Analyse des données massives de source assurantielle de la Mutualité Sociale Agricole, pour la surveillance en santé au travail des travailleurs agricoles en France

Charlotte Maugard

Communauté
UNIVERSITÉ Grenoble Alpes

THÈSE

Pour obtenir le grade de

DOCTEUR DE LA

COMMUNAUTE UNIVERSITE GRENOBLE ALPES

Spécialité : **MBS – Modèles, Méthodes et Algorithmes en Biologie, Santé et Environnement**

Arrêté ministériel : 25 mai 2016

Présentée par

Charlotte MAUGARD

Thèse dirigée par **Pr Vincent BONNETERRE**, Professeur des universités – Praticien hospitalier, CHU Grenoble Alpes, Laboratoire « Techniques de L'Ingénierie Médicale et de la Complexité - Informatique, Mathématiques et Applications » (TIMC-IMAG), Equipe « Environnement et Prédiction de la Santé des Populations » (EPSP)
et codirigée par **Pr Olivier FRANCOIS**, Professeur des universités, Grenoble INP, Laboratoire TIMC-IMAG, Equipe « Biologie Computationnelle et Mathématique » (BCM)

Préparée au sein du **Laboratoire TIMC-IMAG (Université Grenoble Alpes, Institut Polytechnique de Grenoble, CNRS - UMR 5525) – Equipes EPSP & BCM**
dans **l'École Doctorale Ingénierie pour la Santé, la Cognition et l'Environnement (EDISCE)**

Analyse des données massives de source assurantielle de la Mutualité Sociale Agricole, pour la surveillance en santé au travail des travailleurs agricoles en France

Thèse soutenue publiquement le **25 novembre 2019**,
devant le jury composé de :

Dr Pierre LEBAILLY
Maître de conférences, Centre régional de lutte contre le cancer François Baclesse, INSERM - UMR 1086 « ANTICIPE », Rapporteur
Pr Marie ZINS
Professeur des universités - praticien hospitalier, Hôpital Paul Brousse, INSERM - UMS 011 « Cohortes en population », Rapporteur
Dr Florence FORBES
Directrice de recherche, INRIA Grenoble Rhône Alpes, Examinatrice
Dr Rémy SLAMA
Directeur de recherche, Inserm U1209 / CNRS UMR 5309, IAB, UGA, Président du jury

Remerciements

En premier lieu, je souhaite remercier les membres du jury, **Pr Marie Zins** et **Dr Pierre Lebailly** en qualité de rapporteurs, pour la relecture et les corrections apportées au manuscrit ; **Dr Remy Slama** pour ses conseils et son accord pour présider ce jury et enfin, **Dr Florence Forbes**, en qualité d'examinatrice, pour avoir également accepté d'évaluer ce travail.

Puis, je tiens évidemment à remercier mes deux directeurs de thèse qui m'ont donné l'opportunité d'effectuer cette thèse :

- **Pr Vincent Bonneterre** : Merci beaucoup de m'avoir encadrée tout au long de ce travail. Merci pour tous nos échanges et discussions qui m'ont permis d'apprendre beaucoup dans le domaine de la santé au travail. Merci de m'avoir fait confiance, notamment suite au départ de Delphine, pour la reprise du projet. Merci également de m'avoir autant poussée à améliorer mon aisance à l'oral à travers de nombreuses communications car je pense que cela me sera très utile par la suite. Enfin, merci de m'avoir initiée au travail dans le domaine de la recherche.

- **Pr Olivier François** : En tout premier lieu, merci de m'avoir choisie avec Delphine pour faire cette thèse. Merci pour ton encadrement, ta disponibilité, ta réactivité lorsque j'étais en difficulté et tous nos échanges forts instructifs en statistiques et modélisation qui m'ont guidée tout au long de ce travail de thèse et m'ont permis d'acquérir les compétences que je possède aujourd'hui.

De plus, je souhaiterais adresser un remerciement particulier au **Dr Delphine Bosson-Rieutort**, la post-doctorante ayant lancé le projet MSA (Projet de fouille de données de la Mutualité Sociale Agricole) avec Vincent. Mon travail de thèse n'aurait pas été possible sans ta contribution pour mettre le projet sur les rails. Par ailleurs, ton aide et tes précieux conseils lors de ma première année de thèse m'ont permis de bien prendre en main le projet et de le mener jusqu'à aujourd'hui. Un grand merci encore !

Ensuite, je tiens à adresser des remerciements aux différents organismes qui ont permis la mise en place du projet MSA :

- à l'**Agence nationale de sécurité sanitaire de l'alimentation, de l'environnement et du travail (Anses)** qui a financé jusqu'à présent deux conventions de recherche et de développement avec notre équipe de recherche, indispensables au lancement et au développement du projet (financement post-doc, stages) et aussi, pour leur contribution scientifique ;

- à la **MSA** pour nous avoir fourni les données nécessaires dans le cadre du projet et pour leur patience et leur disponibilité vis-à-vis de nos questions sur leurs données ;

- à **Santé Publique France** pour avoir mis à notre disposition leurs matrices culture-expositions (MATPHYTO) mais également pour leur disponibilité et leurs conseils ;

- et enfin, je tiens à remercier l'**Université Grenoble Alpes (UGA)** qui, grâce au programme AGIR-POLE, a financé ma thèse.

Je tiens également à remercier le **laboratoire de recherche TIMC-IMAG (CNRS, UMR 5525)** pour m'avoir accueillie dans ses locaux ainsi que toute l'**équipe EPSP** et spécifiquement :

- Merci au **Pr Anne Maître**, pour m'avoir accueillie au sein de l'équipe et pour tous tes précieux conseils qu'ils soient liés à mon travail de thèse ou non ;

- Merci à **Sylvette Liaudy** pour tous nos échanges très instructifs et tes précieux conseils ;

- Merci à **Franck Balducci** pour être venu à mon secours quand j'avais besoin d'aide en statistiques, mais aussi pour ta bonne humeur et tes blagues aux pauses café ;

- Merci au **Dr Christine Demeilliers**, pour nos nombreux échanges et tes conseils ;

- Enfin, un grand merci au **Dr Pascal Petit**, qui a rejoint le projet MSA lors de ma dernière année de thèse, pour ton soutien, ton aide et tes précieux conseils tout au long de cette dernière année !

Je tiens également à exprimer ma gratitude envers l'**équipe BCM** pour m'avoir laissée participer à votre séminaire du vendredi et notamment envers certains membres, qui m'ont aidée avec R et m'ont donné de précieux conseils pour mon avenir professionnel.

Je souhaiterais aussi remercier l'**équipe de santé au travail du CHU** pour son accueil à deux reprises dans ses locaux et pour toutes nos discussions lors des pauses.

Je voudrais aussi remercier l'ensemble des stagiaires et doctorants qui ont été de passage au cours de ma thèse, et en particulier : **Pauline Achard** et **Camille Eldin** pour leur contribution au projet MSA, leur bonne humeur et nos échanges, qu'ils soient scientifiques ou non ; **Etienne Bourgart**, **Garance Terpant** et **Maguy Basbous** pour tous nos moments de convivialité aux pauses.

J’aimerais adresser un remerciement tout particulier à **Ludivine Taieb** que j’ai eu la chance de côtoyer dans l’équipe pendant quelques mois (période bien trop courte !) lors de sa thèse entre la France et le Canada. Merci pour ton écoute, pour tous nos échanges, notamment sur les galères de la thèse, mais surtout pour tous ces bons souvenirs passés ensemble : nos pauses déjeuner « jeux de société » ou encore nos sorties cinéma… qui me manquent beaucoup. J’espère que nous resterons amies encore longtemps et que tes amis Elliott et Fleur t’aideront à ne jamais m’oublier.

Enfin, j’adresse un énorme remerciement **à mes parents et à ma sœur Pauline**, pour m’avoir soutenue et conseillée tout au long de ces trois années. Je suis fière d’en être arrivée jusqu’ici et c’est grâce à vous.

Sans oublier, un immense remerciement que je souhaite adresser à mon conjoint **Quentin**, qui m’a soutenue, épaulée et remonté le moral tellement de fois pendant cette thèse que je ne les compte plus. Sans ton indéfectible soutien, je ne serais jamais venue au bout de ce travail de thèse. Alors, encore merci de m’avoir poussée et suivie pour que j’en arrive là !

« Il ne faut avoir aucun regret pour le passé, aucun remords pour le présent,
et une confiance inébranlable pour l’avenir. »

Jean Jaurès

Sommaire

Tableaux

Partie 1

Partie 2

Partie 3

Partie 4

Partie 5

Partie 6

Figures

Partie 1

Partie 2

Partie 3

Partie 4

Partie 5

Partie 6

Abréviations

(A)CP	(Analyse en) Composantes Principales
AGRICAN	Cohorte épidémiologique française « AGRIculture et CANcers »
AHS	Cohorte épidémiologique américaine « Agricultural Health Study »
ALD	Affection de Longue Durée
ANSES	Agence Nationale de Sécurité Sanitaire de l'Alimentation, de l'Environnement et du Travail
AT/MP	Accidents du Travail et Maladies Professionnelles
AUC	Aire sous la courbe ROC
AVC	Accident Vasculaire Cérébral
BIC	Bayesian Information Criterion
C2RMP	Comité Régional de Reconnaissance des Maladies Professionnelles
(CC)MSA	(Caisse Centrale de la) Mutualité Sociale Agricole
CCPP	Centre de Consultation de Pathologies Professionnelles
CIM-10	Classification Internationale des Maladies, 10ème révision
CNAP	Cohorte épidémiologique norvégienne « Cancer in the Norwegian Agricultural Population »
CNIL	Commission Nationale de l'Informatique et des Libertés
FAO	Food and Agriculture Organization of the United Nations (Organisation des Nations unies pour l'alimentation et l'agriculture)
HTA	Hypertension artérielle
IC 95%	Intervalle de Confiance à 95%
INSEE	Institut National de la Statistique et des Etudes Economiques
IPP	Incapacité Permanente Partielle
LFMM	Modèles mixtes à facteurs latents (Latent Factor Mixed Models)
NAF	Nomenclature des Activités Française
OR	Odd Ratio
PAN	Périarthrite noueuse
PPV	Phytopharmacovigilance
RL	Régression logistique
RNV3P	Réseau National de Vigilance et de Prévention des Pathologies Professionnelles
ROC	Receiver Operating Characteristic
RSA	Revenu de solidarité active (ex RMI, Revenu Minimum d'Insertion)
ScS	Sclérodermie Systémique
VIH	Virus de l'Immunodéficience Humaine

PARTIE 1

Introduction générale

I. Le secteur agricole, un secteur exigeant, présentant des risques professionnels spécifiques

En 2016, l'agriculture regroupait près d'un tiers de la population active mondiale, soit 28.9% avec notamment 1.5 milliard d'hectares de terres cultivées, soit 11% de la surface des terres émergées de la planète dédiées à la production végétale (1,2). L'agriculture fait partie des secteurs professionnels primordiaux car elle répond aux besoins de base de l'humanité en nourriture, alimentation animale, fibres et combustible. En effet, à titre d'exemple, l'agriculture permet chaque jour la production moyenne de 23.7 millions de tonnes de nourriture (céréales, fruits, légumes, viande, lait, …) tandis que les forêts apportent 9.5 millions de mètres cubes de bois d'œuvre et de combustible (3,4). Cependant, la population mondiale étant en perpétuelle croissance, estimée à 8.3 milliards d'individus en 2030, le secteur agricole se doit d'évoluer afin de relever le défi constant de subvenir aux besoins mondiaux, tout en préservant les ressources naturelles de notre écosystème. D'ailleurs, le Programme de développement durable à l'horizon 2030 de l'Organisation des Nations Unies pour l'alimentation et l'agriculture (FAO) prévoit l'objectif suivant : « Éliminer la faim, assurer la sécurité alimentaire, améliorer la nutrition et promouvoir l'agriculture durable » (4).

De la même façon que la FAO, la France a souhaité développer en 2015 le plan « Agriculture – Innovation 2025 » afin de remplir ces mêmes objectifs au niveau national (5). D'ailleurs, en 2017, la production agricole française était la plus élevée au sein de l'Union Européenne, représentant 16.7% de la production européenne avec une valeur estimée à près de 65 milliards d'euros (2,6). La surface agricole utilisée (SAU) était d'environ 28 millions d'hectares dont 13.1 millions d'hectares dédiés aux grandes cultures (90% de céréales, d'oléagineux et de protéagineux), soit environ 52% de la surface de la France métropolitaine utilisée par le secteur agricole (7,8). Enfin, l'agriculture concernait plus de 1.2 millions d'emplois au régime agricole, regroupant environ 3% de la population active en France. Ce secteur professionnel est d'ailleurs caractérisé par de nombreuses activités, notamment de cultures (céréales, fruits, coquillages, …) ou d'élevages (bovins, caprins, poissons, …) mais il renvoie également, à titre d'exemple, à des entreprises de jardins ou des exploitations forestières. Le monde agricole correspond donc à une grande diversité d'activités, pouvant être réalisées dans des contextes géographiques différents, et qui influencent elles-mêmes les pratiques agricoles (mécanisation, pression de ravageurs, …) (Figure 1). D'ailleurs, ces pratiques évoluent dans le temps. En effet, contrairement à d'autres secteurs professionnels, l'agriculture bénéficie des innovations développées au fil des révolutions technologiques (mécanisation par le passé, numérisation et nano-pesticides aujourd'hui), ce qui peut d'ailleurs engendrer des modifications sociologiques et économiques importantes (9).

Ainsi, selon les activités pratiquées, l'usage de produits phytopharmaceutiques, communément appelés « pesticides », peut varier, ces produits étant destinés à lutter contre les parasites animaux et végétaux nuisibles aux cultures et aux produits récoltés (*Larousse*). D'ailleurs, il est important de noter qu'en 2017, environ 70 tonnes de produits phytopharmaceutiques ont été vendues en France, qu'il s'agisse d'insecticides, d'herbicides ou encore de fongicides, les principales familles chimiques de ces produits (10). La diversité des activités agricoles renvoie alors à autant d'expositions différentes et potentiellement dangereuses qui sont à même d'influer sur l'état de santé des travailleurs de ce secteur. D'ailleurs, que ce soit en France ou dans le monde, le secteur agricole est l'un des secteurs professionnels où nombreux sont les accidents et problèmes de santé liés au travail d'après l'Organisation Internationale du Travail (ILO) (11).

En effet, les risques professionnels auxquels peuvent être exposés les travailleurs agricoles peuvent être d'origine **physique** (contraintes posturales ou de manutention, exposition prolongée aux ultraviolets ou à des vibrations, empoussièrement, risques mécaniques, …), **biologique** (contact avec des animaux, microorganismes, endotoxines, allergènes de source végétale ou animale, …), ou encore **chimique** (produits phytopharmaceutiques, engrais, biocides, …). Cependant, les travailleurs agricoles affichent de faibles taux de mortalité pour la plupart des causes courantes de décès et sont moins enclins à développer des cancers, à la fois dans l'ensemble et certains types de cancers en particulier. Malgré cela, dans la littérature scientifique, l'usage de produits phytopharmaceutiques dans ce secteur a été de nombreuses fois relié au développement de maladies chroniques spécifiques tels que des cancers ou des maladies neurodégénératives. De plus, la situation socio-économique parfois précaire de certains travailleurs agricoles les expose particulièrement à des risques psycho-sociaux, facteurs de risques d'altération de la santé mentale et parfois, de suicides (12–17).

Le secteur agricole est essentiel et doit relever des défis importants afin de subvenir aux besoins d'une population mondiale grandissante, tout en tenant compte des enjeux du développement durable. Ceci se structure au travers de plans tant au niveau mondial (Programme de développement durable à l'horizon 2030 de la FAO) qu'aux échelles nationales (en France, le plan « Agriculture – Innovation 2025 »). La France est particulièrement concernée par ces défis puisqu'elle possède la première production agricole à l'échelle de l'Union Européenne.

Cette production agricole est le fruit du travail d'exploitants et de salariés agricoles, dont l'activité professionnelle est intimement liée à leur santé globale (« bien-être physique, psychique et social » au sens de l'Organisation Mondiale de la Santé), du fait de risques propres voire par le fait que certaines activités conditionnent pour partie leur mode de vie (exemple : ruralité pouvant être liée à l'isolement). Ces activités professionnelles, selon les expositions et les contextes associés, peuvent donc influer positivement ou négativement certains aspects de la santé de ces travailleurs. C'est pourquoi il est important de déchiffrer ces relations travail-santé, afin de mieux les comprendre, pour prévenir autant que possible les risques d'altération de la santé et faire en sorte que ces activités soient avant tout contributrices de bonne santé. En particulier, si certains risques sont identifiés de longue date, il importe de maintenir une vigilance pour rechercher des risques existants qui n'auraient pas été préalablement mis en évidence, et tenter d'identifier précocement des risques émergents pour la santé, dans la mesure où les activités agricoles évoluent dans le temps.

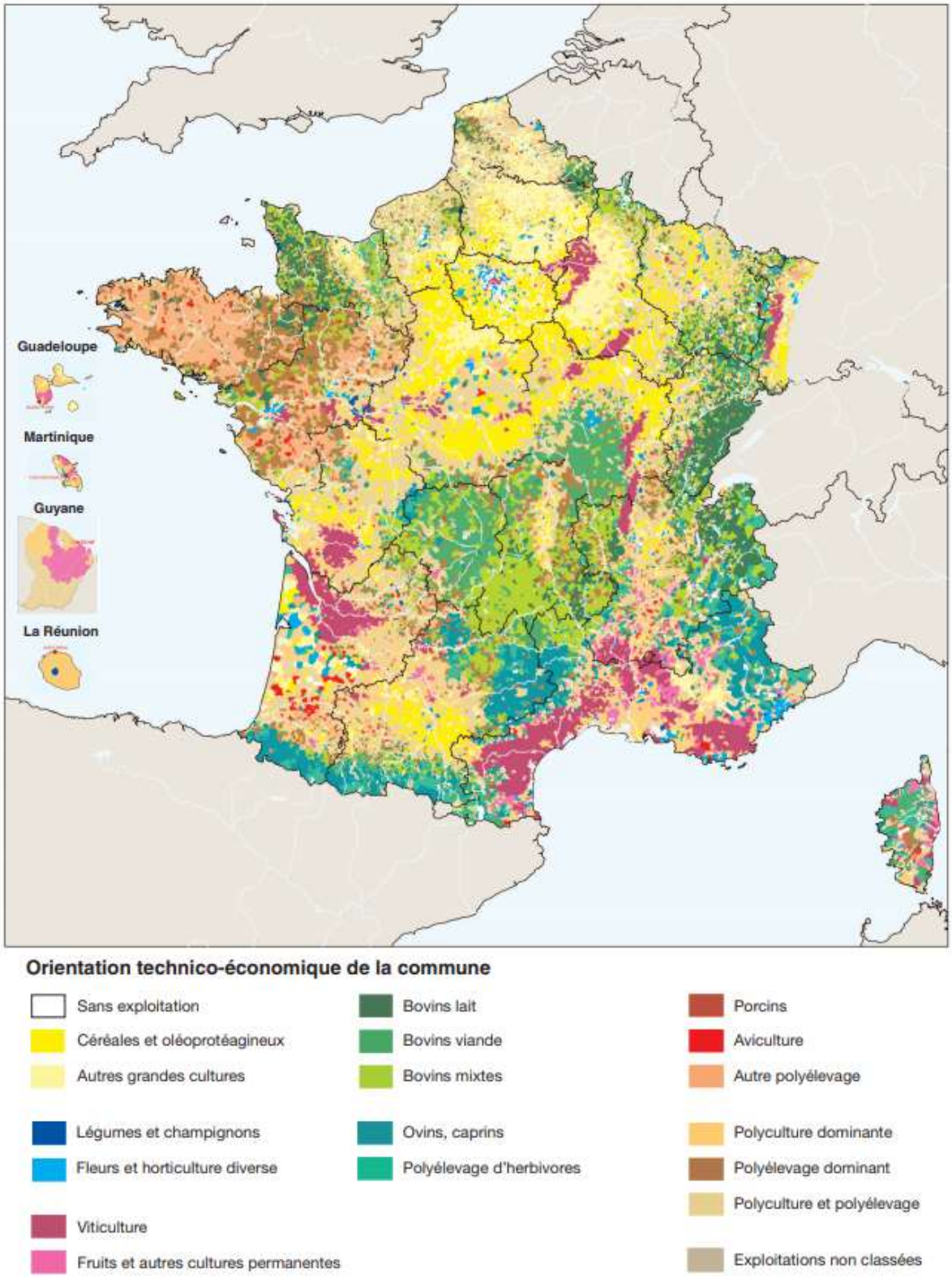


Figure 1 : Carte de répartition des orientations technico-économiques des communes en France au cours du recensement agricole réalisé en 2010 en France (Agreste, Ministère de l'Agriculture et de l'Alimentation)

II. Etude de la santé des travailleurs agricoles

Pour les raisons citées ci-dessus, la population de travailleurs agricoles mérite donc un suivi épidémiologique spécifique. A ce titre, depuis de nombreuses années, plusieurs cohortes ont initié des études sur les effets néfastes des activités agricoles sur la santé des travailleurs de ce secteur. **D'ailleurs, en octobre 2010, la plupart des cohortes agricoles se sont regroupées en un consortium nommé AGRICOH**, initié par le National Cancer Institute (NCI) des Etats-Unis et le Centre International de Recherche sur le Cancer (CIRC[1]). Ce consortium s'est donné pour objectif de mettre en place une collaboration entre différentes cohortes agricoles afin de promouvoir le partage des données pour mieux évaluer les associations entre les expositions agricoles, notamment celles qui peuvent être rares, et leurs effets néfastes sur la santé (18). En Janvier 2016, trente cohortes avaient rejoint le consortium, permettant ainsi de couvrir 5 continents et 12 pays (Figure 2) (19).

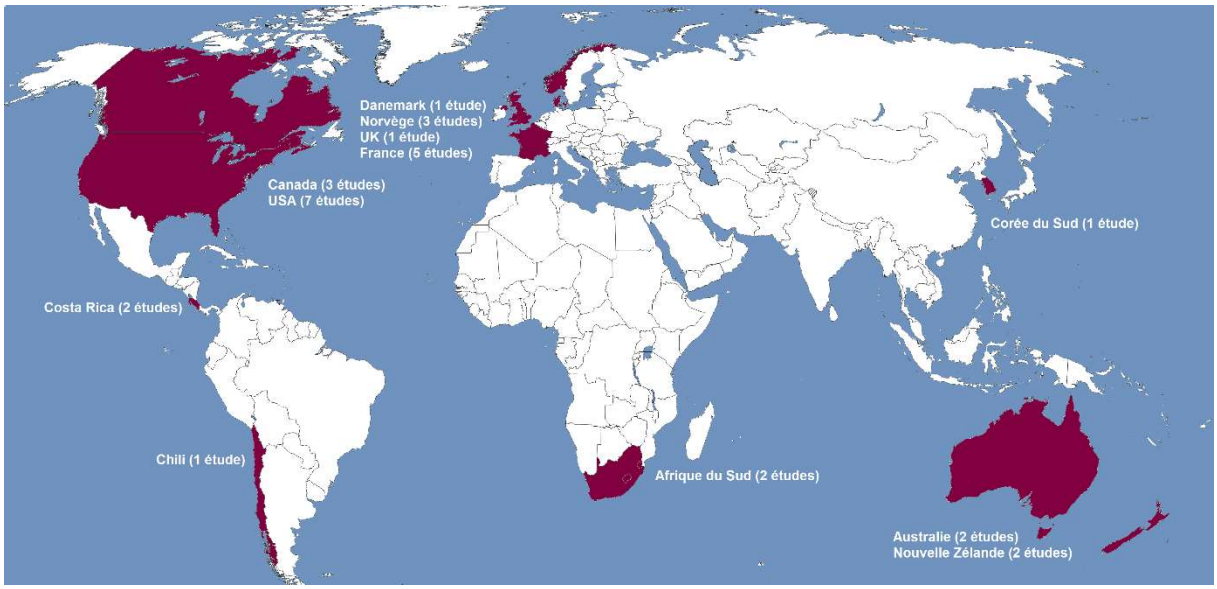


Figure 2 : Carte de répartition des études et cohortes regroupées dans le consortium AGRICOH en 2016

Depuis la création d'AGRICOH, un important travail d'harmonisation des données de trois cohortes (Etats-Unis, Norvège et France) a été mené et la mesure de l'exposition des travailleurs agricoles aux pesticides a été évaluée dans chacune d'entre elles (20). Puis, les données de ces trois cohortes ont permis de rechercher des associations entre l'usage de pesticides et l'apparition de lymphomes non hodgkiniens (21).

[1] En anglais : International Agency for Research on Cancer (IARC)

Parmi les cohortes participantes au consortium AGRICOH, l'une des plus connues est la cohorte américaine **Agricultural Health Study (AHS)**. Elle est pilotée par des institutions américaines : le National Cancer Institute (NCI), le National Institute of Environmental Health Sciences (NIEHS) en collaboration avec l'Environmental Protection Agency (EPA) et le National Institute for Occupational Safety and Health (NIOSH). Il s'agit d'une cohorte prospective lancé en 1993 aux Etats-Unis ayant pour objectif d'étudier les effets néfastes sur la santé des expositions agricoles, notamment ceux liés aux expositions aux pesticides. Elle regroupe environ 89 000 agriculteurs et leurs conjoints recrutés entre 1993 et 1997 et issus de deux états (l'Iowa et la Caroline du Nord) (22). Depuis 1993, les diverses études menées par la cohorte AHS ont permis d'obtenir de nombreux résultats montrant par exemple que l'incidence globale des cancers était moins élevée chez les travailleurs agricoles comparée à la population générale (23). Par ailleurs, certaines expositions aux pesticides seraient suspectées d'augmenter le risque de certaines pathologies telles que le cancer de la prostate (24), l'hypothyroïdisme (25), le diabète (26) ou la maladie de Parkinson (27).

Une autre cohorte agricole faisant partie d'AGRICOH, nommée **Cancer in the Norwegian Agricultural Population (CNAP)**, a été créée en Norvège afin d'étudier les cancers d'origine professionnelle. Des données provenant de recensements agricoles de 1969 à 1989 réalisés par le bureau central des statistiques de Norvège (SSB), contenant des informations sur les usages de pesticides, ainsi que des données de recensement du registre national de la population de Norvège ont été réunies. Ces données ont ensuite été couplées à des données du registre des cancers de Norvège précisant notamment la localisation des cancers via la classification internationale des maladies (CIM). Cette cohorte regroupe alors un total d'environ 136 000 exploitants, 110 000 conjoints et 260 000 enfants (17,28,29). De même que pour la cohorte AHS, les principaux résultats publiés indiquent que les travailleurs agricoles de Norvège sont en meilleure santé que la population générale. Les différences seraient expliquées par une meilleure hygiène de vie en termes de consommation d'alcool et de tabac et par un statut socio-économique potentiellement plus élevé. Par ailleurs, la cohorte CNAP a permis de mettre en évidence une association potentielle entre l'apparition de myélomes multiples et les cultivateurs de pommes de terre (17).

La troisième cohorte à avoir harmonisé ses données dans le cadre d'AGRICOH est la cohorte agricole française **AGRICAN (AGRIculture et CANcers)** qui a été constituée entre 2005 et 2007 dans le but d'étudier l'état de santé des travailleurs agricoles, notamment en termes de cancers par rapport à la population générale. Durant la phase d'inclusion de cette cohorte, plus de 180 000 individus ont été sélectionnés, qu'ils soient actifs ou retraités, tous affiliés au régime de sécurité sociale dédié au secteur agricole, la Mutualité Sociale Agricole (MSA). Diverses données ont été collectées provenant de questionnaires, de la MSA, de l'Institut

National de la Statistique et des Etudes Economiques (INSEE), des registres de cancers et du Centre d'épidémiologie sur les causes médicales de décès (CépiDc) (30,31). De plus, des estimations sur l'utilisation de pesticides sont réalisées à l'aide de la matrice cultures-expositions PESTIMAT (32). Une fois encore, les premières études menées par cette cohorte ont montré que le taux de mortalité par cancer en population agricole est inférieur à celui de la population générale, expliqué par les mêmes raisons qu'en Norvège avec la cohorte CNAP (33). En 2017, de nouvelles analyses ne montrent toujours aucune différence concernant l'incidence de tous les cancers confondus entre la population agricole et la population générale. Cependant, les analyses menées par type de cancer montrent des risques moins élevés de développer des cancers respiratoires dont la principale cause d'apparition serait la consommation de tabac, mais aussi des risques plus élevés de développer des cancers de la prostate, de la peau, de la lèvre, du cerveau, des lymphomes non-hodgkiniens mais aussi des maladies neurodégénératives (34–37).

En France, en complément des résultats actuellement apportés par la cohorte AGRICAN, d'autres dispositifs ont été mis en place afin de mettre en évidence les risques pour la santé liés au travail dans le secteur agricole. Citons, par exemple, le dispositif de toxicovigilance nommé **Phyt'attitude** mis en place par la MSA en 1991 pour le signalement d'effets (aigus, subaigus ou chroniques) de pesticides ou d'autres produits chimiques sur la santé. Ce réseau fonctionne principalement grâce à la participation bénévole de médecins du travail de la MSA et ne permet donc pas un recueil exhaustif. Cependant, il permet à la MSA de mettre en place des actions de prévention relatives à l'usage des produits chimiques (38).

Plus récemment, dans le cadre de la loi d'avenir agricole n°2014-1170 du 13 octobre 2014, un **dispositif de phytopharmacovigilance (PPV)** a été mis en place par l'Agence nationale de sécurité sanitaire de l'alimentation, de l'environnement et du travail (Anses). La phytopharmacovigilance a pour objet la surveillance des effets indésirables des produits phytopharmaceutiques sur la santé humaine, la biodiversité et faune sauvage, la contamination des milieux et l'apparition de résistance. Ce dispositif intègre alors plusieurs composantes dont la santé humaine, la biodiversité et la contamination des milieux, tout en s'appuyant sur les dispositifs prévus par les autres codes (rural, environnement, santé, travail).

De manière complémentaire à la cohorte AGRICAN et à ces dispositifs, une nouvelle cohorte agricole nommée **Coset-MSA** a été lancée par l'agence sanitaire Santé Publique France en partenariat avec la MSA. Cette cohorte a pour objectif d'établir des liens entre les expositions professionnelles du secteur agricole et la survenue de problèmes de santé, à partir de données récoltées grâce à des questionnaires. Par ailleurs, les expositions professionnelles aux produits phytopharmaceutiques seront estimées grâce aux **matrices cultures-expositions**

MATPHYTO développées par cet organisme. Après une première phase pilote menée entre 2010 et 2012, les individus ont été inclus de début 2017 à fin 2018 et les premiers résultats devraient être communiqués en 2019 ou en 2020 (39–41).

Enfin, l'évaluation *a posteriori* des expositions est un vrai challenge, d'autant plus que la littérature sur le sujet des effets sur la santé des pesticides est vaste et de qualité inégale. Signalons que l'Anses a publié en 2016, un rapport assez exhaustif sur les expositions professionnelles aux pesticides des travailleurs agricoles (42) et que l'Institut national de la santé et de la recherche médicale (Inserm) a publié en 2014 un état de l'art sur les effets sur la santé des pesticides (43).

Plusieurs cohortes et dispositifs existent en France et dans le monde afin de caractériser les expositions professionnelles liées au secteur agricole et d'identifier les effets néfastes potentiels sur la santé de ces différentes expositions. Une brève revue de la littérature scientifique a ainsi permis de dresser un aperçu des liens potentiels recensés chez des travailleurs agricoles entre ces expositions et des pathologies (Tableau 1).

Tableau 1 : Aperçu des effets néfastes potentiels avérés et suspectés sur la santé des travailleurs agricoles par type de risque

Risques professionnels liés aux activités agricoles	**Détails des risques**	**Exemples de effets néfastes potentiels sur la santé**
Physiques	Travail physique avec des machines, des véhicules Exposition à un bruit excessif, à des vibrations Exposition prolongée au soleil, à des températures extrêmes ou à des intempéries	Troubles musculosquelettiques et lombalgies (44–47) ; Perte de l'audition (48) ; Cancers de la peau et de la lèvre (49–51) ; Stress thermique (52) ; Maladies de la peau (53)
Biologiques (Allergiques, Toxiniques)	Exposition aux animaux Exposition à des poussières ou autres substances organiques (virus, pathogènes, …)	Broncho-pneumopathie chronique obstructive (54–56) ; Maladies de la peau (53) ; Zoonoses (57–59)
Chimiques	Exposition à des produits phytosanitaires ou autres produits chimiques (solvants, …)	Maladies neurodégénératives (27,37,60–63, R-1) ; Troubles respiratoires (64–67) ; Hypothyroïdisme (25) ; Troubles reproductifs (68) ; Arthrite Rhumatoïde (69) ; Diabètes (26,70) ; Anomalies congénitales (71) ; Maladies de la peau (53) ; Cancers : prostate (24,35,72,73), vessie (74,75), foie (76), poumons (77,78), cerveau (36), autres (23,34,53,79)
Psycho-sociaux	Ruralité, Isolement social Perte de revenus liés aux intempéries ou aux crises économiques	Dépression et troubles de la santé mentale (80,81) ; Suicide (82,83)

Ainsi, les cohortes susmentionnées ont le principal avantage d'avoir des données de grande précision, qui ont permis de mettre en évidence un certain nombre de risques professionnels dans le secteur agricole. Soulignons en particulier la très bonne maîtrise des caractéristiques de la population recrutée, la précision avec laquelle sont renseignées les trajectoires professionnelles des individus ainsi que leurs expositions aux principales familles de pesticides, mais également le renseignement de facteurs de confusion comme l'exposition à d'autres facteurs de risque non professionnels (ex : tabagisme). Cependant, de manière générale, les études de cohortes comportent également un certain nombre d'inconvénients : le coût élevé de mise en place et de suivi, le laps de temps nécessaire à l'inclusion prospective des participants et à la collecte de données, le fait qu'elles n'étudient qu'un échantillon de la population souhaitée, et surtout le fait qu'il existe une période d'inclusion déterminée et limitée. En effet, si le suivi de la population a vocation à être prolongé sur des décennies, les inclusions sont stoppées au-delà d'une certaine date ou d'un certain effectif, ce qui ne permet pas d'envisager une vigilance prospective. Il est tout de même important de noter que dans certains cas, les cohortes peuvent également être ouvertes ou dynamiques si cela peut permettre d'améliorer leur performance. De plus, l'interrogation régulière des personnes suivies permet également une mise à jour des expositions et l'addition d'informations complémentaires.

Ainsi, dans l'optique de disposer d'indicateurs pérennes et utiles à la vigilance, des méthodes complémentaires à ces études épidémiologiques peuvent être développées à partir des données de source assurantielle préexistantes. En effet, l'analyse de ces données médico-administratives, recueillies en routine et très riches en informations, pourrait être un atout pour le suivi épidémiologique en santé au travail de la population agricole (84,85)**.**

III. Analyse de données médico-administratives

La France dispose d'une grande quantité de bases de données médico-sociales et économiques gérées par des organismes publics, qui couvrent de façon quasi-exhaustive et permanente l'ensemble de la population. C'est le cas notamment de l'Assurance maladie qui détient une quantité « massive » de données médico-administratives, créées à des fins de gestion budgétaire, en particulier pour le remboursement et le suivi des dépenses de santé. Ces données présentent un potentiel considérable pour la recherche épidémiologique et pharmaco-épidémiologique du fait de leur disponibilité, de leur exhaustivité et du coût « nul » associé à leur extraction, dans une perspective de réutilisation (R-2). Cependant, elles sont encore largement sous-exploitées même si leur utilisation dans le domaine de la recherche

scientifique s'est largement développée au cours de ces dernières années. On observe une même tendance dans le monde entier comme par exemple à Taiwan (R-3), au Japon (R-4), au Canada (R-5) ou encore aux États-Unis (R-6, R-7, R-8).

En France, depuis quelques années déjà, les agences (Santé Publique France, Haute autorité de Santé, Agence nationale de sécurité du médicament, …) travaillent sur ces données notamment dans le cadre de la surveillance sanitaire (86, R-9). En effet, ces données peuvent être analysées dans le but d'améliorer la santé et le bien-être de la population, de réduire les dépenses de santé, de prévenir l'apparition de maladies ou d'épidémies, de détecter des évènements de santé inhabituels susceptibles de constituer une alerte de santé publique ou encore d'extraire des informations nécessaires à la prise de décisions éclairées par les pouvoirs publics (87,88). D'ailleurs, les données médico-administratives provenant du Système National des Données de Santé (SNDS)[2] ont montré leur intérêt et la pertinence de leur utilisation en contribuant à alimenter des dispositifs pour la surveillance de maladies chroniques ou de maladies infectieuses (89,90). Par exemple, la cohorte française CONSTANCES (cohorte des consultants des centres d'examens de santé), constituée d'un échantillon de 200 000 adultes, procède notamment à un « suivi passif » des individus inclus en recueillant des données de systèmes nationaux dont le SNDS. Ceci en fait un outil utile pour la surveillance épidémiologique, qui se décline au travers de plusieurs projets, en collaboration avec Santé Publique France (R-10, 89).

D'un point de vue technique, l'utilisation des données de source assurantielle à des fins épidémiologiques et de surveillance nécessite un savoir-faire ainsi qu'une bonne connaissance des limites de ces données, du fait de leur complexité, de leur imprécision et de leur volume. De plus, contrairement aux données de registres[3] ou récoltées pour des cohortes, le contrôle des biais liés aux données (qualité, cohérence, …) est bien moins aisé : différences possibles d'accès aux soins au sein de la population couverte, imprécision des données, difficultés pour évaluer la qualité du codage (pouvant varier géographiquement par exemple), absence de données cliniques ou de facteurs de risque comportementaux majeurs (ex : tabagisme), … (R-12, R-13). Un temps important consacré au nettoyage, au contrôle et à la compréhension de ces données pour en vérifier la fiabilité et en mesurer les limites, pour certaines incontournables, est donc un préalable indispensable avant d'entamer toute analyse

[2] Système regroupant l'ensemble des données de l'Assurance Maladie (Système National d'Informations Inter-Régimes de l'Assurance Maladie, « Sniiram ») et des hôpitaux mais aussi les causes médicales de décès (CépiDC) et les données relatives au handicap. Ce système a pour finalité la mise à disposition des données de santé en France afin de favoriser les études contribuant par exemple à la surveillance, à la veille et à la sécurité sanitaire (https://www.snds.gouv.fr/).

[3] Exemple : Dans les pays d'Europe du Nord, le croisement de registre du cancer aves les données de métier issues du recensement avec NOCCA, la « Nordic Occupational Cancer Study » (R-11)

statistique complexe (84, R-14). Il s'agit ensuite de réaliser les jointures et transformations nécessaires de ces données, pouvant être issues de nombreuses tables, pour permettre les analyses. Le choix éclairé des méthodes statistiques ou épidémiologiques classiques et/ou de méthodes issues plus largement des sciences de données et du machine learning est alors un challenge important. Ce choix est fonction tant des objectifs (ex : surveillance temporelle, génération d'hypothèses, etc) que des caractéristiques des données (R-15, R-16).

Ainsi, afin d'extraire des informations nouvelles et utiles à partir de ce type de données « massives », il convient d'utiliser des méthodes de fouille de données ou *data mining*. Le data mining désigne l'ensemble des méthodes et techniques destinées à l'exploration et l'analyse de grandes bases de données. Dans le domaine de la santé, ces méthodes sont déjà largement utilisées que ce soit pour des applications :

- **descriptives,** afin de permettre la mise en évidence d'informations présentes mais cachées par le volume de données (exemples : recherche d'associations entre des expositions environnementales et des maladies chroniques) ;

- ou **prédictives,** afin d'extrapoler de nouvelles informations à partir des données connues (exemple : prédiction du temps de rétablissement après une opération) (91,92, R-16).

Dans certains cas, les deux types d'applications sont utilisées. Par exemple, si le but est de suivre la propagation spatio-temporelle d'une épidémie, il s'agira dans un premier temps de décrire la situation à un instant t mais aussi de prédire l'évolution de la situation afin de mettre en place des actions de prévention. Pour ces différentes applications, la modélisation est souvent utilisée afin de traduire le phénomène voulant être étudié en langage mathématique. Selon la problématique de l'étude, diverses méthodes de modélisation peuvent être utilisées dont les plus connues, de façon générale, sont la régression linéaire, la régression logistique, les modèles de survie, les forêts aléatoires ou les réseaux de neurones.

IV. Présentation du projet de recherche

Sous l'impulsion de son conseil scientifique, la MSA, l'organisme de sécurité sociale dédié à la population agricole, a souhaité développer son activité de vigilance des risques professionnels en exploitant ses bases de données assurantielles, utilisées à des fins de remboursement de prestations de santé. Les bases de données de la MSA sont extrêmement riches et contiennent une grande quantité d'informations sur leurs assurés. En effet, ces données sont séparées en deux types de flux, administratifs (âge, sexe, activité professionnelle, revenus, …) et médico-administratifs (pathologies, dépenses de santé, …). Concernant les flux médico-administratifs, ils sont regroupés au sein de trois bases de données : une pour les déclarations d'Accidents du Travail et maladies professionnelles (AT/MP), une pour les déclarations d'Affections de Longue Durée (ALD) et enfin, une pour les remboursements de médicaments, de soins et de services médicaux appelé « RAAMSES ». Concernant les données des maladies professionnelles, elles ne sont pas en mesure de renseigner de façon globale et robuste les effets sanitaires liés au travail, car les données sont réduites tant quantitativement (sous-déclaration potentielle) que qualitativement (liste limitative d'affections et dans les faits, plus de 90% des affections reconnues sont des troubles musculo-squelettiques). A un niveau plus général, les problématiques de santé des individus, quelle que soit leur étiologie, donnent lieu à des comportements de consommation de soins et de produits de santé qu'il pourrait être intéressant d'investiguer.

A la demande de la MSA, le laboratoire TIMC-IMAG (Techniques de l'Ingénierie Médicale et de la Complexité - Informatique, Mathématiques, et Applications, Grenoble) a proposé un projet de fouille de ses données médico-administratives, pour en explorer l'intérêt et le potentiel, à des fins de génération d'hypothèses pour la vigilance des risques professionnels des travailleurs agricoles, en particulier ceux associés à l'usage des produits phytosanitaires. Le laboratoire TIMC-IMAG réunit aussi bien des scientifiques que des cliniciens autour de l'utilisation de l'informatique et des mathématiques appliquées pour la compréhension et le contrôle des processus normaux et pathologiques en biologie et santé. Un des paradigmes de ce laboratoire est de transformer la donnée en connaissance. Au sein du laboratoire, trois équipes de recherche sont impliquées dans ce projet : Environnement et Prédiction de la Santé des Populations (EPSP), Biologie Computationnelle et Mathématique (BCM) et Techniques pour l'Évaluation et la Modélisation des Actions de la Santé (ThEMAS).

Par ailleurs, dans le cadre de sa nouvelle mission de phytopharmacovigilance présentée ci-dessus, l'Anses s'intéresse aussi aux différentes sources de données pouvant être informatives sur l'impact sanitaire des produits phytopharmaceutiques.

C'est donc en étroite collaboration avec la MSA que ce projet a été initié par le laboratoire TIMC-IMAG notamment grâce aux deux Conventions de Recherche et Développement établies avec l'Anses dans le cadre de leurs missions de PPV. L'objectif de ce projet consiste donc à **exploiter les données de la MSA afin de tester, sans hypothèses préalables, l'existence ou non de liens entre chaque type d'activité professionnelle ou de culture (« proxy » de l'exposition) et chaque pathologie observée,** que cette dernière soit identifiée directement et sans ambiguïté par un code diagnostic (situation spécifique des ALD), ou qu'elles puissent être déduites indirectement d'une consommation de médicaments ou de soins (« proxy ») (Figure 3). De plus, il pourra être possible d'utiliser des sources de données externes à la MSA pour tenter de préciser l'exposition professionnelle, via la Registre Parcellaire Graphique (RPG) ou le Recensement Agricole (RA) par exemple.

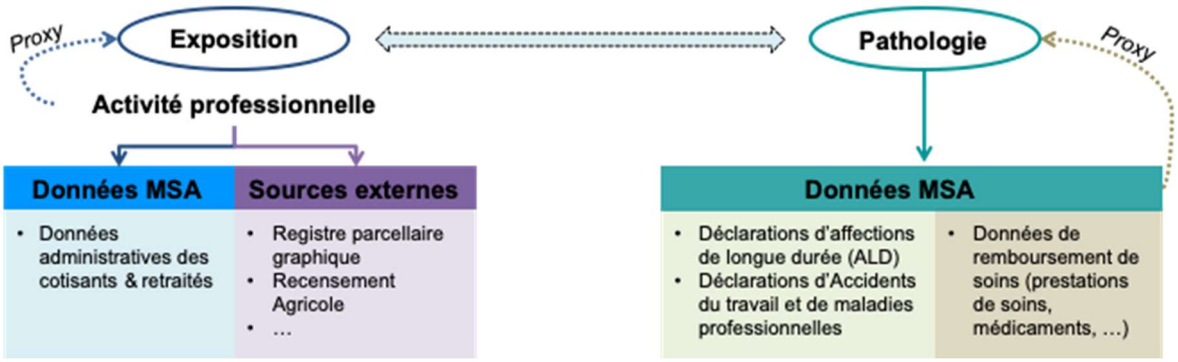


Figure 3 : Schématisation de la problématique du projet de fouille des données médico-administratives de la MSA réalisé par le laboratoire TIMC-IMAG

Les associations mises en évidence entre chaque type d'activité et chaque pathologie, ajustées sur les facteurs de confusion qui pourront être renseignés, devront ensuite être interprétées pour isoler celles qui semblent pertinentes notamment vis-à-vis de la littérature scientifique. L'identification d'associations pourra ensuite conduire à des travaux ciblés (études épidémiologiques, toxicologiques, évaluations des risques, …) visant à confirmer et mieux comprendre les risques professionnels en question afin de mieux les prévenir.

Ce type d'approche, si elle s'avère pertinente, serait à même de produire des informations au fil de l'eau si les analyses sont relancées sur les bases de données actualisées (intérêt pour la vigilance). Il s'agit donc d'une approche complémentaire aux études épidémiologiques basées sur la population agricole. En effet, l'approche proposée est systématique, sans *a priori*, et sans coût associé d'acquisition des données brutes. La puissance de ce type d'étude est considérable puisqu'elle prend en compte toute la population agricole française. Néanmoins, les données de la MSA ne renseignant pas directement les expositions, cette puissance est contrebalancée par le manque de sensibilité ou de spécificité à ce niveau (estimation via les types d'activité et de culture).

V. Cadre et objectifs de la thèse

Dans le cadre de ce travail de thèse, **le premier objectif est d'évaluer la faisabilité du projet, c'est-à-dire, l'utilisation des données médico-administratives de la MSA à des fins de vigilance sanitaire.** Pour cela, dans un premier temps, un croisement des données administratives des cotisants avec leurs données médico-administratives est nécessaire. Ainsi, grâce aux données fusionnées, des méthodes statistiques peuvent être appliquées pour rechercher des associations, sans hypothèses préalables, entre chaque activité professionnelle, considérée alors comme un « proxy » de l'exposition, et chaque pathologie chronique déclarée en tant qu'ALD.

Puis, une fois la faisabilité du projet démontrée, **le deuxième objectif est d'évaluer la pertinence et la robustesse de la méthode statistique utilisée et donc des associations statistiques obtenues.** Des optimisations de la méthode sont ainsi réalisées de sorte à minimiser les biais potentiels, à améliorer la robustesse afin que cette méthode puisse être utilisée en routine sur les données de la MSA, qui seraient idéalement fournies au fil de l'eau. Enfin, les associations statistiques mises en évidence sont comparées à la littérature scientifique. Toutefois, dans le cadre de ce travail, il est à noter que le but n'est pas de discuter la pertinence des signaux d'un point de vue médical ou toxicologique.

Cette thèse a été financée par le programme AGIR-POLE de l'Université Grenoble Alpes qui vise à soutenir des projets scientifiques ou technologiques innovants, des idées novatrices, et le développement de collaborations approfondies entre laboratoires ou entre équipes.

PARTIE 2

Description, traitement et analyses des données de la MSA

I. Description des données de la MSA

a. La Mutualité Sociale Agricole

La Mutualité Sociale Agricole (MSA) est le régime de sécurité sociale dédié à la population agricole. Cette population comprend à la fois des travailleurs actifs (salariés, saisonniers, chefs d'exploitation ou d'entreprise, aidants familiaux, …), des retraités et aussi leurs ayants droit (membres de la famille). Au 1er janvier 2017, 5.6 millions d'individus étaient ressortissants au régime agricole, c'est-à-dire, qu'ils percevaient au moins une prestation au régime agricole (8). Parmi ces derniers, 3.2 millions de personnes étaient protégés au titre de la maladie, c'est-à-dire, qu'ils pouvaient bénéficier de remboursements liés à leurs soins. Parmi ces individus sont comptés aussi bien les ouvrants-droit représentant 2.5 millions d'individus que les ayants droit représentant environ 700 000 individus (conjoints, enfants ou autres membres de la famille) (Figure 4).

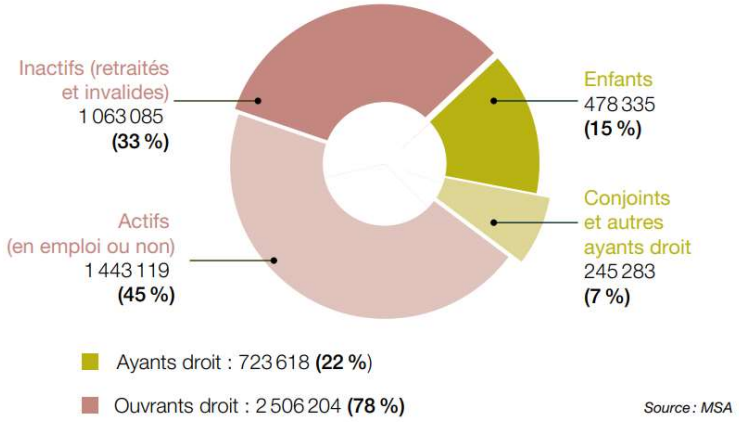


Figure 4 : Distribution des personnes protégées au titre de la maladie à la MSA au 1er janvier 2017

Au sein du régime agricole, la population des actifs, représentant près d'1.2 million d'individus, est divisée en deux sous-groupes : les **salariés** et les **non-salariés**. La population de **non-salariés agricoles** qui incluait 484 600 individus (≈ 42%), se constitue de chefs d'exploitation ou d'entreprise, de conjoints collaborateurs, d'aidants familiaux et de cotisants solidaires. Les cotisants solidaires sont des individus qui dirigent une exploitation ou une entreprise agricole mais leur superficie d'exploitation est inférieure à la surface minimale d'assujettissement[4], ils

[4] Surface d'exploitation départementale fixée par arrêté préfectoral, exprimée en hectares et définie en fonction du type de culture ou d'élevage.

consacrent entre 150 et 1 200 heures par an à leur activité agricole et les revenus perçus liés à leurs activités agricoles sont limités. Quant à la population des **salariés**, elle comptait 678 092 individus qui pouvaient être permanents ou saisonniers, employés par exemple dans des exploitations (cultures ou élevages), dans des entreprises de travaux agricoles, dans le secteur coopératif ou dans d'autres activités du secteur tertiaire (exemple : employés de la banque Crédit Agricole ou des caisses de la MSA).

En effet, étant donné la diversité des activités agricoles, il est important de savoir que l'affiliation des travailleurs, salariés ou non-salariés, au régime de la MSA n'est pas toujours évidente. Pour rappel, les activités agricoles regroupent bien évidemment toutes les formes de cultures et d'élevages mais aussi à titre d'exemple : les activités équestres, les travaux forestiers, les activités de conditionnement et de commercialisation de produits agricoles, les entreprises de travaux agricoles (exemple : travaux d'entretien de parcs et jardins) mais aussi les personnels enseignants des établissements d'enseignement agricoles ou encore les personnes exerçant des activités agro-touristiques (exemple : fermes équestres, campings à la ferme).

Aussi, pour l'ensemble de ses affiliés, la MSA collecte des informations à des fins de gestion administrative et de remboursement de soins. Pour des raisons de confidentialité, ces données administratives et médico-administratives sont structurées et séparées en différentes tables, correspondant chacune à des flux particuliers de données (Figure 5).

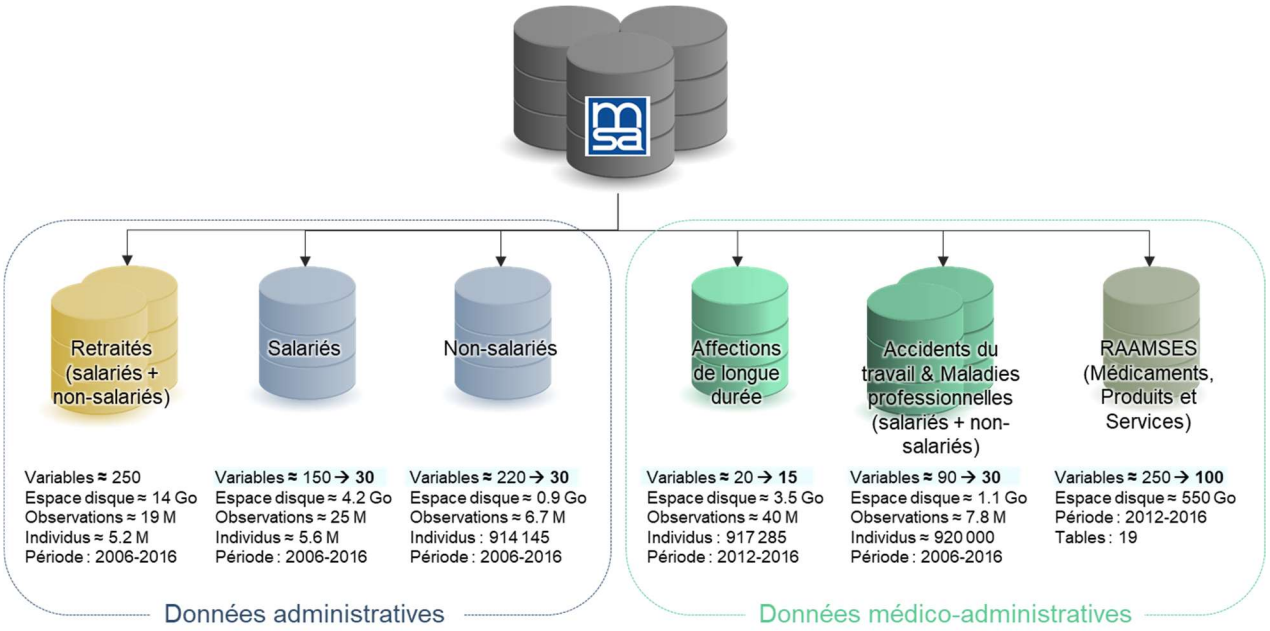


Figure 5 : Structure et caractéristiques des données brutes de la MSA

Ainsi, certains flux sont réservés aux **données administratives** « pures » concernant les cotisants, avec une table dédiée aux non-salariés, une table pour les salariés et deux tables pour les retraités agricoles ayant été salariés ou non-salariés. Les autres flux concernent les **données médico-administratives** ayant pour finalité de procéder aux remboursements des cotisants aussi appelés « liquidations ». Il existe ainsi une table dédiée aux Affections de Longue Durée (ALD), deux autres pour les Accidents du Travail et les Maladies Professionnelles (AT/MP) selon le statut de salarié ou de non-salarié, et un ensemble d'autres tables concernant les remboursements de médicaments, de soins et de services médicaux rassemblées sous la dénomination « RAAMSES ».

D'une part, les données de travail utilisées par la Direction des Etudes Répertoires Statistiques de la Caisse Centrale de la MSA (CCMSA) sont caractérisées par le fait qu'elles possèdent des identifiants différents empêchant jusqu'à présent leur croisement. D'ailleurs, dans le cadre de ce travail et après obtention de l'accord de la Commission Nationale de l'Informatique et des Libertés (CNIL), un identifiant unique a donc été créé par la MSA pour nous permettre de relier les différentes tables de données. D'autre part, il est aussi important de savoir que les données ne possèdent pas non plus les mêmes données d'entrée. Ainsi, dans les données administratives, une observation fait référence à un individu (un employé dans un contrat ou un chef d'exploitation dans une exploitation), tandis que dans les données médico-administratives, les observations font référence à des feuilles de soins ou à des prestations de soins. Par ailleurs, il est également important de préciser qu'il est généré une table de données par année pour la plupart des flux sauf en ce qui concerne les données « RAAMSES » où chaque table de données est générée par trimestre. Il est donc nécessaire de fusionner les données pour obtenir un historique d'informations sur les cotisants.

En outre, étant donné la volumétrie des données et les nombreuses informations assurantielles et financières renseignées au sein des données de la MSA, avant la transmission des données, il a été nécessaire de sélectionner les variables les plus pertinentes pour les analyses menées dans le cadre de ce projet (Figure 5). En effet, les données de la MSA comptent près de 1 000 variables et il a donc été choisi de conserver uniquement celles qui permettaient d'avoir suffisamment d'informations sur les cotisants telles que le sexe, l'année de naissance, les activités professionnelles, les pathologies, les revenus, les dates des événements (exemple : début d'activité, année d'installation de l'exploitation) ainsi que les numéros d'identification individuels. Les variables sont ainsi sélectionnées dans le but de fusionner les différentes tables de données, pour les analyses statistiques ou bien encore pour l'interprétation des résultats.

b. Les flux administratifs

Les flux administratifs concernent les informations sur les cotisants de la MSA ayant le statut d'ouvrants droit, qu'ils soient chefs d'exploitation ou d'entreprise (non-salariés), salariés ou retraités. Pour chaque table de données annuelle, chaque ligne correspond à un contrat d'un salarié, à une exploitation ou une entreprise d'un non-salarié ou à la pension de retraite annuelle d'un retraité. Au sein de ces données, il est ainsi possible de retrouver différents types d'information :

- Individuelle : numéro d'identification, sexe, année de naissance, situation familiale, … ;

- Professionnelle : types d'activités professionnelles (thésaurus interne de la MSA et codes de la Nomenclature d'Activités Française), dates d'activités, type de contrat, année d'installation de l'exploitation, type d'exploitation, superficie d'exploitation, revenus, … ;

- Assurantielle : numéro de département des caisses MSA, type de régime, …

Concernant les informations sur l'activité professionnelle, la MSA utilise deux thésaurus différents : la Nomenclature d'Activités Française (NAF) mais aussi un thésaurus interne. Concernant la NAF, il s'agit d'une nomenclature des activités économiques productives, élaborée par l'INSEE, principalement pour faciliter l'organisation de l'information économique et sociale. Cette nomenclature comporte cinq niveaux emboîtés avec un total de 732 codes renseignant des activités professionnelles spécifiques. Quant au thésaurus interne de la MSA, il s'agit d'un code renseignant une activité professionnelle spécifique plus *à risque* pour laquelle chaque contributeur paie une cotisation. En effet, cette variable, appelée « Risque » par la MSA, correspondant à une catégorie de risque qui doit être déclarée obligatoirement tous les ans, qui est soumise à des contrôles et dont la mise à jour est faite régulièrement contrairement au code NAF. Dans le cas où l'exploitant a de multiples activités, cette variable correspond alors à l'activité professionnelle majoritaire en termes de quotité de travail accordé. Cependant, l'inconvénient majeur de cette variable est le nombre de modalités qu'elle comporte pour décrire l'activité professionnelle (26 et 43 modalités respectivement pour les non-salariés et les salariés). Malgré cet inconvénient, comme le renseignement de cette variable est davantage contrôlé au sein des données administratives de la MSA, son utilisation a été préférée dans le cadre de ce travail. Par la suite, l'activité professionnelle renvoie donc exclusivement à cette variable « Risque » renseignée par la MSA.

Par ailleurs, il est important de préciser que les données des retraités fournies par la MSA, qu'ils soient salariés ou non-salariés, ne comportent aucune information à propos des anciennes activités professionnelles exercées. Cette information étant primordiale pour les analyses statistiques réalisées dans le cadre de ce travail, ces données n'ont pas été considérées. Cependant, en raison de la mise à jour annuelle des données administratives, il est possible que les données des salariés ou des non-salariés puissent contenir des « jeunes retraités ». En effet, si les individus étaient actifs à un instant donné, et deviennent ensuite retraités, ils sont alors présents dans les données des actifs et ensuite, potentiellement dans les données des retraités. Cependant, si un individu n'est plus présent dans les données des actifs, cela ne veut pas nécessairement dire qu'il a pris sa retraite et qu'il est présent dans les données des retraités, il peut avoir changé de régime de sécurité sociale ou être décédé.

Concernant la population des non-salariés, elle est constituée aussi bien de chefs d'exploitation ou d'entreprise, que de conjoints collaborateurs, de cotisants solidaires et d'aidants familiaux. Dans le cadre de ce travail, malgré le statut « actif » des conjoints collaborateurs et des aidants familiaux, il a été choisi d'exclure ces individus du fait d'un problème de rattachement d'une partie des données afférentes. En effet, dans les données administratives des non-salariés, leur numéro d'identification individuel est le même que celui du chef d'exploitation auquel ils sont rattachés alors que dans les données médico-administratives, nous disposons du numéro d'identification individuel. A partir des données transmises par la MSA, récupérer les données médico-administratives des conjoints collaborateurs et des aidants familiaux n'a pas été possible, d'où leur exclusion pour les analyses dans le cadre de ce travail. Cette exclusion a pu être réalisée grâce à une variable « statut » fournie par la MSA et permettant la distinction entre les chefs d'exploitation, les conjoints collaborateurs, les aidants familiaux et les cotisants solidaires. Après l'exclusion de ces individus, chaque individu conservé, qu'il soit cotisant solidaire ou chef d'exploitation, est alors identifié par son propre numéro d'identification individuel. Par ailleurs, du fait de cette complexité due aux nombreux statuts chez les non-salariés, les « jeunes retraités » n'ont pas pu être identifiés au sein des données des non-salariés.

Dans le cadre du projet de recherche, l'historique disponible pour les données administratives est de 11 ans, de 2006 à 2016, que ce soit pour les salariés, les non-salariés ou les retraités. Dans le cadre de ce travail de thèse et du manuscrit, seules les données des non-salariés ont été prises en compte.

c. Les flux médico-administratifs

Les flux médico-administratifs correspondent quant à eux aux déclarations d'ALD et d'AT/MP d'une part, et aux remboursements de médicaments et prestations de soins d'autre part. Au sein de ces flux, on peut retrouver à la fois les ouvrants droit et les ayants droit.

Affections de longue durée, accidents du travail et maladies professionnelles

En France, si un individu développe une pathologie chronique qui nécessite un traitement prolongé et particulièrement onéreux, il lui est possible de demander une reconnaissance de sa pathologie en ALD. Pour cela, la pathologie doit être inscrite sur la liste des ALD conçue par l'assurance maladie (Annexe 1). Une fois l'ALD reconnue, l'individu est pris en charge à 100% concernant les soins et les traitements liés à son ALD.

Par ailleurs, l'individu peut dans certains cas demander que sa maladie soit reconnue et prise en charge au titre des maladies professionnelles. Ceci est avantageux à plusieurs titres : prise en charge des dépenses de soins, prise en charge avantageuse des indemnités journalières et surtout, indemnisation des séquelles sous la forme d'un capital ou d'une rente viagère en fonction du taux d'Incapacité Permanente Partielle (IPP). Il existe différentes possibilités permettant de faire reconnaître une affection comme professionnelle d'un point de vue médico-légal. De manière générale, il faut à la fois que la pathologie et les principaux travaux susceptibles d'exposer au risque de développement de cette pathologie soient inscrits au sein de l'un des tableaux de maladies professionnelles établis par le régime correspondant de sécurité sociale (ici, le régime agricole). Il y a alors présomption d'imputabilité professionnelle si toutes les conditions sont remplies : critères diagnostics, d'exposition, délai de prise en charge entre la fin de l'exposition au risque et la date de diagnostic, et parfois, une durée minimale d'exposition au risque[5]. En dehors de ces situations, et en particulier lorsque la pathologie en question ne fait pas l'objet d'un tableau de maladie professionnelle et qu'elle présente une gravité suffisante (taux d'incapacité permanente partielle prévisible d'au moins 25%), la maladie peut être déclarée et le dossier sera étudié par un comité régional de

[5] A titre d'exemple, depuis mai 2012, la maladie de Parkinson est citée dans le tableau 58 du régime agricole en tant que « maladie de Parkinson provoquée par les pesticides ». Ce dernier tableau propose une liste indicative des principaux travaux susceptibles de provoquer la maladie définie comme suit : « Travaux exposant habituellement aux pesticides : lors de la manipulation ou l'emploi de ces produits, par contact ou par inhalation ; et/ou par contact avec les cultures, les surfaces, les animaux traités ou lors de l'entretien des machines destinées à l'application des pesticides ». Ce tableau fixe un délai de prise en charge de 1 an, et une durée d'exposition de 10 ans au minimum. Si toutes les conditions sont remplies, la maladie de Parkinson sera considérée par défaut comme professionnelle. Si les délais administratifs ne sont pas remplis, le C2RMP se prononcera sur l'imputabilité.

reconnaissance des maladies professionnelles (C2RMP), chargé de se prononcer sur l'existence d'un lien direct et essentiel entre l'affection et l'exposition professionnelle. Le système des AT permet de bénéficier des mêmes droits ; seul le mode d'entrée diffère (lésion accidentelle survenue sur le lieu et le temps du travail).

Pour un individu, une même pathologie ne peut être reconnue en même temps en ALD et en MP. Par ailleurs, la demande de reconnaissance en ALD ou en MP étant différente notamment en termes de prise en charge, la pathologie peut être déclarée en premier lieu en ALD puis en MP l'année suivante. De plus, la réglementation ayant été modifiée au cours du temps, certaines pathologies sont apparues dans les tableaux de maladies professionnelles. C'est le cas de la maladie de Parkinson par exemple qui a pu être prise en charge en tant que MP à partir de mai 2012. De ce fait, pour certaines pathologies, il peut être important de prendre en compte à la fois les déclarations d'ALD et les déclarations de MP.

Au sein du régime agricole, qu'ils soient salariés ou non-salariés, les cotisants ont une assurance obligatoire contre les accidents du travail et les maladies professionnelles. Ainsi, les données relatives aux déclarations d'AT ou de pathologies, qu'elles soient reconnues en ALD ou en MP, sont centralisées à la MSA. Au sein de la table de données des ALD et des tables des AT/MP (salariés et non-salariés), il est alors possible de retrouver les informations suivantes :

- ALD : type d'ALD déclaré, code associé de la Classification Internationale des Maladies (CIM-10)[6], début et fin de prise en charge de l'ALD, type de régime (salarié ou non-salarié), qualité du bénéficiaire (ouvrant droit ou ayant droit), … ;

- AT/MP : type d'AT/MP, circonstances (agent causal) et date de survenue de l'événement, activité professionnelle exercée lors de l'événement, nature et siège de la lésion, …

Dans le cadre du projet de recherche, l'historique disponible pour les données ALD est de 5 ans maximum de manière légale, c'est-à-dire, de 2012 à 2016. Quant aux données AT/MP, l'historique disponible est de 10 ans, de 2006 à 2016. Dans le cadre de ce travail de thèse et du manuscrit, seules les données ALD ont été prises en compte.

[6] Classification Internationale des Maladies, 10ème révision : classification publiée par l'Organisation Mondiale de la Santé permettant de coder les maladies, signes, symptômes, circonstances sociales et causes externes de maladies ou de blessures. La classification possède différents niveaux de précision avec un total de 14 400 codes différents.

« RAAMSES », Médicaments et Prestations de soins

Lorsqu'un individu affilié à la MSA bénéficie d'un remboursement lié à des soins quels qu'ils soient, les données sur les médicaments et prestations de soins sont enregistrées dans différentes entités selon le type d'information et de prestation considérée, le tout rattaché à la dénomination « RAAMSES ». Ces entités sont hiérarchisées et imbriquées autour de l'élément central représenté par une prestation de soin, enregistrée dans l'entité « E400 » (Figure 6).

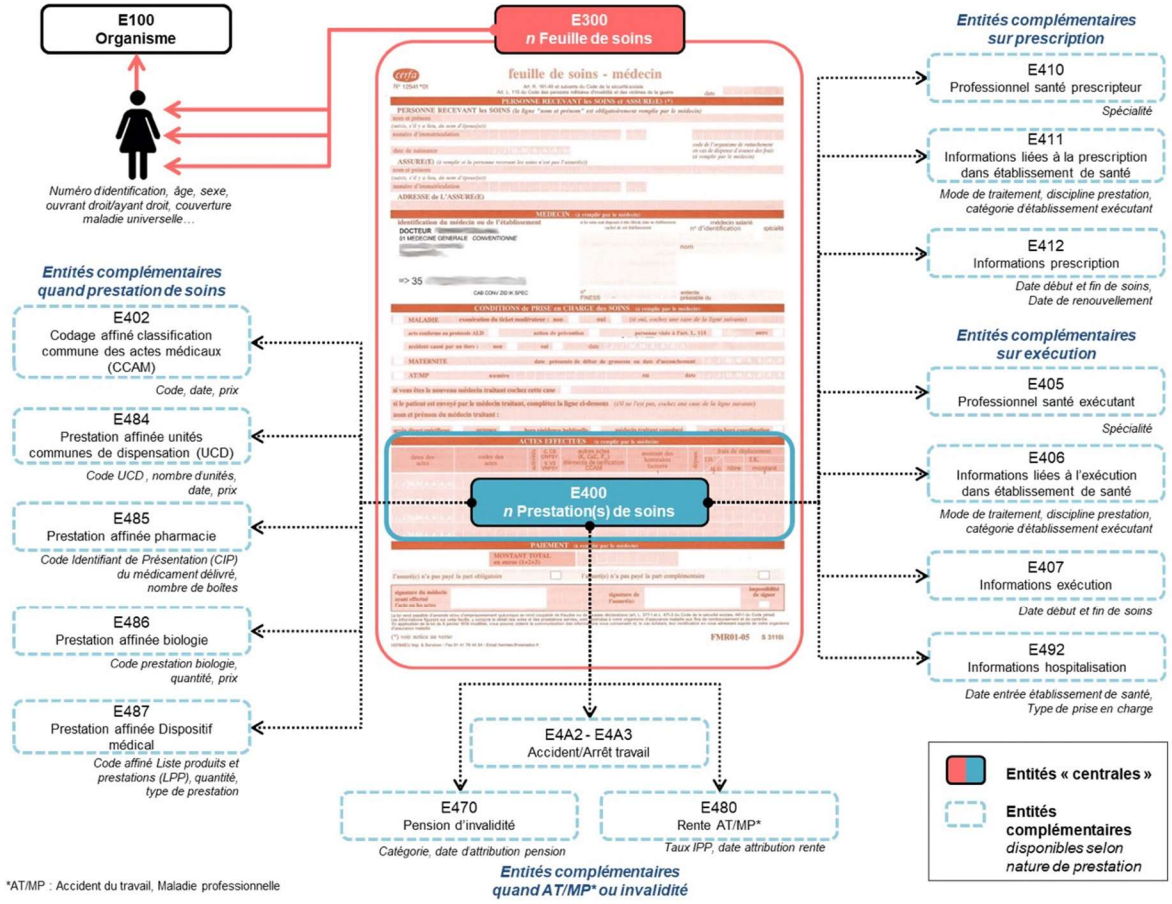


Figure 6 : Hiérarchie et structure des données médico-administratives RAAMSES (médicaments et prestations de soins) de la MSA

Pour chacune des prestations de soins enregistrées dans l'entité « E400 », différentes informations sont associées et stockées dans les différentes entités complémentaires (identifiées par le code E4XX sur la Figure 6). Au sein des données « RAAMSES », on peut retrouver de nombreuses informations sur les prestations de soins (exemples : type de la prestation, dates, spécialité du professionnel de santé) ou sur les médicaments (exemples : code du médicament, doses et/ou nombre de boîtes prescrites).

Dans le cadre du projet de recherche, l'historique disponible pour ces données est de 5 ans, de 2012 à 2016. Dans le cadre de la thèse, ces données n'ont pas été considérées.

II. Traitement des données de la MSA

Dans le cadre de ce manuscrit, le travail s'est focalisé sur les données administratives des non-salariés et sur les données ALD. Ce travail de thèse étant l'une des premières étapes d'un projet plus vaste, l'objectif est de se focaliser en premier lieu sur les non-salariés, une population plus stable car plus encline à conserver une même activité au cours du temps et ainsi, à avoir été exposée à des risques professionnels similaires sur de plus longues périodes. Il a aussi été choisi d'étudier les données ALD car elles sont richement pourvues d'informations sur les pathologies chroniques des cotisants au régime agricole dont les non-salariés et contrairement aux maladies professionnelles, elles ne sont pas explicitement liées aux activités professionnelles.

Par ailleurs, pour réaliser des analyses statistiques sur ces données médico-administratives complexes et non élaborées dans cette optique, il est important d'effectuer un nettoyage minutieux et rigoureux afin d'éviter des erreurs. Une restructuration des données est également nécessaire avant l'application de méthodes statistiques. Cet important travail de compréhension puis de traitement des données a été réalisé en étroite collaboration avec les différents départements de la MSA. Les décisions prises ont été validées en réunion de travail avec des représentants de la MSA et de la PPV de l'Anses.

a. Données administratives des non-salariés

Les données brutes des non-salariés de 2006 à 2016 comportent environ 30 variables, 6.7 millions d'observations pour une volumétrie légèrement inférieure à 1 giga-octet. Grâce au numéro d'identification individuel fourni par la MSA, il a été possible d'identifier 914 145 individus avec au moins un enregistrement au sein des données brutes et au cours de la période d'observation. Le nettoyage et la restructuration des données ont été réalisés en sept étapes majeures comprenant notamment des corrections d'incohérences et des décisions prises et validées par la CCMSA afin que les données soient adaptées pour répondre à nos objectifs (Figure 7).

Figure 7 : Étapes de gestion des données administratives des non-salariés de la MSA de 2006 à 2016

En effet, les données étant générées à une fréquence annuelle, la restructuration des données a été réalisée spécifiquement afin de résumer les informations des non-salariés sur leur période d'observation et ainsi de réduire la complexité des données (**étapes 3 et 5 à 7**).

Concernant l'**étape 1**, il est important de préciser que ces données comportent des individus ayant un numéro d'identification transitoire. En effet, comme il existe un certain délai entre la déclaration et l'enregistrement du numéro d'identification, un numéro d'identification transitoire a pu être attribué à certains individus et il est ensuite impossible de relier le numéro d'identification individuel transitoire au numéro définitif. Pour éviter de comptabiliser des individus plusieurs fois lors des analyses, l'exclusion de ces derniers est primordiale. Ceci a pu être possible grâce à l'ajout d'une variable par la MSA permettant de les identifier.

Concernant l'**étape 3**, il a été décidé de calculer l'âge à partir de l'année de naissance et des années d'observations afin de calculer un âge médian pour chaque individu sur la période d'observation. Ce calcul de l'âge a permis de prendre en compte la période d'observation qui a pu être différente pour chaque individu, puisqu'elle a pu varier de 1 an à un maximum de 11 ans. De plus, en accord avec la MSA, les individus âgés de plus de 100 ans ont été exclus des analyses car il a été estimé qu'ils étaient trop âgés pour être considérés comme des travailleurs « actifs ». Par ailleurs, cette décision a aussi permis d'écarter certains individus dont l'année de naissance a pu être une erreur de saisie (exemple : individu dont l'âge a été calculé à 127 ans).

Concernant l'**étape 6**, les surfaces renseignées pour la superficie d'exploitation ont été additionnées si plusieurs valeurs sont renseignées pour une même année d'observation, puis la médiane de ces valeurs a été calculée pour chaque individu sur sa période d'observation.

Concernant l'**étape 7**, l'assiette brute de revenus professionnels[7] correspond à une unique valeur annuelle renseignée par la MSA pour chaque non-salarié. Les individus avec plusieurs valeurs pour cette variable pour une même année d'observation ont donc été exclus. Puis, de la même façon qu'à l'étape précédente, la médiane des valeurs de ces revenus a été calculée pour chaque individu sur sa période d'observation.

Au cours de ces étapes, certaines variables ont aussi été écartées des données car elles n'ont pas été utilisées pour les analyses réalisées dans le cadre de ce travail (exemple : variable renseignant le code NAF), tandis que d'autres variables ont été créées afin d'être utilisées pour les analyses statistiques. Parmi les variables créées, une variable « nombre d'années

[7] Montant annuel qui donne une indication sur les revenus professionnels des individus et qui sert de base pour calculer les cotisations sociales. Le montant peut être négatif si le non-salarié est en déficit.

d'observations » a été ajoutée pour permettre un ajustement lors de la modélisation sur la période d'observation de chaque individu, pouvant varier de 1 à 11 années. En effet, les individus enregistrés dans les données administratives de la MSA n'ont pas systématiquement des enregistrements chaque année s'ils changent de statut (non-salarié à salarié) ou de régime (activité professionnelle pour laquelle ils sont inscrits au régime général de la sécurité sociale).

Enfin, les données administratives des non-salariés ainsi vérifiées, nettoyées et adaptées pour nos analyses, nous ont permis de réaliser les analyses statistiques sur 899 212 non-salariés.

b. Données des Affections de Longue Durée

Les données brutes des déclarations d'ALD de 2012 à 2016 comportent environ 15 variables, 40 millions d'observations pour une volumétrie égale à environ 3.5 giga-octets. Grâce au numéro d'identification individuel fourni par la MSA, il a été possible d'identifier 917 285 individus qu'ils soient ouvrants droit ou ayants droit, avec au moins une déclaration d'ALD au cours de la période d'observation au sein des données brutes.

Le nettoyage et la restructuration de ces données ont été réalisés en deux étapes majeures comprenant notamment des corrections d'incohérences concernant les codes des pathologies et l'exclusion des ayants droit (conjoints et autres membres de la famille) des analyses, représentant environ 7% des individus (Figure 8). De plus, au cours de ces étapes, les données ont été préparées de sorte à être fusionnées avec les données administratives des non-salariés. Les retraités représentant alors la majorité des individus avec au moins une déclaration d'ALD (≈ 80%), ils ont été exclus lors de l'étape ultérieure de fusion des données.

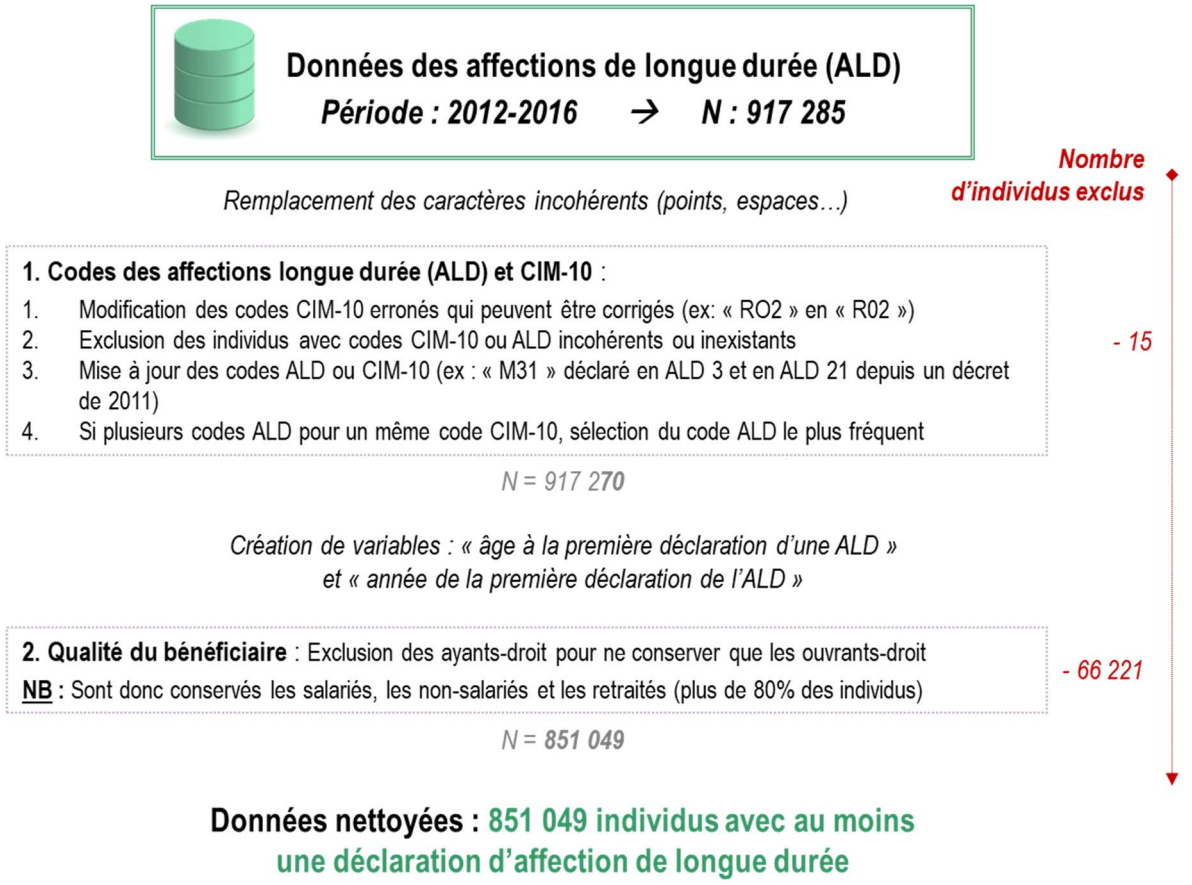


Figure 8 : Étapes de gestion des données des ALD de la MSA de 2012 à 2016

c. Fusion des données

La fusion des données a été effectuée afin de compléter les données administratives des non-salariés, considérées comme les données « source », avec les informations médico-administratives contenues dans les données ALD. Cette fusion des données a été rendue possible par la MSA qui a fourni un numéro d'identification individuel crypté et commun aux données transmises.

Ainsi, les données ALD ont été ajoutées uniquement pour les non-salariés ayant été enregistrés avec une déclaration d'ALD entre 2012 et 2016. Dans le cas où les individus n'ont pas de déclaration d'ALD, la valeur par défaut de « 0 » a été indiquée pour le code ALD et le code de pathologie CIM-10. De plus, dans une optique de mise en évidence de signaux émergents grâce à l'analyse non supervisée de chaque association entre activité professionnelle et pathologie, toutes les combinaisons possibles ont été conservées (Figure 9). Ainsi, si l'individu A est enregistré avec une seule activité professionnelle, mais qu'il a deux déclarations d'ALD, les données fusionnées sont composées de deux observations

correspondantes aux deux associations des activités professionnelles et des ALD. Les données ainsi fusionnées comprennent alors le même nombre d'individus que les données administratives des non-salariés, avec un nombre d'observations plus important dû aux différentes combinaisons possibles.

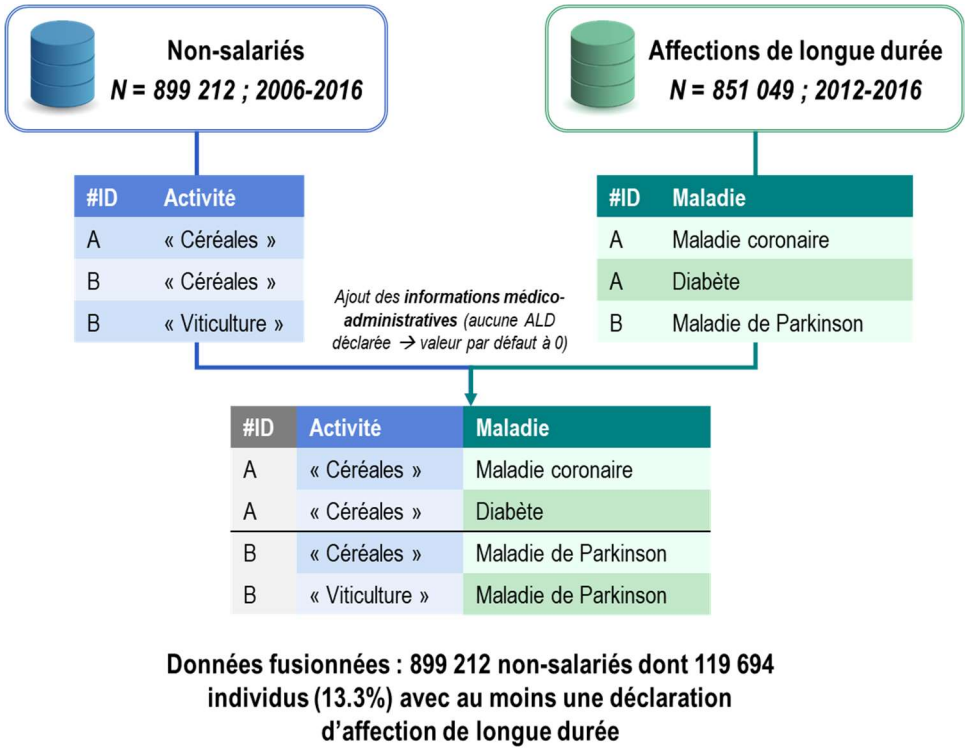


Figure 9 : Fusion des données administratives des non-salariés (2006-2016) et des données des ALD de la MSA (2012-2016)

Cette étape importante de gestion des données a permis d'identifier 119 694 individus ayant eu au moins une déclaration d'ALD au cours de la période d'observation.

Parmi ces derniers, **il est nécessaire d'écarter les individus ayant eu une ALD déclarée avant qu'on dispose d'informations sur les activités professionnelles qu'ils ont exercées** (Figure 10, Cas 4 et 5). En effet, pour ces individus, l'ALD déclarée peut difficilement être associée à l'activité professionnelle. Par exemple, si un individu a une déclaration d'ALD en 1998, comme nous ne disposons d'informations professionnelles qu'à partir de 2006, l'activité professionnelle exercée et déclarée à partir de 2006 a pu évoluer au cours du temps pour cet individu. En effet, même si nous disposons d'une information sur l'année d'installation de l'exploitation ou de l'entreprise du non-salarié, nous n'avons pas voulu considérer par défaut que l'individu a exercé depuis son installation la première activité qui nous est connue, pour éviter d'inférer sur des données non confirmées. Cependant, même si ces individus ne

sont désormais plus considérés comme « malades », une variable binaire (oui/non) a été créée afin d'indiquer s'ils ont eu une déclaration d'ALD antérieure à leur période d'observation. Cette variable sert alors à ajuster sur ce paramètre lors des analyses statistiques.

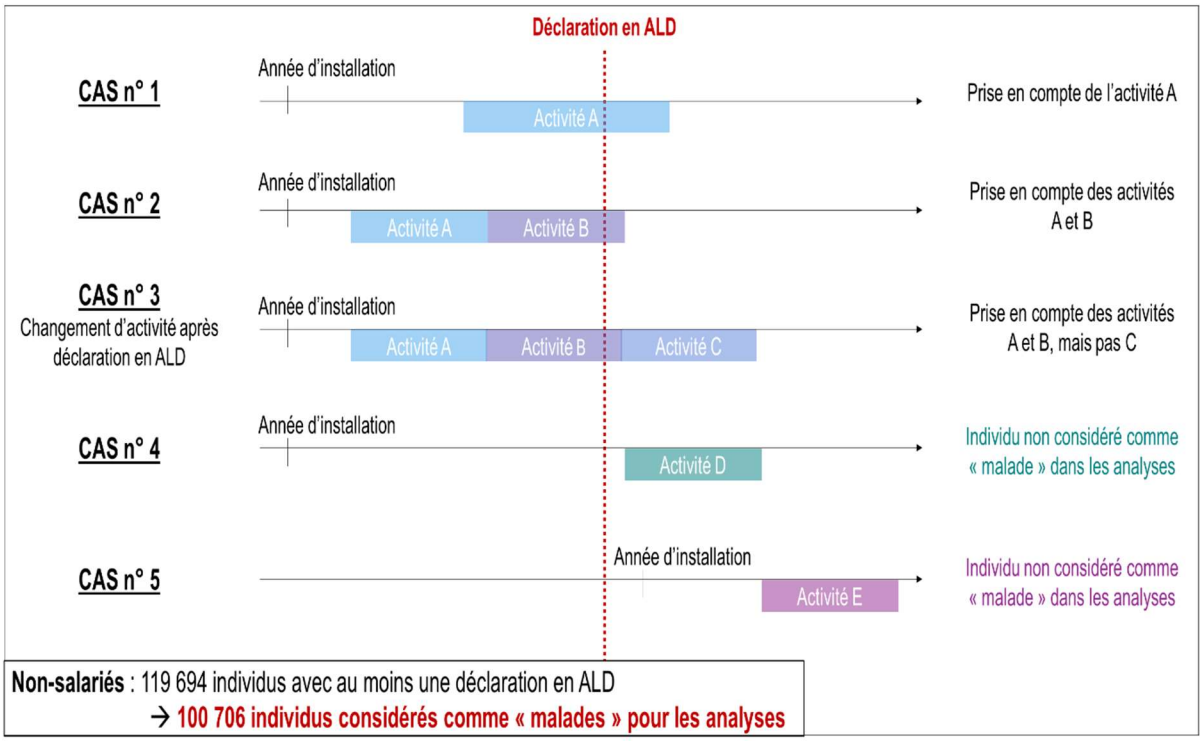


Figure 10 : Règles concernant la prise en compte des non-salariés de la MSA ayant eu une déclaration d'ALD durant la période d'observation de 2012 à 2016

Ainsi, dans le cadre de ce manuscrit, les différentes étapes de gestion des données de la MSA ont permis de réaliser nos analyses statistiques sur une population de non-salariés comptant 899 212 individus dont 100 706 ayant une déclaration d'ALD durant la période d'observation.

III. Analyses descriptives

a. Étude des caractéristiques des non-salariés

Caractéristiques administratives

Les analyses statistiques ont porté sur une population de non-salariés de 899 212 individus, observés en moyenne sur une période d'environ 7 ans avec une proportion d'environ 42% des individus observés sur l'ensemble de la période d'observation, c'est-à-dire, entre 2006 et 2016 (Tableau 2). Cette population est composée d'une majorité d'hommes (70.2%) et la moyenne d'âge est d'environ 50 ans avec une différence significative d'environ 5 ans entre les hommes et les femmes ($p < 2.2^{E-16}$) (Figure 11). Bien que les données administratives des non-salariés ne soient disponibles que de 2006 à 2016, environ 60% des individus se sont installés dans leur exploitation avant les années 2000. Cette population est constituée de près de 78% de chefs d'exploitations et 22% de cotisants solidaires, dont la majorité soit près de 70%, exercent exclusivement des activités relevant du régime agricole (régime maladie d'affiliation de type « exclusif »). Par ailleurs, pour les non-salariés étant mariés, soit environ 59% des individus, seuls 12% de leurs conjoints participent aux travaux de l'exploitation ou de l'entreprise, qu'ils aient le statut de conjoint collaborateur ou non (cf. modalités Tableau 2).

Concernant les activités professionnelles des non-salariés, les individus se concentrent majoritairement dans 5 activités, représentant près de 70% des observations (Tableau 3). Les principales activités sont les cultures céréalières et industrielles (23.0% des observations), les élevages de bovins laitiers (13.8%), les cultures et élevages non spécialisés (11.5%), la viticulture (10.6%) et les élevages de bovins viande (10.3%). Pour plus de 90% des non-salariés, seule une activité professionnelle est enregistrée au cours de leur période d'observation. Cependant, certains individus ont pu avoir changé de métier en ayant jusqu'à 4 (n = 110) ou 5 (n = 1) activités professionnelles différentes au cours de leur période d'observation.

En lien étroit avec leurs activités agricoles, les non-salariés ont déclaré en moyenne une assiette brute de revenus professionnels de 8 845€ sur la période d'observation, ce montant servant administrativement à calculer le montant de leurs cotisations. Parmi les non-salariés, seuls 4.5% d'entre eux ont déclaré une assiette médiane négative montrant qu'ils étaient en déficit sur la période d'observation (Tableau 4). On observe aussi une variation de cette assiette de revenus en fonction des activités professionnelles. En effet, sur la période d'observation, il semble que les non-salariés exerçant des activités dans le secteur équestre aient davantage déclaré des revenus montrant un déficit (Figure 12).

Tableau 2 : Caractéristiques principales des non-salariés de la MSA (2006-2016)

Caractéristiques	**Effectifs de non-salariés**[8]	**%**
Nombre d'individus	**899 212**	-
Nombre d'années d'observations par individu		
Moyenne	7.4	-
Sexe		
Hommes	631 560	70.2
Femmes	267 652	29.8
Moyenne d'âge (années)	50.2	-
Hommes	48.7	-
Femmes	53.8	-
Année d'installation (exploitation/entreprise)		
Médiane	1995	-
Minimum	1940	-
Maximum	2016	-
Statut		
Chefs d'exploitation ou d'entreprise	699 289	77.8
Cotisants solidaires	199 923	22.2
Régime maladie d'affiliation[9]		
Exclusif	625 139	69.5
Principal	42 077	4.7
Non-salarié non agricole	10 463	1.2
Salarié	16 959	1.9
Autres	204 574	22.8
Statut familial[10]		
Célibataire	350 942	39.0[8]
Marié	532 732	59.2
Veuf	37 940	4.2
Divorcé ou Séparé	48 366	5.4
Statut du conjoint		
Non participant aux travaux ou Inexistant	834 344	92.8
Participant aux travaux (hors conjoint collaborateur)	2 856	0.3
Conjoint collaborateur exclusif ou principal	53 520	6
Conjoint collaborateur secondaire[11]	8 492	0.9

[8] Dont chefs d'exploitation et cotisants solidaires.

[9] Régime maladie d'affiliation considéré : **exclusif** si le non-salarié exerce exclusivement des activités relevant du régime agricole ou **principal** si le non-salarié perçoit plus de la moitié du total de ses revenus en exerçant des activités relevant du régime agricole ; **non-salarié non agricole** si le non-salarié exerce une activité de type commerçant, artisan ou une profession libérale ; **salarié** si le non-salarié a une activité principale de salarié agricole ; ou **« autres »** si le non-salarié a une activité principale dans un autre régime (exemple : régime des fonctionnaires, régime des collectivités territoriales…).

[10] Pour le statut familial, un individu peut avoir changé de statut au cours de la période d'observation. Les pourcentages ont été calculés sur le dénominateur du nombre de non-salariés et ne sont pas cumulables à 100%.

[11] L'activité du conjoint collaborateur est considérée comme exclusive, principale ou secondaire en fonction de la quotité de temps de travail sur l'exploitation, faisant ainsi varier ses cotisations à la MSA.

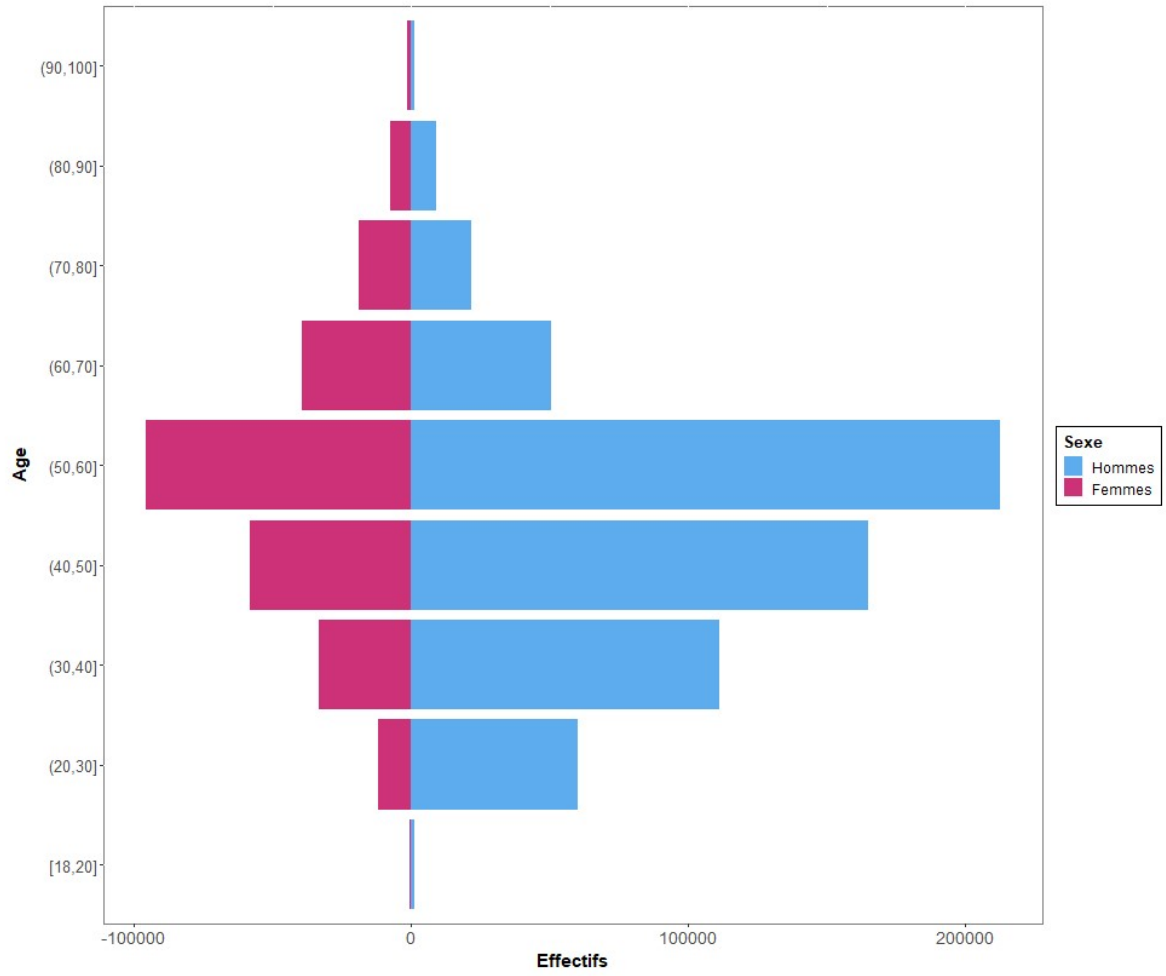


Figure 11 : Distribution de l'âge par sexe chez les non-salariés de la MSA (2006-2016)

Tableau 3 : Caractéristiques professionnelles des non-salariés de la MSA (2006-2016)

Activités professionnelles	**Effectifs de non-salariés**	**%**
Maraîchage, floriculture	36301*	3.7*
Arboriculture fruitière	20934	2.1
Pépinière	4882	0.4
Cultures céréalières et industrielles, « grandes cultures »	226607	23.0
Viticulture	104128	10.6
Sylviculture	1850	0.2
Autres cultures spécialisées	6150	0.6
Élevage bovins-lait	136146	13.8
Élevage bovins-viande	101711	10.3
Élevage bovins mixte	27226	2.8
Élevage ovins, caprins	43038	4.3
Élevage porcin	12004	1.2
Élevage de chevaux	16426	1.7
Autres élevages de gros animaux	2749	0.3
Élevage de volailles, lapins	22004	2.2
Autres élevages de petits animaux	17638	1.8
Entraînement, dressage, haras, clubs hippiques	13612	1.4
Conchyliculture	2967	0.3
Cultures et élevages non spécialisés, polyculture, poly-élevage	112983	11.5
Marais salants	828	0.1
Exploitation de bois	9658	1.0
Scieries fixes	691	0.1
Entreprises de travaux agricoles	13608	1.4
Entreprises de jardins, paysagistes, de reboisement	45254	4.6
Mandataires des sociétés ou caisses locales d'assurances mutuelles agricoles	936	0.1
Artisans ruraux	5369	0.5

**Un individu peut avoir différentes activités professionnelles enregistrées au cours de la période d'observation. Ainsi, la somme des individus dans chaque activité est plus élevée que le nombre d'individus étudiés. Par ailleurs, les pourcentages ont été calculés sur le dénominateur du nombre total d'observations d'activités professionnelles (n = 985 700).*

Tableau 4 : Caractéristiques économiques des non-salariés de la MSA (2006-2016)

Revenus professionnels	**Effectifs de non-salariés**	**%**
Assiette brute de revenus professionnels		
Moyenne	8 845	-
Médiane	5 484	-
Minimum	- 6 933 883	-
Maximum	3 913 218	-
Revenu de solidarité active (RSA, anciennement RMI)		
Individus ayant perçus au moins une fois le RSA au cours de la période d'observation	31 914	3.5
Chômage		
Individus ayant perçus au moins une fois le chômage au cours de la période d'observation	42 545	4.7

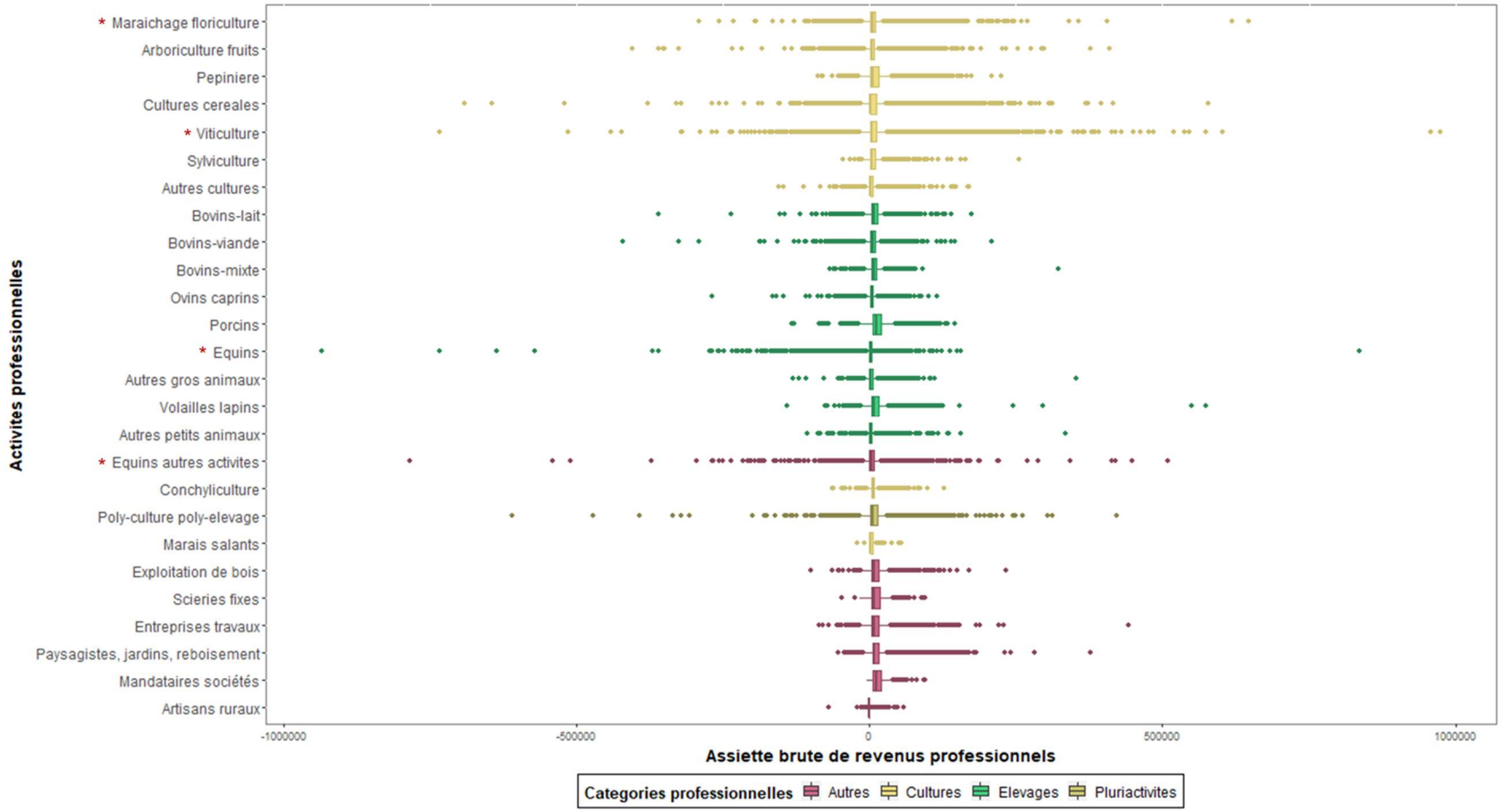


Figure 12 : Distribution de l'assiette brute de revenus professionnels (médiane annuelle en euros) chez les non-salariés de la MSA selon l'activité professionnelle exercée (2006-2016)[12]

[12] Symbole * correspondant aux activités professionnelles pour lesquelles 13 valeurs extrêmes (n = 9 individus) n'ont pas été représentées sur le graphique

Concernant les aides sociales, environ 3% et 5% respectivement des individus ont été bénéficiaires du revenu de solidarité active (RSA, ex RMI) ou de l'allocation chômage au moins une année sur l'ensemble de la période d'observation (Tableau 4). De plus, dans le secteur des entreprises de jardins, paysagistes et de reboisement, il est possible de remarquer que plus de 24% des individus ont bénéficié du chômage au moins une année au cours de la période d'observation. Quant au secteur de la conchyliculture, près de 13% des non-salariés ont bénéficié du RSA au moins une année au cours de la période d'observation (Tableau 5).

Tableau 5 : Répartition des non-salariés de la MSA ayant bénéficié d'aides sociales (chômage ou RSA) au moins une année au cours de la période d'observation (2006-2016)

Activités professionnelles	**Chômage**		**RSA**	
	Effectifs	%	Effectifs	%
Maraîchage, floriculture	2150	5.9	2443	6.7
Arboriculture fruitière	567	2.7	880	4.2
Pépinière	278	5.7	260	5.3
Cultures céréalières et industrielles, « grandes cultures »	3992	1.8	2858	1.3
Viticulture	1862	1.8	3693	3.5
Sylviculture	168	9.1	68	3.7
Autres cultures spécialisées	404	6.6	439	7.1
Élevage bovins-lait	5416	4	4806	3.5
Élevage bovins-viande	3570	3.5	4148	4.1
Élevage bovins mixte	550	2	1017	3.7
Élevage ovins, caprins	2188	5.1	3139	7.3
Élevage porcin	453	3.8	444	3.7
Élevage de chevaux	901	5.5	834	5.1
Autres élevages de gros animaux	188	6.8	164	6
Élevage de volailles. lapins	1446	6.6	1107	5
Autres élevages de petits animaux	1081	6.1	1028	5.8
Entraînement, dressage, haras, clubs hippiques	2243	16.5	1167	8.6
Conchyliculture	251	8.5	381	12.8
Cultures et élevages non spécialisés, polyculture, poly-élevage	3260	2.9	3543	3.1
Marais salants	86	10.4	59	7.1
Exploitation de bois	1406	14.6	575	6
Scieries fixes	25	3.6	12	1.7
Entreprises de travaux agricoles	1342	9.9	441	3.2
Entreprises de jardins, paysagistes, de reboisement	11006	24.3	1362	3
Mandataires des sociétés ou caisses locales d'assurances mutuelles agricoles	15	1.6	10	1.1
Artisans ruraux	188	3.5	8	0.1
Toutes activités confondues	31 914	-	42 545	-

Concernant les caractéristiques des exploitations ou des entreprises, la majorité des non-salariés, soit environ 61% des individus exercent leurs activités agricoles seuls (ou avec les membres de leur famille), c'est-à-dire, dans une exploitation individuelle (forme juridique). Ensuite, les sociétés de type EARL (entreprises agricoles à responsabilité limitée) et GAEC (Groupement agricole d'exploitation en commun) sont les deux formes juridiques les plus utilisées par les non-salariés, représentant respectivement environ 14 et 13% des individus.

Par ailleurs, la plupart des exploitations ou entreprises sont constituées en moyenne 1.3 salarié et d'une superficie moyenne de 35.3 hectares (Tableau 6). Selon les activités professionnelles, il est d'ailleurs possible d'observer une variation importante de la superficie de l'exploitation selon qu'il s'agit d'une activité de type culture, élevage ou d'autres types d'activités. Dans l'ensemble, les activités de type culture ou élevage ont en moyenne une superficie d'exploitation d'environ 40.6 hectares alors que les activités considérées comme « autres » telles que les exploitations de bois ou les artisans ruraux ont en moyenne une superficie d'exploitation ou d'entreprise d'environ 3.6 hectares (Figure 13).

Tableau 6 : Caractéristiques des exploitations des non-salariés de la MSA (2006-2016)

Caractéristiques des exploitations	**Effectifs de non-salariés**	**%**
Type d'exploitation (forme juridique)		
Exploitation ou entreprise individuelle	552 472	61.4
Groupement agricole d'exploitation en commun (GAEC)	119 637	13.3
Entreprise agricole à responsabilité limitée (EARL)	127 133	14.1
Société civile d'exploitation agricole (SCEA)	27 206	3.0
Groupement foncier agricole (GFA)	663	0.1
Société anonyme (SA) ou Société à responsabilité limitée (SARL)	23 138	2.6
Société de fait	6 081	0.7
Autre	4 472	0.5
Pluralité d'exploitation	38 410	4.3
Nombre de salariés		
Moyenne	1.3	-
Minimum	0	-
Maximum	1 608	-
Superficie de l'exploitation (ares)		
Moyenne	3 534	-
Médiane	1 920	
Minimum	0	-
Maximum	200 161	-

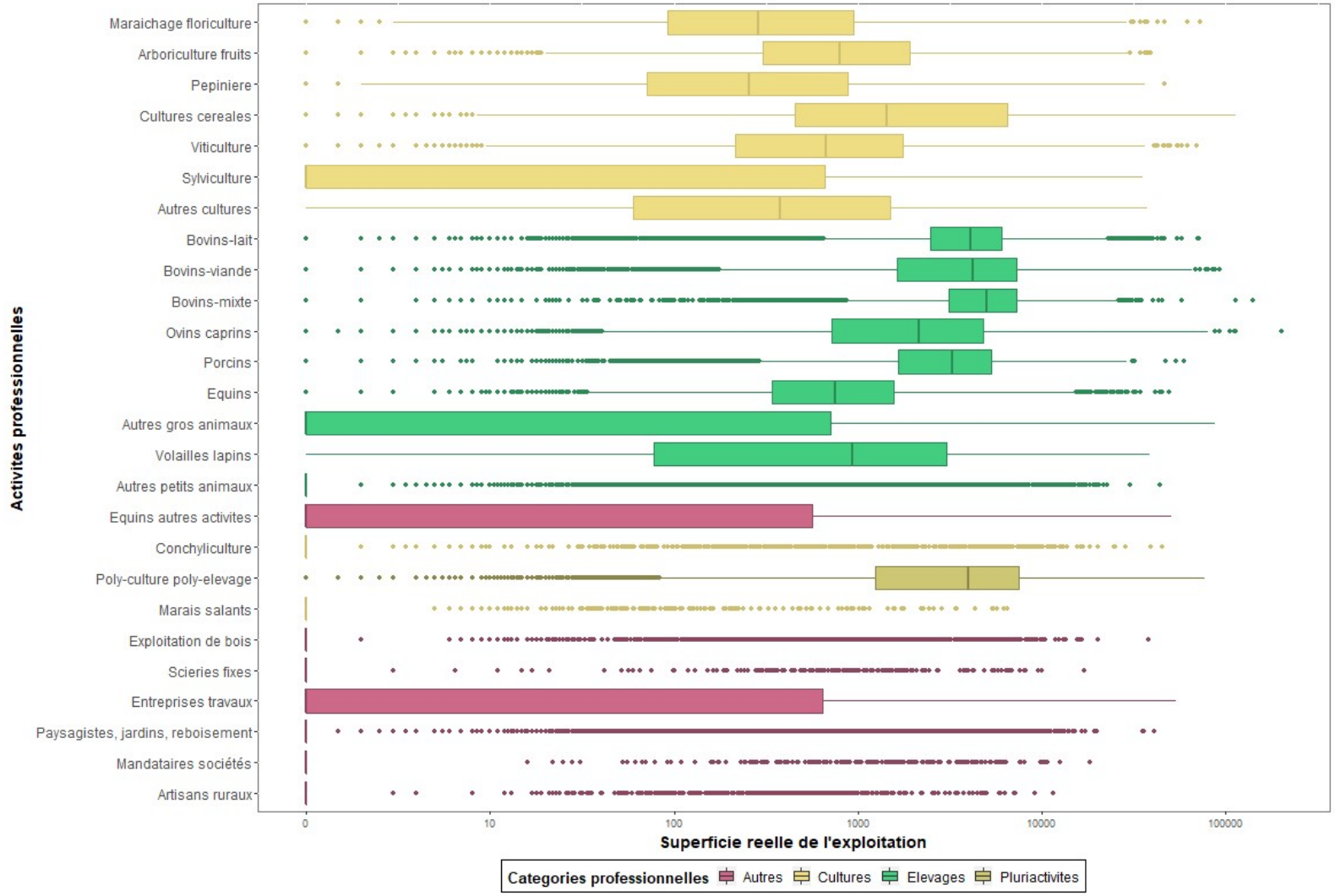


Figure 13 : Distribution de la superficie d'exploitation (médiane annuelle en ares) chez les non-salariés de la MSA selon l'activité professionnelle exercée (2006-2016)

La répartition des non-salariés de la MSA par région administrative montre qu'ils sont davantage concentrés en Nouvelle Aquitaine et en Occitanie qui font partie des régions où la Surface Agricole Utile était la plus importante ces dernières années (93) (Tableau 7).

Tableau 7 : Répartition des non-salariés de la MSA par région en France métropolitaine (2006-2016)

Régions administratives	**Effectifs de non-salariés**	**%**
Auvergne Rhône Alpes	96 499	10.7
Bourgogne Franche Comté	54 751	6.1
Bretagne	67 007	7.4
Centre Val de Loire	41 437	4.6
Grand Est	69 166	7.7
Corse	4 803	0.5
Hauts de France	40 288	4.5
Ile de France	12 797	1.4
Normandie	63 454	7.1
Nouvelle Aquitaine	147 514	16.4
Occitanie	138 648	15.4
Pays de la Loire	69 729	7.8
Provence Alpes Côte d'Azur	93 119	10.4
Total	899 212	100

Caractéristiques médico-administratives

Concernant les ALD, environ 11.2% (n = 100 706) des non-salariés ont au moins une déclaration d'ALD au cours de la période d'observation, c'est-à-dire, entre 2012 et 2016. Pour rappel, il s'agit du nombre de non-salariés que l'on peut considérer comme « malades » car on dispose d'informations sur les activités professionnelles antérieures à leur(s) déclaration(s) d'ALD. Les ALD pour lesquelles il y a le plus de déclarations sont les suivantes : « Tumeurs malignes » (19.9% des déclarations), « Diabètes » (19.4%), « Insuffisances cardiaques graves » (13.1%) et « Maladies coronaires » (10.2%) (Tableau 8). Près de 77% des non-salariés n'ont qu'une seule déclaration d'ALD mais pour environ 5% d'entre eux, il y a plus de deux déclarations d'ALD distinctes. Quant aux non-salariés qui ont une déclaration d'ALD avant la période d'observation, ils représentent environ 3% des individus.

Tableau 8 : Répartition des non-salariés étudiés de la MSA par ALD (2012-2016)

Code	Affections de longue durée (ALD)	Effectifs de non-salariés	%
1	Accident vasculaire cérébral (AVC) invalidant	5288*	4.1*
2	Insuffisances médullaires et autres cytopénies chroniques	284	0.2
3	Artériopathie chronique avec manifestations ischémiques	4942	3.8
4	Bilharziose compliquée	1	0.0
5	Insuffisance cardiaque grave, troubles du rythme graves, cardiopathies valvulaires graves, cardiopathies congénitales graves	16977	13.1
6	Maladies chroniques actives du foie et cirrhoses	1320	1.0
7	Déficit immunitaire primitif grave nécessitant un traitement prolongé et infection par le VIH (Virus de l'immunodéficience humaine)	241	0.2
8	Diabète de type 1 et diabète de type 2	25229	19.4
9	Forme grave des affections neurologiques et musculaires (dont myopathie), épilepsie grave	1927	1.5
10	Hémoglobinopathies, hémolyses, chroniques constitutionnelles et acquises sévères	18	0.0
11	Hémophilies et affections constitutionnelles de l'hémostase graves	257	0.2
12	Hypertension artérielle (HTA) sévère	5678	4.4
13	Maladie coronaire	13210	10.2
14	Insuffisance respiratoire chronique grave	3122	2.4
15	Maladie d'Alzheimer et autres démences	2550	2.0
16	Maladie de Parkinson	1686	1.3
17	Maladies métaboliques héréditaires nécessitant un traitement prolongé spécialisé	968	0.7
18	Mucoviscidose	9	0.0
19	Néphropathie chronique grave et syndrome néphrotique primitif	1676	1.3
20	Paraplégie	176	0.1
21	Périarthrite noueuse (PAN), lupus érythémateux aigu disséminé, sclérodermie généralisée évolutive (ScS)	969	0.8
22	Polyarthrite rhumatoïde évolutive grave	2601	2.0
23	Affections psychiatriques de longue durée	6438	5.0
24	Rectocolite hémorragique et maladie de Crohn évolutives	930	0.7
25	Sclérose en plaques	410	0.3
26	Scoliose structurale évolutive (dont l'angle est égal ou supérieur à 25 degrés) jusqu'à maturation rachidienne	87	0.1
27	Spondylarthrite ankylosante grave	1048	0.8
28	Suites de transplantation d'organe	197	0.1
29	Tuberculose active, Lèpre	52	0.0
30	Tumeur maligne, affection maligne du tissu lymphatique ou hématopoïétique	25934	19.9
31	Affection grave hors liste, nécessitant des soins continus d'une durée prévisible supérieure à 6 mois	5741	4.4
Nombre total d'individus ayant au moins une déclaration d'ALD au cours de la période d'observation		**100 706**	-

**Un individu peut avoir différentes ALD au cours de la période d'observation. Ainsi, la somme des individus dans chaque pathologie est plus élevée que le nombre d'individus étudiés. Les pourcentages ont été calculés sur le dénominateur du nombre de déclarations d'ALD au cours de la période d'observation (n = 129 966).*

Pour rappel, le sexe-ratio est de 2 hommes pour 1 femme dans la population de non-salariés étudiés. Or, selon les ALD, la proportion de femmes diffère de façon importante de la valeur moyenne de 30% (Figure 14). En effet, on observe trois ALD à prédominance féminine : « Alzheimer et autres démences », « Maladies auto-immunes » et « Scoliose structurale ». Ceci est au moins partiellement en lien avec une prévalence de ces affections notoirement plus élevée chez les femmes. Aussi, pour la maladie d'Alzheimer par exemple, le sexe féminin est le deuxième facteur de risque après l'âge, même après prise en compte de leur plus grande longévité, mais surtout pour les formes à début tardif (94). Effectivement, en regardant l'âge et le sexe par ALD, on observe un âge moyen plus élevé chez les femmes que chez les hommes pour l'ALD « Alzheimer et autres démences » (Tableau 9). Pour autant, d'autres facteurs explicatifs, en particulier des biais de sélection, peuvent expliquer ces variations de sexe-ratio.

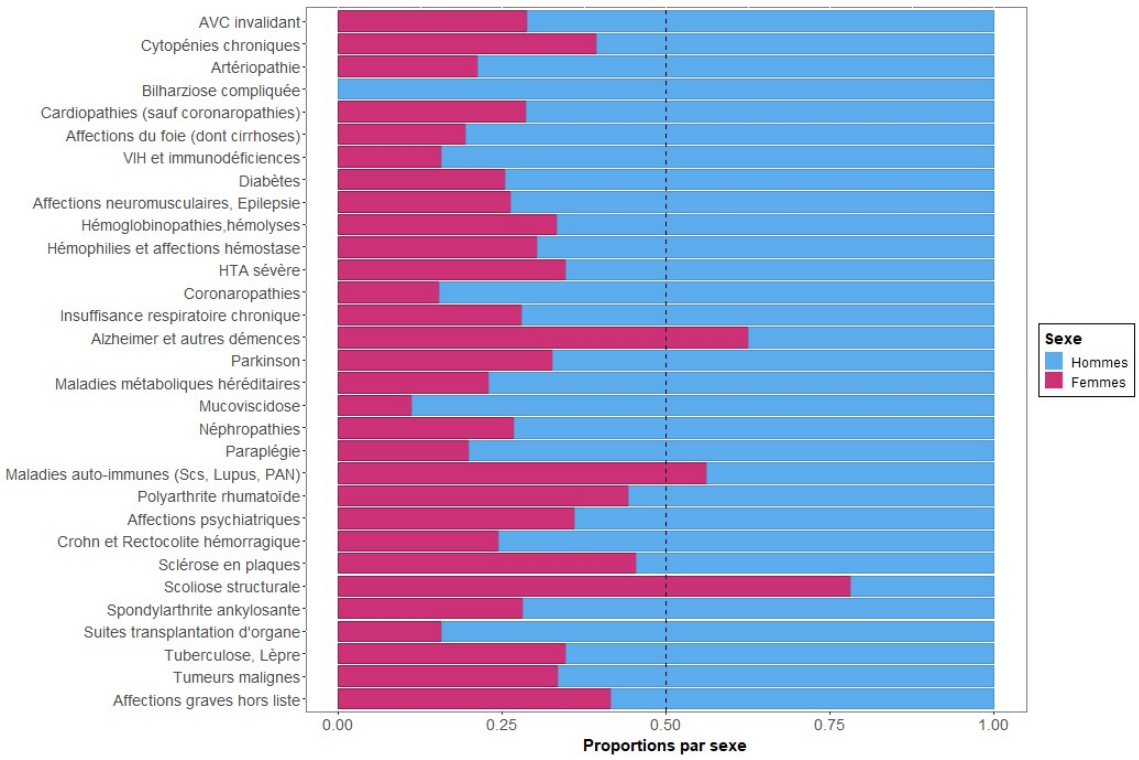


Figure 14 : Proportions d'adultes non-salariés étudiés de la MSA par sexe selon l'ALD déclarée (2012-2016)

Ainsi, pour certaines pathologies dont le sexe-ratio est habituellement proche de 1, la proportion relative observée de femmes et d'hommes est liée à un biais de sélection dû à la population exploitante active considérée. Ainsi pour la mucoviscidose, affection monogénique

portée par un chromosome non sexuel, le sexe-ratio est habituellement proche de 1. Sur les données des non-salariés adultes étudiés, il est très inférieur au sexe-ratio moyen constaté sur la population source dans sa globalité ce qui peut d'abord être expliqué par un nombre de cas très faible (n = 9). En effet, cette affection altérant fortement le pronostic respiratoire depuis l'enfance et le pronostic vital, rares sont les personnes se sachant atteintes de cette affection qui prennent la responsabilité d'une exploitation agricole. De plus, ces individus sont certainement dissuadés par les différentes difficultés liées au secteur agricole (problématique des emprunts, risques respiratoires et sociaux).

Ces éléments de discussion sont là pour rappeler que la base ALD des non-salariés n'est pas en soi une référence épidémiologique, notamment en ce qui concerne les affections survenant dans le jeune âge et en mesure d'influencer considérablement l'orientation professionnelle (contribution au *healthy worker effect*).

Tableau 9 : Répartition des non-salariés étudiés de la MSA et moyenne d'âge en année par ALD et par sexe (2012-2016)

Affections de longue durée (ALD)	**Hommes**			**Femmes**		
	Effectifs	%	Moyenne d'âge	Effectifs	%	Moyenne d'âge
Accident vasculaire cérébral invalidant	3762	5.3	62.5	1526	5.1	69.2
Insuffisances médullaires et autres cytopénies chroniques	172	0.2	63.4	112	0.4	67.1
Artériopathie chronique avec manifestations ischémiques	3888	5.5	63	1054	3.5	71.1
Bilharziose compliquée	1	0	-	0	0	-
Insuffisance cardiaque, troubles du rythme, cardiopathies valvulaires, cardiopathies congénitales graves	12114	17.1	65	4863	16.2	73.5
Maladies chroniques actives du foie et cirrhoses	1064	1.5	55.7	256	0.9	60.3
Déficit immunitaire primitif grave nécessitant un traitement prolongé et infection par le VIH	203	0.3	46.8	38	0.1	50
Diabète de type 1 et diabète de type 2	18795	26.6	59.3	6434	21.4	63.6
Forme grave des affections neurologiques et musculaires (dont myopathie), épilepsie grave	1419	2	55.8	508	1.7	59.9
Hémoglobinopathies, hémolyses, chroniques constitutionnelles et acquises sévères	12	0	49.8	6	0	63.2
Hémophilies et affections constitutionnelles de l'hémostase graves	179	0.3	54	78	0.3	57.2
Hypertension artérielle sévère	3706	5.2	60.7	1972	6.6	67.2
Maladie coronaire	11187	15.8	61.2	2023	6.7	70
Insuffisance respiratoire chronique grave	2247	3.2	62.4	875	2.9	66.1
Maladie d'Alzheimer et autres démences	956	1.4	77.1	1594	5.3	80.3
Maladie de Parkinson	1136	1.6	65.2	550	1.8	71.1
Maladies métaboliques héréditaires nécessitant un traitement prolongé spécialisé	746	1.1	53.1	222	0.7	56.9
Mucoviscidose	8	0	37.9	1	0	36
Néphropathie chronique grave et syndrome néphrotique primitif	1228	1.7	61.4	448	1.5	67.7
Paraplégie	141	0.2	53.1	35	0.1	59.8
Périarthrite noueuse, lupus érythémateux aigu disséminé, sclérodermie généralisée évolutive	424	0.6	61	545	1.8	64.2
Polyarthrite rhumatoïde évolutive grave	1451	2.1	56.4	1150	3.8	60.5
Affections psychiatriques de longue durée	4123	5.8	52.6	2315	7.7	59.1
Rectocolite hémorragique et maladie de Crohn évolutives	702	1	48	228	0.8	50
Sclérose en plaques	224	0.3	46.3	186	0.6	48.4
Scoliose structurale évolutive (dont l'angle est égal ou supérieur à 25 degrés) jusqu'à maturation rachidienne	19	0	50.8	68	0.2	63.9
Spondylarthrite ankylosante grave	753	1.1	48.1	295	1	50.9
Suites de transplantation d'organe	166	0.2	54	31	0.1	57.4
Tuberculose active, Lèpre	34	0	52.4	18	0.1	62.7
Tumeur maligne, affection maligne du tissu lymphatique ou hématopoïétique	17236	24.4	61.5	8698	29	62.3
Affection grave hors liste, nécessitant des soins continus d'une durée prévisible supérieure à 6 mois	3353	4.7	59.7	2388	8	67.6
Toutes ALD confondues	70 700	-	60.7	30 006	-	65.9

Par ailleurs, il existe également d'importantes variations de l'âge en fonction des ALD considérées. À titre d'exemple, il est possible d'observer un âge médian relativement bas pour l'ALD « Mucoviscidose » pour laquelle l'espérance de vie est plus basse que pour les autres ALD, et un âge médian relativement haut pour l'ALD « Alzheimer et autres démences » (Figure 15). Pour autant, l'âge médian de prise en charge en ALD de la mucoviscidose est relativement élevé (39 ans) et proche de la médiane de survie des patients atteints de cette affection en France (34 ans selon les données du Registre Français de la Mucoviscidose en 2015). Ceci est probablement révélateur de phénotypes différents, souvent moins sévères chez les hétérozygotes composites (95).

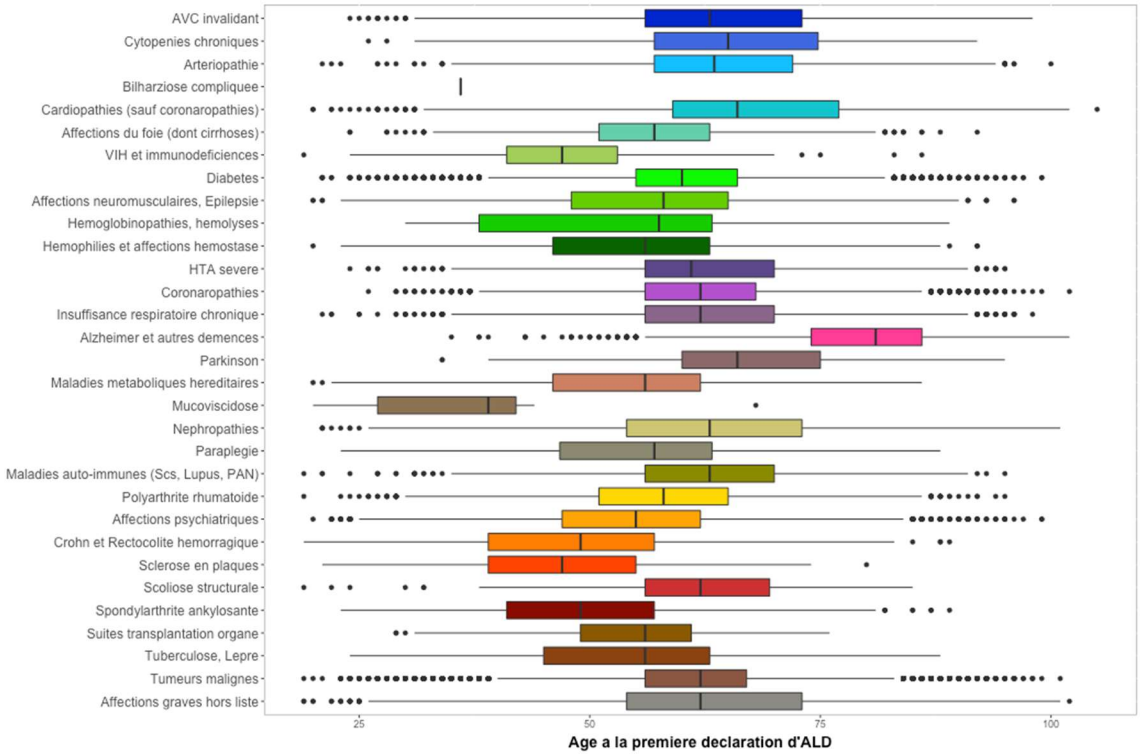


Figure 15 : Distribution de l'âge (en années) des non-salariés étudiés de la MSA, lors de la première déclaration d'ALD au régime agricole, selon l'ALD déclarée (2012-2016)

Pour autant, il est aussi possible que la date de notification de l'ALD ait subi des changements si l'individu a changé de régime de sécurité sociale au cours de son parcours professionnel. En effet, si dans un premier temps, un individu dépend du régime général lors de sa déclaration en ALD et qu'ensuite, il change de régime et dépend du régime agricole, alors, lors du transfert de ses données, la date de notification de l'ALD peut correspondre à sa date d'entrée dans le nouveau régime.

Par ailleurs, à chaque ALD déclarée est associé un code de pathologie de la CIM-10 permettant de décrire plus précisément la pathologie de chaque non-salarié. En moyenne, il y a environ 24 codes CIM-10 par ALD. Cependant, certaines ALD étant plus vastes dans leur dénomination tels que l'ALD « Tumeurs malignes », le nombre de codes CIM-10 peut varier d'un seul à 486 codes pour une même ALD (Figure 16).

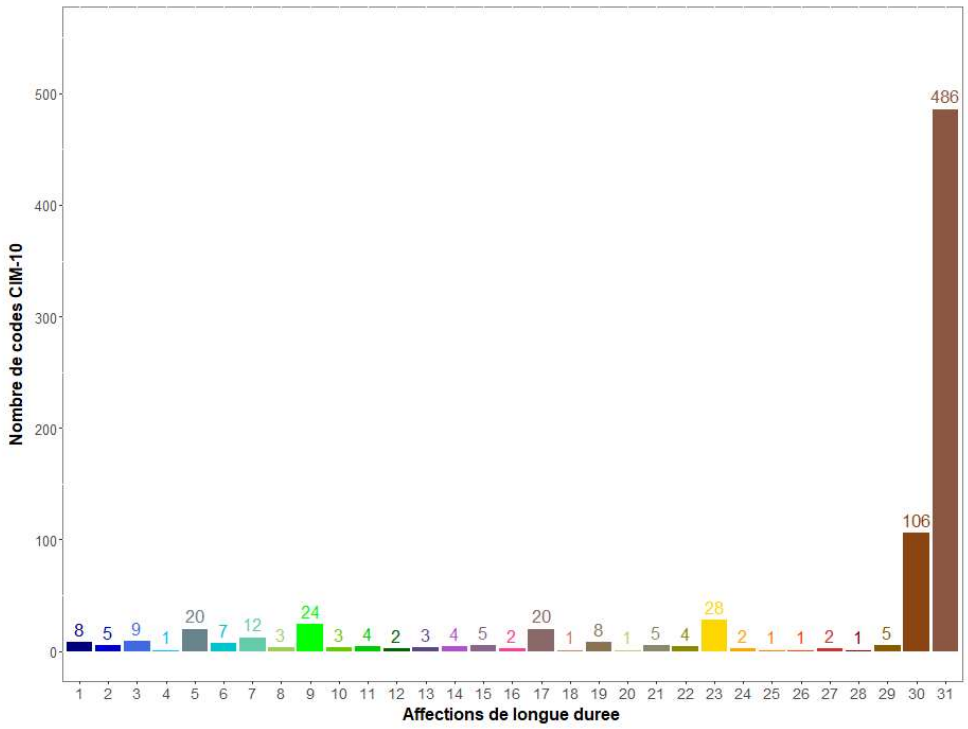


Figure 16 : Nombre de codes différents de pathologies de la CIM-10 par ALD chez les non-salariés étudiés de la MSA (2012-2016)

Par ailleurs, de manière similaire à la répartition des non-salariés par ALD, les trois grandes familles de codes CIM-10 pour lesquelles il y a le plus de déclarations sont les suivantes : « Appareil circulatoire » (32.7% des déclarations), « Maladies endocriniennes, nutritionnelles et métaboliques » (21.3%), et « Tumeurs malignes » (21.3%) (Figure 17).

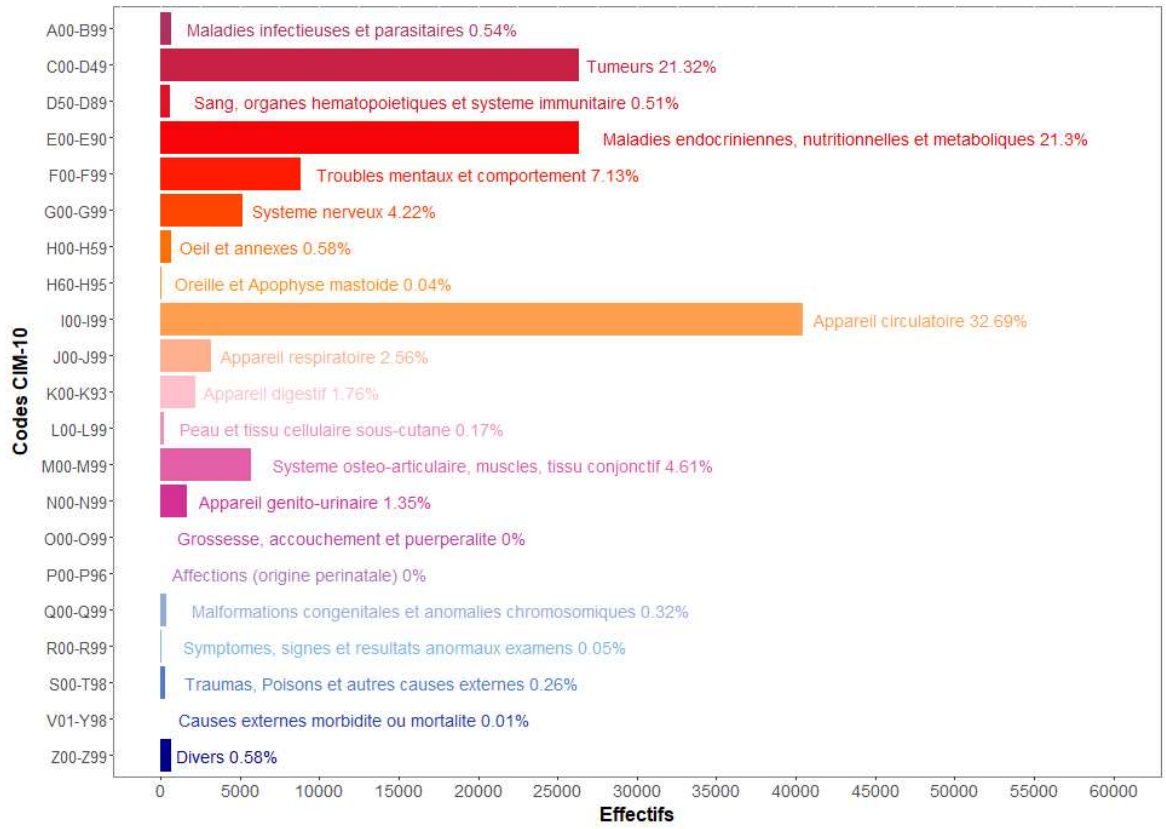


Figure 17 : Répartition des déclarations d'affections longue durée des non-salariés étudiés de la MSA par famille de codes de pathologies de la CIM-10 (2012-2016)

b. Étude des intensités de liaison entre les variables

Étant donné que la majorité des variables de l'étude sont de type catégoriel, la méthode qui a été choisie pour mesurer les liens entre les variables est celle du calcul du **V de Cramer. Il s'agit d'une mesure absolue de l'intensité de la liaison entre deux variables catégorielles qui présente l'avantage d'être indépendante du nombre de modalités et de l'effectif de la population** (92,96). L'intensité de liaison calculée est comprise entre 0 et 1 où plus la valeur est proche de 1, plus l'association entre les variables est importante. Compte tenu du fait que l'ensemble des variables doit être au format catégoriel pour ce calcul d'intensité de liaison, une **discrétisation a été réalisée pour les variables numériques via la méthode des quantiles**, permettant une répartition égale des effectifs dans chacune des classes. En ce qui concerne les seuils choisis pour évaluer ces mesures, l'intensité de liaison entre les variables est considérée comme « très bonne » si la valeur est supérieure à 0.20 et « faible » si la valeur est inférieure à 0.05. De plus, **si la mesure entre deux variables est supérieure à 0.40, on peut considérer que l'intensité de liaison entre ces deux variables est trop importante pour choisir de les inclure toutes les deux dans un même modèle statistique par la suite**. Dans ce cas, il a décidé de ne conserver qu'une seule des deux variables. Ces seuils ont été définis à titre indicatif et utilisés pour les analyses statistiques réalisées dans le cadre de ce travail.

Dans un premier temps, les mesures d'intensité de liaison ont été calculées entre chaque ALD déclarée et chaque activité professionnelle exercée par les non-salariés afin d'observer sans ajustement les liaisons entre les variables (Figure 18). Sur la période d'observation, on n'observe que de très faibles liaisons entre les variables d'intérêt, la liaison la plus importante étant de 0.04 entre l'ALD « Alzheimer et autres démences » et l'activité professionnelle de « Cultures de céréales ».

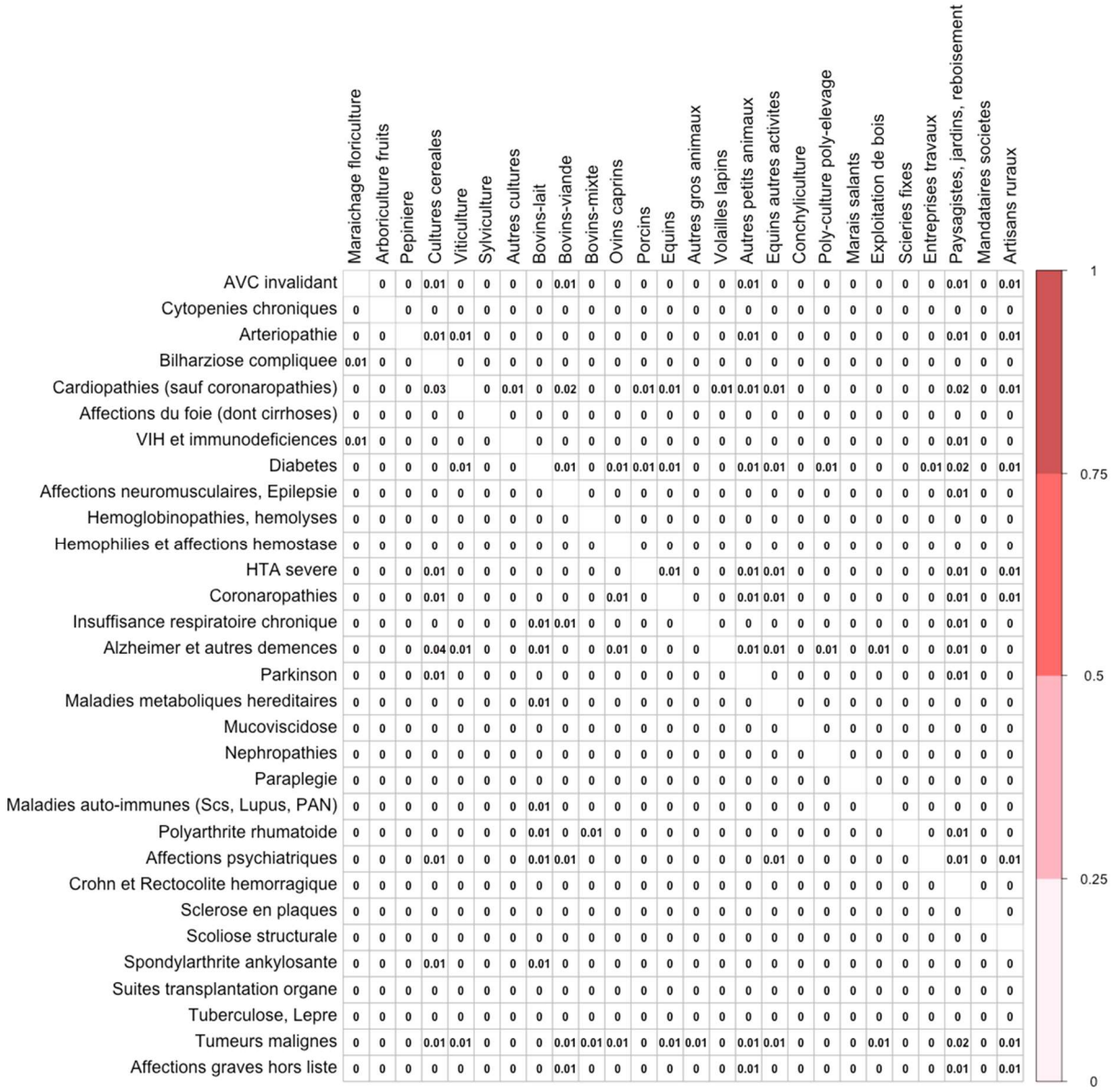


Figure 18 : Mesures des intensités de liaison via la méthode du V de Cramer entre chaque ALD déclarée (2012-2016) et chaque activité professionnelle exercée par les non-salariés de la MSA (2006-2016)

Puis, les intensités de liaison ont aussi été étudiées entre les ALD et les variables d'ajustement potentielles (Figure 19). Sur la période d'observation, on n'observe que des intensités de liaison moyennes, les plus élevées concernant la variable « Âge » et plusieurs ALD. Ces mesures nous indiquent qu'il est primordial de prendre en compte l'âge comme variable d'ajustement dans les analyses statistiques.

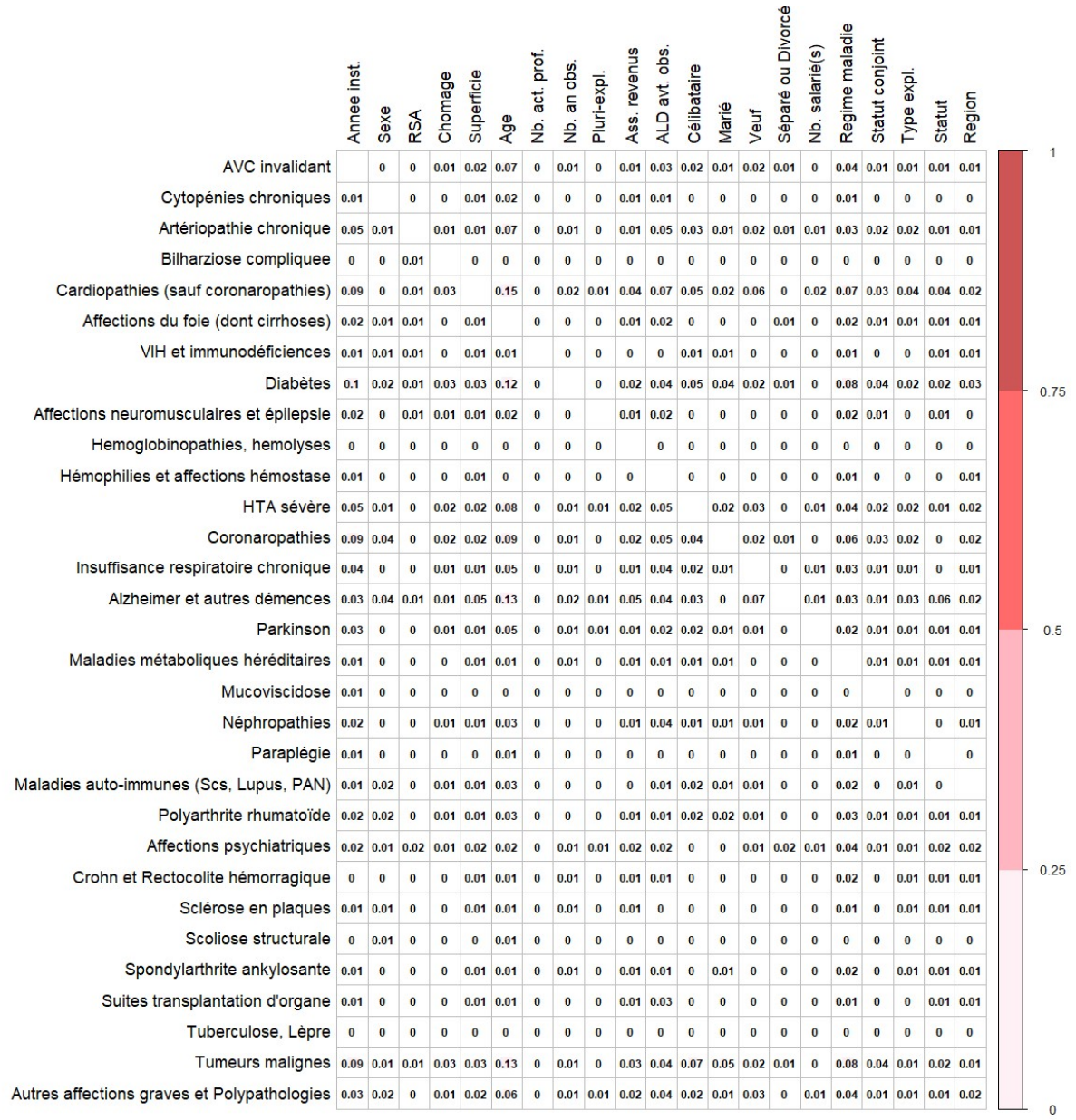


Figure 19 : Mesures des intensités de liaison via la méthode du V de Cramer entre chaque ALD déclarée (2012-2016) et chaque variable d'ajustement potentielle chez les non-salariés de la MSA (2006-2016)

Par ailleurs, avant d'utiliser des méthodes de régression, il est nécessaire d'étudier les liaisons entre les variables d'ajustement potentielles telles que l'âge, le sexe, la localisation géographique (région) ou encore le type d'exploitation (Figure 20). **Les mesures de V de Cramer sont supérieures à 0.40 pour les liaisons suivantes :** la superficie d'exploitation et le statut (chef d'exploitation ou cotisant solidaire) (0.63) ; l'âge et le statut familial « célibataire » (0.49) ; l'âge et le statut (chef d'exploitation ou cotisant solidaire) (0.44) ; le type d'exploitation et la modalité « pluri-exploitation » (0.75) ; les revenus et le statut (chef d'exploitation ou cotisant solidaire) (0.72) ; le statut familial « célibataire » et le statut familial « marié » (0.79) ; le régime maladie et le statut (chef d'exploitation ou cotisant solidaire) (0.62) ; le type d'exploitation et le statut (chef d'exploitation ou cotisant solidaire) (0.42). En conséquence, il a été choisi d'écarter les variables suivantes des analyses statistiques :

- le **statut (chef d'exploitation ou cotisant solidaire)** ayant une forte liaison avec cinq autres variables ;

- le **statut familial « célibataire »** ayant une forte liaison avec le statut familial « marié » et l'âge, cette dernière variable étant primordiale pour les analyses ;

- et la **modalité « pluri-exploitation »** ayant une forte liaison avec le type d'exploitation et n'ayant donc potentiellement aucun intérêt pour les analyses.

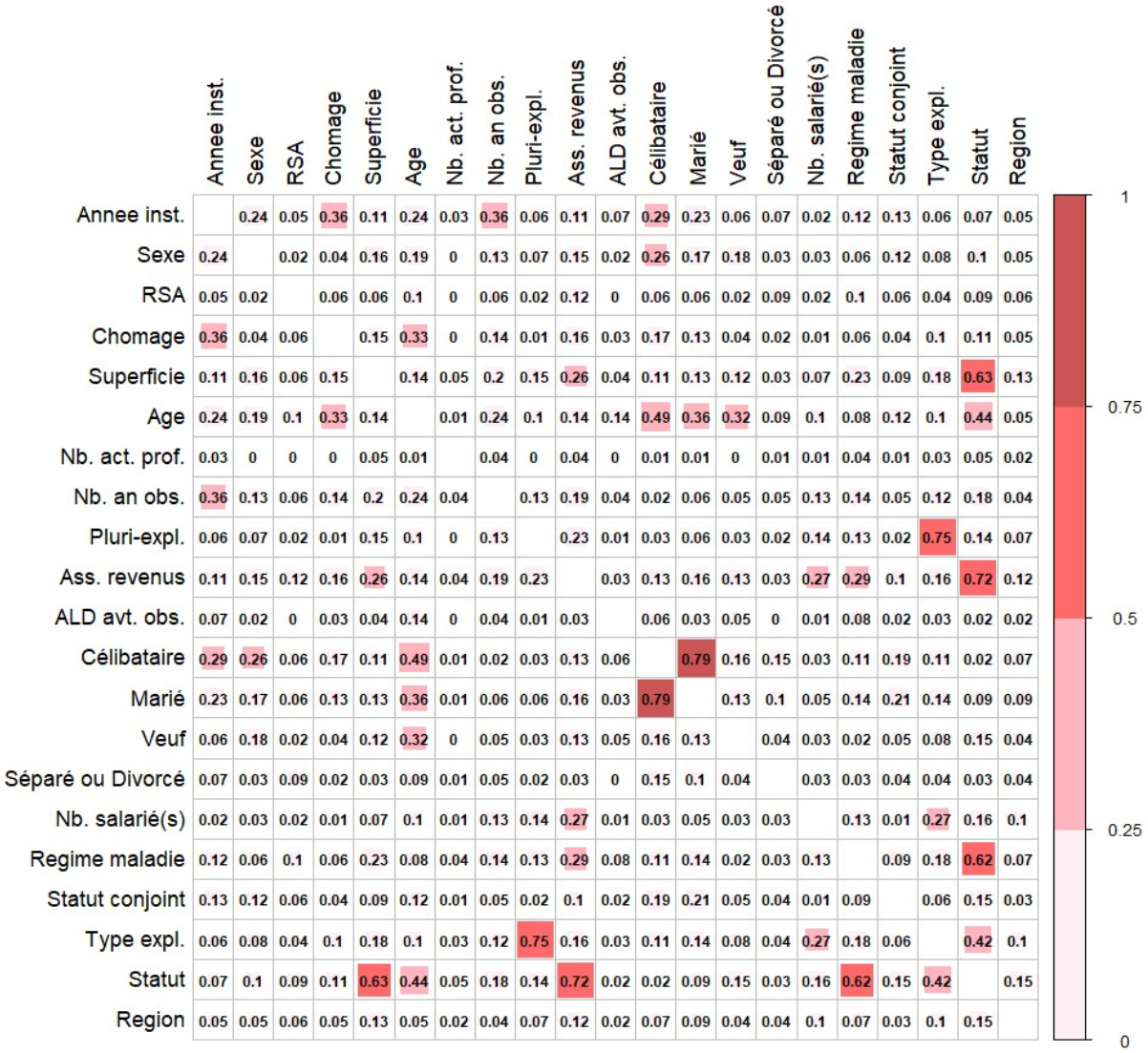


Figure 20 : Mesures des intensités de liaison via la méthode du V de Cramer entre chaque variable d'ajustement potentielle chez les non-salariés de la MSA (2006-2016)

Enfin, les intensités de liaison ont aussi été étudiées entre les activités professionnelles (Figure 21). Sur la période d'observation, on n'observe qu'une seule intensité de liaison considérée comme « très bonne », selon les seuils définis *a priori* ci-dessus, concernant les activités professionnelles « Culture céréalières » et « Elevage de bovins (lait) ». Ainsi, ces mesures nous ont permis de vérifier que les activités professionnelles n'ont donc pas d'intensité de liaison trop importantes entre elles, c'est-à-dire avec une valeur supérieure à 0.40.

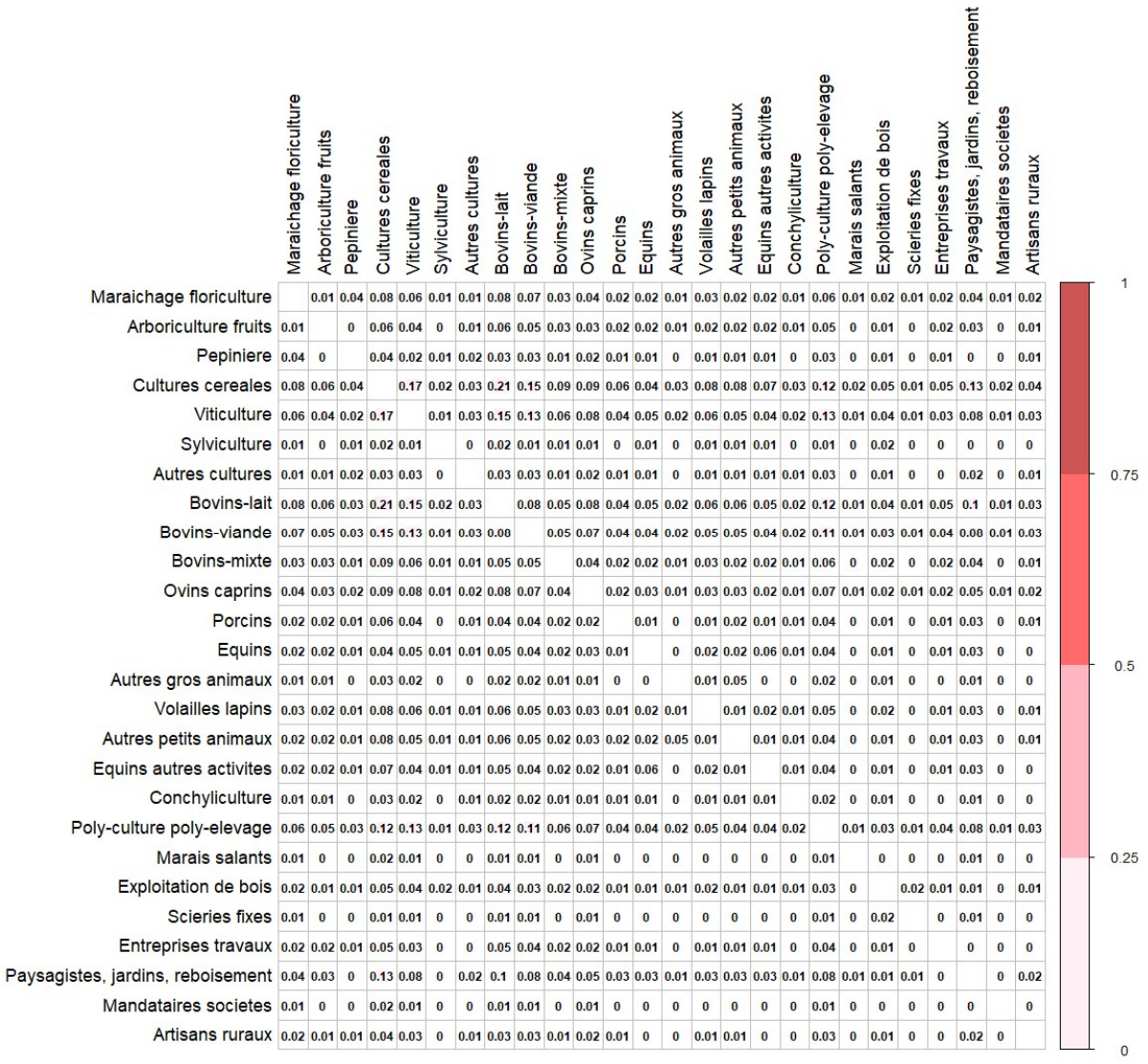


Figure 21 : Mesures des intensités de liaison via la méthode du V de Cramer entre chaque activité professionnelle chez les non-salariés de la MSA (2006-2016)

Résumé

Les différentes étapes de description, compréhension et traitement des données ont été réalisées en étroite collaboration avec les différents départements de la MSA concernés. Ces étapes ont permis de filtrer et structurer les données dans le but d'y appliquer des méthodes statistiques pour remplir les objectifs fixés. Ainsi, les différentes analyses statistiques ont été appliquées à **une population agricole de non-salariés, chefs d'exploitations ou chefs d'entreprises, constituée de 899 212 individus observés entre 2006 et 2016**. Puis, grâce à la fusion des données administratives et médico-administratives des non-salariés, rendue possible grâce à un identifiant unique pour chaque individu fourni par la MSA, il a été possible d'identifier **100 706 individus parmi les non-salariés avec au moins une déclaration d'ALD au cours de la période d'observation de 2012 à 2016 (période disponible pour les données ALD).**

Le jeu de données ainsi préparé a ensuite été analysé via des méthodes d'analyses descriptives usuelles afin d'en permettre une compréhension fine. Ces analyses ont montré que les données de la MSA sont très riches et qu'elles peuvent nous permettre de répondre à nos objectifs. Les données ainsi nettoyées sont prêtes pour l'**application des méthodes statistiques approfondies telles que la régression** qui permet de réaliser des études d'associations tout en tenant compte de variables d'ajustement telles que l'âge ou le sexe.

Dans cette optique, des mesures d'intensités de liaison entre les variables d'ajustement potentielles ont été réalisées afin d'écarter les variables ayant des intensités de liaison trop importante entre elles.

PARTIE 3

Choix et application d'une première méthodologie de modélisation aux données de la MSA

I. Méthodologie

a. Choix d'une approche méthodologique répondant à l'objectif

Dans le cadre de cette thèse, l'objectif est de mettre en évidence l'ensemble des associations entre des activités professionnelles dans le monde agricole et l'apparition de maladies chroniques déclarées en tant qu'ALD. Ainsi, il s'agit d'expliquer et de tester l'influence des activités professionnelles sur l'apparition d'une déclaration d'ALD au cours de la période d'observation. Pour ce faire, nous nous sommes d'abord employés à rechercher une première méthode de régression permettant de mettre en avant le degré de significativité de l'association, tout en en précisant son sens et sa force (« positive » ou « négative »).

Par ailleurs, au vu de la complexité des données, les variables d'intérêt ont dû être restructurées pour permettre l'application d'une première méthodologie. D'une part, les variables renseignant les activités professionnelles ont été considérées en tant que variables binaires pour la prise en compte des activités multiples pour un même individu sur la période d'observation. D'autre part, pour permettre l'analyse indépendante de chaque ALD, la variable renseignant la déclaration des ALD a été scindée en autant de variables binaires que d'ALD étudiées (variables « absence/présence » d'une déclaration d'ALD). Le choix du type de méthodologie dépendant alors de la nature de la variable à expliquer, et les variables ALD étant de nature binaires, nous nous sommes orientés vers la régression logistique, qui fait partie des méthodes classiques de modélisation en épidémiologie (97,98).

b. La régression logistique

La régression logistique est une méthode ancienne qui fut anticipée en 1838 par Pierre-François Verhulst et développée pour des applications biologiques à partir de 1944 par Joseph Berkson (99). Cette méthode est donc depuis longtemps un grand classique dans la classification et de ce fait, dans la pratique quotidienne de la plupart des statisticiens notamment en médecine et en épidémiologie. D'une part, elle permet de traiter des variables à prédire de type binaire avec des variables explicatives pouvant être quantitatives ou qualitatives, et ainsi contrôler en partie les biais de confusion potentiels. D'autre part, les résultats obtenus sous forme d'odds ratios (OR) sont très explicites car ils permettent de mesurer la force de l'association entre les variables explicatives et la variable à expliquer. La régression logistique s'est donc imposée comme une méthode de référence grâce à ces nombreux atouts : fiabilité, généralité et interprétabilité. Cette méthode offre alors un bon compromis entre performance du modèle et pouvoir explicatif (92,97,100).

Dans le cadre de cette méthode, on cherche à modéliser la probabilité de survenue d'un événement (Y) en fonction de variables explicatives (X_1, X_2, …, X_k où k est le nombre de variables), notée « Y | X_1, X_2, …, X_k ». La survenue de l'événement est alors caractérisée par une variable conditionnelle aléatoire suivant une loi de Bernoulli de paramètre : P = P (Y = 1 | X_1, X_2, …, X_k) = E (Y | X_1, X_2, …, X_k) où P est une probabilité comprise entre 0 et 1 et E, l'espérance mathématique. La loi de Bernoulli (appelée aussi épreuve de Bernoulli) correspond à une expérience aléatoire qui n'admet que deux issues différentes, « le succès ou l'échec ». L'espérance mathématique pour la loi de Bernoulli est alors égale à la probabilité de survenue de l'événement et sa valeur doit impérativement être comprise entre 0 et 1. Par ailleurs, cette espérance peut être exprimée comme une combinaison linéaire de variables explicatives si on utilise la transformation « Logit » qui est définie comme le logarithme népérien de la cote : Logit (P) = Ln ($\frac{P}{1-P}$).

Le modèle de régression logistique peut ainsi s'écrire (92,97,101) **:**

$$\textbf{Logit [E (Y | } X_1, X_2, \ldots X_k) \textbf{] = Logit (P) = } \beta_0 + \beta_1 X_1 + \ldots + \beta_k X_k$$

Grâce à cette méthode, il est ainsi possible d'obtenir des coefficients de régression β_k qui, une fois transformés, peuvent être interprétés comme des rapports de cotes ou odds ratios (OR). En effet, si la variable explicative X_j est binaire et indique l'exposition à un facteur donné, β_j représente le logarithme de l'OR du risque de survenue de l'événement étudié chez les exposés au facteur X_j par rapport aux non-exposés, ajusté sur les autres variables explicatives présentes dans le modèle. On obtient alors l'OR mesurant l'association entre X_j et Y en prenant l'exponentielle du coefficient β_j : OR = exp^{β_j}. Pour mesurer la précision de l'estimation de cet OR, il est possible d'en calculer l'intervalle de confiance à 95% (IC95%) grâce à la formule suivante : IC95% = $exp^{\beta_j \pm 1{,}96 * sd\,(\beta_j)}$ (*sd : standard deviation*). L'estimation de l'IC à 95 % des OR permet également de déterminer si la variable explicative correspondante est associée ou non à l'événement étudié, ajustée sur les autres facteurs qui sont dans le modèle : si l'IC à 95 % ne contient pas la valeur 1, l'association est significative pour un risque α de 5 % ; dans le cas contraire, on ne met pas en évidence d'association. Ensuite, si pour la variable explicative, la valeur de l'OR est supérieure ou inférieure à 1, cela indique le sens de l'association, c'est-à-dire, si le risque de survenue de l'événement est plus ou moins important en fonction de la valeur prise par la variable explicative, ajusté sur les autres variables ajoutées au modèle (92,97).

Par ailleurs, il est aussi possible de prendre en compte de possibles interactions entre les variables explicatives au sein du modèle de régression logistique. Cependant, dans le cadre de ce travail, les variables explicatives ont été considérées comme indépendantes afin de ne pas ajouter de complexité au modèle comprenant déjà une quantité importante de variables.

c. Application de la régression logistique

Dans le cadre de cette méthode et de ce travail, on cherche donc à modéliser la probabilité de survenue d'un événement, à savoir une déclaration d'ALD au cours de la période d'observation en fonction de variables explicatives correspondant aux activités professionnelles exercées par les non-salariés et à leurs caractéristiques individuelles telles que l'âge ou le sexe.

Le modèle de régression logistique ainsi appliqué aux données de la MSA et à notre problématique est le suivant :

ALD_i ~ Activité professionnelle $_j$ + sélection de variables explicatives

Où i correspond à chaque ALD (n = 31) et j correspond à chaque activité professionnelle (n = 26).

Par ailleurs, toutes les ALD n'ont pas pu être étudiées dans le cadre de ce travail, à cause des effectifs trop faibles de ces dernières. En effet, il y a moins d'une vingtaine, soit 0.002% des non-salariés ayant une déclaration pour les trois ALD suivantes au cours de la période d'observation : « Bilharziose compliquée » (n = 1), « Hémoglobinopathies, hémolyses, chroniques constitutionnelles et acquises sévères » (n = 18) et « Mucoviscidose » (n = 9). Ainsi, ces proportions trop faibles d'individus n'ont pas permis l'étude des associations entre ces pathologies et les activités professionnelles. Ceci n'est pas en soi une limite car les deux dernières affections correspondent pour l'essentiel à des pathologies constitutionnelles, donc sans lien avec la profession de l'individu.

En ce qui concerne les activités professionnelles, chacune d'entre elles a été testée indépendamment, c'est-à-dire, en n'incluant qu'une seule activité professionnelle par modèle, afin d'éviter les problèmes de multi-colinéarité. En effet, ces problèmes peuvent augmenter la variance des coefficients de régression, les rendant instables et difficiles à interpréter (102). De ce fait, des analyses calculant les facteurs d'inflation de la variance (VIF) ont été réalisées afin d'estimer les évolutions de la variance des coefficients en fonction de leurs relations linéaires. Ces analyses ont effectivement montré qu'il existait bien des problèmes de multi-colinéarité lorsque toutes les activités professionnelles étaient ajoutées aux modèles.

Ainsi, les associations entre 28 ALD et 26 activités professionnelles ont été testées, tout en étant ajustées sur d'autres variables explicatives, représentant un total de 728 modèles de régression logistique.

Pour rappel, en ce qui concerne les variables explicatives, certaines ont été écartées des analyses du fait de liaisons trop importantes entre elles (cf. Partie 2 III. b.). Toutes les variables explicatives ont été simplifiées de sorte à ne conserver qu'une seule observation par individu pour l'ensemble de la période d'observation. De plus, pour inclure les variables quantitatives (âge, assiette de revenus, superficie d'exploitation) dans un modèle de régression logistique, il faut supposer que ces variables ont une relation linéaire avec le Logit. Dans le cadre de ce travail, il a été décidé de choisir l'alternative simple qui consiste à transformer ces variables quantitatives en variables catégorielles. Une discrétisation a alors été réalisée via la méthode des quantiles habituellement utilisée, qui permet une répartition égale des effectifs dans chacune des classes. Le nombre de classes a été fixé à dix (ou moins si certaines classes étaient identiques) pour les trois variables (utilisation des déciles), de sorte à conserver assez de précision pour restreindre la perte d'informations, tout en limitant le nombre de modalités de ces variables, pour éviter une perte de puissance due à la complexité du modèle (92,96,101). Cependant, la perte d'informations ou de puissance pouvant être liée à ces transformations de variables est relativement limitée au vu de l'effectif d'individus considérés.

L'ensemble de ces variables ainsi que le format dans lequel elles ont été utilisées pour chaque étape de la modélisation ont été détaillés dans le Tableau 10.

Tableau 10 : Description et traitement des variables utilisées au cours des analyses statistiques effectuées sur la population des non-salariés de la MSA (2006-2016)

Variables explicatives	Description de la variable et modifications préalables	Format de la variable lors de la sélection du modèle[1]	Format de la variable lors de la modélisation « finale »
Activités professionnelles (« Risque »)	Activité professionnelle majoritaire en termes de temps de travail accordé, considérée comme un « risque » prépondérant et permettant le calcul de cotisations à la MSA (variable catégorielle à 26 modalités) Modification : variable transformée en 26 variables binaires	26 variables binaires imposées lors de la sélection du modèle	26 variables binaires testées une à une pour la recherche d'associations entre chaque ALD et chaque activité professionnelle
Année d'installation	Année d'installation correspondant à l'année de la première affiliation du chef d'exploitation Modifications : - Si plusieurs années d'installations renseignées pour un même individu, année la plus ancienne retenue - Transformée via la méthode des quantiles[2] en variable catégorielle à 9 modalités	9 variables binaires	1 variable catégorielle à 9 modalités
Sexe	Sexe (variable binaire)	1 variable binaire	1 variable binaire
RSA	Bénéficiaire du revenu de solidarité active Modification : Si l'individu a été au moins une fois bénéficiaire du RSA durant la période d'observation, il lui est attribué la valeur « 1 », sinon la valeur « 0 »	1 variable binaire	1 variable binaire
Chômage	Personne bénéficiaire de l'allocation chômage (variable binaire) Modification : Si l'individu a été au moins une fois bénéficiaire de l'allocation chômage durant la période d'observation, il lui est attribué la valeur « 1 », sinon la valeur « 0 »	1 variable binaire	1 variable binaire
Superficie d'exploitation	Superficie annuelle de l'exploitation exprimée en ares (variable quantitative) Modifications : - Calcul de la médiane sur l'ensemble de la période d'observation - Transformée via la méthode des quantiles[2] en variable catégorielle à 9 modalités	9 variables binaires	1 variable catégorielle à 9 modalités

Variables explicatives	Description de la variable et modifications préalables	Format de la variable lors de la sélection du modèle[1]	Format de la variable lors de la modélisation « finale »
Âge	Âge calculé à partir de l'année de naissance de l'individu et de l'année de la dernière déclaration de l'individu dans les données administratives (variable quantitative) Modifications : - Calcul de l'âge médian sur l'ensemble de la période d'observation - Transformée via la méthode des quantiles[2] en variable catégorielle à 10 modalités	10 variables binaires	1 variable catégorielle à 10 modalités
Nombre d'activités professionnelles	Nombre d'activités professionnelles différentes exercées au cours de la période d'observation, pouvant varier de 1 à 5 (variable quantitative) Modification : Transformée en variable catégorielle à 2 modalités (« 1 à 2 activités professionnelles », « 3 à 5 activités professionnelles »)	1 variable binaire	1 variable binaire
Nombre d'années d'observations	Nombre d'années d'observations de chaque individu, pouvant varier de 1 à 11 années (variable quantitative) Modification : Transformée en variable catégorielle à 3 modalités (« 1 à 3 années », « 4 à 7 années », « 8 à 11 années »)	3 variables binaires	1 variable catégorielle à 3 modalités
Assiette brute de revenus professionnels, « Revenus »	Montant d'assiette brute annuelle de revenus professionnels (variable quantitative) Modifications : - Calcul de la médiane sur l'ensemble de la période d'observation - Transformée via la méthode des quantiles[2] en variable catégorielle à 10 modalités	10 variables binaires	1 variable catégorielle à 10 modalités
ALD avant observation	Variable indiquant si l'individu a eu une déclaration d'ALD avant sa période d'observation (variable binaire)	1 variable binaire	1 variable binaire
Statut familial	Statut familial du chef d'exploitation (« marié », « célibataire », « veuf », « séparé ou divorcé »), pouvant évoluer au cours de la période d'observation Modifications : Transformée en 4 variables binaires afin de conserver l'ensemble des statuts pour les individus	4 variables binaires → **3** → 1 variable écartée des analyses (« célibataire ») suite aux études d'intensité de liaison entre les variables	4 variables binaires → **3** → 1 variable écartée des analyses (« célibataire ») suite aux études d'intensité de liaison entre les variables
Nombre de salariés	Nombre de salariés employés sur l'exploitation Modifications : - Calcul du nombre de salariés médian sur l'ensemble de la période d'observation - Transformée en variable catégorielle à 2 modalités	1 variable binaire	1 variable binaire

Variables explicatives	Description de la variable et modifications préalables	Format de la variable lors de la sélection du modèle[1]	Format de la variable lors de la modélisation « finale »
Régime maladie	Régime maladie d'affiliation du chef d'exploitation (variable catégorielle à 5 modalités : « exploitant à titre exclusif », « exploitant à titre principal », « non-salarié non agricole », « régime des salariés agricoles », « autres régimes »)	5 variables binaires	1 variable catégorielle à 5 modalités
Statut du conjoint	Statut du conjoint du chef d'exploitation (sauf si le conjoint lui-même est chef d'exploitation) (variable catégorielle à 4 modalités : « conjoint non participant aux travaux ou inexistant », « conjoint participant aux travaux MSA hors conjoints collaborateurs », « conjoint collaborateur exclusif ou principal », « conjoint collaborateur secondaire »)	4 variables binaires	1 variable catégorielle à 4 modalités
Type d'exploitation	Type de l'exploitation, forme juridique de l'exploitation (9 modalités : « exploitant individuel », « membre d'une société anonyme ou à responsabilité limité (SA/SARL) », « membre d'un groupement foncier agricole (GFA) », …)	9 variables binaires	1 variable catégorielle à 9 modalités
Pluri-exploitation	Indiquant si l'exploitant possède plusieurs exploitations, créée à partir de la variable « Type d'exploitation »	→ Variable écartée des analyses suite aux études d'intensité de liaison entre les variables	→ Variable écartée des analyses suite aux études d'intensité de liaison entre les variables
Statut (chef d'exploitation ou cotisant solidaire)	Distinction entre chef d'exploitation, conjoint collaborateur, aide familial, cotisant solidaire. Dans le cadre de l'étude, les conjoints collaborateurs et les aides familiaux ont été exclus (variable catégorielle à 2 modalités)	→ Variable écartée des analyses suite aux études d'intensité de liaison entre les variables	→ Variable écartée des analyses suite aux études d'intensité de liaison entre les variables
Région	Région administrative, variable crée à partir de la variable département (variable catégorielle à 13 modalités) Modification : Région la plus renseignée en termes d'occurrence pour chaque individu au cours de la période d'observation	13 variables binaires	1 variable catégorielle à 13 modalités

[1]Sélection de variables : toutes les variables ont été mises au format binaire de sorte à réduire le temps de calcul nécessaire à cette étape.
[2]Méthode des quantiles : méthode qui permet de diviser les valeurs des variables en intervalles contenant le même nombre d'individus. Par ailleurs, en l'appliquant sur ces données, il est possible d'obtenir des intervalles identiques pour certaines variables. Certains intervalles peuvent alors contenir plus d'individus.

Ensuite, pour chaque ALD étudiée, une sélection de variables explicatives a été réalisée afin de rechercher le modèle logistique le plus parcimonieux. Pour cela, les données ont d'abord été séparées en deux sous-ensembles pour la construction du modèle : un échantillon d'entraînement (70% des observations) et un échantillon de validation (30%). Pour réaliser le partitionnement, le jeu de données a été stratifié de sorte que la distribution des observations soit uniforme dans chaque échantillon concernant la variable à expliquer (ici, la pathologie déclarée en ALD). Cette méthode permet de limiter l'*overfitting*, c'est-à-dire, le sur-ajustement d'un modèle lorsqu'il est construit sur l'ensemble des observations. En effet, si l'ajustement est trop fort, le modèle capture alors le bruit des données actuelles et risque de ne pas être adapté s'il est appliqué à de nouvelles données. Le principe consiste alors à construire le modèle sur l'échantillon d'entraînement et ensuite à le tester sur l'échantillon de validation au moyen de mesures de robustesse (96,100,103). La sélection de variables a été réalisée pas à pas via l'inclusion (sélection ascendante) et l'exclusion (sélection descendante) de chacune des variables explicatives une à une, selon le critère BIC (*Bayesian information criterion*). Ce critère, basé sur la vraisemblance, est fréquemment utilisé pour comparer des modèles construits sur le même échantillon. Comparé au critère d'information d'Akaike (AIC), il permet davantage de privilégier les modèles les plus parcimonieux et donc plus robustes, puisqu'il pénalise davantage le nombre de variables présentes dans le modèle. Ainsi, le modèle ayant le critère BIC le plus faible est le modèle le plus adapté aux données et est donc celui qui a été choisi (96,104,105). Pour cette étape de sélection de variables, celles renseignant les activités professionnelles ont été imposées de sorte que l'algorithme puisse tenir compte de ces variables à chaque inclusion ou exclusion des autres variables explicatives. **Enfin, comme ce processus a été réalisé pour chaque ALD, 28 sélections de variables ont été réalisées, permettant d'obtenir 28 modèles différents.**

Par la suite, ces modèles ont été évalués sur les échantillons de validation en calculant différentes mesures de robustesse. La première mesure de robustesse calculée est **l'aire sous la courbe ROC (« Receiver Operating Characteristic », AUC)** qui donne une indication sur le pouvoir discriminant d'un modèle. En effet, la courbe ROC permet d'observer la proportion de vrais positifs en ordonnée en fonction de la proportion de faux positifs en abscisse. Plus l'aire sous la courbe ROC est importante, plus le critère AUC est proche de la valeur 1, et meilleure est la qualité du modèle (92,100). Pour une valeur de 0.5, la performance du modèle est identique à celle d'un classificateur aléatoire. Pour compléter cette mesure, les mesures de robustesse suivantes ont aussi été calculées pour chaque modèle :

- la sensibilité, qui est la probabilité que le modèle classe l'individu comme malade si la maladie est présente ;

- la spécificité, qui est la probabilité que le modèle classe l'individu comme non malade si la maladie est absente ;

- et le F1 score, qui combine la sensibilité et la valeur prédictive positive et indique la probabilité que la maladie soit présente lorsque l'individu est classé comme malade par le modèle (100).

Pour chaque ALD, le modèle, ainsi choisi et testé sur l'échantillon de validation, a été utilisé sur l'ensemble des données. Grâce à cette méthode, il a été possible de récupérer les **coefficients de régression β_k pour chaque association d'une ALD avec une activité professionnelle afin de les utiliser pour calculer les odds ratios et leurs intervalles de confiance à 95%.**

Par ailleurs, les p-valeurs de chaque association ont aussi été récupérées. Cependant, comme le nombre de tests d'association réalisés est important (n = 728), les chances de découvrir des associations non pertinentes sont augmentées. **Pour corriger ce biais lié aux tests multiples**, il est donc essentiel d'appliquer une correction aux p-valeurs afin de contrôler les fausses découvertes potentielles (FDR, « false discovery rate »), d'où l'utilisation de la procédure de Benjamini-Hochberg. Cette procédure, inventée en 1995 et connue pour sa fiabilité, a été largement utilisée depuis lors pour le contrôle du FDR (106).

Enfin, certaines associations statistiquement significatives ($p\text{-valeur}_{corrigée} < 0.05$) ont été mises en évidence alors que les effectifs concernés par ces associations étaient inférieurs à 3 individus. De façon similaire au seuil minimal d'individus concernés par des signaux mis en évidence dans le champ de la pharmacovigilance (107), il a été décidé d'écarter les associations statistiquement significatives pour lesquelles les effectifs concernés sont inférieurs à 3 individus, considérant que le « signal » mis en évidence est trop instable et serait de toute façon non prioritaire dans le cadre de la mise en place d'un système de vigilance.

II. Résultats

Des associations entre ALD et activités professionnelles chez les non-salariés de la MSA au cours de la période d'observation ont pu être mises en évidence via la régression logistique. La Figure 22 représente les p-valeurs de chaque association testée, corrigées par la procédure de Benjamini-Hochberg et transformées à l'aide du logarithme décimal ($-\log_{10}$). Les p-valeurs sont ainsi représentée comme en génétique sous forme d'un graphique de type « *Manhattan plot* » permettant de visualiser les p-valeurs les plus faibles. Cette première représentation a permis de montrer que les associations statistiquement significatives ($p\text{-valeur}_{corrigée} < 0.05$) peuvent correspondre à des p-valeurs très petites notamment pour deux associations :

- Diabètes et Entreprise de travaux agricoles : $p\text{-valeur}_{corrigée} = 8.03^{E-17}$;
- Diabètes et Eleveur Bovins-lait : $p\text{-valeur}_{corrigée} = 1.43^{E-16}$.

Par ailleurs, ce type de représentation graphique ne permettant pas de visualiser simplement le sens des associations, une *heatmap* a été réalisée pour les associations statistiquement significatives (Figure 23). Les analyses menées révèlent au total 54 associations statistiquement significatives entre une activité professionnelle et une ALD, dont 35 associations avec un risque plus faible de déclarations des ALD en question et 19 associations avec un risque plus élevé de déclarations des ALD concernées.

Les associations statistiquement significatives avec un risque plus élevé de déclaration d'ALD concernent différents types d'ALD (Figure 24). En effet, sept concernent des ALD cardiovasculaires, trois concernent l'ALD « Diabètes », deux concernent l'ALD « VIH et Immunodéficiences », deux concernent l'ALD « Affections du foie (dont cirrhoses) », une concerne l'ALD « Alzheimer et autres démences », une concerne l'ALD « Maladies auto-immunes (ScS, Lupus, PAN) », une concerne l'ALD « Affections psychiatriques », une concerne l'ALD « Tumeurs malignes » et enfin, une concerne l'ALD « Paraplégie ». Parmi ces associations, six d'entre elles ont des OR supérieurs à 1.5, dont trois avec des OR dépassant la valeur 2.

Au centre de la heatmap, il est possible de remarquer 23 associations statistiquement significatives concernant pour la plupart des éleveurs pratiquant par les activités professionnelles suivantes : « Bovins-lait », « Bovins-viande », « Ovins-caprins », « Porcins », « Equins », « Autres gros animaux », « Volailles, lapins », « Autres petits animaux », « Equins autres ». La majorité des associations concerne la plupart de ces activités professionnelles avec des risques moins élevés de déclarations d'ALD « Diabètes » et d'ALD « HTA sévère » comparativement au reste de la population des non-salariés. A titre d'exemple, on trouve chez

éleveurs de bovins (lait) un risque de déclaration moins élevé d'ALD « Diabètes » (OR = 0.84 [0.81 ;0.87] ; p-valeur$_{corrigée}$ = 1.43^{E-16} ; n = 3 459). De plus, il est possible d'observer quatre associations concernées par des risques de déclarations d'ALD moins élevés et pour lesquelles les OR sont très faibles. A titre d'exemple, c'est le cas pour l'ALD « VIH et Immunodéficiences » chez les éleveurs de bovins (lait) (OR = 0.42 [0.26 ;0.68] ; p-valeur$_{corrigée}$ = 0.002 ; n = 19) et les éleveurs de bovins (viande) (OR = 0.31 [0.16 ;0.59] ; p-valeur$_{corrigée}$ = 0.002 ; n = 10). Par ailleurs, seules trois associations concernent des risques plus élevés de déclarations d'ALD, à savoir : l'ALD « Maladies auto-immunes (ScS, Lupus, PAN) » chez les éleveurs de bovins (lait) (OR = 1.45 [1.24 ;1.69] ; p-valeur$_{corrigée}$ = 6.89^{E-5} ; n = 211) ; et les ALD « Affections psychiatriques » (OR = 1.18 [1.10 ;1.27] ; p-valeur$_{corrigée}$ = 4.11^{E-5} ; n = 993) et « Cardiopathies (sauf coronaropathies) » (OR = 1.08 [1.03 ;1.13] ; p-valeur$_{corrigée}$ = 0.02 ; n = 2 438) chez les éleveurs de bovins (viande).

A l'inverse, certaines activités professionnelles apparaissent avec davantage de risque de déclaration d'ALD, même après ajustement sur les variables disponibles. On trouve notamment **chez les viticulteurs** des risques plus élevés de déclarations pour les ALD suivantes : « Affections du foie (dont cirrhoses) » (OR = 1.31 [1.12 ;1.53] ; p-valeur$_{corrigée}$ = 0.01 ; n = 92), « Diabètes » (OR = 1.16 [1.12 ;1.21] ; p-valeur$_{corrigée}$ = 4.02^{E-13} ; n = 3 402) et « Tumeurs » (OR = 1.11 [1.07 ;1.16] ; p-valeur$_{corrigée}$ = 1.05^{E-6} ; n = 3 375). De la même façon, **chez les maraîchers et les floriculteurs**, comparativement aux autres non-salariés, on trouve des risques plus importants de déclarations pour les ALD suivantes : « Affections du foie (dont cirrhoses) » (OR = 1.40 [1.12 ;1.76] ; p-valeur$_{corrigée}$ = 0.04 ; n = 82), « VIH et immunodéficiences » (OR = 2.88 [1.93 ;4.31] ; p-valeur$_{corrigée}$ = 2.92^{E-6} ; n = 27) et « HTA sévère » (OR = 1.18 [1.04 ;1.34] ; p-valeur$_{corrigée}$ = 0.04 ; n = 262). On trouve également des risques plus élevés de déclarations pour l'ALD « Coronaropathies » chez les individus travaillant dans les exploitations de bois (OR = 1.36 [1.16 ;1.59] ; p-valeur$_{corrigée}$ = 0.002 ; n = 159) et dans le secteur des « paysagistes, jardins, reboisement » (OR = 1.19 [1.07 ;1.31] ; p-valeur$_{corrigée}$ = 0.008 ; n = 434).

Par ailleurs, en ce qui concerne l'**ALD « Insuffisance respiratoire chronique »**, on ne trouve que trois associations concernant des risques moins élevés de déclarations pour cette ALD par rapport au reste de la population des non-salariés. Trois activités professionnelles sont concernées : l'arboriculture fruitière (OR = 0.62 [0.47 ;0.83] ; p-valeur$_{corrigée}$ = 0.02 ; n = 48), le secteur des cultures céréalières et industrielles (OR = 0.88 [0.80 ;0.96] ; p-valeur$_{corrigée}$ = 0.04 ; n = 827) et la viticulture (OR = 0.81 [0.7 ;0.92] ; p-valeur$_{corrigée}$ = 0.02 ; n = 290).

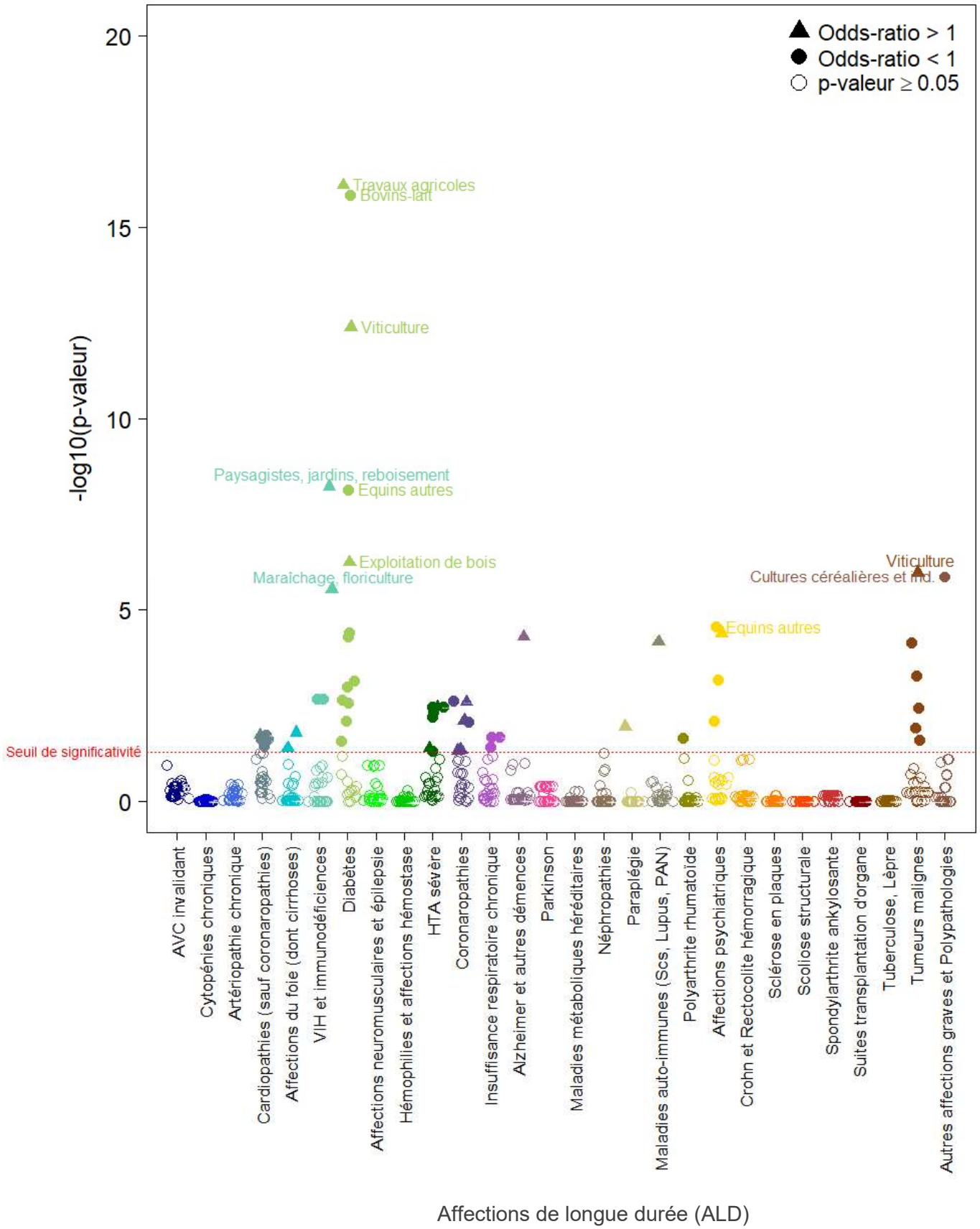


Figure 22 : Représentation graphique des p-valeurs, corrigées par la procédure de Benjamini-Hochberg, obtenues via la **régression logistique** entre chaque combinaison d'ALD et d'activité professionnelle en ajustant sur d'autres paramètres, chez les **non-salariés** de la MSA (2006-2016)

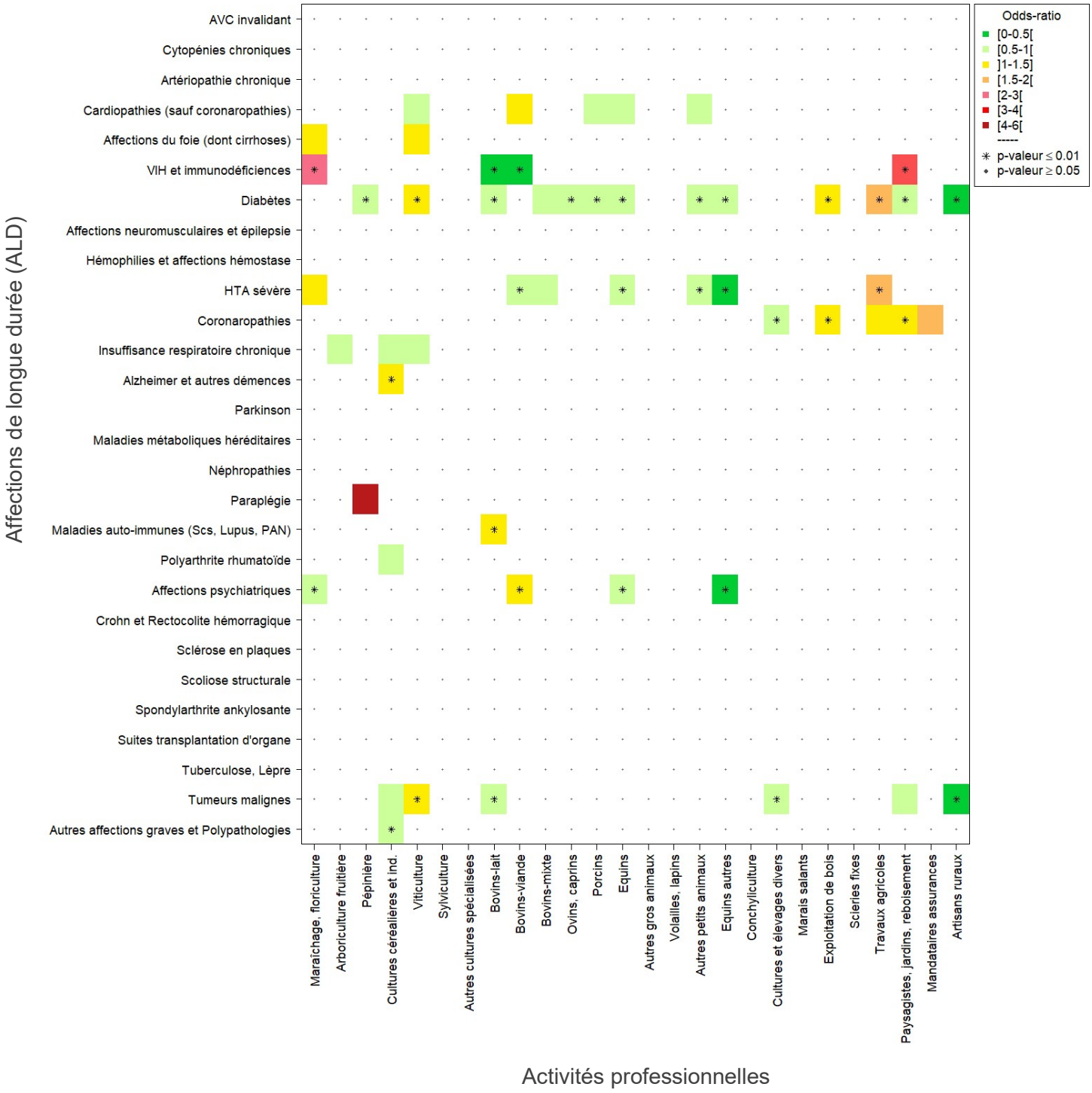


Figure 23 : Représentation graphique des odds ratios obtenus via la régression logistique entre chaque combinaison d'ALD et d'activité professionnelle en ajustant sur d'autres paramètres, chez les non-salariés de la MSA (2006-2016)

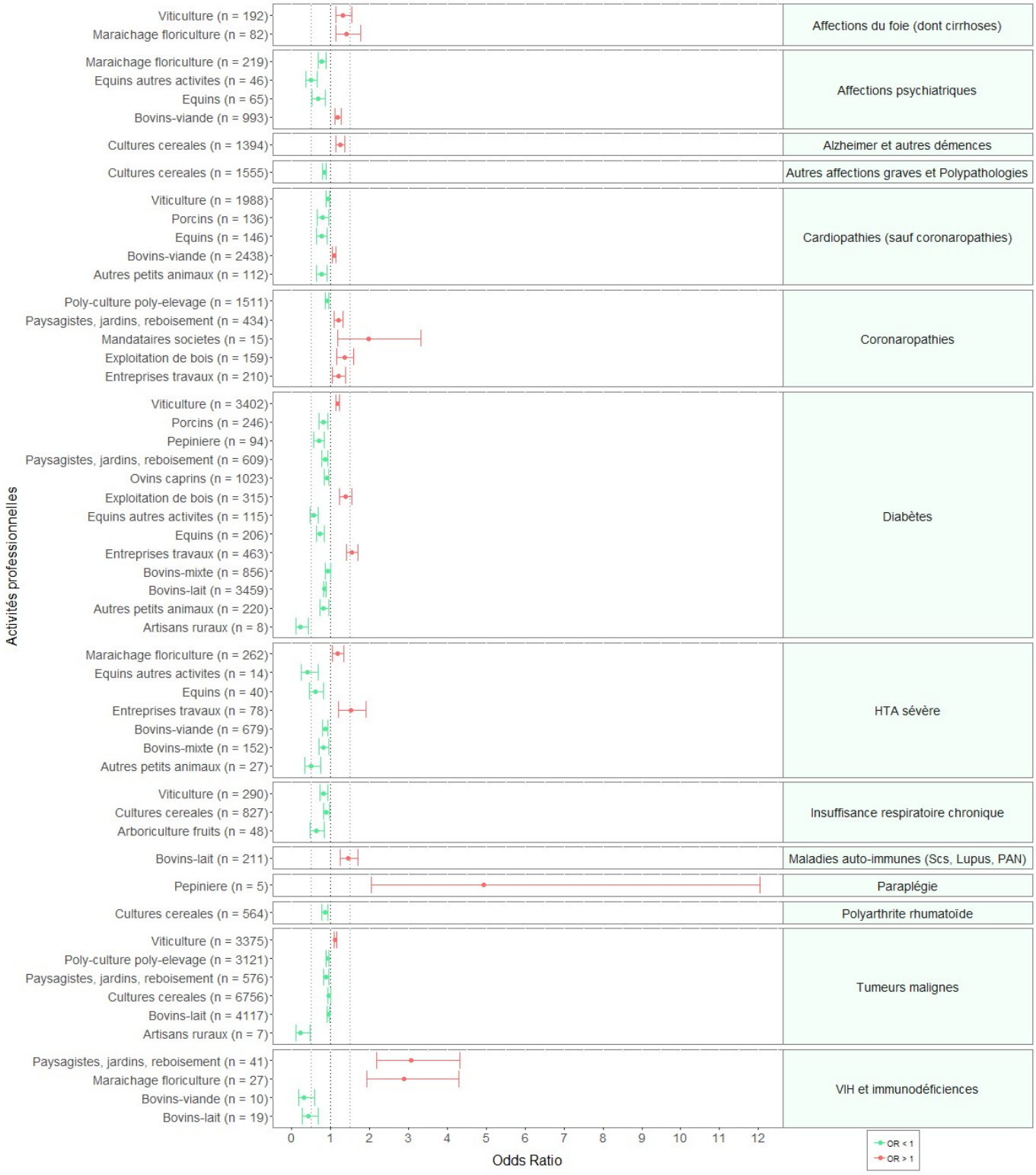


Figure 24 : Représentation graphique des odds ratios et de leurs intervalles de confiance à 95% obtenus via la régression logistique pour chaque association statistiquement significative (p-valeur$_{corrigée}$ < 0.05) entre une ALD et une activité professionnelle en ajustant sur d'autres paramètres, chez les non-salariés de la MSA (2006-2016)

Tableau 11 : Sélection de variables et mesure de l'AUC (échantillon de validation : 30% données) pour chaque ALD qui ont été utilisées lors de la régression logistique effectuée sur les non-salariés de la MSA (2006-2016)

Affections de longue durée (ALD)	Effectifs	Activités prof.	Année d' installation	Sexe	RSA	Chômage	Superficie d' exploitation	Âge	Nb. d' activités prof.	Nb. d' années d' obs.	« Revenus »	ALD avant obs.	« Marié »	« Veuf »	« Séparé ou divorcé »	Nb. de salariés	Régime maladie	Statut conjoint	Type d' exploitation	Région	AUC (échantillon de validation)
AVC invalidant	5288*	X	X					X				X			X		X				0.79
Cytopénies chroniques	284	X						X									X				0.73
Artériopathie chronique	4942	X	X	X			X	X				X			X	X	X				0.82
Cardiopathies (sauf coronaropathies)	16977	X	X	X	X			X			X	X		X			X		X	X	0.84
Affections du foie (dont cirrhoses)	1320	X	X	X	X			X				X			X		X		X		0.73
VIH et Immunodéficiences	241	X						X					X				X				0.81
Diabètes	25229	X	X	X	X	X		X	X		X	X					X		X	X	0.78
Affections neuromusculaires et épilepsie	1927	X	X		X			X				X					X				0.69
Hémophilies et affections hémostase	257	X															X				0.57
HTA sévère	5678	X	X					X			X	X					X		X	X	0.84
Coronaropathies	13210	X	X	X				X				X	X	X	X		X			X	0.81
Insuffisance respiratoire chronique	3122	X	X	X	X			X			X	X				X	X			X	0.81
Alzheimer et autres démences	2550	X	X	X			X	X		X		X		X			X			X	0.94
Parkinson	1686	X	X	X				X		X			X				X				0.85
Maladies métaboliques héréditaires	968	X	X					X				X					X			X	0.70
Néphropathies	1676	X	X	X				X				X					X				0.77
Paraplégie	176	X															X				0.58
Maladies auto-immunes (ScS, Lupus, PAN)	969	X	X	X				X				X					X				0.79
Polyarthrite rhumatoïde	2601	X	X	X				X				X	X				X				0.74
Affections psychiatriques	6438	X	X	X	X			X		X	X	X			X		X		X	X	0.71
Crohn et Rectocolite hémorragique	930	X						X		X		X					X			X	0.65
Sclérose en plaques	410	X		X				X		X							X				0.71
Scoliose structurale	87	X		X													X				0.76
Spondylarthrite ankylosante	1048	X						X		X		X	X				X				0.68
Suites de transplantation d'organe	197	X		X				X				X					X				0.79
Tuberculose, Lèpre	52	X																			0.57
Tumeurs malignes	25934	X	X					X				X	X		X		X			X	0.77
Autres affections graves et Polypathologies	5741	X	X	X				X			X	X		X			X		X	X	0.77

En ce qui concerne les variables sélectionnées pour chaque ALD lors de la régression logistique, il est possible de remarquer que certaines variables sont davantage incluses dans certains modèles que d'autres (Tableau 11) : le Régime maladie (inclus pour 27/28 ALD), l'Âge (inclus pour 24/28 ALD), l'ALD avant la période d'observation (incluse pour 20/28 ALD), l'Année d'installation (incluse pour 18/28 ALD) et enfin, le Sexe (inclus pour 16/28 ALD). Ces résultats montrent que ces cinq variables ont un rôle plus important à jouer dans l'explication générale des ALD que les autres variables.

Par ailleurs, concernant les mesures de robustesse des modèles, leurs aires sous la courbe ROC (AUC) montrent en moyenne un bon pouvoir discriminant (AUC_{moyen} = 0.75). Cependant, pour trois ALD, l'AUC est inférieure à 0.6 montrant un plus mauvais pouvoir discriminant. Par ailleurs, les ALD concernées, à savoir l'ALD « Hémophilies et affections hémostase » (n = 257), l'ALD « Paraplégie » (n = 176) et l'ALD « Tuberculose, Lèpre » (n = 52), ont des effectifs assez faibles vis-à-vis de l'effectif total de non-salariés étudiés. Les formules des modèles pour ces ALD sont les seules à n'inclure, outre l'activité professionnelle, aucune voire une seule variable supplémentaire. De plus, il est intéressant de remarquer que le nombre de variables et l'AUC n'évoluent pas de la même façon selon les ALD. A titre d'exemple, pour l'ALD « VIH et Immunodéficiences », la formule du modèle n'inclut que trois variables en plus de l'activité professionnelle (âge, statut « marié », régime maladie) et l'AUC du modèle est supérieur à 0.8. A l'inverse, pour l'ALD « Affections psychiatriques », la formule du modèle inclut onze variables en plus de l'activité professionnelle et l'AUC du modèle est de 0.71.

Cependant, les mesures de spécificité ($Spécificité_{moyenne}$ = 1), sensibilité ($Sensibilité_{moyenne}$ = 0) et F1 score ($F1_{moyen}$ = 0) montrent que ces modèles arrivent bien à distinguer les « non malades » mais qu'ils ne sont globalement pas performants pour identifier les « malades ».

III. Discussion

Le premier objectif de ce travail qui est d'étudier la faisabilité et la pertinence de l'utilisation des données médico-administratives de la MSA pour de la vigilance sanitaire a été rempli. Cette première partie du travail a démontré qu'il était possible, en collaboration avec la MSA, de sélectionner et restructurer les données de manière pertinente pour des analyses statistiques poussées. En effet, l'application de la régression logistique aux données de la MSA a permis de mettre en évidence 54 associations entre des ALD et des activités professionnelles chez les non-salariés, tout en ajustant sur dix-huit variables sélectionnées spécifiquement pour chaque modèle, notamment des variables de nature démographique (âge, sexe, …), socio-économique (chômage, RSA, assiette brute), géographique (région), ou encore relatives à des spécificités en lien avec l'activité professionnelle (année d'installation, superficie d'exploitation, …).

a. Forces et limites de la méthodologie

L'avantage majeur de ce projet est la puissance due à la **présence de la totalité de la population agricole française, et donc la quasi-totalité de la population des non-salariés dans le cadre de ce travail.** En effet, les études de ce type sont généralement menées sur des échantillons plus ou moins importants de la population cible, qui ne sont pas toujours représentatifs (absence de tirage au sort, défaut de la base de sondage). Ici, la population agricole française étant au complet dans les données de la MSA, la question de la représentativité des non-salariés étudiés ne se pose pas malgré l'exclusion de quelques individus (< 1% de la population des non-salariés).

Cependant, les analyses ne portant que sur la population des non-salariés agricoles, **cette population a été comparée à elle-même** et non à la population générale, ni même à la population couverte en totalité par le régime agricole. Or, il est connu dans la littérature scientifique que les exploitants agricoles partagent un certain nombre de risques communs qu'ils soient éleveurs ou cultivateurs (15,79). Cette limite peut alors avoir pour conséquence principale de masquer certaines associations entre des activités professionnelles et des pathologies. D'ailleurs, il s'agit sûrement de l'une des raisons pour lesquelles, on n'observe aucune association entre un risque professionnel et la maladie de Parkinson dans nos résultats, notamment car le risque d'apparition de cette maladie est diffus au sein de la population agricole (37,62). Par ailleurs, les maladies professionnelles n'ont pas été considérées pour ces analyses et il se trouve que la maladie de Parkinson a été ajoutée au

tableau des maladies professionnelles du régime agricole en 2012. Une analyse spécifique a alors été réalisée en ajoutant les individus ayant exclusivement une déclaration de MP « Parkinson » (n = 36). Cependant, de la même façon, aucune association statistiquement significative pour cette pathologie n'a été mise en évidence.

Par ailleurs, un processus de nettoyage important sur les données a été effectué permettant d'obtenir des données adaptées à nos analyses (erreurs de saisies et incohérences corrigées), mais avec pour conséquence, une sélection de la population **écartant un nombre de non-salariés important (n = 14 933) pour lesquels les données sont soit incomplètes, soit comportent des incohérences.**

Certains choix méthodologiques effectués au cours du nettoyage et de la fusion des données ont pu avoir des conséquences importantes sur la recherche d'associations notamment le choix de ne considérer que les pathologies déclarées en ALD durant la période d'observation, c'est-à-dire, pour lesquelles on dispose d'informations professionnelles antérieures à la déclaration d'ALD. Ce choix a permis de limiter l'hypothèse sur la stabilité professionnelle des non-salariés pour lesquels nous avons considéré qu'ils conservaient la même activité professionnelle entre leur année d'installation dans leur exploitation et leur période d'observation. De ce fait, près de 19 000 non-salariés avec une déclaration d'ALD ont été considérés comme « non-malades » car nous n'avions aucune donnée sur l'activité professionnelle exercée avant la déclaration de leur pathologie. Cela peut avoir pour conséquence de masquer certaines associations concernant des maladies chroniques ayant un délai d'apparition plus important que d'autres.

D'ailleurs, **le délai d'apparition des pathologies n'a pas pu être pris en compte dans ces analyses,** essentiellement car les données à disposition étaient limitées (peu d'années d'antériorité), et dans une moindre mesure du fait qu'il y a peu d'informations dans la littérature scientifique en ce qui concerne le temps de développement de chacune des affections, en particulier en ce qui concerne les cancers et les maladies neurodégénératives. Cependant, dans le cadre de ce travail de thèse, une conceptualisation a été réalisée pour prendre en compte un temps de latence dans le cas où nous disposerions à l'avenir d'une antériorité plus importante concernant les données dans le cadre du projet (Annexe 2). Par ailleurs, ici, ne pas prendre en compte ce temps de latence peut avoir pour conséquence de masquer certaines associations, notamment pour les individus dont l'exposition serait lointaine, ou de façon plus improbable, de mettre en évidence des associations erronées.

De même, **la « durée d'exposition »** (ici, le temps d'exercice d'une activité professionnelle) est en mesure d'influencer l'apparition de maladies chroniques qui seraient liées à un facteur

professionnel. Cependant, cette durée est complexe à estimer pour les non-salariés car nous ne disposons pas de l'ensemble de l'historique professionnel des individus. Cette durée ne peut alors être calculée qu'en fonction des années d'enregistrement car les activités effectuées par un chef d'exploitation entre son année d'installation et la première année d'enregistrement à disposition sont inconnues. Toutefois, en tenant compte de cette durée, calculée de cette façon, nous n'aurions qu'une vision partielle du temps d'exposition des non-salariés. Cela est d'autant plus vrai que, comme vu dans la partie précédente (Tableau 2), environ 23% des non-salariés exercent leur activité principale, en termes de quotité de temps ou de revenus perçus, dans un autre régime que le régime agricole (variable « Régime maladie d'affiliation »). Ce paramètre ajoute également un degré de complexité pour le calcul de cette « durée d'exposition ». Dans le cadre de ce travail, cette durée n'a donc pas été prise en compte et les activités professionnelles ont été traitées en tant que variables binaires (absence ou présence des non-salariés dans les activités professionnelles), ce qui représente un biais supplémentaire qu'il est important de mentionner.

Aussi, **la variable servant à renseigner les activités professionnelles** chez les non-salariés à partir du thésaurus interne à la MSA et utilisée pour approcher les expositions professionnelles des non-salariés ne propose que 26 modalités, dont certaines regroupent diverses activités professionnelles comme la modalité « Cultures céréalières et industrielles ». Cependant, cette variable a été choisie pour être utilisée dans nos analyses car sa fiabilité est plus importante comparée à la variable renseignant le code métier NAF. Ainsi, ce manque de précision sur l'exposition professionnelle ne permet de voir que des associations statistiques à l'échelle macroscopique.

La **régression logistique** a été choisie car cette méthode est très utilisée en épidémiologie pour la recherche d'associations avec des variables à expliquer de type binaire (ici, les pathologies déclarées en ALD) et car elle permet d'obtenir des odds ratios. De plus, cette méthode est connue pour sa fiabilité, et le bon compromis qu'elle permet a priori d'obtenir entre performance du modèle et pouvoir explicatif (92,97,100). Les modèles ont été conçus pour chaque ALD sans *a priori* grâce à une sélection de variables effectuée pas à pas avec la majorité des variables fournies par la MSA, à l'exception de celles ayant des mesures d'intensité de liaison trop importantes entre elles.

D'ailleurs, **la sélection de variables** telle qu'effectuée peut être remise en question du fait que les variables âge et sexe n'ont pas été imposées systématiquement, même quand elles ne sont pas associées à la pathologie étudiée, alors qu'elles sont des déterminants essentiels de la santé. Cependant, il est important de rappeler qu'avec les données ALD de la MSA, nous n'avons pas de vision exhaustive de l'ensemble des individus ayant les pathologies étudiées,

d'autant plus que nous avons dû écarter certains individus « malades » pour nos analyses (cf. ci-dessus). Dans ce contexte, il est possible que des variables comme l'âge ou le sexe ne soient pas associées aux pathologies étudiées et elles n'ont donc pas été ajoutées aux modèles. A titre d'exemple, en ce qui concerne l'ALD « Hémophilies et affections hémostase », la variable sexe n'a pas été incluse dans le modèle alors que l'hémophilie touche exclusivement les garçons sauf cas exceptionnels. Or, il s'agit d'une ALD pour laquelle nous avons peu de déclarations prises en compte dans nos analyses et donc certainement une vision partielle du fait que ces affections sont généralement diagnostiquées dans l'enfance. Ainsi, il est important d'être conscient qu'il peut y avoir des biais liés à la représentativité des individus, dans une plus ou moins grande mesure selon les pathologies étudiées, pouvant alors masquer certaines associations.

En ce qui concerne les mesures de robustesse, on observe des résultats globalement satisfaisants au niveau des AUC. Mais les mesures de spécificité, de sensibilité et de F1 score montrent que les modèles sont peu efficaces pour identifier les individus ayant des déclarations d'ALD. Trois explications peuvent être envisagées :

- soit le nombre d'individus ayant des déclarations d'ALD est trop faible par rapport au nombre total d'individus étudiés (maximum de 2,8% de « malades » si on considère l'ALD « Tumeurs malignes ») (108) ;

- soit les variables à disposition ne suffisent pas à « expliquer » les pathologies déclarées en ALD et il existe des facteurs de confusion résiduels non pris en compte ;

- soit la méthode utilisée pour sélectionner les variables dans chaque modèle ne permet pas d'obtenir une *bonne* sensibilité.

b. Synthèse des principaux résultats

La régression logistique a permis de mettre en évidence des associations statistiques significatives entre des ALD et des activités professionnelles chez les non-salariés de la MSA. **Parmi les 54 associations mises en évidence entre ALD et activités professionnelles, il est possible d'observer que certaines capturent des déterminants de la santé au travail déjà suspectés en agriculture, telles que les associations suivantes :**

- l'association montrant un excès de risque de déclaration de l'ALD « Maladie d'Alzheimer et autres démences » dans le secteur des cultures céréalières et industrielles (OR = 1.24 [1.14 ;1.36] ; p-valeur$_{corrigée}$ = 5.07^{E-5} ; n = 1 394), un secteur agricole particulièrement exposé aux pesticides (109) ;

- l'association montrant un excès de risque de déclaration d'ALD « Tumeurs malignes » dans le secteur de la viticulture (OR = 1.11 [1.07 ;1.16] ; p-valeur$_{corrigée}$ = 1.05^{E-6} ; n = 3 375), ce qui est aussi attendu car d'une part, ce secteur professionnel serait davantage à risque de développer des cancers du poumon (78) et d'autre part, les viticulteurs ont pu être exposés à des pesticides arsenicaux autorisés en France jusque 2001, un facteur de risque majeur dans le développement de cancers de la vessie (75,110) ; cependant, les professionnels du secteur viticole montrent également un risque plus élevé de déclarations d'ALD « Affections du foie (dont cirrhoses) », qui pourrait être en lien avec des facteurs professionnels ou comportementaux (consommation d'alcool) ;

- l'association montrant un risque moins important de déclaration d'ALD « Tumeurs malignes » dans le secteur de l'élevage de bovins laitiers (OR = 0.94 [0.90 ;0.97] ; p-valeur$_{corrigée}$ = 3.55^{E-3} ; n = 4 117) qui a lui aussi déjà été décrit dans la littérature (111) ;

- et également, une association montrant un excès de risque de déclaration d'ALD « Diabètes » dans le secteur de la viticulture (OR = 0.94 [0.90 ;0.97] ; p-valeur$_{corrigée}$ = 3.55^{E-3} ; n = 4 117) a été mise en évidence, comme dans la littérature scientifique où de récentes études ont suggéré une relation entre le diabète et les pesticides (26,112,113), et plus spécifiquement entre l'exposition à l'arsenic même à faible niveau qui semble être un facteur de risque d'insulino-résistance et de diabètes (114,115), même si des investigations supplémentaires sont nécessaires du fait des multiples facteurs étiologiques des diabètes (116).

De plus, on observe deux associations montrant des risques plus élevés de déclarations d'ALD « Coronaropathies » chez les travailleurs des exploitations de bois (OR = 1.36 [1.16 ;1.59] ; p-valeur$_{corrigée}$ = 0.002 ; n = 159) et dans le secteur des « paysagistes, jardins, reboisement » (OR = 1.19 [1.07 ;1.31] ; p-valeur$_{corrigée}$ = 0.008 ; n = 434). Or, dans ces secteurs agricoles, les travailleurs utilisent fréquemment des machines portatives comme les tronçonneuses, exposant particulièrement à des gaz d'échappement comme les fumées de diesel. Dans la littérature scientifique, à titre d'exemple, plusieurs études ont suggéré que l'exposition professionnelle à des fumées de diesel pourrait être associée à des effets néfastes sur la santé cardiovasculaire mais aussi à des cancers du poumon (117). Par ailleurs, au niveau de précision de l'ALD, aucune association n'est mise en évidence pour l'ALD « Tumeurs malignes » pour ces secteurs d'activités. Cependant, cette ALD regroupant un groupe de pathologies aux localisations assez diverses, des analyses sont menées au niveau CIM-10 dans la PARTIE 5 de ce manuscrit de thèse.

Quant à certaines associations, il est possible de les relier à des déterminants sociaux de santé, ce qui est courant en épidémiologie de la santé au travail. A titre d'exemple, l'excès de risque de déclaration d'ALD « Affections psychiatriques » dans le secteur de l'élevage de viande bovine (OR = 1.18 [1.10 ;1.27] ; p-valeur$_{corrigée}$ = 4.11^{E-5} ; n = 993) est concordant avec le taux de suicide relativement élevé observé dans ce secteur professionnel par Santé Publique France entre 2007 et 2009. En effet, à cette même période en France, le secteur de l'élevage bovins était particulièrement affecté par des difficultés financières liées à une crise économique importante (118). Dans le cadre de ce travail, cette association montre aussi qu'il est possible, avec la méthodologie employée, d'observer un signal sanitaire tel que celui-ci de manière plus précoce qu'en étudiant les causes de mortalité.

De même pour l'ALD « VIH et Immunodéficiences », il est possible d'observer un excès de risque de déclaration pour cette ALD dans les secteurs du maraîchage et de la floriculture (OR = 2.88 [1.93 ;4.31] ; p-valeur$_{corrigée}$ = 2.92^{E-6} ; n = 27) mais aussi chez les paysagistes ou dans les entreprises de jardins et de reboisement (OR = 3.06 [2.17 ;4.32] ; p-valeur$_{corrigée}$ = 5.87^{E-9} ; n = 41). Or, en France, dans ces secteurs, il est possible de trouver davantage d'individus ayant été auparavant dans une grande précarité mais qui ont bénéficié d'aides pour favoriser leur réinsertion sociale. Ainsi, on peut faire l'hypothèse d'une prévalence plus élevée de comportements à risque (addictions, maladies sexuellement transmissibles, …) au sein de ces secteurs d'activités. Cette hypothèse peut être corroborée en partie pour les secteurs du maraîchage et de la floriculture pour lesquels on observe également un excès de risque de déclaration d'ALD « Affections du foie (dont cirrhoses) » (OR = 1.40 [1.12 ;1.76] ; p-valeur$_{corrigée}$ = 4.04^{E-2} ; n = 82), où la cirrhose est une pathologie pour laquelle les principaux facteurs de

risque sont une consommation d'alcool excessive et l'hépatite virale chronique (119). A contrario, les éleveurs bovins (lait et viande) présentent des risques plus faibles de déclarations pour cette ALD « VIH et Immunodéficiences ». Il est possible de faire l'hypothèse qu'ils sont moins exposés à ce type de facteurs de risque du fait qu'ils sont plutôt installés en zone rurale, laissant souvent supposer un certain isolement social. Cette hypothèse serait à étayer avec l'aide de sociologues.

Par ailleurs, les analyses statistiques, menées sans hypothèses *a priori*, ont aussi la capacité de générer des hypothèses, révélant des associations qui n'auraient pas nécessairement été explorées auparavant. Par exemple, un excès de risque de déclaration d'ALD « Maladies auto-immunes » dans le secteur de l'élevage de bovins laitiers (OR = 1.45 [1.24 ;1.69] ; p-valeur$_{corrigée}$ = 6.89^{E-5} ; n = 315) a été mis en évidence via notre méthodologie qui, *a priori*, n'a jamais été décrite à ce jour dans la littérature scientifique. Il est intéressant de noter qu'un rôle supposé d'une réponse immunitaire aux protéines du lait de vache a été envisagé dans la pathogenèse de la maladie de Behçet (120), qui fait partie de la pathologie ALD en question (code « M35 : Autres atteintes systémiques du tissu conjonctif » de la CIM-10 : n = 105). Ce type d'association appelle alors à davantage d'investigations.

Enfin, malgré la correction des p-valeurs via la procédure de Benjamini-Hochberg, la pertinence des signaux mis en évidence peut être contrebalancée par les effectifs à la fois très faibles dans certaines ALD ou très importants pour d'autres (sous ou sur-déclarations potentielles). En effet, pour les ALD « Diabètes » et « Tumeurs malignes », pour lesquelles les effectifs étaient importants, il y a davantage d'associations mises en évidence. Les performances de la régression logistique semblent alors être nettement influencées par les effectifs, ce qui peut alors avoir un impact sur la détection d'associations (108). Cependant, il est tout de même possible de voir des associations concernant de faibles effectifs, comme par exemple l'association mise en évidence dans le secteur des pépinières pour l'ALD « Paraplégie » (OR = 4.94 [2.03 ;12.03] ; p-valeur$_{corrigée}$ = 1.12^{E-2} ; n = 5).

En tout état de cause, quelles que soient les caractéristiques des signaux mis en évidence, il est nécessaire de les vérifier et de les valider à l'aide d'autres méthodologies et de les investiguer à l'aide d'un groupe pluridisciplinaire rassemblant des médecins conseils[13] et du travail de la MSA, des épidémiologistes, des toxicologues et des sociologues.

[13] Médecins désignés au sein des organismes de la MSA principalement pour jouer un rôle d'expert du système de santé et contrôler le bien fondé des remboursements de soins des assurés agricoles.

Résumé

La régression logistique a été appliquée avec succès aux données de la MSA. Une sélection spécifique de variables a été réalisée afin de permettre un ajustement aussi bien sur des variables de nature démographique ou socio-économique par exemple, mais aussi dans le but de tenir compte des spécificités de chacune des pathologies ALD étudiées. **Ainsi, 28 sélections de variables ont été réalisées et 728 modèles ont permis de tester chaque association entre une ALD et une activité professionnelle.** De plus, au vu du nombre de tests différents effectués, une correction sur les p-valeurs via la procédure de Benjamini-Hochberg a été utilisée.

La méthodologie employée a alors permis de révéler au total 54 associations statistiquement significatives ($p\text{-valeur}_{corrigée} < 0.05$) entre une activité professionnelle et une ALD, dont 35 associations avec un risque plus faible de déclarations d'ALD spécifiques et 19 associations avec un risque plus élevé de déclarations pour les ALD concernées. **Parmi ces associations, certaines capturent des déterminants de la santé au travail déjà suspectés en agriculture, d'autres peuvent être reliées à des déterminants sociaux de santé et enfin, certaines permettent de générer des hypothèses.**

Par ailleurs, la méthodologie montre certaines limites au travers des mesures de robustesse, notamment une sensibilité nulle des modèles quelle que soit la pathologie étudiée. Des optimisations sont alors réalisées dans les parties suivantes de ce travail, afin de tenter de réduire les difficultés liées aux faibles effectifs de « malades » d'une part, et aux facteurs de confusion résiduels d'autre part. Une autre méthode de sélection de variables est également testée dans la PARTIE 4.

PARTIE 4

Optimisations de la méthodologie de modélisation

A. Estimation de facteurs de confusion

I. Méthodologie

a. Biais de confusion

Dans le cadre d'étude de recherche d'associations, il est fréquent de voir que les associations découvertes peuvent largement varier selon les variables d'ajustement. Il est donc évident d'essayer de limiter ces variations en ajustant au mieux les modèles pour mettre en évidence des associations pertinentes. Par ailleurs, il est important de rappeler que les données utilisées, conçues initialement à des fins de cotisation et de remboursement, ne sont pas optimales pour être utilisées à des fins de vigilance sanitaire et qu'elles ne sont pas aussi complètes que pourraient l'être des données collectées dans le cadre de cohortes à visée épidémiologique. En effet, dans notre étude, nous pouvons faire l'hypothèse qu'il existe un certain nombre de facteurs non observés permettant d'expliquer au mieux les différentes pathologies étudiées mais pouvant aussi être fortement corrélés aux variables observées telles que l'âge ou le sexe. Ainsi, dans ce type d'études de recherche d'associations, on est souvent confronté à des problèmes de causalité dus aux facteurs non observés, aussi appelés facteurs de confusion ou facteurs latents.

En effet, lorsque l'on détecte une association entre deux variables, il n'y a pas nécessairement de relation de causalité car la relation mise en évidence peut aussi impliquer des liens avec des facteurs de confusion. Dans notre contexte, un facteur de confusion peut alors être une variable associée à la fois à l'activité professionnelle, considérée comme le facteur d'exposition et à l'ALD qu'on souhaite expliquer (Figure 25). Ce facteur de confusion peut alors avoir un impact sur l'estimation de l'effet d'une exposition sur une pathologie du fait d'une distribution différente de ce facteur entre les groupes exposés et les groupes non exposés. Un facteur de confusion peut soit entraîner une surestimation ou une sous-estimation de l'effet d'une exposition, soit totalement masquer un effet (121,122). Par ailleurs, il a été reconnu dans la littérature scientifique que le développement de certaines pathologies serait en grande partie lié à une mauvaise hygiène alimentaire, une consommation de tabac ou encore à la sédentarité, autant de paramètres pouvant jouer le rôle de facteurs de confusion lors de nos recherches d'associations et qui ne sont pas disponibles au sein des données de la MSA.

Ainsi, dans ce contexte, deux méthodologies différentes ont été testées pour approcher les facteurs de confusion potentiels, afin de limiter ce type de biais dans la recherche d'associations :

- l'utilisation d'une méthode statistique appelée « LFMM » (*Latent Factor Mixed Models*) permettant l'estimation de facteurs latents à partir des données observées ;

- l'ajout des comorbidités en tant que variables d'ajustement pour les maladies, c'est-à-dire, des ALD survenues avant la date de la déclaration de l'ALD étudiée.

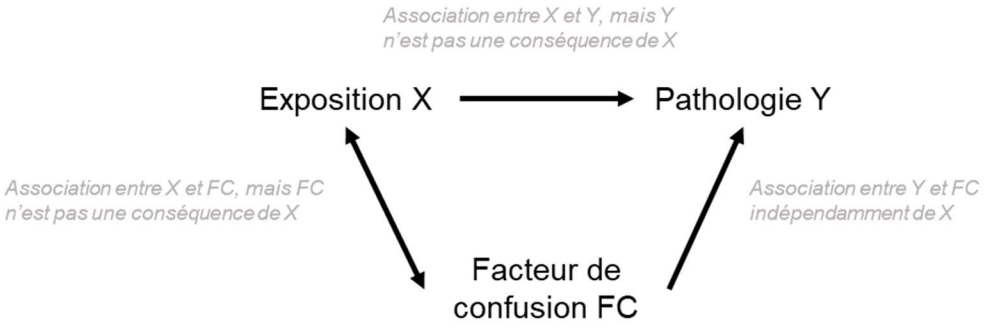


Figure 25 : Exemple d'une association entre une pathologie Y et une exposition X médiée par un facteur de confusion (Organisation Mondiale de la Santé)

b. Estimation de facteurs latents via LFMM

Dans le cadre de cette thèse, nous avons souhaité évaluer la pertinence, vis-à-vis de nos objectifs, de l'utilisation de la méthode LFMM sur les données de la MSA pour estimer les facteurs de confusion potentiels. Cette méthode développée en langage R est généralement utilisée pour corriger les études d'associations. En parcourant la littérature, on s'aperçoit qu'il existe de nombreuses méthodes permettant de prendre en considération des facteurs latents dans les études d'associations. Cependant, aucune méthode ne s'est actuellement imposée comme méthode de référence et la méthode LFMM est d'autant plus adaptée à nos données puisqu'elle est utilisée habituellement en génétique sur des données de grande dimension, comparables à celles de la MSA (123,124).

Comme nous l'avons expliqué dans la partie précédente, l'objectif est d'utiliser la régression logistique afin de mettre en évidence des associations entre la survenue d'un événement (Y) et des variables explicatives (X_k où k est le nombre de variables). Avec l'utilisation de la méthode LFMM, nous ajoutons un paramètre U_m au modèle de régression qui correspond à

la matrice des *m* facteurs latents, calculés en prenant en considération autant les variables à expliquer (pathologies déclarées en ALD) que les variables explicatives (âge, sexe, …). En effet, pour l'estimation des facteurs latents, la méthode utilise d'une part la matrice des variables à expliquer Y de taille *n*i* et la matrice des variables explicatives X de taille *n*k*, où n est le nombre de lignes (individus), i est le nombre d'ALD considérées et k est le nombre de variables explicatives considérées. Les activités professionnelles, étant les principales variables d'intérêt, n'ont pas été utilisées pour estimer les facteurs latents. Ces derniers sont alors estimés d'une part, de sorte à ce que ces facteurs non observés identifient des regroupements d'individus en fonction de leurs pathologies déclarées en ALD (matrice Y) et d'autre part, de sorte à minimiser la perte de variance expliquée par l'ensemble des variables du modèle (matrice X : âge, sexe, … ; et facteurs latents U_m) (125).

Pour définir le nombre de facteurs latents à estimer, la méthode préconisée est l'analyse en composantes principales (ACP), qui a donc été réalisée sur la matrice Y (matrice des ALD composée 28 variables binaires). Cette technique d'analyse des données permet à partir d'une matrice de variables de construire un certain nombre d'autres variables, appelées composantes principales, qui sont des combinaisons linéaires des variables initiales. Ces composantes principales résument alors les informations apportées par les variables initiales, notamment les premières qui restituent le plus d'information (92). Ainsi, à partir de la visualisation graphique de l'ACP réalisée, il a été possible de décider du nombre de facteurs latents à estimer de manière heuristique, en choisissant la valeur pour laquelle l'histogramme des variances présente un « coude ». Ici, la figure nous indique que l'ACP a identifié quatre composantes principales à partir de cette matrice, celles-ci expliquant plus de 60% de la variance totale et donc de l'information apportée par la matrice des ALD (Figure 26). Il a donc été décidé d'estimer quatre facteurs latents via la méthode LFMM.

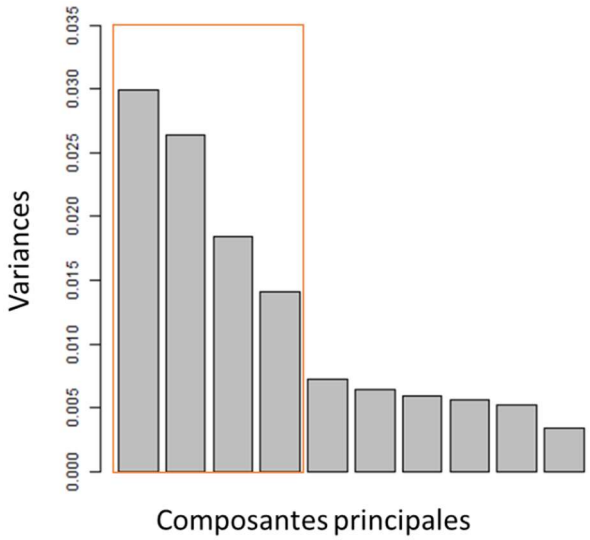


Figure 26 : Représentation graphique de l'analyse en composantes principales réalisée sur la matrice des ALD (28 variables binaires)

Les facteurs latents estimés via LFMM étaient sous forme de variables quantitatives. L'ensemble des variables à expliquer et de variables explicatives étant au format binaire, les facteurs latents ont été normalisés de sorte à prendre des valeurs comprises entre 0 et 1, afin que toutes les variables soient par la suite à la même échelle pour la modélisation.

Une fois les facteurs latents U_m estimés et normalisés, ils ont été ajoutés au modèle de régression logistique précédent (cf. Partie 3 I. c.) avant l'étape de sélection pas à pas des variables explicatives. Ainsi, la méthode de sélection de variables utilisée a permis d'inclure ou non, pour chaque pathologie étudiée, un à quatre facteurs latents en fonction du critère BIC pour obtenir le modèle le plus parcimonieux.

Le modèle utilisé est alors le suivant :

ALD_i ~ Activité professionnelle $_j$ + sélection de variables explicatives X_k et de facteurs latents U_m

où i correspond à chaque ALD (n = 31), j correspond à chaque activité professionnelle (n = 26), k correspond au nombre de variables explicatives ajoutées au modèle et m correspond au nombre de facteurs latents ajoutés au modèle.

c. Ajout des comorbidités

Dans la littérature, il a été démontré qu'un patient développant une pathologie, même spécifique, peut être davantage enclin à développer d'autres pathologies. Par exemple, un patient diabétique a davantage de risques de développer une maladie cardiovasculaire car son diabète peut dans certains cas, accélérer le processus d'athérosclérose, une maladie touchant les artères, elle-même à l'origine d'Accidents Vasculaires Cérébraux (AVC) (126). De plus, le développement de certaines pathologies est lié à des facteurs de confusion potentiels tels que l'alcoolisme ou le tabagisme, variables qui ne sont pas disponibles au sein des données MSA. Par exemple, le tabagisme est la cause principale (80% des cas) de la bronchopneumopathie chronique obstructive (127). Ainsi, pour l'étude d'une pathologie spécifique, nous avons émis l'hypothèse que des pathologies antérieures peuvent soit jouer un rôle dans le développement de cette dernière, soit permettre d'approcher des facteurs de confusion potentiels tels que le tabagisme ou l'alcoolisme.

De ce fait, pour chaque ALD étudiée, il a été possible d'ajouter les variables renseignant les ALD antérieures déclarées en tant que facteurs de confusion. Cependant, comme le nombre d'ALD est important (28 ALD étudiées), il faudrait ajouter autant de variables explicatives aux modèles. Il a donc été décidé de réaliser une analyse en composantes principales de la matrice des ALD antérieures. En effet, l'ACP est une technique couramment utilisée pour diminuer le nombre de dimensions en réduisant le nombre de variables étudiées tout en évitant la perte importante d'information (92). Le premier avantage est que les composantes principales sont alors indépendantes et non corrélées entre elles, contrairement aux variables renseignant les ALD antérieures qui peuvent l'être. Le deuxième avantage est de limiter le nombre de variables à ajouter au modèle pour éviter la perte de puissance statistique car l'ACP permet d'obtenir un nombre de variables moins important mais à fort pouvoir explicatif. Pour chacune des ALD étudiées, une ACP a alors été réalisée sur la matrice des ALD antérieures, de taille $n*(i-1)$ où n est le nombre de lignes (individus) et i est le nombre d'ALD antérieures. Puis, il a été choisi de conserver les quatre premières composantes principales de la même façon qu'il avait été décidé d'estimer seulement quatre facteurs latents via la méthode LFMM. Une matrice des comorbidités C_o a ainsi été obtenue, de taille $n*o$, où o correspond au nombre de composantes principales et n au nombre de lignes (individus).

Ensuite, de la même façon que pour les facteurs latents estimés via LFMM, les quatre variables de la matrice des comorbidités sont sous forme de variables quantitatives. Ces variables ont donc été normalisées de sorte à prendre des valeurs comprises entre 0 et 1, afin que toutes les variables explicatives soient à la même échelle pour la modélisation.

Les variables de la matrice des comorbidités sont ensuite ajoutées au modèle de régression logistique précédent (cf. Partie 3 I. c.) avant l'étape de sélection pas à pas des variables explicatives. Ainsi, la méthode de sélection de variables utilisée a permis d'inclure ou non, pour chaque pathologie étudiée, une à quatre variables de la matrice des comorbidités en fonction du critère BIC pour obtenir le modèle le plus parcimonieux.

Le modèle utilisé est alors le suivant :

$$\mathbf{ALD}_{i} \sim \mathbf{Activité\ professionnelle}_{j} + \mathbf{sélection\ de\ variables\ explicatives}\ X_k\ \mathbf{et\ de\ comorbidités}\ C_o$$

où i correspond à chaque ALD (n = 31), j correspond à chaque activité professionnelle (n = 26), k correspond au nombre de variables explicatives ajoutées au modèle et o correspond au nombre de variables de la matrice des comorbidités ajoutées au modèle.

d. Schéma récapitulatif

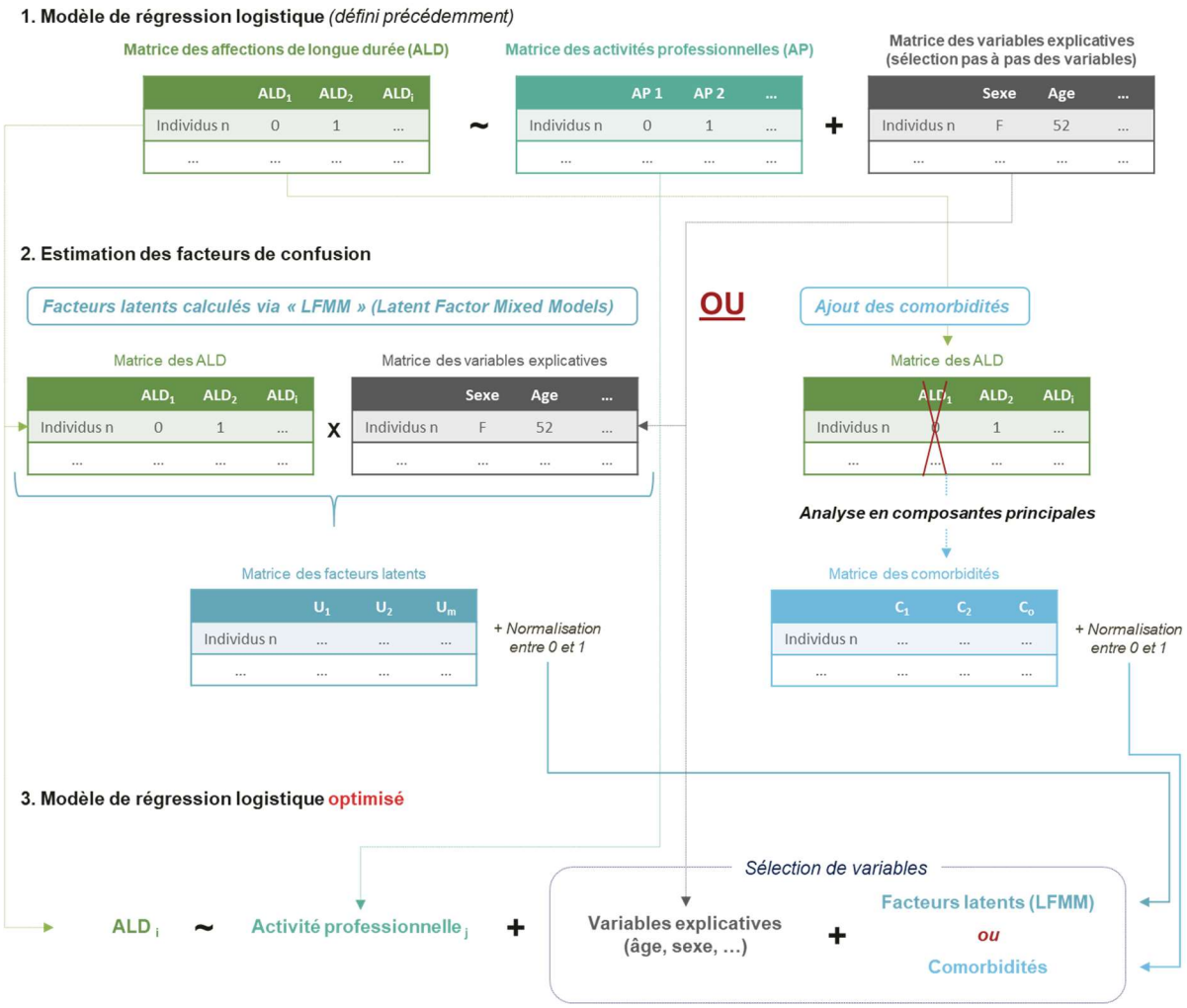


Figure 27 : Schéma récapitulatif des deux méthodologies employées pour l'estimation des facteurs de confusion, utilisés ensuite comme variables d'ajustement additionnelles lors de l'application de la régression logistique aux données de la MSA

II. Résultats

a. Ajout de facteurs latents

De nouveau, des associations entre ALD et activités professionnelles chez les non-salariés de la MSA au cours de la période d'observation ont pu être mises en évidence via la régression logistique en ajoutant les facteurs latents estimés via la méthode LFMM. La Figure 28 représente les p-valeurs de chaque association testée, corrigées par la procédure de Benjamini-Hochberg et transformées à l'aide du logarithme décimal ($-\log_{10}$). Comparé à la Figure 22 représentant les p-valeurs des associations mises en évidence dans la partie précédente de ce travail (Régression logistique « simple »), les p-valeurs des associations mises en évidence en ayant ajouté les facteurs latents estimés via LFMM sont plus élevées et ainsi moins nombreuses à être statistiquement significatives. Contrairement aux 54 associations statistiquement significatives ($\text{p-valeur}_{\text{corrigée}} < 0.05$) mises en évidence précédemment, seules 32 associations entre une ALD et une activité professionnelle ont été mises en évidence. En effet, contrairement aux résultats précédents, il n'y a aucune association mise en évidence pour 4 des 28 ALD : « Cardiopathies (sauf coronaropathies) », « Diabètes », « Coronaropathies » et « Tumeurs malignes ». Pour mémoire, il s'agissait des ALD ayant les effectifs de déclarations les plus importants (cf. Tableau 8) et donc de celles pour lesquelles les facteurs latents sont construits avec le plus d'information.

En complément, une heatmap a été réalisée pour représenter les associations statistiquement significatives (Figure 29). Par rapport aux associations précédemment mises en évidence chez les **viticulteurs**, une seule parmi les trois est maintenue avec un risque de déclaration d'ALD plus élevé pour l'ALD « Affections du foie (dont cirrhoses) » (OR = 1.28 [1.10 ;1.50] ; $\text{p-valeur}_{\text{corrigée}}$ = 0.029 ; n = 192) (Figure 30). Pour ce qui est des associations précédemment mises en évidence chez **les maraîchers et les floriculteurs**, elles sont quasiment similaires en termes d'odds ratios et de p-valeurs avec un risque plus élevé de déclaration d'ALD « Affections du foie (dont cirrhoses) », « VIH et immunodéficiences » et « HTA sévère ». Enfin, pour ce qui est des **éleveurs bovins (lait)**, ils ont toujours un risque de déclaration d'ALD « VIH et immunodéficiences » moins important avec un odds ratio et une p-valeur identiques mais les associations concernant le risque plus important d'ALD « Diabètes » et « Tumeurs malignes » n'ont pas été mises en évidence via cette méthodologie, comme indiqué dans le paragraphe précédent. Pour cette même profession, l'association concernant un risque plus élevé de déclaration d'ALD « Maladies auto-immunes (ScS, Lupus, PAN) » est identique en termes de p-valeur et d'odd ratio.

Néanmoins, malgré un grand nombre d'associations statistiquement significatives qui ont disparu, huit nouvelles sont apparues, pour lesquelles on trouve des :

- Risques plus élevés de déclarations d'ALD « Artériopathie chronique » pour les viticulteurs (OR = 1.16 [1.06 ;1.26] ; p-valeur$_{corrigée}$ = 0.008 ; n = 687), les travailleurs agricoles des exploitations de bois (OR = 1.46 [1.12 ;1.89] ; p-valeur$_{corrigée}$ = 0.03 ; n = 61) et ceux des scieries fixes (OR = 2.59 [1.32 ;5.07] ; p-valeur$_{corrigée}$ = 0.03 ; n = 9) ;

- Risques plus élevés de déclarations d'ALD « Insuffisance respiratoire chronique grave » pour les éleveurs bovins (lait) (OR = 1.15 [1.04 ;1.27] ; p-valeur$_{corrigée}$ = 0.04 ; n = 564) et les éleveurs ovins et caprins (OR = 1.25 [1.06 ;1.46] ; p-valeur$_{corrigée}$ = 0.04 ; n = 176) ;

- Et un risque plus élevé de déclaration d'ALD « Néphropathies » chez les conchyliculteurs (OR = 2.80 [1.49 ;5.26] ; p-valeur$_{corrigée}$ = 0.03 ; n = 10).

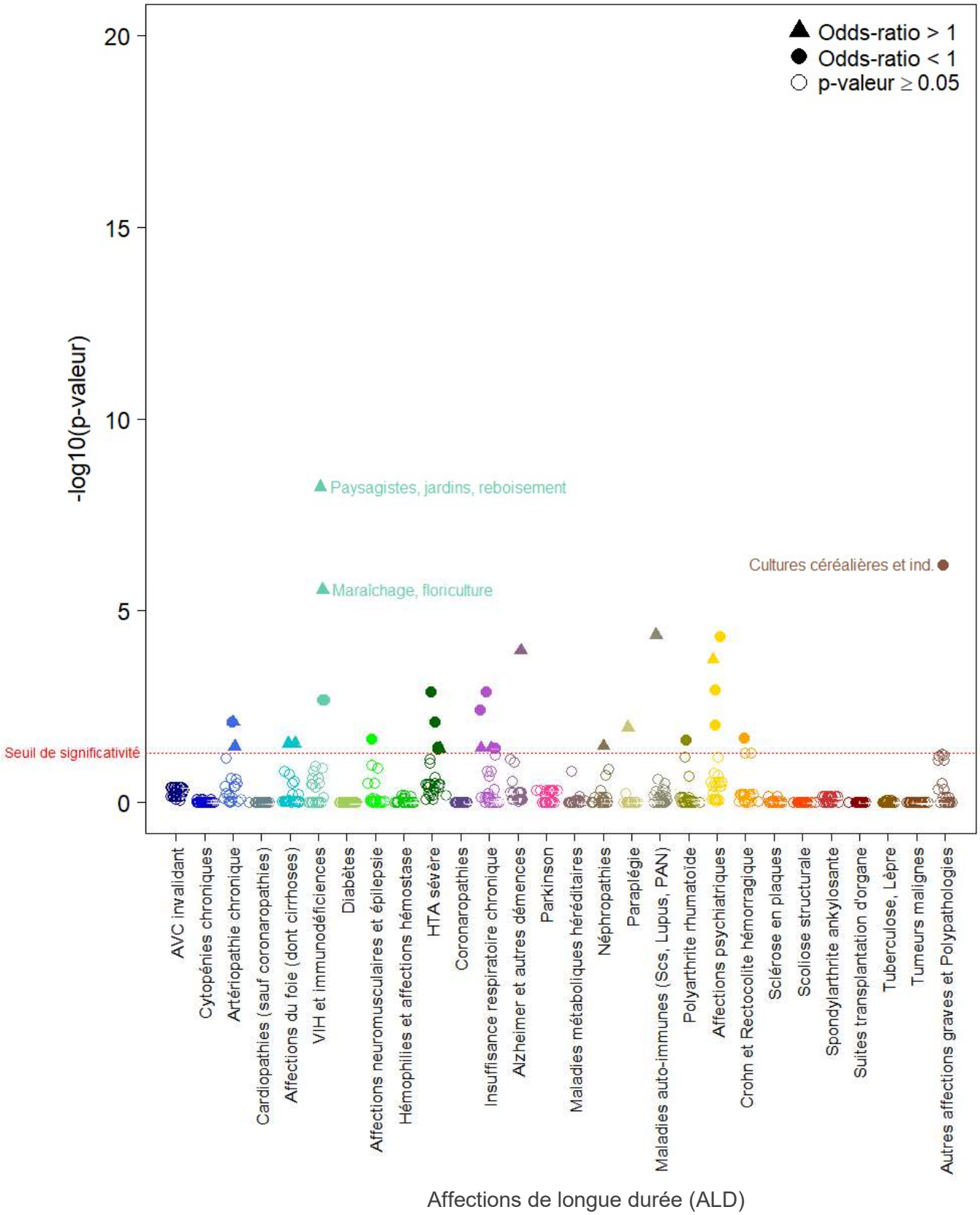


Figure 28 : Représentation graphique des p-valeurs, corrigées par la procédure de Benjamini-Hochberg, obtenues via la régression logistique entre chaque combinaison d'ALD et d'activité professionnelle en ajustant sur d'autres paramètres **notamment des facteurs latents estimés via la méthode « LFMM »,** chez les non-salariés de la MSA (2006-2016)

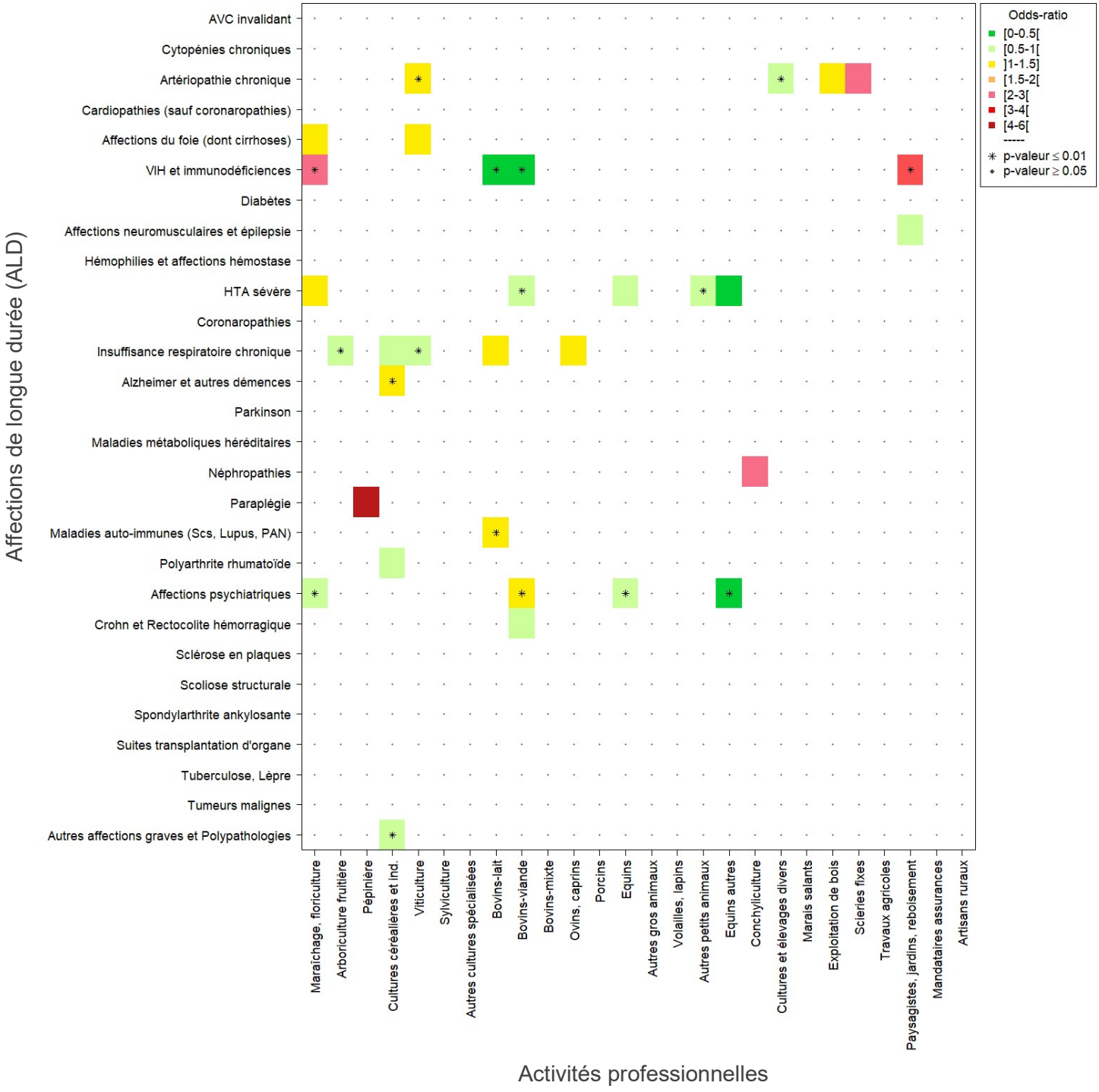


Figure 29 : Représentation graphique des odds ratios obtenus via la régression logistique entre chaque combinaison d'ALD et d'activité professionnelle en ajustant sur d'autres paramètres **notamment des facteurs latents estimés via la méthode « LFMM »,** chez les non-salariés de la MSA (2006-2016)

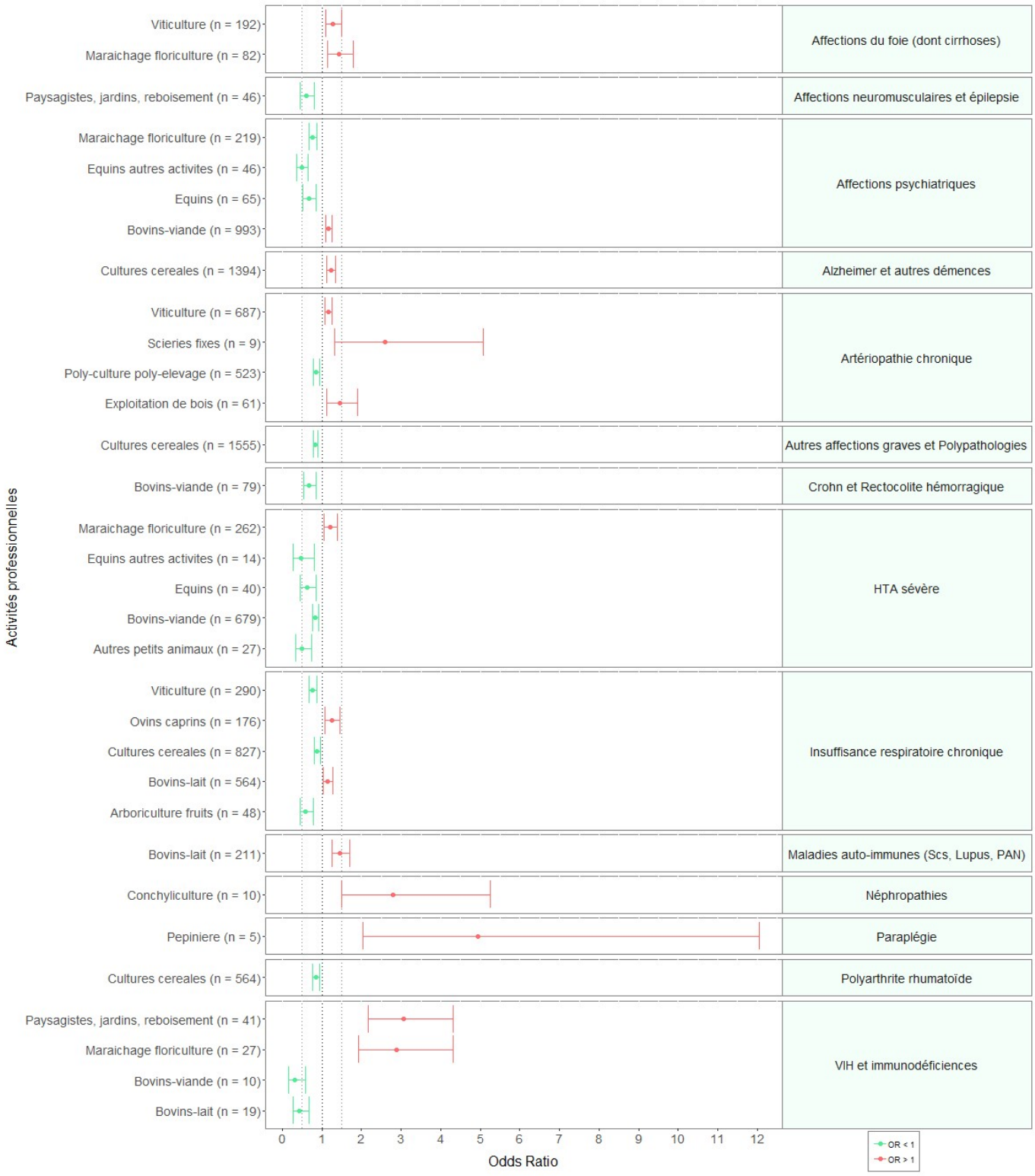


Figure 30 : Représentation graphique des odds ratios et de leurs intervalles de confiance à 95% obtenus via la régression logistique pour chaque association statistiquement significative (p-valeur$_{corrigée}$ < 0.05) entre une ALD et une activité professionnelle en ajustant sur d'autres paramètres **notamment des facteurs latents estimés via la méthode « LFMM »,** chez les non-salariés de la MSA (2006-2016)

Tableau 12 : Sélection de variables et mesure de l'AUC (échantillon de validation : 30% données) pour chaque ALD qui ont été utilisées lors de la régression logistique **avec les facteurs latents estimés via « LFMM »**, effectuée sur les non-salariés de la MSA (2006-2016)

Affections de longue durée (ALD)	Effectifs	Activités prof.	Année d' installation	Sexe	RSA	Chômage	Superficie d' exploitation	Âge	Nb. d' activités prof.	Nb. d' années obs	« Revenus »	ALD avant obs.	« Marié »	« Veuf »	« Séparé-divorcé »	Nb. de salariés	Régime maladie	Statut conjoint	Type d' exploitation	Région	Facteur latent 1	Facteur latent 2	Facteur latent 3	Facteur latent 4	AUC (échantillon de validation)
AVC invalidant	5288*	X	X					X				X			X		X				X	X	X	X	0,842
Cytopénies chroniques	284	X						X				X					X				X				0,738
Artériopathie chronique	4942	X	X	X				X				X					X				X	X	X	X	0,877
Cardiopathies (sauf coronaropathies)	16977	X																					X		1
Affections du foie (dont cirrhoses)	1320	X	X	X	X			X				X			X		X		X		X	X	X		0,776
VIH et Immunodéficiences	241	X						X					X				X								0,809
Diabètes	25229	X																			X	X			1
Affections neuromusculaires et épilepsie	1927	X	X									X					X				X		X		0,708
Hémophilies et affections hémostase	257	X															X						X		0,591
HTA sévère	5678	X	X	X				X			X	X					X		X	X	X	X	X		0,938
Coronaropathies	13210	X																						X	1
Insuffisance respiratoire chronique	3122	X	X					X			X	X					X			X	X	X	X	X	0,854
Alzheimer et autres démences	2550	X	X	X			X	X		X		X		X			X				X		X		0,943
Parkinson	1686	X	X					X		X			X				X				X		X		0,845
Maladies métaboliques héréditaires	968	X	X	X													X			X	X		X		0,694
Néphropathies	1676	X	X					X				X					X				X	X	X	X	0,81
Paraplégie	176	X															X								0,581
Maladies auto-immunes (ScS, Lupus, PAN)	969	X	X	X				X				X					X				X				0,79
Polyarthrite rhumatoïde	2601	X	X	X				X					X				X				X		X	X	0,746
Affections psychiatriques	6438	X	X	X	X			X			X	X				X	X		X	X	X	X	X	X	0,748
Crohn et Rectocolite hémorragique	930	X						X		X		X					X								0,645
Sclérose en plaques	410	X		X				X		X							X								0,706
Scoliose structurale	87	X		X													X								0,763
Spondylarthrite ankylosante	1048	X		X				X		X		X					X								0,682
Suites de transplantation d'organe	197	X		X				X				X					X				X		X		0,811
Tuberculose, Lèpre	52	X																							0,566
Tumeurs malignes	25934	X																			X	X			1
Autres affections graves et Polypathologies	5741	X	X	X				X			X	X					X			X	X	X	X	X	0,809

En ce qui concerne les variables sélectionnées pour chaque ALD lors de la régression logistique en ajoutant les facteurs latents estimés par la méthode LFMM, il est possible de remarquer que certaines variables ont été davantage incluses dans certains modèles que d'autres (Tableau 12) : le Régime maladie (inclus pour 23/28 ALD), l'Âge (inclus pour 18/28 ALD), l'ALD avant la période d'observation (incluse pour 15/28 ALD), l'Année d'installation (incluse pour 14/28 ALD) et enfin, le Sexe (inclus pour 13/28 ALD). Il en va de même pour les facteurs latents : facteur latent n°1 (inclus pour 18/28 ALD), facteur latent n°2 (inclus pour 10/28 ALD), facteur latent n°3 (inclus pour 16/28 ALD), facteur latent n°4 (inclus pour 8/28 ALD). De plus, l'ensemble des quatre facteurs latents ont été inclus pour cinq des ALD : « AVC invalidant », « Artériopathie chronique », « Insuffisance respiratoire chronique », « Néphropathies » et « Affections psychiatriques ». Ces résultats montrent que cinq variables ont un rôle plus important à jouer dans l'explication générale des ALD et que deux des facteurs latents (1 et 3) contribuent également davantage à l'explication des ALD étudiées.

Concernant les mesures de robustesse de ces modèles, leurs aires sous la courbe ROC (AUC) montrent en moyenne un bon pouvoir discriminant (AUC_{moyen} = 0.79), meilleur que sans l'ajout des facteurs latents estimés via la méthode LFMM. De plus, il est à noter qu'on obtient des AUC dont les valeurs sont à 1. Ces valeurs correspondent aux ALD pour lesquelles aucune association n'a été mise en évidence. En outre, de la même façon que les modèles précédents, les mesures de spécificité ($Spécificité_{moyenne}$ = 1), sensibilité ($Sensibilité_{moyenne}$ = 0) et F1 score ($F1_{moyen}$ = 0) montrent une fois de plus que les modèles arrivent bien à distinguer les « non malades » mais qu'ils ne sont toujours pas performants pour identifier les rares « malades » au sein de la population des non-salariés étudiés.

b. Ajout des comorbidités

De nouveau, des associations entre ALD et activités professionnelles chez les non-salariés de la MSA au cours de la période d'observation ont pu être mises en évidence via la régression logistique en ajoutant les comorbidités (composantes principales de l'ACP des ALD dont la survenue est antérieure à la maladie étudiée). La Figure 31 représente une nouvelle fois les p-valeurs de chaque association testée, corrigées par la procédure de Benjamini-Hochberg et transformées à l'aide du logarithme décimal ($-\log_{10}$). Comparé à la Figure 22 représentant les p-valeurs des associations mises en évidence dans la partie précédente de ce travail, en ayant ajouté les comorbidités, les p-valeurs des associations mises en évidence sont similaires. De la même façon, sans l'ajout des comorbidités, 54 associations statistiquement significatives ($p\text{-valeur}_{corrigée} < 0.05$) entre une ALD et une activité professionnelle ont été mises en évidence.

En complément, une heatmap a aussi été réalisée pour permettre la visualisation des associations statistiquement significatives (Figure 32). Parmi les associations précédemment mises en évidence et discutées sans l'ajout des comorbidités, la plupart sont maintenues avec des odds ratios et p-valeurs similaires. **Néanmoins, trois associations statistiquement significatives ont disparu et trois nouvelles sont apparues pour lesquelles on trouve :**

- un risque plus élevé de déclaration d'ALD « Néphropathies » chez les conchyliculteurs (OR = 2.78 [1.48 ;5.21] ; $p\text{-valeur}_{corrigée}$ = 0.03 ; n = 10), qui a également été mis en évidence dans l'analyse précédente avec l'ajout des facteurs latents estimés via LFMM ;

- et un risque plus élevé de déclaration d'ALD « Autres affections graves et Polypathologies » chez les éleveurs ovins et caprins (OR = 1.19 [1.06 ;1.34] ; $p\text{-valeur}_{corrigée}$ = 0.04 ; n = 319).

Par ailleurs, une des associations semble s'être davantage renforcée du fait d'un effet seuil (changement de catégorie de l'OR) par rapport à la valeur de l'OR lui-même. Il s'agit du risque plus important de déclaration d'ALD « Coronaropathies » chez les mandataires d'assurance (OR = 2.08 [1.27 ;3.51] ; $p\text{-valeur}_{corrigée}$ = 0.03 ; n = 15).

Pour ce qui est des trois associations qui ont disparu, elles concernent deux ALD : l'ALD « HTA sévère » et l'ALD « Coronaropathies ». Pour deux de ces associations, il s'agit principalement d'un effet seuil concernant la p-valeur corrigée. Cependant, ce n'est pas le cas de la troisième association, précédemment mise en évidence, montrant un risque plus élevé d'ALD « HTA sévère » dans les entreprises de travaux agricoles (OR = 1.51 [1.20 ;1.90] ; $p\text{-valeur}_{corrigée}$ = 0.003 ; n = 78) (Figure 23), À présent, en ajoutant les comorbidités, la p-valeur de cette

association est située au-dessus du seuil de significativité choisi avec, toutefois un OR supérieur à 1 (OR = 1.30 [1.03 ;1.64] ; p-valeur$_{corrigée}$ = 0.09 ; n = 78).

En ce qui concerne les variables sélectionnées pour chaque ALD lors de la régression logistique en ajoutant les comorbidités, il est possible de remarquer que certaines variables ont été davantage incluses dans certains modèles que d'autres (Tableau 13) : le Régime maladie (inclus pour 27/28 ALD), l'Âge (inclus pour 22/28 ALD), l'ALD avant la période d'observation (incluse pour 19/28 ALD), l'Année d'installation (incluse pour 18/28 ALD) et enfin, le Sexe (inclus pour 14/28 ALD). Quant aux comorbidités, les composantes principales (CP) de l'ACP étaient incorporées ou non aux modèles : CP n°1 (incluse pour 20/28 ALD), CP n°2 (incluse pour 15/28 ALD), CP n°3 (incluse pour 13/28 ALD), CP n°4 (incluse pour 8/28 ALD). De plus, l'ensemble des quatre CP ont été incluses pour sept des ALD : « AVC invalidant », « Artériopathie chronique », « Cardiopathies (sauf coronaropathies) », « Diabètes », « Coronaropathies », « Insuffisance respiratoire chronique », et « Néphropathies ». Ces résultats montrent que cinq variables ont un rôle plus important à jouer dans l'explication générale des ALD et que les deux premières CP contribuent davantage à l'explication des ALD étudiées.

Concernant les mesures de robustesse de ces modèles, leurs aires sous la courbe ROC (AUC) montrent également en moyenne un bon pouvoir discriminant (AUC$_{moyen}$ = 0.76), légèrement meilleur que sans l'ajout des comorbidités. En outre, de la même façon que les modèles précédents, les mesures de spécificité (Spécificité$_{moyenne}$ = 1), sensibilité (Sensibilité$_{moyenne}$ = 0) et F1 score (F1$_{moyen}$ = 0) montrent une fois de plus que les modèles arrivent bien à distinguer les « non malades » mais qu'ils ne sont toujours pas performants pour identifier les « malades ».

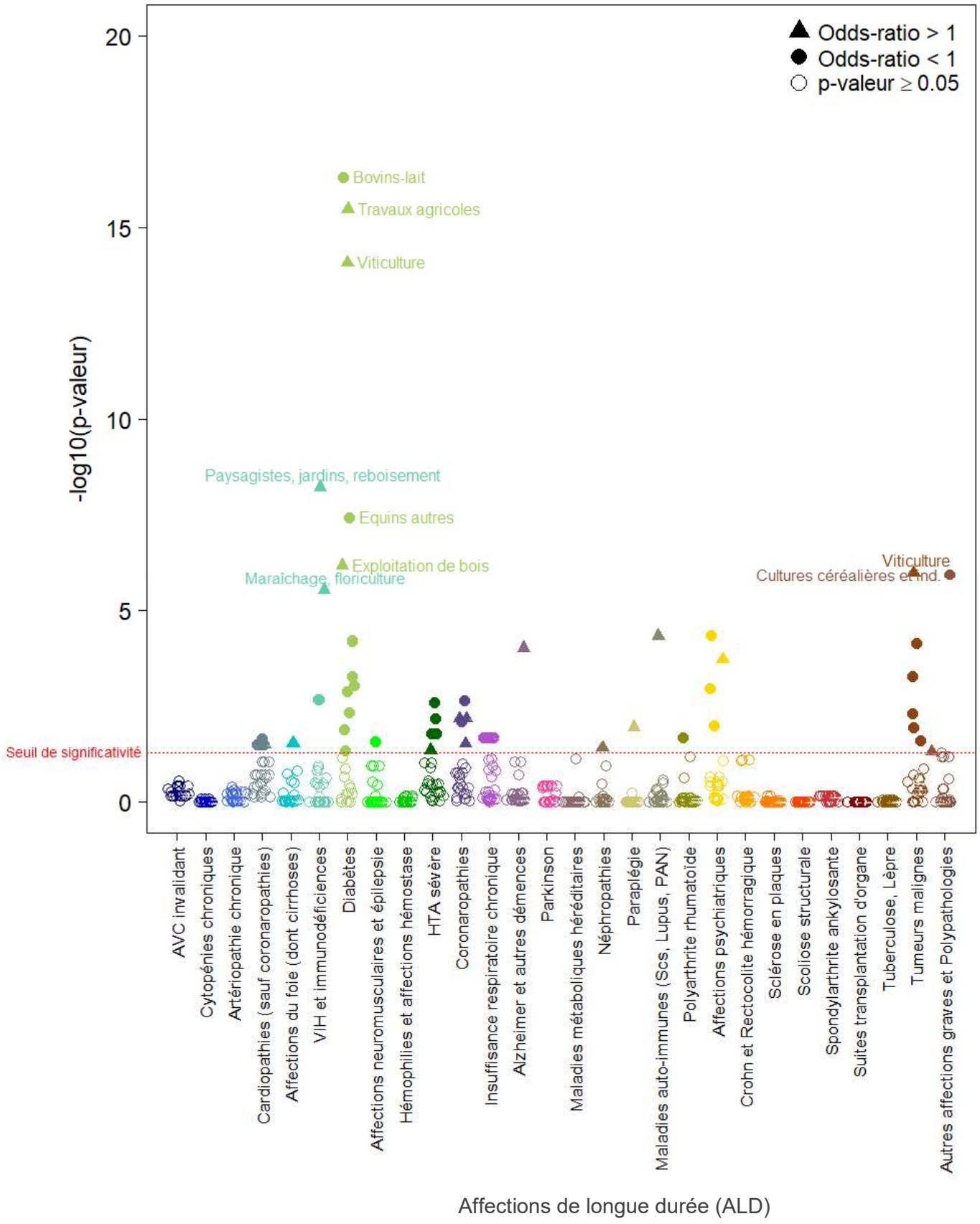


Figure 31 : Représentation graphique des p-valeurs, corrigées par la procédure de Benjamini-Hochberg, obtenues via la régression logistique entre chaque combinaison d'ALD et d'activité professionnelle en ajustant sur d'autres paramètres **notamment les comorbidités,** chez les non-salariés de la MSA (2006-2016)

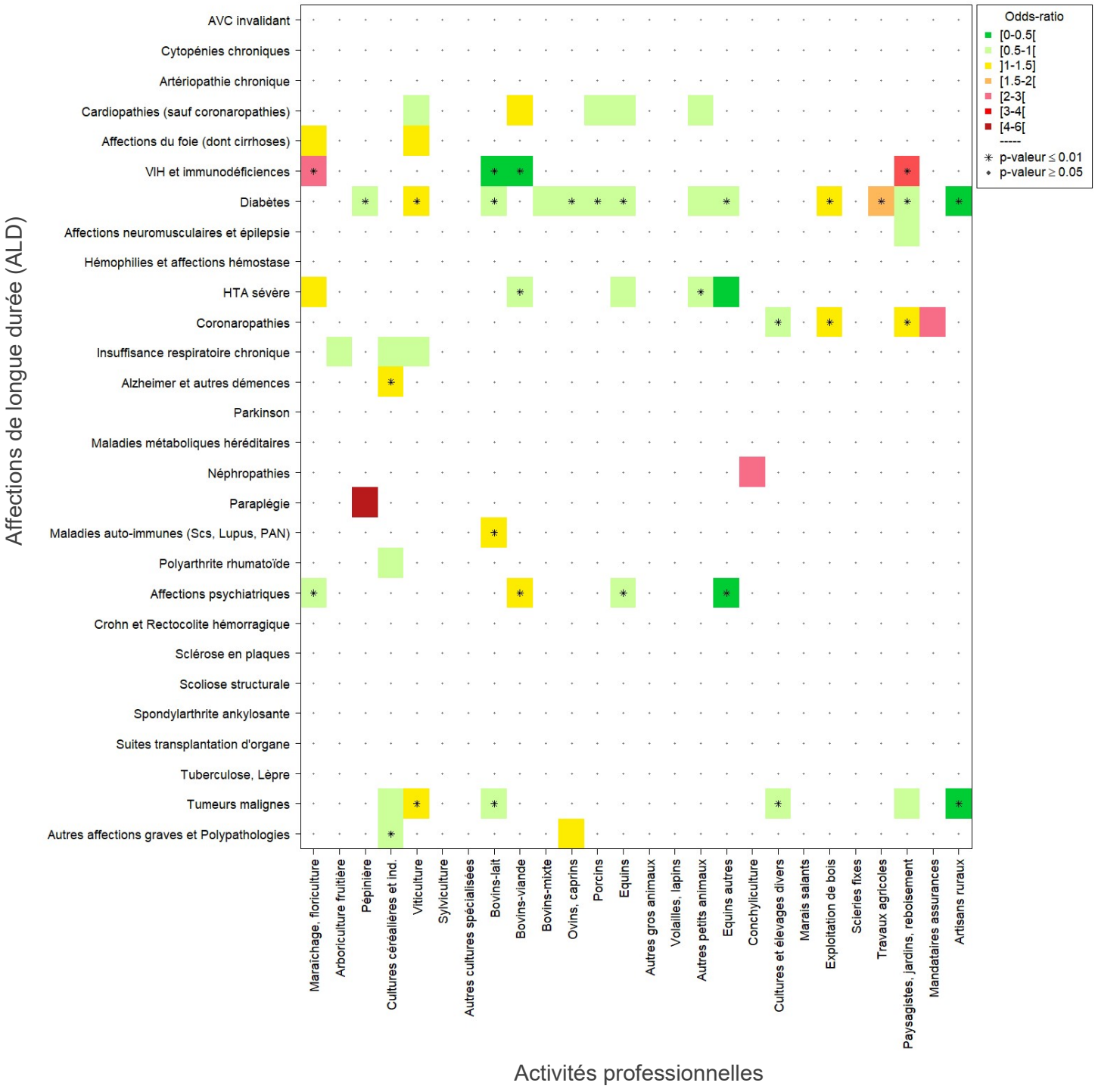


Figure 32 : Représentation graphique des odds ratios obtenus via la régression logistique entre chaque combinaison d'ALD et d'activité professionnelle en ajustant sur d'autres paramètres **notamment les comorbidités,** chez les non-salariés de la MSA (2006-2016)

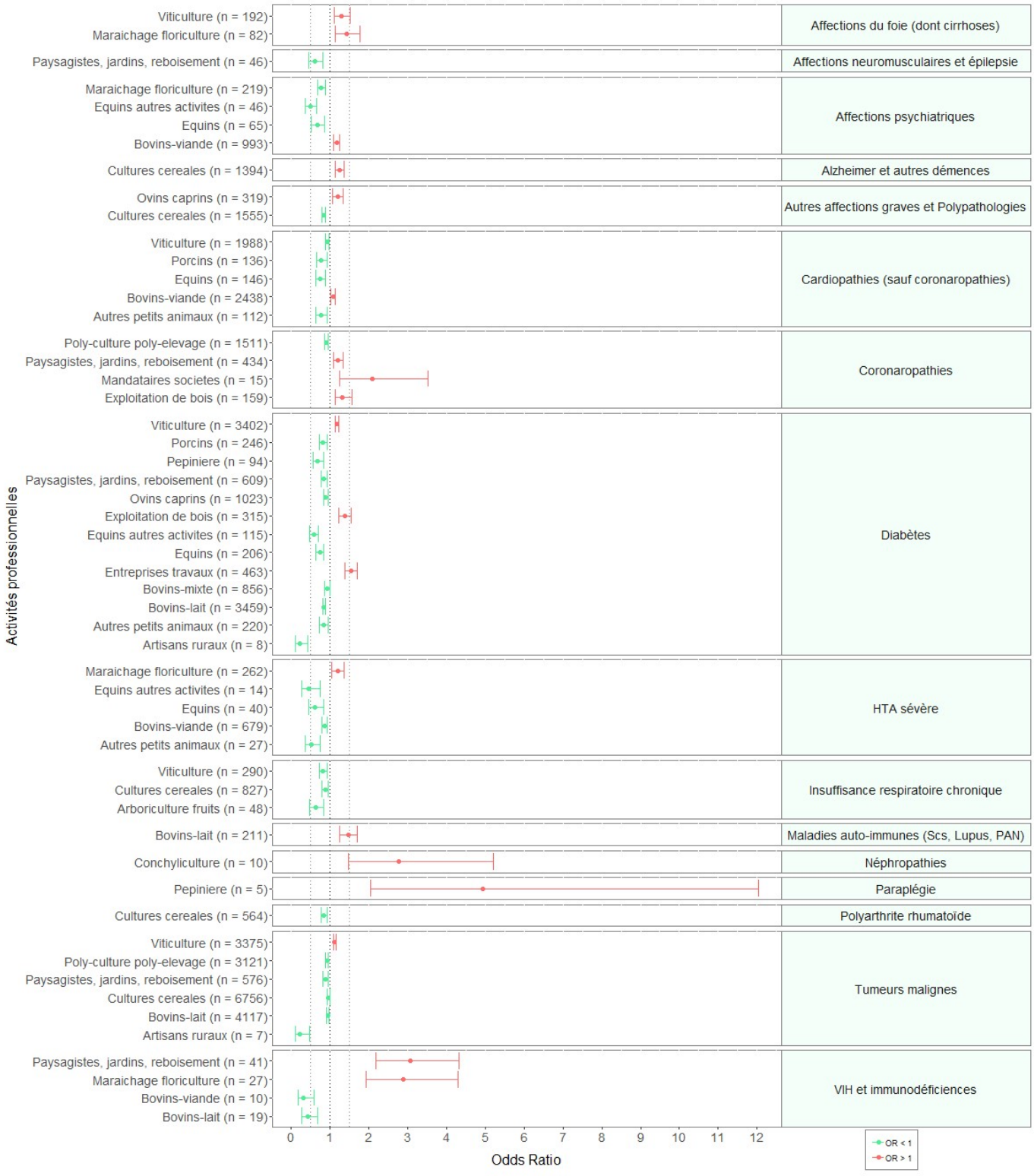


Figure 33 : Représentation graphique des odds ratios et de leurs intervalles de confiance à 95% obtenus via la régression logistique pour chaque association statistiquement significative (p-valeur$_{corrigée}$ < 0.05) entre une ALD et une activité professionnelle en ajustant sur d'autres paramètres **notamment les comorbidités,** chez les non-salariés de la MSA (2006-2016)

Tableau 13 : Sélection de variables et mesure de l'AUC (échantillon de validation : 30% données) pour chaque ALD qui ont été utilisées lors de la régression logistique **avec les composantes principales de l'ACP des comorbidités**, effectuée sur les non-salariés de la MSA (2006-2016)

Affections de longue durée (ALD)	Effectifs	Activités prof.	Année d' installation	Sexe	RSA	Chômage	Superficie d' exploitation	Âge	Nb. d' activités prof	Nb. d' années obs	« Revenus »	ALD avant obs.	« Marié »	« Veuf »	« Séparé- divorcé »	Nb. de salariés	Régime maladie	Statut conjoint	Type d' exploitation	Région	Comorbidité CP 1	Comorbidité CP 2	Comorbidité CP 3	Comorbidité CP 4	AUC (échantillon de validation)
AVC invalidant	5288*	X	X					X				X			X		X				X	X	X	X	0,812
Cytopénies chroniques	284	X						X									X				X				0,732
Artériopathie chronique	4942	X	X	X			X	X				X			X	X	X				X	X	X	X	0,847
Cardiopathies (sauf coronaropathies)	16977	X	X	X				X			X	X		X			X		X		X	X	X	X	0,859
Affections du foie (dont cirrhoses)	1320	X	X	X	X			X				X			X		X		X		X	X	X		0,761
VIH et Immunodéficiences	241	X						X					X				X								0,809
Diabètes	25229	X	X	X	X	X		X	X		X	X					X		X	X	X	X	X	X	0,804
Affections neuromusculaires et épilepsie	1927	X	X		X							X					X				X		X		0,693
Hémophilies et affections hémostase	257	X															X				X				0,59
HTA sévère	5678	X	X					X			X	X					X		X	X	X	X	X		0,889
Coronaropathies	13210	X	X	X				X				X	X	X	X		X			X	X	X	X	X	0,826
Insuffisance respiratoire chronique	3122	X	X		X			X			X	X				X	X			X	X	X	X	X	0,833
Alzheimer et autres démences	2550	X	X	X			X	X		X		X		X			X			X	X	X	X		0,942
Parkinson	1686	X	X	X				X		X			X				X				X				0,846
Maladies métaboliques héréditaires	968	X	X									X					X			X	X	X			0,694
Néphropathies	1676	X	X					X				X					X				X	X	X	X	0,802
Paraplégie	176	X															X								0,581
Maladies auto-immunes (ScS, Lupus, PAN)	969	X	X	X				X				X					X				X				0,785
Polyarthrite rhumatoïde	2601	X	X	X				X				X	X				X					X			0,741
Affections psychiatriques	6438	X	X	X	X			X			X	X			X		X		X	X	X	X	X		0,724
Crohn et Rectocolite hémorragique	930	X						X		X		X					X			X					0,651
Sclérose en plaques	410	X		X				X		X							X								0,706
Scoliose structurale	87	X		X													X								0,763
Spondylarthrite ankylosante	1048	X						X		X		X	X				X								0,682
Suites de transplantation d'organe	197	X		X				X				X					X				X				0,807
Tuberculose, Lèpre	52	X																							0,566
Tumeurs malignes	25934	X	X					X					X		X		X			X	X	X		X	0,776
Autres affections graves et Polypathologies	5741	X	X	X				X			X	X					X			X	X	X	X		0,782

c. Comparaisons entre les modèles

Les résultats obtenus avec les deux méthodologies pour l'estimation de facteurs de confusion ont été comparés avec ceux obtenus via la régression logistique dans la partie précédente de ce travail de thèse. Il est possible de remarquer que ceux obtenus pour la régression logistique effectuée précédemment et ceux obtenus avec l'ajout des comorbidités sont davantage similaires. En ce qui concerne les aires sous la courbe ROC (AUC) calculées sur les échantillons de validation (30% des données), on peut remarquer qu'elles sont en très légère hausse, que l'on ajoute les facteurs latents estimés via la méthodologie LFMM ou que l'on ajoute les comorbidités aux modèles de régression logistique (Figure 34).

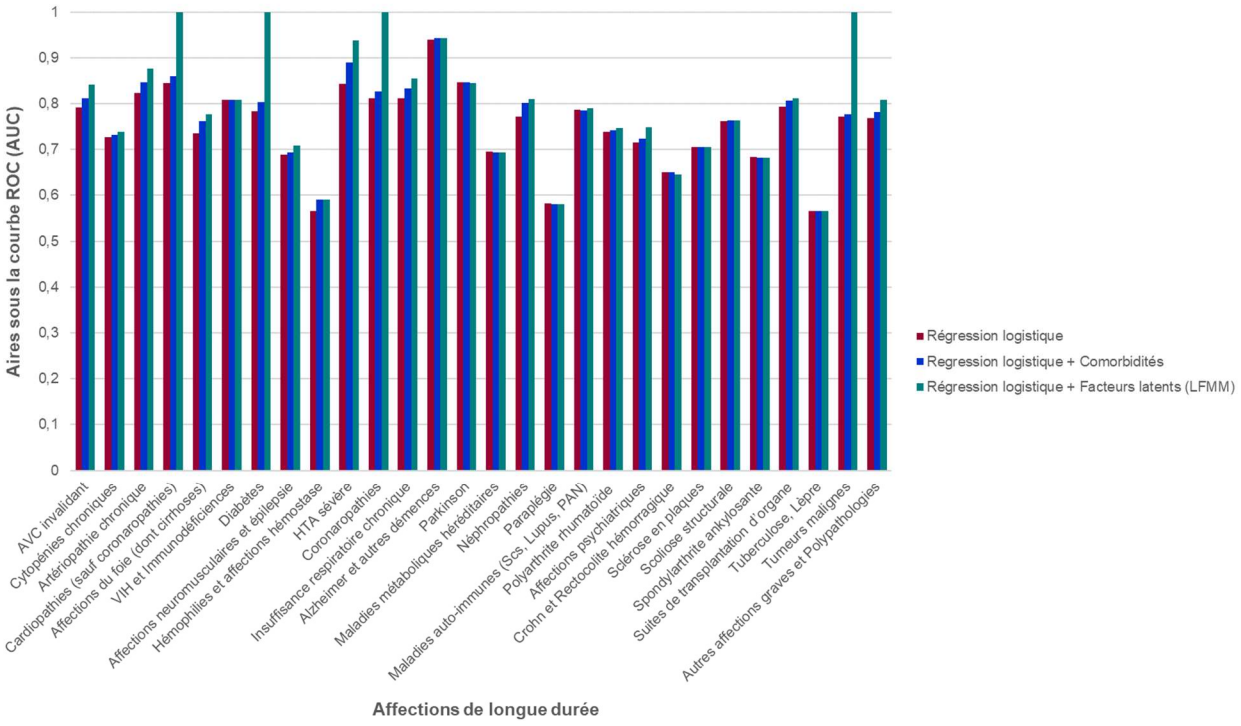


Figure 34 : Comparaison des aires sous la courbe ROC (AUC) calculés sur les échantillons de validation (30% des données) pour chaque méthodologie employée et pour chaque modèle défini par ALD, sur les données des non-salariés de la MSA (2006-2016)

III. Discussion

Cette partie du travail de thèse consiste à utiliser des méthodologies permettant l'estimation de facteurs de confusion ou facteurs latents. En effet, dans la partie précédente, il a été observé que les modèles de régression logistique tels qu'utilisés ne permettent pas d'expliquer suffisamment les ALD et qu'il peut exister des facteurs de confusion non pris en compte dans les analyses. Ne disposant pas de davantage de variables dans les données de la MSA, il a été nécessaire de rechercher des méthodologies statistiques permettant l'estimation de ces facteurs non observés. Ainsi, d'une part, des facteurs latents ont été estimés via la méthode LFMM et ajoutés aux modèles et d'autre part, les comorbidités ont été ajoutées en tant que variables d'ajustement aux modèles.

a. Forces et limites des méthodes utilisées

Concernant l'utilisation de la méthode LFMM, le principal avantage repose sur son estimation de facteurs latents, non seulement à partir de la matrice des ALD mais aussi à partir de la matrice des variables explicatives (123,124). Cette méthode permet alors une estimation de facteurs non observés qui ne sont pas corrélés aux variables explicatives. Cependant, il y a deux inconvénients majeurs à son utilisation. En effet, cette méthode utilise la régression linéaire qui est particulièrement adaptée à des variables à expliquer quantitatives. Or, nous disposons uniquement de variables à expliquer qualitatives dans le cadre de ce travail : les ALD (variables binaires). De plus, bien que cette méthode soit adaptée à des données de grande dimension, elle s'applique davantage à des données où le nombre de variables (colonnes) est bien plus important que le nombre d'observations (lignes). Toutefois, dans le cadre de ce travail, la configuration des matrices est telle que le nombre de variables est bien moins important (inférieur à 100) que le nombre de lignes correspondant au nombre d'individus (n = 899 212 non-salariés). Ces inconvénients peuvent laisser penser que la méthode LFMM peut introduire des biais difficiles à quantifier dans l'estimation des facteurs de confusion à partir de nos données.

Concernant l'ajout des comorbidités, le principal avantage est d'ajouter aux modèles des variables explicatives connues et disponibles au sein des données MSA, ce qui permet d'approcher de potentiels facteurs de confusion. Il a notamment été vérifié qu'il n'y avait pas de difficultés liées aux corrélations trop importantes entre les ALD ou de problèmes de colinéarité dans les modèles. Cependant, les ALD étant nombreuses, il a été nécessaire de trouver une méthode statistique permettant de réduire le nombre de variables à ajouter aux

modèles tout en limitant la perte potentielle de puissance statistique. L'analyse en composantes principales a ainsi été utilisée à cette fin. Néanmoins, l'utilisation d'une telle méthode ajoute une incertitude quand il s'agit de comprendre les sélections de variables à ajouter dans les modèles pour chaque ALD. En effet, même si les composantes principales ont été calculées à partir de la matrice des ALD antérieures, chacune d'entre elles regroupe une partie des ALD sans qu'on sache vraiment les identifier indépendamment.

Par ailleurs, au cours de l'utilisation des deux méthodes, il a été choisi de ne sélectionner que quatre facteurs latents (LFMM) et quatre composantes principales (comorbidités). Ce choix a été pris en fonction de l'ACP réalisée sur l'ensemble de la matrice des ALD pour identifier des regroupements d'individus qui se « ressemblent ». Au final, il a été nécessaire de trouver un équilibre entre ajouter d'un nombre trop important de variables et éviter la perte trop importante d'information. La Figure 26 a ainsi permis de visualiser qu'il y a principalement quatre axes ou « groupes » d'individus, expliquant plus de 60% de la variance et donc de l'information apportée par la matrice des ALD. Choisir davantage d'« axes » et donc de variables à ajouter au modèle aurait pu engendrer une perte de puissance statistique. A contrario, en choisir moins aurait potentiellement pu engendrer une perte trop importante d'informations. Des analyses complémentaires ont donc été réalisées en choisissant moins d'axes et les mesures de robustesse des modèles montrent que ces derniers sont moins performants pour certaines ALD.

b. Synthèse des principaux résultats

Quant aux associations statistiquement significatives mises en évidence, on peut remarquer quelques variations entre la régression logistique sans l'ajout de facteurs de confusion (54 associations) et avec l'ajout de facteurs de confusion, estimés par LFMM (32 associations) ou approchés par l'ajout des comorbidités (54 associations).

En ajoutant les facteurs latents estimés via LFMM par la régression logistique, un grand nombre d'associations statistiquement significatives disparaissent, notamment pour 4 des 28 ALD : « Cardiopathies (sauf coronaropathies) », « Diabètes », « Coronaropathies » et « Tumeurs malignes ». D'ailleurs, il s'agit des ALD ayant les effectifs de déclarations les plus importants (cf. Tableau 8). Une analyse complémentaire des associations entre ces ALD et les facteurs latents, ainsi que les mesures des AUC des modèles de ces ALD (AUC = 1) a montré que les facteurs latents expliquaient davantage les ALD concernées que les activités professionnelles. D'ailleurs, la sélection de variables pour ces ALD montre que seuls des

facteurs latents sont ajoutés aux modèles et qu'à eux seuls, ils suffisent à expliquer la variabilité de ces ALD. On peut alors faire l'hypothèse que soit les associations précédemment mises en évidence ne sont pas attribuables aux activités professionnelles mais uniquement aux facteurs latents lorsqu'ils sont estimés via la méthode LFMM, soit que les résultats sont influencés par des biais liés à l'utilisation de modèles linéaires LFMM pour l'estimation de facteurs latents.

Parmi les nouvelles associations mises en évidence avec l'ajout des facteurs latents estimés par la méthode LFMM, deux associations concernent des risques plus élevés de déclarations d'ALD « Insuffisance respiratoire chronique grave » pour les éleveurs bovins (lait) et les éleveurs ovins et caprins. Or, ces associations sont particulièrement intéressantes car dans la littérature, un risque accru de bronchopneumopathie chronique obstructive (BPCO) et de pneumopathie d'hypersensibilité (128–131) a déjà été décrit au sein de populations d'éleveurs (pour les pneumopathies d'hypersensibilité, en lien principalement avec l'utilisation de fourrage l'hiver, et la contamination de ce dernier par des micro-organismes fongiques). Nous n'avons pas obtenu d'associations statistiquement significatives avec la régression logistique sans l'ajout de facteurs de confusion et avec l'ajout des comorbidités mais les p-valeurs concernées étaient proches de la limite de la significativité.

Des associations montrant des risques plus élevés de déclaration d'ALD « Artériopathie chronique » (principalement, des artériopathies des membres inférieurs) ont été mises en évidence pour les viticulteurs, les travailleurs agricoles des exploitations de bois et ceux des scieries fixes, en ajoutant les facteurs latents estimés via la méthode LFMM. Cela peut être lié notamment à une consommation de tabac, à un comportement davantage sédentaire, comparé aux autres non-salariés, ou encore à des pathologies antérieures telles que le diabète (132). Or, contrairement aux autres méthodologies qui montrent des associations avec des risques plus élevés d'ALD « Diabètes » chez les viticulteurs et les travailleurs agricoles des exploitations de bois, cette méthodologie ne permet pas de mettre en évidence d'associations entre ces secteurs d'activités et l'ALD « Diabètes ». Cela est certainement dû aux biais liés à l'utilisation de la méthode LFMM pour l'estimation de facteurs latents. La mise en évidence de ces associations questionne d'autant plus l'utilisation de cette méthodologie (LFMM) pour l'estimation de facteurs de confusion.

Par ailleurs, le risque plus élevé de déclaration d'ALD « Néphropathies » chez les conchyliculteurs ressort avec les deux méthodologies employées pour l'estimation des facteurs latents. Sans l'ajout de facteurs latents, il est important de souligner que l'association est déjà à la limite de la significativité (OR = 2.67 [1.43 ;5.00] ; p-valeur$_{corrigée}$ = 0.05 ; n = 10). Cette association semble particulièrement intéressante car les conchyliculteurs sont

davantage consommateurs de coquillages, notamment à certaines périodes de l'année. Or, ces coquillages sont une source d'apport de métaux lourds dont le cadmium principalement, mais aussi de plomb et de mercure, toxiques pour les reins (133–135). D'ailleurs, la consommation de coquillages entraine également une exposition à bas bruit à d'autres dérivés métalliques ou métalloïdes. Autant la question de l'implication de l'arsenic dans des pathologies rénales a déjà été posée (136,137), autant celle des organo-étains (issues des peintures « antifouling » des coques de bateaux) et de l'étain en général n'est aujourd'hui, à notre connaissance, pas abordée dans la littérature. Par ailleurs, la consommation de coquillages est aussi associée à une charge sodée (supérieure à la chair des poissons) qui peut intervenir dans l'hypertension, également facteur de risque d'insuffisance rénale (138). Cependant, les trois dernières méthodes de détection d'associations n'ont pas permis de mettre en évidence un lien entre la conchyliculture et l'ALD « Hypertension artérielle sévère ». Cela peut être expliqué soit par le fait que cette affection ne fait plus partie de la liste des ALD depuis 2011 (139) et que les données d'ALD mises à disposition par la MSA s'étendent de 2012 à 2016 (seulement 12 conchyliculteurs avec une déclaration d'ALD « Hypertension artérielle sévère » sur la période d'observation), soit par le fait que la pathologie rénale dans cette population est majoritairement liée à autre chose. Par ailleurs, une analyse de la consommation médicamenteuse de ces individus pourrait peut-être permettre d'avancer dans l'exploration de cette association.

En ce qui concerne l'association mise en évidence suite à l'ajout des comorbidités aux modèles, entre l'ALD « Autres affections graves et Polypathologies » et l'élevage d'ovins et de caprins, il n'est pour le moment pas possible de l'expliciter. En effet, cette « ALD » comprend un certain nombre de pathologies chroniques (n = 486 codes de la CIM-10 pour cette ALD, cf. Figure 16) aux étiologies très différentes. Ce type d'association est à investiguer à un niveau de précision plus fin, c'est-à-dire, en utilisant le codage de la CIM-10.

Par ailleurs, la disparition de l'association entre l'ALD « HTA sévère » et les entreprises de travaux agricoles pose question. En ajustant sur les comorbidités, la p-valeur de cette association passe au-dessus de la limite de significativité ($p\text{-valeur}_{corrigée} = 0.09$) bien que la borne inférieure de l'intervalle de confiance de l'OR reste supérieure à 1 (OR = 1.30 [1.03 ;1.64]). On peut alors faire l'hypothèse qu'en ajustant sur les comorbidités, cette association disparait du fait que la variabilité de cette ALD serait davantage due aux pathologies survenues précédemment, plutôt qu'au travail dans des entreprises de travaux agricoles.

Quant aux mesures d'AUC, elles montrent que l'ajout des facteurs latents (LFMM) permet d'augmenter légèrement la performance des modèles de régression logistique. Néanmoins,

les mesures d'AUC sont égales à 1 (valeur maximale) pour 4 des 28 ALD, ce qui laisse à penser de nouveau que la méthode LFMM n'est pas forcément la plus adaptée pour estimer les facteurs de confusion à partir de nos données.

A propos de la sélection de variables et de l'ajout des facteurs de confusion aux modèles, on peut constater que les comorbidités sont plus fréquemment inclues dans les modèles que les facteurs latents (LFMM). Une analyse des corrélations entre les facteurs latents estimés via la méthode LFMM et les comorbidités a été réalisée et a montré qu'ils sont très corrélés. Ainsi, il est possible de faire l'hypothèse que l'information apportée par les facteurs de confusion serait équivalente, qu'ils soient estimés via LFMM ou approchés par l'ajout des comorbidités.

c. Conclusion

Pour toutes les raisons évoquées précédemment et au vu du fait que l'information portée par les facteurs de confusion est similaire, quelle que soit la méthodologie employée, il a été choisi de conserver le modèle de régression logistique de la partie précédente auquel on ajoute les comorbidités. Cela permet ainsi d'approcher des facteurs de confusion avec une méthodologie davantage adaptée à nos données et qui permet de faire émerger de nouvelles associations à investiguer.

L'estimation de facteurs de confusion a pour objectif de mieux expliquer les ALD et ainsi d'améliorer légèrement les mesures de robustesse. Bien que cette optimisation des modèles de régression logistique ait été nécessaire, les deux méthodologies testées n'ont pas permis d'augmenter sensiblement ces paramètres. Il est donc nécessaire d'essayer de nouvelles méthodologies pour la sélection de variables dans les modèles, afin de limiter les biais liés au faible nombre de « malades » vis-à-vis des témoins, quelle que soit l'ALD étudiée.

B. Sélection de variables via la régression pénalisée *lasso*

I. Méthodologie

Dans les parties précédentes de ce manuscrit, la sélection de variables a été réalisée « pas à pas » à l'aide du critère BIC avant d'utiliser la régression logistique. Cependant, au vu du manque de sensibilité des modèles, il a été décidé de tester une autre méthode de sélection de variables pour tenter d'améliorer ces derniers. **Pour cela, il est possible d'utiliser les méthodes de régression pénalisée afin de contrôler les coefficients de régression, c'est-à-dire, de contrôler les poids affectés à chacune des variables dans le modèle utilisé. On peut alors maitriser la complexité des modèles ainsi construits, notamment lorsque les variables sont très nombreuses** (140).

Il existe principalement deux méthodes de régression pénalisée : *ridge* et *lasso* (acronyme de « *least absolute shrinkage and selection operator »*). La régression *ridge* permet d'éviter le sur-apprentissage en regroupant les variables corrélées, au sens où des variables corrélées auront alors des coefficients similaires (141). Quant à la régression *lasso*, elle cherche à obtenir un modèle parcimonieux plus facilement interprétable, avec un nombre minimum de variables en « poussant » certains coefficients vers 0 (142). Si plusieurs variables sont corrélées entre elles, la régression *lasso* va avoir tendance à choisir une seule d'entre elles (affectant un poids de 0 aux autres), plutôt que répartir les poids équitablement comme dans la régression *ridge* (143). En pratique, la régression *ridge* donne de meilleurs résultats que lasso lorsque les variables sont corrélées entre elles mais la régression *lasso* a l'avantage de simplifier le modèle en réduisant le nombre de variables (100). En effet, la régression *lasso* utilise la norme mathématique L_1 donnée par la somme des valeurs absolues des coefficients avec la relation suivante : $\|\theta\|_{l_1} = |\theta_1| + .. + |\theta_n|$ où *n* est le nombre de variables et θ, les valeurs des « paramètres du modèle ». Cette pénalisation en valeur absolue fait de la régression lasso non seulement une méthode de pénalisation mais aussi, à l'instar de la régression *ridge*, une méthode de sélection de variables (96,100).

Ainsi, dans le cadre de ce travail, étant donné le nombre de variables à disposition et l'objectif de diminuer la complexité du modèle tout en augmentant son pouvoir discriminant, nous avons choisi de tester la régression *lasso* pour la sélection de variables à intégrer dans le modèle de régression logistique. Cette méthode a alors été appliquée aux données de la MSA telles que préparées lors de l'utilisation des autres méthodes statistiques de modélisation. De plus,

comme décidé précédemment, les comorbidités ont été ajoutées à l'ensemble des variables explicatives en tant que facteurs de confusion.

La régression pénalisée *lasso* a pu être réalisée grâce à la librairie « glmnet » conçue pour le logiciel R (144). La première étape de cette méthode consiste alors à définir le paramètre de régularisation λ qui contrôle le niveau de contrainte appliqué aux coefficients (Figure 35). Plus ce paramètre est grand, plus le nombre de coefficients nuls est important (96). Comme précédemment, le modèle a été construit sur un échantillon d'apprentissage, c'est-à-dire 70% des données, et testé ensuite sur un échantillon de validation (30% restants). Pour cette étape, il s'agit alors de fournir l'échantillon d'apprentissage à la librairie avec la variable réponse Y_i et l'ensemble des variables explicatives (âge, sexe, activités professionnelles, comorbidités…) pour chaque ALD. On calcule ensuite chaque aire sous la courbe ROC pour toute une plage de valeurs λ déterminée automatiquement et on conserve le paramètre λ pour lequel la valeur d'AUC est la meilleure (plus proche de 1). La deuxième étape consiste alors simplement à utiliser ce paramètre λ dans le modèle de régression pénalisée *lasso* afin de récupérer les coefficients de régression pour chacune des variables explicatives. Enfin, on obtient une sélection de variables incluses dans chaque modèle (un par ALD) via la valeur des coefficients des variables explicatives : si la valeur du coefficient est nulle, alors la variable n'est pas conservée dans le modèle en question.

Puis, la régression logistique a été réalisée sur l'échantillon d'apprentissage avec les variables retenues pour chaque ALD par la régression pénalisée *lasso*. Les modèles ont ensuite été évalués via le calcul des mesures de robustesse sur l'échantillon de validation (sensibilité, spécificité, AUC, F_1 score). Finalement, les résultats de la modélisation ont été obtenus en réalisant la régression logistique sur l'ensemble des données afin de calculer les odds ratios, leurs intervalles de confiance à 95% et de récupérer les p-valeurs, ensuite corrigées via la méthode de Benjamini-Hochberg comme pour les méthodologies précédentes.

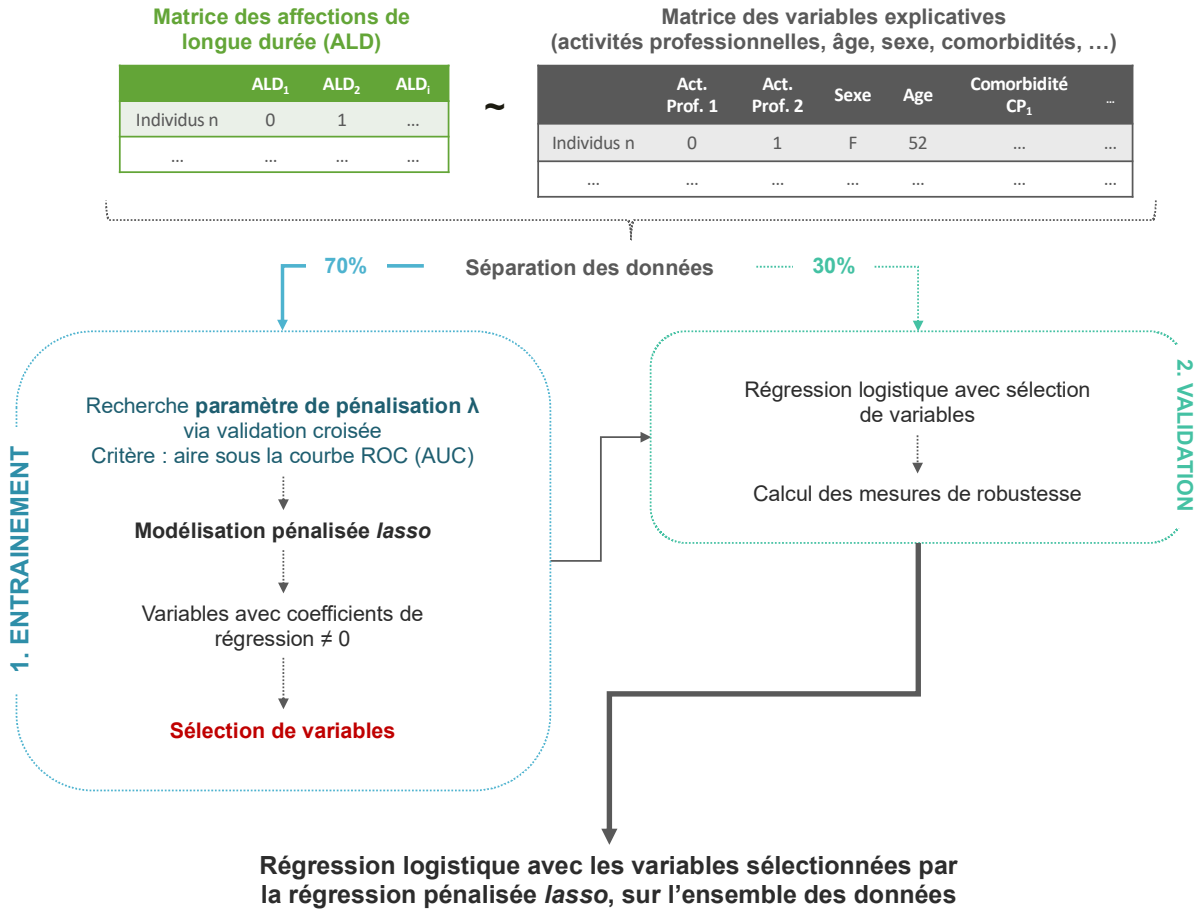


Figure 35 : Application de la régression pénalisée *lasso* aux données de la MSA dans le but de sélectionner les variables pour la régression logistique

II. Résultats

A nouveau, des associations entre ALD et activités professionnelles chez les non-salariés de la MSA au cours de la période d'observation ont pu être mises en évidence via la régression logistique en sélectionnant les variables via la régression pénalisée *lasso*. La Figure 36 représente les p-valeurs de chaque association testée, corrigées par la procédure de Benjamini-Hochberg et transformées à l'aide du logarithme décimal ($-\log_{10}$). Comparé à la Figure 31 représentant les p-valeurs des associations mises en évidence avec la méthode précédente de sélection de variables (critère BIC), les p-valeurs des associations sont d'un point de vue général similaires. **En ce qui concerne le nombre d'associations statistiquement significatives ($p\text{-valeur}_{corrigée} < 0.05$), 39 associations ont été mises en évidence avec cette méthode de sélection de variables en comparaison aux 54 associations mises en évidence précédemment.** Parmi les p-valeurs les plus « significatives », on remarque qu'il y a notamment une différence en ce qui concerne l'ALD « VIH et Immunodéficiences » pour laquelle il y a moins de p-valeurs significatives au-delà de la valeur 10^{-5}.

En effet, lorsqu'on représente sur une heatmap les associations mises en évidence avec la sélection des variables basée sur la régression lasso (Figure 37), on peut observer que **quinze associations ont disparu et qu'une association, précédemment mise en évidence avec la « Régression logistique simple » (Figure 23), est de nouveau statistiquement significative.** Pour mémoire, les associations ayant disparu concernent les ALD suivantes : « VIH et Immunodéficiences », « Cardiopathies », « Affections du foie (dont cirrhoses) », « Affections neuromusculaires et épilepsie », « HTA sévère », « Coronaropathies », « Insuffisance respiratoire chronique », « Néphropathies », « Maladies auto-immunes (ScS, Lupus, PAN) », « Polyarthrite rhumatoïde », « Tumeurs malignes » et « Autres affections graves et Polypathologies ».

Quant à l'association précédemment mise en évidence, elle montre de nouveau un risque plus élevé de déclaration d'ALD « HTA sévère » avec le secteur professionnel des travaux agricoles (OR = 1.48 [1.17 ;1.86] ; $p\text{-valeur}_{corrigée}$ = 0.008 ; n = 78) (Figure 38). Or, cette association avait précédemment été masquée en ajoutant les comorbidités avec une p-valeur proche de la limite de la significativité (régression logistique avec comorbidités : $p\text{-valeur}_{corrigée}$ = 0.09).

En ce qui concerne les variables sélectionnées pour chaque ALD avec cette méthodologie, on peut remarquer que les variables incluses dans les modèles sont plus nombreuses pour la plupart des ALD (Tableau 14, Figure 39). D'ailleurs, pour 4 des ALD,

l'ensemble des variables disponibles ont été sélectionnées pour être utilisées dans les modèles : « Cardiopathies (sauf coronaropathies) », « Affections du foie (dont cirrhoses) », « Affections psychiatriques » et « Spondylarthrite ankylosante ». Cependant, cela n'affecte que deux des associations significatives mises en évidence : « Affections du foie (dont cirrhoses) » et « Cardiopathies (sauf coronaropathies) ».

Concernant les mesures de robustesse de ces modèles, leurs aires sous la courbe ROC (AUC) montrent en moyenne aussi un bon pouvoir discriminant (AUC_{moyen} = 0.75), légèrement moins performant qu'avec la méthode de sélection de variables pas à pas (AUC_{moyen} = 0.76). Si on regarde plus précisément les AUC par ALD, elles sont de façon générale similaires à l'exception de deux ALD : « Parkinson » (AUC_{BIC}= 0.846 ; AUC_{lasso}= 0.5) et « Paraplégie » (AUC_{BIC}= 0.581 ; AUC_{lasso} = 0.5). En outre, de la même façon que les modèles précédents, les mesures de spécificité ($Spécificité_{moyenne}$ = 1), sensibilité ($Sensibilité_{moyenne}$ = 0) et F1 score ($F1_{moyen}$ = 0) montrent à nouveau que les modèles ne sont toujours pas performants pour identifier les rares « malades » au sein de la population des non-salariés étudiés.

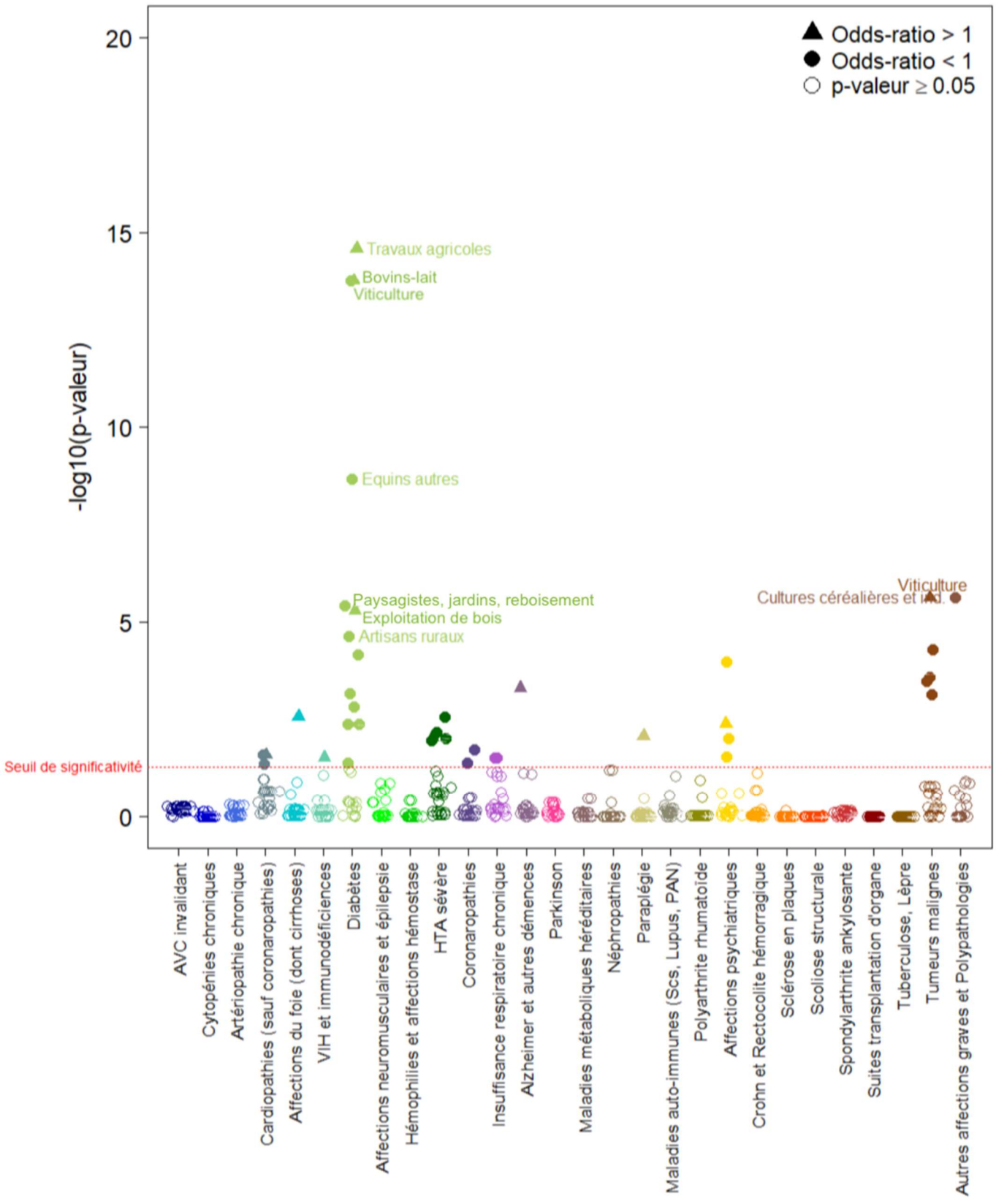


Figure 36 : Représentation graphique des p-valeurs, corrigées par la procédure de Benjamini-Hochberg, obtenues via la régression logistique entre chaque combinaison d'ALD et d'activité professionnelle en ajustant sur d'autres paramètres **sélectionnés par la méthode de régression pénalisée *lasso*,** chez les **non-salariés** de la MSA (2006-2016)

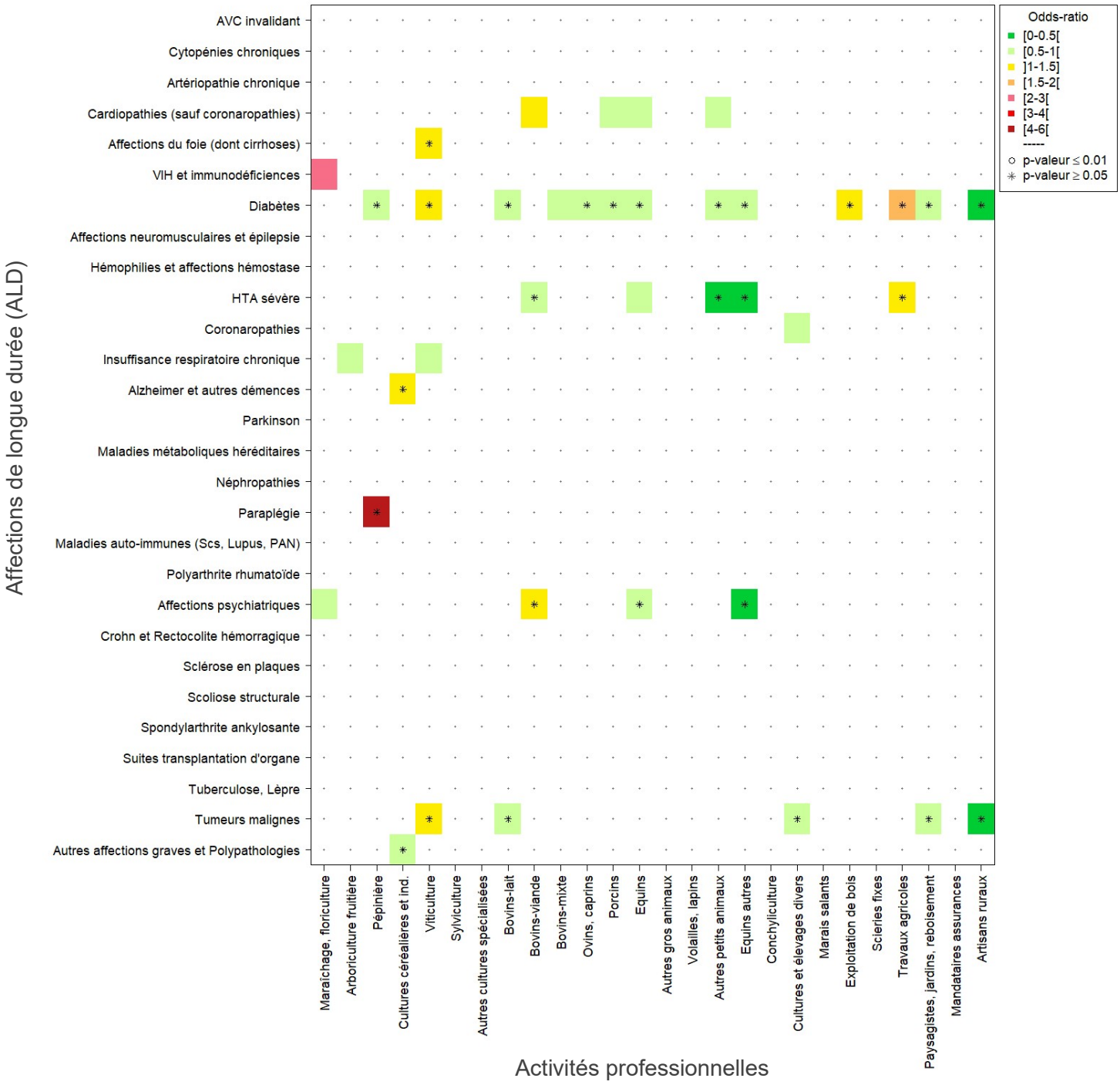


Figure 37 : Représentation graphique des odds ratios obtenus via la régression logistique entre chaque combinaison d'ALD et d'activité professionnelle en ajustant sur d'autres paramètres **sélectionnés par la méthode de régression pénalisée *lasso*,** chez les non-salariés de la MSA (2006-2016)

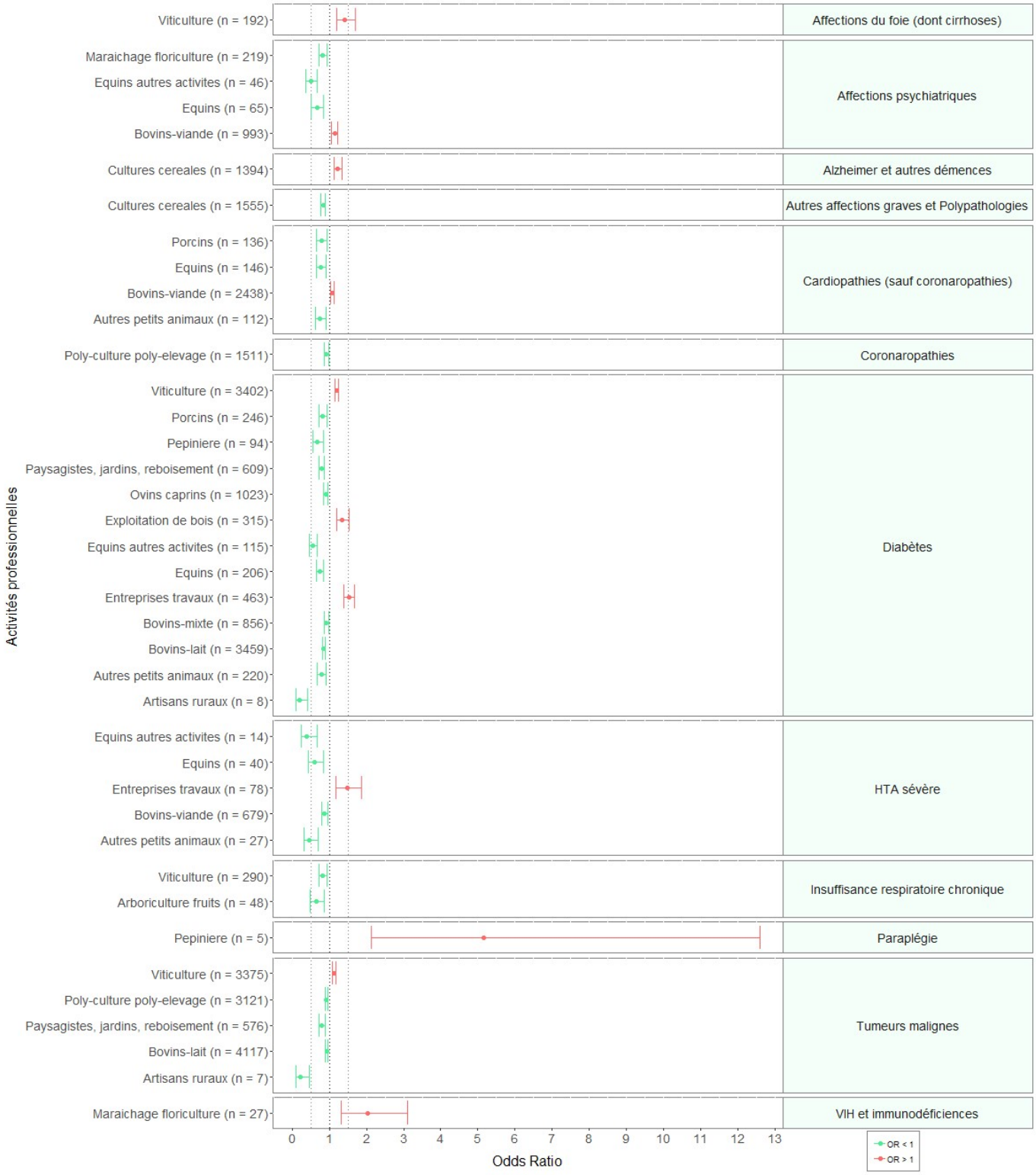


Figure 38 : Représentation graphique des odds ratios et de leurs intervalles de confiance à 95% obtenus via la régression logistique pour chaque association statistiquement significative ($p\text{-valeur}_{corrigée} < 0.05$) entre une ALD et une activité professionnelle en ajustant sur d'autres paramètres **sélectionnés par la méthode de régression pénalisée *lasso*,** chez les non-salariés de la MSA (2006-2016)

Tableau 14 : Sélection de variables et mesure de l'AUC (échantillon de validation : 30% données) pour chaque ALD qui ont été utilisées lors de la régression logistique avec une sélection de variables réalisée via **la méthode de régression pénalisée *lasso***, effectuée sur les non-salariés de la MSA (2006-2016)

Affections de longue durée (ALD)	Effectifs	Activités prof.	Année d' installation	Sexe	RSA	Chômage	Superficie d' exploitation	Age	Nb. d' activités prof.	Nb. d' années obs	« Revenus »	ALD avant obs.	« Marié »	« Veuf »	« Séparé- divorcé »	Nb. de salariés	Régime maladie	Statut conjoint	Type d' exploitation	Région	Comorbidité CP 1	Comorbidité CP 2	Comorbidité CP 3	Comorbidité CP 4	AUC (échantillon de validation)
AVC invalidant	5288*	X	X	X	X	X	X	X		X	X	X	X	X	X		X	X	X	X	X		X	X	0,812
Cytopénies chroniques	284	X	X	X			X	X		X	X	X		X		X	X			X					0,746
Artériopathie chronique	4942	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X		0,847
Cardiopathies (sauf coronaropathies)	16977	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	0,859
Affections du foie (dont cirrhoses)	1320	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	0,761
VIH et Immunodéficiences	241	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X		0,811
Diabètes	25229	X	X	X	X	X	X	X		X	X	X	X	X	X	X	X	X	X	X	X			X	0,804
Affections neuromusculaires et épilepsie	1927	X	X	X	X	X	X	X		X	X	X	X	X	X	X	X	X	X	X	X	X		X	0,694
Hémophilies et affections hémostase	257	X	X				X	X		X	X	X	X		X		X			X					0,588
HTA sévère	5678	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X		X	X	X	0,89
Coronaropathies	13210	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X		0,826
Insuffisance respiratoire chronique	3122	X	X	X	X	X	X	X		X	X	X	X	X	X	X	X	X	X	X		X	X	X	0,833
Alzheimer et autres démences	2550	X	X	X			X	X	X	X	X	X		X	X	X	X	X	X	X	X		X		0,944
Parkinson	1686	X	X	X	X		X	X		X	X	X	X	X	X	X	X	X	X	X	X	X	X		0,5
Maladies métaboliques héréditaires	968	X	X	X	X		X	X		X	X	X	X		X	X	X		X	X					0,694
Néphropathies	1676	X	X	X	X	X	X	X			X	X	X		X		X	X	X	X					0,802
Paraplégie	176	X	X	X							X		X				X	X		X					0,5
Maladies auto-immunes (ScS, Lupus, PAN)	969	X	X	X			X	X		X	X	X	X		X		X	X	X	X					0,784
Polyarthrite rhumatoïde	2601	X	X	X		X	X	X		X	X	X	X	X	X		X	X	X	X					0,741
Affections psychiatriques	6438	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	0,724
Crohn et Rectocolite hémorragique	930	X	X	X		X	X	X		X	X	X	X		X		X		X	X					0,652
Sclérose en plaques	410	X	X	X			X	X		X	X		X	X		X	X		X	X					0,706
Scoliose structurale	87	X	X	X			X	X		X	X						X								0,805
Spondylarthrite ankylosante	1048	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	0,684
Suites de transplantation d'organe	197	X		X				X		X		X					X								0,81
Tuberculose, Lèpre	52	X															X								0,561
Tumeurs malignes	25934	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X		X		X	0,776
Autres affections graves et Polypathologies	5741	X	X	X	X		X	X		X	X	X	X	X	X	X	X	X	X	X	X	X		X	0,782

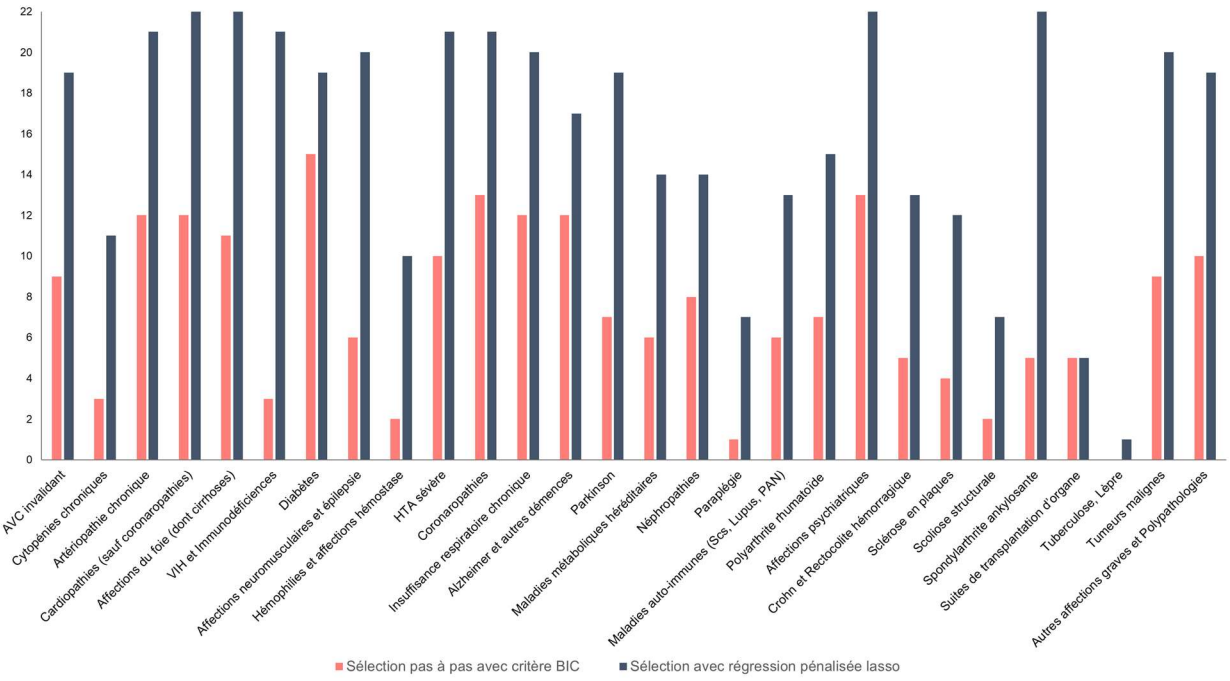


Figure 39 : Comparaison du nombre de variables sélectionnées via les deux méthodologies, **la sélection pas à pas des variables (critère BIC) et la régression pénalisée *lasso*,** pour l'application de la régression logistique aux données de la MSA (2006-2016)

III. Discussion

Comme décrit précédemment, la régression pénalisée *lasso* a été utilisée afin de tester une autre méthodologie de sélection de variables à intégrer aux modèles de régression logistique, dans le but d'améliorer les modèles construits jusque-là avec la sélection de variables pas à pas selon le critère BIC. Au final, on s'aperçoit qu'avec cette méthodologie, le nombre de variables sélectionnées pour la modélisation est bien plus important qu'auparavant, ce qui ajoute un degré de complexité important pour la compréhension des modèles. **De plus, le nombre d'associations statistiquement significatives diminue fortement avec la disparition de 15 associations mises en évidence précédemment, ce qui pourrait suggérer un surajustement.** En effet, pour les ALD concernées par la disparition de signaux, dix variables en moyenne ont été ajoutées aux modèles alors que pour les autres ALD, sept variables ont été ajoutées en moyenne. Ceci pourrait in fine masquer les effets propres des activités professionnelles. D'ailleurs, l'ajout de variables aux modèles n'a pas permis d'améliorer leur pouvoir explicatif et discriminant. Au contraire, les mesures de robustesse calculées sont au mieux, similaires voire inférieures.

Concernant l'association à nouveau mise en évidence en utilisant cette méthodologie de sélection de variables (association entre l'ALD « HTA sévère » et les entreprises de travaux agricoles), il est à noter que cette association (mise en évidence via la « Régression logistique simple », $p\text{-valeur}_{corrigée} = 0.003$) a été masquée en ajoutant les comorbidités avec une p-valeur proche de la limite de la significativité (Régression logistique avec comorbidités : $p\text{-valeur}_{corrigée} = 0.09$). Or, comme la sélection de variables a été réalisée avec une méthodologie différente, il est possible de remarquer que la composante principale n°1 de l'ACP des comorbidités n'a pas été incluse cette fois-ci (Tableau 14). Il est ainsi possible de faire l'hypothèse que cette CP n°1 est davantage porteuse d'information concernant les pathologies antérieures, potentiellement liées à l'apparition de l'ALD concernée, que les autres CP. En effet, lors de l'utilisation de la régression logistique « simple » (sans l'ajout des comorbidités), on observe également des associations montrant des risques plus élevés pour les ALD « Coronaropathies » et « Diabètes ». Or, l'ensemble de ces trois groupes de pathologies partagent des facteurs de risques similaires. La CP n°1 des comorbidités permet alors certainement d'ajuster sur ces autres pathologies, faisant ainsi disparaître l'association en question lorsqu'on l'ajoute à la sélection de variables. Ajuster sur cette CP n°1 est donc *a priori* primordiale, notamment pour cette ALD, ce que ne permet pas la méthodologie de sélection de variables via la régression pénalisée *lasso*.

Ainsi, la régression pénalisée *lasso* utilisée dans ce contexte, pour la sélection de variables, n'a pas apporté les résultats escomptés, à savoir, un meilleur pouvoir discriminant des modèles. Au contraire, pour deux ALD, les AUC montrent des modèles moins discriminants qu'avec la méthode précédente de sélection de variables, c'est-à-dire, via la sélection de variables pas à pas (critère BIC).

Au regard des résultats obtenus, cette méthodologie de sélection de variables n'est pas à privilégier dans le cadre de ce travail. Il a donc été choisi d'utiliser uniquement la méthode de sélection de variables pas à pas via le critère BIC.

C. Correction de biais liés aux événements rares

I. Méthodologie

Pour rappel, dans les analyses statistiques menées jusqu'ici, nous avons un nombre de témoins très important et relativement « démesuré » vis-à-vis du nombre de « malades » considérés quelle que soit la pathologie étudiée. Nous sommes donc face à des évènements rares, pouvant alors expliquer le manque de sensibilité et de discrimination des modèles de régression logistique construits jusque-là. En effet, utiliser la régression logistique dans ce type de situation peut engendrer des biais difficilement quantifiables pouvant entraîner une sous-estimation de la probabilité de survenue des événements rares étudiés (145). Dans la littérature scientifique, ce type de biais a été assez peu mentionné mais des solutions ont tout de même été préconisées, notamment **l'utilisation de la méthode de vraisemblance pénalisée inventée par Firth** en 1993 (108,146,147). Cette approche de vraisemblance pénalisée permet de réduire le biais lié aux événements rares afin d'obtenir des estimations des coefficients de régressions plus précis, qui ne tendent pas vers l'infini (intervalles de confiance plus fiables) et qui seraient ainsi plus plausibles.

Dans le cadre de ce travail, cette approche a été privilégiée pour corriger les biais évoqués car elle a été davantage testée et discutée dans la littérature et son utilisation est relativement aisée via le package « brglm2 » développé pour le logiciel R (148). En effet, la régression logistique est simplement appliquée de la même façon que pour la « Régression logistique avec comorbidités » (Figure 27), avec la même sélection de variables. Cependant, le paramètre « brglm » est ajouté lors de l'application du modèle pour spécifier que la méthode utilisée est celle de Firth. On réalise alors la régression logistique avec ce paramètre supplémentaire sur l'échantillon d'entraînement (70% des données). Puis, le modèle est testé sur l'échantillon de validation (30% des données) via le calcul des mesures de robustesse utilisées jusqu'à présent (sensibilité, spécificité, …). Les modèles ainsi évalués sont appliqués sur l'ensemble des données afin de récupérer les p-valeurs, qui sont ensuite corrigées via la méthode de Benjamini-Hochberg, et les coefficients de régression qui permettent de calculer les odds ratios. Ainsi, on obtient finalement les mesures d'associations entre chaque pathologie (ALD) et chaque activité professionnelle, avec une correction supplémentaire ayant *a priori* permis de réduire les biais liés aux évènements rares.

II. Résultats

De nouveau, des associations entre ALD et activités professionnelles chez les non-salariés de la MSA au cours de la période d'observation ont pu être mises en évidence via la régression logistique en utilisant la méthode de vraisemblance pénalisée de Firth. La Figure 40 représente les p-valeurs de chaque association testée, corrigées par la procédure de Benjamini-Hochberg et transformées à l'aide du logarithme décimal ($-\log_{10}$). **Les p-valeurs des associations sont quasiment identiques à celles mises en évidence précédemment sur la Figure 31 (« régression logistique avec comorbidité »). En ce qui concerne le nombre d'associations statistiquement significatives ($p\text{-valeur}_{corrigée} < 0.05$), 58 associations ont été mises en évidence avec cette méthode, contre 54 précédemment.** En effet, quatre nouvelles associations statistiquement significatives sont apparues que l'on peut visualiser sur la heatmap (Figure 41) et le forest plot (Figure 42) :

- Risques plus élevés de déclarations d'ALD « Crohn et Rectocolite hémorragique » pour les viticulteurs (OR = 1.33 [1.09 ;1.61] ; $p\text{-valeur}_{corrigée}$ = 0.03 ; n = 127), les conchyliculteurs (OR = 2.68 [1.36 ;5.26] ; $p\text{-valeur}_{corrigée}$ = 0.03 ; n = 8) et risque moins élevé de déclaration de cette ALD chez les éleveurs de bovins (viande) (OR = 0.72 [0.57 ;0.91] ; $p\text{-valeur}_{corrigée}$ = 0.03 ; n = 79) ;

- Et un risque moins élevé de déclarations d'ALD « Maladies métaboliques héréditaires » dans le secteur des cultures céréalières et industrielles (OR = 0.77 [0.64 ;0.91] ; $p\text{-valeur}_{corrigée}$ = 0.02 ; n = 176).

Par ailleurs, pour quatre de ces associations, il est important de noter que les p-valeurs avec la méthodologie précédente étaient proches du seuil de significativité ($0.05 < p\text{-valeur}_{corrigée} < 0.10$).

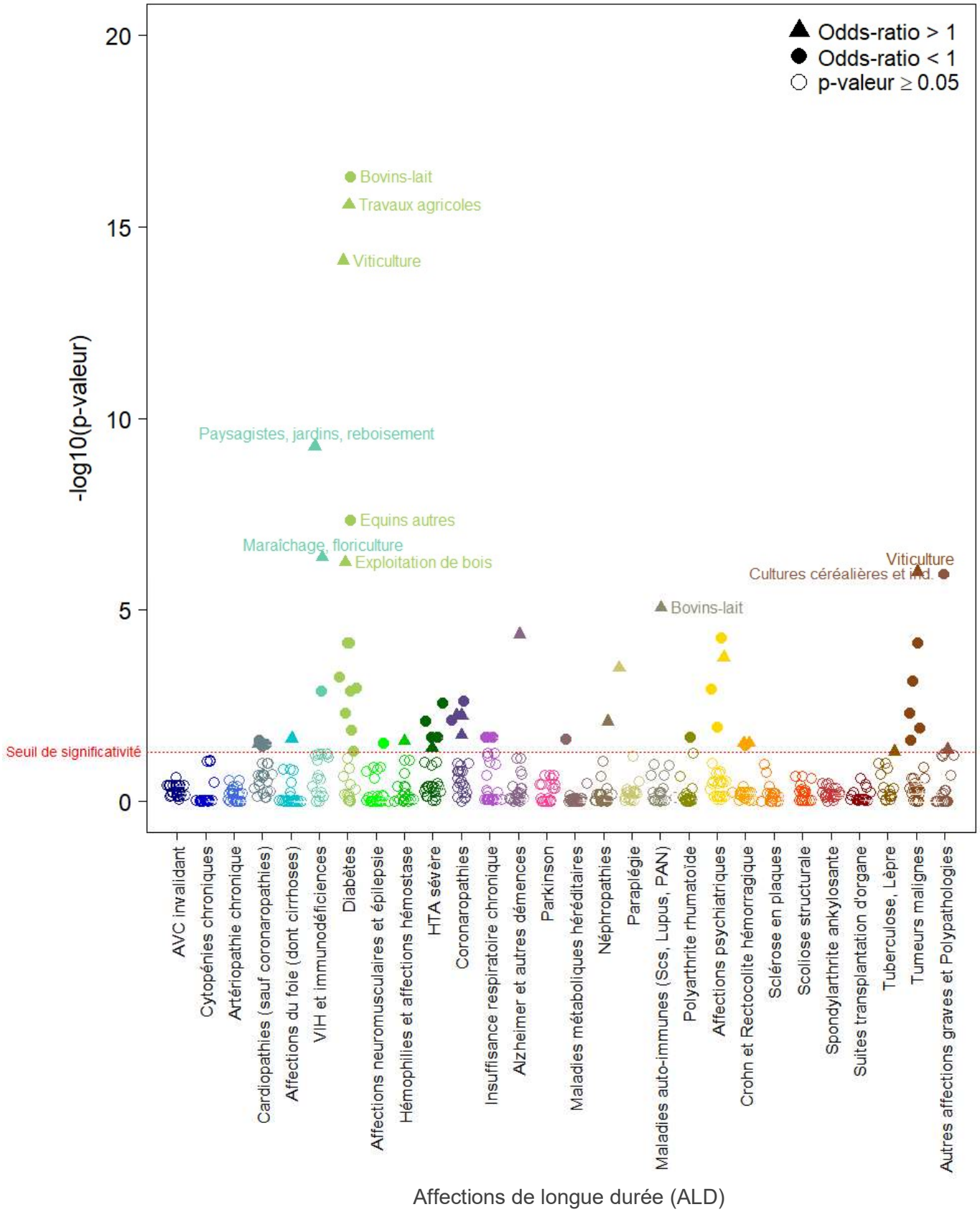


Figure 40 : Représentation graphique des p-valeurs, corrigées par la procédure de Benjamini-Hochberg, obtenues via la régression logistique **avec la correction de biais liés aux événements rares (méthode Firth)** entre chaque combinaison d'ALD et d'activité professionnelle en ajustant sur d'autres paramètres**,** chez les **non-salariés** de la MSA (2006-2016)

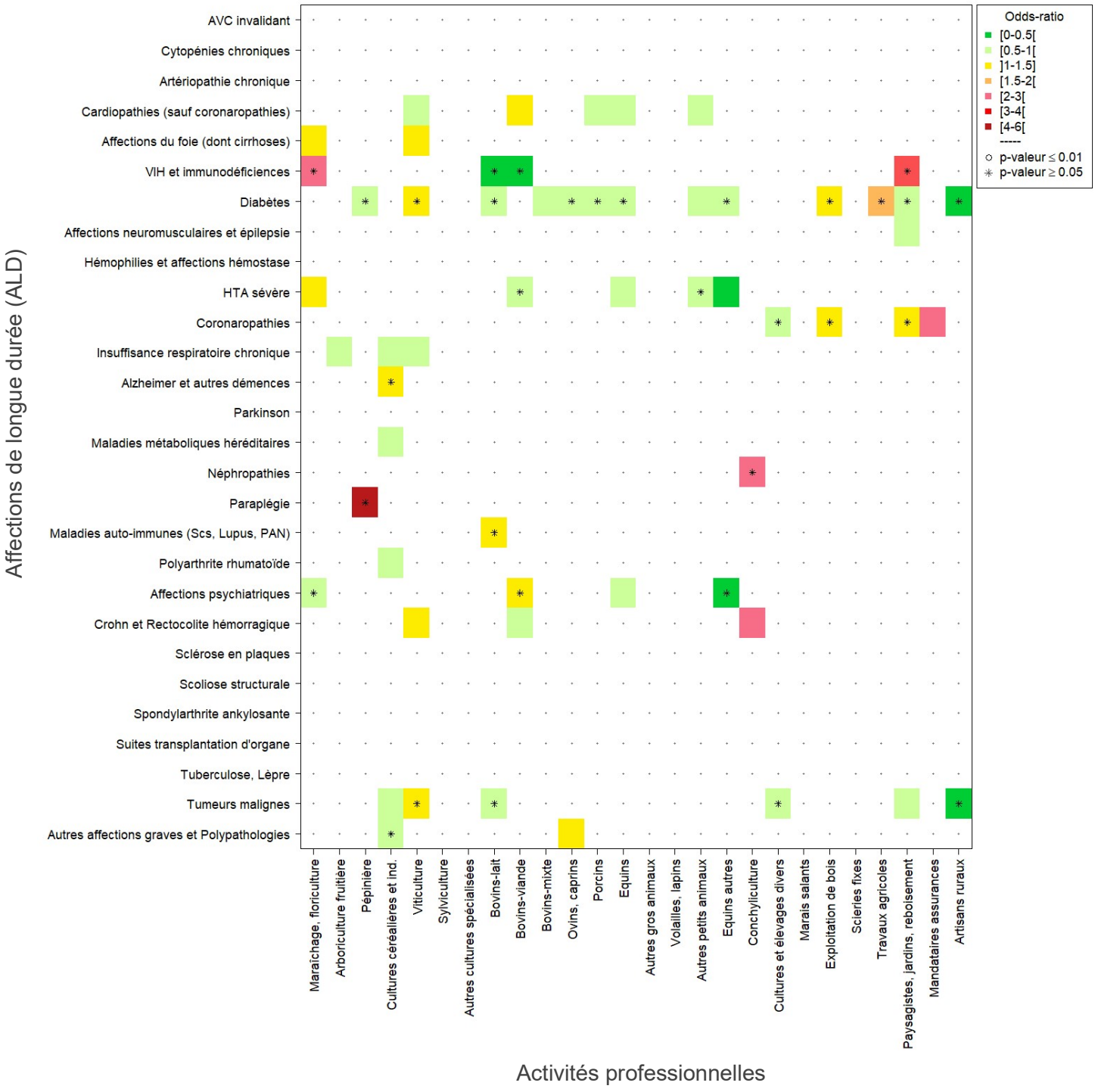


Figure 41 : Représentation graphique des odds ratios obtenus via la régression logistique **avec la correction de biais liés aux événements rares (méthode Firth)** entre chaque combinaison d'ALD et d'activité professionnelle en ajustant sur d'autres paramètres chez les non-salariés de la MSA (2006-2016)

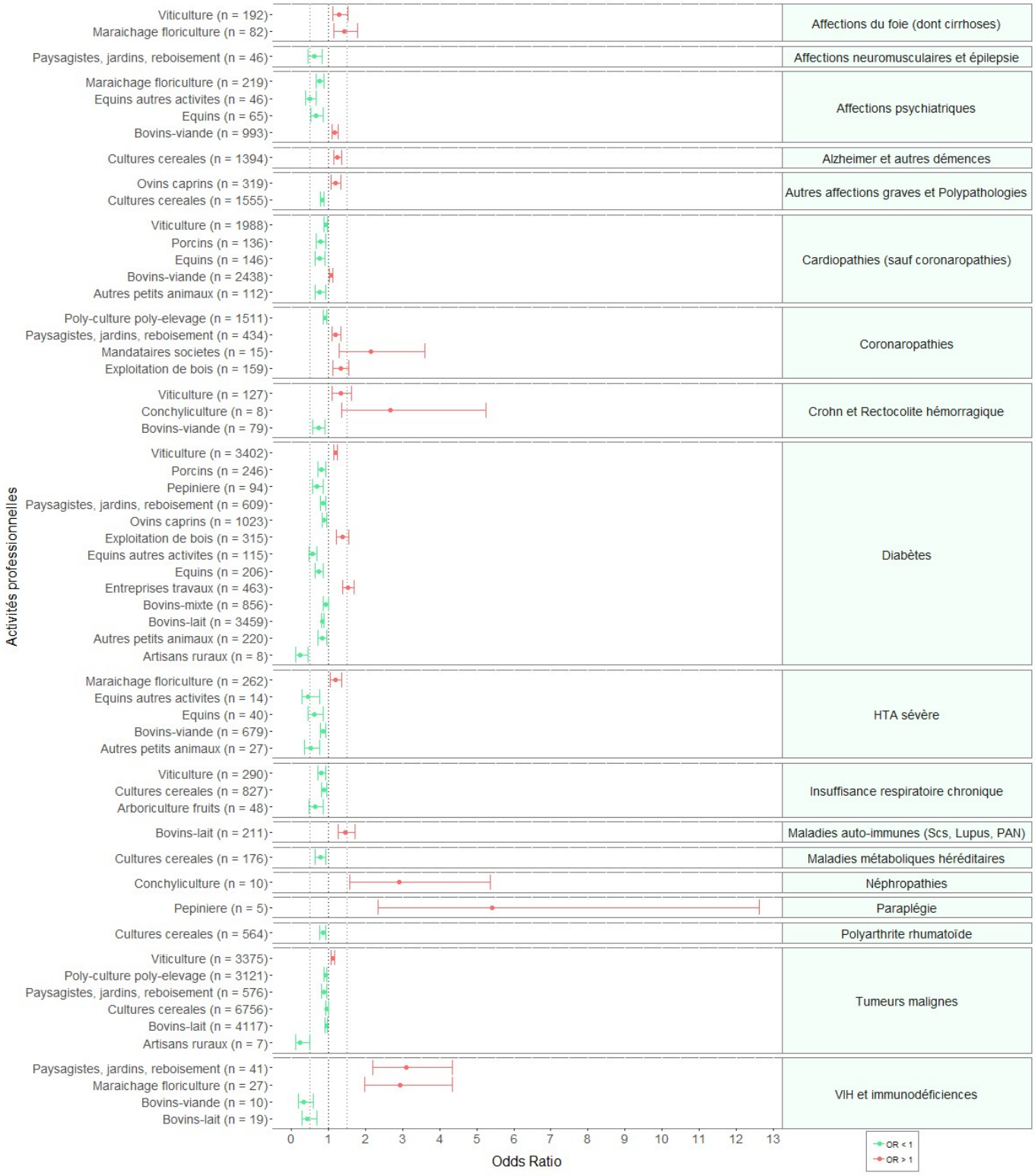


Figure 42 : Représentation graphique des odds ratios et de leurs intervalles de confiance à 95% obtenus via la régression logistique **avec la correction de biais liés aux événements rares (méthode Firth)** pour chaque association statistiquement significative ($p\text{-valeur}_{corrigée} < 0.05$) entre une ALD et une activité professionnelle en ajustant sur d'autres paramètres, chez les non-salariés de la MSA (2006-2016)

Tableau 15 : Sélection de variables et mesure de l'AUC (échantillon de validation : 30% données) pour chaque ALD qui ont été utilisées lors de la régression logistique **avec la correction de biais liés aux événements rares (méthode Firth)**, effectuée sur les non-salariés de la MSA (2006-2016)

Affections de longue durée (ALD)	Effectifs	Activités prof.	Année d' installation	Sexe	RSA	Chômage	Superficie d' exploitation	Age	Nb. d' activités prof.	Nb. d' années obs	« Revenus »	ALD avant obs.	« Marié »	« Veuf »	« Séparé- divorcé »	Nb. de salariés	Régime maladie	Statut conjoint	Type d' exploitation	Région	Comorbidité CP 1	Comorbidité CP 2	Comorbidité CP 3	Comorbidité CP 4	AUC (échantillon de validation)
AVC invalidant	5288*	X	X					X				X			X		X				X	X	X	X	0,793
Cytopénies chroniques	284	X						X									X				X				0,718
Artériopathie chronique	4942	X	X	X			X	X							X	X	X				X	X	X	X	0,824
Cardiopathies (sauf coronaropathies)	16977	X	X	X				X			X	X		X			X		X		X	X	X	X	0,844
Affections du foie (dont cirrhoses)	1320	X	X	X	X			X				X			X		X		X		X	X	X		0,736
VIH et Immunodéficiences	241	X						X					X				X								0,771
Diabètes	25229	X	X	X	X	X		X	X		X	X					X		X	X	X	X	X	X	0,784
Affections neuromusculaires et épilepsie	1927	X	X		X							X					X				X		X		0,698
Hémophilies et affections hémostase	257	X															X				X				0,55
HTA sévère	5678	X	X					X			X	X					X		X	X	X	X	X		0,843
Coronaropathies	13210	X	X	X				X				X	X	X	X		X			X	X	X	X	X	0,812
Insuffisance respiratoire chronique	3122	X	X		X			X			X	X				X	X			X	X	X	X	X	0,812
Alzheimer et autres démences	2550	X	X	X			X	X		X		X		X			X			X	X	X	X		0,941
Parkinson	1686	X	X	X				X		X			X				X				X				0,846
Maladies métaboliques héréditaires	968	X	X									X					X			X	X	X			0,689
Néphropathies	1676	X	X					X				X					X				X	X	X	X	0,765
Paraplégie	176	X															X								0,618
Maladies auto-immunes (ScS, Lupus, PAN)	969	X	X	X				X				X					X				X				0,783
Polyarthrite rhumatoïde	2601	X	X	X				X				X	X				X					X			0,737
Affections psychiatriques	6438	X	X	X	X			X			X	X			X		X		X	X	X	X	X		0,717
Crohn et Rectocolite hémorragique	930	X						X		X		X					X			X					0,652
Sclérose en plaques	410	X		X				X		X							X								0,69
Scoliose structurale	87	X		X													X								0,751
Spondylarthrite ankylosante	1048	X						X		X		X	X				X								0,687
Suites de transplantation d'organe	197	X		X				X				X					X				X				0,798
Tuberculose, Lèpre	52	X																							0,594
Tumeurs malignes	25934	X	X					X					X		X		X			X	X	X		X	0,772
Autres affections graves et Polypathologies	5741	X	X	X				X			X	X					X			X	X	X	X		0,768

Concernant les mesures de robustesse de ces modèles, leurs aires sous la courbe ROC (AUC) montrent en moyenne aussi un bon pouvoir discriminant (AUC_{moyen} = 0.75), légèrement moins performant qu'avec la méthodologie précédente (AUC_{moyen} = 0.76) (Tableau 15). Si on regarde plus précisément les AUC par ALD, elles sont de façon générale similaires à l'exception de deux ALD, où les valeurs sont plus faibles, suggérant des modèles moins performants : « Parkinson » ($AUC_{RL\text{-}comorbidités}$= 0.846 ; AUC_{Firth}= 0.5) et « Paraplégie » ($AUC_{RL\text{-}comorbidités}$= 0.581 ; AUC_{Firth} = 0.5). En outre, du fait que les biais liés aux évènements rares ont *a priori* été corrigés via la méthode de Firth, on obtient des mesures de spécificité ($Spécificité_{moyenne}$ = 1), de sensibilité ($Sensibilité_{moyenne}$ = 0) et de F_1 score ($F1_{moyen}$ = 0) qui montrent que les modèles ne sont pas davantage performants pour identifier les rares « malades » au sein de notre population agricole.

III. Discussion

La méthode de vraisemblance pénalisée de Firth a été utilisée dans le but théorique de limiter les biais liés à l'étude d'événements rares. La méthodologie a pu être appliquée simplement aux données de la MSA, déjà nettoyées et structurées de sorte à permettre l'application de la régression logistique. De manière générale, les résultats montrent des associations statistiquement significatives similaires à celles mises en évidence via la régression logistique avec les comorbidités. Les p-valeurs corrigées et les odds ratios sont relativement similaires. On peut alors faire l'hypothèse que les associations mises en évidence jusque-là par ces deux méthodologies sont relativement stables et robustes, même après correction des biais liés aux événements rares.

Par ailleurs, la méthode de Firth a permis de mettre en évidence quatre nouvelles associations statistiquement significatives. Cependant, il est important de savoir que les p-valeurs corrigées relatives à ces associations étaient en limite de significativité lors de la régression logistique avec comorbidités. Il ne s'agit alors là que d'un effet de seuil. En effet, selon la méthodologie utilisée, il est possible de voir apparaître et disparaître des signaux qui ont une p-valeur corrigée proche du seuil de significativité. Cependant, de tels signaux sont alors peu robustes et peu stables, car dépendants du seuil de significativité fixé et ne seront alors vraisemblablement pas une priorité en termes d'investigation dans le cadre d'un système de vigilance de risques professionnels. A ce titre, dans le cadre du projet, des critères de priorisation des signaux à investiguer, du fait de leur robustesse et de l'importance de l'effet devront être établis ; la question de nouveaux seuils de significativité pour les p-valeurs pourra alors être posée (149).

Enfin, si cette méthode permet *a priori* de corriger les biais liés aux événements rares, elle démontre par ailleurs la robustesse des signaux mis en évidence jusque-là. Cependant, malgré la correction de ces biais, une fois de plus, les mesures de robustesse calculées ne montrent pas de nette amélioration des modèles.

Au regard des résultats obtenus, il n'y a donc pas d'intérêt particulier à favoriser l'utilisation de la méthode de vraisemblance pénalisée de Firth. Il a donc été choisi de conserver la méthode de régression logistique « simple » à laquelle on a ajouté les comorbidités.

D. Comparaison des méthodologies

Différentes méthodologies ont été testées afin de tenter une amélioration des mesures de robustesse et notamment du pouvoir discriminant des modèles de régression logistique. Après évaluation de chacune des méthodologies, il apparaît qu'aucune de ces méthodes ne permet réellement d'augmenter la sensibilité et donc le pouvoir discriminant (Tableau 16). Cependant, l'ajout de comorbidités permet tout de même d'approcher et d'ajouter certains facteurs de confusion potentiels. C'est pourquoi, dans le cadre de ce manuscrit de thèse, la méthode retenue est celle de la régression logistique à laquelle on ajoute les comorbidités. Cependant, malgré la mise en évidence de 54 signaux d'intérêt à partir de cette méthodologie, il est nécessaire d'évaluer d'autres méthodologies sur les données MSA.

Il pourrait d'abord être envisagé de tester des **modèles de survie**. Ces modèles sont très utilisés en épidémiologie, notamment dans les études menées par la cohorte AGRICAN, quand il s'agit d'étudier le délai de survenue d'un événement particulier (ici, la déclaration d'une pathologie en ALD). Pour permettre l'utilisation de ce type de modèles, un travail important de restructuration des données doit être réalisé pour que les données puissent être analysées comme des données longitudinales, c'est-à-dire, en tenant compte de la dimension temporelle. Par ailleurs, dans le cadre de ce travail, il est important de noter qu'une variable « durée d'observation » a été ajoutée à l'ensemble des variables explicatives. Cette variable a permis en quelque sorte de prendre en compte la variabilité interindividuelle du temps d'observation. Néanmoins, les modèles de survie permettraient alors de mieux prendre en compte cette dimension temporelle. Pour chaque pathologie étudiée, il s'agira alors de définir pour chaque individu une date d'origine, ce choix n'étant pas aisé compte tenu de la complexité des données mises à disposition par la MSA, puis de tenir compte de leur date de dernières nouvelles. La durée totale d'observation ne pourra de toute façon pas dépasser la fenêtre temporelle pendant laquelle nous avons des informations sur les cotisants et qui dépend des données mises à disposition par la MSA (2006-2016 : 11 années d'observation). De plus, les données MSA étant incomplètes pour la plupart des individus, il faudra être attentif aux phénomènes de « censure à droite » (l'individu n'a pas eu de déclaration d'ALD à sa date de dernières nouvelles) et de « censure à gauche » (l'individu a déjà une déclaration d'ALD à son inclusion dans l'étude). Ces modèles ont également l'avantage de calculer des « hazards ratios », analogues aux OR obtenus via la régression logistique (101).

Afin de tenir compte de la limite concernant le nombre de « malades » par rapport au nombre très important de témoins, une nouvelle stratégie pourrait être adoptée. En effet, le groupe de témoins pourrait être défini aléatoirement en choisissant par exemple un nombre de 3 témoins pour un « malade » au sein de la population source. Il serait alors possible de renouveler les

analyses plusieurs fois, en changeant à chaque fois la population de témoins. Cette **technique de « boostraping »** permettrait alors d'obtenir des estimateurs avec leurs marges d'incertitude.

Ces méthodologies seront évaluées *a posteriori*, hors du cadre de ce travail de thèse et les résultats seront comparés à ceux obtenus via la régression logistique.

Tableau 16 : Comparaison des forces et limites de chaque méthodologie utilisée sur les données de la MSA (non-salariés, période 2006-2016)

Etape méthodologique	Facteurs de confusion	Méthode de sélection de variables	Correction(s) appliquée(s)	Associations significatives (p-valeur corrigée < 0.05)	Forces principales	Limites principales
Régression logistique « simple »	*Prise en compte des seules variables mises à disposition par la MSA*	Sélection pas à pas des variables (critère BIC)	P-valeurs corrigées par la procédure de Benjamini-Hochberg	54	- Méthode très utilisée en épidémiologie, bon compromis entre fiabilité et lisibilité des résultats (odds ratio) - **Bonne spécificité des modèles (AUC_{moyen} = 0.75)** - Mise en évidence d'associations capturant des déterminants de la santé au travail déjà suspectés en agriculture, OU qu'il est possible de relier à des déterminants sociaux de santé OU encore qui permettent de générer des hypothèses	- **Modèles peu discriminants (sensibilité nulle, modèles peu efficaces pour identifier les individus ayant des déclarations d'ALD)** - Non prise en compte de facteurs de confusion résiduels (non renseignés par les variables existantes) - Méthode n'est pas la plus adaptée à l'étude d'événements rares (ici, faibles proportions d'individus « malades »)
Régression logistique avec estimation de facteurs de confusion	**Ajout de facteurs latents estimés via la méthodologie « LFMM »**			32	- Estimation de facteurs latents à partir des données MSA à disposition, en tenant compte aussi bien des ALD que des variables explicatives (âge, sexe, …) - **Spécificité relativement stable des modèles (AUC_{moyen} = 0.76)**	- Utilisation de la régression linéaire pour l'estimation des facteurs latents alors que nous disposons uniquement de variables catégorielles - Méthodologie peu adaptée à un jeu de données dont le nombre de lignes est bien supérieure au nombre de colonnes (comme c'est le cas ici) - **Modèles peu discriminants (sensibilité nulle)**
	Ajout des comorbidités (ALD antérieures)			54	- Ajout aux modèles de variables explicatives connues et disponibles au sein des données MSA, permettant d'approcher certains facteurs de confusion potentiels - **Spécificité relativement stable des modèles (AUC_{moyen} = 0.76)**	- Utilisation de l'analyse en composantes principales (ACP) pour réduire le nombre de dimensions et donc de variables à ajouter aux modèles → ajout d'une incertitude dans la compréhension de la sélection de variables - **Modèles qui restent peu discriminants (sensibilité nulle)**
Régression logistique avec sélections de variables effectuée via la régression pénalisée lasso		**Régression pénalisée *lasso***		39	- Permet de rechercher des modèles plus parcimonieux - **Spécificité relativement stable des modèles (AUC_{moyen} = 0.75)**	- Nombre de variables sélectionnés importants → ajout de complexité aux modèles, modèles moins parcimonieux, possible surajustements - **Modèles qui restent peu discriminants (sensibilité nulle)**
Régression logistique avec correction par la méthode de vraisemblance pénalisée de Firth		Sélection pas à pas des variables (critère BIC)	P-valeurs corrigées par la procédure de Benjamini-Hochberg **+ Correction des biais liés aux événements rares (méthode de vraisemblance pénalisée de Firth)**	58	- Réduction des biais liés aux événements rares - A permis de montrer la robustesse des signaux mis en évidence jusque-là - **Spécificité relativement stable des modèles (AUC_{moyen} = 0.75)**	- **Modèles qui restent peu discriminants (sensibilité nulle)**

PARTIE 5

Analyses réalisées au niveau de précision de la pathologie CIM-10

I. Méthodologie

Dans les parties précédentes, l'objectif a consisté à montrer qu'il était possible d'appliquer des méthodes statistiques sur les données de la MSA afin de mettre en évidence, sans *a priori*, des associations entre ALD et activités professionnelles du secteur agricole. Maintenant qu'il a été démontré que la méthodologie statistique développée fonctionnait à ce niveau de précision, il est désormais possible de préciser les différents paramètres des modèles, à commencer par la pathologie ALD. En effet, pour chaque ALD déclarée, une pathologie codée via la CIM-10 est renseignée. Ce niveau de précision permet alors de s'affranchir de la liste limitative de pathologies en ALD et de mettre en évidence des associations avec des pathologies plus spécifiques. Pour rappel, cette classification possède différents niveaux de précision avec un total de 14 400 codes différents. Au sein des données de la MSA, nous avons en moyenne 24 codes CIM-10 différents par ALD (Figure 16). Cependant, en utilisant ce niveau de précision, certaines difficultés techniques ont émergé.

D'une part, la quantité de données étant massive, il est difficile d'effectuer une sélection de variables pas à pas via le critère BIC pour chaque pathologie codée via la CIM-10 ($n > 750$), à cause du temps de calcul nécessaire mais aussi des erreurs techniques survenant à cause des effectifs trop faibles de « malades » pour certaines pathologies ($n < 50$). Pour contourner ces contraintes techniques, les sélections de variables effectuées précédemment pour chaque ALD ont été utilisées en fonction des codes CIM-10 étudiés. Par exemple, si on étudie une pathologie ayant un code CIM-10 dans la famille « C00-D49 : Tumeurs », la sélection des variables de l'ALD « Tumeurs malignes » est privilégiée. Cependant, cette sélection de variables peut ne pas être adaptée à chacune des pathologies CIM-10 étudiées et c'est pourquoi, la validité des modèles a tout de même été évaluée avec un calcul de l'AUC sur un échantillon de validation pour chaque pathologie CIM-10 étudiée. Alors, un seuil d'AUC minimal à 0.60 a été choisi afin d'exclure les modèles n'ayant pas un pouvoir discriminant suffisant pour mettre en évidence des associations cohérentes et donc davantage fiables.

D'autre part, la deuxième limite concerne le fait d'étudier des événements bien plus rares qu'en étudiant les ALD, avec un nombre de malades par pathologie encore plus disproportionné par rapport au nombre de témoins. L'évaluation de la méthode de Firth dans la partie précédente n'ayant pas montré de résultats convaincants à son utilisation, la régression logistique « simple » a été utilisée en ajoutant les comorbidités, qui pour rappel, sont les composantes principales de l'ACP des ALD antérieures. Par ailleurs, au regard des biais liés aux événements rares, si des associations étaient mises en évidence avec des effectifs inférieurs à trois individus, elles n'ont pas été prises en compte ni dans le cadre de ce travail, ni dans le cadre d'un système de vigilance des risques professionnels.

Pour ces analyses réalisées au niveau de précision de la pathologie CIM-10, quatre exemples ont été choisis afin de préciser les associations déjà mises en évidence pour les ALD concernées :

- les pathologies CIM-10 de l'ALD « Affections psychiatriques » ;
- les pathologies CIM-10 de l'ALD « Maladies auto-immunes (ScS, Lupus, PAN)[14] » ;
- les pathologies CIM-10 de l'ALD « Alzheimer et autres démences » ;
- et quelques-unes des pathologies CIM-10 de l'ALD « Tumeurs malignes ».

Par ailleurs, comme un individu peut être déclaré pour une pathologie CIM-10 pouvant être classée dans plusieurs ALD différentes, à partir du moment où la pathologie CIM-10 était déclarée au sein de l'ALD étudiée, cet individu a été considéré. Par exemple, si on étudie la pathologie CIM-10 « C50 : Tumeur maligne du sein », l'individu est considéré comme « malade » dès lors qu'il a une déclaration pour la pathologie en question, peu importe qu'elle soit déclarée dans l'ALD « Tumeurs maladies » ou une autre.

En ce qui concerne les modèles, les sélections de variables pour les ALD concernées (cf. parties précédentes) ont été utilisées (Tableau 17).

[14] Maladies auto-immunes, incluant la sclérodermie systémique (également appelée sclérose systémique « ScS »), le lupus systémique (également appelé lupus érythémateux aigu disséminé LEAD), et les vascularites dont la Périartérite noueuse (PAN). Ces maladies sont liées à la production de divers anticorps, en mesure de toucher plusieurs organes (d'où le nom « systémique »), car s'attaquant au tissu conjonctif et/ou à la paroi des vaisseaux sanguins (avec souvent une atteinte préférentielle de la microcirculation). Ces dernières sont alors parfois dénommées vascularites, périartérites, vasculopathies nécrosantes, …

Tableau 17 : Sélection de variables pour chaque ALD étudiée au niveau de précision de la pathologie CIM-10, qui ont été utilisées lors de la régression logistique avec les comorbidités, effectuée sur les non-salariés de la MSA (2006-2016)

ALD	Alzheimer et autres démences	Maladies auto-immunes (ScS, Lupus, PAN)	Affections psychiatriques	Tumeurs malignes
Effectifs	2550	969	6438	25934
Activités professionnelles	X	X	X	X
Année d'installation	X	X	X	X
Sexe	X	X	X	
RSA			X	
Chômage				
Superficie d'exploitation	X			
Âge	X	X	X	X
Nombre d'activités professionnelles				
Nombre d'années d'observation	X			
« Revenus »			X	
ALD avant obs.	X	X	X	
« Marié »				X
« Veuf »	X			
« Séparé-divorcé »			X	X
Nb. de salariés				
Régime maladie	X	X	X	X
Statut conjoint				
Type d'exploitation			X	
Région	X		X	X
Comorbidité CP 1	X	X	X	X
Comorbidité CP 2	X		X	X
Comorbidité CP 3	X		X	
Comorbidité CP 4				X

II. Résultats

a. Affections psychiatriques

L'intérêt de considérer l'ALD « Affections psychiatriques » au niveau de précision de la pathologie CIM-10 est de préciser les associations mises en évidence dans les parties précédentes, notamment celle entre l'ALD « Affections psychiatriques » et l'élevage de bovins (viande) (OR = 1.17 [1.09 ;1.25] ; p-valeur$_{corrigée}$ = 0.0001 ; n = 993). Cela peut également permettre de générer de nouvelles associations, qui sont « diluées » au sein de toutes les pathologies CIM-10 de cette ALD, pour lesquelles les p-valeurs corrigées sont en limite de significativité lorsque les analyses sont menées au niveau ALD.

En effet, au sein de la population des non-salariés étudiés, l'ALD « Affections psychiatriques » compte 28 pathologies CIM-10 différentes, faisant de cette ALD la troisième ayant le plus de pathologies CIM-10 différentes, après les ALD « Autres affections graves et Polypathologies » et « Tumeurs malignes ». En pratique, les non-salariés étudiés sont davantage répartis dans les pathologies CIM-10 « F32 : Episodes dépressifs » (n = 3123), « F31 : Trouble affectif bipolaire » (n = 1193) et « F10 : Troubles mentaux et du comportement liés à l'utilisation d'alcool » (n = 616) (Figure 43).

La régression logistique a alors été effectuée pour chacune des pathologies CIM-10 de l'ALD « Affections psychiatriques » avec la sélection de variables de cette même ALD. **Ces analyses ont alors permis de mettre en évidence des associations d'intérêt, notamment un risque de déclaration plus élevé de pathologie CIM-10 « F32 : Episodes dépressifs » aussi bien dans le secteur de l'élevage de bovins (viande) (OR = 1.28 [1.16 ;1.41] ; p-valeur$_{corrigée}$ = 1.77^{E-5} ; n = 524) que dans celui de l'élevage de bovins (lait) (OR = 1.20 [1.09 ;1.33] ; p-valeur$_{corrigée}$ = 0.002 ; n = 646).**

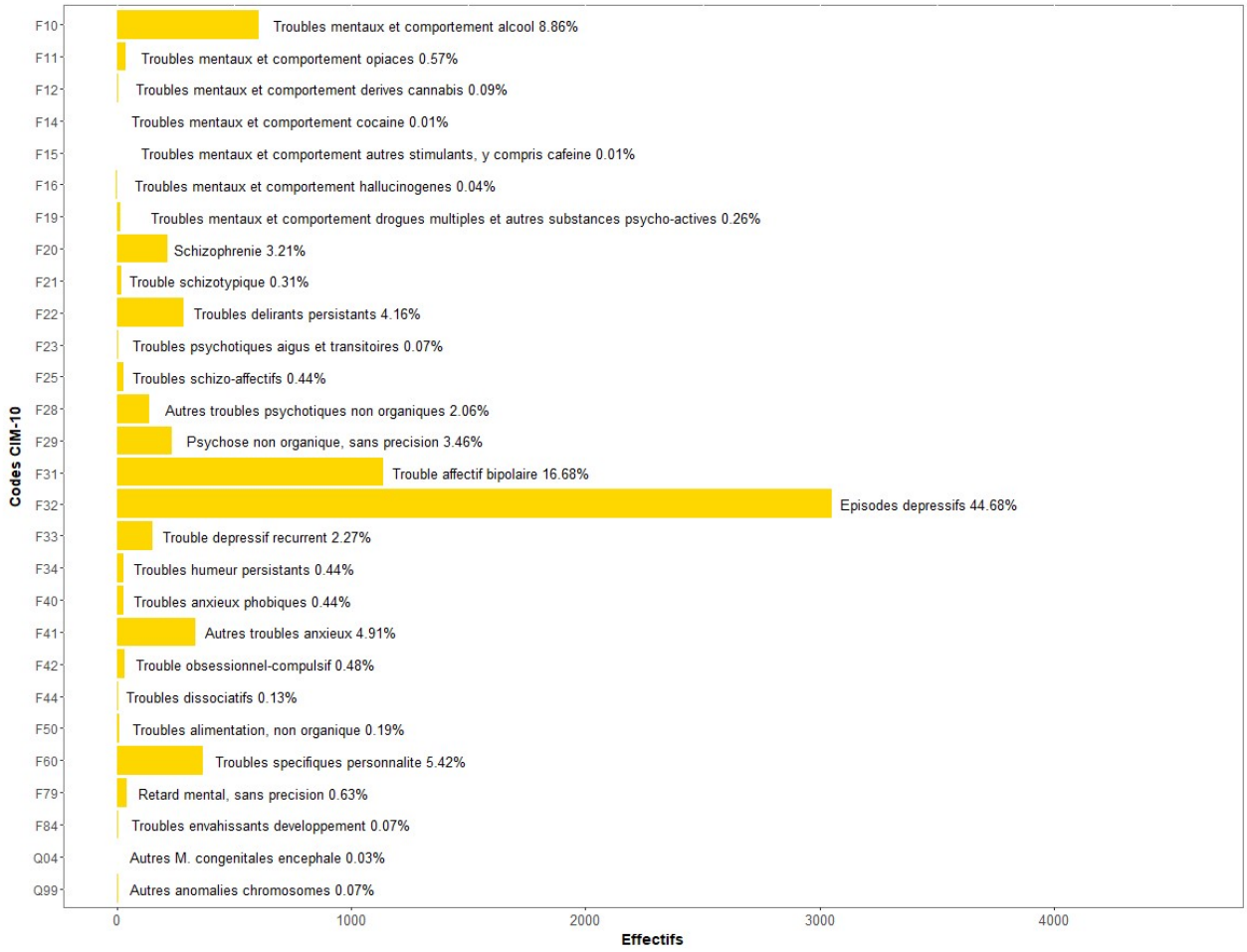


Figure 43 : Répartition des non-salariés de la MSA ayant une déclaration d'ALD « Affections psychiatriques » par pathologie CIM-10 (2012-2016)

b. Maladies auto-immunes (ScS, PAN, Lupus)

Comme précédemment, l'intérêt de ces analyses est notamment de préciser une association mise en évidence précédemment, à savoir, celle entre l'ALD « Maladies auto-immunes (ScS, PAN, Lupus) » et l'élevage de bovins (lait) (OR = 1.46 [1.25 ;1.70] ; p-valeur$_{corrigée}$ = 4.58^{E-5} ; n = 211).

Au sein de la population des non-salariés étudiés, l'ALD « Maladies auto-immunes (ScS, PAN, Lupus) » compte 5 pathologies CIM-10 différentes. En pratique, les non-salariés étudiés sont davantage répartis dans la pathologie CIM-10 « M35 : Autres atteintes systémiques du tissu conjonctif » (n = 453) (Figure 44), comprenant à titre d'exemple, les affections suivantes : les syndromes de Gougerot-Sjögren (« M35.0 »), de Behçet (« M35.2 ») ou d'hypermobilité (« M35.7 »), la polymyalgie rhumatismale (« M35.3 ») ou encore la fibrosclérose multiple

(« M35.5 »). **Néanmoins, n'ayant pas le niveau de précision à 4 digits pour tous les malades au sein des données MSA (le codage étant surtout recommandé à un niveau 3 digits), les analyses ont donc été conduites au niveau de précision de 3 digits**.

La régression logistique a alors été effectuée pour chacune des pathologies CIM-10 de l'ALD « Maladies auto-immunes (ScS, PAN, Lupus) » avec la sélection de variables de cette même ALD. **Ces analyses ont alors permis de mettre en évidence une association avec un risque de déclaration plus élevé de pathologie CIM-10 « M35 : Autres atteintes systémiques du tissu conjonctif » dans le secteur de l'élevage de bovins (lait) (OR = 1.63 [1.31 ;2.04] ; p-valeur$_{corrigée}$ = 0.0004 ; n = 105).**

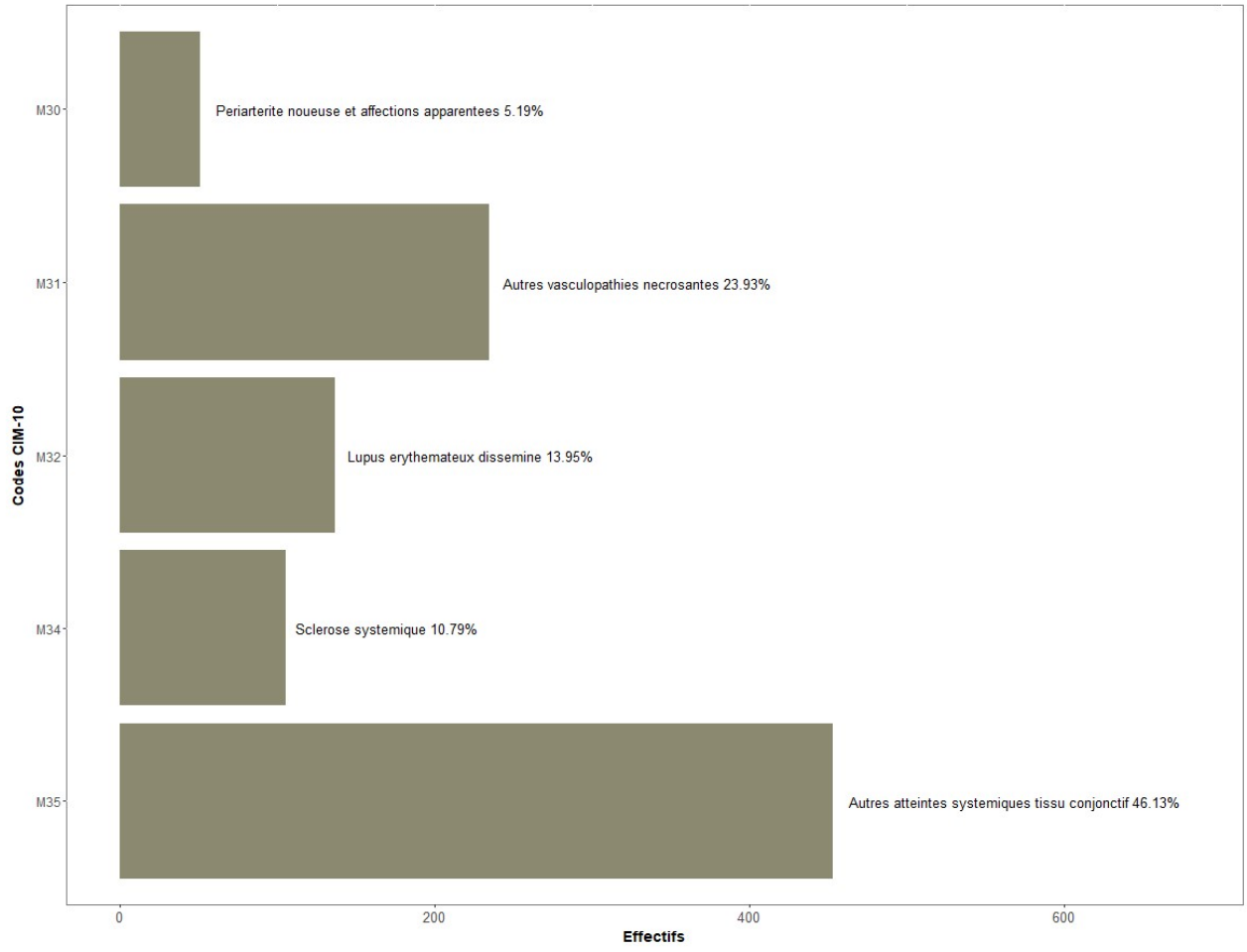


Figure 44 : Répartition des non-salariés de la MSA ayant une déclaration d'ALD « Maladies auto-immunes (ScS, PAN, Lupus) » par pathologie CIM-10 (2012-2016)

c. Alzheimer et autres démences

L'intérêt cette fois-ci est de déterminer si l'association, mise en évidence dans les parties précédentes, entre l'ALD « Alzheimer et autres démences » et le secteur agricole des cultures céréalières et industrielles (OR = 1.24 [1.13 ;1.35] ; p-valeur$_{corrigée}$ = 9.36^{E-5} ; n = 1394), concerne bien la maladie d'Alzheimer.

Au sein de la population des non-salariés étudiés, l'ALD « Alzheimer et autres démences » compte 5 pathologies CIM-10 différentes. En pratique, les non-salariés étudiés sont davantage répartis dans les pathologies CIM-10 « F00 : Démence de la maladie d'Alzheimer » (n = 1413) et « F03 : Démence, sans précision » (n = 885) (Figure 45).

La régression logistique a alors été effectuée pour chacune des pathologies CIM-10 de l'ALD « Alzheimer et autres démences » avec la sélection de variables de cette même ALD. Ces analyses ont alors permis de mettre en évidence deux associations avec un risque de déclaration plus élevé pour les pathologies CIM-10 :

- « F00 : Démence de la maladie d'Alzheimer » dans le secteur des cultures céréalières et industrielles (OR = 1.27 [1.13 ;1.44] ; p-valeur$_{corrigée}$ = 0.002 ; n = 784) ;

- « F01 : Démence vasculaire » dans les entreprises de travaux agricoles (OR = 4.87 [1.95 ;12.14] ; p-valeur$_{corrigée}$ = 0.02 ; n = 5).

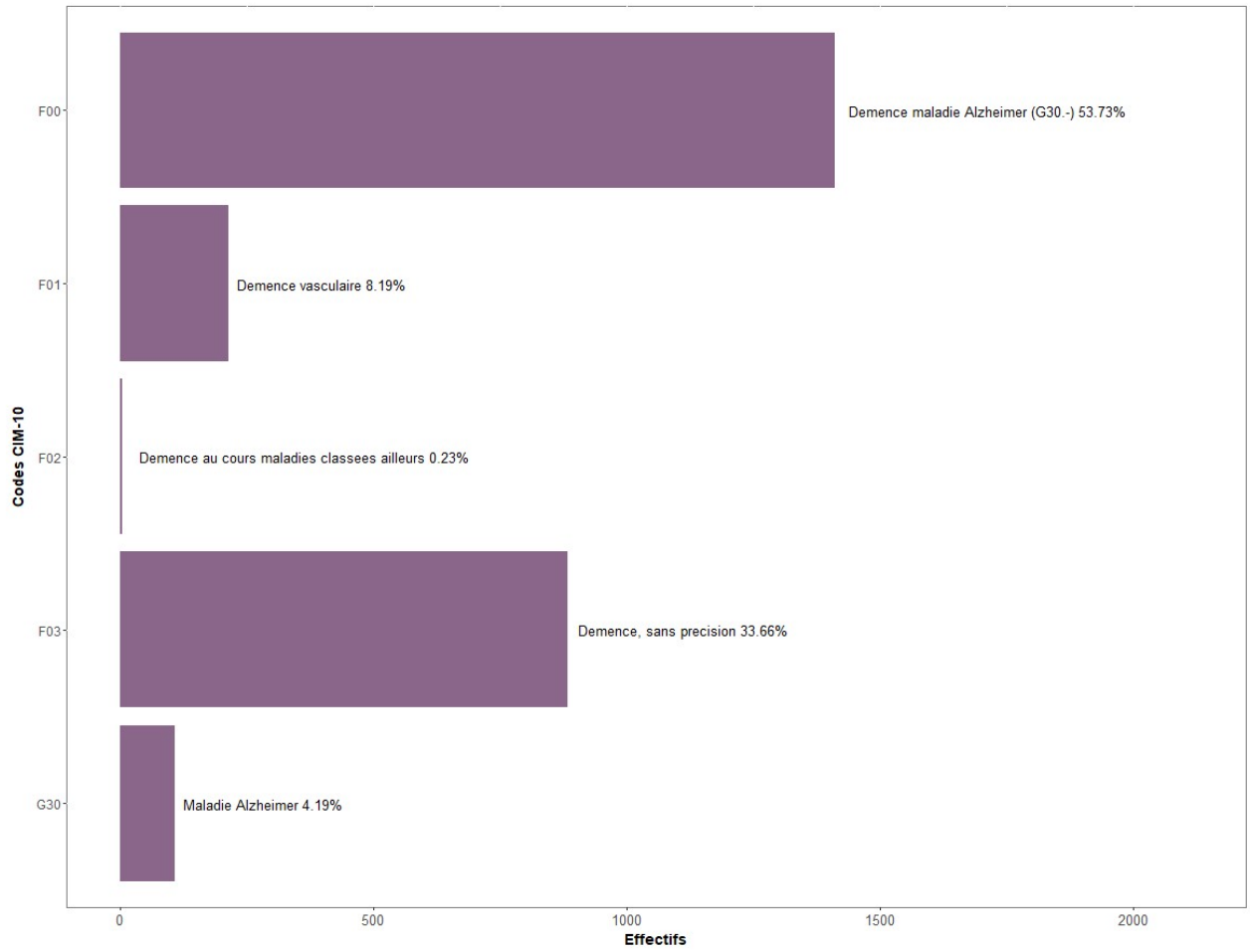


Figure 45 : Répartition des non-salariés de la MSA ayant une déclaration d'ALD « Alzheimer et autres démences » par pathologie CIM-10 (2012-2016)

d. Tumeurs malignes

L'intérêt est ici de préciser les associations mises en évidence dans les parties précédentes, notamment celle entre l'ALD « Tumeurs malignes » et la viticulture (OR = 1.11 [1.07 ;1.16] ; p-valeur$_{corrigée}$ = 1.02^{E-6} ; n = 3375).

Au sein de la population des non-salariés étudiés, l'ALD « Tumeurs malignes » compte 106 pathologies CIM-10 différentes faisant de cette ALD la deuxième ayant le plus de pathologies CIM-10 différentes, après les ALD « Autres affections graves et Polypathologies ». En regardant à un niveau intermédiaire de précision (regroupement de cancers de la CIM-10), on peut remarquer que les non-salariés sont davantage répartis dans les regroupements de pathologies CIM-10 « C60-C63 : Tumeurs malignes des organes génitaux de l'homme » (n = 6816), « C15-C26 : Tumeurs malignes des organes digestifs » (n = 4438) et « C50 : Tumeur maligne du sein » (n = 4125) (Figure 46). Par ailleurs, on observe une représentation relativement faible des pathologies « C00-C14 : Organes respiratoires et intrathoraciques » (n = 708) par rapport à l'ensemble des non-salariés étudiés. Puis, si on regarde par sexe, on remarque qu'on a un sexe-ratio de 2 hommes pour 1 femme pour quasiment toutes les tumeurs confondues, équivalent à celui retrouvé dans l'ensemble des non-salariés étudiés (Tableau 18). De plus, à titre d'exemple, il n'est pas étonnant de voir que les tumeurs malignes du sein (« C50 ») sont davantage retrouvées chez les femmes (98.3%).

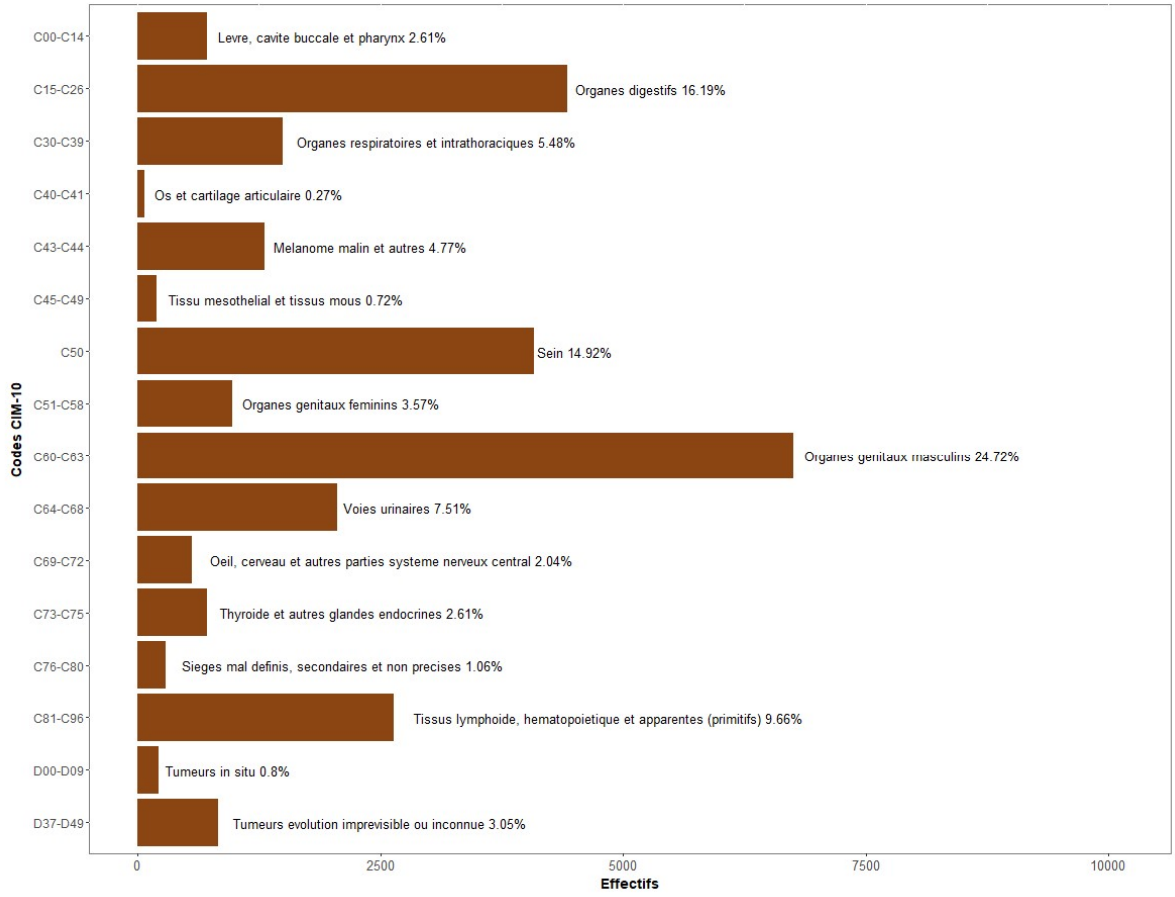


Figure 46 : Répartition des non-salariés de la MSA ayant une déclaration d'ALD « Tumeurs malignes » par regroupement de pathologies CIM-10 (2012-2016)

Tableau 18 : Répartition des non-salariés étudiés de la MSA par regroupement de pathologies CIM-10 (« C00-D49 : Tumeurs ») et par sexe (2012-2016)

Codes de pathologies CIM-10	Localisation et types des tumeurs		Hommes		Femmes		Tous
			Effectifs	%	Effectifs	%	Effectifs
C00-C14	Tumeurs malignes	Lèvre, cavité buccale et pharynx	607	85.7	101	14.3	708
C15-C26		Organes digestifs	3255	73.3	1183	26.7	4438
C30-C39		Organes respiratoires et intrathoraciques	1233	81.4	282	18.6	1515
C40-C41		Os et cartilage articulaire	59	77.6	17	22.4	76
C43-C44		Mélanome malin et autres	834	62.4	503	37.6	1337
C45-C49		Tissu mésothélial et tissus mous	145	70.4	61	29.6	206
C50		Sein	70	1.7	4055	98.3	4125
C51-C58		Organes génitaux féminins	0	0	998	100	998
C60-C63		Organes génitaux masculins	6816	100	0	0	6816
C64-C68		Voies urinaires	1766	84.5	323	15.5	2089
C69-C72		Œil, cerveau et autres parties système nerveux central	396	69.5	174	30.5	570
C73-C75		Thyroïde et autres glandes endocrines	324	44.4	405	55.6	729
C76-C80		Sièges mal définis, secondaires et non précisés	207	68.3	96	31.7	303
C81-C96		Tissus lymphoïde, hématopoïétique et apparentés (primitifs)	1931	73.3	705	26.7	2636
D00-D09	Tumeurs in situ		131	59	91	41	222
D10-D36	Tumeurs bénignes		194	60.6	126	39.4	320
D37-D49	Tumeurs à évolution imprévisible ou inconnue		699	66.4	354	33.6	1053
Total			17789	66.2	9077	33.8	26866

La régression logistique a alors été effectuée pour chacune des pathologies CIM-10 de l'ALD « Tumeurs malignes » avec la sélection de variables de cette même ALD. Ces analyses ont permis de mettre en évidence 33 associations statistiquement significatives ($p\text{-valeur}_{corrigée} < 0.05$) à ce niveau de précision (Figure 47). En ce qui concerne les AUC des modèles utilisés pour chaque regroupement de pathologies, ils montrent en moyenne que les modèles ont un bon pouvoir discriminant avec la sélection de variables de l'ALD « Tumeurs malignes » ($AUC_{moyen} = 0.75$).

Parmi les associations mises en évidence, **quatre concernent un risque de déclaration plus élevé dans le secteur de la viticulture pour les regroupements de pathologies CIM-10 suivants :**

- « C15-C26 : Tumeurs malignes des organes digestifs » (OR = 1.22 [1.12 ;1.34] ; p-valeur$_{corrigée}$ = 0.0002 ; n = 632) ;

- « C30-C39 : Tumeurs malignes des organes respiratoires et intrathoraciques » (OR = 1.23 [1.06 ;1.44] ; p-valeur$_{corrigée}$ = 0.02 ; n = 216) ;

- « C43-C44 : Mélanome malin et autres tumeurs malignes de la peau » (OR = 1.32 [1.13 ;1.57] ; p-valeur$_{corrigée}$ = 0.01 ; n = 195) ;

- « C50 : Tumeur maligne du sein » (OR = 1.13 [1.03 ;1.24] ; p-valeur$_{corrigée}$ = 0.03 ; n = 604).

De même que le secteur de la viticulture, il est possible de remarquer que **le secteur agricole des entreprises de jardins, du paysagisme et du reboisement a un risque de déclaration plus élevé pour trois regroupements de pathologies CIM-10 :**

- « C00-C14 : Tumeurs malignes de la lèvre, de la cavité buccale et du pharynx » (OR = 1.72 [1.19 ;2.47] ; p-valeur$_{corrigée}$ = 0.04 ; n = 33) ;

- « C30-C39 : Tumeurs malignes des organes respiratoires et intrathoraciques » (OR = 1.44 [1.08 ;1.91] ; p-valeur$_{corrigée}$ = 0.04 ; n = 53) ;

- « C60-C63 : Tumeurs malignes des organes génitaux de l'homme » (OR = 1.35 [1.15 ;1.59] ; p-valeur$_{corrigée}$ = 0.004 ; n = 157).

On observe également une association montrant un risque de déclaration plus élevé de pathologies « C30-C39 : Tumeurs malignes des organes respiratoires et intrathoraciques » dans les entreprises de travaux agricoles (OR = 1.77 [1.24 ;2.53] ; p-valeur$_{corrigée}$ = 0.01 ; n = 31).

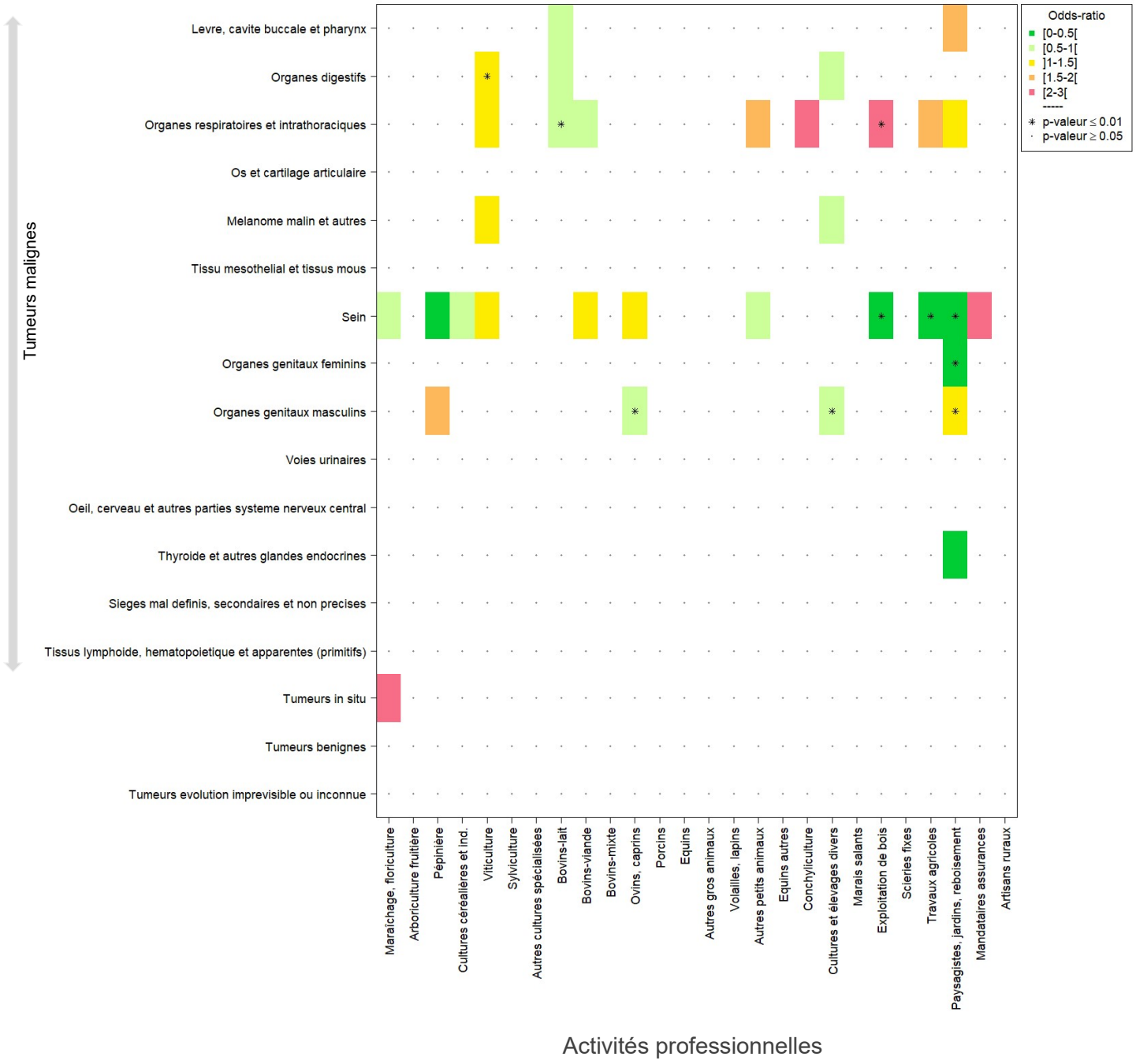


Figure 47 : Représentation graphique des odds ratios obtenus via la régression logistique entre chaque **regroupement de pathologies CIM-10 de la famille « Tumeurs »** et chaque activité professionnelle en ajustant sur d'autres paramètres notamment les comorbidités, chez les non-salariés de la MSA (2006-2016)

De façon complémentaire, quelques analyses ont été réalisées au niveau 3 digits pour cette famille de pathologies. Cependant, il est à noter que la plupart des modèles utilisés pour ces analyses, avec la sélection de variables de l'ALD « Tumeurs malignes » ont des AUC montrant des modèles peu discriminants (AUC proches ou inférieurs à 0.6). Néanmoins, pour certaines maladies, pour lesquelles les modèles ont de bons pouvoirs discriminants (AUC > 0.78), on observe des associations qui montrent par exemple :

- des risques plus élevés de déclaration pour la pathologie « C34 : Tumeurs malignes des bronches et poumons » dans les secteurs de la viticulture (OR = 1.31 [1.10 ;1.54] ; p-valeur$_{corrigée}$ = 0.01 ; n = 178), dans le secteur agricole des entreprises de jardins, du paysagisme et du reboisement (OR = 1.51 [1.11 ;2.07] ; p-valeur$_{corrigée}$ = 0.03 ; n = 43) et dans les entreprises de travaux agricoles (OR = 1.92 [1.30 ;2.84] ; p-valeur$_{corrigée}$ = 0.01 ; n = 26) ;

- un risque plus élevé de déclaration pour la pathologie « C17 : Tumeur maligne de l'intestin grêle » dans le secteur de la viticulture (OR = 2.57 [1.49 ;4.44] ; p-valeur$_{corrigée}$ = 0.02 ; n = 19) ;

- et un risque plus élevé de déclaration pour la pathologie « C61 : Tumeur maligne de la prostate » dans le secteur agricole des entreprises de jardins, du paysagisme et du reboisement (OR = 1.40 [1.17 ;1.67] ; p-valeur$_{corrigée}$ = 0.003 ; n = 130).

III. Discussion

a. Méthodologie

Les différentes analyses menées à titre d'exemple au niveau de précision de la pathologie CIM-10 ont montré qu'il était possible d'affiner les associations mises en évidence dans les parties précédentes de ce travail, entre des ALD et des activités professionnelles. Malgré la diminution du nombre de malades en fonction de la pathologie étudiée et les biais liés aux événements rares abordés précédemment, il a été possible d'obtenir des signaux, c'est-à-dire, des associations statistiquement significatives ($p\text{-valeur}_{corrigée} < 0.05$), même à ce niveau de précision. Cependant, pour des raisons d'effectifs faibles et à cause du pouvoir peu discriminant des modèles, les pathologies CIM-10 « C00-D49 » ont été principalement étudiées de façon regroupée (regroupement de la classification CIM-10). Les associations mises en évidence sur des regroupements de pathologies CIM-10 sont alors tout de même plus précises qu'au niveau de l'ALD. Les résultats montrent alors qu'à ce niveau de précision, il est possible d'obtenir une trentaine d'associations statistiquement significatives au lieu des trois associations mises en évidence au niveau de l'ALD « Tumeurs malignes ».

Par ailleurs, les faibles effectifs de « malades » considérés au niveau des pathologies CIM-10 et le nombre important de pathologies CIM-10 différentes (n > 750) au sein des données étudiées ont entrainé des difficultés techniques en ce qui concerne l'étape de sélection de variables. Cette étape étant particulièrement gourmande en temps de calcul, elle n'est actuellement pas réalisable pour l'ensemble des pathologies CIM-10 avec la méthodologie utilisée. D'autant plus que pour certaines pathologies CIM-10, une erreur technique apparaît, certainement liée aux effectifs trop faibles de la pathologie étudiée (n < 50). Cette difficulté a alors été contournée en utilisant les sélections de variables réalisées dans les parties précédentes pour chacune des ALD. La sélection de variables choisie dépend alors de l'ALD dans laquelle la pathologie CIM-10 étudiée est le plus souvent classée, tout en s'assurant de la cohérence de ce choix. De ce fait, les modèles ont ensuite été évalués à l'aide du calcul de l'aire sous la courbe ROC (AUC) sur l'échantillon de validation (30% des données) et il a été vérifié qu'on obtient pour les modèles utilisés une valeur d'AUC minimal, choisie arbitrairement à 0.60. De plus, étant potentiellement « incertains », les signaux mis en évidence avec un effectif d'individus concernés inférieur à 3 ont été écartés des résultats. En effet, il a été remarqué que pour de tels signaux, les odds ratios sont peu fiables car leurs intervalles de confiance à 95% sont très larges.

Par ailleurs, il est possible de remarquer que la sélection de variables de l'ALD « Tumeurs malignes » ne comprend actuellement pas la variable sexe. Cela peut alors influencer l'émergence ou non de signaux liés à des tumeurs qui seraient davantage présentes chez les hommes ou les femmes (ex : tumeurs malignes du sein). Or, jusqu'ici, nous avons souhaité conserver les sélections de variables construites de façon automatisée pour chaque groupe de pathologie, cette méthode de sélection de variables permettant tout de même de choisir les variables en fonction de leur association à la variable d'intérêt. Il est alors possible de faire l'hypothèse que dans une certaine mesure, la variable sexe est prise en compte dans les autres variables ajoutées aux modèles. Cependant, la sélection de variables a été réalisée au niveau de précision de l'ALD sur l'ensemble des tumeurs malignes. De ce fait, des analyses complémentaires ont été réalisées en incluant la variable sexe. Elles ont montré un gain faible voire nul en termes de pouvoir discriminant du modèle (AUC quasi-identiques) mais surtout, une perte de puissance importante faisant disparaître un grand nombre de signaux d'intérêt. En effet, pour la plupart des signaux disparaissant lors de ces analyses, les p-valeurs corrigées oscillent en réalité entre 0.01 et 0.10, et sont donc proches du seuil de significativité. Cependant, on se rend bien compte de l'importance de la prise en compte de la variable sexe pour les analyses de certaines pathologies comme les tumeurs malignes du sein. La sélection de variables devrait donc être adaptée à chaque pathologie et à l'avenir, dans le cadre du projet, il sera nécessaire de trouver une méthodologie permettant de le faire pour les analyses menées au niveau de précision de la pathologie CIM-10.

Enfin, aucune latence ou temps d'exposition n'ont été pris en compte dans les analyses. Or, il est connu dans la littérature que certaines maladies comme les cancers sont associées à un délai d'apparition important (plusieurs années) après une exposition à une ou plusieurs nuisances. Pour autant, ce délai de latence est complexe à estimer et à prendre en compte dans les analyses. D'une part, aucune donnée précise de latence ou de durée d'exposition n'existe pour chacune des maladies. D'autre part, d'un point de vue pratique, les données de la MSA permettent de n'avoir seulement qu'un historique professionnel et médical sur la période d'observation mise à disposition, ne permettant pas concrètement de réaliser des analyses avec des latences supérieures à 10 ans. Pour mémoire, la prise en compte d'un temps de latence a tout de même été conceptualisée (Annexe 2). Cependant, les hypothèses étant trop importantes (si l'on considère que le travailleur a réalisé la même activité professionnelle depuis son inscription à la MSA) et la puissance trop faible si on se limite aux seules données à disposition dans le cadre de ce travail, les résultats risquent d'être peu fiables. A terme, dans le cadre du projet global, il sera nécessaire d'explorer si d'autres pistes permettent d'inclure ces paramètres dans les analyses.

b. Synthèse des principaux résultats

Les analyses menées au niveau de précision de la pathologie CIM-10 pour les ALD « Affections psychiatriques », « Alzheimer et autres démences » et « Maladies auto-immunes (ScS, Lupus, PAN) » ont permis de mettre en évidence des associations statistiquement significatives plus précises.

En effet, lors de l'analyse des pathologies CIM-10 de l'ALD « Affections psychiatriques », il est possible de remarquer deux associations statistiquement significatives montrant un risque plus important de déclarations d'« Épisodes dépressifs » aussi bien chez les éleveurs de bovins (lait) (OR = 1.20 [1.09 ;1.33] ; $p\text{-valeur}_{corrigée}$ = 0.002 ; n = 646) que chez les éleveurs de bovins (viande) (OR = 1.28 [1.16 ;1.41] ; $p\text{-valeur}_{corrigée}$ = 1.77^{E-5} ; n = 524). Ces associations sont alors d'autant plus cohérentes avec le taux de suicide relativement élevé observé chez les éleveurs bovins, une population confrontée à d'importantes difficultés financières, par Santé Publique France entre 2007 et 2009 (118). Ainsi, à ce niveau de précision, le signal sanitaire est plus précis et d'autant plus intéressant qu'il montre davantage que ces secteurs agricoles sont particulièrement touchés par des pathologies de type « dépression » qui sont un facteur de risque majeur de suicide. En effet, selon l'OMS, 80% des personnes mettant fin à leurs jours présenteraient des signes de dépression (150). Par ailleurs, la question essentielle du suicide dans le milieu agricole a été soulevée encore récemment dans le débat médiatique et politique (151,152) et la MSA propose depuis quelques années des plans de prévention du suicide avec la mise en place d'actions de prévention[15] (153). Toutefois, toutes les dépressions ne sont pas déclarées en ALD et pour repérer plus efficacement et plus précocement les dépressions, il faudrait analyser les données de consommation médicamenteuse, notamment la consommation de psychotropes. En termes de prévention, les analyses ainsi agrémentées seraient un réel atout pour permettre une réaction précoce.

En ce qui concerne les analyses menées sur les pathologies CIM-10 de l'ALD « Alzheimer et autres démences », elles ont permis de confirmer que l'association mise en évidence précédemment concerne bien un risque plus élevé d'avoir des déclarations de démences de la maladie d'Alzheimer dans le secteur agricole des cultures céréalières et industrielles (OR = 1.27 [1.13 ;1.44] ; $p\text{-valeur}_{corrigée}$ = 0.002 ; n = 784), secteur agricole particulièrement exposé aux pesticides. Or, comme évoqué précédemment, un lien probable

[15] Exemples d'actions de prévention mises en place par la MSA : informations (magazine BIMSA), cellules pluridisciplinaires de prévention afin de repérer, d'accompagner et d'orienter les agriculteurs en difficulté, numéro vert Agri'écoute, …

entre l'exposition aux pesticides et l'apparition de la maladie d'Alzheimer a été montré dans la littérature (109) et notamment par une étude de cohorte menée sur des personnes âgées en France (63). Par ailleurs, on trouve également une association montrant un risque plus élevé de pathologies « F01 : Démence vasculaire » dans les entreprises de travaux agricoles (OR = 4.87 [1.95 ;12.14] ; p-valeur$_{corrigée}$ = 0.02 ; n = 5). Or, la démence vasculaire est souvent liée à une hypertension non traitée ou mal contrôlée (154). Pour rappel, au cours des analyses menées au niveau de la pathologie ALD, on trouve une association avec une p-valeur en limite de significativité, montrant un risque plus élevé de déclaration d'ALD « HTA sévère » dans les entreprises de travaux agricoles (OR = 1.30 [1.02 ;1.64] ; p-valeur$_{corrigée}$ = 0.09 ; n = 78).

En ce qui concerne les analyses menées sur les pathologies CIM-10 de l'ALD « Maladies auto-immunes (ScS, Lupus, PAN) », rappelons l'association entre la pathologie « Autres atteintes systémiques du tissu conjonctif » et l'activité professionnelle « élevage de bovins (lait) » (OR = 1.63 [1.31 ;2.04] ; p-valeur$_{corrigée}$ = 0.0004 ; n = 105) qui renforce l'hypothèse précédemment évoquée. En effet, dans la littérature scientifique, il a été évoqué un rôle potentiel d'une réponse immunitaire aux protéines du lait de vache dans la pathogenèse de la maladie de Behçet (120), faisant partie de la pathologie CIM-10 « Autres atteintes systémiques du tissu conjonctif » concernée par l'association. En outre, dans le cadre de ce travail, il n'est pas possible de préciser davantage cette association, n'ayant pas plus de précision au niveau de la pathologie. Par ailleurs, dans le cadre d'un travail complémentaire, il pourrait être envisagé de décrire plus précisément ces individus, notamment leurs consommations médicamenteuses. D'ailleurs, le champ des maladies auto-immunes liées aux expositions professionnelles reste peu exploré et mal connu, en particulier le rôle éventuel du risque biologique. Malgré la complexité du sujet, il est donc d'autant plus intéressant de porter une attention particulière à ces pathologies dans un contexte de vigilance des risques professionnels chez les travailleurs agricoles.

Enfin, pour ce qui est des analyses menées sur les regroupements de pathologies CIM-10 de la famille « C00-D49 : Tumeurs », on remarque un certain nombre d'associations statistiquement significatives concernent deux secteurs agricoles : la viticulture et le secteur du paysagisme, des entreprises de jardins et de reboisement.

En effet, on trouve trois associations dans le secteur de la viticulture avec un risque plus élevé de déclarations de tumeurs malignes des organes respiratoires et intrathoraciques (OR = 1.23 [1.06 ;1.44] ; p-valeur$_{corrigée}$ = 0.02 ; n = 216), de la peau (OR = 1.32 [1.13 ;1.57] ; p-valeur$_{corrigée}$ = 0.01 ; n = 195) et du sein (OR = 1.13 [1.03 ;1.24] ; p-valeur$_{corrigée}$ = 0.03 ; n = 604). Or, comme dit précédemment, ce secteur agricole aurait été exposé à des pesticides arsenicaux jusque 2001, un facteur de risque majeur de développement de cancers

respiratoires, de la peau mais aussi de la vessie (110). Néanmoins, il n'y a pas d'associations entre les tumeurs malignes des voies urinaires et la viticulture. Cela peut suggérer une potentielle sous-déclaration de ces cancers dans ce secteur, associée ou non au fait qu'un autre cancer soit déjà déclaré en ALD, ou que ces cancers soient davantage déclarés en maladies professionnelles (hors données ALD). En effet, en pratique, si la maladie est prise en charge en maladie professionnelle, elle ne figurera pas dans les bases ALD, les prises en charge étant mutuellement exclusives. Aussi, à terme, les analyses devront inclure les données des maladies professionnelles, travaux qui seront menés dans le cadre du projet.

Par ailleurs, en ce qui concerne le risque élevé de déclarations de cancer du sein dans le secteur de la viticulture, il est à noter que la consommation excessive d'alcool ou l'inactivité physique font partis des facteurs de risque d'apparition de ce type de cancer (155). Or, pour rappel, il a été montré dans les parties précédentes des associations entre le secteur de la viticulture et les ALD « Affections du foie (dont cirrhoses) » et « Diabètes », ayant les mêmes facteurs de risques. Pour ce secteur agricole, les associations mises en évidence montrent ainsi qu'on capte aussi bien des facteurs de risques professionnels, que des facteurs de risques comportementaux. Comme pour d'autres associations mises en évidence dans le cadre de ce travail, les investigations relatives à la pertinence des signaux nous portent à des facteurs de risques purement professionnels, et d'autres facteurs potentiellement indirectement liés à la profession, qu'ils soient comportementaux ou socio-économiques. Malheureusement, il n'est actuellement pas possible de pouvoir corriger les analyses avec des facteurs de confusion importants comme le tabac ou l'alcool, paramètres indisponibles au sein des données médico-administratives fournies par la MSA.

Pour ce qui est du secteur agricole du paysagisme et des entreprises de jardins et de reboisement, ils sont concernés par trois associations avec un risque plus élevé de déclarations de tumeurs malignes de la lèvre, de la cavité buccale et du pharynx (OR = 1.72 [1.19 ;2.47] ; p-valeur$_{corrigée}$ = 0.04 ; n = 33), des organes respiratoires et intrathoraciques (OR = 1.44 [1.08 ;1.91] ; p-valeur$_{corrigée}$ = 0.04 ; n = 53) et des organes génitaux de l'homme (OR = 1.35 [1.15 ;1.59] ; p-valeur$_{corrigée}$ = 0.004 ; n = 157). Les travailleurs de ce secteur agricole sont particulièrement exposés au soleil, un facteur de risque d'apparition de cancers de la cavité buccale et de la lèvre (50,156). Cependant, d'une part, aucune association statistiquement significative n'apparaît avec les cancers de la peau dont l'un des facteurs de risque principaux est aussi l'exposition au soleil (51). D'autre part, il a été montré précédemment des associations entre ce secteur agricole et les ALD « Coronaropathies » et « VIH et Immunodéficiences », indiquant que les travailleurs de ce secteur agricole pouvaient avoir davantage de comportements à risque (addictions, alcool, …). Or, une consommation de

tabac ou une consommation excessive d'alcool sont aussi des facteurs de risque d'apparition des cancers de la cavité buccale et de la lèvre, comme le fait d'avoir un système immunitaire affaibli (156). Par ailleurs, de la même façon, la consommation de tabac est aussi un facteur de risque de cancers des bronches et du poumon (157). Néanmoins, en ce qui concerne l'association entre ce secteur agricole et les tumeurs des organes génitaux de l'homme, il est intéressant de remarquer que l'association mise en évidence montre spécifiquement un risque plus élevé de déclarations de cancers de la prostate, au 1er rang des cancers masculins. Or, l'exposition aux pesticides a été suggéré comme une cause possible d'apparition de ce type de cancer dans le monde agricole (24,35,73).

Enfin, afin d'optimiser la méthodologie de génération de signaux, certains regroupements de pathologies CIM-10 seraient pertinents et pourraient être proposés pour des analyses ultérieures. Le groupe de travail chargé d'analyser la pertinence des signaux pourra être sollicité dans cet optique. A titre d'exemple, les pathologies suivantes pourraient être regroupées : « F00 : Démence de la maladie d'Alzheimer » et « G30 : Maladie d'Alzheimer ». Ces regroupements pourraient alors permettre l'émergence de nouveaux signaux d'intérêt.

IV. Conclusion

Ces analyses menées au niveau de précision de la pathologie CIM-10 ont permis de mettre en évidence des associations particulièrement intéressantes bien qu'elles puissent capturer aussi bien des facteurs de risques professionnels que des facteurs de risques comportementaux (exemples : alcool, tabac, sédentarité, …) ou socio-économiques (exemple : difficultés financières) indirectement liés à la profession. Malgré les biais mis en évidence, les modèles ont été conçus de la manière la plus robuste (sélection de variables, validation croisée) par rapport aux données mises à disposition par la MSA, n'étant à l'origine pas construites pour ce type d'analyses. Ainsi, les associations statistiquement significatives pourront et devront être investiguées à l'aide d'un groupe d'experts associant une pluralité de compétences : médico-administratives (médecins conseil de la MSA), médicales et relatives aux réalités professionnelles de terrain (médecins du travail de la MSA), toxicologiques, épidémiologiques et sociales. Par ailleurs, il sera nécessaire de refaire les analyses sur l'ensemble des pathologies CIM-10, idéalement de façon non regroupée. Cependant, une erreur technique ne permet actuellement pas d'effectuer une sélection de variables pour chaque pathologie étudiée, certainement due au fait que les effectifs de malades sont trop faibles à ce niveau de précision (CIM-10). D'autres méthodologies statistiques peuvent être utilisées pour résoudre ce problème (cf. Partie 4 III. D. Comparaison des méthodologies). Par ailleurs, certaines perspectives évoquées dans la partie Discussion générale et perspectives pourraient permettre d'augmenter le nombre de malades. Réaliser une sélection de variables spécifique à chaque CIM-10 permettrait alors d'obtenir des résultats plus robustes.

PARTIE 6

Estimation des expositions aux produits phytosanitaires

I. Introduction

Dans les parties précédentes, l'activité professionnelle des non-salariés de la MSA avait été utilisée en tant que *proxy* de l'exposition professionnelle, n'ayant pas à notre disposition d'informations plus précises sur les expositions professionnelles au sein des données de la MSA. Cependant, il est possible de constater que cette variable ne comporte que 26 modalités dans lesquelles elle s'avère peu précise, notamment pour ce qui est du secteur agricole « culture céréalières et industrielles » (une seule modalité). D'autre part, les expositions ne sont pas identiques pour une même activité, variant notamment selon la localisation géographique, l'année d'exploitation et possiblement la taille de l'exploitation voire les pratiques individuelles. Pour ces raisons, il est nécessaire d'évaluer la possibilité de compléter les données de la MSA par des données renseignant les expositions aux produits phytopharmaceutiques selon le type de culture. L'objectif secondaire consiste ensuite à croiser les expositions aux produits phytopharmaceutiques avec les pathologies chroniques. Néanmoins, pour ce faire, plusieurs étapes ont été nécessaires. La première étape concernant le maillage géographique du territoire national est détaillée ci-dessous.

II. Maillage géographique

a. Méthode de découpage géographique

Dans un premier temps, il est primordial de préciser la localisation géographique de chaque individu afin de pouvoir relier leur activité professionnelle (renseignée au sein des données MSA) à un type de culture et ainsi, a posteriori, à une exposition à des produits phytopharmaceutiques. Cependant, compte tenu de la sensibilité des données de la MSA et afin de réduire les risques de ré-identification des individus, il a été établi que la localisation géographique des individus fournie par les données de la MSA se limiterait à la connaissance du département de résidence des cotisants, bien que leur adresse exacte soit connue par la MSA.

Pour contourner cette limite, une nouvelle variable géographique a été imaginée en collaboration avec le GRICAD (Grenoble Alpes Recherche, Infrastructure de Calcul Intensif et de Données). Le principe est de découper le territoire national sous formes de mailles en fonction d'un seuil de population défini. Il s'agit d'un processus itératif qui divise chaque unité géographique en quatre parties tout en respectant un seuil minimum d'individus dans chaque

maille. Chaque maille possède alors un effectif proche des autres mais la superficie peut être différente d'une maille à une autre (Figure 48).

Ce processus se poursuit jusqu'à ce que chaque maille contienne un nombre homogène d'individus, supérieur ou égal au seuil fixé. Le seuil du secret statistique défini par l'INSEE impose de respecter un nombre minimal de cinq individus par entité (158). Grâce à la mise à disposition par la MSA des prévalences des ALD en population agricole, il a été possible de définir un seuil permettant de minimiser le risque de retrouver moins de cinq individus au sein d'une même maille, lors d'une étude ciblée sur une pathologie d'intérêt. **Ainsi, ce seuil a été fixé à 1 500 travailleurs agricoles par maille.** De plus, l'information sur la commune n'ayant pas été retenue dans les variables déclarées à la CNIL, il était important d'ajouter un autre paramètre à l'algorithme de découpe, de telle sorte que chaque maille contienne au minimum deux communes.

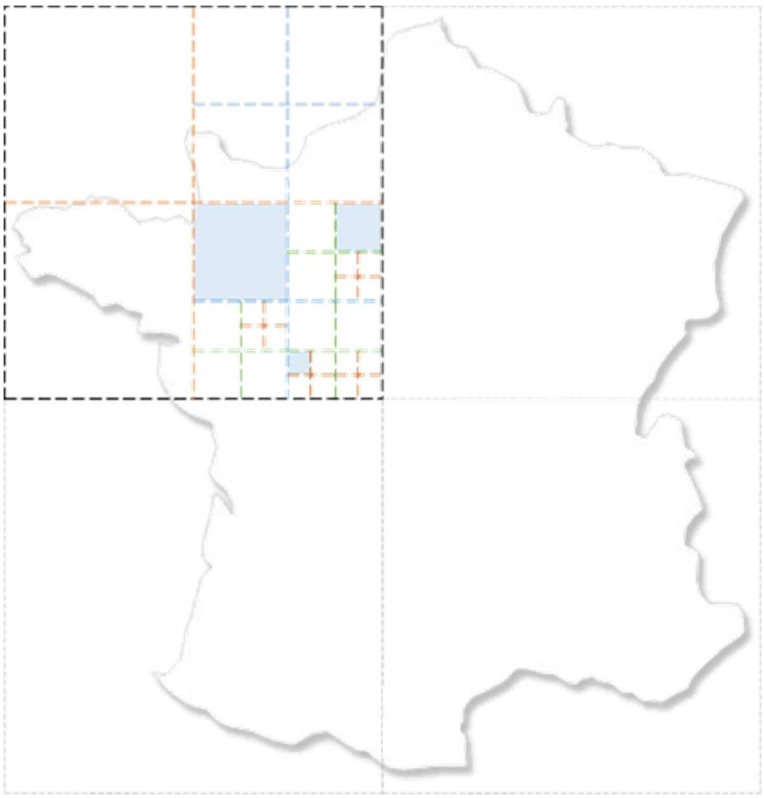
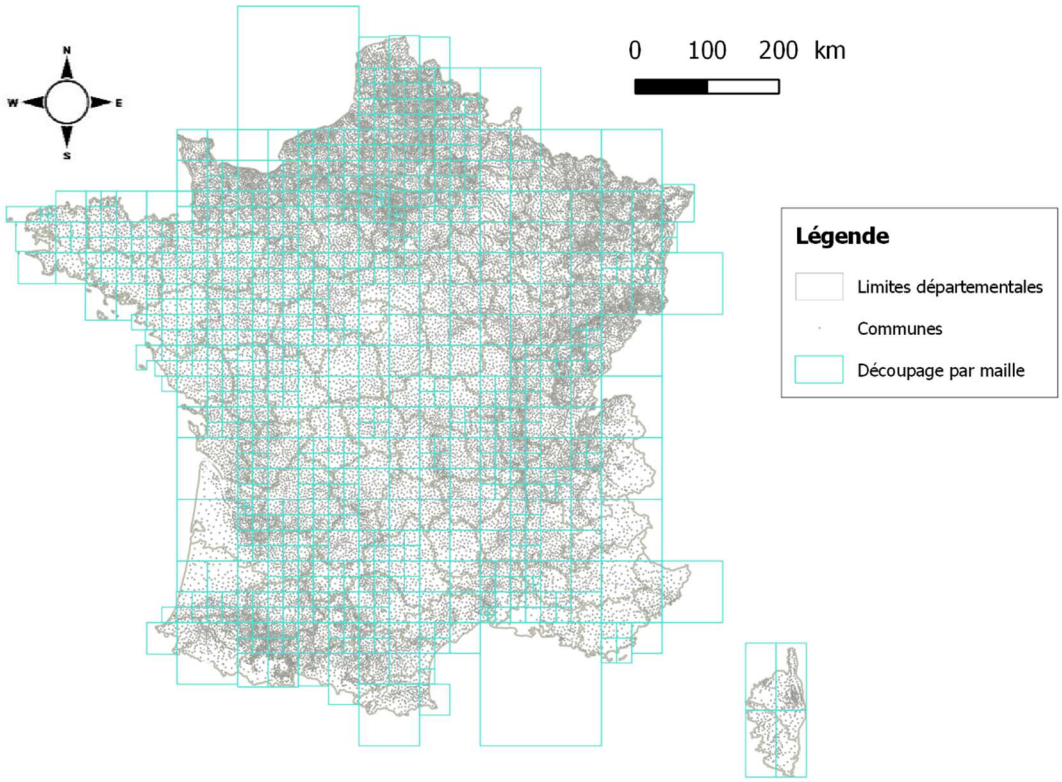


Figure 48 : Illustration du découpage du territoire national sous forme de mailles homogènes en termes de population d'agriculteurs à partir d'un seuil défini au préalable (MSA, 2014)

Il en résulte un découpage du territoire national en 717 mailles, dont les surfaces varient mais qui contiennent toutes au minimum deux communes et 1 500 travailleurs agricoles (Figure 49). Les caractéristiques du maillage obtenu sont présentées dans le Tableau 19. En étudiant la répartition de la population agricole de 2014 fournie par la MSA et utilisée pour définir le seuil de découpe, on observe que la maille contenant le moins d'agriculteurs est située dans le Sud-Est du département de l'Eure-et-Loir (28) et contient 1 528 individus, tandis que celle contenant le plus d'individus est située à l'Ouest. Elle comprend une partie des départements de la Charente-Maritime (17) et de la Gironde (33) et contient 38 903 individus.



Sources : Données MSA 2006-2016 ; Données géographiques GEOFLA® / Réalisation : C. Maugard © - Juillet 2019

Figure 49 : Maillage obtenu après application de l'algorithme de division du territoire national français à la population d'agriculteurs fournie par la MSA (2014)

Tableau 19 : Caractéristiques du maillage obtenu après application de l'algorithme de division du territoire national français à la population d'agriculteurs (MSA, 2014)

	Minimum	Maximum	Moyenne	Ecart-type
Communes par maille	2	479	50.9	47.1
Travailleurs agricoles par maille	1 528	38 903	4 303	3 285.5

Grâce au maillage ainsi généré, un numéro d'identification de maille a été attribué à chaque commune située en France. La table de correspondance a ensuite été envoyée à la MSA pour qu'elle attribue, pour chaque commune de France, le numéro de maille correspondant, et ce, pour chacune des observations enregistrées dans les données administratives fournies.

b. Analyses descriptives

Comme dans les parties précédentes de ce travail de thèse, les analyses descriptives qui suivent se sont focalisées sur la population des non-salariés agricoles affiliés à la MSA entre 2006 et 2016. Ces analyses permettent de visualiser la répartition de ces individus en utilisant le maillage géographique conçu précédemment. En effet, il est possible d'observer que la majorité des non-salariés agricoles étudiés dans ce travail est majoritairement situés dans le Sud-Ouest et le Sud-Est de la France, notamment dans les mailles suivantes : maille 1 (n = 11 599), maille 2 (n = 7 647), maille 3 (n = 6 796), maille 4 (n = 6 470) (Figure 50).

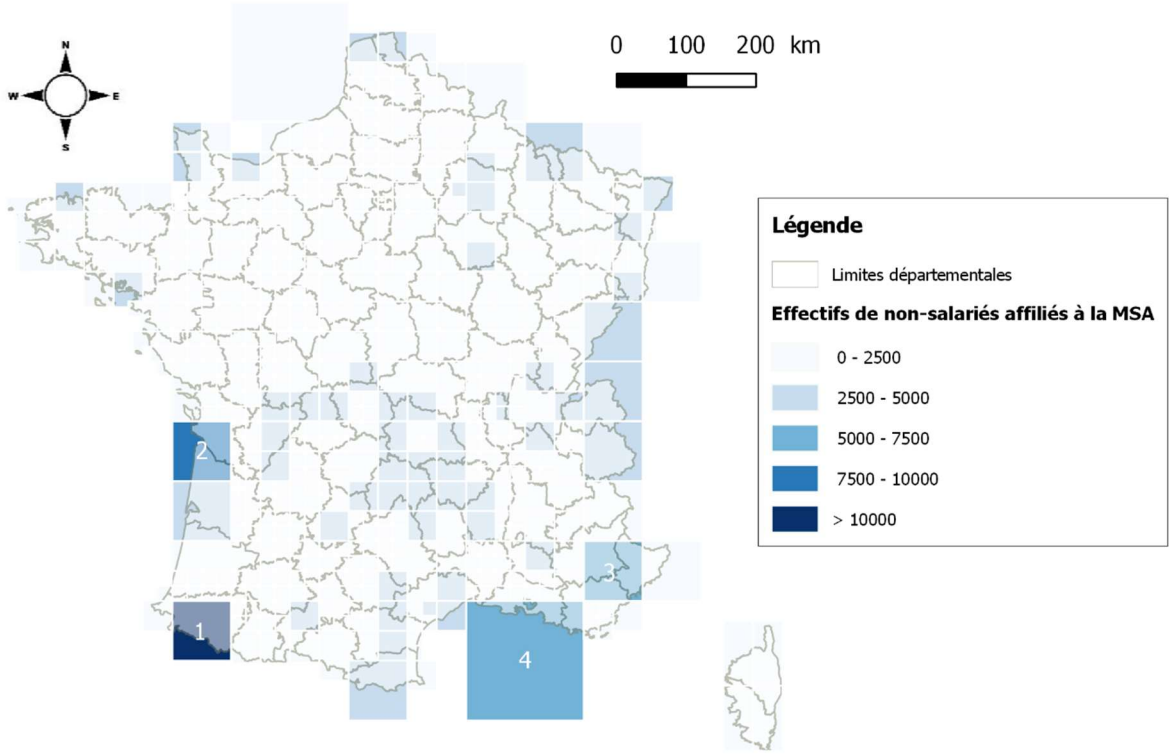


Figure 50 : Répartition géographique par maille des non-salariés affiliés à la MSA entre 2006 et 2016

Par ailleurs, une cartographie des activités professionnelles majoritaires en termes de non-salariés agricoles exerçant la profession a également été réalisée (Figure 51). Il est possible d'identifier des zones où les activités sont plus spécifiques, comme la viticulture située majoritairement au Sud-Ouest et au Sud-Est ou l'élevage de bovins laitiers situé majoritairement au Nord-Ouest et aux frontières à l'Est. D'ailleurs, cette répartition géographique des activités professionnelles par maille est semblable à la carte de répartition des orientations technico-économiques des exploitations des communes en France réalisé lors du recensement agricole de 2010 (Figure 1). **Cette carte montre alors que le maillage géographique créé permet un certain de niveau de précision tout en limitant le risque de réidentification des individus.**

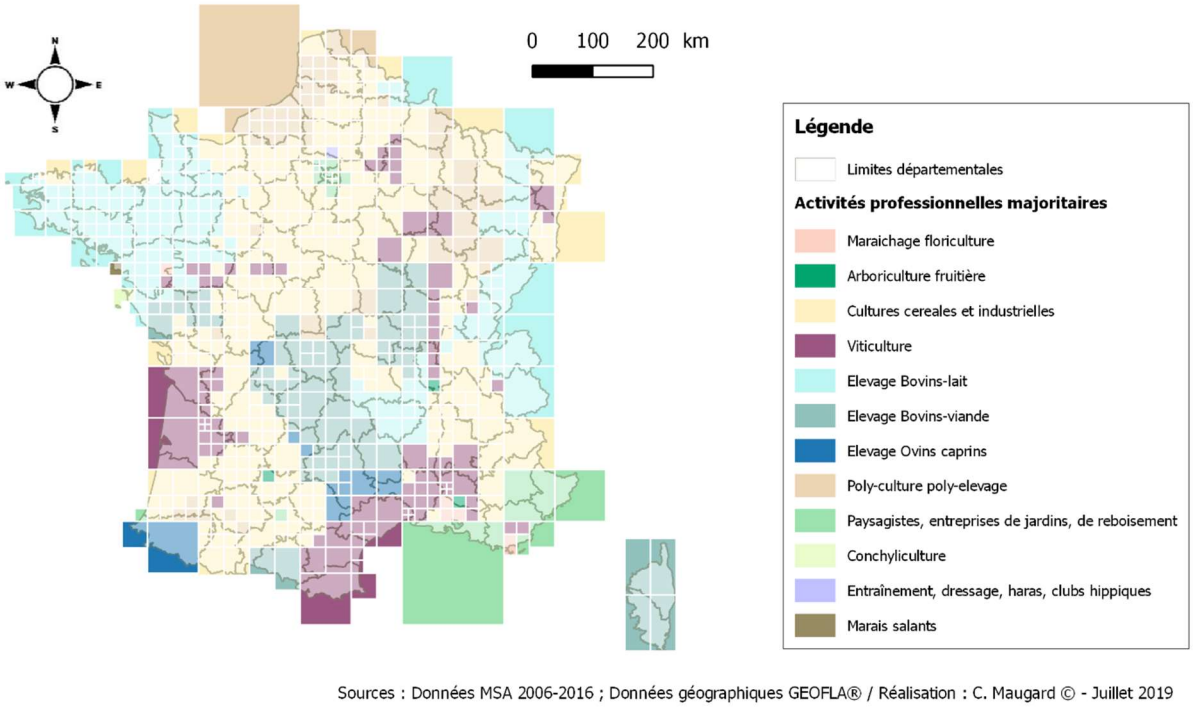


Figure 51 : Répartition géographique par maille des activités professionnelles majoritaires en termes de nombre de non-salariés agricoles exerçant la profession entre 2006 et 2016

Enfin, il est aussi possible de représenter par maille la proportion de non-salariés agricoles ayant eu au moins une déclaration d'ALD entre 2012 et 2016 (données ALD). Cette carte montre des proportions assez homogènes sur l'ensemble du territoire français, mis-à-part dans certaines mailles où la proportion de non-salariés agricoles avec une déclaration d'ALD dépasse les 15% comme au Nord-Est de la Corse (Figure 52).

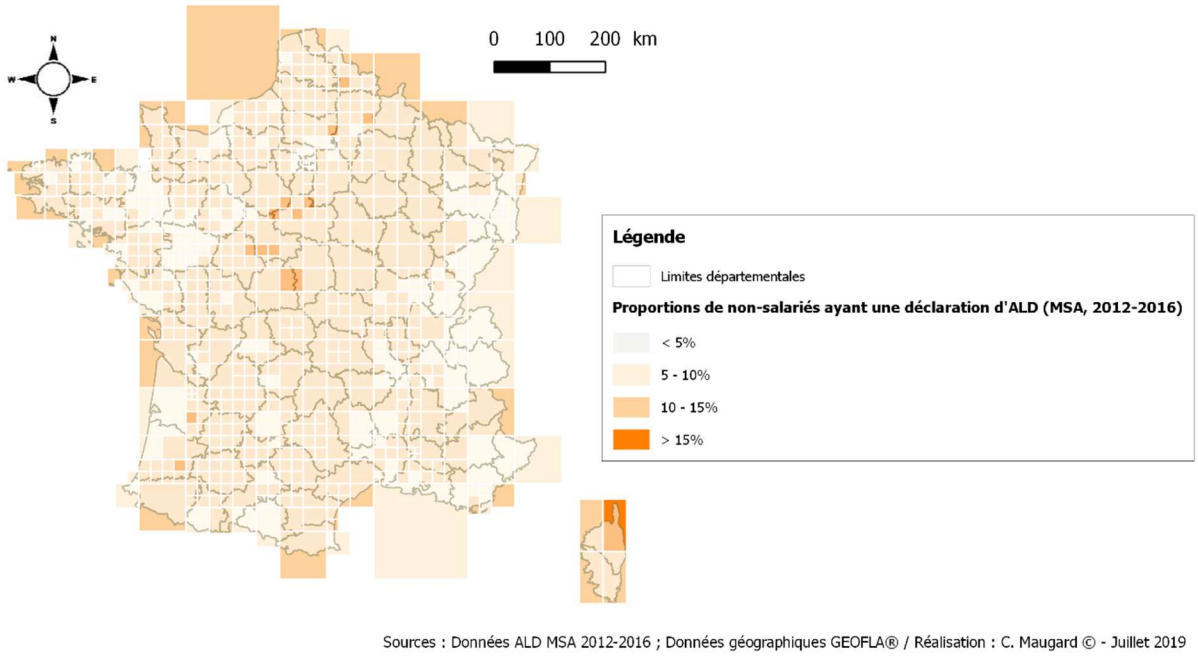


Figure 52 : Répartition géographique par maille des non-salariés agricoles ayant eu au moins une déclaration d'ALD entre 2012 et 2016 en termes de proportions

III. Article "Medico-administrative data combined with agricultural practices data to retrospectively estimate pesticide use by agricultural workers"

Se référer à la partie **Valorisation des travaux**

PARTIE 7

Discussion, perspectives et conclusion

I. Discussion générale et perspectives

Ce travail de thèse a permis de montrer la faisabilité d'utiliser les données de la MSA, qui, après fusion et traitement, ont été analysées à des fins de vigilance sanitaire des risques professionnels liés au monde agricole. En effet, ces analyses ont permis d'effectuer une recherche systématique d'associations statistiquement significatives entre le proxy d'activité professionnelle disponible et les pathologies, en prêtant attention tant à la significativité de l'association qu'à la force de ce lien.

Chacune des parties précédentes ayant déjà fait l'objet d'une discussion et d'un résumé, la discussion générale sera davantage orientée vers les perspectives induites par chacun des éléments de discussion préalablement cités.

a. Méthodes statistiques utilisées

Pour répondre aux objectifs de ce travail de thèse, nous avons choisi la régression logistique à laquelle nous avons ajouté la prise en compte des comorbidités (pathologies ALD déclarées avant la pathologie considérée) en tant que facteurs de confusion potentiels. Qu'elle soit appliquée au niveau de précision de l'ALD ou de la pathologie CIM-10, cette méthode a permis de faire émerger un certain nombre de signaux d'intérêt. Alors que certaines associations mises en évidence entre l'exposition professionnelle et la pathologie étudiée avaient déjà été suggérées par la littérature scientifique, d'autres permettent quant à elles de faire émerger des hypothèses.

Cependant, il est important de garder à l'esprit les **différents biais** évoqués qui peuvent influencer les résultats dans une certaine mesure, difficilement quantifiable. Les principaux biais sont d'ailleurs liés à l'étude d'une faible proportion de « malades » vis-à-vis du nombre de témoins, quelle que soit la pathologie étudiée. Ces biais entrainent alors nécessairement un manque de sensibilité important quand on évalue la méthodologie. Plusieurs solutions ont été proposées dans le cadre de ce travail mais aucune n'a permis d'améliorer significativement la sensibilité et ainsi, le pouvoir discriminant des modèles de régression logistique. Ainsi, pour la suite de ce travail, deux méthodologies pourront être évaluées : les modèles de survie et la technique de « bootstraping » (cf. Partie 4 III. D.). Les résultats qui seront mis en évidence par ces méthodes pourront alors être comparés à ceux obtenus via la régression logistique utilisée dans le cadre de ce travail.

Par ailleurs, contrairement à la régression logistique, les modèles de survie permettraient aussi la prise en compte de la dimension temporelle, c'est-à-dire, le temps d'observation de chaque individu mais aussi potentiellement la durée d'exposition (temps d'exercice connu d'une activité professionnelle) ou encore le temps de latence des pathologies. En effet, ces deux paramètres primordiaux n'ont pas été pris en compte dans les analyses jusqu'à présent, notamment du fait de la complexité technique et de la limite de la période d'observation des données de la MSA (2006-2016). D'ailleurs, pour progresser sur ce sujet, il faudrait pouvoir disposer de davantage d'informations sur l'historique professionnel des non-salariés, notamment entre l'année d'installation de leur exploitation et leur dernière année d'observation (année où ils sont « perdus de vue »).

b. Précision de la maladie

Dans le cadre de ce travail, seules les pathologies reconnues en ALD ont été étudiées sans tenir compte des maladies professionnelles ou des consommations médicamenteuses des individus. Or, il serait particulièrement intéressant d'avoir une vision complète de l'ensemble des individus « malades » au sein de la population des non-salariés de la MSA.

D'ailleurs, des algorithmes tels que ceux déjà développés par le ReDSiam (« Réseau pour l'utilisation des données du système national des Données de Santé ») pourraient être utilisés à cette fin, c'est-à-dire, pour repérer les pathologies à partir des consommations de médicaments renseignées dans les données RAAMSES de la MSA. Pour ce qui est d'intégrer les données de maladies professionnelles, la seule limite technique réside dans le codage des pathologies qui est très différent de celui des ALD. Ce travail est cependant en cours dans le cadre du projet.

Un travail a également été réalisé sur les données « RAAMSES » de la MSA par l'équipe « Techniques pour l'Evaluation et la Modélisation des Actions de la Santé » (ThEMAS) du laboratoire TIMC-IMAG. L'objectif de ce travail réalisé en 2018 était d'étudier l'impact à long terme de l'activité professionnelle sur la consommation d'antibiotiques chez les actifs agricoles, salariés ou non-salariés. Ce travail a d'ailleurs montré une surconsommation d'antibiotiques chez les actifs agricoles comparé à la population générale française et une consommation plus élevée chez les éleveurs comparé aux cultivateurs (159). Ce travail sera approfondi dans le cadre d'une thèse de sciences.

c. Population analysée

Les données des salariés, des retraités et des ayants-droit n'ont pour le moment pas été intégrées aux analyses, comportant chacune des limites à leur utilisation.

Les données des **salariés** sont d'autant plus complexes qu'elles comportent un nombre d'individus cinq fois plus important dont un certain nombre de saisonniers cumulant les contrats courts. Or, comme dit précédemment, chaque observation dans les données correspond à un individu dans chacun de ses contrats ce qui nécessite un travail minutieux et important de compréhension et de nettoyage de ces données.

En ce qui concerne les données des **retraités**, la principale limite concerne le fait que ces dernières ne comportent aucune information sur l'historique professionnel des individus. Nous ne pouvons donc pas à l'heure actuelle étudier ces individus bien que 80% des ouvrants-droit ayant des déclarations d'ALD soient retraités.

Enfin, pour ce qui est des données des **ayants droit**, notamment les conjoints et enfants des non-salariés, il n'est actuellement pas possible de rattacher les ayants droit à l'ouvrant droit, avec les données fournies actuellement par la MSA. Or, on sait par exemple que le conjoint et/ou les enfants peuvent collaborer à l'exploitation du non-salarié sans pour autant avoir le statut à la MSA de « conjoint-collaborateur » ou « aide familial ». Il serait alors possible de faire l'hypothèse qu'ils partagent des expositions professionnelles analogues.

d. Précision de l'activité professionnelle, « proxy » de l'exposition professionnelle

Les expositions professionnelles ont été approchées en utilisant le thésaurus interne de la MSA, renseignant l'activité professionnelle via 26 codes pour les non-salariés. Or, cela nous offre seulement une vision macroscopique de l'exposition professionnelle, notamment du fait que certaines modalités comportent une diversité importante d'activités professionnelles et autant d'expositions qui y sont associées (exemple : modalité « Cultures céréalières et industrielles » qui regroupe aussi bien la culture de blé que la culture de tournesol ou encore de betterave sucrière). Il est donc nécessaire de trouver une méthodologie permettant de préciser cette activité professionnelle. Un premier travail a été réalisé en ce sens dans la Partie 6 de ce manuscrit de thèse et montre qu'il est possible, à partir des données de la MSA et de sources externes de données (notamment le Registre Parcellaire Graphique et des matrices « culture x expositions ») d'estimer les expositions professionnelles et notamment, dans une certaine mesure, les expositions aux produits phytopharmaceutiques.

Par ailleurs, en ce qui concerne les salariés, il serait possible d'obtenir une précision sur leurs activités professionnelles en obtenant le **code « NOSTA »** (nomenclature des situations de travail en agriculture) qui est un code métier informatisé depuis 1999. Ce code est enregistré au sein de bases de données de santé au travail dont la gestion se fait au niveau local, c'est-à-dire au niveau des différentes caisses de la MSA. D'après nos informations, ces données pourraient renseigner aussi d'autres informations pertinentes relatives aux postes, aux nuisances mais aussi possiblement aux expositions (durée, intensité, métrologie, équipements de protection, …) lorsqu'elles sont disponibles. Et surtout, elles pourraient renseigner sur l'historique professionnel (avec un niveau d'antériorité qu'il nous reste à préciser). La quantification de l'exposition associée à chaque nuisance pourrait alors être renseignée avec des ordres de grandeur en termes de durée d'exposition et d'intensité. Le code « NOSTA » est alors renseigné en local par les services de santé au travail de la MSA pour les salariés agricoles (et exceptionnellement pour les non-salariés). Cependant, ces données seraient assez hétérogènes du fait d'un remplissage non optimal sur le territoire et d'une variabilité de remplissage par les médecins. Par ailleurs, ces données sont d'autant plus intéressantes qu'elles contiendraient possiblement des informations sur les facteurs de risques tels que la consommation de tabac, d'alcool ou de drogue mais aussi des données métrologiques (exemples : niveau sonore, concentration atmosphérique de polluants, …). En vue d'intégrer les salariés aux analyses, ajouter ces données serait réellement un atout pour la mise en évidence de signaux. Cependant, pour rendre cela possible, il serait nécessaire d'obtenir l'accord des médecins chefs des différentes caisses de la MSA, de faire en sorte que ces données puissent être centralisées à un niveau national et enfin, d'avoir l'accord de la CNIL. Ce dernier serait demandé au travers d'un avenant détaillant l'utilisation qui serait faite de ces données.

e. Investigation des signaux et utilisation des analyses à des fins de vigilance

On rappelle que l'objectif global du projet de recherche, qui s'inscrit dans une démarche de vigilance, vise à rechercher des informations sur des déterminants professionnels de santé chez les travailleurs agricoles. Ces analyses, menées sans *a priori*, sont là pour générer des hypothèses, en recherchant des associations statistiques ou « signaux ». Dans le cadre de ce travail de thèse, les différents signaux d'intérêt que la méthodologie retenue a permis de mettre en évidence ont alors permis de générer quelques hypothèses. Ces signaux et ces hypothèses seront alors évalués à l'aide d'un groupe d'experts pluridisciplinaires comme évoqué précédemment. Ce groupe d'experts sera notamment composé de :

- médecins conseils de la CCMSA pour leur maitrise des subtilités de l'utilisation et de l'évolution des différents codages internes à la MSA et leurs connaissances relatives aux déclarations des ALD chez les travailleurs agricoles (certains paramètres inconnus pourraient être source de biais) ;

- médecins du travail MSA pour leur connaissance des risques professionnels, du terrain et des spécificités géographiques ;

- d'autres professionnels de la MSA (préventeurs en particulier) qui seront sollicités pour leurs connaissances des pratiques agricoles ou des utilisations de produits phytopharmaceutiques par exemple ;

- et d'experts épidémiologistes, toxicologues et selon les besoins d'autres champs (sociologie du monde agricole ou rural par exemple, dans la mesure où l'on a perçu que certains excès de risque semblaient portés par des populations aux caractéristiques sociales spécifiques).

Les signaux évalués comme les plus pertinents devront faire l'objet d'investigations plus poussées ou d'études spécifiques, afin de mieux caractériser la nature des risques avant d'envisager de possibles actions de prévention.

L'intérêt de la méthodologie développée serait ensuite de pouvoir l'appliquer aux données recueillies en routine, afin d'avoir un « outil » pour montrer aux décideurs et scientifiques, « où regarder », et éviter de « passer à côté » de phénomènes sanitaires qui n'auraient pas encore été mis en évidence. La perspective d'un usage au fil de l'eau, sur des données actualisées, permettrait possiblement une certaine réactivité dans la mise en évidence de ces phénomènes. Cet « outil » permettrait aussi de venir compléter un dispositif de vigilance s'appuyant déjà sur plusieurs outils existants (dispositif de phytopharmacovigilance, Phyt'attitude, Réseau National de Vigilance et de Prévention des Pathologies Professionnelles RNV3P, données épidémiologiques de la cohorte AGRICAN, …).

En particulier, le RNV3P rassemble au sein de ses données, un résumé systématique des consultations réalisées dans les centres de consultation de pathologies professionnelles (CCPP) depuis 2000. Un travail de fouille de données (application de méthodes de pharmacovigilance) est également mené, visant à mettre en évidence de nouvelles associations entre : pathologie et exposition ; pathologie et métier, pathologie et secteur d'activité (160–162). Parmi ces associations, certaines concernent le milieu agricole. Cependant, le renseignement des données du RNV3P repose sur le recrutement des patients

adressés en consultation, et ne permet pas de caractériser la population source, et donc le risque. Il n'est donc pas possible de savoir si l'excès de fréquence d'une association mise en évidence à partir des données du RNV3P, est lié à un biais de recrutement, à un biais d'information, ou à un réel excès de pathologies sur le terrain. Aussi, il sera particulièrement intéressant de savoir si des « signaux » mis en évidence à partir des données du RNV3P, sont confirmés en population générale agricole via la fouille des données de la MSA ; ou à l'inverse, si un signal capté à partir des données de la MSA est traduit par des consultations, voire un signal au sein des CCPP. Dans ce cas, il sera possible de creuser plus en avant les expositions et les hypothèses, en retournant aux dossiers médicaux si besoin.

Si l'utilisation des données de la MSA à des fins de vigilance était validée, il pourrait alors être décidé de faire évoluer les informations et les codages réalisés au sein des données sources MSA, afin qu'elles puissent être davantage informatives pour cette mission (exemples : conserver l'historique des activités professionnelles pour les retraités, apporter une meilleure précision à l'activité professionnelle).

II. Conclusion

Cette thèse s'inscrit dans un projet novateur réunissant la MSA, des chercheurs du laboratoire TIMC-IMAG et des agences sanitaires, l'ANSES et Santé Publique France, à des fins de vigilance des risques professionnels chez les travailleurs agricoles.

Pour la première fois, les bases de données administratives et médico-administratives de la MSA ont été nettoyées pour être fusionnées et restructurées afin d'y appliquer des modèles visant à rechercher de façon systématique des associations entre activités professionnelles (« proxy » de l'exposition) et pathologies (ici, l'ALD considéré comme un « proxy » de la pathologie). La faisabilité a ainsi été démontrée et des améliorations pourront être apportées, tant sur les modèles utilisés que sur les « proxy » d'exposition et de pathologie.

Au total, ce travail a montré qu'il était possible d'utiliser les données de la MSA, non conçues à la base pour l'épidémiologie, pour la mise en évidence d'un certain nombre de signaux d'intérêt, sans faire d'hypothèses *a priori*, entre activités professionnelles et pathologies déclarées en ALD. L'évaluation de la pertinence de ces signaux est une étape supplémentaire à mener par un groupe pluridisciplinaire. Certains de ces signaux nécessiteront probablement des investigations plus poussées, qui pourront donner lieu à des études spécifiques.

Annexes

Liste des Annexes

Annexe 1 : Affections de longue durée

Tableau 20 : Liste des 32 affections de longue durée (ALD) conçue par l'Assurance Maladie

Code	Libellé
1	Accident vasculaire cérébral (AVC) invalidant
2	Insuffisances médullaires et autres cytopénies chroniques
3	Artériopathie chronique avec manifestations ischémiques
4	Bilharziose compliquée
5	Insuffisance cardiaque grave, troubles du rythme graves, cardiopathies valvulaires graves, cardiopathies congénitales graves
6	Maladies chroniques actives du foie et cirrhoses
7	Déficit immunitaire primitif grave nécessitant un traitement prolonge et infection par le VIH (Virus de l'immunodéficience humaine)
8	Diabète de type 1 et diabète de type 2
9	Forme grave des affections neurologiques et musculaires (dont myopathie), épilepsie grave
10	Hémoglobinopathies, hémolyses, chroniques constitutionnelles et acquises sévères
11	Hémophilies et affections constitutionnelles de l'hémostase graves
12	Hypertension artérielle (HTA) sévère
13	Maladie coronaire
14	Insuffisance respiratoire chronique grave
15	Maladie d'Alzheimer et autres démences
16	Maladie de Parkinson
17	Maladies métaboliques héréditaires nécessitant un traitement prolongé spécialisé
18	Mucoviscidose
19	Néphropathie chronique grave et syndrome néphrotique primitif
20	Paraplégie
21	Périarthrite noueuse (PAN), lupus érythémateux aigu disséminé, sclérodermie généralisée évolutive (ScS)
22	Polyarthrite rhumatoïde évolutive grave
23	Affections psychiatriques de longue durée
24	Rectocolite hémorragique et maladie de Crohn évolutive
25	Sclérose en plaques
26	Scoliose structurale évolutive (dont l'angle est égal ou supérieur à 25 degrés) jusqu'à maturation rachidienne
27	Spondylarthrite ankylosante grave
28	Suites de transplantation d'organe
29	Tuberculose active, Lèpre
30	Tumeur maligne, affection maligne du tissu lymphatique ou hématopoïétique
31	Affection grave hors liste, nécessitant des soins continus d'une durée prévisible supérieure à 6 mois
32	Polypathologie, plusieurs affections entrainant un état pathologique invalidant, nécessitant des soins continus d'une durée prévisible supérieure à 6 mois

Annexe 2 : Conceptualisation de la prise en compte de temps de latence

Le temps de latence correspond au délai d'apparition nécessaire après une exposition à une ou des nuisances pour qu'un effet néfaste se manifeste (ex. : apparition d'une pathologie). Par exemple, il est communément admis que le temps de latence est d'environ dix à vingt ans pour une exposition à l'amiante et l'apparition d'une asbestose (163). Le temps de latence est alors un facteur important à prendre en compte dans les analyses statistiques lorsque cela est possible, en particulier pour les cancers ou d'autres maladies chroniques. Cependant, l'intégration de ce facteur dans les analyses se heurte à plusieurs difficultés.

Tout d'abord, la détermination des temps de latence des pathologies nécessite un travail bibliographique important qui ne permettrait pas d'obtenir beaucoup d'informations puisqu'il y a extrêmement peu de données existantes sur le sujet. Il est important de souligner que les temps de latence sont différents d'une pathologie à une autre. De plus, il existe une forte variabilité interindividuelle, due à des facteurs génétiques, environnementaux et comportementaux, qui joue un rôle prépondérant dans le développement et les délais d'apparition des pathologies. Les temps de latence retenus seraient donc nécessairement entachés d'une incertitude importante dont la magnitude est difficilement quantifiable.

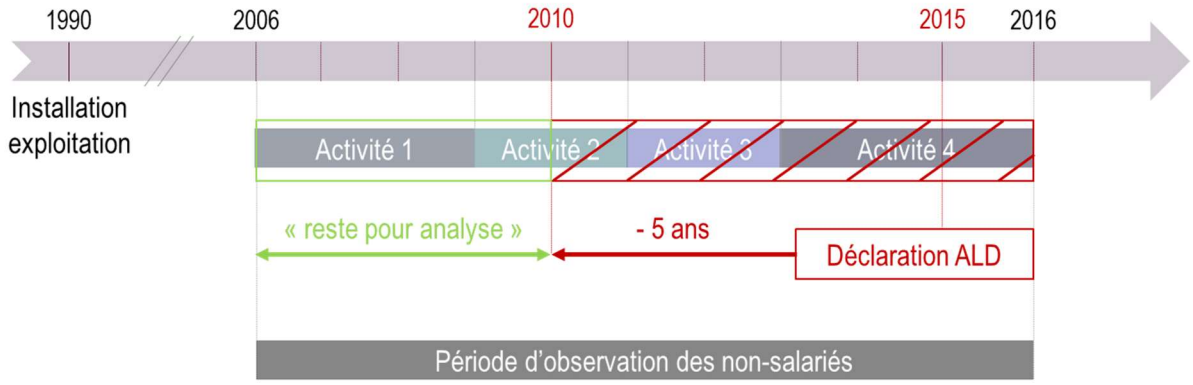
Il faut noter qu'il serait également impossible de prendre en compte des temps de latence d'apparition de pathologie supérieurs à 11 ans avec les données actuellement mises à disposition par la MSA (période s'étendant de 2006 à 2016). Au regard de ces différents obstacles, le potentiel d'intégrer le temps de latence des pathologies dans les analyses parait assez limité.

Bien que son potentiel soit limité dans le cadre de ce travail, une conceptualisation de la méthodologie de la prise en compte de ce paramètre a tout de même été établie. **Dans cet exemple théorique, un temps de latence de 5 ans a été utilisé à titre d'illustration, quelle que soit la pathologie étudiée. Sont alors considérés comme individus « malades », les individus ayant au moins une déclaration d'ALD au cours de la période d'observation. Les autres individus ont été considérés comme « témoins ». Pour rappel, la période couverte par les données ALD se situe entre 2012 et 2016. De plus, les décisions ont été prises de sorte que les modifications de structure de l'emploi agricole n'affectent pas différemment les « malades » et les « témoins ».**

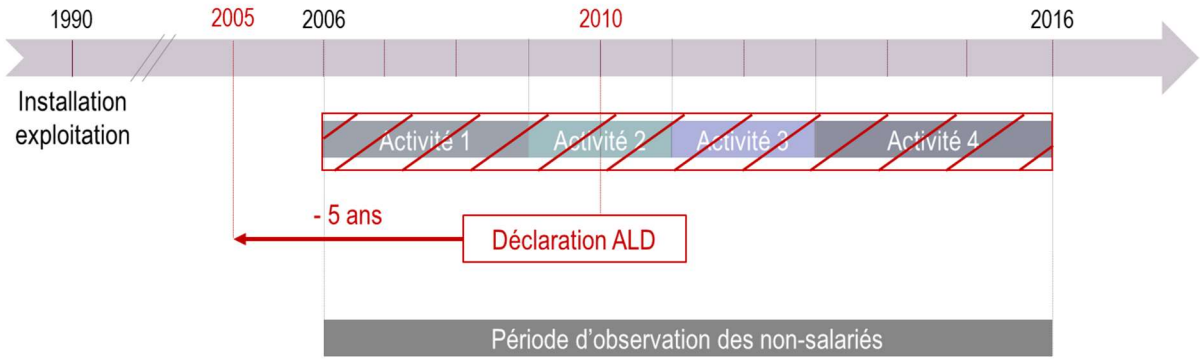
Différents cas de figure peuvent alors être distingués :

- **Pour les individus « malades »**

Cas n°1 : On retranche cinq années d'observation avant l'année de déclaration en ALD. Il « reste » alors quatre années d'observation. Dans ce cas de figure, les activités professionnelles n°1 et n°2 peuvent donc être prises en compte pour les analyses.

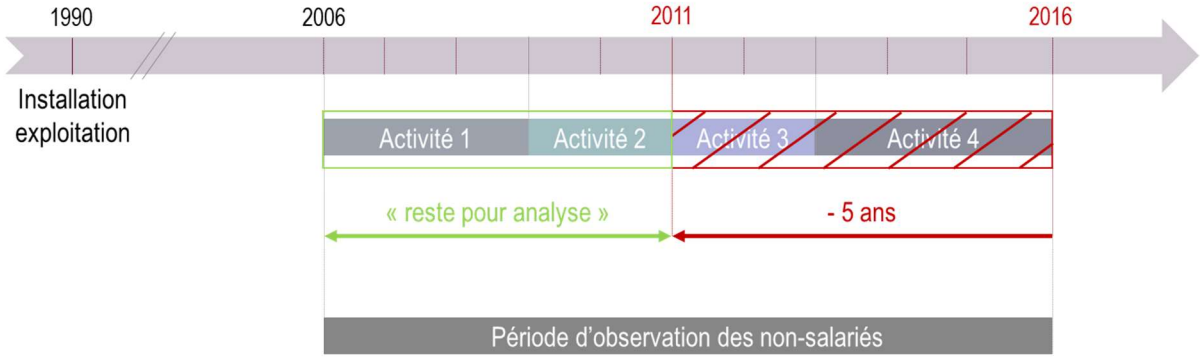


Cas n°2 : On retranche cinq années d'observations avant l'année de déclaration en ALD. Cependant, le curseur se situe alors en dehors de la période d'observation. Dans ce cas de figure, l'individu est alors considéré comme « non malade » car nous ne disposons d'aucune information sur le(s) activités professionnelles exercées par l'individu avant la période d'observation (2006-2016). Ce choix peut potentiellement faire diminuer le nombre de signaux mais évite de faire des suppositions (invention d'information) et évite alors de mettre en évidence des associations erronées.

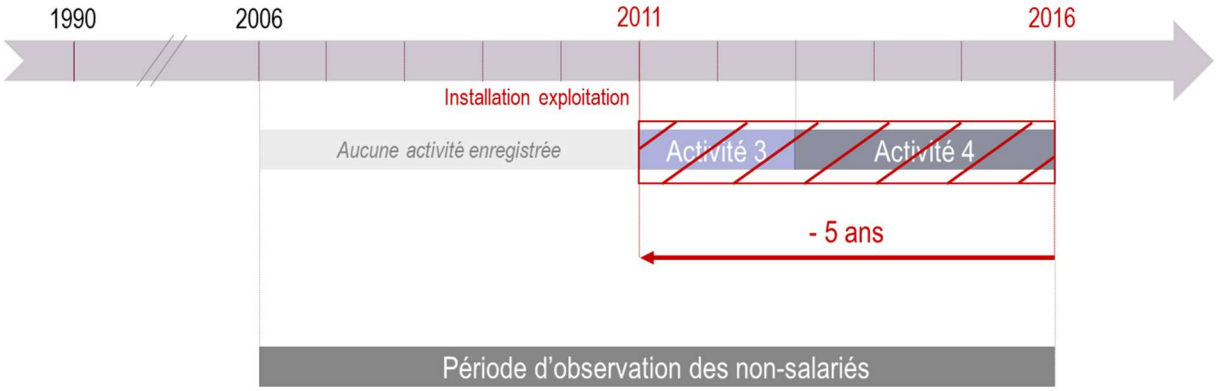


- **Pour les individus « témoins »**

Cas n°1 : On retranche cinq années d'observations avant la date de censure (date à laquelle l'individu est « perdu de vue », qui peut correspondre aussi à la fin de la période d'observation). Il « reste » alors six années d'observation. Dans ce cas de figure, les activités professionnelles n°1 et n°2 peuvent donc être prises en compte pour cet individu dans nos analyses.



Cas n°2 : On retranche cinq années d'observations avant la date de censure. Cependant, dans ce cas de figure, le non-salarié, s'étant installé dans son exploitation à partir de 2011, nous n'avons aucune information sur le(s) activités professionnelles exercées avant cette date. Ces individus « témoins » sont alors exclus des analyses.



Références bibliographiques

1. Food and Agriculture Organization (FAO). FAOSTAT [en ligne]. [consulté le 11 avril 2019]. Disponible sur : http://www.fao.org/faostat/en/#home

2. International Labour Organization (ILO). Employment by sector [en ligne]. 2018, [consulté le 11 avril 2019]. Disponible sur : https://www.ilo.org/ilostat/

3. Food and Agriculture Organization (FAO). L'alimentation et l'agriculture. Les moteurs du Programme pour le développement durable à l'horizon 2030. Rome : FAO, 2017, 40 p.

4. Food and Agriculture Organization (FAO). Construire une vision commune pour une alimentation et une agriculture durables : principes et approches. Rome : FAO, 2014, 56 p.

5. Ministère de l'agriculture et de l'alimentation. 30 projets pour une agriculture compétitive & respectueuse de l'environnement. Paris : Ministère de l'agriculture et de l'alimentation, 2015, 70 p.

6. EUROSTAT. EUROSTAT [en ligne]. [consulté le 15 avril 2019]. Disponible sur : https://ec.europa.eu/eurostat/fr/home

7. Agreste (Ministère de l'agriculture et de l'alimentation). GraphAgri 2018 [en ligne]. 2019, [consulté le 17 août 2019]. Disponible sur : http://agreste.agriculture.gouv.fr/publications/graphagri/article/graphagri-2018

8. Mutualité Sociale Agricole (MSA). Chiffres utiles de la MSA. Edition 2018. Bobigny : MSA, 2018, 42 p.

9. PURSEIGLE F. Le nouveau capitalisme agricole : de la ferme à la firme. Paris : Presses de Sciences Po, 2017, 259 p.

10. Agreste (Ministère de l'agriculture et de l'alimentation). Données de vente des produits phytopharmaceutiques 2016-2017. Chiffres et Données Agriculture, 2019, no 5, 14 p.

11. International Labour Organization (ILO). La sécurité et la santé dans l'agriculture. Genève : ILO, 2011, 402 p.

12. BLAIR A, ZAHM SH, PEARCE NE, et al. Clues to cancer etiology from studies of farmers. Scandinavian Journal of Work, Environment & Health, 1992, vol. 18, no 4, p. 209-215.

13. ALAVANJA MCR, SANDLER DP, LYNCH CF, et al. Cancer incidence in the agricultural health study. Scandinavian Journal of Work, Environment & Health, 2005, vol. 31, no suppl 1, p. 39-45.

14. NGUYEN THY, BERTIN M, BODIN J, et al. Multiple Exposures and Coexposures to Occupational Hazards Among Agricultural Workers: A Systematic Review of Observational Studies. Safety and Health at Work, 2018, vol. 9, no 3, p. 239-248.

15. WHITE G, CESSNA A. Occupational Hazards of Farming. Canadian Family Physician, 1989, vol. 35, p. 2331-2336.

16. SZESZENIA-DABROWSKA N, ŚWIATKOWSKA B, WILCZYNSKA U. Occupational diseases among farmers in Poland. Medycyna Pracy, 2016, vol. 67, no 2, p. 163-171.

17. KRISTENSEN P, ANDERSEN A, IRGENS LM, et al. Incidence and risk factors of cancer among men and women in Norwegian agriculture. Scandinavian Journal of Work, Environment & Health, 1996, vol. 22, no 1, p. 14-26.

18. LEON ME, BEANE FREEMAN LE, DOUWES J, et al. AGRICOH: A Consortium of Agricultural Cohorts. International Journal of Environmental Research and Public Health, 2011, vol. 8, no 5, p. 1341-1357.

19. International Agency for Research on Cancer (IARC). AGRICOH: A Consortium of Agricultural Cohorts [en ligne]. 2019, [consulté le 11 avr 2019]. Disponible sur : http://agricoh.iarc.fr/

20. BROUWER M, SCHINASI L, FREEMAN LEB, et al. Assessment of occupational exposure to pesticides in a pooled analysis of agricultural cohorts within the AGRICOH consortium. Occupational and Environmental Medicine, 2016, vol. 73, no 6, p. 359-367.

21. LEON ME, SCHINASI LH, LEBAILLY P, et al. Pesticide use and risk of non-Hodgkin lymphoid malignancies in agricultural cohorts from France, Norway and the USA: a pooled analysis from the AGRICOH consortium. International Journal of Epidemiology, 2019 March 18. pii: dyz017.

22. ALAVANJA MCR, SANDLER DP, MCMASTER SB, et al. The Agricultural Health Study. Environmental Health Perspectives, 1996, vol. 104, no 4, p. 362-369.

23. LERRO CC, KOUTROS S, ANDREOTTI G, et al. Cancer incidence in the Agricultural Health Study after 20 years of follow-up. Cancer Causes & Control, 2019, vol. 30, no 4, p. 311-322.

24. KOUTROS S, BEANE FREEMAN LE, LUBIN JH, et al. Risk of total and aggressive prostate cancer and pesticide use in the Agricultural Health Study. American Journal of Epidemiology, 2013, vol. 177, no 1, p. 59-74.

25. SHRESTHA S, PARKS CG, GOLDNER WS, et al. Pesticide Use and Incident Hypothyroidism in Pesticide Applicators in the Agricultural Health Study. Environmental Health Perspectives, 2018, vol. 126, no 9, p. 097008.

26. STARLING AP, UMBACH DM, KAMEL F, et al. Pesticide use and incident diabetes among wives of farmers in the Agricultural Health Study. Occupational and Environmental Medicine, 2014, vol. 71, no 9, p. 629-635.

27. KAMEL F, TANNER C, UMBACH D, et al. Pesticide exposure and self-reported Parkinson's disease in the agricultural health study. American Journal of Epidemiology, 2007, vol. 165, no 4, p. 364-374.

28. KRISTENSEN P, ANDERSEN A, IRGENS LM. Hormone-dependent cancer and adverse reproductive outcomes in farmers' families--effects of climatic conditions favoring fungal growth in grain. Scandinavian Journal of Work, Environment & Health, 2000, vol. 26, no 4, p. 331-337.

29. NORDBY KC, ANDERSEN A, KRISTENSEN P. Incidence of lip cancer in the male Norwegian agricultural population. Cancer Causes & Control, 2004, vol. 15, no 6, p. 619-626.

30. TUAL S, CLIN B, LEVEQUE-MORLAIS N, et al. Agricultural exposures and chronic bronchitis: findings from the AGRICAN (AGRIculture and CANcer) cohort. Annals of Epidemiology, 2013, vol. 23, no 9, p. 539-545.

31. TUAL S, LEMARCHAND C, BOULANGER M, et al. Activités agricoles et risque de cancers chez les affiliés du régime agricole. Résultats intermédiaires de l'étude AGRICAN (AGRIculture et CANcers). Innovations Agronomiques, 2015, no 46, p. 136-146.

32. BALDI I, CARLES C, BLANC-LAPIERRE A, et al. A French crop-exposure matrix for use in epidemiological studies on pesticides: PESTIMAT. Journal of Exposure Science & Environmental Epidemiology, 2017, vol. 27, no 1, p. 56-63.

33. LEVEQUE-MORLAIS N, TUAL S, CLIN B, et al. The AGRIculture and CANcer (AGRICAN) cohort study: enrollment and causes of death for the 2005-2009 period. International Archives of Occupational and Environmental Health, 2015, vol. 88, no 1, p. 61-73.

34. LEMARCHAND C, TUAL S, LEVEQUE-MORLAIS N, et al. Cancer incidence in the AGRICAN cohort study (2005-2011). Cancer Epidemiology, 2017, vol. 49, p. 175-185.

35. LEMARCHAND C, TUAL S, BOULANGER M, et al. Prostate cancer risk among French farmers in the AGRICAN cohort. Scandinavian Journal of Work, Environment & Health, 2016, vol. 42, no 2, p. 144-152.

36. PIEL C, POUCHIEU C, TUAL S, et al. Central nervous system tumors and agricultural exposures in the prospective cohort AGRICAN. International Journal of Cancer, 2017, vol. 141, no 9, p. 1771-1782.

37. POUCHIEU C, PIEL C, CARLES C, et al. Pesticide use in agriculture and Parkinson's disease in the AGRICAN cohort study. International Journal of Epidemiology, 2018, vol. 47, no 1, p. 299-310.

38. Mutualité Sociale Agricole (MSA). Phyt'attitude [en ligne]. [consulté le 19 août 2019]. Disponible sur : https://www.msa.fr/lfy/sst/phyt-attitude

39. BENEZET L, SPINOSI J, CHAPERON L, et al. Exposition agricole aux phytosanitaires : croisement d'une matrice culture-expositions de Matphyto avec la cohorte Coset-MSA. Archives des Maladies Professionnelles et de l'Environnement, 2015, vol. 76, no 4, p. 400.

40. GEOFFROY-PEREZ B, BENEZET L, SANTIN G, et al. Programme Coset : Cohortes pour la surveillance épidémiologique en lien avec le travail. Premier bilan de la phase pilote pour la mise en place de la cohorte d'actifs relevant du régime agricole au moment de l'inclusion - cohorte Coset-MSA. Saint Maurice : InVS, 2012, 48 p.

41. Santé Publique France. Coset - Cohortes pour la surveillance épidémiologique en lien avec le travail [en ligne]. [consulté le 19 août 2019]. Disponible sur : http://www.coset.fr/

42. ANSES. Expositions professionnelles aux pesticides en agriculture. 7 volumes. Maisons Alfort : ANSES, 2016, vol. 1 (central), 244 p.

43. INSERM. Pesticides. Effets Sur La Santé. Paris : INSERM, 2014, 1001 p.

44. TAGHAVI SM, MOKARAMI H, AHMADI O, et al. Risk Factors for Developing Work-Related Musculoskeletal Disorders during Dairy Farming. International Journal of Occupational and Environmental Medicine, 2017, vol. 8, no 1, p. 39-45.

45. KHAN MI, BATH B, BODEN C, et al. The association between awkward working posture and low back disorders in farmers: a systematic review. Journal of Agromedicine, 2019, vol. 24, no 1, p. 74-89.

46. OSBORNE A, BLAKE C, FULLEN BM, et al. Prevalence of musculoskeletal disorders among farmers: A systematic review. American Journal of Industrial Medicine, 2012, vol. 55, no 2, p. 143-158.

47. KWAKU ESSIEN S, TRASK C, KHAN M, et al. Association Between Whole-Body Vibration and Low-Back Disorders in Farmers: A Scoping Review. Journal of Agromedicine, 2018, vol. 23, no 1, p. 105-120.

48. LIE A, SKOGSTAD M, JOHANNESSEN HA, et al. Occupational noise exposure and hearing: a systematic review. International Archives of Occupational and Environmental Health, 2016, vol. 89, no 3, p. 351-372.

49. ZINK A, TIZEK L, SCHIELEIN M, et al. Different outdoor professions have different risks - a cross-sectional study comparing non-melanoma skin cancer risk among farmers, gardeners and mountain guides. Journal of the European Academy of Dermatology and Venereology, 2018, vol. 32, no 10, p. 1695-1701.

50. KACHURI L, HARRIS MA, MACLEOD JS, et al. Cancer risks in a population-based study of 70,570 agricultural workers: results from the Canadian census health and Environment cohort (CanCHEC). BMC Cancer, 2017, vol. 17, no 1, p. 343.

51. SZEWCZYK M, PAZDROWSKI J, GOLUSINSKI P, et al. Basal cell carcinoma in farmers: an occupation group at high risk. International Archives of Occupational and Environmental Health, 2016, vol. 89, no 3, p. 497-501.

52. JACKSON LL, ROSENBERG HR. Preventing heat-related illness among agricultural workers. Journal of Agromedicine, 2010, vol. 15, no 3, p. 200-215.

53. SPIEWAK R. Farmers and Farmworkers. Kanerva's Occupational Dermatology. 2e ed. JOHN SM, JOHANSEN JD, RUSTEMEYER T, et al., éditeurs. Berlin : Springer, 2012, p. 1425-1441.

54. FONTANA L, LEE SJ, CAPITANELLI I, et al. Chronic Obstructive Pulmonary Disease in Farmers: A Systematic Review. Journal of Occupational and Environmental Medicine, 2017, vol. 59, no 8, p. 775-788.

55. BARRERA C, ROCCHI S, DEGANO B, et al. Microbial exposure to dairy farmers' dwellings and COPD occurrence. International Journal of Environmental Health Research, 2019, vol. 29, no 4, p. 387-399.

56. EDUARD W, PEARCE N, DOUWES J. Chronic bronchitis, COPD, and lung function in farmers: the role of biological agents. Chest, 2009, vol. 136, no 3, p. 716-725.

57. JONES BA, GRACE D, KOCK R, et al. Zoonosis emergence linked to agricultural intensification and environmental change. Proceedings of the National Academy of Sciences of the United States of America, 2013, vol. 110, no 21, p. 8399-8404.

58. ABADIA G, CAPEK I, ANDRE-FONTAINE G, et al. Étude de séroprévalence de la chlamydiose aviaire chez certains professionnels avicoles en Bretagne et Pays de la Loire, 2001-2002. Bulletin Epidémiologique Hebdomadaire, 2006, no 27-28, p. 204-205.

59. SELEEM MN, BOYLE SM, SRIRANGANATHAN N. Brucellosis: A re-emerging zoonosis. Veterinary Microbiology, 2010, vol. 140, no 3, p. 392-398.

60. KAMEL F, UMBACH DM, BEDLACK RS, et al. Pesticide exposure and amyotrophic lateral sclerosis. Neurotoxicology, 2012, vol. 33, no 3, p. 457-462.

61. PARK RM, SCHULTE PA, BOWMAN JD, et al. Potential occupational risks for neurodegenerative diseases. American Journal of Industrial Medicine, 2005, vol. 48, no 1, p. 63-77.

62. KAB S, SPINOSI J, CHAPERON L, et al. Agricultural activities and the incidence of Parkinson's disease in the general French population. European Journal of Epidemiology, 2017, vol. 32, no 3, p. 203-216.

63. BALDI I, LEBAILLY P, MOHAMMED-BRAHIM B, et al. Neurodegenerative diseases and exposure to pesticides in the elderly. American Journal of Epidemiology, 2003, vol. 157, no 5, p. 409-414.

64. HOPPIN JA, UMBACH DM, LONG S, et al. Pesticides are Associated with Allergic and Non-Allergic Wheeze among Male Farmers. Environmental Health Perspectives, 2017, vol. 125, no 4, p. 535-543.

65. BAUMERT BO, CARNES MU, HOPPIN JA, et al. Sleep apnea and pesticide exposure in a study of US farmers. Sleep Health, 2018, vol. 4, no 1, p. 20-26.

66. CHERRY N, BEACH J, SENTHILSELVAN A, et al. Pesticide Use and Asthma in Alberta Grain Farmers. International Journal of Environmental Research and Public Health, 2018, vol. 15, no 3, p. 526.

67. RINSKY JL, RICHARDSON DB, KREISS K, et al. Animal production, insecticide use and self-reported symptoms and diagnoses of COPD, including chronic bronchitis, in the Agricultural Health Study. Environment International, 2019, vol. 127, p. 764-772.

68. MEHRPOUR O, KARRARI P, ZAMANI N, et al. Occupational exposure to pesticides and consequences on male semen and fertility: a review. Toxicology Letters, 2014, vol. 230, no 2, p. 146-156.

69. PARKS CG, MEYER A, BEANE FREEMAN LE, et al. Farming tasks and the development of rheumatoid arthritis in the agricultural health study. Occupational and Environmental Medicine, 2019, vol. 76, no 4, p. 243-249.

70. PARK S, KIM SK, KIM JY, et al. Exposure to pesticides and the prevalence of diabetes in a rural population in Korea. Neurotoxicology, 2019, vol. 70, p. 12-18.

71. KALLIORA C, MAMOULAKIS C, VASILOPOULOS E, et al. Association of pesticide exposure with human congenital abnormalities. Toxicology and Applied Pharmacology, 2018, vol. 346, p. 58-75.

72. SRITHARAN J, MACLEOD J, HARRIS S, et al. Prostate cancer surveillance by occupation and industry: the Canadian Census Health and Environment Cohort (CanCHEC). Cancer Medicine, 2018, vol. 7, no 4, p. 1468-1478.

73. ALAVANJA MCR, SAMANIC C, DOSEMECI M, et al. Use of agricultural pesticides and prostate cancer risk in the Agricultural Health Study cohort. American Journal of Epidemiology, 2003, vol. 157, no 9, p. 800-814.

74. KOUTROS S, SILVERMAN DT, ALAVANJA MC, et al. Occupational exposure to pesticides and bladder cancer risk. International Journal of Epidemiology, 2016, vol. 45, no 3, p. 792-805.

75. BOULANGER M, TUAL S, LEMARCHAND C, et al. Agricultural exposure and risk of bladder cancer in the AGRIculture and CANcer cohort. International Archives of Occupational and Environmental Health, 2017, vol. 90, no 2, p. 169-178.

76. VOPHAM T, BERTRAND KA, HART JE, et al. Pesticide exposure and liver cancer: a review. Cancer Causes Control, 2017, vol. 28, no 3, p. 177-190.

77. BONNER MR, FREEMAN LEB, HOPPIN JA, et al. Occupational Exposure to Pesticides and the Incidence of Lung Cancer in the Agricultural Health Study. Environmental Health Perspectives, 2017, vol. 125, no 4, p. 544-551.

78. BOULANGER M, TUAL S, LEMARCHAND C, et al. Lung cancer risk and occupational exposures in crop farming: results from the AGRIculture and CANcer (AGRICAN) cohort. Occupational and Environmental Medicine, 2018, vol. 75, no11, p. 776-785.

79. ZHAO G, RONDA E, CEA L, et al. Mortality by cause of death and risk behaviors in farmers versus non-farmers: the importance of avoiding the healthy worker effect. International Archives of Occupational and Environmental Health, 2019, vol. 92, no 4, p. 599-608.

80. BREW B, INDER K, ALLEN J, et al. The health and wellbeing of Australian farmers: a longitudinal cohort study. BMC Public Health, 2016, vol. 16, no 1, p. 988.

81. TELLE-LAMBERTON M, FAYE S, PONTIN F, et al. Trends in work-related mental disorders by sector in France. Occupational Medicine, 2018, vol. 68, no 7, p. 431-437.

82. GIGONZAC V, BREUILLARD E, GUSEVA-CANU I, et al. Caractéristiques associées à la mortalité par suicide parmi les hommes agriculteurs exploitants entre 2007 et 2011. Saint Maurice : Santé Publique France, 2017, 10 p.

83. KLINGELSCHMIDT J, MILNER A, KHIREDDINE-MEDOUNI I, et al. Suicide among agricultural, forestry, and fishery workers: a systematic literature review and meta-analysis. Scandinavian Journal of Work, Environment & Health, 2018, vol. 44, no 1, p. 3-15.

84. Haut Conseil de la Santé Publique. Apport des cohortes à la connaissance de la santé. Actualité et dossier en santé publique. 2012, no 78, 56 p.

85. ROYDA, Desenclos JC. Les bases de données médico-administratives : un nouveau souffle pour la surveillance en santé publique ? Bulletin Epidémiologique Hebdomadaire. 2013, no hors-série, p. 2-3.

86. Haut Conseil de la Santé Publique. Pour une meilleure utilisation des bases de données administratives et médico-administratives nationales pour la santé publique et la recherche. Paris : HCSP , 2012, 56 p.

87. RAGHUPATHI W, RAGHUPATHI V. Big data analytics in healthcare: promise and potential. Health Information Science and Systems, 2014, vol. 2, no 3.

88. BARBIER-FERAUD I, BOSSI MALAFOSSE J, BOUEXEL P, et al. Big Data et Prévention : de la prédiction à la démonstration. International think tank dedicated to big data in healthcare. Paris : Healthcare Data Institute, 2016, 80 p.

89. GREMY I, DOUSSIN A. Surveillance des maladies chroniques en France : la contribution des bases de données médico-administratives. Bulletin Epidémiologique Hebdomadaire, 2013, no hors-série, p. 9-14.

90. FONTENEAU L, LE MEUR N, COHEN-AKENINE A, et al. Apport des bases médico-administratives en épidémiologie et santé publique des maladies infectieuses. Revue d'Épidémiologie et de Santé Publique, 2017, vol. 65, p. S174-S182.

91. KOH HC, TAN G. Data mining applications in healthcare. Journal of Healthcare Information Management, 2005, vol. 19, no 2, p. 64-72.

92. TUFFERY S. Data Mining et statistique décisionnelle : L'intelligence des données. Paris : Technip, 2012, 850 p.

93. Agreste (Ministère de l'agriculture et de l'alimentation). Données en ligne [en ligne]. [consulté le 27 août 2019]. Disponible sur : http://agreste.agriculture.gouv.fr/page-d-accueil/article/agreste-donnees-en-ligne

94. MOSCONI L, BERTI V, QUINN C, et al. Sex differences in Alzheimer risk. Neurology, 2017, vol. 89, no 13, p. 1382-1390.

95. FEREC C, MERCIER B, AUDREZET MP. Les mutations de la mucoviscidose : du génotype au phénotype. Médecine/sciences, 1994, vol. 10, no 6-7, p. 631-639.

96. TUFFERY S. Modélisation prédictive et apprentissage statistique avec R. Paris : Technip, 2015, 434 p.

97. DABIS F, DESENCLOS JC, directeurs. Epidémiologie de terrain : Méthodes et applications. Montrouge : John Libbey Eurotext, 2017, 809 p.

98. INSERM. Epidémiologie. Principes et méthodes quantitatives. Paris : Tec & Doc Lavoisier, 2009, 498 p.

99. BERKSON J. Application of the Logistic Function to Bio-Assay. Journal of the American Statistical Association, 1944, vol. 39, no 227, p. 357-365.

100. BIERNAT E, LUTZ M, LECUN Y. Data science : fondamentaux et études de cas. Machine learning avec Python et R. Paris : Eyrolles, 2015, 311 p.

101. COMMENGES D, JACQMIN-GADDA H. Modèles biostatistiques pour l'épidémiologie. Louvin-la-Neuve : De Boeck Superieur, 2015, 416 p.

102. DE BOURMONT M. La résolution d'un problème de multicolinéarité au sein des études portant sur les déterminants d'une publication volontaire d'informations : proposition d'un algorithme de décision simplifié basé sur les indicateurs de Belsley, Kuh et Welsch (1980). Comptabilités et innovation, Grenoble, 2012. Disponible sur : https://hal.archives-ouvertes.fr/hal-00691156

103. LAUDE H, LAUDE E. Data Scientist et langage R : Guide d'autoformation à l'exploitation intelligente des Big Data. 2e éd. St Herblain : ENI, 2018, 811 p.

104. SCHWARZ G. Estimating the dimension of a model. Annals of Statistics, 1978, vol. 6, no 2, p. 461-464.

105. RIPLEY BD. Model Selection in Complex Classes of Models. Statistical Learning, AMSI Summer School meeting at UNSW (Université de Nouvelle-Galles du Sud), 2003. Disponible sur : http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.67.2617&rep=rep1&type=pdf

106. BENJAMINI Y, HOCHBERG Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. Journal of the Royal Statistical Society: series B (Methodological), 1995, vol. 57, no 1, p. 289–300.

107. EVANS SJ, WALLER PC, DAVIS S. Use of proportional reporting ratios (PRRs) for signal generation from spontaneous adverse drug reaction reports. Pharmacoepidemiology and Drug Safety, 2001, vol. 10, no 6, p. 483-486.

108. WILLIAMS R. Analyzing Rare Events with Logistic Regression. University of Notre Dame, 2018, 5 p.

109. MOSTAFALOU S, ABDOLLAHI M. Pesticides: an update of human exposure and toxicity. Archives of Toxicology, 2017, vol. 91, no 2, p. 549-599.

110. International Agency for Research on Cancer (IARC). Arsenic and arsenic compounds. IARC monographs on the evaluation of carcinogenic risks to humans, volume 100 C. Arsenic, metals, fibres, and dusts. Lyon : IARC, 2012, p. 41-93.

111. TUAL S, LEMARCHAND C, BOULANGER M, et al. Exposure to Farm Animals and Risk of Lung Cancer in the AGRICAN Cohort. American Journal of Epidemiology, 2017, vol. 186, no 4, p. 463-472.

112. EVANGELOU E, NTRITSOS G, CHONDROGIORGI M, et al. Exposure to pesticides and diabetes: A systematic review and meta-analysis. Environment International, 2016, vol. 91, p. 60-68.

113. XIAO X, CLARK JM, PARK Y. Potential contribution of insecticide exposure and development of obesity and type 2 diabetes. Food and Chemical Toxicology, 2017, vol. 105, p. 456-474.

114. RENU K, MADHYASTHA H, MADHYASTHA R, et al. Role of arsenic exposure in adipose tissue dysfunction and its possible implication in diabetes pathophysiology. Toxicology Letters, 2018, vol. 284, p. 86-95.

115. MARTIN EM, STYBLO M, FRY RC. Genetic and epigenetic mechanisms underlying arsenic-associated diabetes mellitus: a perspective of the current evidence. Epigenomics, 2017, vol. 9, no 5, p. 701-710.

116. PATEL CJ, BHATTACHARYA J, BUTTE AJ. An Environment-Wide Association Study (EWAS) on type 2 diabetes mellitus. PLoS One, 2010, vol. 5, no 5, p. e10746.

117. TAXELL P, SANTONEN T. Diesel Engine Exhaust: Basis for Occupational Exposure Limit Value. Toxicological Sciences, 2017, vol. 158, no 2, p. 243-251.

118. KHIREDDINE-MEDOUNI I, BREUILLARD E, BOSSARD C. Surveillance de la mortalité par suicide des agriculteurs exploitants. Situation 2010-2011 et évolution 2007-2011. Saint Maurice : Santé Publique France, 2016, 32 p.

119. Haute Autorité de Santé (HAS). Critères diagnostiques et bilan initial de la cirrhose non compliquée. Recommandations pour la pratique clinique. Paris : HAS, 2008, 5 p.

120. TRIOLO G, ACCARDO-PALUMBO A, DIELI F, et al. Humoral and cell mediated immune response to cow's milk proteins in Behçet's disease. Annals of the Rheumatic Diseases, 2002, vol. 61, no 5, p. 459-462.

121. ANCELLE T. Statistique épidémiologie. 3e éd. Paris : Maloine, 2011, 320 p.

122. BONITA R, BEAGLEHOLE R, KJELLSTROM T. Eléments d'épidémiologie. 2e éd. Genève : OMS, 2010, 233 p.

123. CAYE K, JUMENTIER B, LEPEULE J, et al. LFMM 2: Fast and Accurate Inference of Gene-Environment Associations in Genome-Wide Studies. Molecular Biology and Evolution, 2019, vol. 36, no 4, p. 852-860.

124. CAYE K. Méthodes de factorisation matricielle pour la génomique des populations et les tests d'association. Th : Modèles, méthodes et algorithmes en biologie, santé et environnement ; Grenoble ; 2017, 125 p.

125. FRICHOT E, SCHOVILLE SD, BOUCHARD G, et al. Testing for Associations between Loci and Environmental Gradients Using Latent Factor Mixed Models. Molecular Biology and Evolution, 2013, vol. 30, no 7, p. 1687-1699.

126. INSERM. Diabète de type 2 [en ligne]. [consulté le 20 juin 2019]. Disponible sur : https://www.inserm.fr/information-en-sante/dossiers-information/diabete-type-2

127. INSERM. Bronchopneumopathie chronique obstructive (BPCO) [en ligne]. [consulté le 20 juin 2019]. Disponible sur : https://www.inserm.fr/information-en-sante/dossiers-information/bronchopneumopathie-chronique-obstructive-bpco

128. HOPPIN JA, UMBACH DM, KULLMAN GJ, et al. Pesticides and other agricultural factors associated with self-reported farmer's lung among farm residents in the Agricultural Health Study. Occupational and Environmental Medicine, 2007, vol. 64, no 5, p. 334-341.

129. GUILLIEN A, PUYRAVEAU M, SOUMAGNE T, et al. Prevalence and risk factors for COPD in farmers: a cross-sectional controlled study. European Respiratory Journal, 2016, vol. 47, no 1, p. 95-103.

130. THAON I, REBOUX G, MOULONGUET S, et al. Les pneumopathies d'hypersensibilité en milieu professionnel. Revue des maladies respiratoires, 2006, vol. 23, no 6, p. 705-725.

131. CANO-JIMENEZ E, ACUNA A, BOTANA MI, et al. Farmer's Lung Disease. A Review. Archivos De Bronconeumologia, 2016, vol. 52, no 6, p. 321-328.

132. Aboyans V, Sevestre M-A, Désormais I, et al. Épidémiologie de l'artériopathie des membres inférieurs. La Presse Médicale. 2018, vol. 47, no 1, p. 38-46.

133. Institut français de recherche pour l'exploitation de la mer (Ifremer). Environnement et Ressources des Pertuis Charentais. Le Réseau d'Observation de la Contamination Chimique (ROCCH) [en ligne]. [consulté le 31 août 2019]. Disponible sur : https://wwz.ifremer.fr/lerpc/Activites-et-Missions/Surveillance/ROCCH

134. CHIOCCHETTI G, JADAN-PIEDRA C, VELEZ D, et al. Metal(loid) contamination in seafood products. Critical Reviews in Food Science and Nutrition, 2017, vol. 57, no 17, p. 3715-3728.

135. GUEGUEN M, AMIARD JC, ARNICH N, et al. Shellfish and residual chemical contaminants: hazards, monitoring, and health risk assessment along French coasts. Reviews of Environmental Contamination and Toxicology, 2011, vol. 213, p. 55-111.

136. ROBLES-OSORIO ML, SABATH-SILVA E, SABATH E. Arsenic-mediated nephrotoxicity. Renal Failure, 2015, vol. 37, no 4, p. 542-547.

137. MOODY EC, COCA SG, SANDERS AP. Toxic Metals and Chronic Kidney Disease: a Systematic Review of Recent Literature. Current Environmental Health Reports, 2018, vol. 5, no 4, p. 453-463.

138. INSERM. Insuffisance rénale [en ligne]. [consulté le 20 juin 2019]. Disponible sur : https://www.inserm.fr/information-en-sante/dossiers-information/insuffisance-renale

139. Décret n° 2011-726 du 24 juin 2011 supprimant l'hypertension artérielle sévère de la liste des affections ouvrant droit à la suppression de la participation de l'assuré mentionnée au 3° de l'article L. 322-3 du code de la sécurité sociale. Journal officiel de la République française, 2011, no 0147, texte n°9, p. 10873.

140. HASTIE T, TIBSHIRANI R, FRIEDMAN J. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. 2e éd. New York: Springer, 2009, 745 p.

141. HOERL AE. Application of ridge analysis to regression problems. Chemical Engineering Progress, 1962, vol. 58, n° 3, p. 54-59.

142. TIBSHIRANI R. Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society: Series B (Methodological), 1996, vol. 58, no 1, p. 267–288.

143. AZENCOTT CA. Introduction au Machine Learning. Malakoff : Dunod, 2018, 240 p.

144. HASTIE T, QIAN J. Glmnet vignette. Stanford university, 2016, 42 p.

145. KING G, ZENG L. Logistic regression in rare events data. Political analysis, 2001, vol. 9, no 2, p. 137–163.

146. FIRTH D. Bias reduction of maximum likelihood estimates. Biometrika, 1993, vol. 80, no 1, p. 27–38.

147. ALLISON P. Logistic Regression for Rare Events [en ligne]. Statistical Horizons. 2012, [consulté le 15 juillet 2019]. Disponible sur : https://statisticalhorizons.com/logistic-regression-for-rare-events

148. KOSMIDIS I, PAGUI ECK, SARTORI N. Bias reduction in generalized linear models [en ligne]. 2017, [consulté le 15 juillet 2019]. Disponible sur : https://pdfs.semanticscholar.org/9b6e/b5df5f1de4fe4e74e8678035e509cf301b74.pdf

149. BENJAMIN DJ, BERGER JO, JOHANNESSON M, et al. Redefine statistical significance. Nature Human Behaviour, 2018, vol. 2, no 1, p. 6-10.

150. Observatoire National du Suicide (ONS), Direction de la recherche, des études, de l'évaluation et des statistiques (DREES). Suicide. Etat des lieux et perspectives de recherche. Premier rapport. Paris : ONS, 2014, 221 p.

151. CORNU G. Taux de suicide chez les agriculteurs. Question écrite n°24706, Sénat. Journal officiel du Sénat, 2017, p. 137.

152. Ministère de l'Agriculture, de l'Agroalimentaire et de la Forêt. Taux de suicide chez les agriculteurs. Réponse à la question n°24706 de M. CORNU G, Sénat. Journal officiel du Sénat, 2017, p. 640.

153. Mutualité Sociale Agricole (MSA). Plan national MSA de prévention du suicide 2016-2020. Bobigny : MSA, 2016, 24 p.

154. HANON O. Hypertension artérielle et démences. Annales de cardiologie et d'angéiologie, 2014, vol. 63, no 3, p. 204-208.

155. Institut National du Cancer (INC). Cancer du sein : les facteurs de risque [en ligne]. [consulté le 23 juillet 2019]. Disponible sur : https://www.e-cancer.fr/Patients-et-proches/Les-cancers/Cancer-du-sein/Facteurs-de-risque/Tabac-alcool-et-surpoids

156. Société Canadienne du Cancer (SCC). Facteurs de risque du cancer de la cavité buccale [en ligne]. [consulté le 23 juillet 2019]. Disponible sur : https://www.cancer.ca:443/fr-ca/cancer-information/cancer-type/oral/risks/?region=on

157. Institut National du Cancer (INC). Cancer du poumon : Facteurs de risque [en ligne]. [consulté le 23 juillet 2019]. Disponible sur : https://www.e-cancer.fr/Patients-et-proches/Les-cancers/Cancer-du-poumon/Facteurs-de-risque

158. INSEE. Guide du secret statistique. Paris : INSEE, 2018, 9 p.

159. LOMBARDO D. Impact à long terme de l'activité professionnelle sur la consommation antibiotique chez les actifs agricoles. Sciences pharmaceutiques, 2018, 70 p.

160. BONNETERRE V, BICOUT DJ, DE GAUDEMARIS R. Application of Pharmacovigilance Methods in Occupational Health Surveillance: Comparison of Seven Disproportionality Metrics. Safety and Health at Work, 2012, vol. 3, no 2, p. 92-100.

161. BONNETERRE V, FAISANDIER L, BICOUT D, et al. Programmed health surveillance and detection of emerging diseases in occupational health: contribution of the French national occupational disease surveillance and prevention network (RNV3P). Occupational and Environmental Medicine, 2010, vol. 67, no 3, p. 178-186.

162. Réseau national de vigilance et de prévention des pathologies professionnelles. Méthodes de détection et d'expertise des suspicions de nouvelles pathologies professionnelles (« pathologies émergentes »). Maisons-Alfort : Anses, 2014, 122 p.

163. INRS. Amiante. Effets sur la santé [en ligne]. [consulté le 2 septembre 2019]. Disponible sur : http://www.inrs.fr/risques/amiante/effets-sante.html

R-1. MOISAN F, SPINOSI J, DELABRE L, et al. Association of Parkinson's Disease and Its Subtypes with Agricultural Pesticide Exposures in Men: A Case-Control Study in France. Environmental health perspectives, 2015, vol. 123, no 11, p. 1123-1129.

R-2. SCAILTEUX LM, DROITCOURT C, BALUSSON F, et al. French administrative health care database (SNDS): The value of its enrichment. Therapies, 2019, vol. 74, no 2, p. 215-223.

R-3. CHEN YC, YEH HY, WU JC, et al. Taiwan's National Health Insurance Research Database: administrative health care database as study object in bibliometrics. Scientometrics, 2011, vol. 86, no 2, p 365-380.

R-4. YAMASAKI D, TANABE M, MURAKI Y, et al. The first report of Japanese antimicrobial use measured by national database based on health insurance claims data (2011-2013): comparison with sales data, and trend analysis stratified by antimicrobial category and age group. Infection, 2018, vol. 46, no 2, p. 207-214.

R-5. TU K, CAMPBELL, CHEN ZL, et al. Accuracy of administrative databases in identifying patients with hypertension. Open medicine, 2007, vol. 1, no 1, p. e18-26.

R-6. NICHOLS GA, DESAI J, ELSTON LAFATA J, et al. Construction of a multisite DataLink using electronic health records for the identification, surveillance, prevention, and management of diabetes mellitus: the SUPREME-DM project. Preventing chronic disease, 2012, vol. 9, E110.

R-7. BERNSTEIN AB, SWEENEY MH, Centers for Disease Control and Prevention. Public health surveillance data: legal, policy, ethical, regulatory, and practical issues. MMWR Supplements, 2012, vol. 61, no 3, p. 30-34.

R-8. ERDEM E, KORDA H, HAFFER SC et al. Medicare claims data as public use files: a new tool for public health surveillance. Journal of public health management and practice, 2014, vol. 20, no 4, p. 445-452.

R-9 NDEIKOUNDAM NGANGRO N, VIRIOT D, LUCAS E. Relevance of healthcare reimbursement data to monitor syphilis epidemic: an alternative surveillance through the national health insurance database in France, 2011-2013. BMJ Open, 2018, vol. 8, no 7, p. e020336.

R-10. DELMAS MC, BOUSSAC-ZAREBSKA M, HOUOT M. L'apport des bases médico-administratives dans la surveillance des maladies respiratoires chroniques en France. Bulletin Epidémiologique Hebdomadaire, 2013, no hors-série, p. 30-35.

R-11. BLAIR A. Occupation and cancer in the Nordic countries. Acta Oncologica, 2009, vol. 48, no 5, p. 644-645.

R-12. GEORGES A, BALCAEN T, CARON A, et al. Enhancing Nationwide Medico-Administrative Databases Analysis with SAF4SUHAD: A Statistical Analysis Framework for Secondary Use of Healthcare Administrative Databases. Studies in Health Technology and Informatics, 2018, vol. 255, p 25-29.

R-13. SILENOU BC, AVALOS M, HELMER C, et al. Health administrative data enrichment using cohort information: Comparative evaluation of methods by simulation and application to real data. PLoS One, 2019, vol. 14, no 1, p. e0211118.

R-14. RICE HE, ENGLUM BR, GULACK BC, et al. Use of patient registries and administrative datasets for the study of pediatric cancer. Pediatric blood & cancer, 2015, vol. 62, no 9, p. 1495-1500.

R-15. FLAHAULT A, BAR-HEN A, PARAGIOS N. Public Health and Epidemiology Informatics. Yearbook of medical informatics, 2016, no 1, p. 240-246.

R-16. MOONEY SJ, PEJAVER V. Big Data in Public Health: Terminology, Machine Learning, and Privacy. Annual review of public health, 2018, vol. 39, p. 95-112.

Valorisation des travaux

I. Communications en congrès nationaux et internationaux

a. Communications orales

The 32nd International Congress on Occupational Health (ICOH), Dublin, May 2018

Maugard C, Bosson-Rieutort D, François O, Bonneterre V. Big data and occupational health surveillance: use of french medico-administrative databases for hypothesis generation regarding occupational risks in agriculture. *Occupational and Environmental Medicine*, 2018, vol. 75, n° suppl 2, p. A121-122.

European Congress of Epidemiology, Lyon, July 2018

Maugard C, Bosson-Rieutort D, François O, Bonneterre V. Big-data and occupational health surveillance: screening of occupational determinants of health among French agricultural workers, through data mining of medico-administrative databases. Revue d'Epidémiologie et de Santé Publique, 2018, vol. 66 n° suppl 5, p. S262-263.

b. Communications affichées

Journée de la Recherche Médicale de Grenoble, Juin 2017

Bosson-Rieutort D, Maugard C, Achard P, Duron D, Chanoine S, Bedouch P, François O, Bonneterre V. Le big data pour la santé des travailleurs agricoles : un projet multidisciplinaire. Livret des résumés, 87 p, résumé p. 39.

The 26th International Symposium on Epidemiology in Occupational Health (EPICOH), Edinburgh, August 2017

Maugard C, Cancé C, Achard P, François O, Bonneterre V, Bosson-Rieutort D. How can we avoid re-identification risk in big-data analysis? proposition of a newstrategy of geographical subdivisions using gis tools. *Occupational and Environmental Medicine*, 2017, vol. 74, n° suppl 1, p. A136-137.

Maugard C, Bosson-Rieutort D, François O, Bonneterre V. Big data and occupational health vigilance: use of french medico-administrative databases for hypothesis generation regarding occupational risks in agriculture. *Occupational and Environmental Medicine*, 2017, vol. 74, n° suppl 1, p. A74.

27ème Journées Franco-Suisses de Médecine et de Santé au travail, Annecy, Juin 2019

Maugard C, Rieutort D, François O, Ozenfant D, Bonneterre V. Big Data (données massives) et surveillance en santé-travail : étude des déterminants professionnels de santé chez les travailleurs agricoles français via la fouille de données assurantielles nationales. Carnet des résumés, 93 p, résumé p. 85-86.

II. Articles scientifiques

a. En cours de soumission

1) Maugard C, Bosson-Rieutort D*, Ozenfant D, François O, Bonneterre V. Occupational health surveillance of French agricultural workers using health insurance databases. 2019.

- Article princeps de la thèse : soumission prévue à la revue « *Occupational and Environmental Medicine* » (*co-premier auteur)

2) Vesper-Guillon B, Maugard C, Decaens T, Herve C, Bonneterre V. Digestive cancer risk and occupationnal exposure in French farmers: results from Mutualite Sociale Agricole (MSA) medico-administrative databases. 2019.

- Analyses statistiques réalisées dans ce projet de publication en rapport avec une thèse de médecine soutenue le 27/09/2019 par Baptiste Vesper-Guillon, gastro-entérologue.

3) Article en cours d'écriture sur la présentation des bases de données de la MSA (deuxième-auteur)

b. Publié

Achard P, Maugard C, Cancé C, Ozenfant D, Maitre A, Bosson-Rieutort D, Bonneterre V. Medico-administrative data combined with agricultural practices data to retrospectively estimate pesticide use by agricultural workers. *Journal Of Exposure Science And Environmental Epidemiology*. 2019 Sep 4.

- Participation à ce travail notamment au travers de l'élaboration du maillage géographique puis contribution à l'écriture et à la relecture de cet article.

Occupational health surveillance of French agricultural workers using health insurance databases

Charlotte Maugard[1,*], Delphine Bosson-Rieutort[1,*], Damien Ozenfant[2], Olivier François[1], Vincent Bonneterre[1]

Affiliations

1- Univ. Grenoble Alpes, CNRS, CHU Grenoble Alpes, Grenoble INP, TIMC-IMAG, Grenoble, France

2- Caisse centrale Mutualité Sociale Agricole (CCMSA), 19 rue de Paris, F-93000 Bobigny, France

* These authors contributed equally to this work

Corresponding authors

Vincent Bonneterre: VBonneterre@chu-grenoble.fr

Key words

Medico-administrative databases, Health insurance, Data mining, Occupational risks, Agricultural workers

Contributions

VB, helped by OF and DBR, designed the project, presented it to the former CCMSA scientific committee on occupational risks to get its approval. DBR and VB wrote the request to the French Data Protection Authority (CNIL) to get the authorization to obtain sensitive data. DBR performed the data selection and data curation. DO provided MSA data and technical support on medico-administrative aspects. CM performed data management and data analysis, with the participation of DBR. OF provided technical support on modeling. VB, CM and DBR contributed to the data interpretation. CM, DBR and VB wrote the paper. All authors discussed the results and commented on the manuscript.

Acknowledgments

We thank the Mutualité Sociale Agricole (MSA) for their support, especially Nadia Joubert (current head of MSA statistical department), Alain Pelc (former head of MSA statistical department), Marc Parmentier, Patrick Le Bourhis and Valérie Vincent (Contributors data), Nicolas Sabin (Long Term Diseases Data), Nicolas Viarouge and Sébastien Odiot (occupational accidents and diseases data), Thierry Grech (Retirees data), as well as Prof. William Dab (former chair of MSA scientific committee), Prof. Anne Laure Crémieux (former MSA national physician) and Prof. Jean Marc Soulat (current MSA national physician) for their interest and support for this work. We thank the French National Agency for health safety in food, work and environment ANSES, for the grant which allowed this project to be conducted, and especially Mathilde Merlo, Alexandra Papadopoulos, Fabrizio Botta and Jean Luc Volatier for their technical and scientific support, as well as Prof Gerard Lasfargues (ANSES scientific director). We thank the French Public Health Agency, Santé Publique France, especially Johan Spinosi, Laura Chaperon and Mounia El

Yamani for their technical support on agricultural practices. We also thank the French Data Protection Authority (CNIL) for their agreement in due time, Alison Foote for substantial language editing of the manuscript, as well as DELL society, which, as part of its policy to support medical research, has provided us with the appropriate computer equipment.

Funding

This project was funded and supported by ANSES (grant agreement N°2016-CRD-03_PPV16/ 534B) via the tax on sales of plant protection products. The proceeds of this tax are assigned to ANSES to finance the establishment of a system for monitoring the adverse effects of plant protection products, called 'phytopharmacovigilance' (PPV), established by the French Law on the future of agriculture of 13 October 2014. This project also received funding from Grenoble-Alpes-University (PhD scholarship).

Conflict of interest

The authors declare no conflict of interest

Short running title

Occupational health surveillance of French agricultural workers

List of abbreviations

CNIL: French Authority for Data Protection (Commission Nationale Informatique et Liberté)
OA: Occupational Activity
MSA: "Mutualité Sociale Agricole" (Health insurance system for French agricultural workers)
LTD: Long-Term Disease
ICD-10: International Classification of Diseases (10^{th} version)

Abstract

Background: Beyond classical epidemiological studies, the systematic analysis of massive, routinely collected, individual health insurance data represents an asset for occupational health (OH) surveillance. Both the French health insurance system dedicated to agricultural workers (MSA) and the French Health Safety Agency on Food, Work and Environment (ANSES) have decided to gradually set up a surveillance system based on these databases.

Objective: Describing without any prior assumptions all statistical associations between occupational activities (OA) and illnesses recognized by social security system as long-term disease (LTD) for OH surveillance purposes.

Methods: This study focused on 899,212 self-employed agricultural workers registered between 2006 and 2016 in MSA databases. Among them, 100,706 individuals were identified with at least one declaration of LTD. Logistic regression was used, adjusting for previous diseases and other confounding variables. Associations were characterized by odds ratio and by p-values corrected for multiple testing.

Results: The analysis revealed 54 statistically significant associations between an OA and a LTD, making it possible to capture already known or suspected health determinants, but also to generate interesting new hypotheses. Wine growers, market gardeners and floriculturists were at higher risk of reporting particular LTDs, while most breeders manifested a lower risk.

Conclusions: This work demonstrated the relevance of using health insurance data for OH surveillance. The developed methodology will be extended by using complementary data to refine disease and exposure information.

Key messages

What is already known about this subject?

Information on health consequences of occupational activities and exposures of farmers rely on epidemiological studies focusing on limited time-scales and subpopulations. Complementary methods relying on analysis of health insurance data collected routinely, and covering the target population in a comprehensive way, are emerging as new tool in epidemiological studies, but have never been used for hypotheses generation regarding work-related diseases.

What are the new findings?

This work demonstrates the relevance of analyzing health insurance data for occupational risk surveillance purposes. Without any prior assumptions, the methodology identified statistically significant associations between agricultural activities and diseases recognized as long-term diseases in the entire French agricultural workforce. These associations revealed occupational risks already known or suspected, and raised also interesting new hypotheses.

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

This method has the following advantages: 1) systematic approach, 2) high statistical power, and 3) costless data acquisition. This work will actively contribute to the health surveillance of the entire French agricultural workforce, enriching already existing schemes and cohorts. Hypotheses raised by this work can be the subject of targeted epidemiological or toxicological studies and can guide preventive actions.

Introduction

In order to better highlight occupational determinants of health, some populations such as agricultural workers, might beneficiate from specific surveillance schemes. Indeed, agricultural workers face many different occupational risks, including exposure to a large family of pesticide products implicated in the development of chronic diseases like cancers[1,2] or neurological diseases[3,4]. However, pesticides are not the only threat to the health of agricultural workers, and it is necessary to consider other chemical[5,6], physical[7,8] or biological risks[9] and psychosocial factors[10,11] such as difficult economic context, or remoteness. Until now, cohorts have been set up in order to identify occupational risks related to agriculture, using data collected via questionnaires or interviews[12,13]. Nevertheless, complementary methods based on health insurance data analysis would be useful for early detection of work-related diseases, as they rely on the whole target population (high statistical power), and on continuously updated data.

In France, health insurance covers sickness and most occupational risks, and it ensures the reimbursement of medical expenses, at 70% of most healthcare costs, and at 100% in case of costly and/or recognized long-term diseases (LTD). All workers are required to contribute to a health insurance fund and their immediate family are eligible for benefits as well. This healthcare system, characterized by the World Health Organization as one of the most efficient[14], is composed of five different healthcare funds depending on occupational activity (OA). Agricultural workers are insured by the agricultural health insurance scheme, known as “Mutualité Sociale Agricole” (MSA).

Under the initiative of its occupational health scientific committee, the MSA sought to develop its health surveillance activity regarding agricultural workers’ health by linking and crossing several of its databases, initially designed for health care reimbursement. As a first step in designing this surveillance approach, our goal was to cross administrative and medico-administrative databases and assess, without prior assumptions, associations between each OA and each disease identified through the LTD database. Therefore, this work presents our approach to use MSA databases for early detection of work-related diseases, as well as its main results.

Material and methods

Agreements

This project was approved by the French Authority for Data Protection (CNIL): agreement n°MMS/SBM/AE171001.

Study population

The active population covered by the MSA called “contributors”, is broad, including both self-employed (heads of farms or agricultural enterprises) and employees, working in various fields (crops, livestock, garden and landscaping enterprises, logging, etc). Using the MSA data, we focused on the self-employed population since they are more stable in their activity (employees frequently change activities and contracts, giving a higher noise/signal ratio when linking diseases to previous OAs), and since they have a different coding system for OAs. Considering only the contributor database with self-employed workers, we included, at the individual level, information about OAs, but also demographic (gender, date of birth, etc.) and socio-economic (incomes, social aids, etc.) characteristics. To code OAs, the MSA system uses an internal thesaurus specifying the most-at-risk OA, with 26 modalities (Supplement S1). Note that retirees were not studied because the information regarding their OAs is no longer available in the MSA databases. However, due to the annual update, we were able to include “recent retirees”, as they were identified as “active workers” in databases from previous years. At last, to be included, self-employed

workers had to be 18 years old and affiliated to the MSA, in metropolitan France, between 2006 and 2016 (n = 899,212 eligible subjects).

Illnesses recognized by the French health insurance system as requiring prolonged (and/or expensive) treatment feature in a list of 32 diseases or groups of diseases, known as LTDs (Supplement S2). To identify self-employed worker illnesses, we crossed the contributor database with the medico-administrative database with declarations of LTDs (Figure 1). Each code on this LTDs list can refer to several codes of the 10th revision of the International Classification of Diseases (ICD-10). Any individual with a LTD registered prior to available data will still be present in the database if reimbursement of costs is still going on. To avoid incriminating an exposure that occurred after the disease onset, OA starting after registration of a LTD, were not taken into account. People without information about OAs before LTD registration were not considered with the LTD for analyses concerning that specific LTD. At last, we identified 100,706 individuals with at least one declaration of LTD over the observation period.

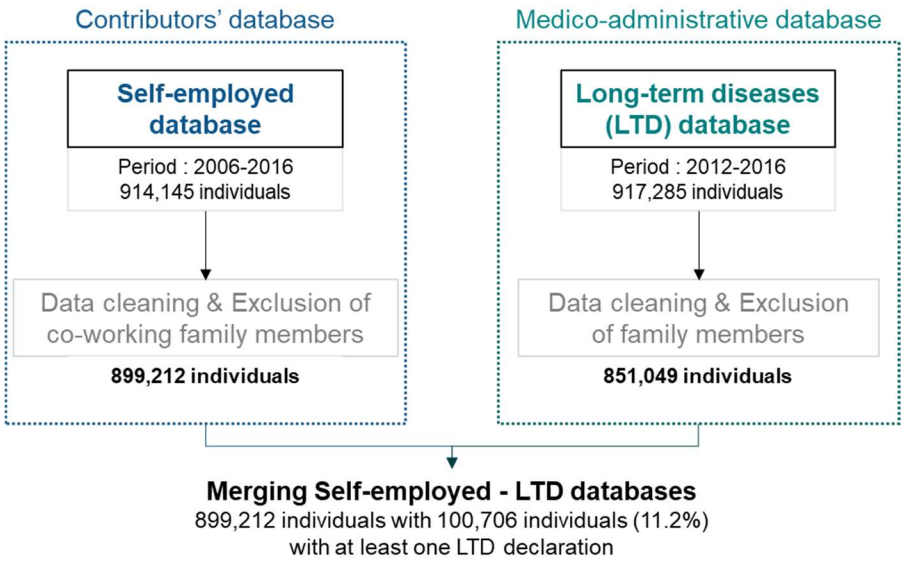


Figure 1: Illustration of the MSA contributors' and medico-administrative databases linkage

Statistical analysis

OAs and LTDs were considered as binary variables. Considering the fact that we had very low prevalence for four LTD, associations for these LTDs were not studied: "Complicated bilharzia" (n = 1), "Hemoglobinopathies" (n = 18), "Cystic fibrosis" (n = 9) and "Poly-diseases" (n = 1). Thus, the aim was to test associations between 26 OAs and 28 LTDs, which means 728 associations. To fulfill this goal, we chose a modeling tool, logistic regression, commonly used to search for associations in epidemiology[15].

As we wanted to adjust for potential confounding variables, we used all variables available in the contributors' database for self-employed workers as installation year, gender, social aids, farm area, age, number of OAs, number of observation years, income, family status, number of employees, professional status, type of the exploitation or the enterprise, administrative region. We took into account previous illnesses for individuals with LTD declaration. To do that, as the number of LTDs is high, a principal component analysis (PCA) was performed to reduce the number of dimensions while avoiding a significant loss of information[16]. We decided to keep only the first four principal components, explaining 60% of the total variance. We added these four components to all other

variables. Then, in order to select a specific set of confounding variables for each LTD, we calculated the strength of associations between these variables and each LTD and between variables, using Cramer's V method[17]. This approach was suitable to calculating the strength of associations between nominal variables, and was known to be independent of the number of modalities of variables and the population size[15]. It allowed us to make a first selection of confounding variables for each LTD.

Then, to limit model overfitting, we used a cross-validation method separating data into training (70%) and validation datasets (30%). In order to define the best fit and most parsimonious model for each LTD, we applied stepwise selection of variables and calculated the Bayesian Information Criterion (BIC) for the inclusion or exclusion of each remaining variable on training datasets. We obtained 28 specific sets of confounding variables, one for each LTD. Then, models were evaluated on validation datasets using measurements of the Area Under the Receiver operating characteristic Curve (AUC). All 728 associations were tested using an appropriate logistic regression model, where LTD was considered as the response variable, OA was considered as the explanatory variable and confounding variables were introduced as covariates in the model. Given the high number of association tests performed, we applied the Benjamini-Hochberg's procedure to correct the p-values[18]. Associations were considered statistically significant if their adjusted p-value was under the $P<0.05$ level, and they were reported as odd ratios (OR) with their 95% confidence intervals (95% CI).

As robustness analyses, we performed other statistical methods: sparse regression to improve variables selection[19], latent factor models to assess and add hidden confounders to models[20] and Firth's method to reduce bias related to rare events in analyses[21]. AUC were calculated for each method.

All data treatment and analyses were performed with R software 3.5.3®[22] for Windows 10©. Packages used on this software were the following: speedglm[23], bigstep[24], ROCR[25], lfmm[20], brglm2[26] and glmnet[27].

Results

Population characteristics

Our study included 899,212 self-employed workers (70.2% men) (Table 1), observed during an average period of seven years, with about 42.0% of individuals observed during the entire observation period (2006-2016). The average age of this population was about 50 years and was significantly different between men and women ($p < 2.2e\text{-}16$). The distribution of self-employed workers by administrative region showed that they were more settled in South-West of France.

Accounted for about 70.0% of records, five OAs (out of 26) predominated: cereals crops (23.0%), dairy farming (13.8%), mixed livestock farming and crop cultivation (11.5%), winegrowing (10.6%) and beef cattle farming (10.3%). Most had started their activity in the '90s. No OA emerged as being predominantly female.

Considering LTDs, 11.2% of farmers declared at least one LTD and were registered as having an agricultural OA before it was registered. The most common LTDs registered, irrespective of gender, were malignant neoplasms (19.9%), diabetes (19.4%), heart diseases (13.1%), coronary diseases (10.2%), and long-term psychiatric disorders (5.0%).

Table 1: Main characteristics of the self-employed population, registered by the MSA between 2006 and 2016

	Number of individuals	%
Population	899,212	
Men	631,560	70.2%
Median age (over the study period)	50.2	
Men	48.7	
Women	53.8	
Family status*		
Single	350,942	39.0%
Married	532,732	59.2%
Widower	37,940	4.2%
Divorced or separated	48,366	5.4%
Family collaborator(s)		
No partner or other adult family member	834,344	92.8%
Number of declared occupational activities (OA)		
1	816,524	90.8%
2	79,000	8.8%
3 +	3,688	0.4%
Professional status		
Working exclusively in agricultural activities	667,674	74.2%
Engaged in a commercial or craft activity or practising a liberal profession	10,463	1.2%
Engaged principally in term of working time as agricultural employees	16,959	1.9%
Main activity in an other scheme than MSA	204,574	22.8%
Type of exploitation		
Individual exploitation or enterprise	552,472	61.4%
Member of a consortium of farmers	309,666	34.3%
Individual having multiple exploitations	38,410	4.3%
Farm or enterprise area (in are)		
Median	1,920	
Number of employees by exploitation		
Mean	1.3	
Installation year		
Median	1995	
Incomes (annual gross basis in euros)		
Median	5,484	
Socials aids		
*Eligible for unemployment benefit** (at least once)*	42,545	4.7%
*Eligible for minimum income benefit** (at least once)*	31,914	3.5%
Number of observation years		
Mean	7.4	
Number of declared long-term diseases (LTD)		
None	798,506	88.8%
≥ 1	100,706	11.2%
1	77,541	77.0%
2	18,089	18.0%
3 +	5,076	5.0%

**Individuals may have changed family status during the study period. Percentages have been calculated on the denominator of the total number of self-employed and cannot be cumulated at 100%; **Provided by the state.*

Multivariate analysis

In view of the numerous associations studied (n = 728), only the significant signals are presented in Figure 2. Full results for each association are given in Supplement S3 and specific selections of variables for each long-term disease, used for logistic regression with AUC calculated on validation datasets are showed in Supplement S4.

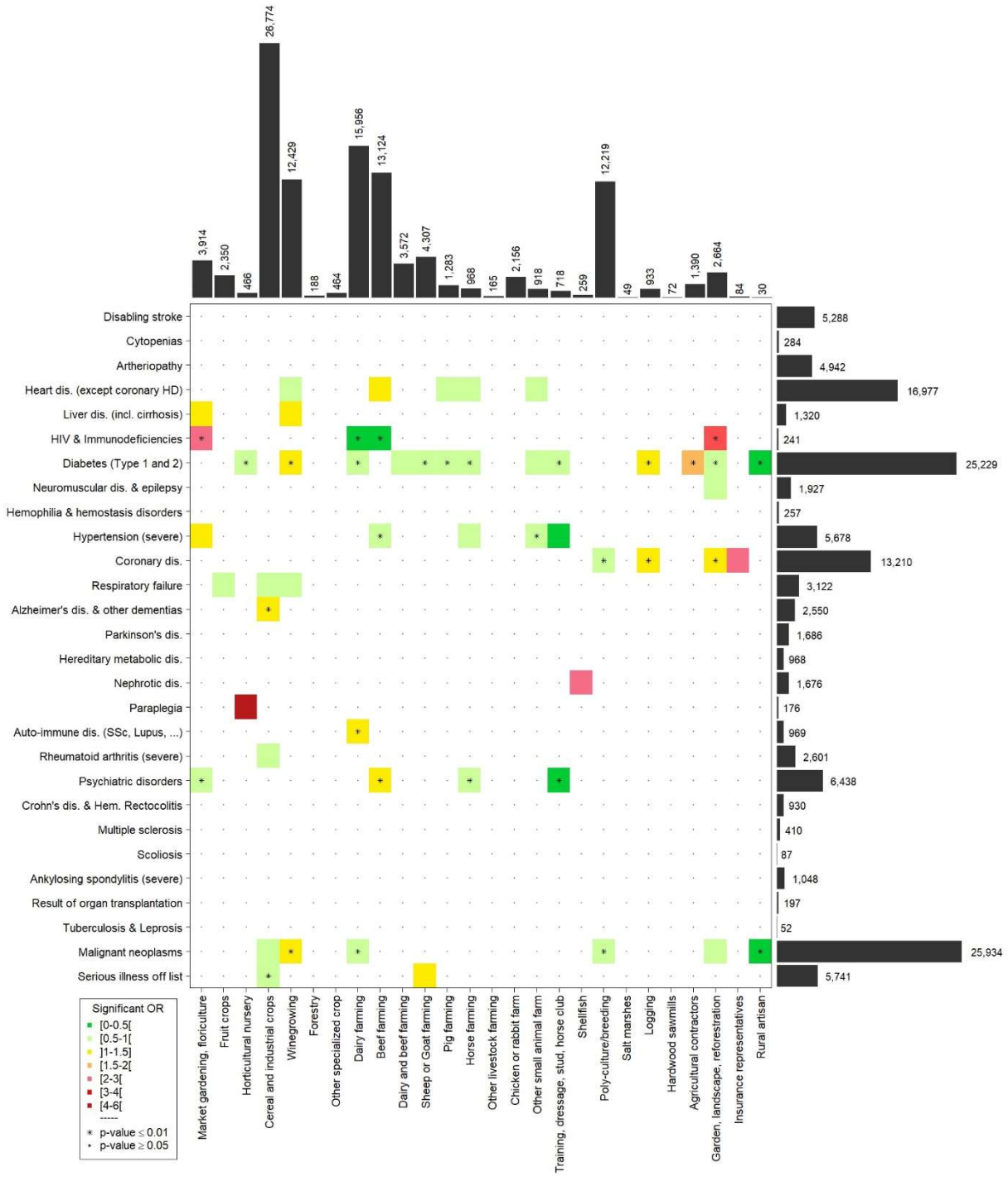


Figure 2: Heatmap showing odds ratios of significant associations between long-term diseases and occupational activities, obtained using logistic regression models, applied to the self-employed workers affiliated to the MSA (2006-2016)

After adjusting for confounding variables, twelve LTDs showed no statistically significant association with an OA. These LTDs included congenital diseases such as hemophilia and hemostasis disorders, acquired childhood diseases (scoliosis), rare infectious diseases (tuberculosis & leprosy), hereditary metabolic diseases as well as disabling stroke, cytopenias, artheriopathy, multiple sclerosis, ankylosing spondylitis, “organ transplantation sequellae”, Crohn’s disease and hemorrhagic colitis, and Parkinson's disease. Analyses revealed 54 statistically significant associations (corrected p-value <0.05) or “signals” between an OA and a LTD, including 35 signals with a lower risk of reporting and 19 with a higher risk of reporting specific LTDs.

The strongest signal we found was between paraplegia and work in horticultural nurseries, with a small number of cases (OR = 4.94 [2.03-12.04]; p = 0.01; n = 5). The next signals were for LTD involving immunodeficiencies, mainly HIV infection, with high risks of reporting this LTD for market gardeners and floriculturists (OR = 2.88 [1.93-4.31]; p = 2.92e-6; n = 27) and those working in gardening, landscaping and reforestation companies (OR = 3.06 [2.17-4.32]; p = 5.88e-9; n = 41). Another particularly interesting strong signal, with a higher risk of reporting nephrotic diseases, concerned shellfish workers (OR = 2.78 [1.48-5.21]; p = 0.04; n = 10).

Breeders (dairy and beef farming, horse farming, etc.) were more concerned by signals with a lower risk of reporting mainly diabetes and hypertension, compared to the rest of the self-employed population. For example, we found a lower risk of reporting diabetes among dairy farmers (OR = 0.84 [0.80-0.87]; p = 4.78e-17; n = 3459). In addition, four associations with lower risk of reporting specifics LTDs with very low OR were observed. For example, it was the case for two associations between HIV and immunodeficiencies and dairy farming (OR = 0.42 [0.26-0.68]; p = 0.002; n = 19) and beef farming (OR = 0.31 [0.17-0.59]; p = 0.002; n = 10). Furthermore, only three associations involved higher risks of reporting specific LTDs among breeders, namely: one for dairy farmers with autoimmune diseases (scleroderma and systemic sclerosis, lupus, etc.) (OR = 1.46 [1.25-1.71]; p = 4.58e-5; n = 211) and two for beef farmers with heart diseases (except coronary heart diseases) (OR = 1.07 [1.02-1.12]; p = 0.03; n = 2438) and psychiatric disorders (OR = 1.17 [1.09-1.25]; p = 1.88e-4; n = 993).

Conversely, specific OAs appeared with a higher risk of reporting LTDs. Winegrowers are particularly concerned by higher risks of reporting three LTDs: liver diseases (including cirrhosis) (OR = 1.29 [1.11-1.51]; p = 0.03; n = 192), diabetes (OR = 1.18 [1.13-1.23]; p = 8.22e-15; n = 3402) and malignant neoplasms (OR = 1.12 [1.07-1.16]; p = 1.02e-6; n = 3375). Similarly, among market gardeners and floriculturists, compared to the rest of the self-employed population, we found higher risks of reporting liver diseases (including cirrhosis) (OR = 1.42 [1.13-1.78]; p = 0.03; n = 82) and hypertension (OR = 1.19 [1.05-1.35]; p = 0.04; n = 262). There were also higher risks of reporting coronary diseases among individuals working in logging (OR = 1.32 [1.12-1.55]; p = 0.01; n = 159) and in gardening, landscaping and reforestation companies (OR = 1.20 [1.08-1.33]; p = 0.01; n = 434).

Concerning respiratory failure, only three associations appeared with lower risks of reporting this LTD. Three OAs are concerned: fruit crops (OR = 0.63 [0.47-0.84]; p = 0.02; n = 48), cereal and industrial crops (OR = 0.87 [0.79-0.95]; p = 0.02; n = 827), winegrowing (OR = 0.81 [0.71-0.92]; p = 0.02; n = 290).

Robustness analysis

Other tested models did not show significant improvement in robustness measures. AUC for each method were compared in Supplement S5. Otherwise, associations highlighted with these methods were very close to those found above. For example, we obtained a significant p-value and a similar OR for the association found above between LTD involving immunodeficiencies and market gardening and floriculture: logistic regression with sparse regression (OR = 2.02 [1.32-3.09]; p = 0.02), logistic regression with added confounders assessed by latent factor models (OR = 2.88 [1.93-4.31]; p = 2.92e-6) and logistic regression using Firth’s method (OR = 2.93 [1.98-4.34]; p = 4.08e-7).

Discussion

This first work on MSA's administrative and medico-administrative databases evaluated statistical associations between occupational activities and long-term diseases without any prior assumptions and with adjustment for available variables including previous illnesses. This was a first step in assessing whether the analysis of the MSA databases could contribute to a system for surveillance of the health of agricultural workers. The presented methodology has the following advantages: 1) a systematic approach relying on exhaustive and up-to-date data gathered for other purposes, 2) strong statistical power and 3) costless data acquisition.

Among the 54 highlighted signals, some capture well-known occupational health determinants, such as the association between malignant neoplasms and winegrowing which occurred in a population exposed to pesticides, notably arsenic compounds (authorized in France for use on vineyards until 2001), whose exposure is notably a risk factor for bladder, lung and skin cancers[28]. Concerning malignant neoplasms, we found a lower risk of reporting this LTD for dairy farmers, consistent with decreased risks of lung cancer found in literature among individuals who had cared for animals and undertaken milking[29]. Some associations, such as the higher risk of reporting diabetes (types 1 and 2) in winegrowers, may be in agreement with recent studies showing a relationship between diabetes and pesticides[30,31], although further investigations are needed due to the multifactorial nature of diabetes. Winegrowers were also at higher risk of reporting chronic liver diseases (including cirrhosis), which could be related to professional or behavioral factors (e.g. alcohol consumption). We also noticed a higher risk or reporting nephrotic diseases for shellfish workers which have an increased consumption of shellfish, a source of heavy metals, including cadmium, but also lead and mercury, all toxic to the kidneys[32,33].

Some signals are probably related to social health determinants, as commonly seen in occupational health epidemiology[34]. For example, we found a signal showing a higher risk of reporting long-term psychiatric disorders in beef farming, consistent with the elevated suicide rate observed for cow breeders by the French public health agency in 2010 and 2011[35]. The economic crisis faced by this population was identified as a significant mental health determinant. Otherwise, signals linking immunodeficiencies, mainly HIV, and market gardening and floriculture might be explained by the social insertion for these sectors in France, of people previously living in important precariousness and presenting health risks behaviors (e.g. addictions), who have set up as independent workers in market and organic gardening. This hypothesis is supported by the higher risk of reporting chronic liver diseases (including cirrhosis) observed in this same population, which raises the question of alcohol consumption levels as well as prevalence of chronic viral hepatitis in this population.

Analyses performed without any prior assumptions, have the capacity to generate hypotheses, revealing associations that would not necessarily have been explored otherwise. For example, we found an association between a LTD gathering autoimmune connective tissues diseases and vasculitis (scleroderma and systemic sclerosis, lupus, etc.) and dairy farming, which, to our knowledge, has never been described. Usually, publications regarding auto-immunity and cow's milk concern children's diet, and the onset of several autoimmune diseases, including type 1 Diabetes[36]. Our finding calls for further investigations regarding auto-immunity in this population. This should lead to a focused epidemiological study documenting precisely related illnesses, as well as occupational exposures, diet habits and past dwellings in this context.

Finally, at a first instance, the absence of any association between an OA and Parkinson's disease may seem surprising. There are three probable explanations for that. First of all, our reference population is made of all self-employed farmers. However, in the literature, the risk of contracting Parkinson's disease among farmers is high overall and homogeneous, as is the case in rural areas[4,37].

Secondly, since 2012, the French health insurance system has recognized the Parkinson's disease as an occupational disease if it is proven that farm workers have been exposed to pesticides. Consequently, such cases are recorded in the "occupational disease" database instead of the LTD database. Thus, some of the cases were not included in the analyses (n = 36 cases in our study period, after exclusion of duplicates). However, analyses merging cases recorded in these two databases, mutually exclusive, did not either showed significant association between an OA and this disease. Finally, we had no access to previous OAs for retirees, an important limitation for this disease whose prevalence is sharply increasing after 65 years old[38].

Limitations

Considering the complexity of the MSA databases, there were some limitations to our approach. First of all, retirees had to be excluded, due to the absence of data on past OAs, implying the non-inclusion of more than 60% of individuals registered with a declaration of LTD. There is currently no certainty that we will be able to recover retirees previous OAs with other databases.

Another limitation of our approach was the lack of accurate information regarding the type of exposure. More precisely, some codes used to describe OAs are broad, for example the OA named "cereals and industrial crops". For that reason, a complementary work was done by Achard *et al.* to precise OAs and retrospectively estimate the probability of pesticide use for each agricultural worker using their OAs, location and external data sources on agricultural practices[39].

As LTD recognition is not mandatory, there might be no signal for some chronic diseases if patients and their physicians did not request 100% reimbursement of healthcare costs. This bias will be investigated and may in part be resolved by integrating drug prescription data. For example, antidepressants could help identify clinically depressed individuals who did not seek LTD recognition.

Finally, another limitation was that in our models the duration of exposure was not considered. Work is ongoing to extract occupational history, which will allow OAs to be weighted according to exposure duration.

Perspectives

The procedure will be refine at the following four levels. The first step is to include employees in analyses; the main related challenge is that this population is highly unstable in terms of occupations (some have more than 100 contracts a year). Next, we have to improve information about occupational activity and probability of pesticides used; our procedure mentioned above will be integrated in further analyses[39]. Another point is to improve the disease proxy. Future work will aim to generate signals directly at the ICD-10 level concerning LTDs (an essential point for neoplasms for example), or groups of ICD-10 codes (for hemomalignancies for instance, as ICD-10 are not relevant in that case). A second way to improve the disease proxy will be to integrate information from and supplementary database recording consumption of medical care and medical goods. This might help to better identify diseases through algorithms of drugs consumption (such as asthma), but also to conduct syndromic surveillance (*e.g.* antibiotics consumptions, psychotropic consumption, etc). Finally, the last aspect is to see to what extend we can better take the time dimension and exposure duration into account with available data and especially through survival analyses.

Conclusion

The WHO Guidelines on Ethical Issues in Public Health Surveillance stated that “Surveillance, when conducted ethically, is the foundation for programs to promote human well-being at the population level” and establish it as one of its objectives[40]. This work was built on collaboration between health insurance, health safety agencies, researchers and occupational physicians. It showed the feasibility and relevance of our methodological approach, which is able, to highlight associations of potential concern regarding health surveillance of farmers, relying on health insurance data flow analysis. A multidisciplinary group of experts will assess signals highlighted through this process and will do a report to MSA scientific committee and ANSES. Relevant associations shall lead to further investigations such as focused epidemiological, toxicological or even sociological studies.

References

1. Lerro CC, Koutros S, Andreotti G, et al. Cancer incidence in the Agricultural Health Study after 20 years of follow-up. *Cancer Causes Control* 2019;**30**(4):311-322.

2. Lemarchand C, Tual S, Leveque-Morlais N, et al. Cancer incidence in the AGRICAN cohort study (2005-2011). *Cancer Epidemiol* 2017;**49**:175-185.

3. Mostafalou S, Abdollahi M. Pesticides: an update of human exposure and toxicity. *Arch Toxicol* 2017; **91**(2):549-599.

4. Kab S, Spinosi J, Chaperon L, et al. Agricultural activities and the incidence of Parkinson's disease in the general French population. *Eur J Epidemiol* 2017;**32**(3):203-216.

5. Parks CG, Meyer A, Beane Freeman LE, et al. Farming tasks and the development of rheumatoid arthritis in the agricultural health study. *Occup Environ Med* 2019;**76**(4):243-249.

6. Tual S, Silverman DT, Koutros S, et al. Use of Dieselized Farm Equipment and Incident Lung Cancer: Findings from the Agricultural Health Study Cohort. *Environ Health Perspect* 2016;**124**:611–618.

7. Szewczyk M, Pazdrowski J, Golusinski P, et al. Basal cell carcinoma in farmers: an occupation group at high risk. *Int Arch Occup Environ Health* 2016;**89**(3):497-501.

8. Khan Mi, Bath B, Boden C, et al. The association between awkward working posture and low back disorders in farmers: a systematic review. *J Agromedicine* 2019;**24**(1):74-89.

9. Fontana L, Lee Sj, Capitanelli I, et al. Chronic Obstructive Pulmonary Disease in Farmers: A Systematic Review. *J Occup Environ Med* 2017;**59**(8):775-788.

10. Brew B, Inder K, Allen J, et al. The health and wellbeing of Australian farmers: a longitudinal cohort study. *BMC Public Health* 2016;**16**(1):988.

11. Klingelschmidt J, Milner A, Khireddine-Medouni I, et al. Suicide among agricultural, forestry, and fishery workers: a systematic literature review and meta-analysis. *Scand J Work Environ Health* 2018;**44**(1):3-15.

12. Alavanja Mcr, Sandler Dp, Mcmaster Sb, et al. The Agricultural Health Study. *Environ Health Perspect* 1996;**104**(4):362-369.

13. Leveque-Morlais N, Tual S, Clin B, et al. The AGRIculture and CANcer (AGRICAN) cohort study: enrollment and causes of death for the 2005-2009 period. *Int Arch Occup Environ Health* 2015;**88**(1):61-73.

14. World Health Organization. The World health report 2000: health systems: improving performance. World Health Organization: Geneva, 2000, 215 p.

15. Berkson J. Application of the Logistic Function to Bio-Assay. *Journal of the American Statistical Association* 1944;**39**(227):357-365.

16. Tufféry S. Data mining et statistique décisionnelle : l'intelligence dans les bases de données. Editions Technip. Paris, 2005.

17. Cramér H. Mathematical methods of statistics. Princeton University Press. Princeton, 1999.

18. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J R Stat Soc Ser B Methodol 1995;**57**:289–300.

19. Tibshirani R. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)* 1996;**58**(1):267–288.

20. Caye K, Jumentier B, Lepeule J, et al. LFMM 2: Fast and Accurate Inference of Gene-Environment Associations in Genome-Wide Studies. *Molecular Biology and Evolution* 2019;**36**(4):852–860.

21. Firth D. Bias reduction of maximum likelihood estimates. *Biometrika* 1993;**80**(1):27–38.

22. R Core Team. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing: Vienna, Austria, 2017. https://www.R-project.org/.

23. Enea M. speedglm: Fitting Linear and Generalized Linear Models to Large Data Sets. R package version 0.3-2 2017. https://cran.r-project.org/package=speedglm

24. Szulc P. bigstep: Stepwise Selection for Large Data Sets. R package version 1.0.0. 2018. https://cran.r-project.org/package=bigstep

25. Sing T, Sander O, Beeernwinkel N et al. ROCR: visualizing classifier performance in R. *Bioinformatics* 2005;**21**:7881.

26. Kosmidis, I. brglm2: Bias reduction in generalized linear models. R package version 0.5.1 2019. https://cran.r-project.org/package=brglm2

27. Friedman J, Hastie T, Tibshirani R. Regularization Paths for Generalized Linear Models via Coordinate Descent. *Journal of Statistical Software* 2010;**33**(1);1-22.

28. International Agency for Research on Cancer. Arsenic and arsenic compounds. In: IARC monographs on the evaluation of carcinogenic risks to humans, volume 100 C, arsenic, metals, fibres, and dusts. IARC: Lyon, 2012,p 41–93.

29. Tual S, Lemarchand C, Boulanger M, et al. Exposure to Farm Animals and Risk of Lung Cancer in the AGRICAN Cohort. *Am J Epidemiol* 2017;**186**:463–472.

30. Evangelou E, Ntritsos G, Chondrogiorgi M, et al. Exposure to pesticides and diabetes: A systematic review and meta-analysis. *Environ Int* 2016;**91**:60–68.

31. Xiao X, Clark JM, Park Y. Potential contribution of insecticide exposure and development of obesity and type 2 diabetes. *Food Chem Toxicol* 2017;**105**:456-474.

32. Chiocchetti G, Jadan-Piedra C, Velez D, et al. Metal(loid) contamination in seafood products. *Crit Rev Food Sci Nutr* 2017;**57**(17):3715-3728.

33. Gueguen M, Amiard Jc, Arnich N, et al. Shellfish and residual chemical contaminants: hazards, monitoring, and health risk assessment along French coasts. *Rev Environ Contam Toxicol* 2011;**213**:55-111.

34. Jessop E. Mortality by occupation: the best basis for actionable results? *Lancet Public Health* 2017;**2**:e486–e487.

35. Khireddine-Medouni I, Breuillard É, Bossard C. Surveillance de la mortalité par suicide des agriculteurs exploitants. Situation 2010-2011 et évolution 2007-2011. Santé Publique France: Saint-Maurice, 2016, 32 p.

36. Garcia-Larsen V, Ierodiakonou D, Jarrold K, et al. Diet during pregnancy and infancy and risk of allergic or autoimmune disease: A systematic review and meta-analysis. *PLoS Med* 2018;**15**(2):e1002507.

37. Pouchieu C, Piel C, Carles C, Gruber A, Helmer C, Tual S et al. Pesticide use in agriculture and Parkinson's disease in the AGRICAN cohort study. Int J Epidemiol 2018; **47**: 299–310.

38. Moisan F, Wanneveich M, Kab S, et al. Fréquence de la maladie de parkinson en France en 2015 et évolution jusqu'en 2030. *Bull Epidémiol Hebd* 2018;(8-9):128-40.

39. Achard P, Maugard C, Cancé C, Ozenfant D, Maitre A, Bosson-Rieutort D, Bonneterre V. Medico-administrative data combined with agricultural practices data to retrospectively estimate pesticide use by agricultural workers. Journal Of Exposure Science And Environmental Epidemiology.

40. World Health Organization. WHO guidelines on ethical issues in public health surveillance. World Health Organization: Geneva, 2017. http://www.who.int/ethics/publications/public-healthsurveillance/en/ (accessed 2 Dec 2018).

Supplement 1: List of the 26 occupational activities from the MSA specific thesaurus dedicated to self-employed workers

Code	Label
1	Market gardening, floriculture
2	Fruit crops
3	Horticultural nursery
4	Cereal and industrial crop, "field crops"
5	Winegrowing
6	Forestry
7	Other specialized crop
8	Dairy farming
9	Beef farming
10	Mixed farming
11	Sheep or goat farming
12	Pig farming
13	Horse farming
14	Other livestock farming
15	Chicken or rabbit farm
16	Other small animal farm
17	Training, dressage, stud, horse club
18	Shellfish
19	Non-specialized crop, polyculture, poly-breeding
20	Salt marshes
21	Logging
22	Hardwood sawmills
23	Agricultural contractors
24	Garden, landscape and reforestation company
25	Agents of the local mutual insurance company
26	Rural artisan

Supplement 2: List of 32 illnesses or group of illnesses recognized by the French health insurance system as requiring prolonged (and/or expensive) treatment feature, known as long-term diseases (LTD)

Code	Precise label	Abreviated label
1	Disabling stroke	Disabling stroke
2	Renal insuffisiency and other chronic cytopenias	Cytopenias
3	Chronic artheriopathy with ischemic manifestations	Arteriopathy
4	Complicated bilharzia	Complicated bilharzia
5	Severe heart failure, severe rythm disorders, severe valvular heart disease, severe congenital heart defects	Heart diseases (except coronary heart diseases)
6	Active chronic diseases of the liver and cirrhosis	Liver diseases (including cirrhosis)
7	Severe primary immunodeficiency requiring prolonged treatment and HIV infection	HIV & Immunodeficiencies
8	Type 1 diabetes (insulin-dependant) and type 2 diabetes (sugar diabetes)	Diabetes (Type 1 and 2)
9	Severe form of neurological and muscular disorders (including myopathy), severe epilepsy	Neuromuscular diseases and epilepsy
10	Hemoglobinopathies, severe constitutional nd acquired chronic hemolysis	Hemoglobinopathies
11	Hemophilia and severe constitutional afflictions of hemostasis	Hemophilia & hemostasis disorders
12	Severe hypertension (high blood pressure)	Hypertension (severe)
13	Coronary disease	Coronary diseases
14	Severe chronic respiratory failure	Respiratory failure
15	Alzheimer's disease and other dementias	Alzheimer’s disease & other dementias
16	Parkinson's disease	Parkinson’s disease
17	Hereditary metabolic diseases requiring prolonged and specific treatment	Hereditary metabolic diseases
18	Cystic fibrosis	Cystic fibrosis
19	Severe chronic nephropathy and primary nephrotic syndrome	Nephrotic diseases
20	Paraplegia	Paraplegia
21	Periarteritis nodosa, systemic lupus erythematosus, active systemic sclerosis	Auto-immune disease (Scleroderma and Systemic Sclerosis, Lupus, …)
22	Severely active rheumatoid arthritis	Rheumatoid arthritis (severe)
23	Long-term psychiatric disorders	Psychiatric disorders
24	Progressive hemorrhagic rectocolitis and Crohn's disease	Crohn disease & Hemorrhagic Colitis
25	Multiple sclerosis	Multiple sclerosis
26	Progressive structural scoliosis (angle equal or greather than 25 degrees) until spinal maturation	Scoliosis
27	Severe ankylosing spondylitis	Ankylosing spondylitis (severe)
28	Result of organ transplantation	Result of organ transplantation
29	Active tuberculosis, leprosy	Tuberculosis & Leprosis
30	Malignant tumor, malignant disease of lymphatic or hematopoietic tissue	Malignant neoplasms
31	Serious illness off list, requiring continuous care for a predictable period of more than 6 months	Serious illness off list
32	Polypathology: several conditions leading to a disabling condition, requiring continuous care for a predictable duration of more than 6 months	Poly-diseases

Supplement 3: Results for each association between a long-term disease and an occupational activity (728 associations), obtained via logistic regression applied on self-employed workers affiliated to the MSA between 2006 and 2016

	Disabling stroke (n = 5288)	Cytopenias (n = 284)	Arteriopathy (n = 4942)	Heart diseases (except coronary heart diseases) (n = 16977)
Market gardening, floriculture (n = 3914)	OR* = 0.89 [0.77-1.04] ; p** = 0.38 ; n = 183	OR = 0.72 [0.36-1.45] ; p = 0.99 ; n = 8	OR = 1.02 [0.89-1.18] ; p = 0.93 ; n = 219	OR = 0.94 [0.87-1.03] ; p = 0.29 ; n = 614
Fruit crops (n = 2350)	OR = 1.01 [0.85-1.21] ; p = 0.90 ; n = 129	OR = 0.70 [0.29-1.69] ; p = 0.99 ; n = 5	OR = 1.01 [0.85-1.21] ; p = 0.95 ; n = 131	OR = 0.89 [0.80-0.99] ; p = 0.09 ; n = 383
Horticultural nursery (n = 466)	OR = 0.64 [0.39-1.05] ; p = 0.38 ; n = 16	-	OR = 0.74 [0.47-1.15] ; p = 0.59 ; n = 20	OR = 0.78 [0.59-1.02] ; p = 0.19 ; n = 56
Cereal and industrial crop, "field crops" (n = 26774)	OR = 0.98 [0.92-1.05] ; p = 0.68 ; n = 1612	OR = 0.92 [0.70-1.21] ; p = 0.99 ; n = 86	OR = 1 [0.93-1.07] ; p = 0.99 ; n = 1486	OR = 0.97 [0.93-1.01] ; p = 0.19 ; n = 5670
Winegrowing (n = 12429)	OR = 0.93 [0.85-1.01] ; p = 0.38 ; n = 613	OR = 0.99 [0.69-1.40] ; p = 0.99 ; n = 36	OR = 1.09 [0.99-1.18] ; p = 0.55 ; n = 687	**OR = 0.93 [0.89-0.98] ; p = 0.03 ; n = 1988**
Forestry (n = 188)	OR = 1.34 [0.77-2.34] ; p = 0.55 ; n = 13	-	OR = 1.02 [0.56-1.86] ; p = 0.99 ; n = 11	OR = 0.93 [0.65-1.35] ; p = 0.73 ; n = 31
Other specialized crop (n = 464)	OR = 1.11 [0.76-1.62] ; p = 0.68 ; n = 27	-	OR = 1.31 [0.92-1.87] ; p = 0.55 ; n = 32	OR = 0.80 [0.62-1.04] ; p = 0.19 ; n = 62
Dairy farming (n = 15956)	OR = 1.03 [0.96-1.11] ; p = 0.62 ; n = 845	OR = 0.98 [0.70-1.37] ; p = 0.99 ; n = 41	OR = 0.99 [0.91-1.07] ; p = 0.93 ; n = 715	OR = 1.04 [0.99-1.08] ; p = 0.23 ; n = 2548
Beef farming (n = 13124)	OR = 1.02 [0.94-1.10] ; p = 0.68 ; n = 703	OR = 0.99 [0.70-1.41] ; p = 0.99 ; n = 36	OR = 0.94 [0.86-1.02] ; p = 0.55 ; n = 583	**OR = 1.07 [1.02-1.12] ; p = 0.03 ; n = 2438**
Mixed farming (n = 3572)	OR = 0.88 [0.75-1.04] ; p = 0.38 ; n = 159	OR = 1.14 [0.61-2.15] ; p = 0.99 ; n = 10	OR = 0.91 [0.77-1.08] ; p = 0.62 ; n = 144	OR = 1.08 [0.99-1.18] ; p = 0.19 ; n = 580
Sheep or goat farming (n = 4307)	OR = 1.04 [0.91-1.20] ; p = 0.68 ; n = 225	OR = 1.51 [0.92-2.47] ; p = 0.81 ; n = 17	OR = 1.09 [0.94-1.25] ; p = 0.62 ; n = 213	OR = 1.03 [0.95-1.11] ; p = 0.62 ; n = 703
Pig farming (n = 1283)	OR = 0.92 [0.71-1.19] ; p = 0.68 ; n = 60	OR = 1.24 [0.46-3.34] ; p = 0.99 ; n = 4	OR = 0.98 [0.76-1.27] ; p = 0.95 ; n = 59	**OR = 0.77 [0.65-0.92] ; p = 0.03 ; n = 136**
Horse farming (n = 968)	OR = 0.86 [0.65-1.13] ; p = 0.55 ; n = 51	-	OR = 1.08 [0.83-1.39] ; p = 0.93 ; n = 62	**OR = 0.75 [0.63-0.89] ; p = 0.02 ; n = 146**
Other livestock farming (n = 165)	OR = 0.81 [0.38-1.70] ; p = 0.68 ; n = 7	-	OR = 0.89 [0.44-1.80] ; p = 0.93 ; n = 8	OR = 0.78 [0.50-1.23] ; p = 0.39 ; n = 20
Chicken or rabbit farm (n = 2156)	OR = 0.90 [0.73-1.10] ; p = 0.55 ; n = 92	OR = 1.32 [0.62-2.81] ; p = 0.99 ; n = 7	OR = 0.88 [0.71-1.09] ; p = 0.62 ; n = 88	OR = 0.95 [0.83-1.07] ; p = 0.51 ; n = 256
Other small animal farm (n = 918)	OR = 0.93 [0.69-1.26] ; p = 0.68 ; n = 44	OR = 1.06 [0.34-3.30] ; p = 0.99 ; n = 3	OR = 0.79 [0.58-1.08] ; p = 0.55 ; n = 43	**OR = 0.76 [0.63-0.93] ; p = 0.03 ; n = 112**
Training, dressage, stud, horse club (n = 718)	OR = 1.27 [0.95-1.70] ; p = 0.38 ; n = 46	-	OR = 0.93 [0.66-1.31] ; p = 0.93 ; n = 34	OR = 0.77 [0.61-0.97] ; p = 0.09 ; n = 73
Shellfish (n = 259)	OR = 1.23 [0.73-2.10] ; p = 0.67 ; n = 14	-	OR = 1.23 [0.75-2.00] ; p = 0.82 ; n = 17	OR = 0.94 [0.66-1.34] ; p = 0.73 ; n = 33
Non-specialized crop, polyculture, poly-breeding (n = 12219)	OR = 0.92 [0.85-1.00] ; p = 0.38 ; n = 609	OR = 1.05 [0.74-1.48] ; p = 0.99 ; n = 36	OR = 0.89 [0.81-0.98] ; p = 0.40 ; n = 523	OR = 0.96 [0.91-1.01] ; p = 0.20 ; n = 1976
Salt marshes (n = 49)	OR = 1.47 [0.47-4.59] ; p = 0.68 ; n = 3	-	OR = 1.25 [0.40-3.95] ; p = 0.93 ; n = 3	-
Logging (n = 933)	OR = 1.24 [0.93-1.65] ; p = 0.38 ; n = 48	-	OR = 1.22 [0.94-1.59] ; p = 0.55 ; n = 61	OR = 1.13 [0.94-1.35] ; p = 0.31 ; n = 127
Hardwood sawmills (n = 72)	OR = 1.71 [0.76-3.85] ; p = 0.46 ; n = 6	-	OR = 2.06 [1.05-4.04] ; p = 0.48 ; n = 9	OR = 0.82 [0.42-1.60] ; p = 0.62 ; n = 9
Agricultural contractors (n = 1390)	OR = 1.25 [0.99-1.58] ; p = 0.38 ; n = 74	-	OR = 0.87 [0.67-1.12] ; p = 0.62 ; n = 62	OR = 1.06 [0.91-1.23] ; p = 0.58 ; n = 189
Garden, landscape and reforestation company (n = 2664)	OR = 1.25 [1.05-1.48] ; p = 0.27 ; n = 146	OR = 0.57 [0.21-1.55] ; p = 0.99 ; n = 4	OR = 0.96 [0.80-1.16] ; p = 0.93 ; n = 140	OR = 0.93 [0.82-1.05] ; p = 0.37 ; n = 294
Agents of the local mutual insurance company (n = 84)	-	-	OR = 0.80 [0.25-2.49] ; p = 0.93 ; n = 3	OR = 0.86 [0.42-1.73] ; p = 0.72 ; n = 8
Rural artisan (n = 30)	-	-	-	OR = 0.41 [0.19-0.87] ; p = 0.09 ; n = 7

Odds ratio with 95% confidence interval; **Corrected p-values (Benjamini-Hochberg procedure); Empty boxes correspond to a number of individuals that are insufficient to allow the search for associations (n ≤ 3); Only results for studied LTDs were showed. No association tests were performed for four LTDs : "Complicated bilharzia" (n = 1), "Hemoglobinopathies" (n = 18), "Cystic fibrosis" (n = 9) and "Poly-diseases" (n = 1).

	Liver diseases (including cirrhosis) (n = 1320)	HIV & Immunodeficiencies (n = 241)	Diabetes (Type 1 and 2) (n = 25229)	Neuromuscular diseases and epilepsy (n = 1927)
Market gardening, floriculture (n = 3914)	**OR = 1.42 [1.13-1.78] ; p = 0.03 ; n = 82**	**OR = 2.88 [1.93-4.31] ; p = 2.92e-6 ; n = 27**	OR = 0.95 [0.89-1.02] ; p = 0.21 ; n = 986	OR = 0.86 [0.67-1.09] ; p = 0.67 ; n = 66
Fruit crops (n = 2350)	OR = 0.77 [0.51-1.15] ; p = 0.64 ; n = 24	OR = 1.39 [0.65-2.94] ; p = 0.74 ; n = 7	OR = 1.00 [0.92-1.09] ; p = 0.99 ; n = 596	OR = 1.06 [0.79-1.43] ; p = 0.96 ; n = 45
Horticultural nursery (n = 466)	OR = 1.00 [0.50-2.00] ; p = 0.99 ; n = 8	-	**OR = 0.68 [0.55-0.84] ; p = 9.21e-4 ; n = 94**	OR = 1.56 [0.95-2.55] ; p = 0.34 ; n = 16
Cereal and industrial crop, "field crops" (n = 26774)	OR = 0.88 [0.76-1.01] ; p = 0.32 ; n = 264	OR = 0.76 [0.53-1.10] ; p = 0.34 ; n = 37	OR = 0.97 [0.94-1.00] ; p = 0.14 ; n = 6504	OR = 1.00 [0.89-1.11] ; p = 0.96 ; n = 447
Winegrowing (n = 12429)	**OR = 1.29 [1.11-1.51] ; p = 0.03 ; n = 192**	OR = 0.97 [0.64-1.47] ; p = 0.99 ; n = 25	**OR = 1.18 [1.13-1.23] ; p = 8.22e-15 ; n = 3402**	OR = 1.01 [0.88-1.17] ; p = 0.96 ; n = 218
Forestry (n = 188)	-	-	OR = 0.85 [0.63-1.16] ; p = 0.43 ; n = 44	OR = 2.06 [0.98-4.34] ; p = 0.30 ; n = 7
Other specialized crop (n = 464)	OR = 1.09 [0.57-2.11] ; p = 0.92 ; n = 9	OR = 2.49 [0.93-6.71] ; p = 0.23 ; n = 4	OR = 0.82 [0.67-0.99] ; p = 0.07 ; n = 112	OR = 0.90 [0.47-1.73] ; p = 0.96 ; n = 9
Dairy farming (n = 15956)	OR = 0.89 [0.76-1.04] ; p = 0.56 ; n = 185	**OR = 0.42 [0.26-0.68] ; p = 0.002 ; n = 19**	**OR = 0.84 [0.80-0.87] ; p = 4.78e-17 ; n = 3459**	OR = 1.04 [0.93-1.17] ; p = 0.96 ; n = 352
Beef farming (n = 13124)	OR = 0.98 [0.83-1.16] ; p = 0.92 ; n = 167	**OR = 0.31 [0.17-0.59] ; p = 0.002 ; n = 10**	OR = 1.00 [0.96-1.04] ; p = 0.99 ; n = 3362	OR = 1.03 [0.90-1.18] ; p = 0.96 ; n = 253
Mixed farming (n = 3572)	OR = 1.12 [0.84-1.48] ; p = 0.92 ; n = 52	OR = 0.94 [0.46-1.89] ; p = 0.99 ; n = 8	**OR = 0.92 [0.86-0.99] ; p = 0.04 ; n = 856**	OR = 1.07 [0.85-1.34] ; p = 0.96 ; n = 78
Sheep or goat farming (n = 4307)	OR = 0.94 [0.73-1.22] ; p = 0.92 ; n = 60	OR = 1.00 [0.56-1.80] ; p = 0.99 ; n = 12	**OR = 0.89 [0.83-0.95] ; p = 0.001 ; n = 1023**	OR = 1.27 [1.05-1.54] ; p = 0.11 ; n = 110
Pig farming (n = 1283)	OR = 0.43 [0.21-0.86] ; p = 0.15 ; n = 8	OR = 1.72 [0.76-3.87] ; p = 0.39 ; n = 6	**OR = 0.81 [0.71-0.93] ; p = 0.004 ; n = 246**	OR = 1.09 [0.77-1.54] ; p = 0.96 ; n = 33
Horse farming (n = 968)	OR = 1.07 [0.68-1.70] ; p = 0.92 ; n = 19	OR = 1.95 [0.92-4.16] ; p = 0.24 ; n = 7	**OR = 0.73 [0.64-0.84] ; p = 6.35e-5 ; n = 206**	OR = 0.47 [0.25-0.88] ; p = 0.11 ; n = 10
Other livestock farming (n = 165)	OR = 1.26 [0.47-3.37] ; p = 0.92 ; n = 4	-	OR = 0.88 [0.64-1.22] ; p = 0.59 ; n = 40	OR = 1.38 [0.57-3.33] ; p = 0.96 ; n = 5
Chicken or rabbit farm (n = 2156)	OR = 0.85 [0.59-1.23] ; p = 0.92 ; n = 29	OR = 1.91 [1.07-3.42] ; p = 0.13 ; n = 12	OR = 1.06 [0.97-1.15] ; p = 0.36 ; n = 540	OR = 0.88 [0.65-1.19] ; p = 0.96 ; n = 42
Other small animal farm (n = 918)	OR = 0.96 [0.60-1.53] ; p = 0.92 ; n = 18	OR = 1.30 [0.53-3.17] ; p = 0.98 ; n = 5	**OR = 0.82 [0.72-0.95] ; p = 0.01 ; n = 220**	OR = 0.73 [0.44-1.22] ; p = 0.67 ; n = 15
Training, dressage, stud, horse club (n = 718)	OR = 0.77 [0.43-1.36] ; p = 0.92 ; n = 12	OR = 2.01 [1.03-3.92] ; p = 0.15 ; n = 9	**OR = 0.57 [0.47-0.69] ; p = 3.70e-8 ; n = 115**	OR = 0.98 [0.63-1.50] ; p = 0.96 ; n = 21
Shellfish (n = 259)	OR = 1.14 [0.51-2.55] ; p = 0.92 ; n = 6	-	OR = 1.05 [0.82-1.35] ; p = 0.80 ; n = 69	OR = 0.76 [0.28-2.03] ; p = 0.96 ; n = 4
Non-specialized crop, polyculture, poly-breeding (n = 12219)	OR = 0.82 [0.69-0.98] ; p = 0.18 ; n = 138	OR = 0.57 [0.35-0.92] ; p = 0.11 ; n = 18	OR = 1.00 [0.97-1.04] ; p = 0.94 ; n = 3317	OR = 1.02 [0.89-1.16] ; p = 0.96 ; n = 251
Salt marshes (n = 49)	OR = 3.06 [0.98-9.60] ; p = 0.28 ; n = 3	-	OR = 0.78 [0.40-1.53] ; p = 0.59 ; n = 9	-
Logging (n = 933)	OR = 1.11 [0.72-1.72] ; p = 0.92 ; n = 21	-	**OR = 1.38 [1.22-1.55] ; p = 6.45e-7 ; n = 315**	OR = 0.41 [0.20-0.82] ; p = 0.11 ; n = 8
Hardwood sawmills (n = 72)	-	-	OR = 1.01 [0.66-1.56] ; p = 0.99 ; n = 22	-
Agricultural contractors (n = 1390)	OR = 0.99 [0.65-1.51] ; p = 0.99 ; n = 23	OR = 1.14 [0.47-2.77] ; p = 0.99 ; n = 5	**OR = 1.53 [1.39-1.69] ; p = 3.34e-16 ; n = 463**	OR = 0.76 [0.50-1.18] ; p = 0.67 ; n = 21
Garden, landscape and reforestation company (n = 2664)	OR = 1.04 [0.79-1.37] ; p = 0.92 ; n = 64	**OR = 3.06 [2.17-4.32] ; p = 5.88e-9 ; n = 41**	**OR = 0.84 [0.77-0.92] ; p = 5.20e-4 ; n = 609**	**OR = 0.61 [0.45-0.82] ; p = 0.03 ; n = 46**
Agents of the local mutual insurance company (n = 84)	-	-	OR = 0.81 [0.49-1.34] ; p = 0.56 ; n = 16	-
Rural artisan (n = 30)	-	-	**OR = 0.21 [0.11-0.43] ; p = 5.93e-5 ; n = 8**	-

Odds ratio with 95% confidence interval; **Corrected p-values (Benjamini-Hochberg procedure); Empty boxes correspond to a number of individuals that are insufficient to allow the search for associations (n ≤ 3); Only results for studied LTDs were showed. No association tests were performed for four LTDs : "Complicated bilharzia" (n = 1), "Hemoglobinopathies" (n = 18), "Cystic fibrosis" (n = 9) and "Poly-diseases" (n = 1).

	Hemophilia & hemostasis disorders (n = 257)	Hypertension (severe) (n = 5678)	Coronary diseases (n = 13210)	Respiratory failure (n = 3122)
Market gardening, floriculture (n = 3914)	OR = 1.25 [0.71-2.18] ; p = 0.99 ; n = 13	**OR = 1.19 [1.05-1.35] ; p = 0.04 ; n = 262**	OR = 1.06 [0.97-1.15] ; p = 0.40 ; n = 551	OR = 0.91 [0.75-1.10] ; p = 0.69 ; n = 112
Fruit crops (n = 2350)	-	OR = 1.01 [0.85-1.19] ; p = 0.94 ; n = 156	OR = 0.99 [0.88-1.11] ; p = 0.90 ; n = 317	**OR = 0.63 [0.47-0.84] ; p = 0.02 ; n = 48**
Horticultural nursery (n = 466)	-	OR = 1.19 [0.81-1.76] ; p = 0.53 ; n = 27	OR = 1.26 [1.01-1.56] ; p = 0.13 ; n = 86	OR = 1.06 [0.63-1.81] ; p = 0.89 ; n = 14
Cereal and industrial crop, "field crops" (n = 26774)	OR = 1.35 [1.02-1.79] ; p = 0.69 ; n = 68	OR = 0.97 [0.91-1.03] ; p = 0.49 ; n = 1815	OR = 0.95 [0.91-0.99] ; p = 0.10 ; n = 3698	**OR = 0.87 [0.79-0.95] ; p = 0.02 ; n = 827**
Winegrowing (n = 12429)	OR = 0.87 [0.58-1.32] ; p = 0.99 ; n = 25	OR = 0.98 [0.90-1.06] ; p = 0.72 ; n = 730	OR = 0.99 [0.94-1.05] ; p = 0.82 ; n = 1580	**OR = 0.81 [0.71-0.92] ; p = 0.02 ; n = 290**
Forestry (n = 188)	-	OR = 1.19 [0.68-2.09] ; p = 0.68 ; n = 13	OR = 1.06 [0.73-1.56] ; p = 0.82 ; n = 28	OR = 0.70 [0.26-1.87] ; p = 0.79 ; n = 4
Other specialized crop (n = 464)	-	OR = 0.68 [0.43-1.08] ; p = 0.28 ; n = 19	OR = 0.89 [0.68-1.16] ; p = 0.52 ; n = 56	OR = 0.83 [0.48-1.44] ; p = 0.79 ; n = 13
Dairy farming (n = 15956)	OR = 0.94 [0.67-1.30] ; p = 0.99 ; n = 43	OR = 0.96 [0.88-1.05] ; p = 0.53 ; n = 747	OR = 0.95 [0.90-1.00] ; p = 0.17 ; n = 1968	OR = 1.12 [1.01-1.24] ; p = 0.11 ; n = 564
Beef farming (n = 13124)	OR = 0.85 [0.57-1.27] ; p = 0.99 ; n = 27	**OR = 0.85 [0.78-0.92] ; p = 0.002 ; n = 679**	OR = 0.95 [0.90-1.00] ; p = 0.17 ; n = 1610	OR = 1.03 [0.93-1.14] ; p = 0.80 ; n = 445
Mixed farming (n = 3572)	OR = 0.97 [0.50-1.88] ; p = 0.99 ; n = 9	OR = 0.82 [0.70-0.97] ; p = 0.10 ; n = 152	OR = 0.92 [0.84-1.02] ; p = 0.24 ; n = 431	OR = 1.18 [0.98-1.41] ; p = 0.22 ; n = 132
Sheep or goat farming (n = 4307)	OR = 0.92 [0.50-1.69] ; p = 0.99 ; n = 11	OR = 1.09 [0.96-1.24] ; p = 0.34 ; n = 267	OR = 0.90 [0.82-0.99] ; p = 0.10 ; n = 489	OR = 1.22 [1.05-1.43] ; p = 0.07 ; n = 176
Pig farming (n = 1283)	OR = 1.5 [0.67-3.37] ; p = 0.99 ; n = 6	OR = 1.02 [0.75-1.38] ; p = 0.94 ; n = 45	OR = 0.92 [0.78-1.07] ; p = 0.42 ; n = 168	OR = 1.08 [0.79-1.47] ; p = 0.81 ; n = 44
Horse farming (n = 968)	OR = 0.96 [0.31-3.00] ; p = 0.99 ; n = 3	**OR = 0.61 [0.44-0.84] ; p = 0.02 ; n = 40**	OR = 0.96 [0.81-1.14] ; p = 0.76 ; n = 134	OR = 0.95 [0.67-1.35] ; p = 0.89 ; n = 33
Other livestock farming (n = 165)	-	OR = 0.52 [0.21-1.26] ; p = 0.32 ; n = 5	OR = 1.23 [0.83-1.83] ; p = 0.43 ; n = 26	OR = 2.1 [1.15-3.83] ; p = 0.08 ; n = 11
Chicken or rabbit farm (n = 2156)	OR = 0.43 [0.14-1.34] ; p = 0.75 ; n = 3	OR = 1.23 [1.01-1.51] ; p = 0.13 ; n = 101	OR = 1.05 [0.92-1.18] ; p = 0.60 ; n = 270	OR = 1 [0.77-1.30] ; p = 0.99 ; n = 60
Other small animal farm (n = 918)	-	**OR = 0.51 [0.34-0.74] ; p = 0.01 ; n = 27**	OR = 1.13 [0.95-1.34] ; p = 0.33 ; n = 139	OR = 0.94 [0.65-1.35] ; p = 0.89 ; n = 29
Training, dressage, stud, horse club (n = 718)	OR = 1.07 [0.40-2.87] ; p = 0.99 ; n = 4	**OR = 0.44 [0.26-0.75] ; p = 0.02 ; n = 14**	OR = 1.08 [0.88-1.32] ; p = 0.60 ; n = 100	OR = 1.05 [0.68-1.62] ; p = 0.89 ; n = 21
Shellfish (n = 259)	-	OR = 1.08 [0.60-1.93] ; p = 0.86 ; n = 12	OR = 1.30 [0.96-1.77] ; p = 0.20 ; n = 44	OR = 0.67 [0.28-1.62] ; p = 0.70 ; n = 5
Non-specialized crop, polyculture, poly-breeding (n = 12219)	OR = 0.99 [0.68-1.43] ; p = 0.99 ; n = 32	OR = 0.91 [0.84-0.99] ; p = 0.10 ; n = 654	**OR = 0.89 [0.85-0.95] ; p = 0.002 ; n = 1511**	OR = 0.94 [0.84-1.05] ; p = 0.54 ; n = 377
Salt marshes (n = 49)	-	-	OR = 1.5 [0.74-3.07] ; p = 0.42 ; n = 8	OR = 3.23 [1.02-10.17] ; p = 0.15 ; n = 3
Logging (n = 933)	OR = 1.01 [0.32-3.17] ; p = 0.99 ; n = 3	OR = 0.93 [0.68-1.27] ; p = 0.72 ; n = 41	**OR = 1.32 [1.12-1.55] ; p = 0.01 ; n = 159**	OR = 0.89 [0.57-1.37] ; p = 0.80 ; n = 21
Hardwood sawmills (n = 72)	-	-	OR = 0.96 [0.52-1.75] ; p = 0.90 ; n = 11	-
Agricultural contractors (n = 1390)	OR = 1.99 [0.99-4.03] ; p = 0.69 ; n = 8	OR = 1.30 [1.03-1.64] ; p = 0.10 ; n = 78	OR = 1.14 [0.99-1.32] ; p = 0.17 ; n = 210	OR = 1.21 [0.89-1.66] ; p = 0.53 ; n = 41
Garden, landscape and reforestation company (n = 2664)	OR = 0.75 [0.40-1.42] ; p = 0.99 ; n = 10	OR = 0.91 [0.73-1.13] ; p = 0.53 ; n = 95	**OR = 1.20 [1.08-1.33] ; p = 0.01 ; n = 434**	OR = 1.27 [1.01-1.60] ; p = 0.14 ; n = 82
Agents of the local mutual insurance company (n = 84)	-	-	**OR = 2.08 [1.24-3.51] ; p = 0.03 ; n = 15**	-
Rural artisan (n = 30)	-	OR = 0.47 [0.15-1.47] ; p = 0.34 ; n = 3	-	-

Odds ratio with 95% confidence interval; **Corrected p-values (Benjamini-Hochberg procedure); Empty boxes correspond to a number of individuals that are insufficient to allow the search for associations (n ≤ 3); Only results for studied LTDs were showed. No association tests were performed for four LTDs : "Complicated bilharzia" (n = 1), "Hemoglobinopathies" (n = 18), "Cystic fibrosis" (n = 9) and "Poly-diseases" (n = 1).

	Alzheimer's disease & other dementias (n = 2550)	Parkinson's disease (n = 1686)	Hereditary metabolic diseases (n = 968)	Nephrotic diseases (n = 1676)
Market gardening, floriculture (n = 3914)	OR = 0.70 [0.54-0.92] ; p = 0.09 ; n = 61	OR = 0.90 [0.70-1.17] ; p = 0.89 ; n = 59	OR = 1.17 [0.86-1.58] ; p = 0.97 ; n = 45	OR = 1.05 [0.82-1.34] ; p = 0.97 ; n = 66
Fruit crops (n = 2350)	OR = 1.06 [0.83-1.36] ; p = 0.86 ; n = 70	OR = 1.27 [0.96-1.67] ; p = 0.37 ; n = 53	OR = 1.24 [0.82-1.88] ; p = 0.97 ; n = 23	OR = 0.58 [0.39-0.87] ; p = 0.11 ; n = 24
Horticultural nursery (n = 466)	OR = 0.94 [0.42-2.12] ; p = 0.93 ; n = 6	-	OR = 0.94 [0.39-2.26] ; p = 0.97 ; n = 5	OR = 1.05 [0.52-2.11] ; p = 0.97 ; n = 8
Cereal and industrial crop, "field crops" (n = 26774)	**OR = 1.24 [1.13-1.35] ; p = 9.36e-5 ; n = 1394**	OR = 0.99 [0.89-1.11] ; p = 0.99 ; n = 551	OR = 0.77 [0.65-0.91] ; p = 0.07 ; n = 176	OR = 0.94 [0.84-1.06] ; p = 0.81 ; n = 489
Winegrowing (n = 12429)	OR = 0.95 [0.84-1.08] ; p = 0.86 ; n = 373	OR = 0.89 [0.76-1.03] ; p = 0.37 ; n = 194	OR = 1.22 [0.99-1.5] ; p = 0.73 ; n = 114	OR = 1.16 [1.01-1.34] ; p = 0.32 ; n = 234
Forestry (n = 188)	OR = 0.90 [0.33-2.43] ; p = 0.93 ; n = 4	OR = 1.96 [0.88-4.40] ; p = 0.37 ; n = 6	-	OR = 1.59 [0.65-3.86] ; p = 0.81 ; n = 5
Other specialized crop (n = 464)	OR = 0.74 [0.33-1.67] ; p = 0.86 ; n = 6	OR = 0.41 [0.13-1.27] ; p = 0.37 ; n = 3	-	OR = 1.03 [0.51-2.08] ; p = 0.97 ; n = 8
Dairy farming (n = 15956)	OR = 0.94 [0.82-1.08] ; p = 0.81 ; n = 269	OR = 0.97 [0.85-1.12] ; p = 0.99 ; n = 247	OR = 1.12 [0.95-1.32] ; p = 0.97 ; n = 216	OR = 0.93 [0.81-1.07] ; p = 0.81 ; n = 243
Beef farming (n = 13124)	OR = 0.93 [0.82-1.06] ; p = 0.65 ; n = 298	OR = 0.96 [0.83-1.11] ; p = 0.99 ; n = 210	OR = 1.02 [0.84-1.25] ; p = 0.97 ; n = 115	OR = 0.91 [0.79-1.06] ; p = 0.81 ; n = 204
Mixed farming (n = 3572)	OR = 0.92 [0.67-1.27] ; p = 0.86 ; n = 40	OR = 1.15 [0.90-1.48] ; p = 0.58 ; n = 64	OR = 0.88 [0.62-1.25] ; p = 0.97 ; n = 33	OR = 1.04 [0.79-1.35] ; p = 0.97 ; n = 58
Sheep or goat farming (n = 4307)	OR = 0.70 [0.54-0.92] ; p = 0.09 ; n = 56	OR = 1.08 [0.85-1.37] ; p = 0.99 ; n = 69	OR = 0.86 [0.61-1.22] ; p = 0.97 ; n = 34	OR = 1 [0.79-1.28] ; p = 0.97 ; n = 69
Pig farming (n = 1283)	OR = 0.62 [0.29-1.31] ; p = 0.62 ; n = 7	OR = 1.02 [0.66-1.59] ; p = 0.99 ; n = 20	OR = 0.76 [0.48-1.21] ; p = 0.97 ; n = 19	OR = 0.91 [0.57-1.45] ; p = 0.97 ; n = 18
Horse farming (n = 968)	OR = 0.97 [0.67-1.39] ; p = 0.93 ; n = 30	OR = 0.95 [0.59-1.54] ; p = 0.99 ; n = 17	OR = 0.91 [0.47-1.76] ; p = 0.97 ; n = 9	OR = 0.92 [0.57-1.49] ; p = 0.97 ; n = 17
Other livestock farming (n = 165)	-	OR = 2.1 [0.87-5.07] ; p = 0.37 ; n = 5	-	-
Chicken or rabbit farm (n = 2156)	OR = 0.90 [0.57-1.43] ; p = 0.86 ; n = 19	OR = 0.62 [0.39-0.99] ; p = 0.37 ; n = 18	OR = 1.02 [0.71-1.46] ; p = 0.97 ; n = 31	OR = 0.92 [0.64-1.32] ; p = 0.97 ; n = 30
Other small animal farm (n = 918)	OR = 0.67 [0.36-1.22] ; p = 0.62 ; n = 11	OR = 0.96 [0.56-1.67] ; p = 0.99 ; n = 13	OR = 1.18 [0.65-2.15] ; p = 0.97 ; n = 11	OR = 1.14 [0.70-1.85] ; p = 0.97 ; n = 17
Training, dressage, stud, horse club (n = 718)	OR = 1.15 [0.57-2.33] ; p = 0.86 ; n = 8	OR = 1.1 [0.59-2.05] ; p = 0.99 ; n = 10	OR = 1.05 [0.57-1.91] ; p = 0.97 ; n = 11	OR = 1.06 [0.61-1.84] ; p = 0.97 ; n = 13
Shellfish (n = 259)	OR = 1.70 [0.69-4.15] ; p = 0.64 ; n = 5	OR = 0.95 [0.30-2.96] ; p = 0.99 ; n = 3	OR = 1.03 [0.33-3.21] ; p = 0.97 ; n = 3	**OR = 2.78 [1.48-5.21] ; p = 0.04 ; n = 10**
Non-specialized crop, polyculture, poly-breeding (n = 12219)	OR = 0.85 [0.73-0.98] ; p = 0.19 ; n = 211	OR = 1.11 [0.96-1.28] ; p = 0.38 ; n = 225	OR = 0.85 [0.69-1.04] ; p = 0.97 ; n = 105	OR = 0.91 [0.78-1.06] ; p = 0.81 ; n = 190
Salt marshes (n = 49)	-	-	-	-
Logging (n = 933)	-	OR = 0.57 [0.26-1.28] ; p = 0.42 ; n = 6	OR = 1.07 [0.55-2.07] ; p = 0.97 ; n = 9	OR = 1.36 [0.86-2.16] ; p = 0.81 ; n = 19
Hardwood sawmills (n = 72)	-	-	-	-
Agricultural contractors (n = 1390)	OR = 1.04 [0.59-1.80] ; p = 0.93 ; n = 13	OR = 0.63 [0.35-1.14] ; p = 0.37 ; n = 11	OR = 0.75 [0.41-1.36] ; p = 0.97 ; n = 11	OR = 1.36 [0.94-1.97] ; p = 0.70 ; n = 29
Garden, landscape and reforestation company (n = 2664)	OR = 1.12 [0.66-1.90] ; p = 0.86 ; n = 15	OR = 0.98 [0.67-1.43] ; p = 0.99 ; n = 28	OR = 0.9 [0.63-1.28] ; p = 0.97 ; n = 33	OR = 0.99 [0.73-1.36] ; p = 0.97 ; n = 43
Agents of the local mutual insurance company (n = 84)	-	OR = 2.6 [0.83-8.15] ; p = 0.37 ; n = 3	-	-
Rural artisan (n = 30)	OR = 2.09 [0.65-6.68] ; p = 0.62 ; n = 3	-	-	-

Odds ratio with 95% confidence interval; **Corrected p-values (Benjamini-Hochberg procedure); Empty boxes correspond to a number of individuals that are insufficient to allow the search for associations (n ≤ 3); Only results for studied LTDs were showed. No association tests were performed for four LTDs : "Complicated bilharzia" (n = 1), "Hemoglobinopathies" (n = 18), "Cystic fibrosis" (n = 9) and "Poly-diseases" (n = 1).

	Paraplegia (n = 176)	Auto-immune disease (Scleroderma and Systemic Sclerosis, Lupus, …) (n = 969)	Rheumatoid arthritis (severe) (n = 2601)	Psychiatric disorders (n = 6438)
Market gardening, floriculture (n = 3914)	OR = 0.98 [0.46-2.08] ; p = 0.99 ; n = 7	OR = 0.78 [0.54-1.14] ; p = 0.64 ; n = 29	OR = 0.97 [0.80-1.19] ; p = 0.97 ; n = 101	**OR = 0.77 [0.67-0.88] ; p = 0.001 ; n = 219**
Fruit crops (n = 2350)	-	OR = 0.74 [0.46-1.20] ; p = 0.64 ; n = 17	OR = 0.87 [0.66-1.15] ; p = 0.90 ; n = 52	OR = 0.84 [0.71-1.00] ; p = 0.21 ; n = 133
Horticultural nursery (n = 466)	**OR = 4.94 [2.03-12.04] ; p = 0.01 ; n = 5**	OR = 1.34 [0.60-3.00] ; p = 0.85 ; n = 6	OR = 0.85 [0.48-1.51] ; p = 0.90 ; n = 12	OR = 0.71 [0.48-1.06] ; p = 0.30 ; n = 25
Cereal and industrial crop, "field crops" (n = 26774)	OR = 1.06 [0.74-1.52] ; p = 0.99 ; n = 39	OR = 0.92 [0.79-1.07] ; p = 0.70 ; n = 267	**OR = 0.84 [0.77-0.93] ; p = 0.02 ; n = 564**	OR = 1.04 [0.98-1.12] ; p = 0.37 ; n = 1350
Winegrowing (n = 12429)	OR = 0.76 [0.45-1.29] ; p = 0.99 ; n = 15	OR = 0.78 [0.63-0.96] ; p = 0.25 ; n = 99	OR = 1.04 [0.92-1.17] ; p = 0.90 ; n = 318	OR = 0.94 [0.86-1.02] ; p = 0.30 ; n = 700
Forestry (n = 188)	-	-	OR = 1.16 [0.48-2.80] ; p = 0.97 ; n = 5	OR = 1.03 [0.58-1.81] ; p = 0.93 ; n = 12
Other specialized crop (n = 464)	-	-	OR = 1.04 [0.61-1.76] ; p = 0.97 ; n = 14	OR = 0.92 [0.67-1.27] ; p = 0.79 ; n = 38
Dairy farming (n = 15956)	OR = 1.01 [0.68-1.49] ; p = 0.99 ; n = 31	**OR = 1.46 [1.25-1.71] ; p = 4.58e-5 ; n = 211**	OR = 1.12 [1.01-1.23] ; p = 0.23 ; n = 500	OR = 1.07 [1.00-1.15] ; p = 0.21 ; n = 1200
Beef farming (n = 13124)	OR = 1.05 [0.67-1.64] ; p = 0.99 ; n = 22	OR = 0.94 [0.78-1.14] ; p = 0.88 ; n = 119	OR = 0.97 [0.86-1.09] ; p = 0.90 ; n = 317	**OR = 1.17 [1.09-1.25] ; p = 1.88e-4 ; n = 993**
Mixed farming (n = 3572)	OR = 0.47 [0.15-1.48] ; p = 0.99 ; n = 3	OR = 1.33 [0.98-1.81] ; p = 0.46 ; n = 43	OR = 1.30 [1.08-1.56] ; p = 0.07 ; n = 123	OR = 0.9 [0.78-1.03] ; p = 0.30 ; n = 213
Sheep or goat farming (n = 4307)	OR = 0.47 [0.18-1.28] ; p = 0.99 ; n = 4	OR = 1.00 [0.73-1.36] ; p = 0.98 ; n = 42	OR = 1.07 [0.89-1.28] ; p = 0.90 ; n = 124	OR = 1.03 [0.92-1.14] ; p = 0.81 ; n = 371
Pig farming (n = 1283)	OR = 1.44 [0.54-3.90] ; p = 0.99 ; n = 4	OR = 1.02 [0.59-1.78] ; p = 0.98 ; n = 13	OR = 1.00 [0.73-1.36] ; p = 0.99 ; n = 41	OR = 0.94 [0.77-1.16] ; p = 0.79 ; n = 98
Horse farming (n = 968)	-	OR = 0.77 [0.40-1.48] ; p = 0.85 ; n = 9	OR = 0.96 [0.66-1.39] ; p = 0.97 ; n = 29	**OR = 0.67 [0.52-0.86] ; p = 0.01 ; n = 65**
Other livestock farming (n = 165)	-	-	-	OR = 1.32 [0.86-2.04] ; p = 0.37 ; n = 21
Chicken or rabbit farm (n = 2156)	OR = 0.83 [0.31-2.24] ; p = 0.99 ; n = 4	OR = 1.46 [1.03-2.09] ; p = 0.31 ; n = 32	OR = 0.85 [0.65-1.10] ; p = 0.76 ; n = 59	OR = 0.89 [0.76-1.04] ; p = 0.32 ; n = 167
Other small animal farm (n = 918)	OR = 1.34 [0.43-4.21] ; p = 0.99 ; n = 3	OR = 1.02 [0.53-1.98] ; p = 0.98 ; n = 9	OR = 1.29 [0.92-1.81] ; p = 0.76 ; n = 34	OR = 0.83 [0.66-1.04] ; p = 0.30 ; n = 78
Training, dressage, stud, horse club (n = 718)	OR = 1.52 [0.56-4.08] ; p = 0.99 ; n = 4	OR = 0.64 [0.27-1.55] ; p = 0.70 ; n = 5	OR = 0.85 [0.55-1.31] ; p = 0.90 ; n = 21	**OR = 0.49 [0.36-0.65] ; p = 4.45e-5 ; n = 46**
Shellfish (n = 259)	-	-	-	OR = 1.07 [0.71-1.61] ; p = 0.84 ; n = 24
Non-specialized crop, polyculture, poly-breeding (n = 12219)	OR = 1.20 [0.79-1.83] ; p = 0.99 ; n = 26	OR = 0.90 [0.74-1.09] ; p = 0.70 ; n = 113	OR = 0.92 [0.81-1.03] ; p = 0.76 ; n = 312	OR = 0.98 [0.91-1.06] ; p = 0.79 ; n = 754
Salt marshes (n = 49)	-	-	-	OR = 0.84 [0.31-2.24] ; p = 0.84 ; n = 4
Logging (n = 933)	OR = 2.50 [1.03-6.07] ; p = 0.57 ; n = 5	-	OR = 0.81 [0.49-1.33] ; p = 0.90 ; n = 16	OR = 0.69 [0.51-0.93] ; p = 0.08 ; n = 43
Hardwood sawmills (n = 72)	-	-	-	OR = 1.09 [0.45-2.65] ; p = 0.88 ; n = 5
Agricultural contractors (n = 1390)	-	OR = 1.58 [0.93-2.69] ; p = 0.48 ; n = 14	OR = 0.83 [0.56-1.23] ; p = 0.90 ; n = 25	OR = 0.92 [0.73-1.16] ; p = 0.73 ; n = 76
Garden, landscape and reforestation company (n = 2664)	OR = 1.20 [0.65-2.22] ; p = 0.99 ; n = 11	OR = 1.01 [0.63-1.63] ; p = 0.98 ; n = 18	OR = 1.00 [0.79-1.27] ; p = 0.99 ; n = 73	OR = 0.90 [0.78-1.05] ; p = 0.37 ; n = 210
Agents of the local mutual insurance company (n = 84)	-	-	-	OR = 0.55 [0.21-1.48] ; p = 0.39 ; n = 4
Rural artisan (n = 30)	-	-	-	-

Odds ratio with 95% confidence interval; **Corrected p-values (Benjamini-Hochberg procedure); Empty boxes correspond to a number of individuals that are insufficient to allow the search for associations (n ≤ 3); Only results for studied LTDs were showed. No association tests were performed for four LTDs : "Complicated bilharzia" (n = 1), "Hemoglobinopathies" (n = 18), "Cystic fibrosis" (n = 9) and "Poly-diseases" (n = 1).

	Crohn disease & Hemorrhagic Colitis (n = 930)	Multiple sclerosis (n = 410)	Scoliosis (n = 87)	Ankylosing spondylitis (severe) (n = 1048)
Market gardening, floriculture (n = 3914)	OR = 0.82 [0.57-1.17] ; p = 0.72 ; n = 32	OR = 1.00 [0.61-1.62] ; p = 0.99 ; n = 17	OR = 1.43 [0.58-3.52] ; p = 0.98 ; n = 5	OR = 0.87 [0.62-1.20] ; p = 0.70 ; n = 37
Fruit crops (n = 2350)	OR = 0.89 [0.55-1.44] ; p = 0.87 ; n = 17	OR = 0.82 [0.39-1.72] ; p = 0.99 ; n = 7	OR = 1.58 [0.50-5.00] ; p = 0.98 ; n = 3	OR = 0.81 [0.51-1.28] ; p = 0.70 ; n = 18
Horticultural nursery (n = 466)	OR = 1.21 [0.58-2.56] ; p = 0.87 ; n = 7	-	-	OR = 1.07 [0.51-2.26] ; p = 0.97 ; n = 7
Cereal and industrial crop, "field crops" (n = 26774)	OR = 1.01 [0.85-1.20] ; p = 0.98 ; n = 190	OR = 1.27 [1.00-1.63] ; p = 0.69 ; n = 85	OR = 1.46 [0.91-2.33] ; p = 0.98 ; n = 25	OR = 0.85 [0.72-1.00] ; p = 0.70 ; n = 178
Winegrowing (n = 12429)	OR = 1.32 [1.09-1.61] ; p = 0.08 ; n = 127	OR = 1.24 [0.93-1.66] ; p = 0.93 ; n = 53	OR = 0.64 [0.29-1.38] ; p = 0.98 ; n = 7	OR = 1.05 [0.87-1.27] ; p = 0.90 ; n = 117
Forestry (n = 188)	-	-	-	-
Other specialized crop (n = 464)	OR = 1.47 [0.73-2.96] ; p = 0.72 ; n = 8	-	-	OR = 0.66 [0.25-1.77] ; p = 0.70 ; n = 4
Dairy farming (n = 15956)	OR = 0.95 [0.79-1.14] ; p = 0.87 ; n = 155	OR = 0.86 [0.67-1.11] ; p = 0.93 ; n = 71	OR = 0.80 [0.44-1.45] ; p = 0.98 ; n = 13	OR = 1.08 [0.93-1.26] ; p = 0.70 ; n = 211
Beef farming (n = 13124)	OR = 0.72 [0.57-0.91] ; p = 0.08 ; n = 79	OR = 1 [0.74-1.35] ; p = 0.99 ; n = 49	OR = 0.84 [0.42-1.67] ; p = 0.98 ; n = 9	OR = 0.92 [0.75-1.11] ; p = 0.70 ; n = 116
Mixed farming (n = 3572)	OR = 0.87 [0.59-1.27] ; p = 0.80 ; n = 28	OR = 0.9 [0.53-1.53] ; p = 0.99 ; n = 14	-	OR = 1.24 [0.93-1.66] ; p = 0.70 ; n = 49
Sheep or goat farming (n = 4307)	OR = 1 [0.74-1.37] ; p = 0.98 ; n = 44	OR = 0.77 [0.47-1.25] ; p = 0.93 ; n = 17	OR = 1.79 [0.86-3.70] ; p = 0.98 ; n = 8	OR = 1.17 [0.90-1.53] ; p = 0.70 ; n = 58
Pig farming (n = 1283)	OR = 0.77 [0.43-1.37] ; p = 0.79 ; n = 12	OR = 1.25 [0.64-2.41] ; p = 0.99 ; n = 9	-	OR = 0.84 [0.51-1.38] ; p = 0.77 ; n = 16
Horse farming (n = 968)	OR = 0.54 [0.24-1.21] ; p = 0.69 ; n = 6	OR = 1.24 [0.61-2.50] ; p = 0.99 ; n = 8	-	OR = 0.95 [0.54-1.69] ; p = 0.97 ; n = 12
Other livestock farming (n = 165)	-	-	-	-
Chicken or rabbit farm (n = 2156)	OR = 1.01 [0.67-1.51] ; p = 0.98 ; n = 25	OR = 0.94 [0.54-1.63] ; p = 0.99 ; n = 13	-	OR = 1.05 [0.74-1.49] ; p = 0.97 ; n = 33
Other small animal farm (n = 918)	OR = 1.46 [0.90-2.37] ; p = 0.69 ; n = 17	OR = 1.88 [1.03-3.44] ; p = 0.69 ; n = 11	-	OR = 1.55 [0.99-2.42] ; p = 0.70 ; n = 20
Training, dressage, stud, horse club (n = 718)	OR = 0.67 [0.36-1.26] ; p = 0.72 ; n = 10	OR = 0.93 [0.46-1.88] ; p = 0.99 ; n = 8	-	OR = 0.58 [0.30-1.13] ; p = 0.70 ; n = 9
Shellfish (n = 259)	OR = 2.52 [1.25-5.08] ; p = 0.08 ; n = 8	-	-	OR = 0.93 [0.30-2.89] ; p = 0.97 ; n = 3
Non-specialized crop, polyculture, poly-breeding (n = 12219)	OR = 0.88 [0.72-1.08] ; p = 0.72 ; n = 116	OR = 0.84 [0.61-1.14] ; p = 0.93 ; n = 45	OR = 0.99 [0.53-1.87] ; p = 0.98 ; n = 11	OR = 0.90 [0.74-1.08] ; p = 0.70 ; n = 122
Salt marshes (n = 49)	-	-	-	-
Logging (n = 933)	OR = 0.63 [0.30-1.34] ; p = 0.72 ; n = 7	-	-	OR = 0.74 [0.38-1.42] ; p = 0.70 ; n = 9
Hardwood sawmills (n = 72)	-	-	-	-
Agricultural contractors (n = 1390)	OR = 1.22 [0.77-1.92] ; p = 0.79 ; n = 19	OR = 0.68 [0.25-1.83] ; p = 0.99 ; n = 4	-	OR = 0.59 [0.32-1.10] ; p = 0.70 ; n = 10
Garden, landscape and reforestation company (n = 2664)	OR = 1.14 [0.87-1.50] ; p = 0.79 ; n = 58	OR = 0.73 [0.43-1.26] ; p = 0.93 ; n = 14	OR = 2.20 [0.78-6.20] ; p = 0.98 ; n = 4	OR = 1.11 [0.85-1.45] ; p = 0.70 ; n = 60
Agents of the local mutual insurance company (n = 84)	-	-	-	-
Rural artisan (n = 30)	-	-	-	-

Odds ratio with 95% confidence interval; **Corrected p-values (Benjamini-Hochberg procedure); Empty boxes correspond to a number of individuals that are insufficient to allow the search for associations (n ≤ 3); Only results for studied LTDs were showed. No association tests were performed for four LTDs : "Complicated bilharzia" (n = 1), "Hemoglobinopathies" (n = 18), "Cystic fibrosis" (n = 9) and "Poly-diseases" (n = 1).

	Result of organ transplantation (n = 197)	Tuberculosis & Leprosis (n = 52)	Malignant neoplasms (n = 25934)	Serious illness off list (n = 5741)
Market gardening, floriculture (n = 3914)	OR = 1.53 [0.85-2.74] ; p = 0.97 ; n = 12	OR = 1.98 [0.71-5.49] ; p = 0.89 ; n = 4	OR = 1 [0.94-1.07] ; p = 0.97 ; n = 1031	OR = 0.90 [0.78-1.04] ; p = 0.44 ; n = 201
Fruit crops (n = 2350)	OR = 0.92 [0.34-2.48] ; p = 0.97 ; n = 4	-	OR = 1.09 [1.01-1.18] ; p = 0.13 ; n = 651	OR = 0.77 [0.63-0.93] ; p = 0.05 ; n = 112
Horticultural nursery (n = 466)	-	-	OR = 0.94 [0.78-1.12] ; p = 0.61 ; n = 126	OR = 0.88 [0.58-1.33] ; p = 0.99 ; n = 23
Cereal and industrial crop, "field crops" (n = 26774)	OR = 1.09 [0.77-1.54] ; p = 0.97 ; n = 42	OR = 0.46 [0.21-1.02] ; p = 0.89 ; n = 7	**OR = 0.96 [0.93-0.99] ; p = 0.02 ; n = 6756**	**OR = 0.83 [0.78-0.89] ; p = 1.15e-6 ; n = 1555**
Winegrowing (n = 12429)	OR = 1.41 [0.95-2.10] ; p = 0.97 ; n = 29	OR = 0.81 [0.32-2.04] ; p = 0.98 ; n = 5	**OR = 1.12 [1.07-1.16] ; p = 1.02e-6 ; n = 3375**	OR = 0.90 [0.82-0.98] ; p = 0.07 ; n = 660
Forestry (n = 188)	-	-	OR = 1.1 [0.83-1.46] ; p = 0.62 ; n = 51	OR = 1.18 [0.66-2.09] ; p = 0.99 ; n = 12
Other specialized crop (n = 464)	-	-	OR = 1.02 [0.86-1.22] ; p = 0.95 ; n = 131	OR = 0.73 [0.47-1.12] ; p = 0.44 ; n = 21
Dairy farming (n = 15956)	OR = 1.04 [0.72-1.49] ; p = 0.97 ; n = 36	OR = 1.33 [0.67-2.66] ; p = 0.98 ; n = 10	**OR = 0.94 [0.91-0.97] ; p = 0.005 ; n = 4117**	OR = 1.10 [1.02-1.19] ; p = 0.07 ; n = 932
Beef farming (n = 13124)	OR = 0.69 [0.42-1.12] ; p = 0.97 ; n = 18	OR = 1.02 [0.44-2.39] ; p = 0.98 ; n = 6	OR = 0.98 [0.94-1.02] ; p = 0.48 ; n = 3250	OR = 1.10 [1.02-1.19] ; p = 0.07 ; n = 827
Mixed farming (n = 3572)	OR = 1.16 [0.59-2.26] ; p = 0.97 ; n = 9	-	OR = 1.00 [0.94-1.07] ; p = 0.97 ; n = 953	OR = 1.08 [0.94-1.25] ; p = 0.76 ; n = 196
Sheep or goat farming (n = 4307)	OR = 0.91 [0.45-1.85] ; p = 0.97 ; n = 8	-	OR = 0.97 [0.91-1.04] ; p = 0.52 ; n = 1056	**OR = 1.19 [1.06-1.34] ; p = 0.04 ; n = 319**
Pig farming (n = 1283)	-	-	OR = 1.09 [0.98-1.21] ; p = 0.25 ; n = 404	OR = 1.11 [0.87-1.40] ; p = 0.88 ; n = 73
Horse farming (n = 968)	OR = 1.61 [0.51-5.07] ; p = 0.97 ; n = 3	-	OR = 0.89 [0.79-1.02] ; p = 0.25 ; n = 254	OR = 1.07 [0.85-1.35] ; p = 0.99 ; n = 75
Other livestock farming (n = 165)	-	-	OR = 0.76 [0.54-1.08] ; p = 0.30 ; n = 34	OR = 1.05 [0.58-1.90] ; p = 0.99 ; n = 11
Chicken or rabbit farm (n = 2156)	-	-	OR = 0.99 [0.91-1.08] ; p = 0.97 ; n = 578	OR = 1.02 [0.85-1.23] ; p = 0.99 ; n = 114
Other small animal farm (n = 918)	-	-	OR = 0.91 [0.80-1.05] ; p = 0.41 ; n = 224	OR = 1.04 [0.80-1.34] ; p = 0.99 ; n = 60
Training, dressage, stud, horse club (n = 718)	-	-	OR = 1.08 [0.94-1.24] ; p = 0.52 ; n = 212	OR = 0.95 [0.71-1.26] ; p = 0.99 ; n = 48
Shellfish (n = 259)	-	-	OR = 0.88 [0.67-1.16] ; p = 0.52 ; n = 53	OR = 1.02 [0.59-1.76] ; p = 0.99 ; n = 13
Non-specialized crop, polyculture, poly-breeding (n = 12219)	OR = 0.92 [0.60-1.40] ; p = 0.97 ; n = 24	OR = 0.91 [0.39-2.13] ; p = 0.98 ; n = 6	**OR = 0.91 [0.88-0.95] ; p = 7.20e-5 ; n = 3121**	OR = 1.00 [0.92-1.08] ; p = 0.99 ; n = 706
Salt marshes (n = 49)	-	-	OR = 1.31 [0.76-2.24] ; p = 0.52 ; n = 14	-
Logging (n = 933)	-	-	OR = 0.91 [0.78-1.06] ; p = 0.44 ; n = 183	OR = 1.30 [0.98-1.72] ; p = 0.24 ; n = 51
Hardwood sawmills (n = 72)	-	-	OR = 0.56 [0.31-1.02] ; p = 0.19 ; n = 11	OR = 1.52 [0.63-3.67] ; p = 0.84 ; n = 5
Agricultural contractors (n = 1390)	-	-	OR = 1.00 [0.89-1.13] ; p = 0.97 ; n = 307	OR = 1.00 [0.77-1.29] ; p = 0.99 ; n = 60
Garden, landscape and reforestation company (n = 2664)	OR = 0.90 [0.44-1.85] ; p = 0.97 ; n = 8	OR = 1.16 [0.36-3.71] ; p = 0.98 ; n = 3	**OR = 0.87 [0.80-0.95] ; p = 0.01 ; n = 576**	OR = 0.99 [0.83-1.18] ; p = 0.99 ; n = 136
Agents of the local mutual insurance company (n = 84)	-	-	OR = 1.19 [0.81-1.76] ; p = 0.52 ; n = 27	OR = 1.00 [0.41-2.42] ; p = 0.99 ; n = 5
Rural artisan (n = 30)	-	-	**OR = 0.22 [0.11-0.46] ; p = 5.35e-4 ; n = 7**	OR = 1.02 [0.45-2.31] ; p = 0.99 ; n = 6

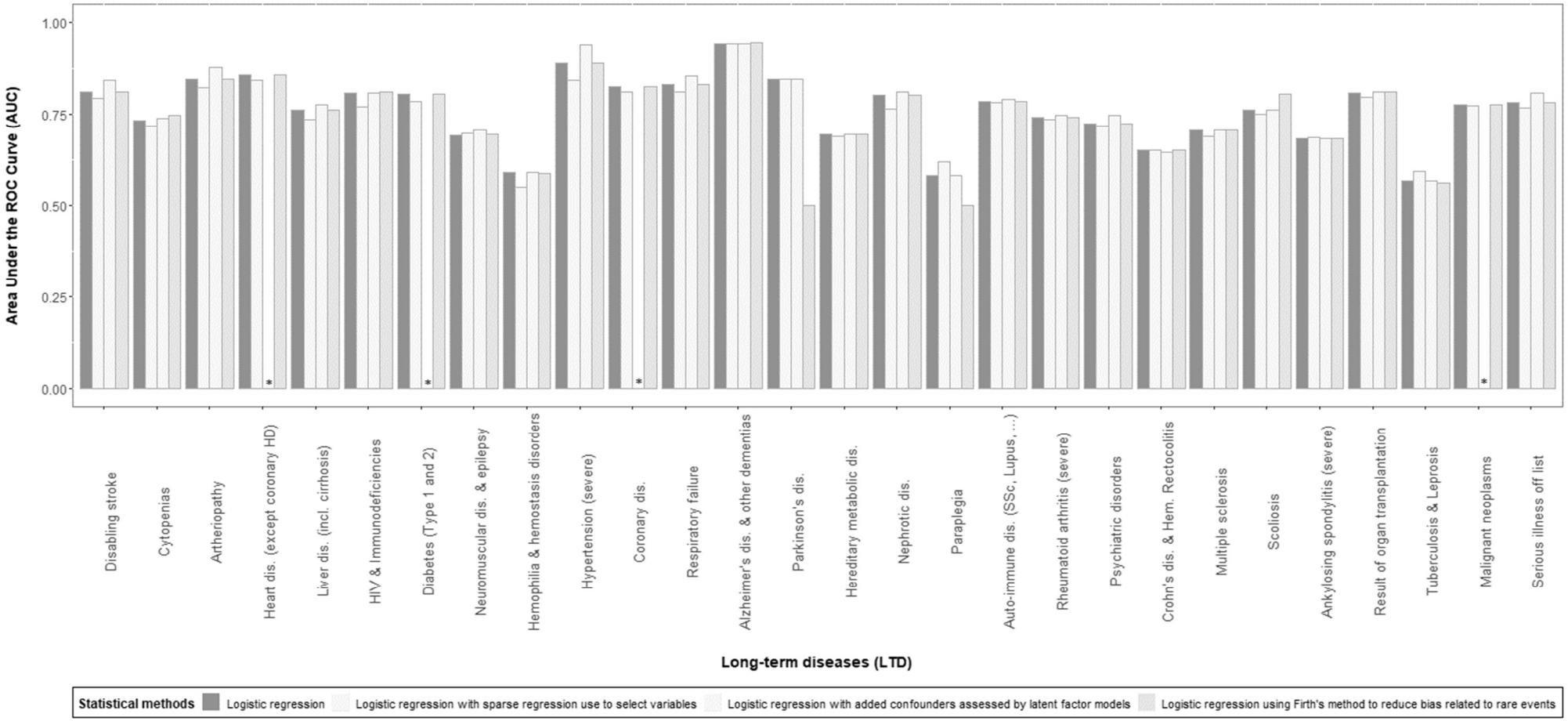
Odds ratio with 95% confidence interval; **Corrected p-values (Benjamini-Hochberg procedure); Empty boxes correspond to a number of individuals that are insufficient to allow the search for associations (n ≤ 3); Only results for studied LTDs were showed. No association tests were performed for four LTDs : "Complicated bilharzia" (n = 1), "Hemoglobinopathies" (n = 18), "Cystic fibrosis" (n = 9) and "Poly-diseases" (n = 1).

Supplement 4: Specific selections of variables for each long-term disease, used for logistic regression applied on self-employed workers affiliated to the MSA between 2006 and 2016

Long-term diseases (LTD)*	N	Occupational activities (OA)	Installation year	Gender	Eligible for minimum income benefit	Eligible for unemployment allowance	Farm or enterprise area	Age	Number of OAs declared during the study period	Number of observation years	Incomes	LTD declared before study period	Family status: « married »	Family status: « widower »	Family status: « divorced or separated »	Number of employees	Professional status	Family collaborator(s)	Type of exploitation	Administrative region	Previous illnesses**	Area under the curve ROC (validation datasets)
Disabling stroke	5288	X	X					X				X			X		X				X	0.81
Cytopenias	284	X						X									X				X	0.73
Artheriopathy	4942	X	X	X			X	X				X			X	X	X				X	0.85
Heart diseases (except coronary heart diseases)	16977	X	X	X				X			X	X		X			X		X		X	0.86
Liver diseases (including cirrhosis)	1320	X	X	X	X			X				X			X		X		X		X	0.76
HIV & Immunodeficiencies	241	X						X					X				X					0.81
Diabetes (Type 1 and 2)	25229	X	X	X	X	X		X	X		X	X					X		X	X	X	0.80
Neuromuscular diseases and epilepsy	1927	X	X		X							X					X				X	0.69
Hemophilia & hemostasis disorders	257	X															X				X	0.59
Hypertension (severe)	5678	X	X					X			X	X					X		X	X	X	0.89
Coronary diseases	13210	X	X	X				X				X	X	X	X		X			X	X	0.83
Respiratory failure	3122	X	X		X			X			X	X				X	X			X	X	0.83
Alzheimer's disease & other dementias	2550	X	X	X			X	X		X		X		X			X			X	X	0.94
Parkinson's disease	1686	X	X	X				X		X			X				X				X	0.85
Hereditary metabolic diseases	968	X	X									X					X			X	X	0.69
Nephrotic diseases	1676	X	X					X				X					X				X	0.80
Paraplegia	176	X															X					0.58
Auto-immune disease (Scleroderma and Systemic Sclerosis, Lupus, …)	969	X	X	X				X				X					X				X	0.78
Rheumatoid arthritis (severe)	2601	X	X	X				X				X	X				X				X	0.74
Psychiatric disorders	6438	X	X	X	X			X			X	X			X		X		X	X	X	0.72
Crohn disease & Hemorrhagic Colitis	930	X						X		X		X					X			X		0.65
Multiple sclerosis	410	X		X				X		X							X					0.71
Scoliosis	87	X		X													X					0.76
Ankylosing spondylitis (severe)	1048	X						X		X		X	X				X					0.68
Result of organ transplantation	197	X		X				X				X					X				X	0.81
Tuberculosis & Leprosis	52	X																				0.57
Malignant neoplasms	25934	X	X					X					X		X		X			X	X	0.78
Serious illness off list	5741	X	X	X				X			X	X					X			X	X	0.78

Only LTDs studied in this work. No association tests were performed on four LTDs : "Complicated bilharzia" (n = 1), "Hemoglobinopathies" (n = 18), "Cystic fibrosis" (n = 9) and "Poly-diseases" (n = 1). **For this variable, the cross means that at least one of the four components of the principal component analysis has been included in the logistic model for this LTD.

Supplement 5: Comparison of the areas under the ROC curve (AUC) for each long-term disease (LTD) of the different statistical methods performed whose aims to test robustness of the chosen method ("regression logistic") in our work, performed on self-employed workers affiliated to the MSA between 2006 and 2016



*For four LTDs, the regression model with added confounders assessed by latent factor models failed to identify any significant association. It was chosen to set the AUC value to 0.

Baptiste VESPER GUILLON

ÉVALUATION DU RISQUE DE CANCERS DIGESTIFS SELON L'EXPOSITION PROFESSIONNELLE CHEZ LES AGRICULTEURS FRANÇAIS, D'APRÈS LES DONNÉES MÉDICO-ADMINISTRATIVES DE LA MUTUALITÉ SOCIALE AGRICOLE (MSA)

RÉSUMÉ

Introduction : Les travailleurs agricoles ont un plus faible risque de cancer digestif. Il s'agit pourtant d'une population fréquemment exposée à de nombreux risques professionnels, parmi lesquels figuraient certains agents classés cancérogènes certains par le CIRC, et figurent toujours certains agents classés cancérogènes probables ou possibles, avec des variabilités d'exposition au sein des différentes activités agricoles. Le but de cette étude était d'évaluer le risque de cancer digestif en fonction de l'exposition professionnelle chez les agriculteurs français.

Méthodes : L'ensemble des exploitants agricoles ayant cotisé au moins une fois à la MSA entre 2006 et 2016 a été étudié. Les déclarations de cancers digestifs en ALD 30, au sein de cette population, ont été relevées de 2012 à 2016, en précisant le sous-type de cancer selon la classification CIM-10. A l'aide de modèles logistiques, permettant l'ajustement sur le sexe, l'âge et d'autres variables, les associations entre déclaration de cancers digestifs et activités agricoles ont été étudiées et exprimées sous forme d'Odds Ratio. Une correction (Benjamini-Hochberg) a été effectuée pour prendre en compte les tests multiples.

Résultats : Sur les 899 212 exploitants agricoles étudiés, 4438 cancers digestifs ont été déclarés. Le risque de cancer digestif était augmenté chez les viticulteurs (OR=1.28 [95% CI 1.16-1.42], p<0.00005), principalement lié à une augmentation du risque de cancer colorectal (OR=1.21 [95% CI 1.09-1.35], p=0.01). Une tendance à l'augmentation du risque de cancer du grêle a également été mise en évidence (OR=2.19 [95% CI 1.31–3.67], proche de la significativité (p=0.07 après application de la correction concernant les tests multiples). Par ailleurs, une diminution du risque de cancer digestif a été retrouvée chez les producteurs céréaliers (OR=0.87 [95% CI 0.8–0.95]), sous-tendue par une diminution de risque de cancer de l'estomac (OR=0.6 [95% CI 0.45–0.81]).

Conclusion : Nos résultats montrent une augmentation de risque de déclaration en ALD de cancer digestif et principalement de cancer colorectal chez les viticulteurs français. Il est également retrouvé une diminution du risque de cancer digestif chez les producteurs céréaliers, principalement une diminution de risque de cancer de l'estomac.

MOTS CLÉS : Cancers digestifs ; Agriculteurs ; Exposition professionnelle ; Facteurs de risque ; Cohorte

FILIÈRE : Gastro-entérologie, hépatologie et oncologie digestive, CHU Grenoble Alpes, France

Baptiste VESPER GUILLON

DIGESTIVE CANCER RISK AND OCCUPATIONAL EXPOSURE IN FRENCH FARMERS: RESULTS FROM MUTUALITE SOCIALE AGRICOLE (MSA) MEDICO-ADMINISTRATIVE DATABASES

ABSTRACT

Introduction - Farmers are at lower risk to present with digestive cancer. However, this population is frequently exposed to carcinogens, with exposure variability depending on the agricultural activity. The aim of our study was to evaluate the risk of digestive cancer according to agricultural activity among French farmers.

Methods - We included all farmers recorded at least once in the social security system of French agricultural workers "Mutualité sociale agricole" (MSA) between January 2006 and December 2016. We recorded the digestive cancers in this population reported as Long Term Disease (LTD) from January 2012 to December 2016. We specified for each case the cancer subtype according to the ICD-10 classification. The association between digestive cancer and agricultural activity was studied with logistic models, and expressed as Odds Ratio

Results - 4,438 digestive cancers were reported among 899,212 included farmers. The risk of presenting with digestive cancer was increased in wine growers (OR=1.28 [95% CI 1.16-1.42]), mainly an increased risk of colorectal cancer (OR=1.21 [95% CI 1.09-1.35]). An increased risk of small bowel cancer was also identified but with a limited statistical significance (OR=2.19 [95% CI 1.31-3.67], p=0.08). Furthermore, a lower risk of digestive

cancer was identified among cereal crop farmers (OR=0.87 [95% CI 0.8-0.95]), supported by a lower risk of stomach cancer (OR=0.6 [95% CI 0.45-0.81]).

Conclusion - Our results suggest an increased risk of digestive cancer among French wine growers, mainly colorectal cancer. We also found a lower risk of digestive cancer among cereal crop farmers, mainly gastric cancer.

KEYWORDS - Gastrointestinal Neoplasms; Farmers; Risk Factors; Occupational Exposure; Cohort Studies

ABRÉVIATIONS

AGRICAN: AGRIculture and CANcer cohort

AHS: Agricultural Health Study

CI: Confidence Interval

IARC: Agency for Research on Cancer

ICD-10: International Classification of WHO Diseases, 10th edition

LTD: Long Term Disease

MCE: Matrice Culture Exposition

MSA: Mutualité Sociale Agricole

OR: Odds Ratio

ARTICLE

DIGESTIVE CANCER RISK AND OCCUPATIONAL EXPOSURE IN FRENCH FARMERS: RESULTS FROM MUTUALITE SOCIALE AGRICOLE (MSA) MEDICO-ADMINISTRATIVE DATABASES

INTRODUCTION

Cancer epidemiology

France has one of the highest incidence of cancer among European countries: 382,000 new cases were reported in 2018 [1]. Globally, it is estimated that one in five men and one in six women will develop cancer during their lifetime [2]. In terms of mortality, 157,000 deaths were cancer related in France, in 2018, making it the leading cause of specific mortality, regardless of gender, before cardiovascular diseases and infectious causes, despite recent and ongoing progress in early detection and treatment [1].

Digestive cancers

The incidence of digestive cancers is high in France, with colorectal cancer ranking first. The aging population, improved screening techniques, but also changes in lifestyle such as increased level of inactivity, have probably led an increased incidence since the 1980s. In 2017, there were an estimated 45,000 new cases per year of colorectal cancer, making it the second most common cancer in women and the third most common in men [3]. It is also one of the deadliest since it is ranked third in men and second in women in terms of mortality [3][4]. A rapid and poorly explained increase in the incidence of pancreatic cancer has also been reported over the past ten years (+ 2.7% per year in men and + 3.8% in women between

1990 and 2018), while these figures have not increased similarly in other industrialized countries [3] [5] [6].

Risk factors and occupational exposures

A study by the International Agency for Research on Cancer (IARC) in 2000 shows that about 35% of incident cancers could be preventable if exposure to risk factors had been at the optimal level or reference level [7]. Numerous published IARC monographs list the various types of known or suspected carcinogens, whether chemical, physical, viruses, dust, ionizing radiation, but also new lifestyle factors such as food quality. Tobacco and alcohol are the best-known and incriminated risk factors, accounting respectively for 20% and 8% of incident cancers [8][9]. Food consumption and obesity both account for 5.4% of the fractions attributable to new cases. Regular physical activity could be a protective factor [10]. 3.6% of incident cancers are attributable to occupational exposure and 12% of employees have been exposed at their workplace to at least one carcinogen (chemical or non-chemical), for occupational medicine [7]. More specifically for digestive cancers, the attributable fractions to occupational exposure were respectively 4% for stomach, 2% for colon-rectum, and 3.7% for liver cancer [9. Studies focusing on a relationship between occupational exposure and cancer risk are valuable tools for prevention campaigns and the legal framework of product use.

Agricultural workers

The relationship between cancer and agriculture has been studied often since the early 1980s and has led to several meta-analyses [11] [12]. The authors of previous major studies have reported a reduction of cancer risk in this population compared to the global population, concerning digestive cancers [13]. Two major agricultural cohorts have allowed confirming

this: the Agricultural Health Study (AHS) among US private pesticide applicators from North Carolina and Iowa [14][15] and the AGRIculture and CANcer (AGRICAN) cohort among French farmers [16]. Possible explanations were a lower rate of tobacco consumption in this population high levels of occupational physical activity, and possibly a safer diet (higher fruits and vegetables intake). In the AGRICAN cohort, however, there was a slightly increased risk of stomach cancer in female farmers and in those who never used pesticides.

However, farmers are frequently exposed to many potentially carcinogenic agents such as pesticides, cristalline silica, ultraviolet rays, and viruses.[17].

Monitoring diseases in this exposed population is crucial for the identification and prevention of the new threats. Using insurance data for the detection of these risks, complementary to conventional epidemiological studies, could improve early detection.

The Mutualité Sociale Agricole (MSA) is a social security system covering all French agricultural workers for the following issues : diseases, family, occupational health and retirement. The MSA has undertaken to use its medico-administrative data for research purposes mainly on occupational risks vigilance. We decided to use these medico-administrative databases in our study to identify, without any prior assumption, any association between various agricultural activities and the occurrence of digestive cancers.

METHODS

Agreements

The massive data mining project in which this study is integrated was approved by the French Data Protection Authority CNIL (MMS/SBM/AE171001).

Study Population

The authors of a pilot study demonstrated the feasibility and relevance of MSA medico-administrative database use to assess risks among agricultural workers [18]. We used this cohort by focusing specifically on digestive cancers. MSA is the French national health insurance for every agricultural worker. It groups two different statuses of agricultural workers: farm owners and farm salaried workers, whether active or retired. They can be farmers, breeders, but also less frequent occupations such as beekeepers, foresters, shellfish farmers, and even tertiary workers serving the agricultural population (bank, insurance and MSA workers themselves. In 2016, 3.3 million people were affiliated with the MSA. We included all adults 18 of age or more, men and women, active farm owners, recorded in the MSA at least once between 2006 and 2016 (Table 1). Farm workers (employees) were not included in the analysis as a large proportion of them have either seasonal jobs, or move shortly from one contract to another, which introduce a higher variability in their occupations, and makes the assessment of occupational activity more complex and less accurate, so less relevant for our analysis. We did not take into account farm owners retired before 2006 since we have no information about occupational activity for them.

Data collection

Data collected from the MSA records was divided in two different categories: administrative database and medico-administrative database. The administrative database was implemented in 2006 and included sex, date of birth, region of residence ("Departement" administrative division), socio-economic status, first year of activity and agricultural activity as defined by the MSA thesaurus (Table 2). Each individual had to report his agricultural activity once a year and was regularly controlled.

The medico-administrative database was implemented in 2012. We identified all Long-Term Disease (LTD) declarations and the corresponding disease sub-category following the International Classification of WHO Diseases (ICD-10) (Table 3) for every year in this database. Chronic diseases such as cancers are reported to the health insurance system, in France, by family practitioners as LTDs so that patients can have free medical care adapted to the disease. LTD 30 is the code for malignant tumors. We collected all ICD-10 codes from C15 to C26 (digestive cancers) in the database, reported between January 2012 and December 2016. We decided to group some ICD-10 codes because of their clinical and epidemiological similarity, with the aim to increase statistical power. Malignant neoplasms of the colon were grouped with rectosigmoid junction and rectum neoplasms (C18, C19, and C20 respectively); gallbladder neoplasms were grouped with other and unspecified biliary tract neoplasms (C23 and C24 respectively).

Statistical analysis

Two matrices were generated with ICD-10 codes on one hand and occupational activities on the other hand, studied as binary variables. Each association between one ICD-10 code and one occupational activity was studied with a specific model build with stepwise selection of variables. Because of the great number of tests, it was essential to control the potentially false identifications and highlight the relevant associations. We chose the Benjamini-Hochberg procedure, stable but little conservative, to adjust p-values. We calculated p-values for each association, with an alpha risk of 0.05. P-values <0.05, after correction for multiple testing, were considered as statistically significant. Results were computed as Odds Ratio (OR) expressed with a 95% confidence interval (CI), adjusted on age, sex, year of first installation, matrimonial status, affiliation status (according to percentage of incomes from agriculture : 100% incomes from agriculture ; > 50% incomes from agriculture ; < 50% incomes from

agriculture), exploitation administrative area, and history of any LTD. All statistical analyses were performed using R statistical software.

RESULTS

Study Population

The main characteristics of the study population are listed in Table 1. 899,212 farm-owners were identified; mostly men (70.2%) with a median age of 50.2 years (± 12.9 for men and ± 13.5 for women), median year of first installation in 1995 (1940 to 2016), and 69.5% were exclusively working in agriculture. We identified, for the five main activities, 226,607 (23%) people working in cereal crops, 136,146 (13.8%) in dairy cattle, 104,128 (10.6%) in winegrowing, 101,711 (10.3%) in beef cattle, and 112,983 (11.5%) in mixed crops and breeding.

Occupational exposure and cancer risk

We identified 4,438 digestive cancers (Table 4) between January 2012 and December 2016. Colorectal cancer was the most frequent subtype (n=2,930; 66%), followed by pancreatic cancer (n=434; 9.8%), gastric cancer (n=314; 7%), esophagus cancer (n=290; 6.5%), and liver and intrahepatic bile ducts cancer (n=260; 5.9%). The associations between occupational activities and occurrence of digestive cancer are listed in Table 5 and Table 6

Winegrowing

Winegrowing was associated with an overall increased risk of digestive cancer of nearly 30% (OR=1.28, 95%CI 1.16–1.42; p<0.00005). The risk increase was preponderant for colorectal cancer, nearly 20%, resistant to correction of multiple tests (OR=1.21, 95%CI 1.09–1.35; p=0.01). We found a higher risk of small intestine cancer for winegrowing before multiple

test correction (OR=2.19, 95%CI 1.31–3.67), still nearly significant, after Benjamini-Hochberg correction (p=0.07). The high level of statistical significance for all cancers OR estimate is partly due to the fact that, except for esophagus, all cancer risks estimates were above 1, even if none of them, except for colorectal cancer, was statistically significant by itself.

Cereal crop

We found an overall lower risk for all digestive cancers in cereal crop farmers, of nearly 13% (OR=0.87, 95%CI 0.8–0.95; p=0.02) and more specifically a significantly lower risk of 40% of stomach cancer (OR=0.6, 95%CI 0.45–0.81; p=0.02).

Other occupational activities

No significant difference in risk was identified for other cancer sites such as esophagus, liver and bile ducts, pancreas, anal canal, regardless of occupational activity.

DISCUSSION

This large-scale study highlighted an increased risk of digestive cancer and more specifically of colorectal cancer among French winegrowers. It also highlighted a reduction in the risk of stomach cancer in cereal crop farmers.

Limitation and strength of our study

The first strength is the completeness of data collection concerning farm-owners since the MSA is the one and only social scheme for these professionals in France.

The follow-up period was long, since the median year of installation was 1995 and the first recorded agricultural activity went back to 1940. Cancer declarations have been recorded since 2012, allowing for a delay between professional exposure and occurrence of cancer consistent with the current data on carcinogenic processes [19][20][21][22][23].

We performed a robustness test, to avoid the risk of over-adjustment, by limiting the number of adjustment variables. We kept only age, sex, year of installation, health regimen, and geographic region. The results obtained were the same as in the main analysis, reinforcing the power of our study.

Our study however, had several limitations. The occupational activity was assessed as an exposed / not exposed dichotomous form. Our data prevent us to take into account the duration of activity and thus the duration of exposure, whereas we suppose that the real risk depend on cumulative exposure, and that exposure might have change over successive periods (change in pesticides use, etc).

We were unable to include tobacco abuse status in the adjustment variables since this data was not present in the MSA records. However, it is well known that tobacco abuse among farmers is lower than in the global population (21% vs. 34% respectively), which would tend to support our results [24].

We chose to include only farm-owners because they were considered as more stable in their professional activity, whereas it was difficult to establish a work time threshold for farm-salaried-workers that could be qualified as "stable". In the Maugard pilot study, 75% of employees had to be excluded because they did not meet the criteria for stable employment

(>1,014h / year) [18]. The AGRICAN cohort, however, showed a slightly increased cancer risk in the farm-worker subgroup compared to farm-owners (+ 14%) [16]. Not taking this subpopulation into account was therefore a limitation of our study, but a database review is ongoing to include them in the near future.

Differences between observed and predicted incidence using the LTD variable were assessed in French administrative subdivision using a cancer registry [25]. This method was chosen to estimate the overall incidence of colorectal cancer in France from data in administrative subdivision with a cancer registry, with a relative error of prediction inferior to 15%. The authors showed that the use of French National Hospital Information System (PMSI, Programme de Médicalisation des Systèmes d'Information) data combined with LTD allowed improving the quality of incidence prediction for pancreatic, esophagus, and stomach cancer. It is therefore possible that incidence for these locations are higher than the occurrence found in our study, which could decrease the risk estimate for these cancers. Completing our study with hospital PMSI data could improve our results. Furthermore, the MSA medico-administrative database also contains data on medicine consumption. Integrating this data with the LTD declaration could help identify more patients with cancer.

Winegrowing

Published data concerning associations between winegrowing and digestive cancer risk is scarce. Some authors have shown an increased risk of rectosigmoid junction carcinoma among residents of a German winegrowing community (area under wine cultivation >20% of community area), but with no excess of cancer of either colon, rectum, or colon and sigmoid and rectum all together [26]. Thus, this result might be due to chance. This population was also characterized by an overall lower risk of digestive cancers. Winegrowing is the

agricultural activity in which pesticides are the most used in France (20% of pesticides for only 3% of the country's agricultural area)[27]. One of the main reasons for this increased risk could be the use of arsenic. Arsenical pesticides have been banned since 1973, except for winegrowing where they were allowed until 2001. It is estimated that between 1979 and 2001, from 60,000 to 100,000 winegrowers used this type of pesticide [28]. Inorganic arsenic is classified as carcinogenic to humans (group 1) by IARC and several authors have well documented the link between its use and the risk of skin, lung, bladder, prostate, and liver cancer [29]. A decreased incidence of colorectal cancer was reported after decreasing arsenic levels in drinking water but an increased risk of cancer was reported in individuals exposed to high levels of arsenic in drinking water [30][31]. Several animal and in vitro models support the potential carcinogenic role of arsenic in colorectal cancer [32][33]. In contrast, research has highlighted the therapeutic potential of arsenical derivatives in the treatment of digestive cancers [34][35].

Cereal crops

We observed a lower risk of digestive cancers and mainly gastric cancers in cereal crop growers. The AGRICAN cohort has already proved a lower risk of gastric cancer among male farm-owners. The link with occupational exposure is not established and it is likely that the environment, especially dietary patterns, play a role in this risk variability. Several protective associations for gastric cancer have already been demonstrated such as the consumption of fruits, vegetables, and dietary flavonoids [36][37]. In contrast, high salt intake or red and processed meat consumption may be associated with an increased risk of gastric cancer [38][39].

Research perspectives

This exploratory preliminary study could be extended by identification of phytosanitary product use by farmers. The Graphical Terrain Register (Registre Parcellaire Graphique) lists all French agricultural areas and their main cultures. Furthermore, the Crop Exposure Matrices (MCE) of the Matphyto project of the National Public Health Agency (*Santé Publique France*) allows the evaluation of retrospective phytosanitary product exposure according to agricultural specialty [40]. We could evaluate the association between phytosanitary product use and digestive cancer by associating these tools with our study models [41]. We could thus improve vigilance for phytosanitary product use among farmers.

CONCLUSION

We found a positive association between winegrowing and the occurrence of LTD digestive cancers among French farm-owners, most often colorectal cancer. We question the use of arsenic as a pesticide until 2001 in this population to be one of the possible risk factors. We also highlighted a reduction in the risk of stomach cancer in cereal crop farmers.

The evaluation of occupational exposure is a major challenge to detect new risks and protect agricultural workers. Further studying the associations between phytosanitary product use and cancer risk could provide crucial answers to these topical issues.

Table 1
Main characteristics of study population made of French self-employed agricultural workers

Characteristics	N	%
Overall	899,212	-
Gender		
Men	631,560	70.2
Women	267,652	29.8
Mean age	50.2	
Men	48.7	-
Women	53.8	-
Year of first installation		
Median	1995	-
Minimum	1940	-
Maximum	2016	-
Health regimen		
Exclusive	625,139	69.5
Principal	42,077	4.7
Liberal	10,463	1.2
Employees	16,959	1.9
Others	204,574	22.8
Matrimonial status		
Single	350,942	39.0
Married	532,732	59.2
Widowed	37,940	4.2
Divorced	48,366	5.4

[1]Health regimen: exclusive: 100% incomes from agriculture. Principal: > 50% incomes from agriculture. Liberal: < 50% incomes from agriculture. Others: simultaneous assignment to another health regimen

Table 2
Agricultural activities as defined by MSA thesaurus for self-employed workers

Value	Name
1	Vegetables, flowers
2	Fruit trees
3	Plant nursery
4	Cereals
5	Winegrowing
6	Forestry
7	Other specialized crops
8	Dairy cattle farmers
9	Beef farmers
10	Mixed dairy and beef cattle
11	Sheep and goat
12	Pigs
13	Horses
14	Other big animals
15	Poultry and rabbits
16	Other small animals
17	Horse riding club
18	Shellfish farming
19	Non-specialized and/or mixed crops and breeding
20	Salt marshes
21	Wood working
22	Sawmills
23	Agricultural company
24	Landscaping
25	Local agricultural health insurance agents

Table 3
ICD-10 classification for malignant digestive neoplasms

Code	Name
ICD-15	Malignant neoplasm of esophagus
ICD-16	Malignant neoplasm of stomach
ICD-17	Malignant neoplasm of small intestine
ICD-18	Malignant neoplasm of colon
ICD-19	Malignant neoplasm of rectosigmoid junction
ICD-20	Malignant neoplasm of rectum
ICD-21	Malignant neoplasm of anus and anal canal
ICD-22	Malignant neoplasm of liver and intrahepatic bile ducts
ICD-23	Malignant neoplasm of gallbladder
ICD-24	Malignant neoplasm of other and unspecified parts of biliary tract
ICD-25	Malignant neoplasm of pancreas
ICD-26	Malignant neoplasm of other and ill-defined digestive organs

Table 4 - Number of cancer cases following occupational exposures

		Vegetables, flowers	Fruit trees	Plant nursery	Cereal	Winegrowing	Forestry	Other specialized crops	Dairy cattle farmers	Beef farmers	Mixed dairy and beef	Sheep and goat	Pigs	Horses	Other big animals	Poultry and rabbits	Other small animals	Horse riding club	Shellfish	mixed crops and breeding	Salt marshes	Wood working	Sawmill	Agricultural company	Landscaping	Insurance agents	Rural artisans
	TOTAL	**185**	**125**	**16**	**1211**	**644**	**12**	**26**	**661**	**596**	**153**	**177**	**74**	**44**	**6**	**88**	**40**	**28**	**7**	**517**	**2**	**38**	**1**	**62**	**116**	**6**	**1**
ICD-15 esophagus	**290**	18	6	1	75	30	1	3	38	35	12	9	6	7	2	10	0	1	0	39	1	2	1	2	8	0	0
ICD-16 stomach	**314**	8	9	2	63	39	1	3	59	51	12	19	5	2	0	9	4	2	1	35	0	5	0	2	10	1	0
ICD-17 small intestine	**83**	3	3	0	18	19	0	0	13	15	3	1	0	2	0	1	1	1	0	6	0	0	0	0	2	1	0
ICD-18 colon	**1935**	71	54	7	560	275	4	10	293	240	58	76	34	16	2	32	18	12	3	220	0	12	0	30	45	2	0
ICD-19 rectosigmoid junction	**220**	9	7	0	61	42	0	3	28	30	8	4	3	4	0	3	3	1	0	20	0	3	0	5	6	0	0
ICD-20 rectum	**824**	43	19	2	221	117	6	3	114	119	25	26	15	3	0	21	7	3	1	95	0	7	0	11	18	1	1
ICD-21 Anus and anal canal	**54**	2	1	0	14	8	0	0	4	4	4	2	1	0	0	0	0	2	0	9	0	1	0	1	2	1	0
ICD-22 liver and intrahepatic bile ducts	**260**	9	7	1	66	36	0	0	35	31	13	11	3	3	0	6	1	2	2	32	1	6	0	5	6	0	0
ICD-23 gallbladder	**24**	1	0	0	9	4	0	0	3	6	0	1	0	2	0	0	0	0	0	1	0	0	0	0	1	0	0
ICD-24 other and unspecified parts of biliary tract	**63**	1	1	0	13	8	0	0	11	14	3	2	1	1	0	1	0	1	0	6	0	0	0	1	3	0	0
ICD-25 Pancreas	**434**	19	16	3	108	63	0	4	60	47	14	26	5	4	2	5	6	3	0	53	0	2	0	4	14	0	0
ICD-26 other and ill-defined digestive organs	**20**	1	2	0	3	3	0	0	3	4	1	0	1	0	0	0	0	0	0	1	0	0	0	1	1	0	0

Table 5 - Corrected p-values (Benjamini-Hochberg test) for associations between occupational exposures and digestive cancers

	ICD-15	ICD-16	ICD-17	ICD-18-19-20	ICD-21	ICD-22	ICD-23-24	ICD-25	ICD-26	ICD-15-26
Vegetables, flowers	2.52E-01	6.53E-01	0.998	0.791555556	0.998	0.998	0.998	0.984	0.999	0.788666667
Fruit trees	9.95E-01	9.64E-01	0.998	0.791555556	0.998	0.998	0.998	0.972	0.999	0.1235
Plant nursery	9.95E-01	9.64E-01		0.462222222		0.998		0.984		0.537333333
Cereal	9.95E-01	**1.85E-02**	0.998	0.496	0.998	0.998	0.998	0.984	0.999	**0.02158**
Winegrowing	9.31E-01	9.64E-01	0.07748	**0.010972**	0.998	0.998	0.998	0.972	0.999	**0.0000455**
Forestry	0.9308	0.964		0.182866667						0.432545455
Other specialized crops	0.51025	0.65288889		0.889473684		0.972				0.432545455
Dairy cattle farmers	5.10E-01	5.10E-01	0.998	0.481	0.7605	0.998	0.998	0.972	0.999	0.2496
Beef farmers	9.31E-01	5.10E-01	0.875333333	0.902	0.998	0.998	0.13754	0.972	0.999	0.869916667
Mixed dairy and beef	9.95E-01	9.64E-01	0.998	0.4732	0.7605	0.998	0.998	0.984	0.999	0.87
Sheep and goat	9.31E-01	5.10E-01	0.875333333	0.434571429	0.998	0.998	0.998	0.972		0.432545455
Pigs	9.31E-01	9.64E-01		0.434571429	0.998	0.998	0.998	0.984	0.999	0.2496
Horses	2.52E-01	9.53E-01	0.875333333	0.23595		0.998	0.5668	0.984		0.865913043
Other big animals	0.25168		0.434571429		0.972					0.537333333
Poultry and rabbits	2.52E-01	7.25E-01	0.998	0.902		0.998	0.998	0.972		0.437666667
Other small animals		9.64E-01	0.998	0.902		0.998		0.984		0.87
Horse riding club	9.31E-01	9.64E-01	0.998	0.7176	0.4251	0.998	0.998	0.984		0.537333333
Shellfish		0.964		0.599857143		0.998				0.432545455
Mixed crops and breeding	9.95E-01	9.64E-01	0.875333333	0.182866667	0.998	0.998	0.998	0.984	0.999	0.068986667
Salt marshes	0.25168		0.2977							0.6682
Wood working	0.995	0.65288889		0.902	0.998	0.2977		0.984		0.537333333
Sawmill	0.51025									0.537333333
Agricultural company	0.768444444	0.65288889		0.462222222	0.998	0.998	0.998	0.974	0.999	0.587052632
Landscaping	9.95E-01	9.64E-01	0.998	0.902	0.998	0.998	0.998	0.984	0.999	0.865913043
Insurance agents		0.5564	0.10647	0.791555556	0.11596					0.537333333
Rural artisans				0.481						0.432545455

Table 6 - Non-corrected p-values for associations between occupational exposures and digestive cancers

	ICD-15	ICD-16	ICD-17	ICD-18-19-20	ICD-21	ICD-22	ICD-23-24	ICD-25	ICD-26	ICD-15-26
Vegetables, flowers	**4.84E-02**	0.192	0.839	0.548	0.873	0.63	0.428	0.688	0.834	0.637
Fruit trees	6.66E-01	0.764	0.469	0.529	0.755	0.998	0.438	0.132	**0.0479**	**0.019**
Plant nursery	6.63E-01	0.855		0.16		0.726		0.667		0.372
Cereal	8.28E-01	**0.00071**	0.577	0.248	0.924	0.518	0.575	0.809	0.317	**0.00166**
Winegrowing	4.95E-01	0.822	**0.00298**	**0.000422**	0.628	0.429	0.699	0.17	0.751	**0.00000175**
Forestry	0.46	0.59		**0.0211**						0.152
Other specialized crops	0.13	0.226		0.65		0.257		0.154		
Dairy cattle farmers	1.57E-01	0.0785	0.721	0.209	0.0971	0.54	0.904	0.241	0.875	**0.0492**
Beef farmers	5.37E-01	**0.0427**	0.169	0.706	0.262	0.664	**0.00529**	0.267	0.318	0.803
Mixed dairy and beef	7.57E-01	0.713	0.89	0.182	0.117	0.158	0.951	0.709	0.786	0.87
Sheep and goat	4.16E-01	0.0761	0.202	0.117	0.829	0.91	0.735	0.0914		0.178
Pigs	3.98E-01	0.701		0.0843	0.767	0.823	0.88	0.56	0.248	0.0576
Horses	**1.38E-02**	0.403	0.158	**0.0363**		0.869	**0.0436**	0.801		0.766
Other big animals	**0.048**		0.109		0.299		0.372			
Poultry and rabbits	**4.51E-02**	0.279	0.531	0.779		0.463	0.579	0.125		0.202
Other small animals		0.687	0.766	0.856		0.304		0.459		0.837
Horse riding club	5.35E-01	0.768	0.563	0.414	**0.0327**	0.938	0.6	0.84		0.279
Shellfish		0.889		0.323		0.184		0.123		
Mixed crops and breeding	8.67E-01	0.447	0.121	**0.0193**	0.421	0.757	0.18	0.638	0.301	**0.00796**
Salt marshes	**0.0249**		**0.0229**		0.514					
Wood working	0.701	0.187		0.771	0.362	**0.0155**		0.464		0.343
Sawmill	0.146		0.346							
Agricultural company	0.266	0.217		0.155	0.652	0.777	0.931	0.337	0.178	0.429
Landscaping	9.59E-01	0.939	0.967	0.887	0.617	0.392	0.145	0.623	0.494	0.759
Insurance agents		0.107	**0.00819**	0.523	**0.00446**		0.335			
Rural artisans		0.222		0.183						

Table 7 - Odds Ratio for associations between occupational exposures and digestive cancers

	ICD-15	ICD-16	ICD-17	ICD-18-19-20	ICD-21	ICD-22	ICD-23-24	ICD-25	ICD-26	ICD-15-26
Vegetables, flowers	1.63 [1-2.64]	0.63 [0.31-1.27]	0.89 [0.28-2.81]	1.06 [0.88-1.28]	0.89 [0.22-3.66]	0.85 [0.43-1.67]	0.57 [0.14-2.31]	1.1 [0.69-1.75]	1.24 [0.17-9.29]	1.04 [0.87-1.24]
Fruit trees	0.83 [0.37-1.89]	1.11 [0.57-2.17]	1.53 [0.48-4.87]	1.08 [0.85-1.37]	0.73 [0.1-5.29]	1 [0.46-2.17]	0.46 [0.06-3.29]	1.48 [0.89-2.48]	4.42 [1.01-19.31]	1.28 [1.04-1.58]
Plant nursery	0.64 [0.09-4.66]	1.14 [0.28-4.65]		0.61 [0.31-1.21]		0.7 [0.09-5.19]		1.29 [0.4-4.11]		0.78 [0.45-1.35]
Cereal	0.97 [0.73-1.28]	**0.6 [0.45-0.81]**	0.86 [0.5-1.48]	0.95 [0.87-1.04]	1.03 [0.54-1.95]	0.9 [0.67-1.23]	0.86 [0.52-1.44]	0.97 [0.77-1.23]	0.52 [0.14-1.87]	**0.87 [0.8-0.95]**
Winegrowing	0.87 [0.6-1.28]	1.04 [0.74-1.46]	2.19 [1.31-3.67]	**1.21 [1.09-1.35]**	1.21 [0.57-2.57]	1.16 [0.81-1.66]	1.13 [0.61-2.08]	1.21 [0.92-1.59]	1.22 [0.36-4.2]	**1.28 [1.16-1.42]**
Forestry	2.1 [0.29-15.02]	1.72 [0.24-12.26]		2.09 [1.12-3.9]			1.62 [0.84-3.12]			
Other specialized crops	2.42 [0.77-7.57]	2.02 [0.65-6.33]		1.13 [0.67-1.9]			1.8 [0.65-5.01]		1.37 [0.89-2.11]	
Dairy cattle farmers	0.78 [0.55-1.1]	1.3 [0.97-1.73]	0.9 [0.49-1.63]	0.93 [0.84-1.04]	0.42 [0.15-1.17]	0.89 [0.62-1.29]	1.04 [0.58-1.85]	0.85 [0.64-1.12]	0.91 [0.26-3.15]	0.91 [0.82-1]
Beef farmers	0.89 [0.62-1.28]	1.37 [1.01-1.85]	1.48 [0.85-2.6]	1.02 [0.91-1.14]	0.56 [0.2-1.55]	0.92 [0.62-1.35]	2.05 [1.24-3.38]	0.84 [0.62-1.14]	1.75 [0.58-5.27]	1.01 [0.91-1.13]
Mixed dairy and beef	1.1 [0.61-1.97]	1.12 [0.62-1.99]	0.92 [0.29-2.93]	0.86 [0.69-1.07]	2.27 [0.81-6.34]	1.51 [0.85-2.66]	0.96 [0.3-3.06]	0.9 [0.53-1.55]	1.32 [0.18-9.98]	0.98 [0.82-1.19]
Sheep and goat	0.76 [0.39-1.48]	1.53 [0.96-2.44]	0.28 [0.04-1.99]	0.85 [0.69-1.04]	0.86 [0.21-3.52]	1.04 [0.56-1.92]	0.82 [0.26-2.6]	1.42 [0.95-2.13]		0.88 [0.73-1.06]
Pigs	1.42 [0.63-3.21]	1.19 [0.49-2.89]		1.29 [0.97-1.72]	1.35 [0.19-9.84]	0.88 [0.28-2.77]	0.86 [0.12-6.2]	0.77 [0.31-1.87]	3.32 [0.43-25.37]	1.29 [0.99-1.67]
Horses	2.64 [1.22-5.73]	0.55 [0.14-2.23]	2.77 [0.67-11.38]	0.62 [0.39-0.97]		1.1 [0.35-3.51]	3.32 [1.04-10.61]	0.88 [0.33-2.38]		0.95 [0.66-1.35]
Other big animals	4.24 [1.01-17.73]		0.29 [0.07-1.31]				2.16 [0.5-9.3]		0.6 [0.19-1.86]	
Poultry and rabbits	1.92 [1.01-3.63]	1.45 [0.74-2.82]	0.53 [0.07-3.84]	0.96 [0.73-1.26]		1.36 [0.6-3.08]	0.57 [0.08-4.13]	0.5 [0.2-1.21]		0.84 [0.65-1.1]
Other small animals		1.23 [0.45-3.33]	1.35 [0.19-9.84]	1.04 [0.7-1.53]		0.35 [0.05-2.57]		1.37 [0.6-3.13]		0.96 [0.66-1.4]
Horse riding club	0.54 [0.07-3.84]	0.81 [0.2-3.28]	1.79 [0.25-13.04]	0.81 [0.49-1.34]	4.78 [1.14-20.11]	1.06 [0.26-4.34]	1.7 [0.23-12.33]	0.89 [0.28-2.8]		0.77 [0.49-1.23]
Shellfish		1.15 [0.16-8.33]		0.6 [0.22-1.65]		2.71 [0.62-11.79]		0.41 [0.13-1.27]		
Mixed crops and breeding	1.03 [0.73-1.45]	0.87 [0.61-1.24]	0.52 [0.23-1.19]	0.87 [0.77-0.98]	1.34 [0.65-2.75]	0.94 [0.64-1.38]	0.59 [0.27-1.28]	0.93 [0.7-1.25]	0.35 [0.05-2.59]	0.86 [0.77-0.96]
Salt marshes	9.59 [1.33-69.17]					9.91 [1.37-71.53]		1.59 [0.4-6.41]		
Wood working	0.76 [0.19-3.08]	1.82 [0.75-4.44]		1.07 [0.69-1.64]	2.54 [0.34-18.81]	2.78 [1.21-6.35]		0.59 [0.15-2.4]		1.19 [0.83-1.71]
Sawmill	4.31 [0.6-30.88]						0.39 [0.05-2.77]			
Agricultural company	0.45 [0.11-1.83]	0.41 [0.1-1.68]		1.26 [0.92-1.73]	1.58 [0.22-11.55]	1.14 [0.45-2.88]	0.92 [0.13-6.64]	0.61 [0.22-1.68]	4.07 [0.53-31.34]	1.13 [0.83-1.54]
Landscaping	0.98 [0.47-2.03]	1.03 [0.53-1.97]	0.97 [0.23-4.06]	1.02 [0.79-1.31]	1.45 [0.34-6.31]	0.69 [0.3-1.61]	2.17 [0.77-6.17]	1.15 [0.66-2.01]	2.07 [0.26-16.6]	1.04 [0.83-1.29]
Insurance agents		5.06 [0.7-36.33]	14.75 [2.01-108.38]	1.45 [0.46-4.52]	18.37 [2.47-136.58]		1.62 [0.61-4.36]			
Rural artisans		0.29 [0.04-2.1]					0.26 [0.04-1.88]			

REFERENCES

[1] Defossez G, Le Guyader‐Peyrou S, Uhry Z, Grosclaude P, Colonna M, Dantony E, et al. Estimations nationales de l'incidence et de la mortalité par cancer en France métropolitaine entre 1990 et 2018. Synthèse. Saint‐Maurice : Santé publique France, 2019. 20 p.

[2] Bray F, Ferlay J, Soerjomataram I, Siegel RL, Torre LA, Jemal A. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J Clin. 2018;68(6):394-424.

[3] Jéhannin-Ligier K, Dantony E, Bossard N, Molinié F, Defossez G, Daubisse-Marliac L, Remontet L, Uhry Z. Projection de l'incidence et de la mortalité par cancer en France métropolitaine en 2017. Rapport technique. Saint-Maurice : Santé publique France, 2017. 80 p.

[4] Santé publique France. Taux de participation au programme de dépistage organisé du cancer colorectal 2017-2018. 2019.
URL :https://www.santepubliquefrance.fr/maladies-et-traumatismes/cancers/cancer-du-colon-rectum/articles/taux-de-participation-au-programme-de-depistage-organise-du-cancer-colorectal-2017-2018

[5] Ilic M, Ilic I. Epidemiology of pancreatic cancer. World J Gastroenterol. 28 nov 2016;22(44):9694-705.

[6] Bouvier A-M, Uhry Z, Jooste V, Drouillard A, Remontet L, Launoy G, et al. Focus on an unusual rise in pancreatic cancer incidence in France. Int J Epidemiol. 01 2017;46(6):1764-72.

[7] World Health Organization, International Agency for Research on Cancer (2007). Attributable Causes of Cancer in France in the Year 2000. IARC Working Group Reports, Vol. 3. Lyon, France: International Agency for Research on Cancer.

[8] Marant Micallef C, Shield K, Vignat J, Hill C, Rogel A, Menvielle G, et al. Nombre et fractions de cancers attribuables au mode de vie et à l'environnement en France métropolitaine en 2015 : résultats principaux. Bulletin Epidemiologique Hebdomadaire. 26 juin 2018;

[9] Marant Micallef C, Shield KD, Baldi I, Charbotel B, Fervers B, Gilg Soit Ilg A, et al. Occupational exposures and cancer: a review of agents and relative risk estimates. Occup Environ Med. août 2018;75(8):604-14.

[10] Neugut AI, Terry MB, Hocking G, Mosca L, Garbowski GC, Forde KA, et al. Leisure and occupational physical activity and risk of colorectal adenomatous polyps. Int J Cancer. 11 déc 1996;68(6):744-8.

[11] Acquavella J, Olsen G, Cole P, Ireland B, Kaneene J, Schuman S, et al. Cancer among farmers: a meta-analysis. Ann Epidemiol. janv 1998;8(1):64-74. PMID : 9465996.

[12] Blair A, Zahm SH, Pearce NE, Heineman EF, Fraumeni JF. Clues to cancer etiology from studies of farmers. Scand J Work Environ Health. août 1992;18(4):209-15. PMID : 1411362.

[13] Pukkala E, Martinsen JI, Lynge E, Gunnarsdottir HK, Sparén P, Tryggvadottir L, et al. Occupation and cancer - follow-up of 15 million people in five Nordic countries. Acta Oncol. 2009;48(5):646-790.

[14] Alavanja MCR, Sandler DP, Lynch CF, Knott C, Lubin JH, Tarone R, et al. Cancer incidence in the agricultural health study. Scand J Work Environ Health. 2005;31 Suppl 1:39-45; discussion 5-7. PMID : 16190148.

[15] Lerro CC, Koutros S, Andreotti G, Sandler DP, Lynch CF, Louis LM, et al. Cancer incidence in the Agricultural Health Study after 20 years of follow-up. Cancer Causes Control. avr 2019;30(4):311-22.

[16] Lemarchand C, Tual S, Levêque-Morlais N, Perrier S, Belot A, Velten M, et al. Cancer incidence in the AGRICAN cohort study (2005-2011). Cancer Epidemiol. 2017;49:175-85.

[17] Darcey E, Carey RN, Reid A, Driscoll T, Glass DC, Benke GP, et al. Prevalence of exposure to occupational carcinogens among farmers. Rural Remote Health. 2018;18(3):4348.

[18] Maugard C, Rieutort DB, Ozenfant D, François O, Bonneterre V. Big-data and occupational health surveillance: Screening of occupational determinants of health among French agricultural workers, through data mining of medico-administrative databases. Revue d'Épidémiologie et de Santé Publique. 1 juill 2018;66:S262-3.

[19] Hisabe T, Hirai F, Matsui T. Development and progression of colorectal cancer based on follow-up analysis. Dig Endosc. avr 2014;26 Suppl 2:73-7.

[20] Muto T, Bussey HJ, Morson BC. The evolution of cancer of the colon and rectum. Cancer. déc 1975;36(6):2251-70.

[21] O'Connell B, Hafiz N, Crockett S. The Serrated Polyp Pathway: Is It Time to Alter Surveillance Guidelines? Curr Gastroenterol Rep. 29 août 2017;19(10):52.

[22] Garnett MJ, Marais R. Guilty as charged: B-RAF is a human oncogene. Cancer Cell. oct 2004;6(4):313-9.

[23] Brosens LAA, Hackeng WM, Offerhaus GJ, Hruban RH, Wood LD. Pancreatic adenocarcinoma pathology: changing « landscape ». J Gastrointest Oncol. août 2015;6(4):358-74.

[24] Lauzeille D, Marchand J L, Ferrand M. Consommation de tabac par catégorie socioprofessionnelle et secteur d'activité – Outil méthodologique pour l'épidémiologie. Saint-Maurice (Fra) : Institut de veille sanitaire, juillet 2019, 208 p.

[25] Chatignoux É, Remontet L, Colonna M, Grosclaude P, Decool E, Uhry Z. Estimations régionales et départementales d'incidence et de mortalité par cancers en France, 2007-2016. Évaluation de l'utilisation des données médico-administratives pour estimer l'incidence départementale : comparaison de l'incidence observée et prédite dans les registres sur la période 2007-2014. Saint-Maurice : Santé publique France, 2019. 106 p.

[26] Seidler A, Hammer GP, Husmann G, König J, Krtschil A, Schmidtmann I, et al. Cancer risk among residents of Rhineland-Palatinate winegrowing communities: a cancer-registry based ecological study. J Occup Med Toxicol. 6 juin 2008;3:12.

[27] INSERM. Pesticides. Effets sur la santé. Collection expertise collective, Inserm, Paris, 2013

[28] Spinosi J, Févotte J, Vial G. éléments techniques sur l'exposition professionnelle aux pesticides arsenicaux. Matrice cultures - expositions aux pesticides arsenicaux. Saint-Maurice (Fra) : Institut de veille sanitaire, avril 2009, 19 p.

[29] International Agency on Research on Cancer, Arsenic and arsenic compounds, IARC Monographs on the Evaluation of Carcinogenic Risks to Humans Volume 100C, 2012. 527 p.
http://publications.iarc.fr/120

[30] Yang C-Y, Chang C-C, Ho S-C, Chiu H-F. Is colon cancer mortality related to arsenic exposure? J Toxicol Environ Health Part A. 2008;71(8):533-8.

[31] Tsai SM, Wang TN, Ko YC. Mortality for certain diseases in areas with high levels of arsenic in drinking water. Arch Environ Health. juin 1999;54(3):186-93.

[32] Cholpraipimolrat W, Suriyo T, Rangkadilok N, Nookabkaew S, Satayavivad J. Hijiki and sodium arsenite stimulate growth of human colorectal adenocarcinoma cells through ERK1/2 activation. Food Chem Toxicol. déc 2017;110:33-41.

[33] Eyvani H, Moghaddaskho F, Kabuli M, Zekri A, Momeny M, Tavakkoly-Bazzaz J, et al. Arsenic trioxide induces cell cycle arrest and alters DNA methylation patterns of cell cycle regulatory genes in colorectal cancer cells. Life Sci. 15 déc 2016;167:67-77.

[34] Ma Z-B, Xu H-Y, Jiang M, Yang Y-L, Liu L-X, Li Y-H. Arsenic trioxide induces apoptosis of human gastrointestinal cancer cells. World J Gastroenterol. 14 mai 2014;20(18):5505-10.

[35] Lv X-H, Wang C-H, Xie Y. Arsenic trioxide combined with transarterial chemoembolization for primary liver cancer: A meta-analysis. J Gastroenterol Hepatol. sept 2017;32(9):1540-7.

[36] Woo HD, Lee J, Choi IJ, Kim CG, Lee JY, Kwon O, et al. Dietary flavonoids and gastric cancer risk in a Korean population. Nutrients. 10 nov 2014;6(11):4961-73

[37] Gonzalez CA, Lujan-Barroso L, Bueno-de-Mesquita HB, Jenab M, Duell EJ, Agudo A, et al. Fruit and vegetable intake and the risk of gastric adenocarcinoma: a reanalysis of the European Prospective Investigation into Cancer and Nutrition (EPIC-EURGAST) study after a longer follow-up. Int J Cancer. 15 déc 2012;131(12):2910-9.

[38] Abnet CC, Corley DA, Freedman ND, Kamangar F. Diet and upper gastrointestinal malignancies. Gastroenterology. mai 2015;148(6):1234-1243.e4.

[39] González CA, Jakszyn P, Pera G, Agudo A, Bingham S, Palli D, et al. Meat intake and risk of stomach and esophageal adenocarcinoma within the European Prospective Investigation Into Cancer and Nutrition (EPIC). J Natl Cancer Inst. 1 mars 2006;98(5):345-54.

[40] Spinosi J, Févotte J. Le programme Matphyto–Matrices cultures-expositions aux produits phytosanitaires. Saint-Maurice (Fra): Institut de veille sanitaire, juin 2008, 16 p.

[41] Achard P, Maugard C, Cancé C, Spinosi J, Ozenfant D, Maître A, et al. Medico-administrative data combined with agricultural practices data to retrospectively estimate pesticide use by agricultural workers. J Expo Sci Environ Epidemiol. 2019.

Journal of Exposure Science & Environmental Epidemiology
https://doi.org/10.1038/s41370-019-0166-x

ARTICLE

Medico-administrative data combined with agricultural practices data to retrospectively estimate pesticide use by agricultural workers

Pauline Achard[1] · Charlotte Maugard[1] · Christophe Cancé[2] · Johan Spinosi[3,4] · Damien Ozenfant[5] · Anne Maître[1] · Delphine Bosson-Rieutort[1] · Vincent Bonneterre[1]

Received: 9 January 2019 / Revised: 10 June 2019 / Accepted: 17 June 2019

Abstract

This work is part of a global project aiming to use medico-administrative big data from the whole French agricultural population (~3 millions), collected through their mandatory health insurance system (Mutualité Sociale Agricole), to highlight associations between chronic diseases and agricultural activities. At the request of the French Agency for Food, Environmental and Occupational Health & Safety (ANSES), our objective was to estimate which pesticides were probably used by each agricultural worker, in order to include this information in our analyses and search for association with diseases. We selected five databases to achieve this objective: the Graphical Land Parcel Registration (RPG), the French Agricultural Census, "Cultivation Practice" surveys from the Agriculture ministry, the MATPHYTO crop-exposure matrix and the Compilation of Phytosanitary Indexes from the French Public Health Agency. A geographical grid was designed to use geographical location while maintaining worker anonymity, dividing France into square tracts of variable surface each containing a minimum of 1500 agricultural workers. We developed an automated algorithm to predict each individual potential exposure by crossing her/his occupational activity, the geographical grid and the RPG to deduce cultivation practices and use it as a gateway to estimate pesticides use. This approach allowed drawing, from administrative data, a list of substances potentially used by each agricultural worker throughout France. Results of the algorithm are illustrated at collective level (descriptive statistics for the whole population), as well as at individual level (some workers taken as examples). The generalization of this method in other national contexts is discussed. By linking this information with the health insurance databases, this approach could contribute to the agricultural workers health surveillance.

Keywords Occupational exposure · big-data · health insurance data · agricultural practices · France

Introduction

Pesticides, especially plant protection products, are used to control or to fight harmful organisms. They include a large number of molecules, whose chemical properties and toxic effects are very diverse. Numerous studies have shown that chronic exposure to plant protection products may be associated with an excess risk of chronic diseases such as certain cancers or Parkinson's disease [1–4]. Cohort studies conducted among agricultural workers seem to confirm this, particularly the Agricultural Health Study [5–8] conducted in the United States or more recently the AGRICAN (Agriculture and Cancer) cohort in France [9, 10]. For simplicity, hereinafter, plant protection products will be referred to as the broader term "pesticides".

These authors contributed equally: Delphine Bosson-Rieutort, Vincent Bonneterre

✉ Vincent Bonneterre
VBonneterre@chu-grenoble.fr

[1] Univ. Grenoble Alpes, CNRS, CHU Grenoble Alpes, Grenoble INP, TIMC-IMAG (EPSP Team), F-38000 Grenoble, France

[2] Univ. Grenoble Alpes, (UGA)/UMS GRICAD, F-38000 Grenoble, France

[3] Santé publique France, 12 Rue du Val d'Osne, F-94410 Saint-Maurice, France

[4] University of Lyon, Université Claude Bernard Lyon 1, Ifsttar, UMR T_9405, F-69373 Lyon, France

[5] Caisse centrale Mutualité Sociale Agricole (CCMSA), 19 rue de Paris, F-93000 Bobigny, France

Published online: 04 September 2019

In France, all agricultural workers are covered by a compulsory healthcare insurance system, the "Mutualité Sociale Agricole" (MSA). The MSA routinely collects large amounts of medico-administrative data on this entire population of approximately 3 million individuals [11], for health insurance purposes only.

This work was part of a project on surveillance concerning occupational health risks in agricultural workers, based on exhaustive health data from the French agricultural population. Nevertheless, if occupational activity of workers is coded in MSA administrative databases, no information concerning a possible use of pesticides is indicated. In addition, the extremely large number of individuals included does not allow the direct application of commonly used strategies to retrospectively estimate pesticide exposure such as questionnaires or expert judgment. Thus to go further to estimate pesticide use, it seemed necessary to cross administrative data from the MSA with already existing external sources of information about farming practices, such as job-exposure matrices, crop-exposures matrices or census. This would make it possible to obtain more detailed information on occupational activities in order to deduce the practices of pesticide use. As these data sources are structured to inform all practices concerning a specific crop, and the information on the crops is only partial within the MSA database, two families of data sources are needed. The first data sources are those which allow, thanks to the geographical location of the individual, to translate and specify the occupational activity in terms of the crops cultivated. Then this information might serve as a gateway to another category of data sources able to provide, from the crop, a list of potentially used pesticides.

The objective of this work was to identify useful potential external sources of expert data and develop a strategy of combined analysis of MSA's medical-administrative databases and those external sources, in order to deduce in an automated way, from individual information on occupational activity and the geographical location, the crops cultivated and the chemical substances potentially used by each individual. This new strategy will then be used to refine the modeling of the relationship between the use of phytosanitary substances and the long-term diseases.

Material and methods

The massive data mining project in which this study is integrated was approved by the French Data Protection Authority CNIL (MMS/SBM/AE171001).

MSA population and data

The Mutualité Sociale Agricole (MSA) is the healthcare insurance system covering the whole French agricultural population, i.e. employees or self-employed workers (including farm managers, landowning farmers, contractors etc.), agricultural retirees, as well as "affiliates of beneficiaries" (family members and other relatives). Extensive administrative data and medico-administrative data are routinely collected by the MSA. Medico-administrative data include declared long-term diseases (LTDs), occupational accidents and diseases and also healthcare related expenses (drugs, medical consultations, complementary examinations etc.). The administrative data contain an identification variable specific to each individual (an anonymized number based on social security numbers), demographic information (age, sex), workplace locations, as well as the occupational activities performed by each individual during the study period, including the start and end dates of the employment contract for employees, or the year of installation of the exploitation or enterprise for the self-employed (Table 1). The nature of the occupational activity is indicated by two "activity codes": a NAF code (French Activities Nomenclature) and an MSA internal "risk" code, used to calculate the amount of the annual contributions, based both on the most risky occupational activity undertaken and the activity for which working time is the longest.

The administrative data regarding active agricultural workers are contained into two databases that are annually updated. The first one concerns agricultural employees and

Table 1 Structure and content of the administrative data routinely collected by the MSA

	Employee database	Self-employed database
Annual input	Contracts	Entreprises/ individuals
Number of observations	22,864,682	6,699,447
Number of distinct individuals	5,238,619	1,042,951
Available variable		
Identification number	✓	✓
Age	✓	✓
Gender	✓	✓
Location	✓	✓
Contract start date	✓	
Contract end date	✓	
Year of installation		✓
NAF code	✓	✓
Risk code (internal)	✓	✓

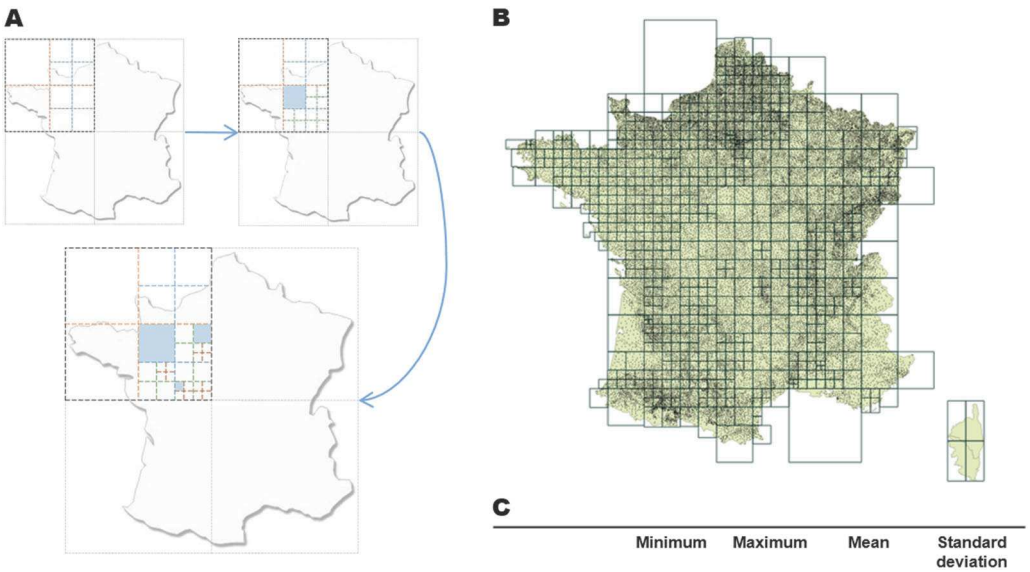
MSA Mutualité Sociale Agricole, *NAF* French Activities Nomenclature

the second is for self-employed agricultural workers. These two databases, although containing similar information, have different structures: within the employee database each individual is registered as many times as he has short-term contracts during the year, while in the self-employed database, each individual is registered annually, for as long as he owns at least one agricultural exploitation or enterprise. In this respect, over the period of the study, the cumulative databases can include, for each individual, several lines corresponding to either different activities or the continuity of these activities. From 2006 to 2015, the population covered and registered in these administrative databases accounted for 6,281,571 distinct agricultural workers, of which over 80% were employees.

Segmentation of the French territory

Considering the variability in the types of cultivation across the country, the geographical location is an essential parameter to link an activity sector to a defined range of crops, and thus to focus on probable pesticide uses. However, given both the sensitivity of health insurance data, the size of the population concerned and the significant risk of re-identification of individuals, it was decided that the geographical location provided by the MSA would be limited to the administrative subdivision called "départements" (96 relatively large geographical units about the size of a county), although the exact address of the worker is registered in the original databases. It therefore seemed necessary to include a spatial variable whose size had to be both sufficiently "wide" in order to minimize the risk of re-identification, but also precise enough to differentiate local cultivation practices.

For this purpose, a strategy of dividing the territory into a geographic grid via an iterative process has been developed. At first, the territory was divided into four square geographical units of equal area, then each of these units was again divided orthogonally into four new squares. The process stops when one or more of the four "daughter units" (subunits) obtained no longer meets one of the criteria set below; and in this case only the "mother unit" (original undivided unit) is preserved (Fig. 1a). As part of this study, one of the criteria was to impose a minimum threshold of 1,500 individuals (active agricultural workers) per unit as well as a minimum of 2 towns. This minimum threshold was calculated both based on the prevalence of LTDs and on the recommendation of the Secure Data Access Center (https://www.casd.eu/en/le-centre-dacces-securise-aux-donnees-casd/le-casd/) regarding the outcome presentation of studies presenting a risk of re-identification of individuals. We then calculated the threshold in order to minimize the risk of finding less than five individuals presenting the



C

	Minimum	Maximum	Mean	Standard deviation
Town by unit	2	479	50,88	47,15
Agricultural workers by unit	1 528	38 903	4 302,99	3 285,51

Fig. 1 Segmentation of the French territory into geographical units. **a** Procedure followed by the algorithm to subdivide the territory: each unit is divided into 4 square subunits of equal size, except when one of the subunits produced no longer meet the imposed criteria. In this case, the algorithm stops and only the preceding unit is retained. **b** Geographical subdivision of French territory obtained after application of the algorithm with a minimum threshold of 1500 agricultural workers and 2 towns per unit. **c** Unit characteristics

same LTD within the same unit. For example, for Parkinson's disease with a LTD prevalence of 430 cases per 100,000 individuals among agricultural workers [12], a minimum of 1,163 individuals within a unit would be required to identify at least five workers having the disease. Finally, the minimum of two towns per unit was set to take into account of extreme cases, such as an individual declaring a rare disease and/or a rare occupational activity in the area. Likewise, this was meant to limit the re-identification risk and to avoid associating any recognizable individual within a specific town, even if it has more than 1,500 agricultural workers. The result was a division of France into 720 units of variable area, each containing at least two municipalities and at least 1,500 agricultural workers (Fig. 1b). Through a correspondence table, listing all towns in each unit, the MSA was able to assign to each individual the identification number of the corresponding unit, instead of identifying the town.

External data sources on agricultural practices

Five external data sources on crop distribution and pesticide use were identified. Firstly, those allowing to specify the type of crop by crossing occupational activity with geographical location (unit as previously described); and secondly, those allowing to list all the substances used in the identified crops. For each of these sources, its conditions of accessibility, its completeness, as well as its advantages and limitations concerning both the content and the possibilities of integrating it into an automated algorithm were investigated, and are summarized in Table 2.

The sources allowing to specify the type of cultivation by crossing occupational activity with geographical location, are the Graphical Land Parcel Registration (RPG) and the Agricultural Census (AC). In line with the EU Common Agricultural Policy, the RPG is a geographical information system (GIS), listing nearly all parcels across the territory,

Table 2 Comparison of external sources of data for the estimation of the use of phytosanitary substances in France

	Geographical distribution of crop		Phytosanitary products used on crop		
Data source	Graphical Land Parcel Registration ("RPG")	Agricultural census (AC)	Crop Exposure Matrix "Matphyto"	Compilation of Phytosanitary Indexes ("CIPA")	Cultivation Practice surveys (CP)
Holder	Administrative body in liaison with the EU Common Agricultural Policy	Agriculture Ministry	French Public Health Agency (Santé Publique France)	Agriculture Ministry	
Database creation	2007	1970	2008	2008	1994
Update frequency	1 year	10 year	n/a	nk	5 year
Availability	Open-source, online	Virtual environment, CASD	Agreement	Online open-source, on request	virtual environment, CASD
Sources	Agricultural worker statements	Questionnaire	Diversified	Agricultural Technical Coordination Association (ACTA)	Questionnaire
Information	Graphic representation of the parcels (GIS)	Inventory of farms/ exploitations and their production	Probability, frequency and intensity of use of substances by crop (in France)	Marketed substances	Substances used in France, by crop
Available historic since	2002	1970	1960	1961	1994
Geographical accuracy	Parcel	Town	Variable according to the crop	Whole territory	Town
Number of crop categories	28 (open data)	258	n/a	n/a	n/a
Nature of listed substances	n/a	n/a	Frequently used toxic substances	Marketed substances registered in the ACTA index	All substances used
Number of subtances	n/a	n/a	~100 per crop	943	Variable

CASD Secure Data Access Center, *nk* not known, *n/a* not applicable

the crop grown on each of them, and the agricultural worker cultivating them. According to the RPG data in 2012, around 6 million agricultural parcels ($n = 5{,}941{,}743$) were declared in France, referring to 27.4 million hectares divided into 26 categories of cultivation type. The RPG is updated each year. Part of the information collected is available online on the French government's open- access data site (*data.gouv.fr*). The AC is a comprehensive survey of French agricultural production, targeting all farms in the territory, requested by the Agriculture Ministry [13]. An AC is conducted every ten years: the first computerized AC, dated 1970, was followed by four others in 1979, 1988, 2000 and 2010. The data collected during these surveys cover seven major themes including information on crops and areas cultivated as well as livestock. In the agricultural census of 2010, about half a million exploitations ($n = 518{,}925$) were counted and 258 distinct crops identified. However, given the detail and data sensitivity, access to the entirety of the data is restricted and only available within a dedicated virtual environment that does not allow one to perform analyzes crossing massive data from other sources.

The data sources allowing one to list all the substances used on the identified crops are the French crop-exposure matrices MATPHYTO, the Compilation of Phytosanitary Indexes (CIPA), and cultivation practices surveys conducted by the Agriculture ministry (CP). The crop-exposure matrices MATPHYTO have been developed by the French Public Health agency (Santé Publique France) [14]. To date, seven matrices have been developed, six of which are crop-specific (maize, cereals, potatoes, vines, bananas and sugar cane), and one specific to arsenic-based pesticide use. Tables by crop give information on the geographical area, the period, the phytosanitary substances commonly used and quantitative information concerning their use, each year since 1960. The phytosanitary index of ACTA (French network of agricultural technical institutes), which catalogs all phytosanitary substances marketed in France and their approved uses (including for crops), has been published each year since 1961. Cumulative data is freely available directly online in the Compilation of Phytosanitary Indexes (CIPA) [15, 16]. From 1961 to 2014, 943 phytosanitary substances were listed for 58 possible uses. Finally, "cultivation practice" (CP) surveys (14), collected about every ten years, give information on agricultural practices for "field crops" (cereals, oilseeds, pulses, potatoes, and industrial beet), viticulture, market gardening and vegetable growing, and arboriculture. For each of the selected parcels, all the interventions carried out are listed. The collected data can be classified into nine categories, including fertilization, growth regulator use, weed-killer use and use of other phytosanitary substances. For each use of a chemical substance, the dates of treatment, the surface treated and the quantity used are recorded. These surveys, which are complex to carry out, are supplemented by "phytosanitary" surveys focusing only on the use of phytosanitary substances. To date, nine "cultivation practices" and "phytosanitary" surveys have been carried out. These data are accessible under the same conditions as previously presented for the Agricultural Census.

Strategy developed to estimate pesticide use for each worker

The MSA administrative data contain three variables indirectly related to the use of phytosanitary substances: occupational activity (including the sector of activity, described by a NAF code), the period of activity and the geographical location. The deduction of the use of pesticides is thus carried out for each combination NAF code – year – geographical unit.

Figure 2 illustrates the strategy we developed for automatically determining, from MSA data and external data sources if they were all available, a list of substances possibly used for each NAF code-year-geographical unit combination. All individuals with the same combination are considered to make similar use of pesticides. The NAF code firstly makes it possible to draw up a related list of agricultural productions (Step 1A). Crossing the RPG with the national territorial grid allows one to obtain an inventory of the cultivated parcels that are declared therein and thus to determine the number of parcels and the surface of each type of crop (Step 1B). In the same way, linking the Agricultural Census with the unit of the place of exploitation, could allow one to draw up an inventory of present exploitations. The exhaustive information available makes it possible to calculate the number of farms producing each crop as well as the associated areas. From the list of crops established using the NAF code, only crops inventoried both by crossing the RPG and the AC with the geographic grid were included (Step 2).

The crop represents a gateway variable between the occupational activity and a possible exposure to phytosanitary substances (Step 3). In particular, it allows the use of crop-exposure matrices and pragmatic crop surveys (Step 4A), which provide information on the most frequently used substances by crop, while taking into account the variations in their use related to the period and the geographical location. Firstly, these tools allow one to list for each NAF-year-geographic unit combination, and considering the previously established list of crops, the most commonly used substances, and therefore those agricultural workers are most likely to have been exposed. The CIPA database provides only qualitative information on the substances marketed per period, without specifying their quantity, but can be used as complementary information when the crop of

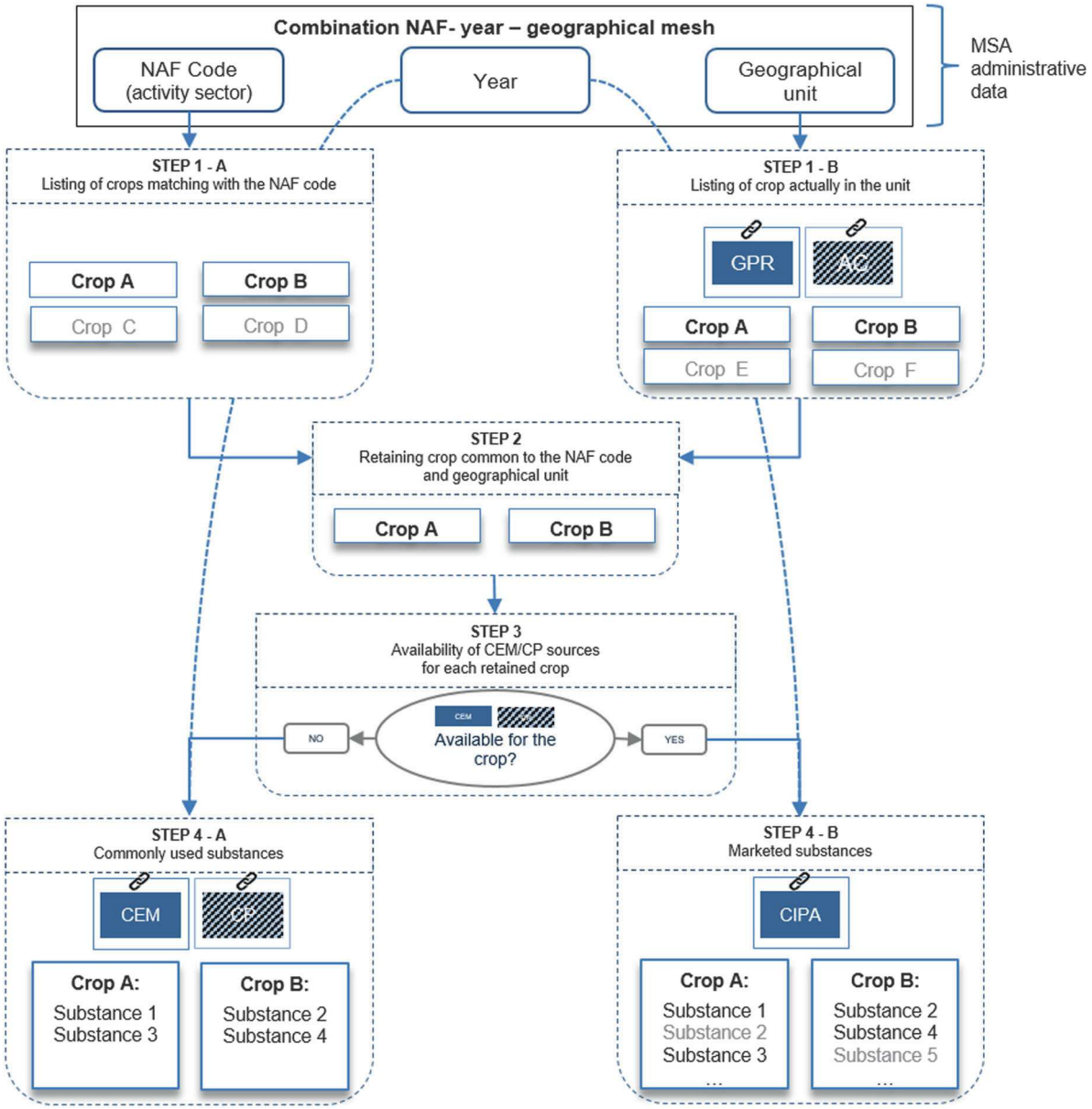


Fig. 2 Illustration of application of the automated algorithm using different data sources to deduce the phytosanitary substances used by French agricultural workers. GPR Graphical Parcel Register, AC Agricultural Census, CEM Crop-exposure matrices, CP "Cultivation Practices" and "Phytosanitary" surveys, CIPA Compilation of ACTA Phytosanitary Indexes

interest has not been the subject of a crop-exposure matrix or a cultivation practices survey (Step 4B). The expected result is, for each combination of NAF-year-geographic unit, a list of potentially used substances, which depends on the crops cultivated, years of activity and geographical locations. Figure 3 is an illustrative example of the process presented in Fig. 2, for a specific unit and a specific NAF code (Fig. 3a). When matching RPG data within the geographical grid (Fig. 3b), it appears that four kinds of crops were particularly important in terms of the number of parcels and surface covered: permanent grassland (42.0% of the parcels in the grid), nuts (32.2%), maize (8.8%), and temporary grassland (7.6%) (Fig. 3c). The analysis of the AC data shows that three major crop categories are more abundant than the others: grasslands making up 72% of agricultural exploitations in the geographic grid and representing more than 60% of the agricultural surface, nuts (66.8% of agricultural exploitations number and 25.6% of agricultural surface) and cereal crops (37.2% of agricultural exploitations, 8.3% of agricultural surface) (Fig. 3d). The RPG and AC data were similar with close percentages due to different groupings (e.g. "all grassland" versus

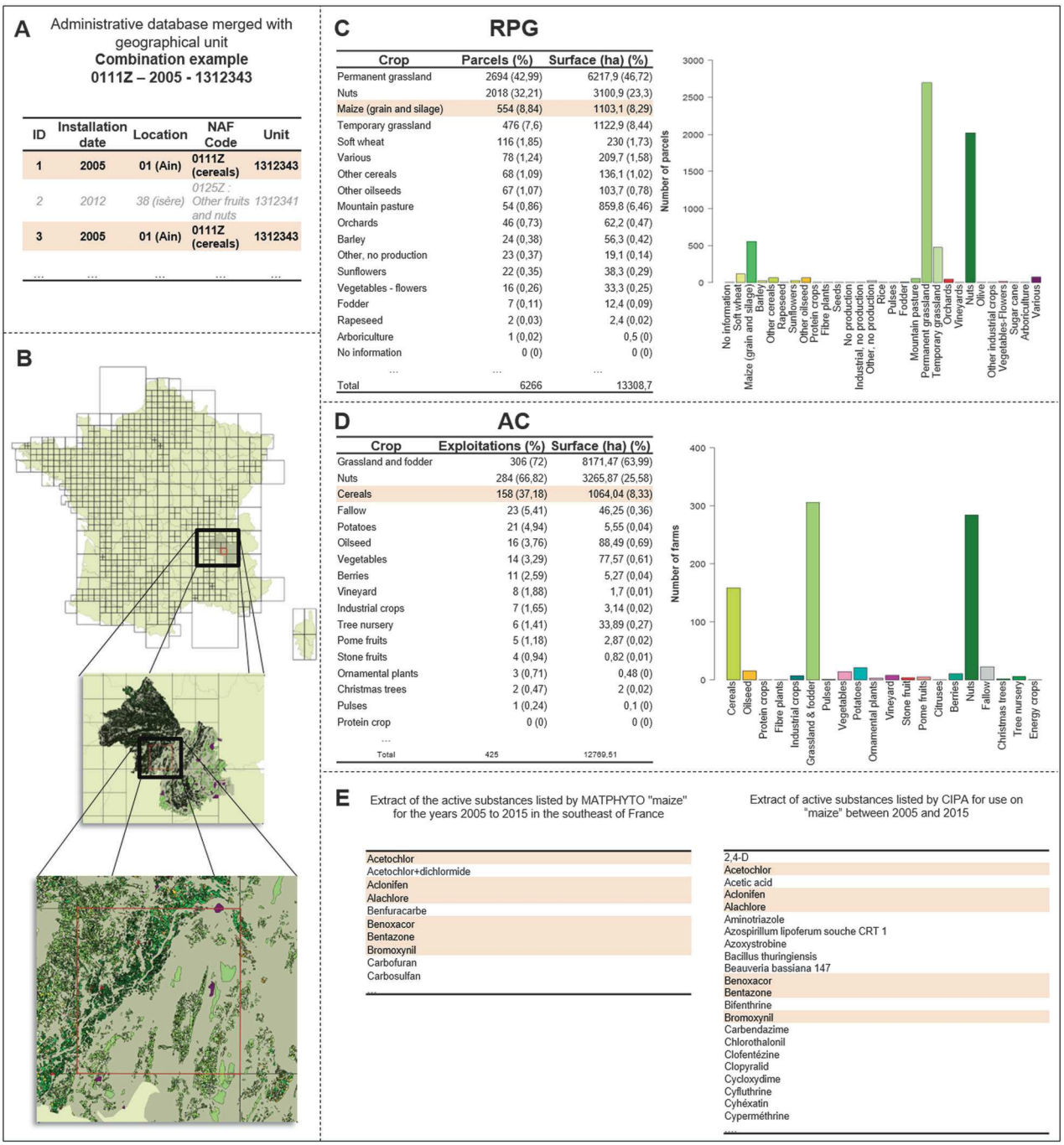


Fig. 3 Example of pesticide use estimation from MSA administrative data crossed with different external data sources. **a** Example of the administrative information retained for a combination: NAF—Year –Geographical unit; **b** Location of the geographical unit taken as the example and graphical representation of the parcels from the RPG 2012; **c** 2012 RPG data: Number of parcels present per crop category within the unit; **d** 2010 AC data: Number of crop exploitations in each crop category within the mesh; **e** Extract of the active substances listed by the crop-exposure matrix and CIPA

"permanent" and "temporary" grasslands in the example above). The related substances are then extracted from the crop-exposures matrices and/or CIPA compilation (Fig. 3e).

Even if data from Agriculture Ministry (AC and cultivation practices survey) might have been useful for our purposes (see illustration Fig. 3), the virtual environment in which they currently have to be processed, prevented any use with our automated algorithm, or AC for one of the geographical units is presented

The analysis and the algorithm development were performed with R software, version 3.3.2 [17]; and QGIS software, version 2.18.3 [18] was used for geographic data.

Results

Illustration of results given by the algorithm at the collective level

Although we deployed the algorithm on the overall agricultural population, we choose here to present a focus on one specific year (2012) in order to illustrate the outcomes as clearly as possible. External sources used by the algorithm were the RPG, the available MATPHYTO matrices for cereals and maize, and the CIPA compilation to complement information when matrices are not applicable (see Fig. 2). We restricted the data from the matrices to substances whose probability of use in the agricultural worker's area was higher or equal to 50% probability. In total, we applied the algorithm on 1,049,680 agricultural workers (1,357,216 declarations) for whom the combination NAF code – year – geographical unit was available, but also for whom the NAF code crossed with at least one RPG code. From the 720 initial units, 718 have been successfully merged with RPG data, meaning that at least one crop was declared in 2012 in 718 areas. Descriptive statistics about crops and pesticides for both units and individuals are summarized in Table 3.

In 2012, there were between 3 and 73,165 distinct parcels by unit (mean = 8829, IQR = 5073), for an average surface of 13,152 hectares (IQR = 5210 ha) and 18 RPG codes (IQR = 15). For each unit, the average number of farmers working on crops was 3,351 (IQR = 2427). More than 50% of parcels were related to three RPG codes: "permanent grassland" (28.7%), "temporal grassland" (15%) and "cereals" (13.4%). Considering the NAF codes, 50% of declaration were related to "viticulture" (19.5%), "cereals, legumes, oilseeds" (14.2%), "Technical or vocational secondary education" (6.3%), "landscaping" (4.6%), "pome fruits and stone fruits" (4%) and "vegetables, melons, roots and tubers" (3.7%). Focusing only on NAF code matching with RPG codes, "viticulture", "cereals, legumes, oilseeds", "pome fruits and stone fruits" and "vegetables, melons, roots and tubers" accounted for more than 80% of observations. Thanks to the algorithm, we were able to identify a list of pesticides potentially used by each worker, depending on her/his combination NAF code – year – geographical unit. An average of 527 substances were identified in the units (divided into 168 chemical families), but at the individual level, we identified an average of 77 substances used (mean family = 34), for an average period of 7.9 years.

Illustration of results given by the algorithm at the individual level

Table 4 illustrates the difference that could occur between individuals, depending on their administrative characteristics such as status, location and occupational activities. The first part of the table refers to the MSA administrative

Table 3 Algorithm results about crops and pesticides by units and workers for the MSA population in 2012, using RPG, MATPHYTO matrices and CIPA data

	Mean	Sd	Median	Minimum	Maximum	IQR
Cross between RPG data/Units						
Number of parcels by unit	8,829.8	7,737.8	6,572	3	73,165	5,073.8
Number of distinct codes of RPG by unit	18.3	3.1	18	2	25	5.0
Total surface (hectares)	13,152.2	11,359.1	10,283	1	91,921	5,210.0
Cross MSA data/Units						
Number of distinct NAF code by unit	41.9	12.7	39	17	118	15.0
Number of agricultural workers[a] by unit	3,351.5	3,757.3	2,198	500	40,066	2,427.3
Results of algorithm/unit						
Number of distinct potential substances by unit	526.8	70.0	536	0	644	61.0
Number of distinct potential chemical family[b] by unit	168.4	16.7	172	0	189	12.0
Results of algorithm/MSA individuals						
Number of distinct potential substances by individual	76.8	113.7	13	0	626	102.0
Number of distinct potential chemical family[b] by individual	34.1	40.4	10	0	182	51.0
Period of study	7.9	9.9	1	1	66	13.0

RPG Graphical Land Parcel Registration, *MSA* Mutualité Sociale Agricole, *CIPA* Compilation of Phytosanitary Indexes

[a]Restricted on the population of farmers working on crops

[b]Chemical family as defined in MATPHYTO matrices and CIPA

Table 4 Results of the algorithm for five different individuals, illustrating different contexts as status, location and availability of crop-exposure matrices

	Individual #1	Individual #2	Individual #3	Individual #4	Individual #5
Administrative data from MSA					
Status	Self-employed	Self-employed	Self-employed	Self-employed	Employee
NAF code	0111Z - Cereals, legumes, oilseeds	0111Z - Cereals, legumes, oilseeds	0111Z - Cereals, legumes, oilseeds	0121Z - Viticulture	0121Z - Viticulture
Location of the unit	Northeast	Southwest	Southwest	Northeast	Southeast
Year of installation / Contract	2002	2006	2006	2009	2005
Cross with RPG data to list of crops (Step 1 and 2)					
Label of RPG code(s) found in the unit	Common wheat, Maize (grain and forage), Barley, Other cereals, Pasture, Permanent grassland, Temporary grassland, Other	Common wheat, Maize (grain and forage), Barley, Other cereals, Other set-aside, Forage, Pasture, Permanent grassland, Temporary grassland, Orchard, Nuts, Vegetables-Flowers, Arboriculture, Other	Maize (grain and forage), Other cereals, Sunflower, Other set-aside, Pasture, Permanent grassland, Temporary grassland, Orchard, Nuts, Industrial crops, Vegetables-Flowers, Other, No information on the crop	Common wheat, Maize (grain and forage), Barley, Other cereals, Rapeseed, Sunflower, Other oilseed, Protein crops, Flowering plants, Seed, Other set-aside, Leguminous seed, Forage, Permanent grassland, Temporary grassland, Orchard, Viticulture, Industrial crops, Vegetables-Flowers, Other, No information on the crop	Common wheat, Maize (grain and forage), Barley, Other cereals, Rapeseed, Sunflower, Protein crops, Seed, Other set-aside, Leguminous seed, Forage, Pasture, Permanent grassland, Temporary grassland, Orchard, Viticulture, Nuts, Olive tree, Industrial crops, Vegetables-Flowers, Arboriculture, Other, No information on the crop
Label of RPG code(s) related to NAF code (intersection of RPG et NAF codes)	Common wheat, Maize (grain and forage), Barley, Other cereals, Rapeseed, Sunflower, Other oilseed, Protein crops, Seed, Leguminous seed	Common wheat, Maize (grain and forage), Barley, Other cereals, Rapeseed, Sunflower, Other oilseed, Protein crops, Seed, Leguminous seed	Common wheat, Maize (grain and forage), Barley, Other cereals, Rapeseed, Sunflower, Other oilseed, Protein crops, Seed, Leguminous seed	Viticulture	Viticulture
Common labels between NAF and RPG in unit	Common wheat, Maize (grain and forage), Barley, Other cereals	Common wheat, Maize (grain and forage), Barley, Other cereals	Maize (grain and forage), Other cereals, Sunflower	Viticulture	Viticulture
Cross-referencing with crop-exposure matrices and/or CIPA to list substances (Step 3)					
Reporting period (until 2012)	11 years	7 years	7 years	4 years	8 years
Use of available crop-exposures matrices	Yes	Yes	Yes	No	No
Use of CIPA	No	No	Yes	Yes	Yes
Results: Pesticides potentially used over the period (Step 4)					
Number (families)	8 (8)	7 (7)	43 (31)	37 (31)	69 (50)
List of potentially used pesticides	Any fungicide, Any herbicide, Any insecticide, Any pyrethroid, Any strobilurin, Any triazole, Atrazine, Nicosulfuron	Any fungicide, Any herbicide, Any insecticide, Any strobilurin, Any triazole, Mesotrione, Nicosulfuron	Acetic acid, Alachlor, Any fungicide, Any herbicide, Any insecticide, Any strobilurin, Any triazole, Bifenthrin, Boscalid, Carbendazim, Chlormequat chlorine, Chlorothalonil, Chlorpyriphos ethyl, Dicofol, Dimethenamid, Endosulfan, Fluroxypyr, Flusilazole, Fomesafen, Haloxyfop-r, Imazamox, Iprodione, Malathion, Maneb, Mesotrione, Methomyl, Nicosulfuron, Nonyl phenol ethoxylates, Oxydemeton-methyl, Paclobutrazol, Paraquat, Phosalone, Procymidone, Prothioconazole, Pyraflufen-ethyl, Quinoxyfen, Sulfosate, Thiaclopride, Thiodicarb, Thiophanate-methyl, Triazamate, Trifluralin, Vinclozolin	Acetic acid, Ametoctradin, Benalaxyl-M, Bifenthrin, Butralin, Captan, Chlorantraniliprole, Polymeric wax, Hydrogene cyanamide, Cyazofamid, Cyhexatin, Dichlobenil, Dicofol, Dinocap, Disodium phosphonate, Emamectin benzoate, Flufenoxuron, Fluopicolide, Fluroxypyr, Flusilazole, Heptamaloxyglucan, Insecticide white oil, Lufenuron, Mandipropamid, Meptyldinocap, Methoxyfenozide, Oxyquinoline, Penoxsulam, Potassium phosphate, Propargite, Propineb, Proquinazid, Pyraflufen-ethyl, Thiram, Trichoderma atroviride strain I-237, Trinexapac-ethyl, Valiphenal	Ametoctradin, Acetic acid, Beta-indolebutyric acid, Azinphos methyl, Azocyclotin, Bacillus subtilis, Benalaxyl-M, Benthiavalicarb isopropyl, Bifenthrin, Boscalid, Butralin, Captan, Carbaryl, Carbendazim, Carfentrazone-ethyl, Chlorantraniliprole, Polymeric wax, Hydrogene cyanamide, Cyazofamid, Cyhexatin, Diazinon, Dichlobenil, Dichlorvos, Dicofol, Diethofencarb, Dinocap, Disodium phosphonate, Diuron, Emamectin benzoate, Fenugreek extract, Fluopicolide, Fluroxypyr, Fenarimol, Fenbutatin-oxide, Fenitrothion, Flufenoxuron, Flusilazole, Hexaconazole, Petroleum oil, Empyreumatic oil, Heptamaloxyglucan, Insecticide white oil, Lufenuron, Malathion, Mandipropamid, Mepanipyrim, Meptyldinocap, Methomyl, Methoxyfenozide, Metrafenone, Oxyquinoline, Paraquat, Penoxsulam, Phosalone, Potassium phosphate, Procymidone, Propargite, Propineb, Proquinazid, Pyraflufen-ethyl, Rotenone, Spinosad, Sulfosate, Thiodicarb, Thiram, Trichoderma atroviride strain I-237, Trinexapac-ethyl, Valiphenal, Vinclozolin

data; the second refers to the merge between the geographical subdivision and the RPG; the third refers to the link with the external sources of pesticides uses, either the MATPHYTO matrices or the CIPA compilation; and the fourth refers to the outcome of the algorithm.

Therefore, individuals #1 to #4 were self-employed, installed since 2002 (#1), 2006 (#2 and #3) and 2009 (#4); and located in Northeast of France (#1 and #4) and Southwest (#2 and #/3). Individual #5 was a salaried worker since 2005 in the Southeast. For the individual #1, #2 and #3 working in the same occupational activity "Cereals, legumes, oilseeds", but in two different areas, we draw up a NAF-related list of 10 potential RPG codes, however, only 4 (#1 and #2) and 3 (#3) were also inventoried in their unit. We used the matrices or CIPA, depending on the combination between the NAF code and RPG codes. For two of them the matrices were available, for the individual #3 the CIPA data were used. Based on the period between the installation date and the declaration date, the algorithm listed all substances used for this kind of crop, filtering on a probability of 50% for the matrices. According to the outcome, the individual #1 has potentially used 8 substances (from 8 families), over a period of 11 years. In comparison, individuals #2 potentially used 7 substances over 7 years and individual #3 43 substances over 7 years. Individual #4 and #5 worked in the same occupational activity "viticulture" but in two different areas and with different statuses, self-employed and employee. Only one specific RPG code was related to the viticulture, and the list of potential substances used by both of them contained respectively 37 and 69 substances over 4 and 8 years.

Discussion

We have developed and implemented a strategy to estimate the past use of phytosanitary substances by all French agricultural workers, in order to include it in a wider project involving data mining of medico-administrative data of these workers (chronic diseases and medical expenses) so as to generate hypotheses on occupational health determinants. In addition to the MSA databases, five sources of information were identified. Two of them, the RPG and the AC, coupled with a geographical grid conceived for this project, can allow us to make an inventory of the crops cultivated; while the others, especially the crop-exposure matrices allowed us to list the phytosanitary substances potentially used for each of these crops in a given geographical area (unit) and period. The major point of this method is to be able to generate, *a posteriori*, a list of substances potentially used, based on individual administrative data (NAF code, year and location).

In this paper, we applied the algorithm to the population of 1,049,680 farmers working on crops who declared an activity in 2012. We were able to list, for each individual, the list of the potential substances used during her/his occupational activities, based on her/his administrative data as single starting point. When looking at the geographical unit itself, the number of substances potentially used was substantial (mean = 522), but when refining the results by the NAF code and the period of activity this number decreases to 69 substances. This number may seem relatively high; however, it is important to note that this corresponds to the list of the substances potentially used by a worker, depending on her/his different activities, over a period of activity from 1 to several decades. The examples in Table 4 illustrate the variability between different individuals, according to their administrative information. The list of substances potentially used varied for individual in the same area but in different occupational activity and conversely. An important point that can be raised is the major advantage of having expert data of crop-exposure matrices, since the number of potentially used substances is getting bigger when using CIPA data.

Although we identified several different external sources, some data sources were better suited than others for use with an automated algorithm, intended for the automated production of a list of substances potentially used by each agricultural worker. Thus, for the automated algorithm, it was easier to use RPG data sources, crop-exposure matrices and the CIPA database to generate a list of phytosanitary substance use. The structure of the other data sources does not currently allow them to be integrated in an automated system, although their use at a later stage is not excluded (e.g. in second intention for the exploration of signals). The RPG and the AC each have advantages and disadvantages. The advantages of the RPG are its accessibility and the high periodicity of its update, which makes it possible to take into account rotations of crops (biennial, triennial, etc.) using the RPG from several consecutive years. On the other hand, certain crops, such as vines, are less well informed because they are not the subject of a request for subsidies from the European Common Agricultural Policy, which is the primary purpose of the RPG. The AC is exhaustive and the crop categories are much more detailed, however, the 10-year period between censuses does not take into account crop rotations. In addition, access to it via a virtual network with restrictions complicates its integration in an automated system and crossover with other data available outside this secure virtual space.

The three other sources of data allow one to estimate from the crops listed, the use of phytosanitary substances by an individual worker, and potentially, his/her exposure to these substances. The MATPHYTO crop-exposure matrices list the most commonly used toxic substances for four crop categories (two available at the time of the study), providing exposure parameters for a given period and geographical

location. However, because data about substance use are retrospectively collected, the accuracy of this information decreases with time. In addition, crop-exposure matrices are not available for all crops; only the most commonly used and toxic substances are identified and other matrices are under construction. It is also important to note that another crop-exposure matrix, the PESTIMAT project [19] could be used. To date, this matrix does not cover the whole of France and crops and it was therefore decided not to consider it to start with. Further work and collaborations are planned to evaluate the interest of integrating this crop-exposure matrix in the algorithm. While the CIPA database compiles all phytosanitary substances used in agriculture since 1961, the classification is such that certain substances cannot be directly associated with a specific crop (the case for soil or seed treatments). The number of substances used per crop may therefore be underestimated. Moreover, no quantitative information is provided within CIPA. Finally, agricultural workers not involved in crop production, such as those working in animal husbandry, are not currently included in our algorithm due to lack of job exposure matrices for this population even though they may be exposed to pesticides.

If our initial approach allows one to estimate the number of phytosanitary substances potentially used by individual worker, it does not currently take into account quantitative information, such as the abundance of a particular crop within in each geographical unit or the proportion of users of each substance (except for the fixed threshold at 50%). However, as the surface exploited is available within the MSA database, it is envisaged to take these factors into account in a subsequent stage of the project, so as to refine the probabilities of use for each substance. Also, it should be noted that the list of substances generated by the algorithm provides only an estimate of occupational exposure. While exposure depends on factors such as the crop, the geographical location, the year of exploitation or the type of pesticides approved for use, it also depends on individual factors such as the tasks performed (e.g. preparation of the mixture, application, re-entry into the parcel etc.), the technique and equipment used, or the use of personal or collective protection [20–22] that cannot be included in our approach because the information is unavailable. Finally, the succession of steps used to create the list of pesticides may generate bias. Indeed, the list of crops potentially used by an individual worker is derived from the NAF code and the geographical location of the exploitation, although in practice the worker probably did not cultivate all the listed crops and did not use the whole range of authorized substances. At last, no information is available as to whether the phytosanitary substance of interest was applied by the exploitation owner or by an employee. Given the large number of individuals potentially exposed over the entire duration of the whole study (more than 6 million), it appears difficult to deal with these biases. Nevertheless, while this approach to exposure may seem to lack precision at the individual level, it remains relevant at a population level in the context of sanitary vigilance. Our next step will be to validate our approach to identifying the crops and the phytosanitary substances used on them by comparison with real data from sample farms in different sectors of activity and different geographical units. We also want to evaluate the possibility of adding other expert data sources, as previously stated.

Importantly, our work has a potential of generalization to other national contexts. First, the principle of the geographical grid with cells size adapted to the number of agricultural workers (to avoid re-identification) can easily be generalised with census data concerning agricultural workers, available in many countries (but also for other categories of workers). For instance, in European Union, thanks to the Common Agricultural Policy, there is a unique identification of farmers (the so-called farmer's register) that could be used to create such a grid at EU level (this register belongs to the Integrated Administration and Control System described below). The second point is then to match this grid with data regarding cultures and then cultural practices within the grid cells. At the EU scale again, the EU Common Agricultural Policy (CAP) allows direct aid to farmers according to several criteria related to their parcels, cultures, number of animals, etc. (but also some practices such as organic vs traditional cultures). In order to ensure that this direct aid is correctly allocated to the right farmers, Member States should operate since 1992 a system for the management and control of payments to farmers (the "Integrated Administration and Control System" (IACS) which include an IT solution [23]. Technically, IACS includes several computerized and interconnected databases, including a system for the identification of all agricultural parcels in Member States (Land Parcel Identification System or LPIS) [24], which is the generic name of the "Graphical Land Parcel Registration" we presented. More precisely, among the 28 Member States, there were in 2015, 44 national or regional LPISs in operation, containing over 135 million referenced parcels [24]. Therefore, our method should be generalizable, at least in these main features, in EU countries, which count around 22 million people working regularly in the agricultural sector, and 11 million farms. The generalisation in non EU-countries could be further investigated.

Conclusion

Conventional epidemiology, which deals with population samples, attempts to estimate exposure in different ways either by directly soliciting the memory of the interested

parties (active file of a cohort, case-controls), or in a probabilistic way (job-exposure matrices or expert opinions). However, this is not compatible with the processing of massive amounts of exhaustive data from whole national agricultural population. We show that the combination of medico-administrative data on agricultural workers, from their healthcare insurance system, and external sources of information created for other purposes, constitutes an innovative and interesting approach to estimate the use of phytosanitary substances by agricultural workers throughout France, and that this method has a potential of generalisation at international level. This strategy will soon be used as part of a large-scale occupational risk surveillance project covering the entire French agricultural population, by combining the activity and probability of pesticide use data with diseases and health expenses recorded by the agricultural workers' healthcare insurance system.

Acknowledgements We thank the Mutualité Sociale Agricole (MSA) for their support, especially Nadia Joubert (current head of MSA statistical department), Alain Pelc (former head of MSA statistical department), the Medical prestation data department supervised by Damien Ozenfant, as well as privileged contact, Marc Parmentier, Patrick Le Bourhis and Valérie Vincent (Contributors data), Nicolas Sabin (Long Term Diseases Data), Nicolas Viarouge and Sébastien Odiot (occupational accidents and diseases data), Thierry Grech (Retirees data), as well as Prof William Dab (former chair of MSA scientific committee), Prof Anne Laure Crémieux (former MSA national physician) and Prof Jean Marc Soulat (current MSA national physician) for their interest and support for this work. We thank the French Agency for Food, Environmental and Occupational Health & Safety ANSES, for the grant which allowed this project to be conducted, and especially Mathilde Merlo, Alexandra Papadopoulos, Fabrizio Botta, and Jean Luc Volatier for their technical and scientific support. We thank the French Public Health Agency, Santé Publique France, especially Johan Spinosi and Laura Chaperon (agronomists) and Mounia El Yamani (head of the occupational exposure assessment Unit) to have made their crop-exposure matrix available for this work, and for their technical support on agricultural practices. Finally, we thank Alison Foote (Grenoble Alpes University Hospital) for editing the manuscript.

Funding This project was funded and supported by ANSES (grant agreement 2016-CRD-03_PPV16) via the tax on sales of plant protection products. The proceeds of this tax are assigned to ANSES to finance the establishment of the system for monitoring the adverse effects of plant protection products, called 'phytopharmacovigilance' (PPV), established by the French Act on the Future of Agriculture of 13 October 2014.

Author contributions VB and DBR designed the project and acquired the necessary funding and collaborations. AP and DBR performed data management, data analysis and algorithm development. CC provided expertise in GIS and performed the geographical grid. JS provided crop exposures matrices and technical support on cultural practices. DO provided MSA data and technical support on medico-administrative aspects. VB, DBR, AP and CM contributed to the data interpretation. AP, CM, DBR and VB wrote the manuscript. All authors discussed the results and commented on the manuscript.

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflict of interest.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

References

1. Alavanja MCR, Bonner MR. Occupational pesticide exposures and cancer risk: a review. J Toxicol Environ Health Part B. 2012;15:238–63. https://doi.org/10.1080/10937404.2012.632358.
2. INSERM. Pesticides. Effets sur la santé. Paris: Inserm; 2013.
3. Gangemi S, Miozzi E, Teodoro M, Briguglio G, De Luca A, Alibrando C, et al. Occupational exposure to pesticides as a possible risk factor for the development of chronic diseases in humans. Mol Med Rep. 2016;14:4475–88. https://doi.org/10.3892/mmr.2016.5817.
4. Gunnarsson L-G, Bodin L. Parkinson's disease and occupational exposures: a systematic literature review and meta-analyses. Scand J Work Environ Health. 2017;43:197–209. https://doi.org/10.5271/sjweh.3641.
5. Alavanja MCR, Samanic C, Dosemeci M, Lubin J, Tarone R, Lynch CF, et al. Use of agricultural pesticides and prostate cancer risk in the Agricultural Health Study cohort. Am J Epidemiol. 2003;157:800–14.
6. Bonner MR, Freeman LEB, Hoppin JA, Koutros S, Sandler DP, Lynch CF, et al. Occupational exposure to pesticides and the incidence of lung cancer in the Agricultural Health Study. Environ Health Perspect. 2017;125:544–51. https://doi.org/10.1289/EHP456.
7. Lerro CC, Koutros S, Andreotti G, Friesen MC, Alavanja MC, Blair A, et al. Organophosphate insecticide use and cancer incidence among spouses of pesticide applicators in the Agricultural Health Study. Occup Environ Med. 2015;72:736–44. https://doi.org/10.1136/oemed-2014-102798.
8. Brouwer M, Schinasi L, Beane Freeman LE, Baldi I, Lebailly P, Ferro G, et al. Assessment of occupational exposure to pesticides in a pooled analysis of agricultural cohorts within the AGRICOH consortium. Occup Environ Med. 2016;73:359–67. https://doi.org/10.1136/oemed-2015-103319.
9. Boulanger M, Tual S, Lemarchand C, Guizard A-V, Velten M, Marcotullio E, et al. Agricultural exposure and risk of bladder cancer in the AGRIculture and CANcer cohort. Int Arch Occup Environ Health. 2017;90:169–78. https://doi.org/10.1007/s00420-016-1182-y.
10. Lemarchand C, Tual S, Boulanger M, Levêque-Morlais N, Perrier S, Clin B, et al. Prostate cancer risk among French farmers in the AGRICAN cohort. Scand J Work Environ Health. 2016;42:144–52. https://doi.org/10.5271/sjweh.3552.
11. MSA. Les chiffres utiles de la MSA. Bagnolet: Mutualité Sociale Agricole; 2017.
12. MSA. Tableau de bord ALD. Bagnolet: MSA; 2015. Incidence 2015 et Prévalence au 31 décembre.
13. Ministère de l'Agriculture et de l'alimentation. AGRESTE - La statistique, l'évaluation et la prospective agricole - Structure des exploitations - recensements. 2009. http://agreste.agriculture.gouv.fr/enquetes/structure-des-exploitations-964/ (accessed 26 Jul 2017).
14. Spinosi J, Févotte J. Le programme Matphyto - Matrices cultures-expositions aux produits phytosanitaires. Saint-Maurice (Fra): Institut de Veille Sanitaire; 2008. http://invs.santepubliquefrance.fr/publications/2008/matphyto/rapp_sci_matphyto.pdf (accessed 16 Jul 2018).

15. Chaperon L, Perrier L, Spinosi J, El Yamani M. Eléments techniques sur la compilation des index phytosanitaires Acta. Saint Maurice: Institut de Veille Sanitaire; 2016. http://invs.santepubliquefrance.fr/Publications-et-outils/Rapports-et-syntheses/Travail-et-sante/2016/Elements-techniques-sur-la-compilation-des-index-phytosanitaires-Acta (accessed 16 Jul 2018)a.
16. Santé Publique France. Compilation des index ACTA. Compil. Index ACTA Un Outil Programme MatPhyto. 2016. http://matphyto.acta-informatique.fr/Accueil (accessed 16 Jul 2018).
17. R Core Team. R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing; 2017. https://www.R-project.org/.
18. QGIS Development Team. QGIS Geographic Information System. Chicago (USA): Open Source Geospatial Foundation; 2016.
19. The PESTIMAT Group, Baldi I, Carles C, Blanc-Lapierre A, Fabbro-Peray P, Druet-Cabanac M, et al. A French crop-exposure matrix for use in epidemiological studies on pesticides: PESTIMAT. J Expo Sci Environ Epidemiol. 2017;27:56–63. https://doi.org/10.1038/jes.2015.72.
20. Baldi I, Lebailly P, Rondeau V, Bouchart V, Blanc-Lapierre A, Bouvier G, et al. Levels and determinants of pesticide exposure in operators involved in treatment of vineyards: results of the PESTEXPO Study. J Expo Sci Environ Epidemiol. 2012;22:593–600. https://doi.org/10.1038/jes.2012.82.
21. Baldi I, Lebailly P, Bouvier G, Rondeau V, Kientz-Bouchart V, Canal-Raffin M, et al. Levels and determinants of pesticide exposure in re-entry workers in vineyards: Results of the PESTEXPO study. Environ Res. 2014;132:360–9. https://doi.org/10.1016/j.envres.2014.04.035.
22. Anses. Expositions aux pesticides des utilisateurs et des travailleurs agricoles. Maison-Alfort: Agence nationale de sécurité sanitaire de l'alimentation, de l'environnement et du travail; 2016.
23. European Commission. Integrated Administration and Control System (IACS). Agric. Rural Dev. - Eur. Comm. 2012. https://ec.europa.eu/agriculture/direct-support/iacs_en (accessed 29 Apr 2019).
24. European Court of Auditors. The Land Parcel Identification System: a useful tool to determine the eligibility of agricultural land – but its management could be further improved. Luxembourg: Publications Office of the European Union:: European Union; 2016. https://www.eca.europa.eu/Lists/News/NEWS1610_25/SR_LPIS_EN.pdf.

TITRE : Analyse des données massives de source assurantielle de la Mutualité Sociale Agricole, pour la surveillance en santé au travail des travailleurs agricoles en France

RESUME

Introduction : La surveillance sanitaire et la vigilance (identification de nouveaux risques en particulier) représentent un enjeu majeur dans le champ santé-travail. En complément des études épidémiologiques classiques, l'analyse systématique, sans a priori, de données collectées en routine pourrait être un atout pour la détection précoce de pathologies en lien avec le travail. Dans ce contexte, la Mutualité Sociale Agricole (MSA), le régime de protection sociale dédié aux travailleurs agricoles français, a souhaité développer son activité de vigilance en exploitant ses données médico-administratives, utilisées pour le remboursement de prestations de santé. En partenariat avec l'Agence nationale de sécurité sanitaire de l'alimentation, de l'environnement et du travail (Anses), un projet de fouille des données a donc été mis en place dans lequel ce travail de thèse s'inscrit. L'objectif de la thèse consiste plus précisément à tester, sans hypothèses préalables, l'existence ou non d'associations entre les activités agricoles et les pathologies reconnues en tant qu'affection de longue durée (ALD). **Méthode** : Les travaux présentés ont été menés sur la population de non-salariés (chefs d'exploitation ou d'entreprise) affiliés à la MSA, en disposant d'une part de données de cotisations, renseignant au niveau individuel, les activités professionnelles, caractéristiques démographiques et socio-économiques, et d'autre part, de données médico-administratives renseignant les déclarations de pathologies reconnues en ALD et informations associées dont la pathologie codée en CIM-10. Grâce à l'accord de la CNIL, un identifiant unique a été créé pour que, pour la première fois, ces données administratives et médico-administratives puissent être fusionnées et restructurées afin de permettre l'application de modèles. Des modèles de régression logistique ont été utilisés, en adaptant la sélection de variables pour chaque ALD et en utilisant la validation croisée afin de limiter le surajustement des modèles. Plusieurs méthodes ont été testées pour mieux prendre en compte les facteurs de confusion potentiels. Ces différents modèles ont ensuite été évalués via des mesures de robustesse et appliqués aux données à deux niveaux de précision pour la pathologie (ALD et CIM-10). Les associations statistiques entre chaque combinaison d'activité professionnelle et de pathologie ont été caractérisées par leur p-valeur, corrigées pour les tests multiples, et la valeur de l'odds ratio correspondant. **Résultats** : Le traitement des données a permis d'étudier une population constituée de 899 212 non-salariés affiliés entre 2006 et 2016. Au sein de cette population, il a été possible d'identifier 100 706 individus avec au moins une déclaration d'ALD sur la période d'observation. La méthodologie appliquée a mis en évidence 54 associations statistiquement significatives entre une activité professionnelle et une ALD, permettant à la fois de capturer des déterminants de santé déjà connus ou suspectés mais aussi de générer des hypothèses intéressantes. Après ajustement sur des facteurs de confusion, les secteurs agricoles les plus associés à des pathologies, faisant l'objet d'ALD chez les non-salariés, sont la viticulture, l'exploitation de bois, le paysagisme, et les entreprises de jardins ou de reboisement. **Discussion** : Ce travail de thèse apporte une première démonstration de la faisabilité et de la pertinence de l'analyse systématique des données collectées en routine à des fins assurantielles, sur l'ensemble de la population agricole, pour rechercher des risques sanitaires associés aux diverses activités professionnelles. Les « signaux » ainsi mis en évidence seront investigués à l'aide d'un groupe d'experts. D'autres modèles pourront être testés, au premier rang desquels les modèles de survie. Cette approche pourra ainsi constituer un outil précieux contribuant au dispositif de vigilance sanitaire des risques professionnels agricoles.

Mots clés : Bases de données administratives, Assurance Maladie, Surveillance épidémiologique, Fouille de données, Risques professionnels, Travailleurs agricoles

TITLE: Health insurance data analysis for occupational health surveillance of French agricultural workers

ABSTRACT

Introduction: Health surveillance and vigilance (identification of new risks in particular) represent a major challenge in the field of occupational health. In addition to classical epidemiological studies, the systematic analysis, without a priori, of data collected routinely could be an asset for the early detection of diseases related to work. In this context, the social protection scheme dedicated to French agricultural workers, known as "Mutualité Sociale Agricole" (MSA), wanted to develop its vigilance activity by exploiting its medico-administrative data, used for the reimbursement of health expenditures. In partnership with the French Agency for Food, Environmental and Occupational Health & Safety (ANSES), a data mining project has been set up in which this thesis work fits. The aim of the thesis is, more precisely, to test, without any prior assumptions, the existence of associations between agricultural activities and pathologies recognized as long-term disease (LTD). **Method**: The work presented was conducted on self-employed population (heads of farms or enterprises) affiliated to the MSA. It relied on the one hand on a contributors' database which includes, at the individual level, information about occupational activities, demographic and socio-economic characteristics, and on the other hand, on a medico-administrative database with declarations of long-term diseases (LTD) and associated information like ICD-10 diseases. Thanks to the agreement of the French Data Protection Authority (CNIL), a unique identifier was created so that, for the first time, these administrative and medico-administrative data could be merged and restructured to allow the application of models. Logistic regression models were performed, adapting variable selection for each LTD and using cross-validation to limit over-fitting of models. Several methods have been tested to better take into account potential confounders. These different models were evaluated via robustness measures and applied at two-level of precision for pathology (LTD and ICD-10). The statistical associations between each combination of occupational activity and LTD were characterized by p-values, corrected for multiple tests, and odds ratio. **Results**: Data management allowed us to consider a population of 899 212 self-employed affiliated between 2006 and 2016. Among them, it was possible to identify 100 706 individuals with at least one declaration of LTD over the observation period. The applied methodology revealed 54 statistically significant associations between an occupational activity and a LTD, making it possible to capture already known or suspected health determinants but also to generate interesting hypotheses. After adjusting for confounding factors, the agricultural sectors most associated with LTD, among the self-employed, are viticulture, timber exploitations, landscaping and gardening or reforestation. **Discussion**: This thesis provides a first demonstration of the feasibility and relevance of the systematic analysis of data collected routinely for insurance purposes, concerning the overall agricultural population, to search for health risks associated with occupational activities. The statistical "signals" thus highlighted will then be investigated by a group of experts from different scientific and occupational fields. Other models should be tested like survival models. This approach may thus be a valuable tool contributing to the health surveillance system dedicated to agricultural workers.

Keywords: Medico-administrative databases, Health insurance, Epidemiologic surveillance, Data mining, Occupational risks, Agricultural workers