainty questions the validity of statistical inference in occupational cohort studies in radiation epidemiology (Thomas et al. 1993; Kim et al. 2006; Physick et al. 2007). When it is not or only poorly accounted for, exposure measurement error may cause bias in health risk estimates, a distortion of the exposure-risk relationship and a loss of statistical power (Carroll 2005; Carroll et al. 2006). Actually, its exact consequences on statistical inference depend on the magnitude and the type of error (e.g., classical/Berkson, shared/unshared, differential/non differential, …) but also on the type of disease model (e.g., Poisson regression, logistic regression, survival models, …) (Stefanski and Carroll 1985; Flegal et al. 1991; Reeves et al. 1998; Heid et al. 2002; Richardson and Loomis 2004; Hoffmann et al. 2018). Despite its deleterious consequences and despite its ubiquity in observational research, exposure measurement error is only rarely accounted for in the estimation of risk coefficients in radiation epidemiology (Kreuzer et al. 2013; Leuraud et al. 2015; Rage et al. 2015; Zablotska et al. 2018). One of the main reasons why measurement error is often discussed, but rarely accounted for in radiation epidemiology, may be that dealing with complex patterns of occupational exposure uncertainty implies to use sophisticated statistical approaches for which no user-friendly software exists. In radiation epidemiology, a single imputation method is often used to replace left-censored exposure data by zero, the DL or half of the DL of the dosimeters used (Ron 1998; Gilbert et al. 2006; Laurent et al. 2016; Yoder et al. 2018). Even if the DL is small, especially in the latest periods of exposure, this naive approach may cause both a substantial underestimation of the variability of exposure and a possible under or over estimation of the cumulative exposure received by a worker over time (Lubin et al. 2004; Xue et al. 2006), which is the main covariate of interest when modelling exposure-risk relationships in radiation epidemiology. Moreover, in cases where an association exists between cumulative exposure and a given risk, this approach may cause bias in risk estimates (Lubin et al. 2004; Xue et al. 2006).

Up to the authors' knowledge, the issue of missing and left-censored exposure data prone to measurement error has never been dealt with simultaneously in radiation epidemiology. The censoring process, measurement error and missing exposure are generally accounted for separately. Xue et al. (2006) studied the impact of classical measurement errors and left-censored radiation exposure data due to DL on all-cause death risk estimates. Using a Cox model, they found that, when there is a true association between exposure and risk, the risk estimate is biased towards zero and the statistical power decreases as the magnitude of measurement error and the proportion of censored exposure data increase. However, when there is no association between exposure and risk, the risk estimate is unbiased neither by measurement error nor by censoring whatever the level of measurement error and the proportion of censored exposure data are. Hoffmann et al. (2017) proposed a Bayesian hierarchical approach to account for exposure measurement

error when estimating the risk of lung cancer associated with occupational radon exposure in the French cohort of uranium miners (Vacquier et al. 2008; Rage et al. 2015). They illustrated that the Bayesian hierarchical approach is arguably a very flexible approach to deal with complex patterns of measurement error and multiple sources of uncertainty in occupational cohort studies in radiation epidemiology. Nevertheless, they considered zero and missing exposure data recorded in the database as the true mark of no exposure to radon.

In this paper, three Bayesian hierarchical models are described, to account for both missing and left-censored exposure data prone to classical measurement error when estimating health risks. It is assumed that the censoring indicators may be wrong due to exposure measurement error. Thanks to its well-known flexibility, using a Bayesian hierarchical approach made it possible to describe several sources of exposure uncertainty (i.e., missing exposures, left-censored exposures and measurement errors) and to estimate jointly all unknown quantities, including risk estimates and "true" exposures, in a unique model.

To illustrate the point, this paper focuses on the specific problem of accounting for the three above sources of uncertainty when estimating the association between occupational exposure to low levels of γ-radiation and lung cancer mortality in the French cohort of uranium miners. This association was previously estimated in Rage et al. (2015) by fitting a Poisson regression model without accounting for the measurement errors related to occupational γ-ray exposure (Allodji 2011). Moreover, all the zeros and missing personal dose equivalents due to external γ-ray exposure recorded in the database were considered by default as true zeros. An excess relative risk of 0.74 per 100 mSv (95% CI: 0.23–1.73) was estimated when using the internal sub-cohort of non-exposed miners as a reference. A further aim of the present study was then to test the robustness of this result when accounting for these different sources of uncertainty. Note that external γ-ray exposures were recorded individually in the database, with personal dose equivalents. That is why "dose" uncertainty us used hereafter, instead of exposure uncertainty, although this work can be more generally applied to exposure data.

Materials and methods

Study population

The study population is a sub-cohort of the French cohort of uranium miners. The characteristics, sources of data and methods of data collection (e.g., vital status, causes of death,…) of this retrospective occupational cohort were described previously (Rage et al. 2015). Briefly, the last update included 5086 males who were employed as uranium miners for at least 1 year in the CEA-COGEMA group between 1946 and 1990 and who were followed from 1946 to December 31, 2007.

The routine recording of external γ-ray exposures by individual dosimeters only began in 1956 in the French mines, following the introduction of radiation protection measures like the introduction of forced ventilation. In this paper, the study population is thus restricted to the so-called post-55 sub-cohort which includes 3377 miners from the original cohort who were first employed after December 31, 1955. At the end of follow-up, 94 miners had died of lung cancer. An age limitation of 85 years for follow-up is fixed due to the imprecision in determining the exact cause of death in those occurring after the 85th birthday (Bouvier-Colle et al. 1990). Main characteristics of the post-55 sub-cohort are shown in Table 1 (Rage et al. 2015).

γ-ray dose measurements

Personal dose equivalents (simply called "doses" hereafter) due to external γ-ray exposure were recorded individually using two different types of personal dosimeters, depending on the calendar period: personal film badge dosimeters (CEA PS1 type) from 1956 to 1985 and personal thermoluminescence dosimeters (TLDs) integrated to the individual system of integrated dosimetry (ISID) from 1986 onwards. The DL of TLDs and film badge dosimeters were extracted from Allodji (2011). The DL of TLDs was 0.55 mSv and the DL of film badge dosimeters was 2.2 mSv. Owing to technological progress, the TLDs were then more accurate than film badge dosimeters. Moreover, it is reasonable to assume that the variance of measurement error was lower after 1986

Table 1 Main characteristics of the post-55 French sub-cohort of uranium miners

No. of miners	3377
Age at entry into study, mean [min, max]	28.3 [16.9, 57.7]
Duration of work in years, mean [min, max]	16.7 [1.0, 40.9]
Duration of follow-up in years, mean [min, max]	32.8 [0.1, 51.0]
Vital status, *n* (%)	
Alive < 85 years old	2412 (71.4)
Alive ≥ 85 years old	74 (2.2)
Death from lung cancer	94 (2.8)
Death from another cause	777 (23.0)
Lost to follow-up	20 (0.6)
External γ-ray exposure*	
Exposed miners, *n* (%)	3240 (95.9)
Duration of exposure (in years), mean [min, max]	13.2 [1.0, 36.0]
Cumulative exposure (in mSv), mean [min, max]	54.9 [0.2, 470.1]

*Results only on measured γ-ray exposures

given that the personal dosimeters were changed in 1986 and provided more precise γ-ray dose measurements. The annual individual dose values, expressed in mSv, were computerized from paper archives for the entire study period for the purpose of the epidemiological study.

Figure 1 displays the boxplot of annual personal dose equivalents from γ-ray exposure for the years 1956–2007, for exposed miners of the post-55 sub-cohort. The annual means of exposed miners increased between 1956 and 1961 and decreased between 1961 and 2007. Moreover, the distribution of γ-ray doses strongly shrank from 1986 meaning that the variance of γ-ray doses strongly decreased from this year on. This may be due to changes in personal dosimeters making γ-ray dose values more precise from 1986 onwards, due to the use of personal TLDs.

Zero and missing γ-ray dose measurements

In the post-55 sub-cohort of French uranium miners, 7.60% of γ-ray dose values are recorded as zero values while 2.94% of these values are recorded as missing values. Nevertheless, some of these zero and missing values can reasonably be assumed to be strictly positive, even if potentially close to zero. This can be seen as a specific type of measurement error that was created during the import process of the dose values to the database. In this context, a first step of correction of this measurement error consisted in classifying these zero and missing γ-ray doses in: (a) true zeros; (b) γ-ray doses that were so close to zero that the dosimeter could not measure them (reported as false zeros or missing values in the database); typically, this corresponds to dose values that were strictly positive but smaller than the DL of the dosimeter used; (c) γ-ray doses that were strictly positive, without any restriction. During their occupational activity, miners were simultaneously exposed to radon gas (and its short-lived progeny), external γ-rays and long-lived radionuclides (LLR) of uranium ore dust (Vacquier et al. 2011). To make relevant assumptions about zero and missing γ-ray exposures in the post-55 sub-cohort of French uranium miners, the radiological co-exposure of miners as well as their job position and exposure history were used. Table 2 summarizes the assumptions made regarding zero and missing γ-ray doses in the post-55 French sub-cohort of uranium miners. It also indicates the relative frequency (expressed in percentage) of γ-ray doses in each category. In the following paragraphs, the different assumptions regarding zero and missing γ-ray doses are justified, according to the available information. These assumptions will be used in the Bayesian hierarchical models described thereafter.

The specific situation of radiological co-exposure of uranium miners was used to differentiate true zeros (i.e., no true exposure to external γ-rays) and false zeros. On the one hand, any zero γ-ray dose that was associated to zero exposure to radon and LLR in the database was assumed to be a true zero. On the other hand, any zero γ-ray dose that was associated to strictly positive exposure values to radon or LLR in the database was assumed to be a false zero, which represents the most frequent category (4.85%) in Table 2. Each false zero was assumed to be strictly positive but smaller than the DL of the dosimeter used.

Regarding missing γ-ray doses, two situations were distinguished depending on the availability of information about radiological co-exposures to radon and LLR. If information on the radiological co-exposure of a given uranium miner was available, the following assumptions were made:

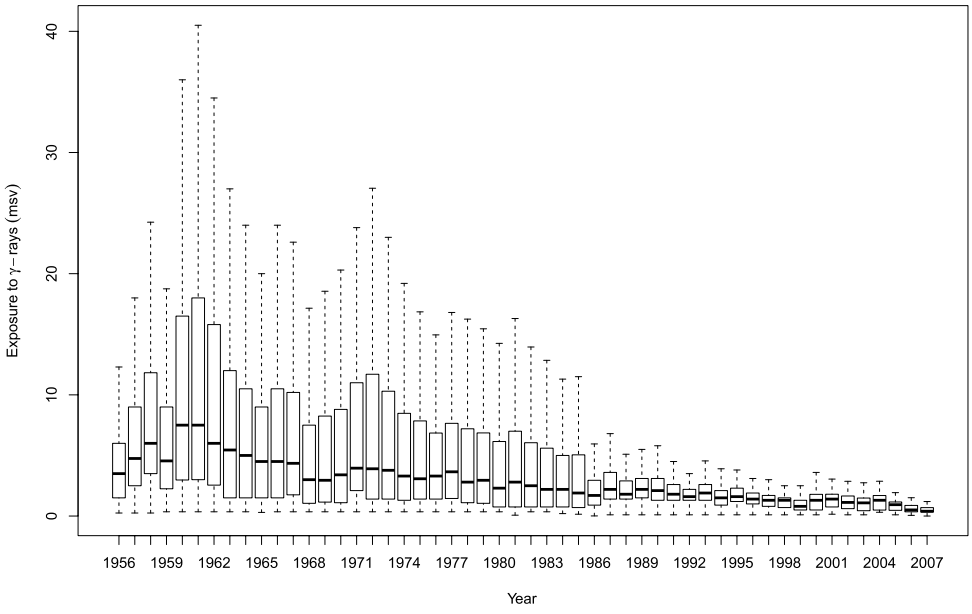

Fig. 1 Boxplot of annual doses (i.e., personal dose equivalents) from γ-ray exposure (in mSv) for the years 1956–2007 in the post-55 French sub-cohort of uranium miners. Non-exposed uranium miners were excluded

Table 2 Summary of assumptions about zero and missing γ-ray dose values made for the post-55 sub-cohort of French uranium miners

γ-ray dose value recorded	Co-exposure measurement	Miner's job position	Assumptions regarding γ-ray dose value	Relative frequency of γ-ray dose values (in %)
0	Radon = 0 and LLR = 0	Regular	0	2.75
	Radon > 0 or LLR > 0	Regular	< DL**	4.85
NA*	Radon = 0 and LLR = 0 Radon = NA* and LLR = 0 Radon = 0 and LLR = NA*	Regular	0	0.50
	Radon > 0 or LLR > 0	Regular	< DL**	0.45
	Radon = NA* and LLR = NA*	Regular	< DL**	1.41
		Expatriate	> 0	0.30
		Other***	= 0	0.28

*Missing dose value

**Denotes the detection limit of the dosimeter

***Denotes any miner's job position that is different from "Regular position" or "Expatriate". "Military service", "medical leave" or "fired" are non-exhaustive examples of other miner's job positions

a) Missing γ-ray doses were assumed to be equal to zero when the exposures to radon and LLR were equal to zeros;
b) Missing γ-ray doses were assumed to be equal to zero when the exposure to radon (resp. LLR) was missing and the exposure to LLR (resp. radon) was equal to zero;
c) Missing γ-ray doses were assumed to be strictly positive measured values but lower than the DL when the exposures to radon or LLR were strictly positive. Indeed, in this case, the associated values of radon and LLR co-exposures were mainly lower than the first quartile of their respective distribution suggesting low values of γ-ray exposures. More elaborate assumptions could have been made using the correlation between the three radiological sources but, given that this category only concerned 0.45% of γ-ray dose values, this simple assumption was used.

If information on the radiological co-exposure of a given uranium miner was not available at a given time t (i.e., the annual exposures to radon, LLR and γ-ray were all missing at time t) but the annual exposure to γ-rays of this miner was strictly positive for at least 1 year before t and 1 year after t, then the job position indicated in the administrative files of this uranium miner was used to distinguish three situations and the following assumptions were made:

a) If the miner's job position at time t was recorded as "regular", his missing γ-ray dose value at time t was assumed to be a strictly positive value but lower than the DL. Indeed, in this case, it was assumed to be unlikely that a miner in a "regular" job position and who was at least one time exposed to γ-rays before year t or after year t was not exposed (i.e., γ-ray dose truly equal to zero) at time t. Moreover, given that the three co-exposures were missing, the most probable reason is that the three co-exposures that were measured by different devices were all lower than the respective DL of these devices.
b) If the miner's job position at time t was recorded as "expatriate", the individual was assumed to have been exposed as a uranium miner in a foreign mine and his missing γ-ray dose value at time t was assumed to be strictly positive but not necessarily lower than the DL since the exposure conditions to γ-rays are unknown in foreign mines. It was simply assumed that the exposure conditions were similar as in the post-55 sub-cohort of French uranium miners. This means that the missing γ-ray dose values for miners who were expatriated were assumed to follow the same probability distribution (with same parameters) as the one used to describe the strictly positive γ-ray dose values in the post-55 sub-cohort of French uranium miners. There was then no restriction to the exposure range.
c) If the miner's job position indicates that the miner was not working in a mine for a while (for example, "military service", "medical leave" or "fired"), his γ-ray dose was reasonably assumed to be a true zero. Note that if the situation of radiological co-exposure of a given uranium miner was not available at a given time t and his annual exposure to γ-rays was never strictly positive before and after t then it was simply assumed that the miner was not exposed to γ-rays at time t.

Model formulation

Several Bayesian hierarchical models are proposed here to account for left-censored and missing dose values prone to measurement error when modelling the association between external γ-ray exposure and lung cancer mortality in the post-55 sub-cohort of French uranium miners.

Strongly inspired by the Bayesian hierarchical model that was proposed to estimate a corrected risk of death by lung cancer due to radon exposure in the French cohort of uranium miners (Hoffmann et al. 2017), these models are composed of three sub-models:

- The disease sub-model: this model relates the outcome of interest of each miner to his true cumulative and occupational γ-ray dose;
- The measurement sub-model: this model describes the association between the true (and unknown) annual γ-ray dose and the measured annual γ-ray dose;
- The exposure sub-model: this model describes the probability distribution of the true (and unknown) annual γ-ray dose in the post-55 sub-cohort.

Following Hoffmann et al. (2017), it is assumed that the measurement error related to γ-ray exposure is non-differential in the post-55 sub-cohort of French uranium miners. Thus, the disease sub-model, measurement sub-model and exposure sub-model are linked via conditional independence assumptions (Richardson and Gilks 1993).

Disease sub-model

Let T_i be the age (in days) at death by lung cancer of miner i, $i \in \{1, 2,\ldots, n\}$ where n is the total number of miners. Let C_i be the right-censored age defined as the earliest age of miner i among age at death by a cause other than lung cancer; age on December 31, 2007; age in days corresponding to his 85th birthday and age until loss to follow-up. For each miner i, the observed outcome of interest can therefore be represented by the non-negative continuous variable $Y_i = \min(T_i, C_i)$ and the binary variable δ_i, where $\delta_i = 1$ if $T_i \leq C_i$ (i.e., miner i died of lung cancer at age $Y_i = T_i$) and $\delta_i = 0$ if $T_i > C_i$ (i.e., miner i "would have died of lung cancer" after age C_i).

Let $X_i^{\text{cum}}(t-5)$ be the true and unknown cumulative occupational γ-ray dose of miner i at age t, lagged by 5 years. Indeed, a latency period of 5 years between a received exposure and its potential impact on lung cancer mortality is assumed (Langholz et al. 1999; Rage et al. 2015).

It is proposed to describe the relationship between the cumulative occupational γ-ray doses and the age at death by lung cancer of miner i with a survival sub-model defined by the following instantaneous hazard rate function:

$$h_i(t;\beta) = h_0(t)\left(1 + \text{EHR}_i(t;\beta)\right), \tag{1}$$

where $\text{EHR}_i(t;\beta)$ is the excess hazard ratio (EHR) of death by lung cancer potentially associated to γ-ray exposure for miner i at age t.

As commonly assumed when modelling the association between solid cancer mortality and exposure to γ-rays (Vacquier et al. 2011; Rage et al. 2015), a linear structure in cumulative dose is assumed, with no effect modification:

$$\text{EHR}_i(t;\beta) = \beta X_i^{\text{cum}}(t-5). \tag{2}$$

Thus, β is the unknown risk coefficient of interest, subject to the constraint $\beta X_i^{\text{cum}}(t-5) > -1\ \forall t \forall i$ to ensure the positivity of $h_i(t;\beta)$. Finally, $h_0(t)$ corresponds to the instantaneous baseline hazard rate of death by lung cancer at age t for an unexposed miner.

The baseline hazard rate $h_0(t)$ is assumed to be piecewise constant and given by:

$$h_0(t) = \lambda_j \forall t \in \left(s_{j-1}, s_j\right], \tag{3}$$

with cut-points of the time axis fixed at: $s_0 = 0$, $s_1 = 40$, $s_2 = 55$, $s_3 = 70$ and $s_4 = 85$ years old. Thus, four age intervals are considered for which the values of the baseline hazard λ_j are assumed to be constant. In the following, $\lambda = (\lambda_1, \lambda_2, \lambda_3, \lambda_4)$ denotes the vector of unknown baseline hazard parameters.

Measurement sub-model

This sub-model describes the association between a true γ-ray dose and a γ-ray dose prone to measurement error. The latter can be either a strictly positive measured dose or the unknown dose that would have been measured if non-missing or non-censored. Here, it is assumed that missing exposure data are missing at random, meaning that their absence could be compensated by strictly positive γ-ray doses and other relevant covariates like calendar period, localization of the mine, job type.

All the γ-ray doses assumed to be zero according to Table 2 are assumed to be true zeros in the following. Consequently, they are assumed not to be prone to measurement error and, consequently, they are not included in the following measurement sub-model. Additionally, some missing γ-ray doses and all those assumed to be lower than the DL are treated as latent random variables (with restrictions detailed in Table 2) whose uncertainty is described by the same probability distribution (i.e., with the same parameters) as the one used to describe the strictly positive γ-ray doses measured at a given time, for the exposed miners of the post-55 sub-cohort.

Let $X_i(t)$ be the true γ-ray dose of miner i at time t. Let $Z_i(t)$ be either the strictly positive measured γ-ray dose of miner i at time t (if higher than the DL) or the unknown γ-ray dose that would have been measured for miner i if non-missing at time t or if the personal dosimeter did not have a DL at time t (denoted with DL(t) hereafter). $Z_i(t)$ is assumed to be prone to measurement error. Moreover,

$X_i(t)$ and $Z_i(t)$ are treated as latent variables, except when $Z_i(t)$ corresponds to a strictly positive measured γ-ray dose. If $Z_i(t)$ is not measured, it is assumed to follow a specific probability distribution (described in the following) and, additionally, to be smaller than DL(t) if it is left-censored. Note that, here, the censoring process is deterministic since DL(t) is fixed at each time t: DL(t) = 2.2 mSv for t between 1956 and 1985, and DL(t) = 0.55 mSv for t between 1986 and 2007. The outcome of interest of the measurement sub-model is defined by $(W_i(t), \delta_{it}^{W})$ where $W_i(t) = \max(Z_i(t), \mathrm{DL}(t))$ and $\delta_{it}{}^{W}$ indicates whether $Z_i(t)$ exceeds the DL or not (i.e., $\delta_{it}{}^{W} = 1$ if $Z_i(t) \geq DL(t)$ and $\delta_{it}{}^{W} = 0$ if $Z_i(t) < DL(t)$). If the miner's job position is recorded as "regular" with an assumed exposure smaller than DL (see Table 2), then it is assumed that $W_i(t) = DL(t)$ and $\delta_{it}{}^{W} = 0$. If the miner's job position is recorded as "expatriate" (see Table 2), then it is assumed that $W_i(t) = \max(Z_i(t), \mathrm{DL}(t))$ and $\delta_{it}{}^{W} = 0$ or 1.

Given that γ-ray dose values are based on personal dosimeters during the whole exposure period (1956–2007) in the post-55 sub-cohort of French uranium miners, a classical error is naturally assumed for γ-ray doses in order to reflect the lack of precision of personal dosimeters. A shared measurement error occurs when the discrepancy between true and observed exposures depends on error components that affect similarly several miners or several exposure times of a same miner. In the specific context of personal dosimetry, there is no reason to assume a shared error component in the measurement sub-model, for the whole exposure period.

Consistent with much of the literature suggesting that a multiplicative measurement error sub-model may be more realistic than an additive one in occupational epidemiology in general (Armstrong 1998), a lognormal and multiplicative error structure is postulated. Nevertheless, given that γ-ray doses were measured with film badge dosimeters between 1956 and 1985 and then with TLDs between 1986 and 2007, the variance of measurement error changed over time. Therefore, the following classical measurement error sub-model is assumed:

$$Z_i(t) = X_i(t)U_i(t). \tag{4}$$

The measurement error terms $U_i(t)$ are supposed to be independent from each other and to follow a lognormal distribution with a geometric mean of $-\sigma^2{}_{U,q(t)}/2$ and a geometric variance of $\sigma^2{}_{U,q(t)}$:

$$U_i(t) \sim \log N\left(-\frac{\sigma^2_{U,q(t)}}{2}, \sigma^2_{U,q(t)}\right), \tag{5}$$

where $q(t)$ equals 1 for exposure period 1956–1985 and 2 for exposure period 1986–2007.

Assuming a lognormal distribution implies that $E(U_i(t)) = 1$ for all i and for all t and thus, that $E(Z_i(t)|X_i(t)) = X_i(t)$. $Z_i(t)$ is thus an unbiased estimate of the true dose $X_i(t)$. Note that, since $Z_i(t)$ is prone to classical measurement error, $X_i(t)$ might be greater than DL(t) even if $Z_i(t) < \mathrm{DL}(t)$.

Due to the lack of validation data to estimate the true variance of the measurement error, the geometric standard deviation of the lognormal error terms $U_i(t)$ was extracted from Allodji (2011) for each calendar period of interest: $\sigma_{U,1} = 0.245$ between 1956 and 1985 and $\sigma_{U,2} = 0.16$ between 1986 and 2007. To estimate these values, Allodji (2011) first listed several potential reasons for the existence of measurement error in γ-ray doses in the post-55 sub-cohort: accuracy of measurement devices for the medical, radiological and environmental radiation fields, loss of record keeping, and issues in data transcription. Then, the impact of these reasons on the global measurement error was estimated from the studies published by Brady (1985) and Gilbert et al. (1996).

Exposure sub-model

The correction of classical measurement error requires the specification of the probability distribution for the true latent γ-ray dose $X_i(t)$. As occupational exposures are often assumed to follow a lognormal distribution (Steenland et al. 2015), and following Allodji (2011), a lognormal distribution is assumed:

$$\log\left(X_i(t)\right) \sim N\left(\mu_{x,i}(t), \sigma^2_{x,i}(t)\right), \tag{6}$$

where $\mu_{x,i}(t)$ and $\sigma_{x,i}(t)$ denote the expected value and the standard deviation of the log-transformed true γ-ray dose of miner i at time t.

For the sake of parsimony, different modelling assumptions are made on the structure of $\mu_{x,i}(t)$ and $\sigma_{x,i}(t)$, in order to avoid having as many unknown parameters as the total number of years of exposure for all the exposed uranium miners of the post-55 sub-cohort of French uranium miners. These assumptions also allow taking advantage of the information brought by all the miners over time, depending on the calendar period or the type of mine.

Previous statistical analyses (results not shown) showed that the geometric standard deviation $\sigma_{x,i}(t)$ can reasonably be assumed to be constant over time and for all miners. Therefore, in the following, it is assumed that $\sigma_{x,i}(t) = \sigma_x$ for all i and all t.

To allow the expected value of the log-transformed true γ-ray dose $\mu_{x,i}(t)$ to vary over time, three sub-models are proposed. The first exposure sub-model, noted M_1, assumed $\mu_{x,i}(t)$ to be the same piecewise-constant function over time for all miners:

$$\mu_{x,i}(t) = \mu_{x,p(t)} \forall i \forall t, \tag{7}$$

where the variable $p(t)$ takes values in {1, 2, 3, 4, 5} corresponding to five exposure periods of approximately 10 years and accounting for the change in personal dosimeters that

occurred in 1986: 1956–1965, 1966–1975, 1976–1985, 1986–1995 and 1996–2007. Therefore, $\mu_x = (\mu_{x,1}, \mu_{x,2}, \mu_{x,3}, \mu_{x,4}, \mu_{x,5})$ is the vector of unknown expected values of the log-transformed true doses for the five exposure periods. The vector of six unknown parameters of the sub-model M_1 is then given by (μ_x, σ_x).

The second exposure sub-model, referred to as M_2, is a hierarchical model that describes the uncertainty on parameters $\mu_{x,i}(t)$ through a normal distribution with the mean to be defined as a linear function over time with an unknown slope a and intercept b. This linear function is assumed to be the same for all miners:

$$\mu_{x,i}(t) \sim N\left(a \cdot f(t) + b, \sigma_\mu^2\right), \text{ where } f(t) = t - 1956. \quad (8)$$

The parameter b can be interpreted as the expected value of the log-transformed true γ-ray dose in 1956 in the post-55 sub-cohort of French uranium miners. For the sake of parsimony, the standard deviation σ_μ of the expected value of the log-transformed true γ-ray dose is assumed to be constant over time. The vector of four unknown parameters of the sub-model M_2 is $(a, b, \sigma_x, \sigma_\mu)$.

Finally, it is also supposed that the temporal trend of log-transformed true γ-ray doses is different in underground mines and in open-pit ones. The third exposure sub-model, referred to as M_3 and which is also a hierarchical model, accounts for this assumption:

$$\begin{aligned} &\mu_{x,i}(t) = \mu_{x,m_i(t)}(t) \text{ with} \\ &\mu_{x,J}(t) \sim N\left(a_J \cdot f(t) + b_J, \sigma_\mu^2\right) \text{ and} \\ &\mu_{x,F}(t) \sim N\left(a_F \cdot f(t) + b_F, \sigma_\mu^2\right), \end{aligned} \quad (9)$$

where $m_i(t) = \{\mathrm{F}, \mathrm{J}\}$ is the type of mine where the miner i worked at time t.

Here, the sub-model depends on the miner through the type of mine, as the type of mine is time and miner specific. In this sub-model, the slope and the intercept of the linear temporal trend depend on the type of mine. When information on the type of mine $m_i(t)$ is missing for miner i at time t (percentage of missing values in the post-55 sub-cohort = 11%), a Bernoulli probability distribution is assigned to the variable $m_i(t)$. Its parameter is p_F which denotes the unknown probability for a miner to work in an underground mine. F and J correspond to an underground mine and an open-pit mine, respectively. The vector of seven unknown parameters of the sub-model M_3 is $(a_J, a_F, b_J, b_F, \sigma_x, \sigma_\mu, p_F)$.

Prior choice and Bayesian inference

Figure 2 shows the directed acyclic graph for the full hierarchical model combining the disease sub-model, the measurement sub-model and the exposure sub-model M_3. To fit the three hierarchical sub-models under the Bayesian paradigm, prior distributions must be assigned to all the unknown parameters, which are assumed to be independent.

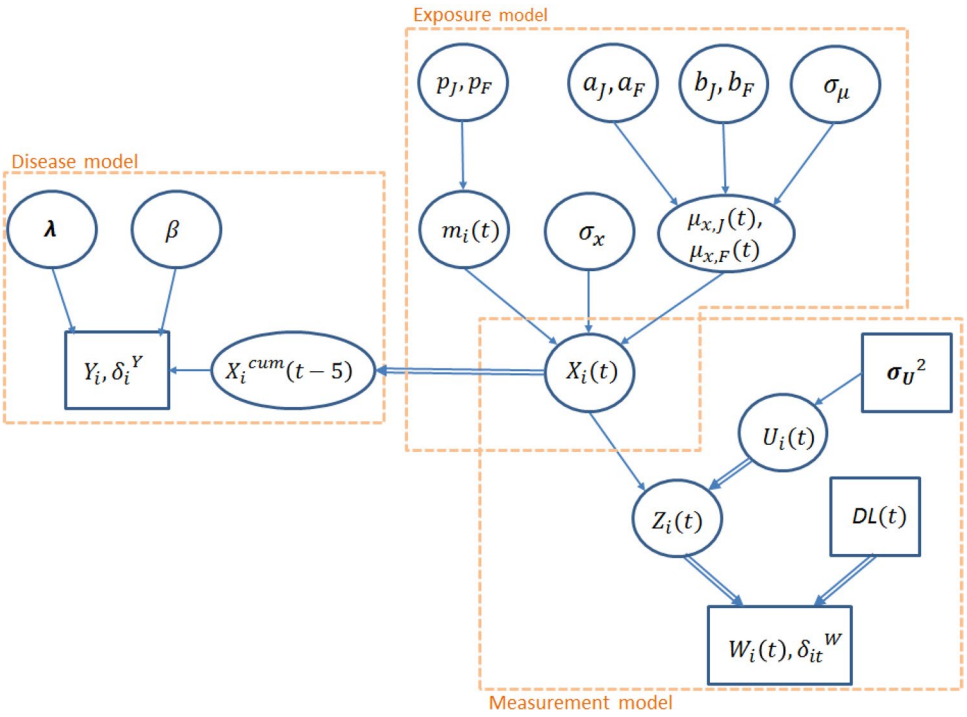

Fig. 2 Directed acyclic graph for the full hierarchical model based on the exposure sub-model M_3. Circles indicate unknown quantities and rectangles indicate observed variables or fixed parameters. Single arrows indicate oriented probabilistic links between two quantities and double arrows indicate oriented deterministic links between two quantities

The prior for the risk coefficient β is a vague prior in the form of a centered normal distribution with large variance (10^6). The prior is not left truncated but the positivity of the instantaneous hazard $h_i(t)$ is guaranteed during the inferential process, when computing the sub-model likelihood. Following Hoffmann et al. (2017), an informative gamma prior is assigned with shape parameter 23.66 and scale parameter 4.90 10^8 for the baseline hazard rate λ_1 (corresponding to the age interval [0; 40]) and flat uniform distributions between 0 and 100 for the baseline hazard rates λ_2, λ_3 and λ_4. Actually, since only one miner died of lung cancer before the age of 40 in the post-55 sub-cohort, the information contained in the data about λ_1 is very poor, making a reliable estimation of this parameter impossible from data. Consequently, it would have been hazardous to estimate the baseline hazard rate λ_1 from a flat uniform prior distribution and this justified assigning an informative prior to this parameter. The hyperparameters of the gamma prior distribution for λ_1 were derived from external data on the French national mortality rates by lung cancer between 1968 and 2005 (Hoffmann et al. 2017).

Finally, the prior distributions for the parameters of the three exposure sub-models are inspired from non-informative Jeffreys' prior distributions (Jeffreys 1998). They are indicated in Table 3. The goal is here to formalize the absence of knowledge about the distribution of the log-transformed true γ-ray doses in the post-55 sub-cohort of French uranium miners and to use prior probability distributions that are invariant by reparametrization.

A Markov Chain Monte Carlo (MCMC) algorithm was implemented in the programming language Python (for more details see: https://www.python.org/doc/) to sample from the joint posterior distribution of the unknown parameters and latent variables of each hierarchical model, combining the disease sub-model, the measurement sub-model and one of the three exposure sub-models. A metropolis-within-Gibbs algorithm (Roberts and Rosenthal 2009) was adopted to conduct the Bayesian inference, as full conditional distributions were analytically intractable except for the parameters of the exposure sub-model. In the adaptive phase of Metropolis–Hastings steps, the variance of each proposal distribution was calibrated to target an acceptance rate of 40% for single parameters and 20% for vectors (Roberts and Rosenthal 2009), in order to improve the efficiency and the convergence of the algorithm. Parameters and latent variables were independently updated. Three chains with different initial values were run. After 100 cycles of 100 iterations for the adaptive phase, the first 10,000 iterations were discarded as burn-in phase, and 55,000 additional iterations were run for each model. To decrease intra-chains autocorrelation, the sample was thinned storing only every 30 iterations. The posterior samples included then 5500 values. Trace plots of the Markov chains and the Gelman–Rubin statistics (Gelman and Rubin 1992) were used to check that there were no convergence issues (results not shown). Finally, the Monte-Carlo (MC) error (Gilks et al. (1995)) of the risk coefficient β was computed and showed that the size of the posterior samples was large enough to provide accurate estimates of β (results not shown).

Competing exposure sub-models were compared via the Watanabe–Akaike, also called widely applicable information criterion (Watanabe 2010). WAIC quantifies the predictive accuracy of a fitted model to data while accounting for its effective number of parameters to adjust for overfitting. WAIC is based on the log-transformed pointwise posterior predictive distribution of the outcomes of interest that can be approximated by simulations from the posterior sample derived from the MCMC algorithm. A smaller WAIC value indicates a better predictive accuracy for a given model.

Table 3 Jeffreys prior probability distributions for the parameters of the three exposure sub-models

Exposure sub-model	Parameters	Prior distribution (up to a constant)
M_1	$\mu_{x,1}, \mu_{x,2}, \mu_{x,3}, \mu_{x,4}, \mu_{x,5}$	1
	τ_x*	$1/\tau_x$
M_2	a, b	1
	τ_x	$1/\tau_x$
	τ_μ**	$1/\tau_\mu$
M_3	a_J, b_J, a_F, b_F	1
	τ_x	$1/\tau_x$
	τ_μ	$1/\tau_\mu$
	p_F	Beta(1/2, 1/2)

*$\tau_x = 1/\sigma^2$; τ_x is called the precision parameter

**$\tau_\mu = 1/\sigma_\mu^2$; τ_μ is called the precision parameter

Results

Sensitivity to the exposure sub-model

As described earlier, three different exposure sub-models were implemented to describe the uncertainty of the true latent γ-ray doses in the post-55 sub-cohort. The posterior medians of the excess hazard ratio (EHR) per 100 mSv and the associated 95% credible intervals (95% CI) as well as the DIC and the WAIC values for the different exposure sub-models are given in Table 4.

In general, there are no substantial differences between the risk estimates, 95% CI, DIC and WAIC values for the three exposure sub-models. This shows that the risk estimates are robust to the exposure sub-model choice. Given that the sub-model M_3 provides more information than the other sub-models on the temporal trend of log-transformed

Table 4 Posterior medians and 95% credible intervals (CI) of the excess hazard ratio (EHR) (per 100 mSv) for lung cancer mortality in the post-55 sub-cohort of French uranium miners, assuming three different exposure sub-models

Exposure Sub-model	EHR per 100 mSv	WAIC	DIC
M_1	0.82 [0.29; 1.71]	2462.0	2461.8
M_2	0.81 [0.30; 1.75]	2461.9	2461.7
M_3	0.81 [0.28; 1.75]	2461.7	2461.5

Watanabe–Akaike information criterion (WAIC) and deviance information criterion (DIC) for the three exposure sub-models

true γ-ray doses according to the type of mine and the smallest estimated geometric standard deviation σ_x of the true latent γ-ray doses, in the following focus will be placed on this sub-model. Results of the Bayesian hierarchical models based on the exposure sub-models M_1 and M_2 are provided in the supplementary materials.

Results for the hierarchical model with exposure sub-model M_3

Table 5 gives the posterior medians and 95% CI of the EHR (per 100 mSv) for lung cancer mortality, the baseline hazards and the parameters of the full hierarchical model based on the exposure sub-model M_3 (see Fig. 2).

The corrected risk coefficient for death by lung cancer associated with cumulative γ-ray doses in the post-55 French sub-cohort of uranium miners was estimated to be 0.81 per 100 mSv (95% CI: [0.28; 1.75]). Therefore, after accounting for γ-ray dose uncertainty, there is still a statistically significant positive association between γ-ray doses and the risk of death by lung cancer in the post-55 sub-cohort. The slope parameters a_J and a_F defining the temporal trend of the log-transformed γ-ray doses are significantly negative for both types of mine. This confirms that the expected value of the log-transformed true γ-ray dose decreased over time in the French uranium mines (Fig. 1). This downward trend is stronger in underground mines than in open-pit ones, as displayed in Fig. 3. As expected, the intercept parameter b_F that denotes the expected log-transformed true γ-ray dose in 1956 is higher for the underground mines than the intercept parameter b_J for the open-pit mines. Interestingly, Fig. 3 also shows that, from 1996, there is no significant difference between the annual expected true γ-ray dose in underground mines and in open-pit mines. This does not mean that the levels of γ-ray exposure were the same in underground and open-pit mines, but rather that the number of exposed miners was too low from 1996 to highlight any significant difference, should it exist. Indeed, French uranium mines closed shortly after this date.

Table 5 Posterior medians and 95% credible intervals (CI) of the parameters of the full hierarchical model combining the disease sub-model, the measurement sub-model and the exposure sub-model M_3

Parameter	Posterior median	95% CI
EHR per 100 mSv	0.81	[0.28; 1.75]
$\lambda_1\,(10^{-6})$	0.05	[0.03; 0.07]
$\lambda_2\,(10^{-6})$	0.78	[0.45; 1.27]
$\lambda_3\,(10^{-6})$	4.24	[2.72; 6.22]
$\lambda_4\,(10^{-6})$	7.33	[4.18; 11.88]
a		
Open-pit (J)	−0.025	[−0.034; −0.016]
Underground (F)	−0.047	[−0.054; −0.039]
b		
Open-pit (J)	−1.59	[−1.81; −1.37]
Underground (F)	−0.05	[−0.26; 0.16]
p		
Underground (F)	0.67	[0.66; 0.68]
σ_μ	0.33	[0.27; 0.45]
σ_x	0.93	[0.90; 0.96]

EHR excess hazard ratio

Sensitivity to the detection limit of dosimeters

Given that the values fixed for the DL of the dosimeters might also be uncertain, the robustness of the estimated

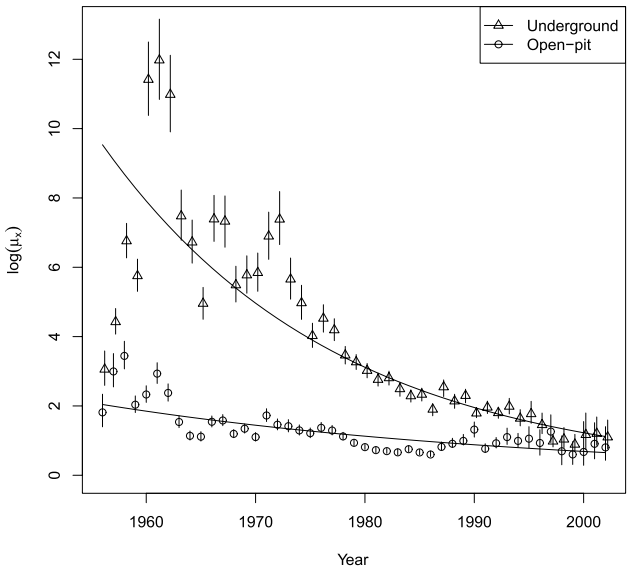

Fig. 3 Expected log-transformed true γ-ray exposures in underground mines (i.e., $\log(\mu_{x,F}(t))$) and open-pit mines (i.e., $\log(\mu_{x,J}(t))$) over time, in the post-55 sub-cohort of French uranium miners. The circles and triangles represent the posterior medians of $\log(\mu_{x,J}(t))$ and $\log(\mu_{x,F}(t))$, respectively, and the segments indicate the 95% credible intervals. Solid lines show $\log(a_J \bullet f(t) + b_J)$ and $\log(a_F \bullet f(t) + b_F)$ that are estimated from the posterior medians of a_J, a_F, b_J and b_F

EHR to these values was also tested. For this, the Bayesian hierarchical model including exposure sub-model M_3 was fitted after considering two alternative scenarios. First, the previously defined DLs were increased by 50% (3.3 mSv before 1986 and 0.825 mSv after 1986). To do this, any dose value between the original DL and the increased DL was assumed to be left-censored. In other words, $Z_i(t)$—that was observed with the original DL—was assumed to be unknown but smaller than the DL(t), and was also assumed to follow the lognormal distribution given by Eq. (4). The posterior median of the EHR was then estimated to be 0.80 per 100 mSv with a 95% CI of [0.28; 1.73]. Second, the previously defined DLs were increased by 75% (3.85 mSv before 1986 and 0.9625 mSv after 1986). The posterior median of the EHR was then estimated to be 0.83 per 100 mSv with a 95% CI of [0.29; 1.75]. Therefore, it is observed that increasing the value of the DL does not substantially change the risk estimates.

Impact of the different sources of γ-ray dose uncertainty on risk estimate

Finally, the disease sub-model was fitted without accounting for exposure measurement error and after replacing all left-censored and missing γ-ray dose values by zero. The uncorrected EHR of death by lung cancer due to occupational γ-ray exposure was then estimated to be 0.78 per 100 mSv with a 95% CI of [0.28; 1.64]. This result is similar to the uncorrected excess relative risk estimated from a Poisson regression model which was 0.74 per 100 mSv with a 95% CI of [0.23; 1.73] (Rage et al. 2015). Given that the corrected EHR estimate is 0.81 per 100 mSv with a 95% CI of [0.28; 1.75], it is concluded that no substantial change of the estimated EHR per 100 mSv is observed when accounting for dose uncertainty. The associated 95% CI is larger when accounting for dose uncertainty but does not include zero meaning that the positive association between lung cancer mortality and γ-ray exposure remains statistically significant.

In order to distinguish the impact on risk estimate of both classical measurement errors and left-censored or missing dose values, a hierarchical model was fitted accounting for measurement error with exposure sub-model M_3 but after replacing all left-censored and missing exposure data by zero. In other words, classical measurement errors were accounted for but the existence of left-censored or missing exposure data was neglected. The EHR of death by lung cancer due to occupational γ-ray exposure was then estimated to be 0.80 per 100 mSv (95% CI: [0.29; 1.70]). Thus, when only accounting for classical measurement error, again no substantial change in the estimated EHR per 100 mSv was observed. Finally, a hierarchical model was fitted accounting for the existence of left-censored and missing dose values but neglecting the existence of classical measurement errors. The EHR of death by lung cancer due to occupational γ-ray exposure was then estimated to be 0.80 per 100 mSv (95% CI: [0.28; 1.75]). Again, when only accounting for left-censored and missing exposure data, no substantial change of the EHR per 100 mSv was observed. Nevertheless, the associated 95% CI is slightly wider when this source of uncertainty is accounted for by the model, as compared to the 95% CI obtained when only accounting for classical measurement error. In all cases, the 95% CI of the EHR per 100 mSv did not include 0 and, consequently, the positive association between lung cancer mortality and γ-ray exposure remained statistically significant. Figure 4 displays the prior and posterior densities of the risk coefficient (i.e., β) for these different models.

Sensitivity to the assumptions regarding the magnitude of the measurement error

In order to test the robustness of the results to the geometric standard deviations fixed for the lognormal measurement error components (see subsection on measurement sub-model), the full Bayesian hierarchical model including exposure sub-model M_3 was fitted after increasing the corresponding standard deviations by 50% and 100%. When assuming an increase of 50%, the EHR per 100 mSv remained positive and statistically significant with a posterior median equal to 0.86 per 100 mSv and a 95% CI of [0.33; 1.84]. When assuming an increase of 100%, the estimated posterior median of the EHR was 0.90 per 100 mSv with a 95% CI of [0.34, 1.89]. As expected, the results are

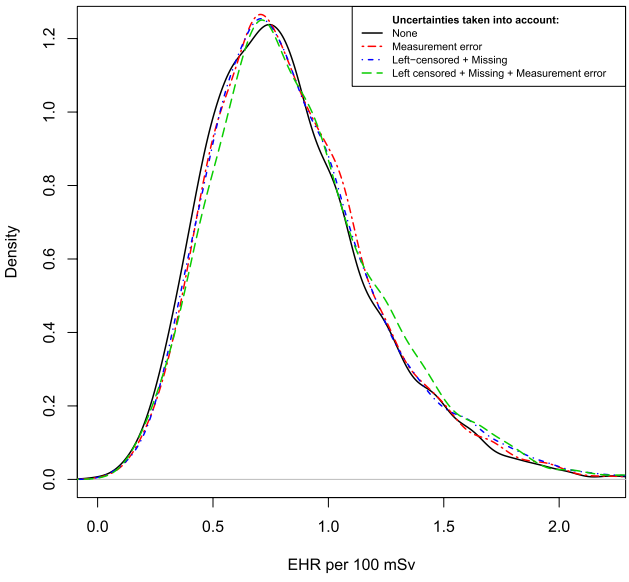

Fig. 4 Posterior densities of the EHR per 100 mSv of death by lung cancer due to occupational γ-ray exposure in the post-55 sub-cohort of French uranium miners, depending on the sources of exposure uncertainty that were accounted for

sensitive to the magnitude of the measurement error. They also show that, in cases where this parameter is under-estimated (resp. over-estimated), the risk of death by lung cancer may be under-estimated (resp. over-estimated) as well, as its estimation uncertainty.

Discussion

In this work, three Bayesian hierarchical models based on different exposure sub-models were developed, fitted and compared, to explicitly account for missing and left-censored γ-ray dose measurements prone to classical measurement error, when estimating the risk of death by lung cancer, in the post-55 sub-cohort of French uranium miners. Advantage was taken of the flexibility of hierarchical modelling to simultaneously describe several sources of exposure uncertainty, i.e., classical measurement errors, the deterministic censoring process due to the detection limit of the dosimeters, and missing data. In contrast to classical functional approaches such as SIMEX (Cook and Stefanski 1994) or regression-calibration (Stefanski and Carroll 1985), where several consecutive steps are used to impute left-censored and missing exposure data and to estimate true exposure and unknown risk parameters, fitting these hierarchical models under the Bayesian paradigm allowed for a simultaneous estimation—through a coherent and valid inferential framework—of risk parameters, true doses, and assumed doses (in case of false zeros or missing data). This implies that the estimation of uncertainty of each unknown quantity—which may be large and/or far from a Gaussian distribution—is accounted for when estimating other unknown quantities that depend on it. In this work, for instance, the estimation of uncertainty of the true latent γ-ray doses $X_i(t)$ is accounted for when estimating the EHR of death by lung cancer. Sampling from the joint posterior distribution of the assumed dose values (i.e., $Z_i(t)$) for false zeros and missing γ-ray dose values only) and the true doses (i.e. $X_i(t)$) was carried out through the MCMC algorithm. This allowed to explicitly account for the associated uncertainty when computing the cumulative dose of each miner. At this point, it is also important to note that the proposed Bayesian hierarchical models allowed taking advantage of all information available through the strictly positive and true zero γ-ray dose values at a given period and/or in a given type of mine (i.e., open-pit, underground) to learn about all the unknown strictly positive and potentially left-censored γ-ray dose values in the same period and/or type of mine. Finally, another advantage of using Bayesian statistics was the possibility to assign an informative prior distribution for unknown parameters that were only poorly informed by the data, in order to improve their estimation. In the present study, this was particularly useful for the estimation of the baseline hazard of death by lung cancer for uranium miners younger than 40 years old (i.e. parameter λ_1) for which poor information was available in the post-55 sub-cohort of French uranium miners.

In this paper, emphasis was placed on the problem of accounting for three specific sources of dose uncertainty when estimating the association between occupational exposure to low levels of γ-radiation and lung cancer mortality in the post-55 sub-cohort of French uranium miners. Missing and left-censored γ-ray dose values prone to unshared classical measurement error were modelled on an individual level, using an excess hazard ratio (EHR) survival sub-model. Given the modelling assumptions, no substantial impact of the considered sources of dose uncertainty on the risk of death by lung cancer due to cumulative γ-ray exposure was found. Recently, Hoffmann et al. showed, through a simulation study, that the impact of unshared classical measurement error (multiplicative and log-normally distributed) on an EHR survival sub-model was very small for a geometric variance of the measurement error equal to 0.1 (Hoffmann et al. 2018). Yet here, and according to the previous work by Allodji (2011), it was assumed that the geometric variance of the measurement error was always smaller than 0.06 for the whole period of γ-ray exposure, which might explain why measurement error had no substantial impact on risk estimates. Additionally, the small proportion (about 7%) of left-censored and missing γ-ray exposures in the post-55 sub-cohort of French uranium miners explained why replacing all left-censored and missing γ-ray exposures by zero had no significant impact on risk estimates either. A sensitivity analysis showed that the results obtained are robust to an increase of the detection limit of the dosimeters by 75%. The corrected risk coefficient of death by lung cancer was estimated to be 0.81 per 100 mSv with a 95% CI of [0.28; 1.75]. This confirms the robustness of the risk estimate obtained from a survival EHR which does not account for measurement error and where all left-censored and missing γ-ray dose values were replaced by zero (EHR of 0.78 per 100 mSv with a 95% CI of [0.28; 1.64]). Even if the 95% credible interval of the EHR is wider after accounting for exposure uncertainty, a statistically significant positive association remained between γ-ray exposure and the risk of death by lung cancer. It was also shown that the results obtained are robust to the different assumptions made to describe the temporal trend of the expected true γ-ray doses.

The robustness of the results could also have been tested with other assumptions on the expected true dose such as non-linear trends or temporal correlations. However, the estimated risk of death by lung cancer associated to γ-ray exposure cannot, for now, be compared to other risk estimates in comparable populations. Indeed, up to our knowledge, this risk has not been calculated in other cohorts of uranium miners. An analysis of the health risks associated to γ-ray exposure is planned in the pooled uranium miners

analysis (PUMA) (Rage et al. 2020), which will allow a comparison with our results.

Interestingly, the Bayesian hierarchical model based on the exposure sub-model M_3 allowed estimating a statistically significant decrease of the true log-transformed γ-ray doses over time in underground and open-pit mines, and showed that this downward trend was stronger in underground mines than in open-pit ones. This decreasing trend was also observed in the German uranium miners' cohort (Kreuzer et al. 2013). This trend could be due to intensive mining operations that may have led to lower uranium concentration and thus a reduced level of γ-ray exposure, and to changes in radiation protection practice like a reduction in exposure time of uranium miners.

In this study, since there were no validation data available to estimate the magnitude of measurement error, the measurement error was fixed according to the calendar period following the work by Allodji (2011). Nevertheless, as expected, the risk estimate of death by lung cancer was sensitive to this parameter showing that a careful assessment of this crucial parameter must be made to ensure the validity of the risk estimate. Here, the corrected risk estimate of death by lung cancer (0.81 per 100 mSv with a 95% CI of [0.28; 1.75]) must be considered with prudence, assuming that the geometric standard deviation is 0.245 between 1956 and 1985 and 0.16 between 1986 and 2007 when considering a multiplicative error structure.

This study did not account for the tobacco consumption of miners, although smoking is known to be the most important cause of lung cancer. Unfortunately, the smoking status is only available for about 4% of the miners in the post-55 subcohort of French uranium miners. This major lack of information makes it very hazardous to adjust for smoking status when estimating the risk of death by lung cancer due to γ-ray exposure among this cohort. Moreover, previous analyses on the impact of smoking in occupational cohort studies of uranium miners suggested that smoking was not a source of confounding in these studies (Richardson et al. 2014; Keil et al. 2015). Additionally, analyses on sub-cohorts for which the smoking history of workers was available observed a significant association between radon exposure and lung cancer mortality for both smokers and for non-smokers with an estimated risk coefficient that tended to be higher for non-smokers than for smokers (Tomasek 2002; Tirmarche et al. 2012; Kreuzer et al. 2017). Finally, case–control studies nested in the French cohort of uranium miners (Leuraud et al. 2007) and in two other European cohorts of uranium miners (Leuraud et al. 2011) found that, when adjusting for smoking, the effect of radon exposure on lung cancer risk persisted and the adjustment did not substantially alter the estimated risk coefficient associated with radon exposure.

Radon was classified as a pulmonary carcinogen in humans by the International Agency for Research on Cancer in 1988 (International Agency for Research on Cancer 1988), and is considered as the second cause of death by lung cancer after tobacco consumption (Samet 1989; Birchall and Marsh 2005). Given that radon and γ-ray exposure measurements are highly correlated in the French cohort of uranium miners (Pearson correlation coefficient = 0.90) (Vacquier et al. 2011) and given that radon is not included in the hierarchical models used here, due to multi-collinearity issues, the statistically significant positive association found between γ-ray exposure and the risk of death by lung cancer must be interpreted with caution. Indeed, the results obtained here could reflect either a real specific association between γ-ray exposure and lung cancer mortality or, more probably, a spurious association due to radon which acts as a confounding factor in this relationship.

Because of its flexible and modular nature, the Bayesian hierarchical models described in this paper could be easily extended to account for missing and left-censored dose values or more generally exposure data, prone to more complex patterns of measurement error, for which a substantial impact on risk estimates may be suspected. This should be particularly the case for exposure uncertainty that is shared between and/or within individuals (i.e., shared for several years of exposure for an individual) (Hoffmann et al. 2018). The possibility to include additional information on uncertain parameters like the standard deviation of the measurement error or exposure sub-model parameters could also help reducing uncertainty on risk coefficient estimates. Finally, to better highlight the potential impact on risk estimates of replacing missing and left-censored dose values or exposure data by zero, it would be very interesting to apply the Bayesian hierarchical models used here on cohorts of occupationally exposed uranium workers affected by a very high proportion of missing and left-censored dose values or exposure data.

Conclusion

This paper highlights the flexibility and relevance of the Bayesian hierarchical approach to account for both missing and left-censored radiological exposure data that are prone to measurement error, when estimating radiation-related risks. Up to the authors' knowledge, this is the first time these three sources of uncertainty are dealt with simultaneously in radiation epidemiology. Regarding the specific problem of estimating the risk of death by lung cancer due to external γ-rays exposure among French uranium miners, the impact of these three sources of uncertainty on the risk estimate is found to be of marginal importance (given the modelling assumptions). The corrected EHR is 0.81 per 100 mSv (95% credible interval: [0.28; 1.75]). Interestingly, this paper shows that, even if the 95%

credible interval of the corrected EHR is wider than the uncorrected one, a statistically significant positive association remains between γ-ray exposure and the risk of death by lung cancer, after accounting for dose uncertainty. This could reflect either a real specific association or, more probably, a spurious association due to radon which acts as a confounding factor in this relationship. Finally, because of its flexible and modular nature, the Bayesian hierarchical models described in this paper could be easily extended to account for higher proportions of missing and left-censored exposure data than in the post-55 sub-cohort of French uranium miners and more complex patterns of measurement error, for which a substantial impact on risk estimates may be suspected. This should be particularly the case for exposure uncertainty that is shared between and/or within individuals.

Acknowledgements This work was partially supported by ORANO in the framework of a bilateral agreement between IRSN and ORANO. We thank the two anonymous reviewers and the associated editor for their very constructive comments.

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflict of interest.

References

Allodji SR (2011) Prise en compte des erreurs de mesure dans l'analyse du risque associe a l'exposition aux rayonnements ionisants dans une cohorte professionnelle: application à la cohorte française des mineurs d'uranium, Université Paris Sud-Paris XI

Allodji RS, Thiebaut AC, Leuraud K, Rage E, Henry S, Laurier D et al (2012) The performance of functional methods for correcting non-Gaussian measurement error within Poisson regression: corrected excess risk of lung cancer mortality in relation to radon exposure among French uranium miners. Stat Med 31(30):4428–4443

Armstrong BG (1998) Effect of measurement error on epidemiological studies of environmental and occupational exposures. Occup Environ Med 55(10):651–656

Birchall A, Marsh J (2005) Radon dosimetry and its implication for risk. International Congress Series, Elsevier

Bouvier-Colle M-H, Vallin J, Hatton F (1990) Mortalité et causes de décès en France, Doin

Brady W (1985) Radiac instruments and film badges used at atmospheric nuclear tests. Defense Nuclear Agency

Carroll RJ (2005) Measurement error in epidemiologic studies. Encyclopedia of biostatistics 5

Carroll RJ, Ruppert D, Crainiceanu CM, Stefanski LA (2006) Measurement error in nonlinear models: a modern perspective. Chapman and Hall/CRC, New York

Clement CH, Tirmarche M, Harrison J, Laurier D, Paquet F, Blanchardon E et al (2010) Lung cancer risk from radon and progeny and statement on radon. Ann ICRP 40(1):1–64

Cook JR, Stefanski LA (1994) Simulation-extrapolation estimation in parametric measurement error models. J Am Stat Assoc 89(428):1314–1328

Flegal KM, Keyl PM, Nieto FJ (1991) Differential misclassification arising from nondifferential errors in exposure measurement. Am J Epidemiol 134(10):1233–1246

Fournier L (2017) Effets sanitaires d'une exposition chronique à de faibles doses de rayonnements ionisants: contribution à l'estimation des risques radio-induits de cancers dans une cohorte française de travailleurs du nucléaire. Université Paris-Saclay

Gelman A, Rubin DB (1992) Inference from iterative simulation using multiple sequences. Stat Sci 7(4):457–472

Gilbert E, Fix J, Baumgartner W (1996) An approach to evaluating bias and uncertainty in estimates of external dose obtained from personal dosimeters. Health Phys 70(3):336–345

Gilbert E, Thierry-Chef I, Cardis E, Fix J, Marshall M (2006) External dose estimation for nuclear worker studies. Radiat Res 166(1):168–173

Gilks WR, Richardson S, Spiegelhalter D (1995) Markov chain Monte Carlo in practice. Chapman & Hall/CRC interdisciplinary statistics. CRC Press, New York. ISBN 1482214970, 9781482214970

Heid I, Küchenhoff H, Wellmann J, Gerken M, Kreienbrock L, Wichmann H (2002) On the potential of measurement error to induce differential bias on odds ratio estimates: an example from radon epidemiology. Stat Med 21(21):3261–3278

Hoffmann S, Rage E, Laurier D, Laroche P, Guihenneuc C, Ancelet S (2017) Accounting for Berkson and classical measurement error in radon exposure using a bayesian structural approach in the analysis of lung cancer mortality in the French cohort of uranium miners. Radiat Res 187(2):196–209

Hoffmann S, Laurier D, Rage E, Guihenneuc C, Ancelet S (2018) Shared and unshared exposure measurement error in occupational cohort studies and their effects on statistical inference in proportional hazards models. PLoS ONE 13(2):e0190792

International Agency for Research on Cancer (1988) Man-made mineral fibres and radon. IARC monographs on the evaluation of carcinogenic risks to humans 43

Jeffreys H (1998) The theory of probability. OUP Oxford, Oxford

Keil AP, Richardson DB, Troester MA (2015) Healthy worker survivor bias in the Colorado plateau uranium miners cohort. Am J Epidemiol 181(10):762–770

Kim H-M, Yasui Y, Burstyn I (2006) Attenuation in risk estimates in logistic and Cox proportional-hazards models due to group-based exposure assessment strategy. Ann Occup Hyg 50(6):623–635

Kreuzer M, Dufey F, Sogl M, Schnelzer M, Walsh L (2013) External gamma radiation and mortality from cardiovascular diseases in the German WISMUT uranium miners cohort study, 1946–2008. Radiat Environ Biophys 52(1):37–46

Kreuzer M, Sobotzki C, Schnelzer M, Fenske N (2017) Factors modifying the radon-related lung cancer risk at low exposures and exposure rates among German uranium miners. Radiat Res 189(2):165–176

Langholz B, Thomas D, Xiang A, Stram D (1999) Latency analysis in epidemiologic studies of occupational exposures: application to the Colorado plateau uranium miners cohort. Am J Ind Med 35(3):246–256

Laurent O, Gomolka M, Haylock R, Blanchardon E, Giussani A, Atkinson W et al (2016) Concerted uranium research in Europe (CURE): toward a collaborative project integrating dosimetry, epidemiology and radiobiology to study the effects of occupational uranium exposure. J Radiol Prot 36(2):319

Leuraud K, Billon S, Bergot D, Tirmarche M, Caër S, Quesne B et al (2007) Lung cancer risk associated to exposure to radon and smoking in a case-control study of French uranium miners. Health Phys 92(4):371–378

Leuraud K, Schnelzer M, Tomasek L, Hunter N, Timarche M, Grosche B et al (2011) Radon, smoking and lung cancer risk: results of a joint analysis of three European case-control studies among uranium miners. Radiat Res 176(3):375–387

Leuraud K, Richardson DB, Cardis E, Daniels RD, Gillies M, O'hagan JA et al (2015) Ionising radiation and risk of death from leukaemia and lymphoma in radiation-monitored workers (INWORKS): an international cohort study. Lancet Haematol 2(7):e276–e281

Lubin JH, Colt JS, Camann D, Davis S, Cerhan JR, Severson RK et al (2004) Epidemiologic evaluation of measurement data in the presence of detection limits. Environ Health Perspect 112(17):1691–1696

Physick WL, Cope ME, Lee S, Hurley PJ (2007) An approach for estimating exposure to ambient concentrations. J Expo Sci Environ Epidemiol 17(1):76

Rage E, Caër-Lorho S, Drubay D, Ancelet S, Laroche P, Laurier D (2015) Mortality analyses in the updated French cohort of uranium miners (1946–2007). Int Arch Occup Environ Health 88(6):717–730

Rage E, Richardson DB, Demers PA, Do M, Fenske N, Kreuzer M et al (2020) PUMA–pooled uranium miners analysis: cohort profile. Occup Environ Med 77(3):194–200

Reeves GK, Cox DR, Darby SC, Whitley E (1998) Some aspects of measurement error in explanatory variables for continuous and binary regression models. Stat Med 17:2157–2177

Richardson S, Gilks WR (1993) A Bayesian approach to measurement error problems in epidemiology using conditional independence models. Am J Epidemiol 138(6):430–442

Richardson DB, Loomis D (2004) The impact of exposure categorisation for grouped analyses of cohort data. Occup Environ Med 61(11):930–935

Richardson DB, Laurier D, Schubauer-Berigan MK, Tchetgen ET, Cole SR (2014) Assessment and indirect adjustment for confounding by smoking in cohort studies using relative hazards models. Am J Epidemiol 180(9):933–940

Roberts GO, Rosenthal JS (2009) Examples of adaptive MCMC. J Comput Graph Stat 18(2):349–367

Ron E (1998) Ionizing radiation and cancer risk: evidence from epidemiology. Radiat Res 150(5s):S30–S41

Samet JM (1989) Radon and lung cancer. J Natl Cancer Inst 81(10):745–758

Steenland K, Karnes C, Darrow L, Barry V (2015) Attenuation of exposure-response rate ratios at higher exposures: a simulation study focusing on frailty and measurement error. Epidemiology 26(3):395–401

Stefanski LA, Carroll RJ (1985) Covariate measurement error in logistic regression. Ann Stat 13(4):1335–1351

Thomas D, Stram D, Dwyer J (1993) Exposure measurement error: influence on exposure-disease relationships and methods of correction. Annu Rev Public Health 14(1):69–93

Tirmarche M, Harrison J, Laurier D, Blanchardon E, Paquet F, Marsh J (2012) Risk of lung cancer from radon exposure: contribution of recently published studies of uranium miners. Ann ICRP 41(3–4):368–377

Tomasek L (2002) Czech miner studies of lung cancer risk from radon. J Radiol Prot 22(3A):A107

Vacquier B, Caer S, Rogel A, Feurprier M, Tirmarche M, Luccioni C et al (2008) Mortality risk in the French cohort of uranium miners: extended follow-up 1946–1999. Occup Environ Med 65(9):597–604

Vacquier B, Rage E, Leuraud K, Caër-Lorho S, Houot J, Acker A et al (2011) The influence of multiple types of occupational exposure to radon, gamma rays and long-lived radionuclides on mortality risk in the French "post-55" subcohort of uranium miners: 1956–1999. Radiat Res 176(6):796–806

Watanabe S (2010) Asymptotic equivalence of Bayes cross validation and widely applicable information criterion in singular learning theory. J Mach Learn Res 11:3571–3594

Xue X, Kim MY, Shore RE (2006) Estimation of health risks associated with occupational radiation exposure: addressing measurement error and minimum detectable exposure level. Health Phys 91(6):582–591

Yoder RC, Dauer LT, Balter S, Boice JD Jr, Grogan HA, Mumma MT et al (2018) Dosimetry for the study of medical radiation workers with a focus on the mean absorbed dose to the lung, brain and other organs. Int J Radiat Biol. https://doi.org/10.1080/09553002.2018.1549756

Zablotska LB, Fenske N, Schnelzer M, Zhivin S, Laurier D, Kreuzer M (2018) Analysis of mortality in a pooled cohort of Canadian and German uranium processing workers with no mining experience. Int Arch Occup Environ Health 91(1):91–103

C.2 Bayesian profile regression to deal with multiple highly correlated exposures and a censored survival outcome. First application in ionizing radiation epidemiology

frontiers in Public Health

ORIGINAL RESEARCH
published: 27 October 2020
doi: 10.3389/fpubh.2020.557006

Bayesian Profile Regression to Deal With Multiple Highly Correlated Exposures and a Censored Survival Outcome. First Application in Ionizing Radiation Epidemiology

Marion Belloni[1]*, *Olivier Laurent*[1], *Chantal Guihenneuc*[2†] *and Sophie Ancelet*[1†]

[1] PSE-SANTE/SESANE/LEPID, Institut de Radioprotection et de Sûreté Nucléaire, Paris, France, [2] Université de Paris, Unité de Recherche "Biostatistique, Traitement et Modélisation des données biologiques" BioSTM - UR 7537, Paris, France

OPEN ACCESS

Edited by:
Marc Chadeau-Hyam,
Imperial College London,
United Kingdom

Reviewed by:
Xavier Basagaña,
Instituto Salud Global Barcelona (ISGlobal), Spain
Erica Ponzi,
University of Oslo, Norway

***Correspondence:**
Marion Belloni
marion.belloni@irsn.fr

[†]These authors have contributed equally to this work

Specialty section:
This article was submitted to Exposome, a section of the journal Frontiers in Public Health

Received: 29 April 2020
Accepted: 21 September 2020
Published: 27 October 2020

Citation:
Belloni M, Laurent O, Guihenneuc C and Ancelet S (2020) Bayesian Profile Regression to Deal With Multiple Highly Correlated Exposures and a Censored Survival Outcome. First Application in Ionizing Radiation Epidemiology. Front. Public Health 8:557006. doi: 10.3389/fpubh.2020.557006

As multifactorial and chronic diseases, cancers are among these pathologies for which the exposome concept is essential to gain more insight into the associated etiology and, ultimately, lead to better primary prevention strategies for public health. Indeed, cancers result from the combined influence of many genetic, environmental and behavioral stressors that may occur simultaneously and interact. It is thus important to properly account for multifactorial exposure patterns when estimating specific cancer risks at individual or population level. Nevertheless, the risk factors, especially environmental, are still too often considered in isolation in epidemiological studies. Moreover, major statistical difficulties occur when exposures to several factors are highly correlated due, for instance, to common sources shared by several pollutants. Suitable statistical methods must then be used to deal with these multicollinearity issues. In this work, we focused on the specific problem of estimating a disease risk from highly correlated environmental exposure covariates and a censored survival outcome. We extended Bayesian profile regression mixture (PRM) models to this context by assuming an instantaneous excess hazard ratio disease sub-model. The proposed hierarchical model incorporates an underlying truncated Dirichlet process mixture as an attribution sub-model. A specific adaptive Metropolis-Within-Gibbs algorithm—including label switching moves—was implemented to infer the model. This allows simultaneously clustering individuals with similar risks and similar exposure characteristics and estimating the associated risk for each group. Our Bayesian PRM model was applied to the estimation of the risk of death by lung cancer in a cohort of French uranium miners who were chronically and occupationally exposed to multiple and correlated sources of ionizing radiation. Several groups of uranium miners with high risk and low risk of death by lung cancer were identified and characterized by specific exposure profiles. Interestingly, our case study illustrates a limit of MCMC algorithms to fit full Bayesian PRM models

even if the updating schemes for the cluster labels incorporate label-switching moves. Then, although this paper shows that Bayesian PRM models are promising tools for exposome research, it also opens new avenues for methodological research in this class of probabilistic models.

Keywords: Bayesian inference, ionizing radiation, lung cancer, multicollinearity, profile regression, survival data, truncated Dirichlet process mixture

1. INTRODUCTION

Over the last decade, the human exposome has emerged as a novel and promising research paradigm in epidemiology, biomedical, and environmental health sciences (1–3). Originally proposed by Dr. Christopher Wild in 2005 (4), it encompasses the totality of human environmental (meaning all non-genetic) exposures throughout life—from conception to death. This concept, that argues for a holistic and integrated consideration of all environmental exposures simultaneously (5, 6), is the key complement to the genome in terms of understanding human health. Its initial aim is to decipher how complex environmental exposure situations lead to disease development. Its final aims are to gain more insight into the etiology of multifactorial and chronic pathologies, and, ultimately, to lead to better primary prevention strategies for public health. Obviously, cancers are among these pathologies for which the exposome concept is essential, as they result from the combined influence of many genetic, environmental (i.e., physical, biological, chemical) and behavioral stressors that may occur simultaneously and interact (7–11).

In epidemiological studies, it is thus important to properly account for multifactorial exposure patterns when estimating (or predicting) specific cancer risks at individual or population level. However, historically, epidemiological studies linking the adverse effects of environmental stressors and human health have mostly focused on characterizing the effect of a single stressor. This one is typically considered of "main interest" for investigation (12, 13). A few additional risk factors, including other environmental stressors, are usually considered, but this is most frequently because of their feared role as potential confounders. They are therefore adjusted for in regression models, in order to estimate the effect of the "main environmental stressor of interest" but independently from the potential influence of the other risk factors (14, 15). Only a few studies aim to estimate the interaction between exposure to an environmental stressor and other risk factors (e.g., smoking and asbestos or radon) (16, 17), and, even more rarely, the joint effects of exposure to several environmental stressors (e.g., ambient particles and ozone) (18). In the specific field of protection against the effects of ionizing radiation—that will be of interest in this paper—estimating radiation-related cancer risks and its uncertainty has been a key objective for decades, for the purpose of setting exposure limits (19). However, although ionizing radiation epidemiology has successfully reached that goal, the question of estimating how simultaneous environmental exposures to multiple radiological stressors of different nature potentially affect cancer risks has not yet been investigated thoroughly (20).

Estimating cancer risks due to simultaneous exposures to multiple environmental stressors may be challenging for several reasons, which are detailed elsewhere (21, 22). Particularly, major statistical difficulties occur when exposure-based risk factors are highly correlated. This occurs when collecting data on multiple environmental stressors during life. This may be also the case, for instance, when a worker is simultaneously exposed to many chemical and physical stressors in the course of his occupational activity. This situation will be referred to as co-exposure in the following. In this context, it is well-recognized that applying standard multiple regression models—in which at least two highly correlated predictors are assessed simultaneously- may lead to unstable risk coefficient estimates with high variance. Therefore, this approach may lead to misleading conclusions and unrealistic interpretations about the effect of each of the collinear predictors on the outcome variable (23–25). More sophisticated statistical methods must then be used to deal with this multicollinearity issue.

Although not yet widely used in practice (26), several statistical methods have been proposed to deal with multicollinearity and then, to potentially investigate the combined effect on health outcomes of highly correlated environmental stressors. Many previous studies relied on an environment-wide association approach (EWAS) where, in its simplest version, the association between each single exposure factor and the outcome was estimated separately (27, 28). Even if potentially useful to discover priority risk factors, this approach is mainly considered in an exploratory research phase and leads to limited investigations of an health-exposome association. Other approaches that have been proposed in this specific context mainly rely on: (a) variable selection in a regression context using, for instance, the elastic net criterion (29) or the Graphical Unit Evolutionary Stochastic Search (30); (b) data-driven dimension reduction using regression on principal components (31) or the sparse partial least squares regression (32, 33); (c) machine learning algorithms like recursive partitioning using random forests (34); and (d) clustering approaches to profile multiple correlated data (35) like k-means, the latent class analysis (LCA) (36) and the Bayesian profile regression mixture (PRM) models (37). Variable selection approaches are very interesting tools to identify a small subset of environmental stressors that are the "true villain" most responsible for affecting the health outcome of interest. They are particularly adapted when a huge number of stressors is considered. However, when only a few highly correlated exposure covariates are available,

the idea is not to omit some of them in the study but rather to estimate an exposure-risk relationship using all available covariates and appropriate statistical methods to deal with multicollinearity issues. They may also be limited in their ability to efficiently differentiate true predictors from correlated covariates when the latter are very highly correlated (38). Data-driven dimension reduction aims at constructing summary latent variables as linear combinations of the original exposure covariates and then, to include these new uncorrelated variables in a multiple regression model (39). One major drawback is that these variables are constructed without considering the disease outcome of interest in principal component regression (PCR). Even if the sPLS (32) corrects for this by constructing uncorrelated latent variables as linear combinations of the original covariates and response variables, another drawback of data-driven dimension reduction approaches concerns the uncertainties related to this construction. Indeed, given that the disease risks are estimated in a second disjoint step, the loss of information about the uncertainty associated to this construction may lead to misleading interpretation of risk estimates. Finally, machine learning algorithms are both relevant and efficient approaches to deal with a huge number of stressors.

In this work, we focused on the specific problem of estimating the combined health effect—in terms of disease excess risk—of a few but highly correlated environmental exposure covariates, from a censored survival outcome. We opted for the PRM models. They are infinite mixture models that link a disease outcome to a set of correlated covariates through cluster membership. They are based on a Dirichlet process mixture as an attribution sub-model. By capturing the heterogeneity among the covariates, the PRM models allow both identifying specific patterns of covariate values—called covariate profiles—that are representative of a subpopulation (i.e., a cluster) and associating them with the disease outcome via a regression model. Then, inferring this probabilistic model allows both simultaneously identifying fine exposure profiles based on several correlated covariates, clustering individuals with similar risks and similar exposure characteristics and estimating the associated risk for each cluster. This joint modeling approach allows to rigorously capture uncertainty on all estimated parameters included in the different submodels. Compared to LCA and k-means algorithm, one of the principal motivations for PRM models is that the disease outcome influences cluster membership so that they can inform each other. Thus, the disease outcome may guide inference toward the most relevant clustering structures and is not only used during post-treatments. Another motivation for PRM models is that the number of clusters is unknown and informed by the data. Moreover, fitting PRM models under the Bayesian paradigm offers additional advantages. First, it allows dealing with the numerous latent variables included in these complex models and getting probabilistic answers to the studied question. Second, all uncertainty, including uncertainty associated with the clustering of the individuals, is reflected in credible intervals of risk parameters. Third, it provides the possibility to include external information on parameters in the form of prior distributions which is particularly useful when some unknown quantities of interest are not or only poorly informed by the data. Finally, it allows predicting the disease risk of a multi-exposed individual while conserving the uncertainty of estimated parameters. These models have already been employed in a variety of fields including genetics (40), environmental epidemiology (37, 41–44) and occupational epidemiology (45, 46) but never in ionizing radiation epidemiology. Note that an R package called PReMiuM (47) implements the Bayesian inference of PRM models for Gaussian, binary, ordinal, categorical, Poisson, and censored survival outcomes based on a Weibull distribution.

We extended the class of PRM models to deal with a censored survival outcome following an instantaneous excess hazard ratio model. This class of survival models is commonly used to estimate cancer risks in ionizing radiation epidemiology (48) but is not implemented in the PreMiuM package. The Bayesian inference of the proposed PRM model is conducted with a specific adaptive Metropolis-Within-Gibbs algorithm, implemented in Python and including three label switching moves. To illustrate our point, we applied our PRM model to the specific problem of estimating the risk of death by lung cancer among multi-exposed French uranium miners. Indeed, in the context of their work, underground uranium miners are simultaneously exposed to radon, external γ-ray and uranium dust (as well as other chemical and physical agents). Interestingly, these three sources of radiation exposure are highly correlated to each other in the French cohort of uranium miners. Actually, they are associated with the same initial phenomenon of disintegration of uranium, which is ubiquitous in uranium mines (49). Moreover, at this stage, an additive or synergic effect of co-exposure to these various radiological components on lung cancer risks cannot be excluded. Until now, most of the epidemiological studies on the French cohort of uranium miners have focused on studying the association between a chronic and low-dose exposure to radon and lung cancer mortality, as if radon—that is considered to be the second leading cause of lung cancer after smoking (50)—had an isolated effect. An EWAS approach was performed where the association between each single source of ionizing radiation and the risk of death by lung cancer was estimated separately, using a Poisson regression model. It showed that each source of ionizing radiation was significantly associated to a higher risk of death by lung cancer, in the French cohort of uranium miners (28). We propose to treat the multicollinearity issue in this case study, using our proposed Bayesian PRM model. Up to our knowledge, this is the first application of Bayesian PRM models to deal with highly correlated co-exposure in ionizing radiation epidemiology.

2. MATERIALS AND METHODS

2.1. Study Population

The French cohort of uranium miners is a retrospective cohort whose characteristics, sources of data and methods of data collection (e.g., vital status, causes of death, ...) have been described previously (28). Briefly, the last update included 5,086 males who were employed as uranium miners for at least 1 year in the CEA-COGEMA group between 1946 and 1990 and who were followed from 1946 to December 31, 2007. Uranium

TABLE 1 | Main characteristics of the post-55 French cohort of uranium miners.

No. of miners	3,377
Age at entry into study, mean [min, max]	28.3 [16.9, 57.7]
Duration of work in years, mean [min, max]	16.7 [1.0, 40.9]
Duration of follow-up in years, mean [min, max]	32.8 [0.1, 51.0]
Vital status, *n* (%)	
Alive <85 years old	2,412 (71.4)
Alive ≥85 years old	74 (2.2)
Death from lung cancer	94 (2.8)
Death from another cause	777 (23.0)
Lost to follow-up	20 (0.6)
Exposure to radon*	
Exposed miners, *n* (%)	2,910 (86.2)
Duration of exposure (in years), mean [min, max]	12.9 [1.0, 35.0]
Cumulative exposure (in WLM), mean [min, max]	17.8 [0.003, 128.4]
Exposure to γ-rays*	
Exposed miners, *n* (%)	3,240 (95.9)
Duration of exposure (in years), mean [min, max]	13.2 [1.0, 36.0]
Cumulative exposure (in mSv), mean [min, max]	54.9 [0.2, 470.1]
Exposure to uranium dusts*	
Exposed miners, *n* (%)	2,746 (81.3)
Duration of exposure (in years), mean [min, max]	12.9 [1.0-35.0]
Cumulative exposure (in kBq·m^{-3}·h), mean [min, max]	1.64 [0.01, 10.4]

**Results only on measured exposures.*

miners are simultaneously exposed to three sources of ionizing radiation: radon and its short-lived decay products (simply called radon hereafter), external γ-ray and uranium dusts. In the French cohort of uranium miners, the annual exposures to radon were assessed from 1946. On the other hand, the routine recording of occupational annual exposures to external γ-ray and uranium dust only began in 1956 in the French mines, following the introduction of radiation protection measures like the introduction of forced ventilation. In this paper, the study population was thus restricted to the so-called post-55 subcohort, in order to have simultaneous exposure measurements for the three sources of ionizing radiation. This subcohort included 3,377 miners from the original cohort who were first employed after December 31, 1955. At the end of follow-up, 94 miners had died of lung cancer. An age limitation of 85 years for follow-up was fixed due to the imprecision in determining the exact cause of death in those occurring after the 85th birthday (28). Main characteristics of the post-55 subcohort are recorded in **Table 1**.

2.2. Multiple Exposure Assessment, Proxy Variables, and Multicollinearity

In the French cohort of uranium miners, information on radon, external γ-ray exposures and uranium dusts exposure was assessed individually for each year of employment, but the method of measurement changed over time. Between 1946 and 1955, there was no systematic exposure assessment in the French uranium mines. Therefore, the annual radon exposure, expressed in working level months (WLM), was retrospectively reconstructed by a group of experts for this period, based on environmental measurements performed in the mines and information concerning the miners' type of work and location. Then, from 1956, the individual radon exposure was recorded systematically, following the new radiation protection measures which were set up at this date. More specifically, from 1956 to 1982, individual radon exposure was assessed from monthly ambient concentration measurements and information about the miners' activity (i.e., job type, location, and time spent at each location). From 1983, annual radon exposure was individually recorded, using personal dosimeters integrated to the Individual System of Integrated Dosimetry (ISID). Personal dose equivalents due to γ-ray exposures, expressed in millisieverts (mSv), were recorded individually since 1956, using two different types of personal dosimeters, depending on the calendar period: personal film badge dosimeters (CEA PS1 type) from 1956 to 1985 and personal thermoluminescence dosimeters (TLDs) integrated to the ISID from 1986 onwards. Finally, the annual exposure to long-lived radionuclides arising from uranium ore dust, expressed in Becquerels per cubic meter hour (Bq·m^{-3}·h), was retrospectively reconstructed for the period 1956–1958 (51). It was then assessed from monthly ambient measurements from 1959 to 1982. From 1983, individual measurements were collected with the ISID.

Potentially relevant proxy variables are also available in the French cohort of uranium miners to reflect the uranium miners' working conditions and any other occupational exposures. First, there are the job types of French uranium miners which are classified into five categories : (1) hewers before mechanization, (2) hewers after mechanization, (3) other underground work before mechanization, (4) other underground work after mechanization, and (5) surface worker. The mechanization of work in the French uranium mines began in 1977, with the introduction of trucks. Thus, from 1977 onwards, the uranium miners' working conditions can be assumed to be less physically demanding compared to the period before mechanization. But on the other hand, an additional occupational exposure to diesel, recognized as a lung carcinogen (52), appeared in the mines at the same period. Finally, hewers were assumed to have a more physically demanding labor and harsher working conditions than other underground and open-pit uranium miners. An additional proxy for uranium miners' working conditions is the working location which includes four different mining districts: (1) Vendée, (2) Crouzille, (3) Forez, and (4) Hérault. Actually, the type of uranium deposits (i.e., granitic, sedimentary) has an impact on the undergrounds galleries of the uranium mines and then, on the miners' working conditions. Note that the type of uranium deposits depends on the mining district. It is granitic for the districts of Vendée, Crouzille, and Forez and sedimentary for the district of Hérault (53).

An estimation of Pearson's correlation coefficients, using all the available pairs of cumulative exposures to two different sources of ionizing radiation, clearly shows that the assessed values of occupational exposures to radon, γ-ray and uranium dusts are highly correlated in the post-55 subcohort of French uranium miners. Indeed, the estimated coefficients are pretty high. It is equal to 0.90 between radon and γ-ray, to 0.82 between uranium dusts and γ-ray and to 0.78 between radon and

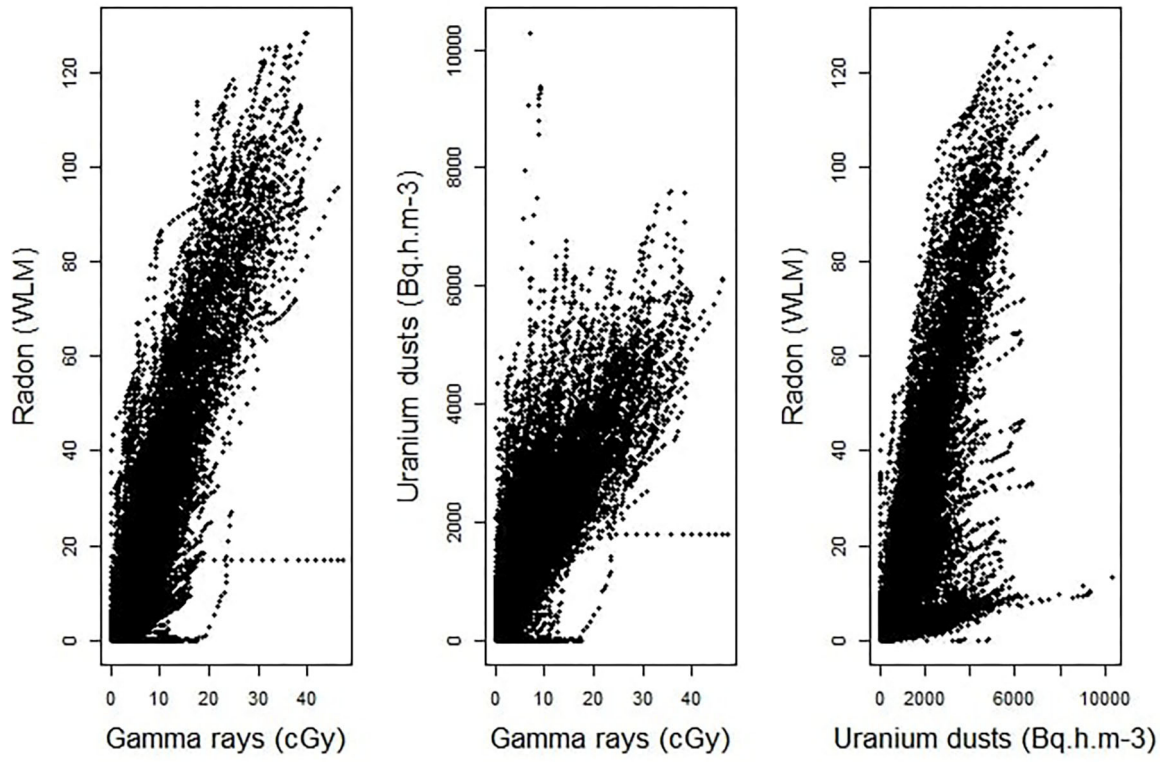

FIGURE 1 | Scatter plots of the observed cumulative exposures to γ-rays and radon (left-hand panel), γ-rays and uranium dust (at the center), radon and uranium dust (right-hand panel).

uranium dusts. **Figure 1** displays the scatter plots of the observed cumulative exposures to the three sources of ionizing radiation. It clearly confirms that we are faced with a multicollinearity issue, requiring the use of a suitable statistical approach to estimate the combined effect of these three radiological exposures, the job type and the localization of the mine on the risk of death by lung cancer in the post-55 subcohort of French uranium miners.

2.3. Model Formulation

To deal with multicollinearity in the specific context of estimating the combined effect of a few but highly correlated exposure covariates, we opted for a Bayesian profile regression mixture PRM model. In this approach, three submodels must be specified and linked, through conditional independence assumptions: the disease, the exposure and the attribution submodels. A Bayesian PRM model is a hierarchical model that allows jointly describing: (a) the association between a disease outcome (e.g., the age at death by lung cancer of a miner) and an exposure profile (disease sub-model); (b) the probability distribution of the different covariates of interest in each cluster, in order to characterize specific exposure profiles (exposure sub-model); and (c) the random assignment of an individual to a given profile (or cluster) (attribution sub-model).

The disease sub-model conventionally used in radiation epidemiology is an Excess Hazard Ratio (EHR) model. Let S_i be the age (in days) at death by lung cancer of miner i, $i \in 1, 2, ..., n$ where n is the total number of miners. Let R_i be the right-censored age defined as the earliest age of miner i among age at death by a cause other than lung cancer; age on December 31, 2007; age in days corresponding to his 85th birthday and age until loss to follow up. For each miner i, the observed outcome of interest can therefore be represented by the non-negative continuous variable $Y_i = min(S_i, R_i)$ and the binary variable δ_i where $\delta_i = 1$ if $S_i \leq R_i$ (i.e., miner i died of lung cancer at age $Y_i = S_i$) and $\delta_i = 0$ if $S_i > R_i$ (i.e., miner i "would have died of lung cancer" after age R_i). The instantaneous hazard rate of death by lung cancer of miner i at age t, noted $h_i(t)$ is defined by

$$h_i(t) = h_0(t) \cdot (1 + \beta_{C_i}) \tag{1}$$

Baseline hazard $h_0(t)$ is the instantaneous risk of death by lung cancer at age t by not exposed profile (the reference cluster of miners not exposed to ionizing radiation), C_i is the cluster label of miner i and β_c is the instantaneous excess risk of death by lung cancer of the cluster c. Thus, two miners belonging to the same cluster c have the same risk of death by lung cancer. Note that $\forall c$, β_c is subject to the constraint $\beta_c > -1$ to ensure the positivity of $h_i(t)$.

Following Hoffmann et al. (48), $h_0(t)$ is assumed to be piece-wise constant on four age intervals for which values of baseline hazard are assumed to be constant. These four intervals correspond to a partition of age axis defined by before 40 years

old, between 40 and 55, between 55 and 70 and finally after 70 years old. The corresponding four constants of baseline hazard are denoted by $\lambda_1, \lambda_2, \lambda_3$ and λ_4.

When modeling lung cancer mortality in the French cohort of uranium miners, we considered the age at death by lung cancer of each miner as disease outcome. Indeed, Kleinbaum suggested to favor age as time-scale whenever age at event is likely to have a larger effect on the hazard than time-on-study (54). Moreover, based on previous findings on cohorts of uranium miners, we can assume that, contrary to the attained age of a miner, the timing of study initiation has no inherent meaning in terms of the risk of lung cancer mortality in the cohort. Finally, several authors recommend to favor age as time-scale whenever possible since the modeling of the effect of age can be complex and prone to misspecification errors. Based on these arguments, we chose attained age as time scale. Thus, age is still accounted for in the disease model.

The exposure sub-model defines clusters based on covariates levels and on a similar risk to lung cancer death. Probability distribution of the covariates conditionally to a cluster is introduced. The different covariates considered for clusters include cumulative radiation exposures and other characteristics of miners. Details on these covariates are the following:

- Cumulative exposure of occupational radon X_i^R, γ-rays X_i^G, and uranium dust X_i^D during the whole following up period of miner i;
- Job type J_i most occupied by miner i. This categorical variable have five modalities: (1) hewers before mechanization, (2) hewers after mechanization, (3) other underground work before mechanization, (4) other underground work after mechanization, and (5) surface work;
- Age at first exposure A_i of miner i. Sensibility of radiation can be function of age of exposure (55);
- Localization of the mine M_i. We distinguished Hérault mine and the others based on the deposit's type;
- Exposure duration T_i of the miner i. Four duration periods with similar number of miners are considered: miners who were exposed 5 years and less, 6–12 years, 13–18 years, and finally those who have been exposed for at least 19 years.

The probability distribution of each covariate depends on parameters which are function of the cluster c. We assumed lognormal distributions $\text{LogN}(\mu_c^X, \sigma_c^X)$ for positive and continuous variables and multinomial distributions $\text{Multinomial}(p_c^X)$ for categorical variables.

The different distributions are the following:

$$\begin{cases} X_i^R|C_i = c, \mu_c^R, \sigma_c^R & \sim LogN(\mu_c^R, \sigma_c^R) \\ X_i^G|C_i = c, \mu_c^G, \sigma_c^G & \sim LogN(\mu_c^G, \sigma_c^G) \\ X_i^P|C_i = c, \mu_c^P, \sigma_c^P & \sim LogN(\mu_c^P, \sigma_c^P) \\ A_i|C_i = c, \mu_c^A, \sigma_c^A & \sim LogN(\mu_c^A, \sigma_c^A) \\ J_i|C_i = c, p_c^J & \sim Multinomial(p_c^J) \\ M_i|C_i = c, p_c^M & \sim Multinomial(p_c^M) \\ T_i|C_i = c, p_c^T & \sim Multinomial(p_c^T) \end{cases} \quad (2)$$

The attribution sub-model associates miner i to a cluster C_i based on the probability ϕ_c of belonging to the cluster c. Let C_{max} be the maximum number of clusters, $\phi = (\phi_1, \phi_2, ..., \phi_{C_{max}})$ defines the vector of the probabilities of assignment to each cluster among the C_{max} ones. The parameter vector ϕ follows a Dirichlet process. Due to the Dirichlet process, the number of non-empty groups is not arbitrarily fixed but estimated, only the maximum number of clusters C_{max} is given. The construction of these mixing weights $\phi = (\phi_1, \phi_2, ..., \phi_{C_{max}})$, also called "stick-breaking," is the following:

$$V_c \sim Beta(1, \alpha), c \in \{1, ..., C_{max} - 1\} \quad (3)$$

$$\phi_c = V_c \cdot (1 - \sum_{k=1}^{c-1} \phi_k), c \in \{1, ..., C_{max} - 1\} \quad (4)$$

$$\phi_{C_{max}} = 1 - \sum_{k=1}^{C_{max}-1} \phi_k \quad (5)$$

The number of non-empty clusters is guided by α. A small value of α reduces the probability to have a large number of non-empty clusters, and respectively. This "stick-breaking" construction approximates the infinite cluster model with a finite one. The value of C_{max} has to be chosen large enough to give a good approximation but small enough to avoid unnecessary calculations. C_{max} should be set so that the probability $\phi_{C_{max}}$ is expected to be small (56). The choice of C_{max} is highly affected by the value of α, and for α up to 10, the probability $\phi_{C_{max}}$ is negligible with C_{max} equals to 50 (57). Some guidelines and more detailed description are given in Molitor et al. (37).

2.4. Prior Distributions and Bayesian Inference

2.4.1. Prior Distributions

Prior distributions are chosen poorly informative except for parameters involved in baseline hazard, in stick-breaking prior as well as means of exposure for which external information were available.

Thus, normal centered distributions with large variance were considered for the risk parameters β_c and for the means of age at first exposure μ_c^A (on log scale) in each group $c, c = 1, ..., C_{max}$. Large Uniform distributions were considered for the geometric standard deviation parameters of the lognormal distributions σ_c^R, σ_c^G, σ_c^P, and σ_c^A. Dirichlet prior distributions with parameters equal to 1/2 were considered for the parameters of multinomial distributions, namely p_c^J, p_c^M, and p_c^T.

Concerning the mean of γ-rays μ_c^G, radon μ_c^R, and uranium dust μ_c^P exposures (on log scale), information are available from German uranium miner cohort (58). Normal prior were considered for μ_c^G, μ_c^R and μ_c^P with means and variances based on exposure levels of this cohort.

As parameters involved in baseline hazard are poorly informed by data in particular for young miners, external

TABLE 2 | Prior probability distributions assigned to the unknown parameters of a Bayesian PRM model including the disease sub-model, the exposure sub-model and the attribution sub-model.

	Parameter	Family	
Disease sub-model	β_C	Normal	N (0, 10^6)
	λ_1	Gamma	G (23.7, $4.9 \cdot 10^8$)
	λ_2	Gamma	G (35.5, $2.6 \cdot 10^7$)
	λ_3	Gamma	G (88.1, $1.6 \cdot 10^7$)
	λ_4	Gamma	G (29.7, $3.2 \cdot 10^6$)
Exposure sub-model	μ_c^G	Normal	N (0.10, 2.25)
	μ_c^R	Normal	N (−2.3, 8.08)
	μ_c^P	Normal	N (1.01, 11.79)
	μ_c^A	Normal	N (0, 10^6)
	$\sigma_c^G, \sigma_c^R, \sigma_c^P, \sigma_c^A$	Uniform	U [0,100]
	p_c^J, p_c^M, p_c^T	Dirichlet	D [0.5, …, 0.5]
Attribution sub-model	α	Uniform	U [0.3, 10]

data on lung cancer mortality among men in France between 1968 and 2005 were used to specify the informative prior gamma distributions on the parameters λ_1, λ_2, λ_3, and λ_4 defining the baseline risk of death by lung cancer among French uranium miners (assumed constant by age intervals). Finally, as recommended by Molitor et al. (37), we used a uniform distribution on the interval [0.3, 10] for the parameter α which influences the number of non-empty clusters *a posteriori*. All details are given in **Table 2**.

2.4.2. Bayesian Inference

Figure 2 shows the directed acyclic graph for the full hierarchical model combining the disease sub-model, the exposure sub-model and the attribution sub-model. R package "PReMiuM" already exists to implement the Bayesian profile regression (47) for Bernoulli, Binomial, Poisson, Normal, categorical response as well as Weibull survival model. Unfortunately, the EHR survival model is not a possible option in this package. Thereby, a Markov Chain Monte Carlo (MCMC) algorithm was implemented in Python to sample from the joint posterior distribution of all unknown parameters and latent variables. Simulations were performed in order to validate the code, results of these simulations can be found in the **Supplementary Material**. We used a Metropolis-within-Gibbs algorithm (59) to conduct the Bayesian inference, as full conditional distributions were not always analytically tractable. An adaptive phase of Metropolis-Hastings steps, which is necessary to improve the convergence and the efficiency of the algorithm, updates the variance of each proposal distribution to target an acceptance rate of 40% for single parameters and 20% for vectors (59). The parameters and the latent variables were updated separately. We ran 100 steps of 100 iterations for the adaptive phase, then 10,000 iterations were dropped for the burn-in phase and finally 150,000 additional iterations were run. To decrease within-chain autocorrelations, we thinned the sample by storing only every 20 iterations. Posterior sample of each unknown quantity therefore contains 7,500 values. A particular attention was done on the convergence toward local modes by considering different initial values for parameter α directly linked to the number of non-empty clusters. Moreover, as suggested by Liverani et al. (47), we introduced three label switching moves in order to try to best avoid convergence to local mode (60, 61). The use of this three label switching moves is justified by the weak identifiably of the clusters labels leading to multiple modes of the posterior distributions of the ϕ_c's. To explore multimodal posterior distributions, Papaspiliopoulos and Roberts (60) introduce two label switching moves which allow moves particularly at the beginning of MCMC algorithm. To improve ability of moves, Hastie et al. (61) add a third one . The basic idea of moves is to switch two labels j and k according to a probability min (1,r_{jk}). Details on r_{jk} are given in **Table 3**. Main characteristics of these three moves are the following. The first move has high acceptance probability of switching j and k when weights ϕ_j and ϕ_k are close. On the other hand, two clusters with similar number of miners are rarely switched. The two other moves propose only switch between two neighboring clusters namely j and $j+1$. When label switching is accepted according to the second or third switching moves, the respective beta components V involved in the stick-breaking procedure are simultaneously modified (and consequently the weights ϕ). The second move corresponds to high acceptance probability for neighboring clusters including different number of miners. For the third label switching, the respective beta components V are modified so that the corresponding weights ϕ_j and ϕ_{j+1} are close to their expectation conditional on these new labels. Details on r and V are given in **Table 3**. For the three switching procedures, corresponding excess risks β and other cluster specific parameters are simply exchanged when move is accepted.

2.5. Post-treatment

As described in Molitor et al. (37) and in Liverani et al. (47), the post-treatment is realized after running the MCMC algorithm. We chose to determine an optimal partition corresponding to a partition sampled from our MCMC algorithm. The main advantage of using a sampled partition is to avoid difficult problems linked to clusters labels which could be different between iterations. There are different techniques to obtain this optimal partition. We decided to use the post-processing approach based on a posterior similarity matrix. Another possibility could have been to use the MAP estimate corresponding to the partition leading to the highest value of the marginal posterior distribution. As mentioned by Liverani et al. (47), the MAP estimate is more sensitive to the Monte Carlo error than the techniques based on the similarity matrix. If K is the number of iterations, K binary square matrices S_k of dimension $n \times n$ are determined at each iteration k where $S_k(i,j) = 1$ if miners i and j share the same cluster at iteration k of the MCMC sampler, and 0 if not. The mean S of these K matrices $(S_1, ..., S_K)$ thus contains the proportion that two miners belong to the same cluster during MCMC sampler. The estimated best partition called C^{best} is the one that minimizes the least-squared distance to matrix S. C^{best} is a vector such

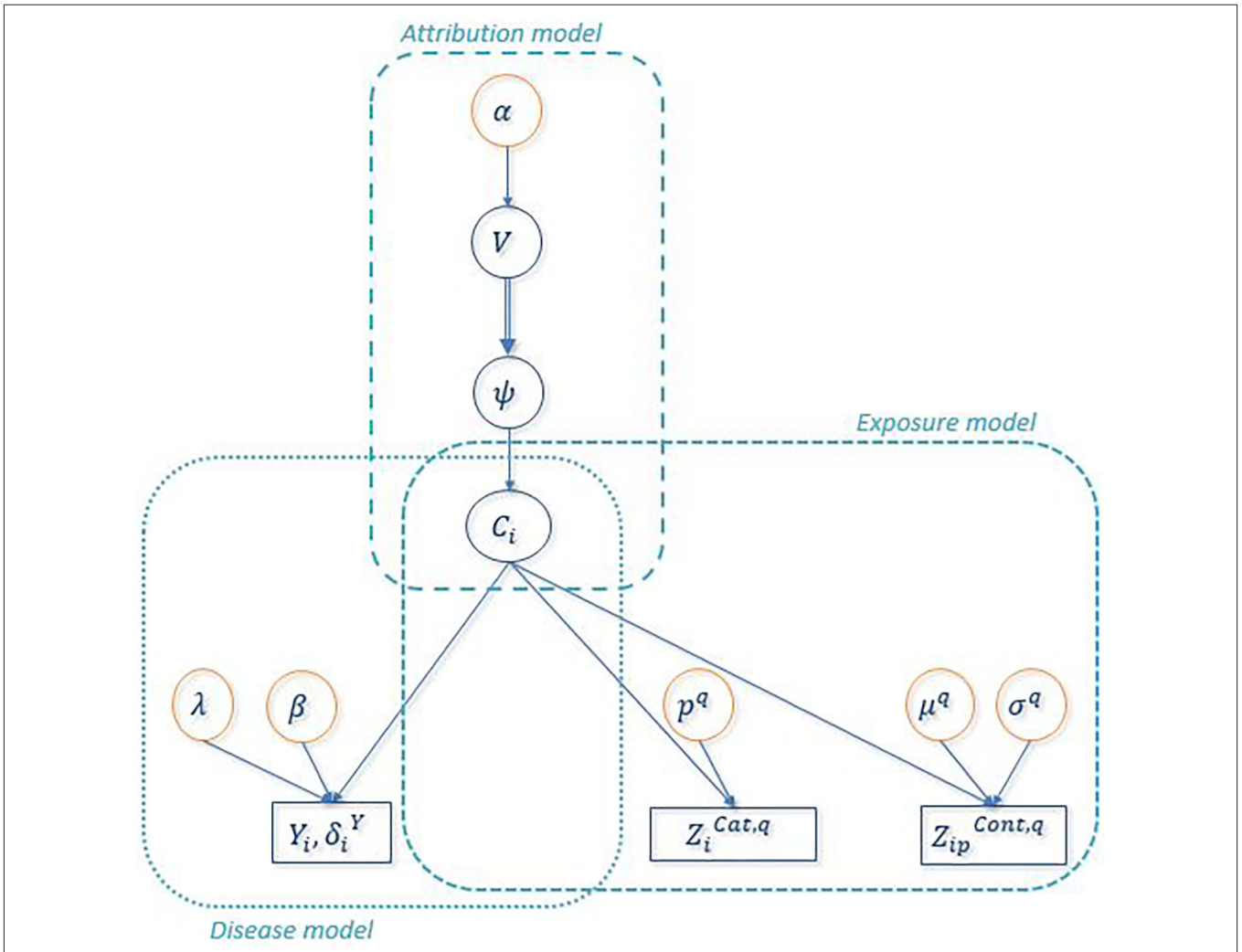

FIGURE 2 | Directed Acyclic Graph associated to the full Bayesian PRM model. Circles indicate unknown quantities and rectangles indicate observed variables. Single arrows indicate oriented probabilistic links between two quantities and double arrows indicate oriented deterministic links between two quantities. $Z_i^{Cat,q}$ denotes the observed value of any categorical covariate q for uranium miner i and $Z_i^{Cont,q}$ denotes the observed value of any continuous covariate q of uranium miner i.

that c_i^{best} is equal to the cluster label of miner i in this optimal partition.

Posterior distributions of parameters are obtained conditionally to the best partition C^{best}. Generally speaking, if θ_c denotes a parameter depending of cluster c, a sample from posterior distribution of parameter θ_c conditionally to partition C^{best} is $\{\bar{\theta}_{c,k}, k = 1, ..., K\}$ such that

$$\bar{\theta}_{c,k} = \frac{1}{n_c} \sum_{i\,:\,c_i^{best}=c} \theta_{c_i^k,k} \tag{6}$$

with n_c the number of uranium miners in cluster c and c_i^k the cluster of miner i at iteration k. This post-processing procedure is apply on all parameters depending on cluster label involved in the three sub-models that is $(\beta, \mu^R, \sigma^R, \mu^G, \sigma^G, \mu^P, \sigma^P, \mu^A, \sigma^A, p^J, p^M, p^T)$ as well as the weights ϕ of clusters.

3. RESULTS

3.1. Univariate and Multivariate EHR Model Without Clustering

In order to assess impact of multicollinearity on EHR model, classical Excess Hazard Ratio model was implemented without clustering procedure. The instantaneous hazard rate of death by lung cancer of miner i at time t $h_i(t)$ is here directly function of exposures, without taking into account of multicollinearity. A first approach consists in considering each radiation source separately and secondly, to include simultaneously the three ones. Posterior median and 95% credible interval of β are obtained in each case. With only one exposure, $h_i(t)$ is then

TABLE 3 | Label switching moves.

	Move 1	Move 2	Move 3
r_{jk}	$(\frac{\phi_j}{\phi_k})^{n_k-n_j}$	$\begin{cases} \frac{(1-V_{j+1})^{n_j}}{(1-V_j)^{n_{j+1}}} & k=j+1 \\ 0 & \text{otherwise} \end{cases}$	$\begin{cases} (\frac{\phi^+}{\phi_{c+1}R_1+\phi_c R_2})^{n_j+n_{j+1}} R_1^{n_{j+1}} R_2^{n_j} & k=j+1 \\ 0 & \text{otherwise} \end{cases}$
V'_l	V_l	$\begin{cases} V_{j+1} & l=j \\ V_j & l=j+1 \\ V_l & \text{otherwise} \end{cases}$	$\begin{cases} \frac{\phi'_j}{\prod_{k<j}(1-V_k)} & l=j \\ \frac{\phi'_{j+1}}{(1-V'_j)\prod_{k<j}(1-V_k)} & l=j+1 \\ V_l & \text{otherwise} \end{cases}$

The switching between labels j and k is accepted with probability min(1, r_{jk}). If label switching is accepted, V′ is the new value of beta-component V. n_j is the number of miners in cluster j.

$\phi^+ = \phi_j + \phi_k$, $\phi' = \phi_{j+1}\frac{E(\phi_j|C',\alpha)}{E(\phi_{j+1}|C,\alpha)} + \phi_j\frac{E(\phi_{j+1}|C',\alpha)}{E(\phi_j|C,\alpha)}$,

$R_1 = \frac{1+\alpha+n_{j+1}+\sum_{l>j+1} n_l}{\alpha+n_{j+1}+\sum_{l>j+1} n_l}$ and $R_2 = \frac{\alpha+n_j+\sum_{l>j+1} n_l}{1+\alpha+n_j+\sum_{l>j+1} n_l}$

defined by $h_i(t) = h_0(t) \cdot (1 + \beta \cdot X_i)$ where baseline hazard h_0 is assumed piece-wise constant as previously, β the excess risk of death by lung cancer associated to cumulative exposure X and X_i the cumulative exposure of miner *i*. When considering single exposure, X can be X^R, X^G or X^D for respectively radon, γ-rays and uranium dust. Posterior medians of β and associated 95% credible intervals are 2.7 [1.1 , 5.2], 0.78 [0.28 , 1.67], and 3.34 10^{-2} [1.07 10^{-2} , 7.00 10^{-2}] for respectively radon, γ-rays and uranium dust. As zero is excluded from each credible interval, the excess risk of death by lung cancer is strictly positive for each exposure. When considering simultaneously the three exposures of ionizing radiations, then $h_i(t) = h_0(t) \cdot (1 + \beta_R X_i^R + \beta_G X_i^G + \beta_D X_i^D)$. Posterior medians of β_R, β_G and β_D with associated credible intervals are now 2.7 [−0.2 , 5.8], 0.00 [−0.39 , 1.17], and −0.15 10^{-2} [−1.66 10^{-2}, 3.81 10^{-2}], respectively. None of the exposures were significantly associated to the risk of death by lung cancer anymore. This result highlights the issue of multicollinearity of the exposures in our case. When considering exposure one per one, the values of estimated risks are difficult to interpret because could also be due to confusing effect from the other radiation sources which are both correlated with death by lung cancer and with studied exposure. As expected, introduction of simultaneous exposures leads to huge imprecision and consequently to no significant associations for some radiological exposures.

3.2. Convergence Toward Local Mode Under PRM Model

PRM model as defined in section 2.3 is implemented on the post-55 sub-cohort. As already mentioned in Liverani et al. (47), parameter α in Equation (3) is directly linked to the number of non-empty clusters. Under PRM model, this number is also estimated (only the maximum number of clusters C_{max} is fixed) and a particular attention has to be made on local convergence issue even if label switching moves are introduced. To assess a convergence toward a local mode, MCMC samplers were run from different initial values of α. Initial values are chosen from 0.5 to 9.5 covering the prior support of α. For a given initial value, the number of non-empty clusters systematically converges to a single value without moves during the sampler, while there is no convergence issue for the other parameters. Results are presented in **Figure 3** where the number of non-empty clusters takes four possible values from 5 to 8 (including the cluster of non-exposed uranium miners) according to the different initial values of α. Local convergence issue is also clearly suspected despite the three label switching procedure. An explanation could be the low proportion of miners died from lung cancer. This proportion is indeed near 3% giving a low signal to infer the risk between clusters and lung cancer. Consequently, a restricted profile regression mixture RPRM model is considered where the number K of non-empty clusters is fixed. The attribution sub-model defined section 2.3 is then simplified where the weights ϕ have now a fixed number K of component. We ran MCMC algorithm from two different sets of initial values. A solution to choose K could have been to choose one value among the four values suggested by **Figure 3**. Deviance information criteria (DIC) (62) as well as Watanabe-Akaike, also called Widely Applicable, Information Criterion (WAIC) (63) are presented in **Table 4** for K from 5 to 8. These two criteria are concordant in favor of 8 non-empty clusters. As penalized deviance is well-known to possibly select most complex models, we prefer to present results with K equal to 8 non-empty clusters but also to compare with the three other RPRM models corresponding to 5, 6, or 7 non-empty clusters (results given in the **Supplementary Material**). Note that when the number of non-empty clusters is fixed, no convergence issue was found for all other parameters.

3.3. Results With Fixed Number of Non-empty Clusters

Results for eight clusters model are summarized on **Figures 4**, **5** while results for the 5–7 clusters can be found in the **Supplementary Material**.

On the left of **Figure 4**, number of miners (top) and number of cases (bottom) per cluster are represented except for the cluster of non-exposed uranium miners. The seven resulting clusters are denoted by A to G. The order of clusters representation follows the order of the associated estimated risk of each cluster. Thus, cluster A corresponds to the lowest estimated posterior median of β and cluster G to the higher one. The number of miners varies from 285 to 633 and the number of cases per cluster from 4 to 30. On the right of **Figure 4**, results on the excess risk of death by lung cancer of each cluster (β_A to β_G) are given. Boxes correspond to the posterior quartiles of β and the whiskers extend to the posterior 2.5% and 97.5% quantiles illustrating 95% credible interval of β. Colors indicate whether posterior 95% credible interval of β is greater than zero (red) or include zero (blue). A cluster is called "significant high risk cluster" (or respectively "significant low risk cluster") if whiskers are >0 (respectively lower than 0). Two significant high risk clusters are here identified, namely clusters F and G. The posterior median excess risk of cluster G is estimated to 1.14 and to 0.66 for cluster F. Note that an excess risk of 1.14 means that miners belonging to this cluster have a risk multiplied by 2.14 compared to non-exposed uranium miners.

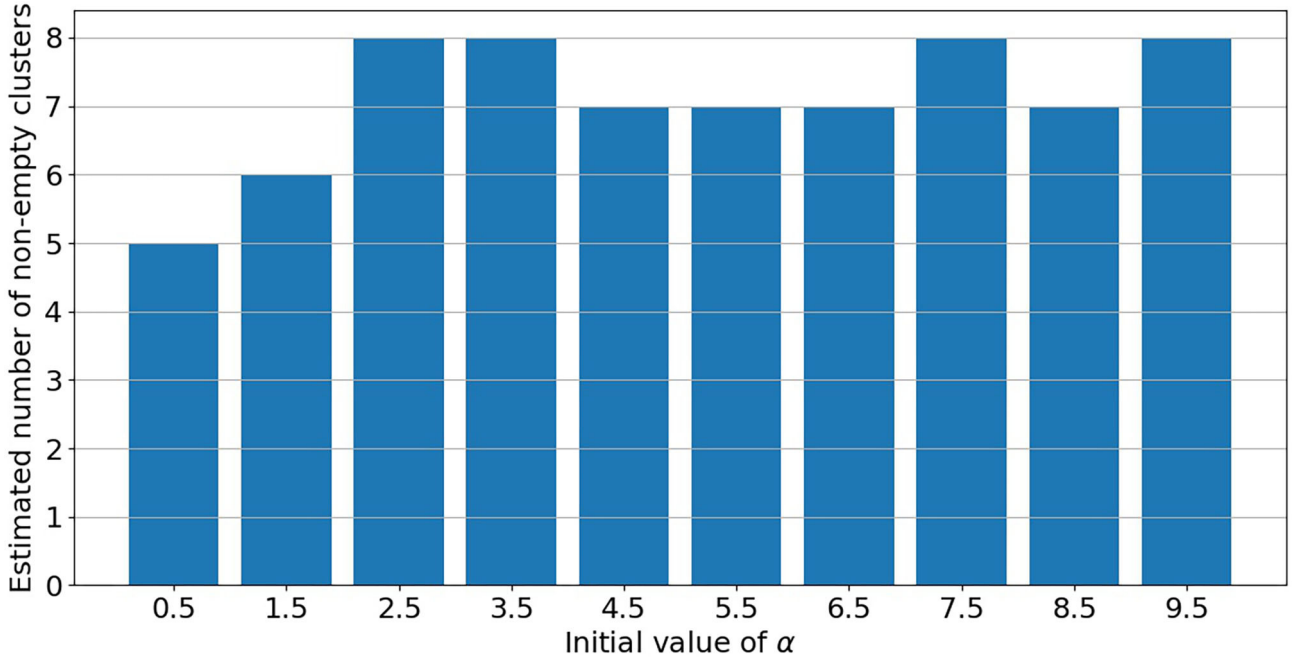

FIGURE 3 | Estimated number of non-empty clusters according to the initial value of α.

TABLE 4 | DIC and WAIC of Bayesian PRM model according to the fixed number K of non empty clusters.

Number K of non-empty clusters	DIC	WAIC
5	146,345	110,872
6	136,714	108,773
7	118,602	107,004
8	104,566	105,704

Characterization of each cluster in terms of covariates is illustrated on **Figure 5**. Each column corresponds to one covariate, cluster labels being specify on horizontal axis. For continuous covariates, such as cumulative exposures and age at first exposure, results on medians (e^{μ}) are on the top while results on standard deviation (on log scale) on bottom. For categorical covariates, such as Job type, Mine and Exposure duration, posterior distribution of probability of each category is shown. Boxes and whiskers are defined as previously. The two different colors, green and red, correspond to a 95% credible interval, respectively under or upper the global median on all miners (whatever the cluster) while blue color shows no particular values of the covariate for this cluster.

The cluster G with the highest risk of death by lung cancer corresponds to the most exposed uranium miners as credible intervals of the mean for cumulative radon, γ-rays and uranium dust exposure are high. They were mainly working before mechanization or as hewer after mechanization, not in Herault's mine, pretty old when they started working compared to the other groups and being exposed during long time (longer than 19 years). This cluster corresponds to the most difficult working conditions. This high risk cluster is found for 5, 6, or 7 non-empty clusters (see **Supplementary Material**). Its systematic identification is reassuring in terms of model validity since it is consistent with standard assumptions in the field.

The cluster F associated to the second highest risk of death by lung cancer is characterized by miners who were also highly exposed but less than in cluster G, worked as hewer after mechanization or other underground job before mechanization, not working in Hérault's mine, were young when they started working compared to the other groups and exposed more than 13 years. Working conditions of this second cluster can also be considered as difficult but less than those of cluster G in particular concerning hewer before or after mechanization and the duration of exposure a little lower. On the other hand, this second cluster highlights risk profile of miners who started to work early compared to the other groups. Results concerning this second cluster differ slightly depending on the fixed number of non-empty clusters (see **Supplementary Material**). Indeed, this cluster is associated to a positive excess risk which is significant for RPRM with K = 7, nearly significant with K = 6 but not significant with K = 5. Posterior medians of β_F and β_G as well as characteristics of these two clusters are very similar with K = 6 and K = 7 to those already found with K = 8. Concerning RPRM model with 5 non-empty clusters, results on cluster G are similar while posterior median of excess risk β_F and characteristics of cluster F are different. Indeed, this second cluster F not contains exactly the same number of uranium miners for different values of K. Almost 630 common miners belong the second cluster for all fixed number K except for K = 5 where there are approximately 250 miners more (cluster F in **Supplementary Figures 1, 2**). When comparing the 630 commons miners to these 250 miners,

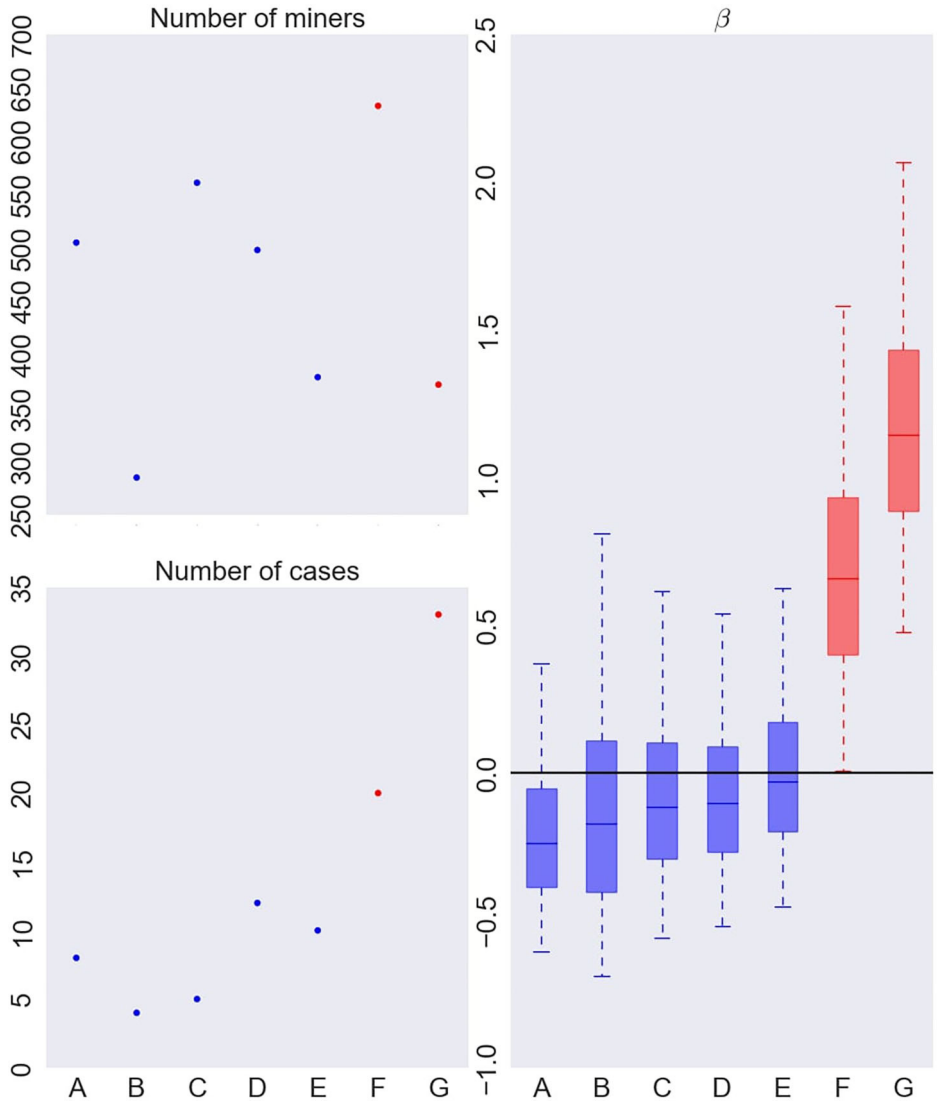

FIGURE 4 | Number of French uranium miners (top left), number of deaths by lung cancer (bottom left) and instantaneous excess hazard ratio (per 100 WLM) of death by lung cancer (β) in each cluster (right), when fitting a Bayesian RPRM model assuming 8 non-empty clusters from the French cohort of uranium miners. The cluster including non-exposed miners is not displayed. The boxes represent the three quartiles (1st quartile, median, and 3rd quartile) of the posterior distribution of β and the whiskers of the boxplots show the 95% credible interval of the posterior distribution for each group.

we notice that common miners received a higher cumulative exposure to radon and they were all working in other mines than Hérault's one. The 250 miners who differed with K = 5, have lower cumulative exposure to radon and slightly more than half of them worked in Hérault's mine. Finally, there are only two cases of lung cancer death among these 250 miners. The risk associated to the 630 common miners is also higher than that associated to 880 miners belonging the second cluster with the partition in 5 non-empty clusters. Consequently, this second cluster F is again significant or nearly significant with partitions in 6 or 7 clusters but not with the partition in 5 non-empty clusters. The posterior median of β_F is estimated near to the same value for 6 and 7 clusters than 8 clusters but near 0.3 for model with 5 clusters. Despite these differences, this second high risk cluster exists for all models with very near characteristics, in particular with less important cumulative exposures to radon, γ-rays and uranium dust exposure but with young age at the start of work.

We do not systematically observe an increasing risk corresponding to increasing exposure levels. It is particularly the case when focusing on cluster B (**Figure 5**). This cluster is associated to the second lowest risk whereas the miners in this cluster are highly exposed. The main differences compared to other clusters are the important proportion of uranium miners working in Hérault's mine and the period after mechanization. Modeling association between profiles and mortality allows to obtain finer interpretation of effect of exposure levels than studies including direct associations with exposures could not have done.

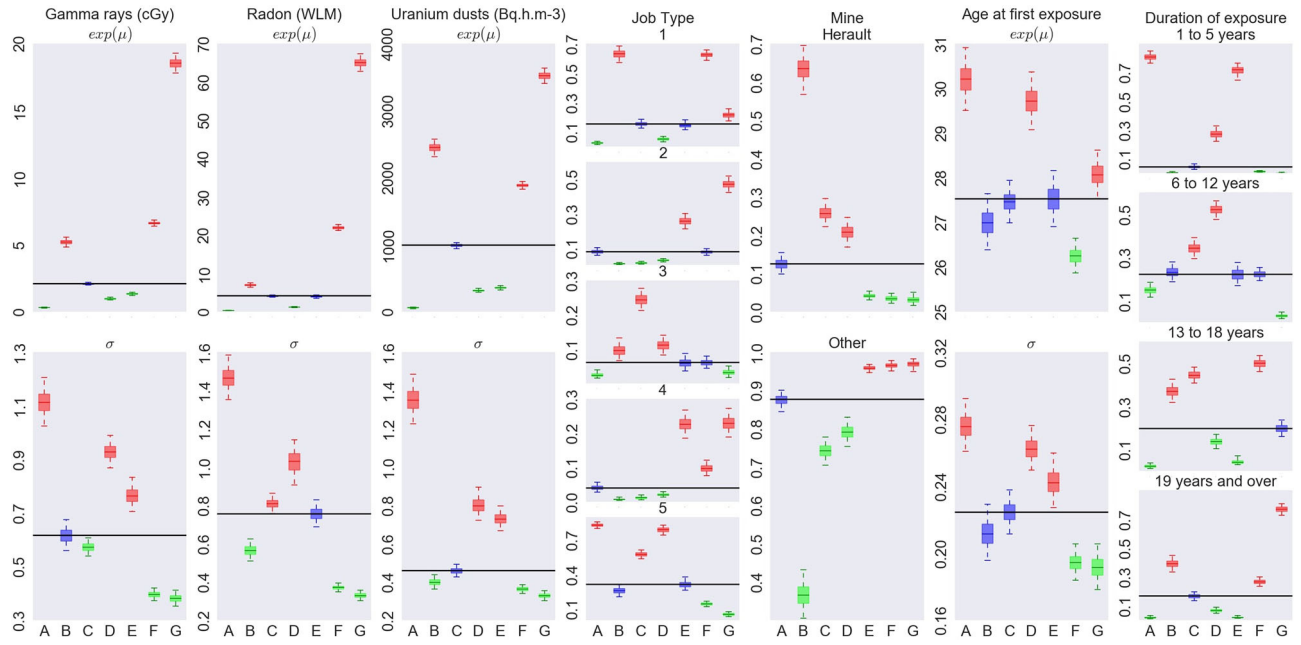

FIGURE 5 | Characterization of the exposure profiles associated to each cluster, when fitting a Bayesian RPRM model assuming 8 non-empty clusters. The cluster including non-exposed miners is not displayed.

4. DISCUSSION

In this work, we developed an original Bayesian PRM model based on an instantaneous excess hazard ratio model as disease submodel and a truncated Dirichlet process mixture as attribution submodel. This model was applied to the estimation of the lung cancer mortality associated with multiple cumulative exposures to ionizing radiations as well as any other occupational exposures through proxy variables (i.e., job types and localization of the mines). An adaptive Metropolis-Within-Gibbs algorithm, including three label switching moves, was implemented in Python to sample from the joint posterior distribution of all the unknown parameters and latent variables. Simulations were performed in order to validate the implemented algorithm (Results can be found in the **Supplementary Material**).

After fitting our full Bayesian PRM model to the post-55 sub-cohort of French uranium miners, the target posterior distribution was suspected to be highly multi-modal and our MCMC algorithm to converge to local modes. Consequently, Bayesian RPRM models were also fitted to the post-55 sub-cohort, where the number K of non-empty clusters was fixed to 5, 6, 7, and 8. In this paper, we focused on the results provided by the Bayesian RPRM with 8 non-empty clusters (including the cluster of non-exposed miners) that led to very interesting clusters of miners. Two of them were associated with a strictly positive and statistically significant EHR of death by lung cancer. The first group (EHR = 1.4, 95%IC = [0.60, 2.60]) corresponded to the miners the most highly exposed to radon, gamma rays and uranium dust and for more than 19 years (mainly before mechanization or as hewer after mechanization not in the mine located at Herault). The second group (EHR = 1.2, 95%IC = [0.17, 2.80]) corresponded to the miners who were very young when first exposed and who were highly exposed to radon, gamma rays and uranium dust for more than 13 years (mainly hewer after mechanization or other underground job before mechanization). Finally, the model showed that the group of miners who worked after the mechanization and mainly in the mine located at Herault (the only included uranium mine with sedimentary soil) had the second lowest risk whereas the miners in this cluster were highly exposed. Thus, this Bayesian RPRM model allowed providing an original, rich and fine interpretation of the potential association between the risk of death by lung cancer and specific radiation exposure profiles of French uranium miners, especially by modulating the effect of radiation co-exposures by other information, such as age at first exposure and duration of exposure. Results with the three other possible values of K from 5 to 7 are described in **Supplementary Material**.

Unfortunately, the target posterior distribution of our full Bayesian PRM model was suspected to be highly multi-modal, given the data available in the post-55 sub-cohort of French uranium miners. This could be due to a lack of signal in the database avoiding to strongly highlight, if it exists, an "optimal" partition of uranium miners (i.e., with the highest posterior probability). Additionally, the Bayesian PRM models have a large number of parameters and latent variables and, thus, in the specific context of a lack of signal in the available data, applying a MCMC algorithm might not be the most suitable Bayesian inference. As illustrated by Gelman et al. (64), due to the random walk of Gibbs sampler and Metropolis algorithm, the simulations can take a long time before moving to the target distribution. Particularly, for complex models with high dimensional target distribution, a random walk can remain local. Betancourt and Girolami (65) also illustrated that Gibbs samplers and Metropolis-Hastings algorithms explore the target

distribution slowly, and it get worse when the number of groups or levels increases. Although difficult to tune, Hamiltonian Monte Carlo (HMC) (66) algorithms may be more efficient than adaptative Metropolis-Within-Gibbs algorithms to fit Bayesian PRM models (65).

Other limitations, which are specific to our case study, open new avenues for methodological research in Bayesian PRM models. First, in this paper, we only considered the sum of exposure measurements collected for each covariate, over the entire career of each miner. The Bayesian PRM models could be extended to take into account the temporal dynamics of multiple exposures. Each individual could be assigned to a unique cluster that would depend on his whole trajectory of exposure. Alternatively, the class label of each individual could change over time depending on the temporal dynamics of his exposures. Secondly, this study does not account for the tobacco consumption of miners whereas it is known to be the most important cause of lung cancer. The smoking status is only available for 4.2% of the miners in the post-55 sub-cohort of French uranium miners. This major lack of information makes it very unreliable to adjust for smoking status when estimating the risk of death by lung cancer due to multiple exposures. It makes it also very unreliable to impute about 96% of smoking status given that no potential predictors for smoking status are available in the French cohort of uranium miners. Actually, if tobacco consumption is the main responsible for the excess hazard ratio of death by lung cancer in the French cohort of uranium miners then a higher proportion of smokers should be observed in the clusters with high excess hazard ratio compared to the ones with low excess hazard ratio (and reciprocally). Given the available data, this does not appear to be the case. The ratios between the number of smokers and the number of non-smokers for clusters A, B ,C, D, E, F, G (defined in **Figure 4**) are 12/3, 7/0, 14/5, 17/4, 5/5, 16/8, 34/12, respectively, where clusters F and G have the highest excess hazard ratios of death by lung cancer. The associated proportions of smokers for clusters A, B ,C, D, E, F, G are 0.8, 1.0, 0.74, 0.81, 0.50, 0.67, 0.74, respectively. Of course, these estimated ratios must be interpreted with caution given the limited available data (i.e., 142 miners with smoking status data). Nevertheless, previous analyses on the impact of smoking in occupational cohort studies of uranium miners suggested that smoking was not a source of confounding in these studies (67). This is not surprising since there is actually no strong reason to think that the smoking status is strongly associated with occupational exposure levels. Interestingly, if the proportion of missing smoking status was reasonable (about 30%). The Bayesian PRM models could deal with these missing covariates while accounting for their associated uncertainty to identify exposure profiles. Note that our results should be interpreted with caution given the small number of death by lung cancer in the post-55 French cohort of uranium miners and the lack of data about the tobacco consumption of French uranium miners. As a third limitation of our study, exposure measurement error on radon, γ-rays and uranium dust was not accounted for when identifying the clusters and estimating the associated risks of death by lung cancer. However, complex structures of measurement error were identified in the French cohort of uranium miners (48, 53, 68). It is also well-known that exposure measurement error questions the validity of statistical inference in epidemiological studies (69, 70). When it is not or only poorly accounted for, it may lead to biased risk estimates, a loss in statistical power and a distortion of the exposure-response relationship. Owing to their hierarchical structure, the Bayesian PRM models could be extended to account for exposure measurement error which is, with multicollinearity, one of the most important issues when assessing exposome-health associations (21).

Defining and monitoring the human exposome is a strongly difficult task, given the wide variety of environmental factors, biological endpoints and gene-environment interactions (4, 6, 22). Wild suggested that measuring exposure in any one of the following broad exposure categories—internal (e.g., hormones, microflora), specific external (e.g., toxicants) and general external (e.g., social, psychological)—can reflect certain aspects of the overall exposome (5). Moreover, following Bennett et al. (71), it can be advantageous for the development of statistical methods to narrow the focus of the exposome to a particular class of exposures or/and specific life stages as a way to improve and validate them to apply them later to the broader exposome concepts in a risk assessment or regulatory framework. This was the case in this work that focused on occupational exposure to several types of ionizing radiations of French uranium miners, considering only a small number (i.e., 7) of exposure covariates. This paper shows that the PRM models are promising for exposome research in this context. Interestingly, they could also guide some extensions for higher dimensional data. A great number of covariates including environmental and genetic risk factors could be included in the PRM models in order to study, for instance, gene-environment interactions but the performances of the PRM models should then be assessed in this more challenging context.

DATA AVAILABILITY STATEMENT

The data analyzed in this study is subject to the following licenses/restrictions: data availability is restricted due to subject anonymity. Requests to access these datasets should be directed to klervi.leuraud@irsn.fr.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by Comité Consultatif sur le Traitement de l'Information en matière de Recherche dans le domaine de la Santé (CCTIRS) and Commission Nationale de l'Informatique et des Libertés (CNIL). Written informed consent for participation was not required for this study in accordance with the national legislation and the institutional requirements.

AUTHOR CONTRIBUTIONS

SA and CG contributed the conception and design of the study. MB performed the statistical analysis under the guidance of SA and CG. MB wrote the first draft of the manuscript. MB, OL, SA, and CG contributed to the results evaluation and interpretation.

MB, CG, and SA wrote the sections of the manuscript. All authors contributed to manuscript revision, read, and approved the submitted version.

FUNDING

This work was partially supported by ORANO in the framework of a bilateral agreement between IRSN and ORANO. ORANO had no role in study design, data analysis, or in the interpretation of the results.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpubh.2020.557006/full#supplementary-material

REFERENCES

1. Buck Louis GM, Yeung E, Sundaram R, Laughon SK, Zhang C. The exposome-exciting opportunities for discoveries in reproductive and perinatal epidemiology. *Paediatr Perinat Epidemiol*. (2013) 27:229–36. doi: 10.1111/ppe.12040
2. Rappaport SM, Barupal DK, Wishart D, Vineis P, Scalbert A. The blood exposome and its role in discovering causes of disease. *Environ Health Perspect*. (2014) 122:769–74. doi: 10.1289/ehp.1308015
3. Vrijheid M. The exposome: a new paradigm to study the impact of environment on health. *Thorax*. (2014) 69:876–8. doi: 10.1136/thoraxjnl-2013-204949
4. Wild CP. Complementing the genome with an "exposome": the outstanding challenge of environmental exposure measurement in molecular epidemiology. *Cancer Epidemiol Biomark Prev*. (2005) 14:1847–50. doi: 10.1158/1055-9965.EPI-05-0456
5. Wild CP. The exposome: from concept to utility. *Int J Epidemiol*. (2012) 41:24–32. doi: 10.1093/ije/dyr236
6. Rappaport SM, Smith MT. Environment and disease risks. *Science*. (2010) 330:460–1. doi: 10.1126/science.1192603
7. Eiband JD, Elias EG, Suter CM, Gray WC, Didolkar MS. Prognostic factors in squamous cell carcinoma of the larynx. *Am J Surg*. (1989) 158:314–7. doi: 10.1016/0002-9610(89)90123-2
8. Krejs GJ. Gastric cancer: epidemiology and risk factors. *Digest Dis*. (2010) 28:600–3. doi: 10.1159/000320277
9. Lin RS, Kessler II. A multifactorial model for pancreatic cancer in man: epidemiologic evidence. *JAMA*. (1981) 245:147–52. doi: 10.1001/jama.245.2.147
10. Zabaleta J. Multifactorial etiology of gastric cancer. *Methods Mol Biol*. (2012) 863:411–35. doi: 10.1007/978-1-61779-612-8_26
11. Steliga MA, Dresler CM. Epidemiology of lung cancer: smoking, secondhand smoke, and genetics. *Surg Oncol Clin*. (2011) 20:605–18. doi: 10.1016/j.soc.2011.07.003
12. Li N, Chen G, Liu F, Mao S, Liu Y, Mao Z, et al. Associations between long-term exposure to air pollution and blood pressure and effect modifications by behavioral factors. *Environ Res*. (2020) 182:109. doi: 10.1016/j.envres.2019.109109
13. Dominici F, Peng RD, Barr CD, Bell ML. Protecting human health from air pollution: shifting from a single-pollutant to a multi-pollutant approach. *Epidemiology*. (2010) 21:187–94. doi: 10.1097/EDE.0b013e3181cc86e8
14. Greenland S, Pearl J, Robins JM. Causal diagrams for epidemiologic research. *Epidemiology*. (1999) 10:37–48. doi: 10.1097/00001648-199901000-00008
15. Bell ML, Kim JY, Dominici F. Potential confounding of particulate matter on the short-term association between ozone and mortality in multisite time-series studies. *Environ Health Perspect*. (2006) 115:1591–5. doi: 10.1289/ehp.10108
16. Klebe S, Leigh J, Henderson D, Nurminen M. Asbestos, smoking and lung cancer: an update. *Int J Environ Res Public Health*. (2019) 17:258. doi: 10.3390/ijerph17010258
17. Leuraud K, Schnelzer M, Tomasek L, Hunter N, Timarche M, Grosche B, et al. Radon, smoking and lung cancer risk: results of a joint analysis of three European case-control studies among uranium miners. *Radiat Res*. (2011) 176:375–87. doi: 10.1667/RR2377.1
18. Lin H, Guo Y, Ruan Z, Yang Y, Chen Y, Zheng Y, et al. Ambient PM2.5 and O3 and their combined effects on prevalence of presbyopia among the elderly: a cross-sectional study in six low- and middle-income countries. *Sci Total Environ*. (2019) 655:168–73. doi: 10.1016/j.scitotenv.2018.11.239
19. United Nations Scientific Committee on the Effects of Atomic Radiation. *Sources, Effects and Risks of Ionizing Radiation. UNSCEAR 2017 Report with Scientific Annexes A and B*. New York, NY: United Nations (2017).
20. National Research Council N. *Heath Risks From Exposure to Low Levels of Ionizing Radiation. BEIR VII Phase 2*. Washington, DC: The National Academies Press (2006).
21. Billionnet C, Sherrill D, Annesi-Maesano I, GERIE S. Estimating the health effects of exposure to multi-pollutant mixture. *Ann Epidemiol*. (2012) 22:126–41. doi: 10.1016/j.annepidem.2011.11.004
22. Slama R, Vrijheid M. Some challenges of studies aiming to relate the exposome to human health. *Occup Environ Med*. (2015) 72:383–4. doi: 10.1136/oemed-2014-102546
23. Farrar DE, Glauber RR. Multicollinearity in regression analysis: the problem revisited. *Rev Econ Stat*. (1967) 49:92–107. doi: 10.2307/1937887
24. Mela CF, Kopalle PK. The impact of collinearity on analysis: the asymmetric effect of negative and positive correlations. *Appl Econ*. (2002) 34:667–77. doi: 10.1080/00036840110058482
25. Tu YK GM Clerehugh V. Collinearity in linear regression is a serious problem in oral health research. *Eur J Oral Sci*. (2004) 112:389–97. doi: 10.1111/j.1600-0722.2004.00160.x
26. Vatcheva KP, Lee, McCormick JB, Rahbar MH. Multicollinearity in regression analyses conducted in epidemiologic studies. *Epidemiology*. (2016) 6:227. doi: 10.4172/2161-1165.1000227
27. Patel J C J Bhattacharya, Butte AJ. An Environment-Wide Association Study (EWAS) on type 2 diabetes mellitus. *PLoS ONE*. (2010) 5:e10746. doi: 10.1371/journal.pone.0010746
28. Rage E, Caër-Lorho S, Drubay D, Ancelet S, Laroche P, Laurier D. Mortality analyses in the updated French cohort of uranium miners (1946–2007). *Int Archiv Occupat Environ Health*. (2015) 88:717–30. doi: 10.1007/s00420-014-0998-6
29. Lenters V, Portengen L, Rignell-Hydbom A, Jönsson BA, Lindh CH, Piersma AH, et al. Prenatal phthalate, perfluoroalkyl acid, and organochlorine exposures and term birth weight in three birth cohorts: multi-pollutant models based on elastic net regression. *Environ Health Perspect*. (2016) 124:365–72. doi: 10.1289/ehp.1408933
30. Bottolo L, Richardson S. Evolutionary stochastic search for Bayesian model exploration. *Bayesian Anal*. (2010) 5:583–618. doi: 10.1214/10-BA523
31. Massy WF. Principal components regression in exploratory statistical research. *J Am Stat Assoc*. (1965) 60:234–56. doi: 10.1080/01621459.1965.10480787
32. Jain P, Vineis P, Liquet B, Vlaanderen J, Bodinier B, Van Veldhoven K, et al. A multivariate approach to investigate the combined biological effects of multiple exposures. *J Epidemiol Community Health*. (2018) 72:564–71. doi: 10.1136/jech-2017-210061
33. Wold S, Ruhe A, Wold H, Dunn W III. The collinearity problem in linear regression. The partial least squares (PLS) approach to generalized inverses. *SIAM J Sci Stat Comput*. (1984) 5:735–43. doi: 10.1137/0905052
34. Marshall R. The use of classification and regression trees in clinical epidemiology. *J Clin Epidemiol*. (2001) 54:603–9. doi: 10.1016/S0895-4356(00)00344-9

35. Forgy E. Cluster analysis of multivariate data: efficiency vs interpretability of classifications. *Biometrics*. (1965) 21:768–9.
36. Patterson BH, Dayton CM, Graubard BI. Latent class analysis of complex sample survey data: application to dietary data. *J Am Stat Assoc*. (2002) 97:721–8. doi: 10.1198/016214502388618465
37. Molitor J, Papathomas M, Jerrett M, Richardson S. Bayesian profile regression with an application to the National Survey of Children's Health. *Biostatistics*. (2010) 11:484–98. doi: 10.1093/biostatistics/kxq013
38. Agier L, Portengen L, Chadeau-Hyam M, Basagana X, Giorgis-Allemand L, Siroux V, et al. A systematic comparison of linear regression-based statistical methods to assess exposome-health associations. *Environ Health Perspect*. (2016) 124:1848–56. doi: 10.1289/EHP172
39. Wold S, Esbensen K, Geladi P. Principal component analysis. *Chemometr Intell Lab Syst*. (1987) 2:37–52. doi: 10.1016/0169-7439(87)80084-9
40. Papathomas M, Molitor J, Hoggart D, Hastie D, Richardson S. Exploring data from genetic association studies using Bayesian variable selection and the Dirichlet process: application to searching for gene × gene patterns. *Genet Epidemiol*. (2012) 36:663–74. doi: 10.1002/gepi.21661
41. Papathomas M, Molitor J, Richardson S, Riboli E, Vineis P. Examining the joint effect of multiple risk factors using exposure risk profiles: lung cancer in nonsmokers. *Environ Health Perspect*. (2011) 119:84–91. doi: 10.1289/ehp.1002118
42. Coker E, Liverani S, Su J, Molitor J. Multi-pollutant modeling through examination of susceptible subpopulations using profile regression. *Curr Environ Health Rep*. (2018) 5:59–69. doi: 10.1007/s40572-018-0177-0
43. Pirani M, Best N, Blangiardo M, Liverani S, Atkinson R, Fuller G. Analysing the health effects of simultaneous exposure to physical and chemical properties of airborne particles. *Environ Int*. (2015) 79:56–64. doi: 10.1016/j.envint.2015.02.010
44. Liverani S, Lavigne A, Blangiardo M. Modelling collinear and spatially correlated data. *Spatial Spatiotemp Epidemiol*. (2016) 18:63–73. doi: 10.1016/j.sste.2016.04.003
45. Hastie D, Liverani S, Azizi L, Richardson S, Stucker I. A semi-parametric approach to estimate risk functions associated with multi-dimensional exposure profiles: application to smoking and lung cancer. *BMC Med Res Methodol*. (2013) 13:129. doi: 10.1186/1471-2288-13-129
46. Mattei F, Liverani S, Guida F, Matrat M, Cenee S, Azizi L, et al. Multidimensional analysis of the effect of occupational exposure to organic solvents on lung cancer risk: the ICARE study. *Occupat Environ Med*. (2016) 73:368–77. doi: 10.1136/oemed-2015-103177
47. Liverani S, Hastie DI, Azizi L, Papathomas M, Richardson S. PReMiuM: an R package for profile regression mixture models using Dirichlet processes. *J Stat Softw*. (2015) 64:1. doi: 10.18637/jss.v064.i07
48. Hoffmann S, Rage E, Laurier D, Laroche P, Guihenneuc C, Ancelet S. Accounting for Berkson and classical measurement error in radon exposure using a Bayesian structural approach in the analysis of lung cancer mortality in the French cohort of uranium miners. *Radiat Res*. (2017) 187:196–209. doi: 10.1667/RR14467.1
49. Vacquier B, Rage E, Leuraud K, Caër-Lorho S, Houot J, Acker A, et al. The influence of multiple types of occupational exposure to radon, gamma rays and long-lived radionuclides on mortality risk in the French "post-55" sub-cohort of uranium miners: 1956–1999. *Radiat Res*. (2011) 176:796–806. doi: 10.1667/RR2558.1
50. IARC. *IARC Working Group on the Evaluation of Carcinogenic Risks to Humans Which Met in Lyon. Man-Made Mineral Fibres and Radon*. Lyon: IARC (1988).
51. Rage E, Vacquier B, Blanchardon E, Allodji RS, Marsh JW, Caer-Lorho S, et al. Risk of lung cancer mortality in relation to lung doses among French uranium miners: follow-up 1956-1999. *Radiat Res*. (2012) 177:288–97. doi: 10.1667/RR2689.1
52. Bhatia R, Lopipero P, Smith AH. Diesel exhaust exposure and lung cancer. *Epidemiology*. (1998) 9:84–91. doi: 10.1097/00001648-199801000-00017
53. Allodji SR. *Prise en compte des erreurs de mesure dans l'analyse du risque associe a l'exposition aux rayonnements ionisants dans une cohorte professionnelle: application á la cohorte française des mineurs d'uranium*. Paris: Université Paris-Sud (2011).
54. Kleinbaum DG, Klein M. *Survival Analysis*. New York, NY: Springer (2010).
55. Crosfill M, Lindop PJ, Rotblat J. Variation of sensitivity to ionizing radiation with age. *Nature*. (1959) 183:1729. doi: 10.1038/1831729a0
56. Ohlssen DI, Sharples LD, Spiegelhalter DJ. Flexible random-effects models using Bayesian semi-parametric models: applications to institutional comparisons. *Stat Med*. (2007) 26:2088–112. doi: 10.1002/sim.2666
57. Ishwaran H, Zarepour M. Markov chain Monte Carlo in approximate Dirichlet and beta two-parameter process hierarchical models. *Biometrika*. (2000) 87:371–90. doi: 10.1093/biomet/87.2.371
58. Kreuzer M, Sobotzki C, Fenske N, Marsh JW, Schnelzer M. Leukaemia mortality and low-dose ionising radiation in the WISMUT uranium miner cohort (1946–2013). *Occupat Environ Med*. (2017) 74:252–8. doi: 10.1136/oemed-2016-103795
59. Roberts GO, Rosenthal JS. Examples of adaptive MCMC. *J Comput Graph Stat*. (2009) 18:349–67. doi: 10.1198/jcgs.2009.06134
60. Papaspiliopoulos O, Roberts GO. Retrospective Markov chain Monte Carlo methods for Dirichlet process hierarchical models. *Biometrika*. (2008) 95:169–86. doi: 10.1093/biomet/asm086
61. Hastie DI, Liverani S, Richardson S. Sampling from Dirichlet process mixture models with unknown concentration parameter: mixing issues in large data implementations. *Stat Comput*. (2015) 25:1023–37. doi: 10.1007/s11222-014-9471-3
62. Spiegelhalter DJ, Best NG, Carlin BP, Van Der Linde A. Bayesian measures of model complexity and fit. *J R Stat Soc B Stat Methodol*. (2002) 64:583–639. doi: 10.1111/1467-9868.00353
63. Watanabe S. Asymptotic equivalence of Bayes cross validation and widely applicable information criterion in singular learning theory. *J Mach Learn Res*. (2010) 11:3571–94. Available online at: https://jmlr.org/papers/v11/watanabe10a.html
64. Gelman A, Carlin JB, Stern HS, Dunson DB, Vehtari A, Rubin DB. *Bayesian Data Analysis*. New York, NY: Chapman and Hall; CRC (2013).
65. Betancourt M, Girolami M. Hamiltonian Monte Carlo for hierarchical models. *Curr Trends Bayes Methodol Appl*. (2015) 79:30. doi: 10.1201/b18502-5
66. Neal RM. An improved acceptance procedure for the hybrid Monte Carlo algorithm. *J Comput Phys*. (1994) 111:194–203. doi: 10.1006/jcph.1994.1054
67. Richardson DB, Laurier D, Schubauer-Berigan MK, Tchetgen ET, Cole SR. Assessment and indirect adjustment for confounding by smoking in cohort studies using relative hazards models. *Am J Epidemiol*. (2014) 180:933–40. doi: 10.1093/aje/kwu211
68. Allodji RS, Leuraud K, Bernhard S, Henry S, Bénichou J, Laurier D. Assessment of uncertainty associated with measuring exposure to radon and decay products in the French uranium miners cohort. *J Radiol Protec*. (2012) 32:85. doi: 10.1088/0952-4746/32/1/85
69. Gustafson P. *Measurement Error and Misclassification in Statistics and Epidemiology: Impacts and Bayesian Adjustments*. New York, NY: Chapman and Hall; CRC (2003).
70. Armstrong BG. Effect of measurement error on epidemiological studies of environmental and occupational exposures. *Occupat Environ Med*. (1998) 55:651–6. doi: 10.1136/oem.55.10.651
71. Bennett B, Workman T, Smith MN, Griffith WC, Thompson B, Faustman EM. Characterizing the neurodevelopmental pesticide exposome in a children's agricultural cohort. *Int J Environ Res Public Health*. (2020) 17:1479. doi: 10.3390/ijerph17051479

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

université PARIS-SACLAY | ÉCOLE DOCTORALE Santé Publique (EDSP)

Titre : Approche hiérarchique bayésienne pour l'estimation du risque de cancers radio-induits en situation d'expositions professionnelles multiples et incertaines : Application aux travailleurs du cycle du combustible nucléaire

Mots clés : Co-expositions, erreurs de mesure, modélisation hiérarchique, statistique bayésienne, cancer, épidémiologie des rayonnements ionisants.

Résumé : Les travailleurs du cycle du combustible nucléaire sont chroniquement exposés à de multiples sources radiologiques. À ce jour, les risques de cancers associés à une prise en compte simultanée de ces expositions, souvent corrélées, sont peu étudiés et les incertitudes de mesure sur ces expositions peu considérées. L'objectif de ce travail est de promouvoir l'utilisation de l'approche hiérarchique bayésienne pour dépasser ces limites actuelles sur l'estimation des risques. Des modèles hiérarchiques bayésiens ont été proposés pour tenir compte de l'impact potentiel d'expositions sujettes à erreurs de mesure et censurées à gauche (conséquence des limites de détection des dosimètres). Une estimation corrigée du risque de décès par cancer du poumon associé à une exposition chronique aux rayonnements γ a été obtenue à partir de la cohorte française des mineurs d'uranium. Afin d'apporter une réponse souple et élégante au problème des co-expositions en épidémiologie des rayonnements ionisants, les modèles de mélange par régression bayésienne sur profils d'exposition ont été étendus aux modèles de survie en excès de risque instantané. Une étude par simulations a été réalisée et des groupes de mineurs aux profils d'expositions radiologiques à haut risque de décès par cancer du poumon identifiés.

Title : Bayesian hierarchical approach to estimate the risk of radiation-induced cancers in the situation of multiple and uncertain occupational exposures: Application to workers in the nuclear fuel cycle

Keywords : Coexposure, measurement error, hierarchical modelling, Bayesian statistics, cancer, radiation epidemiology.

Abstract : Nuclear fuel cycle workers are chronically exposed to multiple radiological sources. To date, the cancer risks associated with simultaneous and often correlated exposures have been rarely estimated and the measurement uncertainties on these exposures rarely considered. The aim of this work is to promote the Bayesian hierarchical approach to overcome these current limitations on risk estimates. Bayesian hierarchical models were proposed to account for the potential impact of exposures subject to measurement errors and left censored (due to dosimeter detection limits). A corrected estimation of the risk of death by lung cancer due to chronic exposure to γ-rays was obtained from the French cohort of uranium miners. To provide a flexible and elegant answer to the challenge of coexposure in radiation epidemiology, the Bayesian profile regression mixture models were extended to the excess hazard ratio survival models. A simulation study was performed and some groups of uranium miners with radiological exposure profiles at high risk of death by lung cancer were identified.

Maison du doctorat de l'Université Paris-Saclay
2ème étage aile ouest, Ecole normale supérieure Paris-Saclay
4 avenue des Sciences,
91190 Gif sur Yvette, France