HAL
open science

Facteurs de risque professionnels des troubles musculo–squelettiques aux coudes et aux genoux

Eleonore Herquelot

To cite this version:

Eleonore Herquelot. Facteurs de risque professionnels des troubles musculo–squelettiques aux coudes et aux genoux. Médecine humaine et pathologie. Université de Versailles-Saint Quentin en Yvelines, 2015. Français. NNT : 2015VERS004V . tel-01179196

HAL Id: tel-01179196

https://theses.hal.science/tel-01179196v1

Submitted on 22 Jul 2015

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Université de Versailles St Quentin-en-Yvelines

Ecole Doctorale 420 de Santé Publique

THÈSE

Pour l'obtention

Du **DIPLÔME DE DOCTORAT**

de l'Université de Versailles Saint-Quentin-en-Yvelines

en **Epidémiologie – Santé Publique**

Eléonore HERQUELOT

Soutenue le 21 janvier 2015

Facteurs de risque professionnels des troubles musculo-squelettiques aux coudes et aux genoux

Directeur de thèse : **Alexis DESCATHA**

JURY

Joël ANKRI	Président
Cécile PROUST LIMA	Rapporteur
Philippe MAIRIAUX	Rapporteur
Jean-Claude PAIRON	Examinateur
Thierry THOMAS	Examinateur
Alexis DESCATHA	Directeur de thèse

Résumé de la thèse

Les troubles musculo-squelettiques (TMS) sont la principale cause d'absentéisme au travail. Ils représentent un coût économique important, mais ils ont aussi de graves conséquences au niveau individuel – douleurs persistantes, limitations fonctionnelles ou perte d'emploi. De nombreuses expositions professionnelles ont déjà été mises en évidence pour expliquer la présence de ces symptômes, mais certaines relations restent encore à confirmer.

L'objectif de ce travail de thèse était d'étudier l'association entre les facteurs professionnels, en particulier les facteurs physiques, et les TMS au niveau des coudes et des genoux. Les TMS au niveau des coudes ont été étudiés à travers la prévalence des symptômes aux coudes et des épicondylites, et l'incidence des épicondylites. Les douleurs aux genoux ont été étudiées à travers l'incidence des douleurs de courte ou de longue durée.

Ce travail a nécessité l'utilisation de méthodologies spécifiques, en particulier les diagrammes causaux et les méthodes de gestion des données manquantes qui seront explicitées dans une partie théorique.

La population dans ce travail de thèse était une cohorte de 3 710 sujets représentatifs des actifs des Pays de la Loire. Ils ont été recrutés entre 2002 et 2005 et ont été suivis entre 2007 et 2010. A chaque phase, un questionnaire sur les conditions de travail a été rempli par les participants et un examen clinique qui évaluait la présence de troubles musculo-squelettiques a été réalisé par des médecins du travail volontaires.

En conclusion, les facteurs professionnels mis en évidence étaient globalement des facteurs de mouvements répétitifs impliquant les articulations étudiées (torsion des poignets et flexion/extension des coudes ou le fait de s'agenouiller). Le travail en force (manipulation de charges, efforts physiques importants) et les tâches répétitives ont également été mis en évidence comme prédicteurs de TMS ultérieurs.

Mots clefs : Troubles musculo-squelettiques, épicondylites latérales, douleurs aux genoux, expositions professionnelles, imputations multiples

Thesis summary:

Occupational risk factors of musculoskeletal disorders at the elbow and knee level

Musculoskeletal disorders (MSDs) are a leading cause of absenteeism from work. In addition to their major economic impact, musculoskeletal disorders have important consequences on individuals with the persistence of pain, disability and potential job loss. Many occupational exposures were highlighted to explain these symptoms, but certain relationships between occupational exposures and MSDs should be examined further.

The aim of this thesis is to investigate the association between occupational exposures, especially biomechanical exposures, and elbow or knee MSDs. The elbow MSDs are studied with the prevalence of symptoms and lateral epicondylitis, and the incidence of lateral epicondylitis. The incidence of knee pain is examined according to its duration.

The methodologies and theories appropriate to this type of work is briefly presented - a section focuses on the causal diagrams and methods of handling missing data.

The population used is a cohort of 3710 subjects representative of the French workforce. Subjects were selected from workers undergoing a mandatory annual health examination between 2002 and 2005 and were followed between 2007 and 2010. For each phase of the study, a self-administered questionnaire on working conditions was completed and a clinical examination was performed in order to evaluate the presence of musculoskeletal disorders.

In conclusion, repetitive movements involving the joints studied such as wrist-twisting, flexion/extension of the elbows or kneeling are associated with MSDs. The notion of force, such as load handling and physical exertion, and the notion of repetitive tasks are also identified as predictors of subsequent MSDs.

Keywords : Musculoskeletal disorders, lateral epicondylitis, knee pain, occupational exposures, multiple imputations

Remerciements

Je tiens à remercier toutes les personnes qui ont pu rendre ce travail de thèse possible.

Tout d'abord, je suis très reconnaissante à Marie Zins, responsable de l'UMS 011, et à Marcel Goldberg de m'avoir accueilli au sein de l'équipe et en particulier pour leur appui précieux et leurs nombreux conseils au cours de ces 3 années de thèse.

Je remercie Alexis Descatha, mon directeur de thèse, d'avoir su me transmettre, avec une grande générosité, ses idées et son enthousiasme et de m'avoir permis de travailler sur des projets captivants. Il a toujours été d'un grand soutien.

Je remercie Joël Ankri d'avoir d'accepté d'être président du jury. Je tiens à remercier les deux rapporteurs, Philippe Mairiaux, pour ses remarques judicieuses et Cécile Proust-Lima, pour les questions intéressantes qu'elle soulève. Je suis reconnaissante à Jean-Claude Pairon et Thierry Thomas d'avoir accepté de faire partie de ce jury de thèse.

Je remercie également Annette Leclerc qui, avec une rigueur scientifique et une franchise sans pareil, a toujours su me poser les bonnes questions pour améliorer mon travail. Je la remercie également pour ses nombreuses relectures.

Je remercie Alice Guéguen et Gäelle Santin pour leurs conseils éclairés dans la bonne humeur et leur relecture du manuscrit de thèse qui ont permis une grande clarification des méthodes présentées dans ce travail.

Je remercie les médecins du travail du réseau de surveillance des troubles musculo-squelettiques dans les Pays de la Loire pour leur engagement afin d'améliorer les connaissances scientifiques. Je remercie le Département Santé Travail de l'Institut de Veille Sanitaire et l'équipe du LEEST coordonnée par le Pr Yves Roquelaure d'avoir permis à ce travail d'exister en m'autorisant à utiliser les données recueillies dans le cadre du réseau. Je remercie également Julie Bodin pour sa disponibilité et sa gentillesse.

Je remercie l'ED420 et son directeur, Jean Bouyer, pour m'avoir fourni un cadre de formation intéressante afin de mener à bien ma thèse. Je remercie également Audrey Bourgeois pour sa diligence et sa bienveillance.

Je remercie toute l'équipe UMS 011 pour leur bonne humeur en toutes circonstances, en particulier, pour les fous rires à la cafétéria. J'écris une mention spéciale pour mes collègues de bureau, Iris Hourani et Julie Gourmelen, autant pour leur soutien sur la thèse que pour tous les à-côtés.

Je tiens à remercier ceux qui ont croisé ma route durant cette thèse, en particulier Kevin Jean, Pascale Lissouba, Audrey Bonaventure et Fanny Leroy pour leur amitié et leur écoute durant toutes ces années.

Pour finir, je remercie ma famille sans qui rien ne serait possible…

Articles de la thèse

Articles acceptés

Work-related risk factors for lateral epicondylitis and other cause of elbow pain in the working population. Eleonore Herquelot, Julie Bodin, Yves Roquelaure, Catherine Ha, Annette Leclerc, Marcel Goldberg, Marie Zins, Alexis Descatha. American Journal of Industrial Medecine. Avril 2013

Work-related risk factors for incidence of lateral epicondylitis in a large working population. Eleonore Herquelot, Alice Guéguen, Yves Roquelaure, Julie Bodin, Celine Sérazin, Catherine Ha, Annette Leclerc, Marcel Goldberg, Marie Zins, Alexis Descatha. Scandinavian Journal of Work, Environment & Health. Aout 2013

Incidence of chronic and other knee pain and related occupational risk factors in a large working population. Eleonore Herquelot, Julie Bodin, Yves Roquelaure, Audrey Petit, Catherine Ha, Annette Leclerc, Marcel Goldberg, Marie Zins, Alexis Descatha. Annals of Occupational Hygiene. Janvier 2015.

Communications

Etude des facteurs de risque professionnels d'épicondylite latérale et d'autres types de douleurs de coude dans une population de salariés – Colloque Aderest. Mars 2012 (oral)

Average treatment effect estimation with a rare outcome: an example and simulations – 33rd Annual Conference of the International Society for Clinical Biostatistics. Aout 2012 (poster)

Interest of using multiple imputations. An example on assessment of work-related risk factors of the incidence of lateral epicondylitis – EPICOH. Juin 2013 (poster)

Natural history of lateral epicondylitis among French Workers – PREMUS. Juillet 2013 (oral)

Incidence des épicondylites latérales et facteurs de risque associés – Colloque Aderest. Octobre 2013 (poster)

Incidence of knee pain and its work-related risk factors in a large working population – EPICOH. Juin 2014 (oral)

Incidence des douleurs aux genoux et les facteurs professionnels associés dans une population d'actifs représentative – Adelf. Septembre 2014 (oral)

Autres travaux

Use of Propensity Scores in Occupational Health? Alexis Descatha, Annette Leclerc, et Eleonore Herquelot. Journal of Occupational and Environmental Medicine / American College of Occupational and Environmental Medicine 55, no 5. Mai 2013

Self-Reported Physical Exposure Association with Medial and Lateral Epicondylitis Incidence in a Large Longitudinal Study. Alexis Descatha, Ann Marie Dale, Lisa Jaegers, Eléonore Herquelot, et Bradley Evanoff. Occupational and Environmental Medicine 70, no 9. Septembre 2013

Description of Musculoskeletal Disorders and Occupational Exposure from a Field Pilot Study of Large Population-Based Cohort (CONSTANCES). Matthieu Carton, Annette Leclerc, Sandrine Plouvier, Eleonore Herquelot, Marie Zins, Marcel Goldberg, Yves Roquelaure, et Alexis Descatha. Journal of Occupational and Environmental Medicine / American College of Occupational and Environmental Medicine 55, no 8. Août 2013

Épicondylalgies latérales dans une cohorte de salariés ligériens : évolution et déterminants. Alexis Descatha, Eleonore Herquelot, Zakia Mediouni, Audrey Petit, Catherine Ha, Annette Leclerc, et Yves Roquelaure. Revue du Rhumatisme 81, no 4. Juillet 2014

Long-Term Persistence of Knee Pain and Occupational Exposure in Two Large Prospective Cohorts of Workers. Herquelot, Eléonore, Julie Bodin, Audrey Petit, Catherine Ha, Annette Leclerc, Marcel Goldberg, Marie Zins, Yves Roquelaure, et Alexis Descatha. *BMC Musculoskeletal Disorders* 15, n° 1. Septembre 2014

Table des matières

Résumé de la thèse 1
Thesis summary: 2
Remerciements 3
Articles de la thèse 5
Communications 5
Autres travaux 6

1 Contexte et objectifs 11

1.1 Contexte et généralités 11
1.1.1 TMS et douleurs articulaires 11
1.1.2 Expositions professionnelles 14
1.1.3 Modèle conceptuel d'apparition des douleurs et des TMS 15
1.1.4 Forme du lien entre les facteurs professionnels et les TMS 17

1.2 Etat de la littérature 18
1.2.1 Sur les symptômes aux coudes 18
1.2.1.1 Prévalence 19
1.2.1.2 Incidence 20
1.2.1.3 Associations avec les facteurs professionnels 20
1.2.2 Sur les symptômes aux genoux 21
1.2.2.1 Incidence 21
1.2.2.2 Associations avec les expositions professionnelles 22

1.3 Objectifs et hypothèses 22

1.4 Plan 25

2 Outils méthodologiques 26

2.1 Mesures utilisées 26
2.1.1 Calcul des taux d'incidence 26
2.1.2 Modèle logistique 28
2.1.2.1 Variable à expliquer binaire 28
2.1.2.2 Variable à expliquer catégorielle 29
2.1.3 Modèle log-linéaire de Poisson 30
2.1.3.1 Généralités 30
2.1.3.2 Utilisation avec des données de survie 30
2.1.4 Fractions attribuables parmi les exposés 31

2.2 Directed Acyclic Graph (DAG) 32
2.2.1 Définitions et lien avec les données 32
2.2.2 Utilisation des DAGs 34

2.3 Gestion des données manquantes 35
2.3.1 Exemple introductif et notations 35
2.3.2 Caractérisation des mécanismes de données manquantes 38
2.3.2.1 Mécanismes MCAR, MAR et MNAR 38
2.3.2.1.1 Définitions et exemples 38
2.3.2.1.2 Implications pratiques 40
2.3.2.2 Mécanismes MAR séquentiels 41
2.3.2.3 « Recoverability » 42

2.3.2.3.1 Définition et exemples ... 42
2.3.2.3.2 Propriété de *recoverability* des mécanismes définis ... 43
2.3.2.3.3 La *recoverability* en pratique ... 44
2.3.3 Les mécanismes de données manquantes et les DAGs ... 45
2.3.3.1 Représentations ... 45
2.3.3.2 Utilisation des DAGs en pratique et premiers exemples ... 45
2.3.3.3 Construction d'un DAG utilisé dans les analyses ... 48
2.3.4 Les différentes méthodes d'estimation ... 52
2.3.4.1 Analyse sur cas complets avec pondérations ... 54
2.3.4.1.1 Principes généraux ... 54
2.3.4.1.2 Exemple ... 54
2.3.4.1.3 Conclusion ... 55
2.3.4.2 Les imputations multiples ... 57
2.3.4.2.1 Principes généraux ... 57
2.3.4.2.2 Modèle d'imputation gaussien ... 59
2.3.4.2.2.1 Principe ... 59
2.3.4.2.2.2 Exemple ... 59
2.3.4.2.2.3 Conclusion sur la méthode d'imputation gaussienne ... 60
2.3.4.2.3 Méthode d'imputation par « chained equation » (MICE) ... 61
2.3.4.2.3.1 Principe ... 61
2.3.4.2.3.2 Exemple ... 62
2.3.4.2.3.2.1 Exemple MAR ... 62
2.3.4.2.3.2.2 Exemple MAR séquentiel ... 63
2.3.4.2.3.3 Conclusion sur la méthode d'imputation MICE ... 63
2.3.4.3 Conclusion sur la gestion des données manquantes ... 64
2.3.5 Analyses de sensibilité pour l'hypothèse MAR ... 65
2.3.5.1 Modèles de sélection ... 66
2.3.5.2 Modèles « pattern-mixture » ... 67
2.3.5.2.1 Cas d'une variable gaussienne ... 67
2.3.5.2.2 Cas d'une variable binaire ... 68
2.3.5.2.3 Choix du paramètre δ ... 68
2.3.5.2.4 Cas de plusieurs variables manquantes ... 69
2.3.6 Conclusion générale sur la gestion des données manquantes ... 69

3 Matériel et méthodes ... 70

3.1 Cohorte des Salariés Ligériens (Cosali) ... 70
3.1.1 Phase initiale : Réseau expérimental de surveillance épidémiologique des troubles musculo-squelettiques dans les Pays de la Loire ... 70
3.1.1.1 Le réseau de médecins du travail ... 70
3.1.1.2 Les participants recrutés ... 71
3.1.1.3 L'auto-questionnaire initial ... 72
3.1.1.3.1 Facteurs de risque personnels ... 72
3.1.1.3.2 Les caractéristiques socio-professionnelles ... 72
3.1.1.3.3 Facteurs de risque professionnels ... 73
3.1.1.3.4 Les symptômes ... 74
3.1.1.4 L'examen clinique initial ... 74
3.1.2 Phase de suivi : Cosali ... 75
3.1.2.1 L'auto-questionnaire de suivi ... 75
3.1.2.2 L'examen clinique de suivi ... 77

3.2 Les populations d'étude sélectionnées ... 80

3.3 Facteurs de risque utilisés dans les analyses ... 81

3.4 TMS et douleurs ... 82

3.4.1 La prévalence des épicondylites et des douleurs de coudes ... 82
3.4.2 L'incidence des épicondylites ... 83
3.4.3 L'incidence des douleurs aux genoux ... 83

3.5 Choix d'analyses dans toutes les études ... 83

3.5.1 Analyses séparées hommes/femmes ... 83
3.5.2 Sélection des modèles multivariés ... 83
3.5.3 Logiciels utilisés ... 84

4 Etudes épidémiologiques et résultats ... 85

4.1 Prévalence des épicondylites latérales et des symptômes aux coudes ... 85

4.1.1 Analyses statistiques ... 85
4.1.2 Résultats ... 86
4.1.3 Discussion ... 93

4.2 Incidence des épicondylites latérales ... 95

4.2.1 Analyses statistiques ... 95
4.2.2 Gestion des données manquantes ... 97
4.2.2.1 Analyses sous l'hypothèse MAR séquentiel ... 98
4.2.2.1.1 Hypothèse MAR séquentiel ... 98
4.2.2.1.2 Les modèles d'imputations ... 100
4.2.2.2 Les données imputées ... 101
4.2.2.3 Analyses de sensibilité ... 102
4.2.2.3.1 Premiers scénarios d'analyses ... 102
4.2.2.3.1.1 Hypothèses MNAR ... 102
4.2.2.3.1.2 Prévalences imputées sous les hypothèses MNAR ... 103
4.2.2.3.2 Deuxièmes scénarios d'analyses ... 104
4.2.2.3.2.1 Hypothèses MNAR ... 104
4.2.2.3.2.2 Prévalences imputées sous les hypothèses MNAR ... 104
4.2.3 Résultats ... 105
4.2.3.1 Taux d'incidence ... 105
4.2.3.2 Distribution des expositions en fonction de la profession ... 105
4.2.3.3 Analyses préliminaires ... 105
4.2.3.4 Analyses principales (données complètes et données imputées MAR) ... 106
4.2.3.5 Analyses additionnelles (données imputées MNAR) ... 110
4.2.3.5.1 Premiers scénarios ... 110
4.2.3.5.2 Deuxièmes scénarios ... 111
4.2.4 Discussion ... 112

4.3 Incidence des douleurs aux genoux ... 114

4.3.1 Analyses statistiques ... 114
4.3.2 Résultats ... 115
4.3.2.1 Données manquantes et score de propension à répondre ... 115
4.3.2.2 Incidence des douleurs aux genoux ... 119
4.3.2.3 Les expositions en fonction de la catégorie socio-professionnelle ... 119
4.3.2.4 Les facteurs de risque d'incidence des douleurs aux genoux ... 119
4.3.2.5 Les fractions attribuables parmi les exposés ... 124

4.3.2.6 Les analyses sur des sous-populations 124
4.3.2.6.1 Participants sans douleurs initialement 124
4.3.2.6.2 Participants sans changement de poste 124
4.3.2.7 Analyses pondérées 125
4.3.3 Discussion 126

5 Discussion générale 129

5.1 Synthèse des résultats 129
5.1.1 Epicondylites latérales et symptômes aux coudes 129
5.1.2 Incidence des douleurs aux genoux 131

5.2 Limites 131

5.3 Applications de ce travail 133
5.3.1 Pour le chercheur 133
5.3.1.1 En santé publique 133
5.3.1.2 Sur les hypothèses de recherche épidémiologique 133
5.3.1.3 Sur les études avec des données manquantes 135
5.3.2 Pour le praticien 137

5.4 Perspectives 138
5.4.1 Evaluation des expositions professionnelles 138
5.4.2 Importance de la définition des cas 139
5.4.3 Etude des liens entre les facteurs psychosociaux, les facteurs biomécaniques et les TMS 140

6 Bibliographie 142

Annexe A Exemples de DAG avec plus de deux variables manquantes 152
Annexe B Procédure d'imputation pour un exemple de mécanisme MAR séquentiel 156
Annexe C Schéma causal complet pour l'analyse des épicondylites incidentes 158
Annexe D Analyses de sensibilité sur l'incidence des épicondylites – détails techniques 159
Annexe E Questionnaire Nordique 163

1 Contexte et objectifs

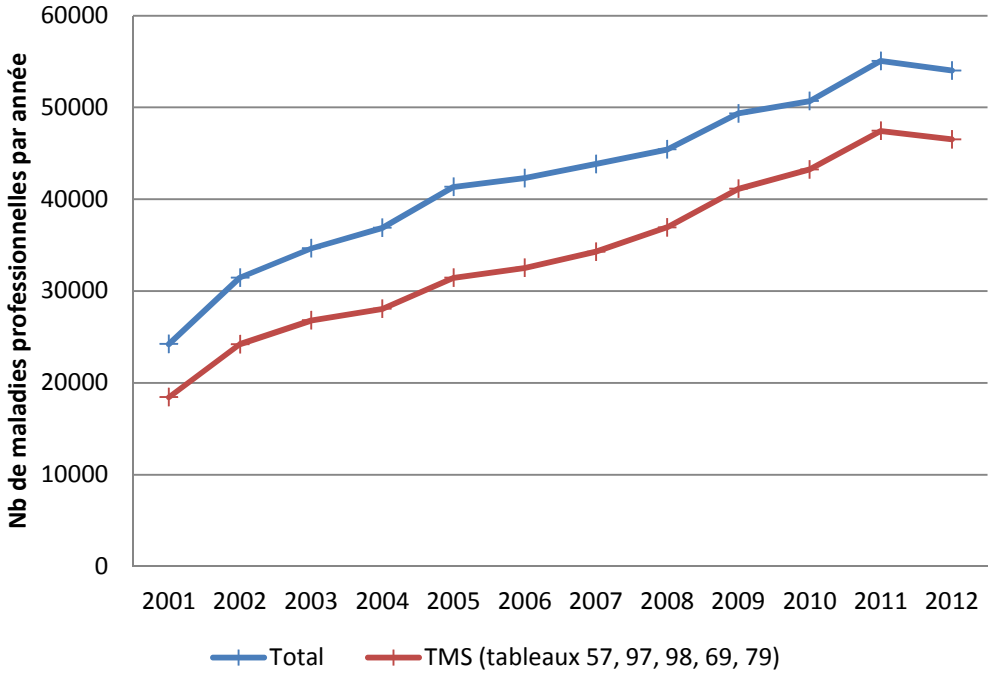
1.1 Contexte et généralités

1.1.1 TMS et douleurs articulaires

Les troubles musculo-squelettiques (TMS) regroupent des affections touchant les muscles, les tendons, les nerfs, les articulations (cartilages, ménisques...) et se traduisent principalement par des douleurs et des limitations fonctionnelles. Ils résultent d'un déséquilibre entre les capacités fonctionnelles des personnes et les sollicitations extérieures [1]. Ils peuvent apparaitre dans un contexte de travail, mais également dans d'autres contextes (sport de loisir...). Ce travail s'intéresse aux TMS d'origine professionnelle.

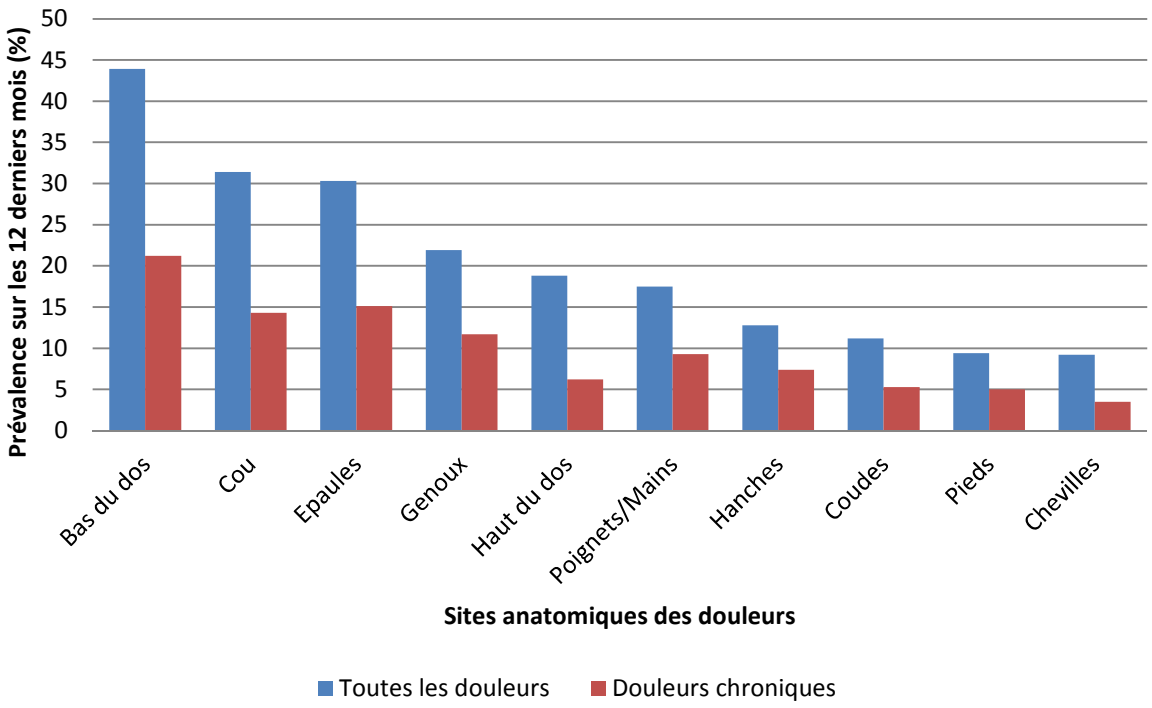
Ces pathologies peuvent être reconnues comme maladies professionnelles dans le cadre de tableaux annexés au code de Sécurité Sociale (tableaux 57, 69, 79, 97 et 98). Les TMS constituent ainsi la première cause de morbidité liée au travail, avec 46 538 affections péri-articulaires soit 86% des maladies professionnelles reconnues en 2012. Ils constituent également la première cause d'arrêts de travail pour incapacité temporaire, correspondant à 10 416 649 jours non travaillés (soit 96% des jours non travaillés pour incapacité temporaire). Le nombre de TMS reconnus comme maladies professionnelles augmente globalement chaque année (cf. Figure 1). Ces chiffres ne sont pas le reflet exact du nombre de TMS en France et sont probablement sous-estimés, puisque tous les TMS d'origine professionnelle ne sont pas reconnus en maladies professionnelles [2].

Figure 1 – Evolution du nombre de maladies professionnelles reconnues chaque année entre 2001 et 2012



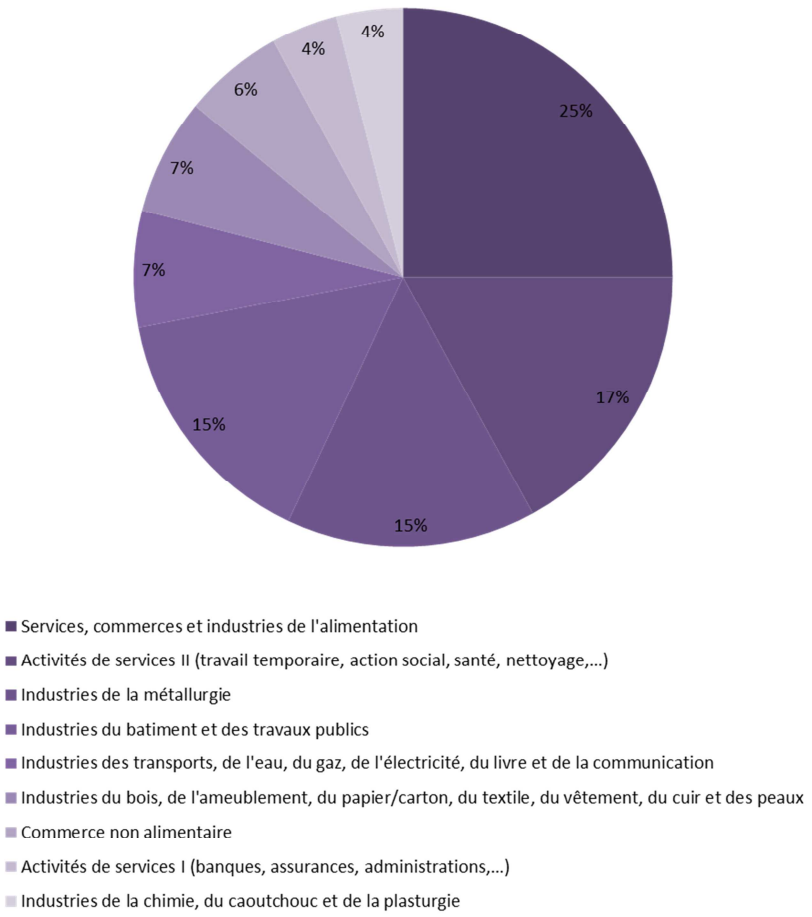
Un autre indicateur de la prévalence des TMS est la prévalence des douleurs articulaires en population générale. Selon une étude sur 15 pays Européens en 2006, 19% des individus déclarent des douleurs chroniques (douleurs intenses durant plus de 6 mois) et 15% en France [3]. Une étude réalisée en 2003 sur les douleurs chroniques déclarées au Pays-Bas indique les prévalences des douleurs par site anatomique (cf. Figure 2) et montre une forte prévalence des douleurs articulaires en particulier pour le bas du dos [4].

Figure 2 - Prévalence des douleurs articulaires sur les 12 derniers mois par site anatomique



Si on reprend les chiffres de 2012 sur les reconnaissances de maladies professionnelles par secteur d'activité, on voit que certains secteurs ont plus de TMS que d'autres : ainsi, l'industrie de l'alimentation, les activités de service (travail temporaire, action sociale, santé, nettoyage), l'industrie de la métallurgie et les industries du bâtiment et des travaux publics représentent 72% des TMS reconnus en 2012 (Figure 3). Ces secteurs semblent correspondre à certains métiers exposés à des contraintes fortes.

Figure 3 - Répartition des TMS reconnus en 2012 par secteur d'activité



1.1.2 Expositions professionnelles

Les facteurs professionnels correspondent à tout ce qui est lié aux tâches demandées au travail, mais également à l'environnement de travail. Les facteurs professionnels sont souvent classés en 3 composantes principales :

- Les facteurs dit « organisationnels » : ce sont les facteurs qui ont trait à la façon dont le travail est organisé, encadré et effectué. Ils concernent l'organisation structurelle du travail et sont donc une composante objective du travail [5,6].

  Exemples : nombre d'heures de travail, style d'encadrement, culture d'entreprise
- Les facteurs dit « psycho-sociaux » : ce sont les facteurs qui se réfèrent à la qualité de l'environnement de travail tel qu'il est perçu par les travailleurs. Ils sont donc une composante subjective du travail [6].

  Exemples : relations avec les collègues, stress perçu
- Les facteurs biomécaniques ou physiques : ce sont les facteurs qui portent sur les contraintes physiques au travail. Ce sont en principe des facteurs objectifs qui peuvent être mesurés à l'aide d'études ergonomiques mais ils sont en pratique souvent mesurés de manière déclarative [7].

  Exemples : mouvements répétés au niveau d'une articulation, port de charges

La distinction entre ces différentes composantes n'est pas toujours évidente en pratique, en particulier dans le cas de données uniquement déclarées par les travailleurs. En effet, certains facteurs déclarés peuvent relever de plusieurs composantes du travail à la fois. Par exemple, le travail répétitif peut relever de l'organisation à la chaine (composante organisationnelle), d'un jugement de valeur du travailleur sur son environnement de travail (composante psycho-sociale) ou de mouvements répétitifs (composante physique).

Pour être exposé aux facteurs professionnels, un individu doit être actif. Or, la population active est une population sélectionnée sur des critères d'âge, de qualification et d'état de santé. En effet, un individu peut être exclu du marché du travail pour des raisons de santé, à l'embauche et au cours de son parcours professionnel. A l'inverse, un individu qui reste en activité et longtemps exposé à une même exposition est probablement plus résistant à cette exposition que ceux exposés moins longtemps ou jamais exposés. Ce lien entre l'état de santé et les expositions observées peut potentiellement biaiser les relations étudiées dans le sens où il sous-estime l'effet des expositions professionnelles. Ce biais est appelé biais du travailleur sain (« healthy worker effect »). L'amplitude de ce biais est peu connue, mais il est possible que ce biais soit peu important lorsque les travailleurs en mauvaise santé continuent à être exposés (en cas d'impossibilité de reconversion par exemple).

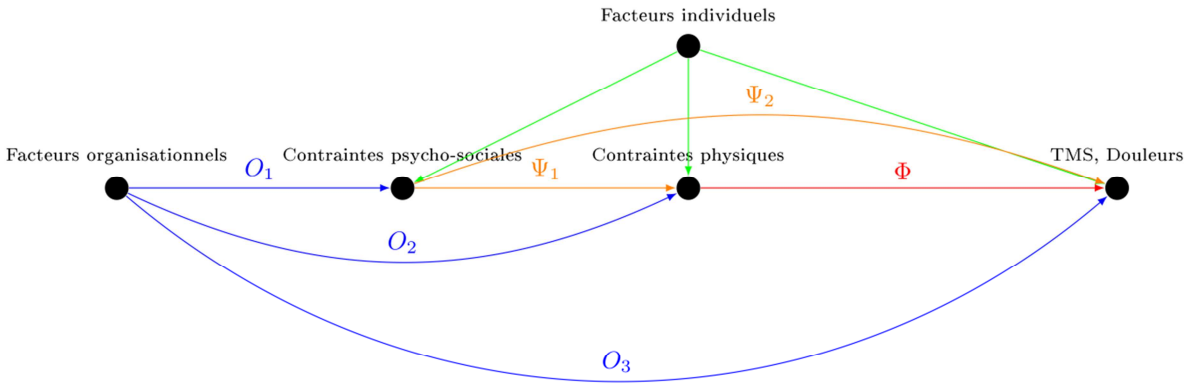
Pour conclure, une exposition professionnelle à un temps donné est la résultante du parcours professionnel et de l'état de santé d'un individu. Ainsi, il est essentiel de prendre en compte plusieurs mesures des expositions professionnelles au cours de la carrière ou à défaut, de discuter ce que peut signifier être exposé professionnellement ou non à un temps donné sur le parcours professionnel antérieur (cf. paragraphe 4.2.4, p 112).

1.1.3 Modèle conceptuel d'apparition des douleurs et des TMS

Les TMS et les douleurs articulaires ont initialement été étudiées en fonction des contraintes physiques professionnelles [8]. Le contexte d'organisation de l'entreprise et les contraintes psycho-sociales sont maintenant également prises en compte [9].

Un article de Marras en 2004 indique l'apport que pourraient avoir les modèles causaux pour la compréhension des TMS [10]. La Figure 4 propose un modèle conceptuel des relations causales entre les différentes composantes des facteurs de risque et les TMS.

Figure 4 - Schéma causal des liens entre les différentes contraintes professionnelles et les troubles musculo-squelettiques



Ce modèle s'inspire des précédents travaux menés sur les modèles conceptuels d'apparition des TMS en fonction des facteurs organisationnels, psycho-sociaux et biomécaniques [9,11–15], dont on peut trouver une synthèse pour les modèles antérieurs à 2002 [6]. Dans les paragraphes suivants, on détaille quels mécanismes expliquent les effets causaux illustrés sur la Figure 4.

Les facteurs individuels influencent les contraintes psycho-sociales et physiques et les TMS et douleurs. Les « facteurs individuels » correspondent à un découpage assez arbitraire. Ils peuvent inclure au minimum les facteurs génétiques, l'âge et le sexe. Certains mécanismes décrits dans les paragraphes suivants incluent eux-mêmes certains facteurs qui pourraient être décrits comme des facteurs individuels. Ils sont surtout figurés ici pour indiquer qu'il peut y avoir des phénomènes de confusion faisant intervenir des facteurs d'origine non professionnelle.

Les contraintes physiques causent les TMS et les douleurs, cette dernière relation ayant été particulièrement étudiée [8].

Les facteurs organisationnels causent certaines contraintes psycho-sociales et physiques (flèches O_1 et O_2) [6,12]. Par exemple, le contact avec le public peut impliquer une exposition à de la violence verbale (contrainte psycho-sociale) tandis que le travail à la chaine peut causer une tension musculaire importante (contrainte physique).

Les facteurs organisationnels peuvent causer directement des TMS et des douleurs articulaires (flèche O_3) par différents mécanismes :

- Les facteurs organisationnels peuvent impliquer une récupération insuffisante des tissus musculaires (par exemple, l'absence de temps de pause) qui peut causer les douleurs articulaires [9,15],
- Les facteurs organisationnels peuvent causer des comportements de santé délétères (absence d'exercice physique, alimentation anarchique...) qui eux-mêmes peuvent causer des douleurs articulaires [9,14],
- Les facteurs organisationnels peuvent impliquer une reconnaissance et une déclaration différente des symptômes musculo-squelettiques [12].

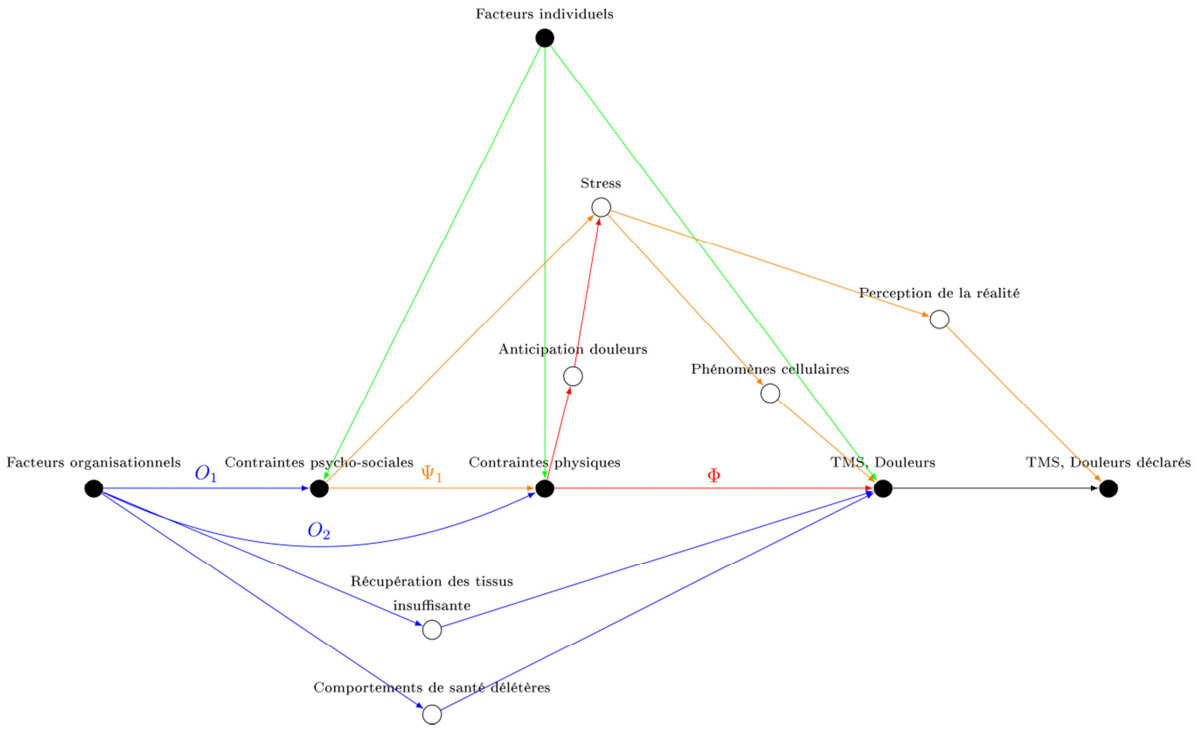
Les contraintes psycho-sociales peuvent causer ou accentuer les contraintes physiques (flèche Ψ_1) [11]. Par exemple, le faible soutien social des collègues peut se traduire par l'absence d'aide pour porter certaines charges et donc une charge physique plus importante. Globalement, les facteurs dits « psycho-sociaux » peuvent accroitre les contraintes physiques et donc la tension musculaire.

Les contraintes psycho-sociales peuvent aussi causer directement des TMS par plusieurs mécanismes (flèche Ψ_2) :

- Les contraintes psycho-sociales augmentent le niveau de stress et ce stress peut causer les TMS (on suppose un mécanisme hormonal ou par hyperventilation par exemple) [11,13],
- Les contraintes psycho-sociales augmentent le niveau de stress et ce stress peut engendrer des perceptions de la réalité et de la douleur différentes [11,12].

Le fait que l'exposition à des facteurs psycho-sociaux engendre la production de certaines hormones de stress a notamment été mis en évidence dans une étude de 2008 [16]. On peut résumer les mécanismes décrits avec la Figure 5 qui détaille les effets causaux de la Figure 4.

Figure 5 - Schéma causal des liens détaillés entre les différentes contraintes professionnelles et les troubles musculo-squelettiques



1.1.4 Forme du lien entre les facteurs professionnels et les TMS

Les schémas conceptuels précédents ne font aucune hypothèse sur la forme du lien causal entre les contraintes professionnelles et les TMS et les douleurs. C'est l'objet de ce paragraphe.

Les troubles musculo-squelettiques à un moment donné sont potentiellement le résultat des expositions professionnelles sur l'ensemble de la carrière. Plusieurs hypothèses peuvent être faites sur ce lien temporel :

- Un effet à court terme : seules les expositions professionnelles récentes ont un effet sur les TMS, ce qui pourrait se décrire simplement en « les articulations n'ont pas de mémoire »,
- Un effet à long terme : les expositions professionnelles récentes n'ont pas d'effet sur les TMS, ce sont les expositions anciennes qui jouent, il y a un temps de latence entre l'exposition et l'apparition TMS,
- Un effet cumulatif monotone : c'est la durée d'exposition sur toute la carrière qui joue sur les TMS,
- D'autres hypothèses plus ou moins complexes : des effets de seuils, des modifications d'effets selon les facteurs en présence…

L'hypothèse la plus fréquemment retenue pour les TMS est celle d'un effet cumulatif, sans période de latence [14,17]. Cependant, les hypothèses sur la forme du lien dépendent du TMS étudié,

puisque les TMS regroupent des troubles variés, avec des relations au temps qui peuvent être différentes. Par exemple, dans le cas des tendinites aux coudes, on aura tendance à considérer des effets à court terme puisque cette pathologie a une évolution rapide avec une guérison à court terme pour la majorité des cas. A l'inverse, dans le cas de l'arthrose aux genoux, on considérera plutôt des effets à long terme et des effets cumulatifs puisque c'est une pathologie à évolution lente et en principe irréversible. De la même façon, pour certains TMS, quand plusieurs facteurs de risque sont étudiés à la fois, il peut y avoir des sur-risques importants ou des modifications des risques associés au cumul de différents facteurs de risque professionnels [8].

1.2 Etat de la littérature

Les TMS sont largement étudiés dans la littérature [8,18]. A la lumière des critères proposés par Hill en 1965 pour évaluer si une association pouvait être considérée comme causale [19], les revues de la littérature concluent globalement à un faisceau de preuves raisonnable à insuffisant selon les expositions professionnelles et selon les pathologies [8,18]. Dans ce travail, on s'intéresse spécifiquement à deux articulations particulières : les coudes et les genoux (cf. Figure 2, p 12). Ces deux articulations sont intermédiaires (ni terminales, ni attachées au tronc) et les TMS associés ont des causes plutôt différentes. Quand l'origine des TMS aux coudes semble plutôt multifactorielle, les TMS aux genoux semblent causés par un facteur prépondérant, à savoir la charge qui porte sur l'articulation.

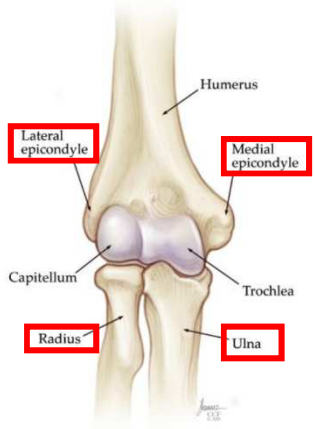
En particulier, pour les TMS au niveau des coudes, une revue de 2010 conclut à un niveau de preuves raisonnable pour les postures inconfortables, le travail répétitif et l'âge et un niveau de preuves limité pour l'indice de masse corporelle (IMC), le travail physiquement exigeant et la présence d'autres TMS des membres supérieurs [18]. Pour les TMS au niveau des genoux, la revue de 2010 conclut à un niveau de preuve raisonnable pour les postures inconfortables, le port de charges, le travail répétitif et à un niveau de preuves limité pour les facteurs psychologiques, le travail physiquement exigeant et l'IMC. Aucun facteur professionnel n'est reconnu à un niveau de preuve fort (c'est-à-dire qu'aucun ne présentait plus de 4 critères de causalité parmi les 5 suivants : (1) un risque relatif plus grand que 3 ; (2) plus d'une étude documente cette association ; (3) dont au moins une prospective ; (4) au moins une étude trouve une relation de dose-réponse ; et (5) le facteur étudié correspond à une hypothèse en cohérence avec la physiopathologie de la maladie).

1.2.1 Sur les symptômes aux coudes

De manière générale, il n'existe pas de consensus sur l'évaluation des troubles musculo-squelettiques du membre supérieur [20]. Un article de 2001 de Sluiter et al. proposait un ensemble de critères standardisés pour déterminer la présence de certains troubles musculo-squelettiques aux

membres supérieurs et en particulier ceux aux coudes [20]. Cependant, le débat reste ouvert sur ces critères, en particulier certains auteurs pensent que la définition des cas doit être adaptée à l'objectif de chaque étude [21].

Figure 6 - Anatomie du coude - vue antérieure des os, tiré de [22]



1.2.1.1 Prévalence

La prévalence des douleurs aux coudes est estimée à environ 10% en population générale [3,23]. Les douleurs aux coudes peuvent être localisées à différentes parties de l'articulation : au niveau des épicondyles latéraux et médiaux, du nerf ulnaire ou du nerf radial (cf. Figure 6). Selon la localisation et la nature de la douleur, les douleurs aux coudes peuvent correspondre à des épicondylites latérales ou médiales, des syndromes du nerf radial, des syndromes du nerf ulnaire ou des arthroses du coude.

Les syndromes du nerf radial (ou ulnaire) correspondent à la compression du nerf cubital (ou ulnaire) à la face postérieure et interne du coude (cf. Figure 6). Ils sont peu fréquents : 5% des cas de douleurs aux coudes dans une population à fortes contraintes professionnelles pour les syndromes du nerf radial, 3% des individus actifs avec une forte exposition aux facteurs de risque professionnels pour les syndromes du nerf ulnaire [24–27]. L'arthrose résulte de l'usure du cartilage des articulations mais est peu fréquente au niveau des coudes (20–23).

Les épicondylites sont une inflammation ou une irritation de l'insertion du tendon [28].

Les épicondylites médiales sont plus fréquemment observées que les pathologies précédentes (prévalence de 0,4% en population générale) et moins fréquemment observées que les épicondylites latérales [25,27,29,30]. Elles sont connues sous le nom « golfer elbow » car elles sont fréquentes chez les joueurs de golf.

La prévalence des épicondylites latérales est estimée entre 0,7 et 4,0% en population générale [27,31]. Cette prévalence parmi les populations au travail avec un travail répétitif monte entre 1,3 et 12,2% [25,26,32–37]. Elles sont connues sous le nom « tennis elbow » car elles sont fréquentes chez les joueurs de tennis.

1.2.1.2 Incidence

Pour des raisons de faibles fréquences d'apparition, les études sur les symptômes aux coudes autres que les épicondylites sont quasi-inexistantes dans la littérature [24].

Peu d'études sur l'incidence des épicondylites latérales sont disponibles dans la littérature [30,38–43]. Elle est estimée dans plusieurs contextes : en consultation de médecine générale [30,43], sur la base des demandes d'indemnisation [41], à l'hôpital [39] et dans des secteurs d'activité à risque [38,40,42].

Le taux d'incidence est estimé entre 0,3 et 1,1 pour 100 patients-années en médecine générale et entre 2,0 et 11,3 pour 100 travailleurs-années dans les secteurs d'activités à risque.

1.2.1.3 Associations avec les facteurs professionnels

Des études transversales ont exploré le lien entre facteurs biomécaniques professionnels et épicondylites [23,25,34–37,44–46]. Peu d'études reposant sur des données longitudinales sont disponibles [38,40]. Plusieurs revues de la littérature font l'état des lieux de ces associations [26,27,47]. Les facteurs de risque biomécaniques suivants sont mis en évidence comme résumé dans l'article de 2009 de Van Rijn et al. :

- la force : deux études transversales trouvent une association avec la manipulation de charges [25,37],
- la répétition : quatre études transversales trouvent une association avec les gestes répétitifs [25,36,37] en particulier avec des gestes répétitifs de la main et du bras [36],
- les vibrations : les résultats sont contrastés sur les vibrations, la plupart des études ne trouvent pas d'association [25], mais certaines études trouvent une association chez les hommes [37],
- les postures et les gestes : les épicondylites latérales sont associées à certaines postures, en particulier le facteur « tourner comme pour visser » dans une étude longitudinale [40], les facteurs « bras en l'air à l'avant », « flexion ou torsion des poignets » et « mouvements précis durant une partie de la journée de travail » dans une étude transversale [37].

L'effet de la combinaison de plusieurs facteurs biomécaniques est peu étudié, certains sur-risques ont été trouvés pour la combinaison de facteurs de force, de répétition et de posture [25,34,37]. Les

combinaisons impliquant spécifiquement des mouvements des coudes n'ont jamais été étudiées [26].

Parmi les facteurs psycho-sociaux, un faible soutien social au travail et une faible latitude décisionnelle sont associés aux épicondylites latérales [37]. D'autres facteurs psychologiques comme la détresse psychologique ou les symptômes dépressifs sont associés aux épicondylites latérales [23,40].

Une revue de la littérature en 2009 conclut sur le manque d'études longitudinales sur les épicondylites latérales qui permettraient de renforcer le faisceau de preuves sur les facteurs cités dans ce paragraphe [26]. Elle soulève également le problème des évaluations hétérogènes de l'existence d'épicondylites (par auto-évaluation, examen clinique ou examen électrophysiologique). Ce problème est d'autant plus crucial que les résultats peuvent être différents en fonction du type d'évaluation choisi.

1.2.2 Sur les symptômes aux genoux

Traditionnellement, les symptômes aux genoux étaient étudiés via l'arthrose aux genoux diagnostiquée par des critères radiographiques [48]. Cependant, il s'avère que les critères radiographiques ne sont pas fortement associés aux symptômes cliniques aux genoux [49,50]. Considérer les douleurs aux genoux revient à considérer de nombreux diagnostics possibles – arthrose, blessure, hygroma ou douleur irradiant d'autres articulations. Cependant, les douleurs aux genoux sont des déterminants importants de limitations fonctionnelles ultérieures, meilleurs que les critères radiographiques [3,51,52]. Les douleurs aux genoux peuvent également être associées à l'arthrose aux genoux. D'ailleurs, parmi les personnes de plus de 50 ans, la cause la plus fréquente des douleurs aux genoux est l'arthrose, cette cause étant de plus en plus fréquente avec l'âge [50].

1.2.2.1 Incidence

On trouve souvent dans la littérature une gradation des douleurs en fonction de leur durée (plus ou moins 30 jours ; ou plus ou moins 3 mois dans l'année qui précède) ou de leur intensité (intensité faible, modérée ou forte du dernier épisode de douleurs) dans l'objectif d'affiner la définition des symptômes et d'isoler les cas les plus graves et les plus susceptibles de persistance ou de récidives [4,53–55].

L'incidence des douleurs aux genoux a été étudiée en consultations de médecine générale, dans des secteurs d'activité spécifiques et parmi des travailleurs nouvellement employés [55–58]. Elle est estimée entre 3 et 10 cas incidents pour 100 personnes-années. Toutes les études précédemment citées portent sur des populations avec des professions manuelles, par exemple des pompiers, des

constructeurs de bateaux, des travailleurs forestiers ou des travailleurs de la santé [55,56,58] ou sur des populations de plus de 40 ans [57].

1.2.2.2 Associations avec les expositions professionnelles

Les douleurs aux genoux sont associées à l'âge, à des susceptibilités génétiques, à l'indice de masse corporelle (IMC) et aux antécédents de blessures au niveau des genoux [59,60].

De nombreuses publications sont disponibles sur le lien entre les douleurs ou l'arthrose aux genoux et les expositions professionnelles [48,61–63]. Cette littérature abondante a donné lieu à un grand nombre de revues de la littérature et de méta-analyses [64–70].

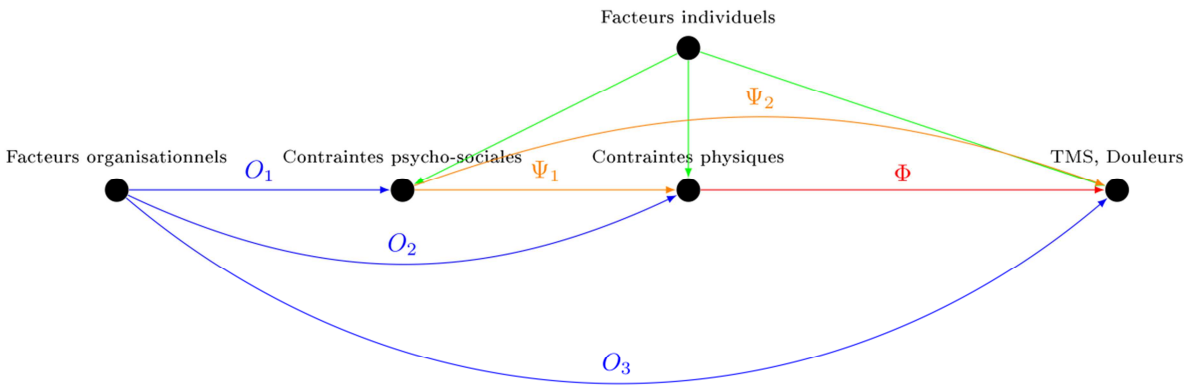
Les facteurs de risque mis en évidence dans de précédentes études portent sur des notions de gestes et de postures. Le port fréquent de charges lourdes est associé aux douleurs ou à l'arthrose des genoux dans de nombreuses études transversales [62,63,71,72], mais n'est pas retrouvé dans les études longitudinales [48]. Le fait de s'agenouiller (ou d'être accroupi) est associé aux douleurs ou à l'arthrose des genoux [56,62,72–74]. Monter des escaliers ou des échelles est aussi associé aux douleurs aux genoux dans quelques études transversales [62,71], de même que se tenir debout [63].

Les revues de la littérature concluent à un faisceau de preuves limité à modéré de ces relations (avec un niveau de preuves plus faible pour les femmes) puisqu'elles ne sont pas retrouvées dans toutes les études, en particulier dans les études longitudinales. Selon plusieurs auteurs, d'autres études prospectives sont nécessaires pour améliorer les connaissances sur le lien entre les facteurs professionnels et les douleurs aux genoux, en particulier pour les femmes [66,67,70]. En effet, malgré un grand nombre d'articles avec une bonne qualité méthodologique, il n'y a pas d'étude prospective récente sur l'incidence des douleurs aux genoux et son lien avec les expositions professionnelles dans la population active [70].

1.3 Objectifs et hypothèses

L'objectif de ce travail de thèse était d'étudier l'association entre les facteurs professionnels et les troubles musculo-squelettiques au niveau des coudes et des genoux. Cela sera fait au travers de 3 études différentes. Elles s'intéressaient principalement à la relation causale entre les facteurs physiques professionnels et les TMS et les douleurs des articulations (flèche Φ, Figure 4). Les facteurs psycho-sociaux étaient vus comme des facteurs de confusion de la relation avec les facteurs physiques.

Figure 4 (duplicata) - Schéma causal des liens entre les différentes contraintes professionnelles et les troubles musculo-squelettiques



Pour des raisons de fréquence d'apparition (pour les syndromes du nerf radial, les syndromes du nerf ulnaire et l'arthrose aux coudes) et de disponibilité des données (pour les épicondylites médiales), certaines pathologies n'étaient pas spécifiquement étudiées dans ce travail et ont été regroupées dans une catégorie composite de « symptômes aux coudes ». Les épicondylites latérales étaient étudiées plus précisément. Les troubles musculo-squelettiques au niveau des coudes étaient décrits dans deux études : une sur la prévalence des symptômes aux coudes et des épicondylites latérales, et une autre sur l'incidence des épicondylites latérales.

Pour l'étude des épicondylites et des symptômes aux coudes en général, une attention particulière a été portée aux expositions professionnelles biomécaniques suivantes : l'effort physique important au travail, les tâches répétitives, les mouvements répétés de flexion et extension des coudes et de torsion des poignets.

Dans l'analyse sur la prévalence des symptômes aux coudes et des épicondylites, les associations avec les symptômes seuls et avec les épicondylites ont été comparées afin d'évaluer la sensibilité des résultats à des définitions différentes des troubles articulaires des coudes [21]. On a supposé que les expositions professionnelles étudiées étaient associées à la présence de symptômes et d'épicondylites, avec une association plus forte pour les épicondylites. L'hypothèse sous-jacente était que d'une part, les symptômes aux coudes précédaient les épicondylites latérales et d'autre part, que plus la durée d'exposition était importante, plus les travailleurs pouvaient être susceptibles de développer des épicondylites au lieu de symptômes seuls.

Pour des raisons d'effectif, l'analyse sur l'incidence des épicondylites regroupait une définition plus large des cas (les cas symptomatiques non avérés sont inclus) et regardait l'association avec la première mesure de ces expositions professionnelles puis avec les deux mesures successives de ces expositions professionnelles. On a supposé que les expositions professionnelles avaient un effet

cumulé sur l'incidence des épicondylites, en particulier qu'une exposition de longue durée avait un effet plus important. L'hypothèse sous-jacente était que, selon leur durée d'exposition plus ou moins longue, les travailleurs étaient plus ou moins à risque de développer des épicondylites.

Une troisième étude a porté sur l'incidence des douleurs aux genoux. Les expositions professionnelles considérées étaient : l'effort physique important au travail, la manipulation récurrente de charges lourdes, le fait de s'agenouiller de manière répétée, et être assis. Les symptômes aux genoux ont été séparés en douleurs de longue durée (plus de 30 jours dans les 12 derniers mois) et de courte durée (1 à 29 jours dans les 12 derniers mois) afin de regrouper les cas les plus graves. On a supposé que les expositions professionnelles considérées étaient associées à l'incidence des douleurs aux genoux et plus particulièrement aux épisodes longs, même après la prise en compte des facteurs personnels importants comme l'âge ou l'IMC.

1.4 Plan

Ce travail a nécessité l'utilisation de méthodologies particulières, notamment les diagrammes causaux et des méthodes pour gérer les données manquantes. La partie 2 précise les principaux outils méthodologiques utilisés dans ce travail. Cette partie n'est pas indispensable à la compréhension des résultats épidémiologiques, mais permet d'approfondir la théorie associée aux méthodes statistiques utilisées. Elle est rédigée dans un objectif opérationnel et servira de référence pour justifier les choix méthodologiques et les interprétations des résultats dans les parties suivantes.

La partie 3 décrit la population utilisée dans ces études et les éléments communs aux différentes analyses statistiques effectuées.

La partie 4 décrit séparément chacune des trois études présentées avec pour chacune : les analyses statistiques spécifiques, la gestion des données manquantes (si pertinent), les résultats et la discussion de ces résultats.

La partie 5 donne une synthèse des résultats avec leurs forces et leurs limites. Elle précise les applications et les perspectives ouvertes par ce travail, tant d'un point de vue méthodologique qu'au niveau épidémiologique et pratique.

2 Outils méthodologiques

Dans cette partie, les outils méthodologiques utilisés dans les analyses ultérieures sont présentés.

Une première partie indique brièvement le principe et l'interprétation des modèles utilisés dans les analyses. Une deuxième partie détaille les « Directed Acyclic Graphs » (DAGs) utilisés dans ce travail de thèse pour illustrer les hypothèses. La troisième partie développe les méthodologies disponibles pour les analyses dans les cas où des données sont manquantes, et justifie les choix effectués dans les analyses sur l'incidence.

De manière générique dans cette partie, on écrira $P(X)$ pour parler de la distribution de la variable aléatoire X. Dans le cas où X est binaire, $P(X)$ représente la probabilité que $X = 1$ ou que $X = 0$. Dans le cas où X est continue, $P(X)$ représente la fonction densité de X, même s'il s'agit d'un abus de notation.

2.1 Mesures utilisées

Les analyses concernent des variables à expliquer catégorielles. Les paragraphes suivants sont présentés dans un objectif opérationnel. Pour une description plus complète de ces modèles classiquement utilisés en épidémiologie, nous renvoyons le lecteur au livre d'Agresti sur l'analyse des variables catégorielles [75].

2.1.1 Calcul des taux d'incidence

Définition de l'incidence [76]

> *L'incidence est une « mesure de la survenue de nouveaux cas. Elle recouvre les notions de risque, ou incidence cumulée, et de taux d'incidence. Le risque, pour une période de temps donnée, est la probabilité qu'a un sujet indemne au début de la période d'être atteint par la maladie au cours de la période. Le taux d'incidence est une mesure de la fréquence de survenue de nouveaux cas par unité de temps. »*

La notion d'incidence est particulièrement adaptée aux évènements dont on connait précisément la date de survenue et qui n'ont pas lieu plusieurs fois pour une même personne. Dans le cas d'évènements rares, le risque et le taux d'incidence sont très proches sur une même période d'observation.

Calcul des taux d'incidence [77]

> *« Pour calculer un taux d'incidence (TI), nous devons suivre une cohorte de sujets, compter le nombre de nouveaux cas (ou cas incidents), I, d'une maladie dans cette cohorte et calculer le temps total durant lequel les individus indemnes sont observés sur la période (ou personnes-temps), PT. » On obtient alors un taux d'incidence moyen sur la période.*
>
> $$TI = \frac{I}{PT} = \frac{Nb\ de\ cas\ incidents}{\sum_{personnes\ a\ risque} Temps\ de\ suivi}$$

L'unité de personnes-temps choisie est très souvent le nombre de personnes-années.

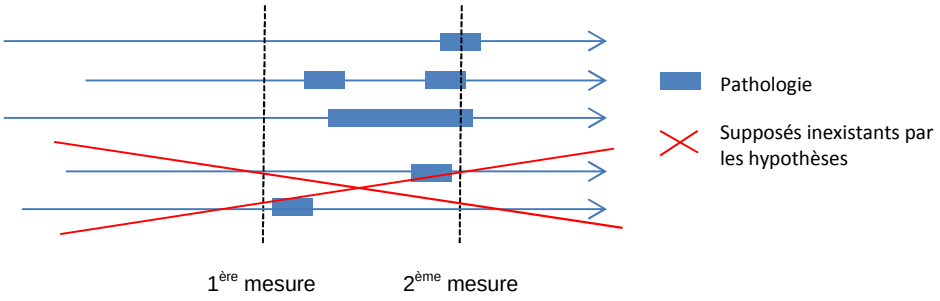
Dans ce travail, on dispose uniquement de deux temps de mesure. Il a été décidé d'approximer le taux d'incidence par :

$$TI = \frac{Nb\ de\ personnes\ sans\ la\ pathologie\ initialement\ et\ avec\ la\ pathologie\ sur\ le\ suivi}{\sum_{personnes\ sans\ la\ pathologie\ initialement} Duree\ entre\ les\ 2\ mesures}$$

Cette approximation est valable sous les hypothèses suivantes :

- Tous les cas incidents entre les deux mesures sont encore malades à la deuxième mesure

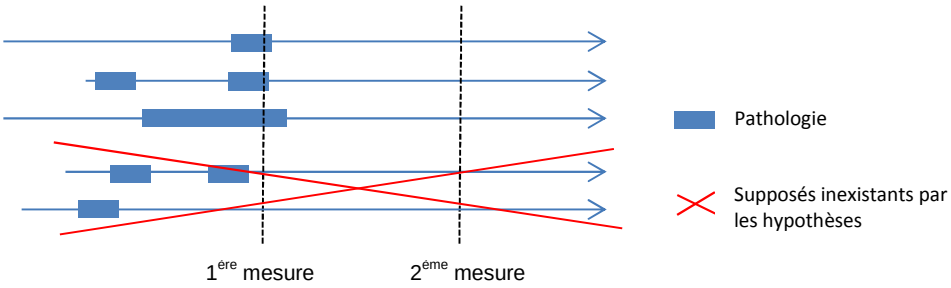
Figure 7 - Exemples d'histoires naturelles de pathologie (1)



Si cette hypothèse n'est pas vérifiée, elle conduit à une sous-estimation du taux d'incidence

- Tous les cas ayant déjà eu la pathologie à la première mesure ou avant sont exclus des analyses. Ainsi, on suppose que ceux pour lesquels la pathologie n'est pas présente à la première mesure n'ont pas eu d'épisodes antérieurs de cette pathologie.

Figure 8 - Exemples d'histoires naturelles de pathologie (2)



Ces hypothèses ne sont probablement pas toujours vérifiées en pratique puisque certains TMS sont réversibles et que les temps de mesures seront indépendants de la présence ou non de la pathologie ou des épisodes antérieurs de celle-ci.

On remarque que la formule précédente porte au dénominateur sur le temps écoulé entre les deux mesures et non le temps de suivi en étant « indemne » de la pathologie. Cette approximation aura peu de conséquences dans ce travail puisque les pathologies étudiées sont peu fréquentes et donc le

choix du temps de survenue de la pathologie aura peu d'impact sur le calcul du nombre de personnes-années total (la plupart des individus étant indemnes à la fin du suivi).

Définition des rapports d'incidence

> *On définit les rapports d'incidence comme le rapport entre le taux d'incidence dans une catégorie fixée et le taux d'incidence dans une catégorie de référence. On écrit ce rapport de la manière suivante :*
>
> $$IRR_j = \frac{TI_{catégorie\ j}}{TI_{catégorie\ référence}}$$

2.1.2 Modèle logistique

2.1.2.1 Variable à expliquer binaire

Soit $(Y_i)_{i=1\ldots n}$ une variable binaire à expliquer pour les unités i, $Y_i \in \{0,1\}$. Soit $\left(X_{ij}\right)_{i=1\ldots n, j=1\ldots, k}$ un ensemble de k variables explicatives pour les unités i.

Définition du modèle logistique

> *On définit le modèle logistique par la relation suivante en fonction du vecteur de paramètres* $\alpha = \left(\alpha_j\right)_{j=1\ldots k}$ *de taille k:*
>
> - Y_i *suit une loi de Bernoulli de paramètre* π_i *c'est-à-dire* $P(Y_i = 1|X_{i1}\ldots, X_{ik}) = \pi_i$
> - $\forall i = 1\ldots, n\ \ log\frac{P(Y_i=1|X_{i1}\ldots,X_{ik})}{P(Y_i=0|X_{i1}\ldots,X_{ik})} = logit(\pi_i) = \sum_{j=1}^{k} \alpha_j X_{ij}$

Des algorithmes classiques de maximisation de la vraisemblance peuvent être utilisés pour estimer α [75].

Définition des odds-ratios ajustés

> *On définit les odds-ratios (ou rapport des cotes) par la multiplication sur les odds (ou les cotes) si le facteur* X_j *est augmenté d'une unité avec les autres facteurs fixés à des valeurs données :*
>
> $$\forall j = 1\ldots, k$$
>
> $$OR_j = \frac{P(Y_i=1|X_{i1}=x_1\ldots,X_{ij}=x_j+1\ldots,X_{ik}=x_k)}{P(Y_i=0|X_{i1}=x_1\ldots,X_{ij}=x_j+1\ldots,X_{ik}=x_k)} \Big/ \frac{P(Y_i=1|X_{i1}=x_1\ldots,X_{ij}=x_j\ldots,X_{ik}=x_k)}{P(Y_i=0|X_{i1}=x_1\ldots,X_{ij}=x_j\ldots,X_{ik}=x_k)}$$

Définition des risques relatifs ajustés

> *On définit les risques relatifs comme la multiplication de la probabilité d'avoir l'évènement d'intérêt si le facteur X_j est augmenté d'une unité avec les autres facteurs fixés à des valeurs données :*
>
> $$RR_j = \frac{P\left(Y_i = 1 \middle| X_{i1} = x_1 \ldots, X_{ij} = x_j + 1 \ldots, X_{ik} = x_k\right)}{P\left(Y_i = 1 \middle| X_{i1} = x_1 \ldots, X_{ij} = x_j \ldots, X_{ik} = x_k\right)}$$

Dans le cas d'un évènement rare, on a $P(Y_i = 0) \cong 1$ et donc $OR_j \cong RR_j$ [75].

Interprétation des coefficients :

Les estimations $\left(\alpha_j\right)_{j=1\ldots k}$ du modèle logistique sont souvent exprimées en terme d'odds-ratio, puisque $OR_j = e^{\alpha_j}$.

2.1.2.2 Variable à expliquer catégorielle

Soit $(Y_i)_{i=1\ldots n}$ une variable catégorielle à expliquer pour les unités i, $Y_i \in \{0 \ldots, L\}$.

Définition du modèle logistique multinomial

> *On a la relation suivante pour un modèle logistique multinomial avec pour référence la catégorie 0*
>
> $$\forall l \in \{1 \ldots, L\} \ log \frac{P(Y_i = l | X_{i1} \ldots, X_{ik})}{P(Y_i = 0 | X_{i1} \ldots, X_{ik})} = \sum_{j=1}^{k} \alpha_{jl} X_{ij}$$

Les calculs et l'algorithme de maximisation du cas binaire se généralisent facilement au cas catégoriel. L'interprétation des coefficients est identique avec un odd-ratio par catégorie d'évènement d'intérêt.

2.1.3 Modèle log-linéaire de Poisson

2.1.3.1 Généralités

Le modèle log-linéaire de Poisson est utilisé fréquemment pour des variables de comptage. Il sera en particulier utilisé dans les analyses sur des taux, c'est-à-dire une variable de comptage rapportée à un index de temps.

Soit $(Y_i)_{i=1\ldots n}$ des variables de comptage à expliquer pour les unités i ($\forall i, Y_i \in \{0,1,2,3\ldots\}$)

Soit $(T_i)_{i=1\ldots n}$ des index de temps correspondants

<u>Définition du modèle de Poisson</u>

> *On définit le modèle de Poisson par les deux hypothèses suivantes :*
> • *Y_i suit une loi de Poisson de paramètre λ_i c'est-à-dire $P(Y_i = k) = e^{-\lambda_i}\frac{\lambda_i^k}{k!}$*
> • $log\big(E(Y_i|X_{i1}\ldots,X_{ik})\big) = log(\lambda_i) = \sum_{j=1}^{k}\alpha_j X_{ij} + log(T_i)$
> *T_i est désigné sous le terme d'« offset ».*

Une restriction importante de ce modèle est qu'il impose $E(Y_i) = Var(Y_i) = \lambda$, ce qui peut ne pas être vérifié en pratique. C'est pour cette raison qu'on introduit souvent une contrainte de sur- ou de sous-dispersion afin de corriger les variances : $Var(Y_i) = \phi E(Y_i)$. Ce paramètre est estimé et permet de corriger les variances des estimateurs initiaux.

2.1.3.2 Utilisation avec des données de survie

Soit Y_i une variable binaire qui indique la survenue d'un évènement au temps T_i.

On peut montrer qu'utiliser un modèle de Poisson dans ce cadre revient à utiliser un modèle de survie avec un risque instantané μ_i constant au cours du temps [75]. Dans ce cas, le risque instantané μ_i est aussi égal au taux d'incidence .Ainsi, on a $\lambda_i = \mu_i * T_i$ et donc $\log(\mu_i) = \sum_{j=1}^{k}\alpha_j X_{ij}$.

<u>Interprétation des coefficients :</u>

Les estimations $(\alpha_j)_{j=1\ldots k}$ sont souvent exprimées en termes de rapport d'incidence puisque $IRR_j = e^{\alpha_j}$.

2.1.4 Fractions attribuables parmi les exposés

Définition des fractions attribuables parmi les exposés

> *On définit la fraction attribuable à une exposition (AF) comme la proportion de cas qui n'auraient pas eu lieu en l'absence de cette exposition [78,79].*
> *Quand l'exposition est binaire, l'AF parmi les exposés se calcule de la façon suivante :*
> $$AF = \frac{I_1 - I_0}{I_1} = \frac{RR - 1}{RR}$$
> *Où I_1 et I_0 sont respectivement les incidences parmi les exposés et les non exposés, et RR le risque relatif brut associé à l'exposition.*

Pour la fraction attribuable parmi les exposés, on peut utiliser les risques relatifs ajustés [78]. Il est également possible de remplacer les risques relatifs par des odds-ratios quand les cas sont rares.

Interprétation :

L'AF parmi les exposés s'interprète toujours comme l'écart de l'incidence chez les exposés et les non exposés rapporté à l'incidence chez les exposés. L'AF peut, dans certains cas précis, s'interpréter comme l'impact potentiel que peut avoir une intervention sur une proportion de cas [80]. Par exemple, l'exposition aux cancérogènes comme l'amiante peut être contrôlée par la législation et une intervention qui réduirait l'exposition à néant est tout à fait envisageable. Cependant, d'autres types d'expositions professionnelles comme le travail répétitif pour un ouvrier travaillant à la chaîne ou l'exposition aux fumées pour un pompier laissent peu d'espoir quant à la possibilité d'une intervention (l'exposition pouvant être intrinsèque au métier considéré et de fait non contrôlable).

2.2 Directed Acyclic Graph (DAG)

Ce paragraphe introduit un outil graphique, les « Directed Acyclic Graphs » (DAGs), qui permet de représenter les liens causaux entre les différents éléments considérés [81]. De manière générale, les DAGs ont pour objectif de présenter le cadre théorique de recherche et de clarifier les hypothèses sous-jacentes posées avant les analyses. Ils ont d'ailleurs été considérés dans la partie d'introduction (cf. paragraphe 1.1.3, Figure 4-Figure 5, p 15-17).

2.2.1 Définitions et lien avec les données

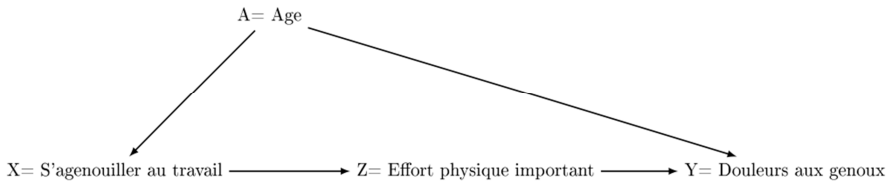
Définition

> *Les DAGs sont composés de nœuds (variables d'intérêt) et de flèches (effets causaux). Ils obéissent à des critères simples :*
>
> • *« directed » : une flèche orientée indique une relation de cause à effet*
>
> *On ne doit pas faire figurer des doubles flèches dans le graphe*
>
> • *« acyclic » : les relations cycliques ne sont pas autorisées*
>
> *La notion de temporalité (une cause précède un effet) interdit les cycles*
>
> *Un DAG contient toutes les variables qui sont causes communes de plusieurs variables du graphe.*

Exemple de DAG

Un exemple simple de DAG en Figure 9 indique que A, X, Z et Y sont des nœuds et que A cause X et Y ; X cause Z ; Z cause Y.

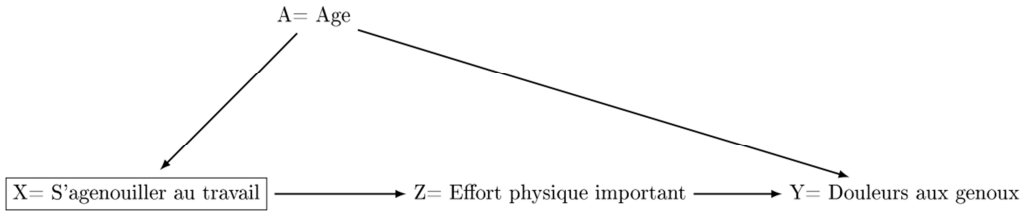
Figure 9 - Exemple de DAG simple



Le DAG permet de visualiser les associations entre les variables puisque quand il y a un ensemble de traits entre deux variables (hors conséquences communes), alors ces variables sont associées. Dans l'exemple de la Figure 9, A et Z sont associés par X.

Dans un DAG, on parlera de « conditionnement » sur une variable de manière générale pour indiquer qu'on prend en compte une variable. Ce conditionnement est figuré en encadrant la variable prise en compte (cf. Figure 10).

Figure 10 - Exemple de DAG simple avec conditionnement



En pratique, ce « conditionnement » sur une variable peut indiquer différentes méthodes, comme la stratification (on se place à une valeur donnée de cette variable), la pondération (on pondère les individus par leur probabilité d'avoir leur valeur observée pour cette variable) ou la standardisation (on développe les éléments étudiés pour toutes les valeurs prises par cette variable). Ces notions seront utiles dans la suite pour indiquer sur quelles variables il est nécessaire de « conditionner ».

Définition préalable : Indépendance entre deux variables

> *On dit que deux variables aléatoires X et Y sont indépendantes si et seulement si $P(X,Y) = P(X)P(Y)$.*
> *On écrira de manière équivalente :*
> $$X \text{ et } Y \text{ sont indépendantes} \leftrightarrow X \coprod Y \leftrightarrow Y \coprod X \leftrightarrow P(Y|X) = P(Y)$$

Ces définitions et ces notations pourront aussi être utilisées conditionnellement à une troisième variable Z : $X \coprod Y \,|Z \leftrightarrow \mathrm{Y} \coprod \mathrm{X} \,|Z \leftrightarrow \mathrm{P}(Y|X,Z) = P(Y|Z)$.

On utilisera dans la suite la propriété suivante : si R est une variable binaire et si $Y \coprod R \,|Z$, alors $\mathrm{P}(Y|Z) = P(Y|Z, R = 1)$.

Propriété : Indépendance entre deux variables dans un DAG [81]

> *Deux variables aléatoires X et Y sont indépendantes si :*
> *(0) Aucune flèche ne relie X et Y ou*
> *(1) Les flèches qui relient X et Y contiennent au moins une conséquence commune sur laquelle on n'a pas conditionné*
> *(2) Les flèches qui relient X et Y ne contiennent aucune conséquences communes et on peut conditionner sur une variable du chemin*
> *(3) On a des flèches comme décrites en (1) et des flèches comme décrites en (2)*

Exemple de DAG (suite)

Dans cet exemple, il y a deux flèches qui relient A et Z : celle qui passe par X et celle qui passe par Y (conséquence commune). D'après la propriété précédente, on a donc A et Z qui sont indépendants conditionnellement à X (soit $A \coprod Z|X$) avec la condition (3) : la flèche qui passe par X correspond à

la condition (2) puisqu'on conditionne sur X et la flèche qui passe par Y correspond à la condition (1) puisqu'il y a une conséquence commune sur ce chemin (ici, Y).

L'idée générale est donc que les DAGs permettent d'évaluer les relations de dépendances et d'indépendances entre les variables. Ces éléments sont assez généraux et visent uniquement à la compréhension des éléments de la thèse, des articles plus approfondis sur ce sujet existent dans la littérature [81].

2.2.2 Utilisation des DAGs

Les DAGs ne sont pas construits grâce aux données disponibles, mais grâce aux hypothèses de recherche *a priori*. Ainsi, certaines relations d'intérêt causales (ou « flèches ») du DAG vont être testées dans les analyses en supposant que les autres relations causales du DAG sont exactes. Les DAGs n'ont pas vocation à décrire la modélisation précise des relations entre les variables (du type interactions, fonction de lien...).

Les DAGs seront créés ici en considérant tout d'abord les variables intervenant dans les relations d'intérêt, puis en ajoutant les variables liées à au moins deux variables du DAG (par exemple, les facteurs de confusion ou les facteurs intermédiaires).

Les DAGs sont utilisés pour faciliter le protocole d'une étude, pour sélectionner les variables d'ajustement [82,83] ou pour classifier les sources de biais [84,85].

Dans ce travail, les DAGs sont utilisés en deux temps :

- dans un premier temps, pour figurer les hypothèses retenues et justifier les variables d'ajustement utilisées,
- dans un deuxième temps, pour indiquer un DAG « simplifié » cohérent avec les résultats trouvés (où seules les relations confirmées dans ces études sont conservées).

Plusieurs DAGs pourront être proposés pour un même problème afin de vérifier la robustesse des résultats à différentes hypothèses causales, en particulier pour les analyses avec des données manquantes.

A noter que d'autres utilisations sont proposées dans la littérature, comme la figuration des modifications d'effets et les DAGs « signés » indiquant les sens des effets causaux (positifs ou négatifs) [86,87].

2.3 Gestion des données manquantes

Ce paragraphe présente en détail les méthodologies fréquemment utilisées pour gérer les données manquantes.

2.3.1 Exemple introductif et notations

On présente tout d'abord un exemple qui introduit les notations utilisées.

<u>Exemple 1 :</u>
Soit une étude en deux phases, avec des travailleurs invités à répondre à un questionnaire par phase sur les TMS et les expositions biomécaniques professionnelles. On veut étudier l'association $Q(x) = P(Y = 1|X = x)$ entre les expositions de la première phase (X) et les TMS de la deuxième phase (Y). Certains participants n'ont pas été suivis pour la deuxième phase donc on ne mesure qu'une partie des TMS Y^* (avec un mécanisme de données manquantes qui génère R_Y). On considère également la présence d'un gène U qui n'est pas mesuré dans cette étude mais qui est liée à la présence de TMS.

Individu	Expositions (X)	TMS au suivi (Y)	TMS au suivi observés (Y^*)	Données manquantes Oui=1/Non=0 (R_Y)	Caractéristiques génétiques U	Caractéristiques génétiques U observées
1	Faible	Non	Non	0	Gène	?
2	Forte	Oui	Oui	0	Pas de gène	?
3	Modérée	Oui	?	1	Gène	?
4	Faible	Non	Non	0	Gène	?
5	Modérée	Non	Non	0	Pas de gène	?
6	Forte	Oui	?	1	Pas de gène	?

En pratique, on n'a donc pas accès à la variable Y complète mais seulement à sa version partiellement observée Y^*.

Voici les notations dans ce paragraphe :

Soit $V^o = \left(V_j^o\right)_{j=1\ldots,k}$ l'ensemble des *k* variables complètement observées (« *o* » pour observé).

Soit $V^{m*} = \left(V_j^{m*}\right)_{j=1\ldots,l}$ l'ensemble des *l* variables avec les valeurs partiellement observées, c'est-à-dire les versions partiellement observées de $V^m = \left(V_j^m\right)_{j=1\ldots,l}$ complètes auxquelles on n'a pas accès (« m » pour manquant).

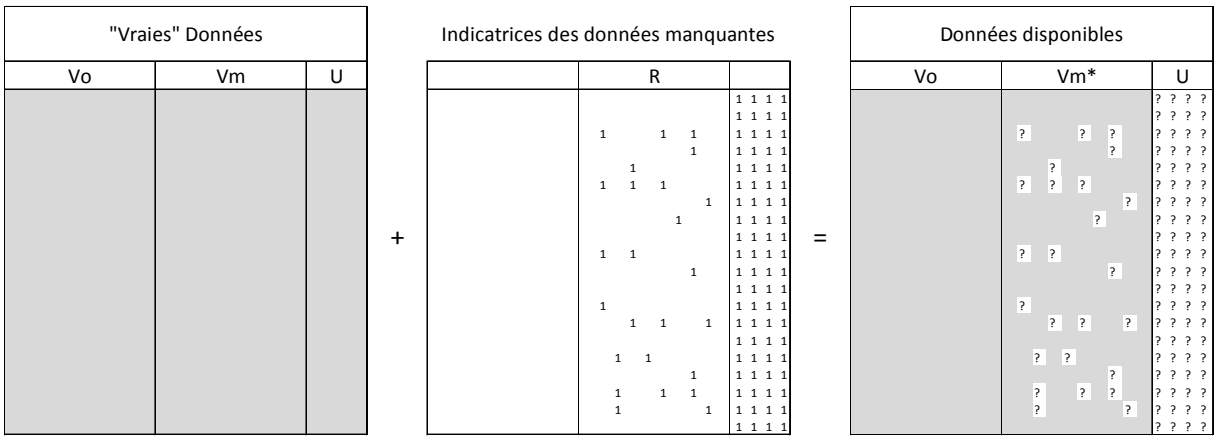
Soit $R = \left(R_j\right)_{j=1\ldots,l}$ les indicatrices de la présence de données manquantes telles que $R_{ij} = 1$ si V_{ij}^{m*}est manquant pour l'unité i et $R_{ij} = 0$ si $V_{ij}^{m*} = V_{ij}^m$.

On se place dans le cas où on veut estimer une relation $Q = f\left(P(V^o, V^m)\right)$, qui dépend de l'ensemble des données V^o, V^m.

Soit U l'ensemble des variables non disponibles et associées aux variables précédentes. Ces variables sont en pratique les variables qui pourraient être utiles pour l'étude de Q mais qui ne sont pas mesurées dans les études.

La Figure 11 résume la représentation des données dans la suite (sur les données disponibles, en gris= présence d'une observation, en blanc= absence d'observation).

Figure 11 - Représentation des données disponibles



<u>Exemple 1 (suite) :</u>
Dans cet exemple, les notations introduites correspondent à :

- $X = V_1^o$*, les expositions professionnelles complètement observées*
- $Y = V_1^m$*, les TMS au suivi*
- $Y^* = V_1^{m*}$*, les TMS au suivi partiellement observés*
- $Q(x) = P(V_1^m = 1 | V_1^o = x)$*, la relation d'intérêt, l'association entre les TMS et les expositions professionnelles*

En théorie, Q peut représenter tout type de relation : une prévalence, une association, une densité... Mais, en pratique, nous serons intéressés par la mesure d'une association dans un modèle paramétrique entre deux variables. Deux cas principaux seront étudiés :

- Une variable est complètement observée $\mathrm{V}_1^{\mathrm{o}}$ et l'autre variable est partiellement observée $\mathrm{V}_1^{\mathrm{m}}$ (cas de l'exemple 1). On définit le modèle paramétrique correspondant par :

$$g(E(\mathrm{V}_1^{\mathrm{m}} | \mathrm{V}_1^{\mathrm{o}}, V_2^o \dots V_k^o)) = \sum_{j \geq 2} \alpha_j V_j^o + \beta V_1^o$$

Avec g la fonction de lien qui dépend de la nature de V_1^m (ex : V_1^m binaire et g=logit), $(\alpha_j)_j$ les paramètres associés aux facteurs d'ajustement, β le paramètre associé au facteur d'intérêt.

- Les deux variables sont partiellement observées $\mathrm{V}_1^{\mathrm{m}}, V_2^m$. On définit le modèle paramétrique correspondant par :

$$g(E(\mathrm{V}_1^{\mathrm{m}} | \mathrm{V}_1^{\mathrm{o}}, V_2^o \dots V_k^o, V_2^m)) = \sum_{j \geq 1} \alpha_j V_j^o + \beta V_2^m$$

Avec g la fonction de lien qui dépend de la nature de V_1^m (ex : V_1^m binaire et g=logit), $(\alpha_j)_j$ les paramètres associés aux facteurs d'ajustement, β le paramètre associé au facteur d'intérêt.

La relation d'intérêt s'exprimera dans les deux cas comme une fonction de β, par exemple, dans le cas binaire, par $Q = e^{\beta} = OR$.

Dans les exemples et les méthodologies présentés, on ne se placera pas automatiquement dans le cadre d'une modélisation paramétrique. On le précisera explicitement quand ce sera le cas.

2.3.2 Caractérisation des mécanismes de données manquantes

2.3.2.1 Mécanismes MCAR, MAR et MNAR

2.3.2.1.1 Définitions et exemples

Définition des mécanismes de génération des données manquantes

> *On classe habituellement les mécanismes de données manquantes pour les données V^m (qui est associé à R) en trois catégories distinctes :*
>
> *- « Missing Completely At Random » (MCAR) : les mécanismes des données manquantes sont indépendants de toutes les variables considérées*
>
> $R \coprod V^o, V^m, U$
>
> $\leftrightarrow P(R|V^o, V^m, U) = P(R)$
>
> $\leftrightarrow P(V^o, V^m) = P(V^o, V^m|R = 0) = P(V^o, V^{m*}|R = 0)$
>
> *- « Missing At Random » (MAR) : les mécanismes des données manquantes dépendent uniquement des variables complètement observées*
>
> $R \coprod V^m, U|V^o$
>
> $\leftrightarrow P(R|V^o, V^m, U) = P(R|V^o)$
>
> $\leftrightarrow P(V^o, V^m) = P(V^m|V^o)P(V^o) = P(V^m|V^o, R = 0)P(V^o) = P(V^{m*}|V^o, R = 0)P(V^o)$
>
> *- « Missing Not At Random » (MNAR) : les mécanismes de données manquantes ne sont ni MCAR ni MAR*

Nous utilisons ici la définition des MAR basée sur les variables [88–90], qui diffère légèrement de la définition basée sur les observations initialement proposée par Little et Rubin [91]. Cette définition plus restrictive permettra d'appréhender plus clairement les enjeux dans le cas de variables partiellement observées. Elle permettra également une exploitation claire des DAG pour les données manquantes qui sera expliquée par la suite (cf. 2.3.3, p 45).

Ces définitions sont basées sur l'ensemble de variables prédéfinies V^o, V^m, U ; un mécanisme pour les données V^m est donc MAR conditionnellement aux variables complètement observées V^o.

<u>Exemple pratique avec différents mécanismes de données manquantes</u>

Une étude en santé au travail se propose d'évaluer les expositions professionnelles avec 3 mesures différentes : les expositions auto-rapportées par questionnaire (par AQ, notées V_1^m), les expositions évaluées à l'aide de l'intitulé d'emploi avec une matrice emploi-expositions (par MEE, notées V_2^m) et les expositions évaluées par des ergonomes sur le lieu de travail (par E, notées V_3^m). Pour tous les participants, un intitulé d'emploi et certaines caractéristiques individuelles sont disponibles (sans données manquantes, notées V^o). Certaines données sont manquantes pour les expositions :

- *par AQ : quand les participants ont choisi de ne pas répondre (R_1),*
- *par MEE : quand l'intitulé d'emploi n'a pas pu être mis en relation avec ceux de la MEE (R_2),*
- *par E : pour des raisons de coûts, ces mesures n'ont pu être effectuées que sur les individus avec des métiers à risque élevé de TMS (R_3).*

D'après les éléments précédents, voici ce qu'on peut être amené à conclure pour les données manquantes :

- *sur les expositions par MEE : le fait que l'exposition soit renseignée semble être aléatoire dans ce cas. On peut supposer qu'on est en présence d'un mécanisme MCAR ($R_2 \coprod V^o, V_1^m, V_2^m, V_3^m$)*
- *sur les expositions par E : la présence d'une évaluation dépend uniquement de l'intitulé d'emploi. C'est un mécanisme MAR ($R_3 \coprod V_1^m, V_2^m, V_3^m | V^o$)*
- *sur les expositions par AQ : la réponse ou non à un AQ est déterminée par des caractéristiques individuelles et l'emploi. On peut supposer que le mécanisme est MAR puisqu'on dispose de ces informations (c'est-à-dire $R_1 \coprod V_1^m, V_2^m, V_3^m | V^o$). Cependant, il est possible que les caractéristiques individuelles disponibles ne soient pas suffisantes pour déterminer le choix des individus de répondre ou non (traits de personnalité, conditions de travail...notées U). Dans ce cas, le mécanisme sera MNAR.*

<u>Exemple 2 :</u>

Un exemple classique de données manquantes MNAR est lorsque la variable $Y^* = V_1^{m*}$ est partiellement observée et son mécanisme de données manquantes qui génère R_1 est causé par $Y = V_1^m$. Par exemple, si Y est le revenu et Y^* le revenu partiellement observé par déclaration, il est plausible que les hauts revenus ne déclarent pas leurs revenus. Ainsi, on a un mécanisme de données manquantes MNAR puisqu'il est causé par la valeur même du revenu non observé.

2.3.2.1.2 Implications pratiques

Le traitement des données manquantes dans les analyses dépend du type de mécanisme considéré :

- Dans le cas où le mécanisme de données manquantes est MCAR, l'analyse sur les données complètes est suffisante (bien que sous-efficace puisqu'on se restreint à une partie de la population) pour obtenir des estimations sans biais.
- Dans le cas où le mécanisme de données manquantes est MAR, de nombreuses méthodologies sont disponibles pour obtenir des estimations sans biais (par exemple les imputations multiples ou les pondérations).
- Dans le cas où le mécanisme de données manquantes est MNAR, les choses sont plus complexes, il n'existe pas de méthodologie « type » et on doit travailler au cas par cas en fonction des hypothèses qu'on peut faire sur le mécanisme. Il peut être nécessaire de modéliser la distribution conjointe des données (V^o, V^m, U) et du mécanisme des données manquantes qui génère R.

Quand on est dans le cas MCAR ou MAR, on parle de mécanisme de données manquantes *ignorable.*

Puisque ces mécanismes impliquent des façons de gérer les données et des analyses différentes, la question de la possibilité de tester ces mécanismes est essentielle.

On peut tester si un mécanisme est MCAR versus MAR, mais on ne peut pas tester si un mécanisme est MAR versus MNAR puisque, par définition, les variables partiellement observées V^{m*} ou non observées U ne sont pas disponibles pour étudier le mécanisme. En d'autres termes, on peut avoir les mêmes données disponibles que le mécanisme de données manquantes soit MAR ou MNAR. Quelques travaux proposent de tester à partir des données disponibles certaines hypothèses impliquées par le mécanisme MAR [88,90,92], mais les implémentations pratiques de ces tests ne sont pas disponibles dans les logiciels. Ce travail de thèse ne teste pas ces hypothèses pour cette raison. Cependant, l'hypothèse MAR étant forte, elle sera discutée et des analyses de sensibilité seront effectuées.

Avant de continuer cette partie avec un dernier exemple, on va introduire un sous-type de mécanisme MNAR, qu'on appellera mécanisme MAR séquentiel.

2.3.2.2 Mécanismes MAR séquentiels

<u>Définition des MAR séquentiels – Cas particulier de deux variables manquantes</u> [93,94]

> *On dit que le mécanisme de données manquantes de (V_1^m, V_2^m) associé aux indicatrices (R_1, R_2) est MAR séquentiel s'il existe une séquence de variables manquantes pour lesquelles chaque mécanisme est MAR conditionnellement aux variables suivantes [93,94]. En d'autres termes, on peut ordonner les variables pour obtenir (V_1^m, V_2^m) telles que :*
> *(1) V_2^m est MAR conditionnellement à $V^o \leftrightarrow V_2^m \coprod R_2 \,|V^o$*
> *(2) V_1^m est MAR conditionnellement à $(V_2^m, V^o) \leftrightarrow V_1^m \coprod R_1 \,|V^o, V_2^m$*
> *(3) V_1^m et R_2 sont indépendants conditionnellement à $(V_2^m, V^o) \leftrightarrow V_1^m \coprod R_2 \,|V^o, V_2^m$*

<u>Définition des MAR séquentiels</u> [93,94]

> *On dit que le mécanisme de données manquantes de $(V_1^m, V_2^m \ldots, V_l^m)$ associé aux indicatrices $(R_1, R_2 \ldots, R_l)$ est MAR séquentiel s'il existe une séquence de variables manquantes pour lesquelles chaque mécanisme est MAR conditionnellement aux variables suivantes. En d'autres termes, on peut ordonner les variables pour obtenir $(V_1^m, V_2^m \ldots, V_l^m)$ tel que :*
> *(1) V_l^m est MAR conditionnellement à $V^o \leftrightarrow V_l^m \coprod R_l \;|V^o$*
> *(2) $\forall i = 1 \ldots, l \quad V_i^m$ est MAR conditionnellement à $V^o, V_{i+1}^m \ldots, V_l^m$*
> *$\leftrightarrow \forall i = 1 \ldots, l \; V_i^m \coprod R_i \;|V^o, V_{i+1}^m \ldots, V_l^m$*
> *(3) $\forall i = 1 \ldots, l \quad V_i^m$ et $(R_{i+1} \ldots R_l)$ sont indépendants conditionnellement à $V^o, V_{i+1}^m \ldots, V_l^m$*
> *$\leftrightarrow \forall i = 1 \ldots, l \; V_i^m \coprod (R_{i+1} \ldots R_l) \;|V^o, V_{i+1}^m \ldots, V_l^m$*

D'après la définition d'un mécanisme MNAR choisie, les mécanismes MAR séquentiels forment donc un sous-ensemble des mécanismes MNAR.

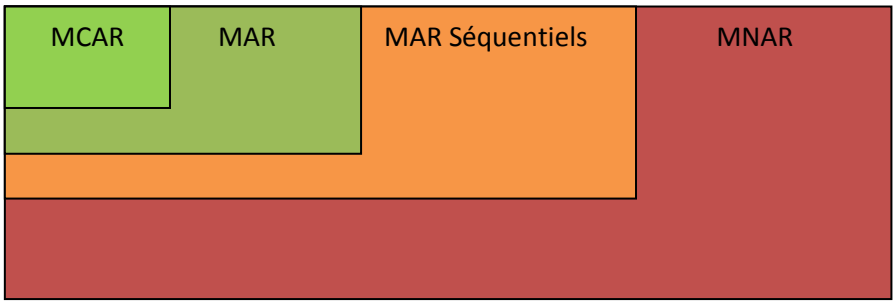
> <u>*Exemple 3 :*</u>
>
> *On part de l'exemple 1, c'est-à-dire une étude en deux phases, avec des travailleurs invités à répondre à un questionnaire par phase sur les TMS et les expositions biomécaniques professionnelles. On veut étudier l'association $Q(x) = P(Y = 1|X = x)$ entre les expositions de la première phase (X) et les TMS de la deuxième phase (Y). Si on suppose que la présence d'un suivi est causée exclusivement par les expositions professionnelles observées en 1ère phase (X), on est en présence d'un mécanisme MAR (conditionnellement à X). Dans ce cas, on peut estimer sans biais la relation d'intérêt sur les seules données complètes ($Q(x) = P(Y = 1|X = x) = P(Y^* = 1|X = x, R_Y = 0)$).*
>
> *Supposons qu'en plus des TMS de la deuxième phase (Y), les expositions de la première phase (X) sont partiellement observées (données manquantes indiquées par R_X), car certains questionnaires de la première phase ont été aléatoirement perdus. On est dans un cas MCAR pour les expositions de la première phase.*
>
> *Au final, on se retrouve dans un cas MNAR pour les TMS de la deuxième phase, puisque R_Y dépendent des expositions de la première phase partiellement observées. D'après les définitions précédentes, on voit qu'on est dans un mécanisme MAR séquentiel, puisqu'on a : (1) $R_X \coprod X$, (2) $Y \coprod R_Y|X$ et (3) $Y \coprod R_X|X$.*

Cependant, bien que le mécanisme ne soit pas MAR, on montre qu'on peut estimer l'association recherchée sur les données complètes. En effet, par les deux relations d'indépendance $Y \perp\!\!\!\perp R_Y | X$ et $Y \perp\!\!\!\perp R_X | X$, on a :

$$\begin{aligned} Q(x) &= P(Y = 1 | X = x) \\ &= P(Y = 1 | X = x, R_Y = 0) \\ &= P(Y = 1 | X = x, R_Y = 0, R_X = 0) \\ &= P(Y^* = 1 | X^* = x, R_Y = 0, R_X = 0) \end{aligned}$$

On peut schématiser la partition des mécanismes de données manquantes comme indiqué dans la Figure 12.

Figure 12 - Types de mécanismes de données manquantes



2.3.2.3 « Recoverability »

L'exemple 3 montre qu'une propriété qui nous intéresse est de savoir si avec les données disponibles, il existe une procédure qui permet d'obtenir des estimations non biaisées. C'est la notion de *recoverability* (ou récupérabilité) introduite par Pearl et Mohan [89,95]. Cette propriété, bien que peu exploitée dans la partie pratique, permet de clarifier les propriétés des MAR séquentiels, puisqu'ils sont peu fréquemment évoqués dans la littérature [95].

2.3.2.3.1 Définition et exemples

Définition de la recoverability d'une relation Q

> *Une relation d'intérêt $Q = f\big(P(V^o, V^m)\big)$ est dite recoverable si et seulement si Q peut être exprimée comme une fonction de (V^0, V^{m*}, R), c'est-à-dire que Q s'exprime avec l'ensemble des éléments observables.*

Les notions MCAR, MAR et MNAR ne concernent que le mécanisme qui génère les données manquantes tandis que la notion de recoverability concerne en plus la relation étudiée.

Exemple 1 (suite) :

Dans cet exemple simple, on a $Q(x) = P(Y = 1 | X = x)$ la relation d'intérêt (association entre les expositions de la première phase (X) et les TMS de la deuxième phase (Y)) avec Y partiellement observé et le mécanisme MAR qui génère les données manquantes R_Y correspondant. D'après la définition générale de la recoverability, si Q s'exprime en fonction de X, Y^* (les TMS partiellement observés) et R_Y, alors Q est recoverable. C'est bien le cas ici, puisqu'on peut écrire avec

l'hypothèse MAR que : $Q(x) = P(Y = 1|X = x)$
$= P(Y = 1|X = x, R_Y = 0)$
$= P(Y^* = 1|X = x, R_Y = 0)$

Exemple 3 (suite) :

Dans cet exemple, on a vu que $(x) = P(Y = 1|X = x) = P(Y^* = 1|X^* = x, R_Y = 0, R_X = 0)$. *D'après la définition générale de la recoverability, Q s'exprime en fonction de* X^*, Y^* *(les expositions et TMS partiellement observés) et* R_X, R_Y *, et est donc recoverable.*

Propriété des relations recoverables [90]

> *Si une relation* Q *est recoverable, alors il existe un estimateur consistant de* Q.

Afin de simplifier les notations dans les parties suivantes, on introduit une dernière définition sur la recoverability.

Définition de recoverability d'un mécanisme

> *On dit qu'un mécanisme de données manquantes est recoverable si l'ensemble des relations d'intérêt* Q *sous ce mécanisme sont recoverables.*

NB : un mécanisme *ignorable* est *recoverable* puisque toutes les relations peuvent s'exprimer en fonction de $(\mathrm{V}^0, \mathrm{V}^{\mathrm{m}*})$. A l'inverse, certaines relations Q sont *recoverables* et nécessitent de modéliser le mécanisme des données manquantes et sont associées à des mécanismes de données manquantes *non-ignorables* (cf. exemple A2 en Annexe A , p 152).

2.3.2.3.2 Propriété de *recoverability* des mécanismes définis

Les mécanismes MAR (et par conséquent MCAR) sont *recoverables* car on peut écrire, en conditionnant sur les variables complètement observées V^o :

$$P(V^o, V^m) = P(V^m|V^o)P(V^o) = P(V^m|V^o, R = 0)P(V^o) = P(V^{m*}|V^o, R = 0)P(V^o)$$

Ainsi, toutes les relations d'intérêt Q pourront s'exprimer en fonction des données disponibles, puisque c'est le cas de la distribution des données (cf. exemple 1).

Les MAR séquentiels sont *recoverables* (cf. exemple 3). On peut le démontrer dans le cas de deux variables V_1^m *et* V_2^m partiellement observées avec un mécanisme MAR séquentiel. On a par définition $(V_1^m \coprod R_1 | V^o, V_2^m)$, $(V_1^m \coprod R_2 | V^o, V_2^m)$ et $(V_2^m \coprod R_2 | V^o)$.

Ainsi, on peut écrire :

$$P(V_1^m, V_2^m, V^o) = P(V_1^m|V_2^m, V^o)P(V_2^m|V^o)P(V^o)$$

$$= \underbrace{P(V_1^m|V_2^m, V^o, R_1 = 0, R_2 = 0)}\underbrace{P(V_2^m|V^o, R_2 = 0)}P(V^o)$$

car $V_1^m \coprod R_1 \,|V^o, V_2^m$ et car $V_2^m \coprod R_2 \,|V^o$

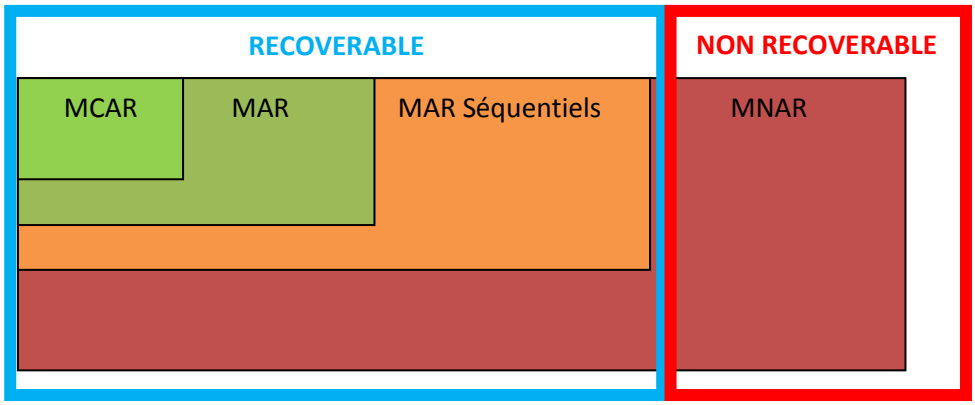
$V_1^m \coprod R_2 \,|V^o, V_2^m$

Cette propriété est généralisable au cas de plus de deux variables avec un mécanisme MAR séquentiel.

Certains mécanismes MNAR sont recoverables (exemple 3) et d'autres sont non *recoverables* (exemple 2).

Ainsi, on résume les propriétés précédentes avec le schéma suivant :

Figure 13 - Types de mécanismes de données manquantes et propriété de *recoverability*



2.3.2.3.3 La *recoverability* en pratique

On voit que la *recoverability* permet d'exprimer la possibilité d'estimer sans biais la relation d'intérêt. Elle fournit également une décomposition qui indique quelles distributions peuvent être estimées sans biais et comment les combiner.

En pratique, cette décomposition donne les conditionnements minimum à prendre en compte pour avoir des estimations sans biais. Selon les choix de méthodologies en pratique, il pourra être intéressant de prendre en compte des variables supplémentaires (par exemple, pour diminuer la variance des estimations).

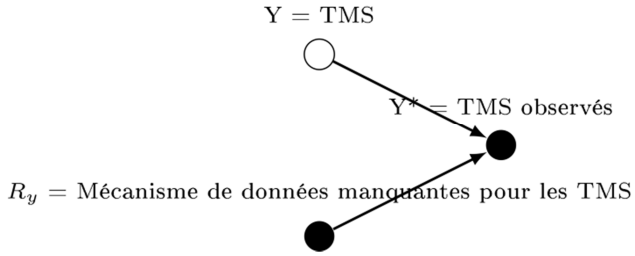
Enfin, la *recoverability* étant une notion récente, avoir un mécanisme *recoverable* n'implique pas qu'il y ait automatiquement une procédure implémentée en pratique pour trouver une estimation pour la relation d'intérêt, même si en théorie il en existe une.

2.3.3 Les mécanismes de données manquantes et les DAGs

2.3.3.1 Représentations

De récents travaux ont mis en avant l'apport que les DAGs décrits dans le paragraphe 2.2 peuvent avoir sur la compréhension des mécanismes de données manquantes et les stratégies d'analyse en présence de données manquantes [89,95,96]. En effet, on peut étoffer les DAGs avec le mécanisme de données manquantes et ses causes afin de figurer les relations d'indépendance entre les variables partiellement observées et la présence de données. Plus précisément, au lieu d'avoir une variable Y observée, on représente la variable partiellement observée Y^*, l'indicatrice de données manquantes R_Y et la variable Y correspondante. Les variables Y et R_Y sont causes de la variable Y^*. La variable Y étant partiellement observée, elle est représentée par un cercle vide dans le DAG. Il faut de plus introduire dans le graphe les causes de R_Y (Figure 14). On remarque que R_Y ne suit pas la dernière assertion de la définition d'un DAG classique car il est représenté même s'il n'est pas associé à au moins deux variables du DAG (cf.2.2.1, p 32).

Figure 14 - Représentation des données partiellement observées



2.3.3.2 Utilisation des DAGs en pratique et premiers exemples

Les DAGs fournissent des informations très précises sur la caractérisation des données manquantes, en particulier sur les notions de MCAR, MAR et MNAR et la notion de *recoverability*. En effet, dans les DAGs, on indique toutes les hypothèses sur un problème donné, avec les causes des variables considérées et les causes de leurs mécanismes de données manquantes.

On peut savoir si un mécanisme est MCAR, MAR ou MNAR à partir d'un DAG avec les règles suivantes :

- Mécanisme MCAR : aucune variable du DAG ne cause les indicatrices de données manquantes R,
- Mécanisme MAR : les variables qui causent les indicatrices de données manquantes R sont complètement observées (cercle plein),

- Mécanisme MNAR : une partie des variables qui causent les indicatrices de données manquantes R sont partiellement observées voire (complètement) non observées (cercle vide).

Pour savoir si un mécanisme est *recoverable* ou non (quand il est MNAR), on utilisera les relations d'indépendance qu'on peut lire sur le DAG afin d'essayer de trouver une expression de la relation d'intérêt en fonction des données uniquement observées.

En présence de données manquantes, on va donc procéder en plusieurs étapes :

- On crée d'abord le DAG incluant les variables de la relation d'intérêt, les facteurs de confusion et les facteurs intermédiaires (selon le type de relation d'intérêt) comme s'il n'y avait pas de données manquantes.

- On inclut ensuite les indicatrices des données manquantes pour chaque variable partiellement observée et on pose les liens causaux avec les variables représentées. On ajoute si nécessaire d'autres variables qui causent le fait que les données sont manquantes et qui sont associées à d'autres variables du DAG.

- Sur ce DAG global, on regarde la nature des mécanismes des données manquantes ; on cherche à voir s'il s'agit d'un mécanisme MCAR, MAR ou MAR séquentiel :

    - Quand le mécanisme est MCAR ou MAR, on sait que la relation d'intérêt est *recoverable* et il existe des méthodologies adaptées qui seront détaillées dans les paragraphes suivants (cf. paragraphe 2.3.4, p 52) ;

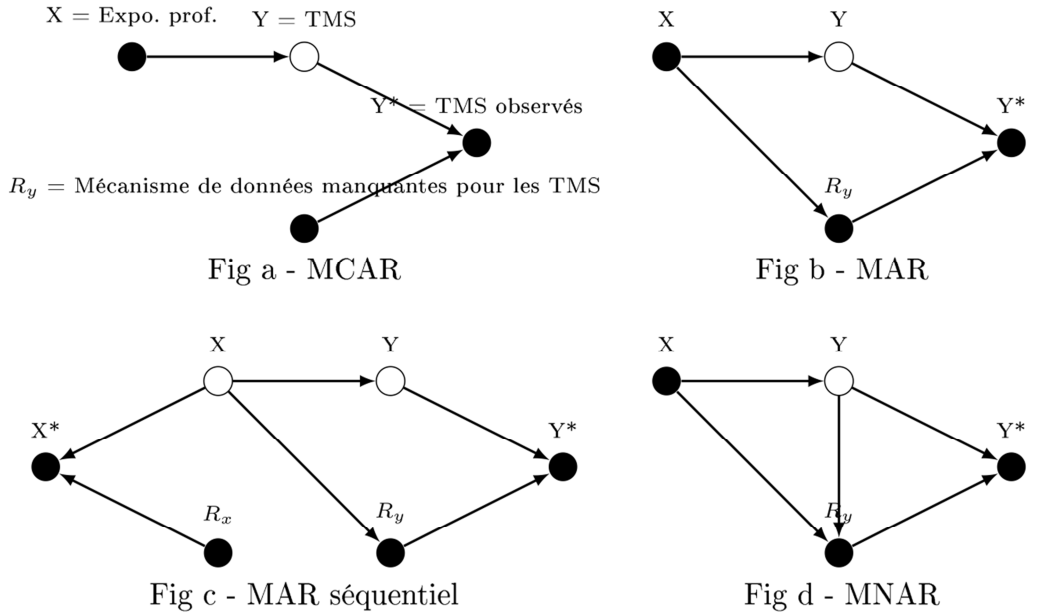
    - Dans le cas où le mécanisme est MAR séquentiel, on sait que la relation d'intérêt est *recoverable*, mais il peut être utile de trouver l'ordre dans lequel le mécanisme MAR séquentiel opère et sur quel ensemble de variables on peut conditionner pour chaque mécanisme « MAR » successif. On citera dans la suite une méthodologie adaptée dans ce cas (cf. paragraphe 2.3.4.2.3, p 61) ;

    - Dans le cas où le mécanisme est MNAR, on regarde si la relation est *recoverable*. La décomposition qui peut être trouvée dans le cas MNAR *recoverable* donne des indices sur la procédure à suivre pour réussir à estimer la relation d'intérêt sans biais. Les aspects analytiques dans le cas MNAR ne seront développés dans la suite que pour les analyses de sensibilité (cf. paragraphe 2.3.5, p 65).

En conclusion, le DAG permet d'exhiber les caractéristiques du mécanisme de données manquantes et d'en voir ses propriétés. Il permet aussi d'appréhender quelles distributions seront estimées sans biais et donc donne des indications sur la procédure d'estimation de la relation d'intérêt.

Pour mieux se rendre compte de l'utilisation des DAGs en pratique, on présente les DAGs correspondant à des exemples de différents mécanismes de données manquantes (Figure 15).

Figure 15 – Exemples de DAGs avec données manquantes



La Figure 15a est une situation MCAR, puisqu'aucune variable ne cause R_y : on a $R_y \coprod (X, Y)$. Ici, on suppose que le fait de mesurer ou non les TMS ne dépend pas des expositions professionnelles ou de la présence ou non de TMS ; cela peut par exemple être le cas dans un recueil par questionnaire si certains questionnaires ont été perdus « au hasard ».

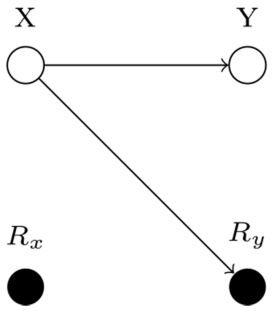
La Figure 15b est une situation MAR, puisque X cause R_y : on a $R_y \coprod Y|X$. Ici, on suppose que le fait de mesurer ou non les TMS dépend des expositions professionnelles. Par exemple, dans le cadre d'un service de médecine du travail, on peut avoir un suivi plus rapproché pour les individus ayant certaines expositions professionnelles et donc la présence d'une mesure peut indiquer une plus forte exposition.

La Figure 15c est la situation MAR séquentiel de l'exemple 3. On voit que R_x et X sont indépendants. X cause Y et R_y. Ainsi, conditionnellement à X, Y est indépendant de R_y et R_x. Ces éléments d'indépendance correspondent à la définition d'un mécanisme MAR.

La Figure 15d est la situation MNAR de l'exemple 2. En effet, on a Y qui cause R_y, ce qui est une situation MNAR classique car on ne peut pas trouver un conditionnement qui rende Y et R_y indépendants comme dans un mécanisme MAR.

On peut simplifier les diagrammes en ne faisant pas figurer les X^*, Y^*. Par exemple, pour la Figure 15c, on obtient la Figure 16. On écrira les DAGs de cette façon dans la suite pour plus de lisibilité.

Figure 16 - DAG avec données manquantes simplifié (figure 15c)



2.3.3.3 Construction d'un DAG utilisé dans les analyses

Les exemples proposés dans ce paragraphe construisent progressivement le DAG utilisé dans les analyses (Figure 34, p 114). On souhaite mettre en évidence le lien entre les hypothèses sur les mécanismes de données manquantes, les DAGs et les ajustements nécessaires et suffisants à la prise en compte correcte des données manquantes selon la relation étudiée

Un lecteur à l'aise avec les notions et les exemples précédents peut directement passer directement au dernier exemple de cette partie (exemple 7) ou à ceux de l'Annexe A (p 152).

On se place dans le cas où on a une exposition professionnelle X binaire (les tâches répétitives au travail par exemple) et on veut étudier l'association avec les TMS Y (la présence de douleurs aux coudes par exemple). X et Y ont des données manquantes par deux mécanismes qui génèrent R_X et R_Y. La relation d'intérêt est le risque relatif de présenter un TMS pour les exposés par rapport aux non exposés :

$$Q = \frac{P(Y = 1|X = 1)}{P(Y = 1|X = 0)}$$

Différents exemples sont figurés par des DAGs afin de comprendre quels sont leurs implications en pratique.

> *<u>Exemple 4 :</u>*
>
> *On suppose ici qu'on se trouve dans un cas simple sans facteur de confusion.*
> *On dispose de l'exposition professionnelle X et des TMS Y pour une partie*

des individus. Les mécanismes de données manquantes ne dépendent d'aucun facteur particulier, on est dans un cas MCAR.

Figure 17 - DAG avec mécanisme MCAR

Sur le graphique, on voit que $Y \coprod R_X$ *et* $Y \coprod R_Y$*, ainsi la distribution* $P(Y|X)$ *peut s'écrire :*

$$P(Y|X) = P(Y|X, R_X = 0, R_Y = 0) = P(Y^*|X^*, R_X = 0, R_Y = 0)$$

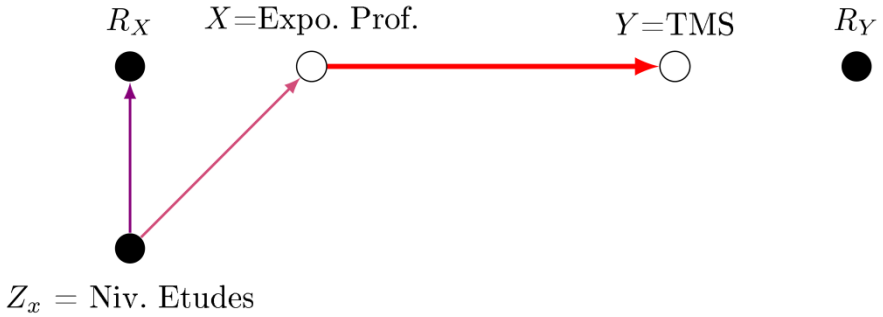
Ainsi, on peut écrire : $Q = \frac{P(Y^*=1|X^*=1,R_X=0,R_Y=0)}{P(Y^*=1|X^*=0,R_X=0,R_Y=0)}$.

Le risque relatif peut donc être estimé sans biais à partir des individus pour lesquels les données sont complètes.

<u>Exemple 5 :</u>

On se place dans le cas où l'exposition professionnelle X est recueillie par auto-questionnaire et la présence d'une réponse à celui-ci dépend du niveau d'études (Z_x)*. Les TMS Y sont évalués lors d'une visite médicale du travail obligatoire et la présence d'une évaluation est complètement aléatoire. On a une cause de mécanisme de données manquantes et on suppose qu'aucun facteur de confusion ne rentre en jeu.*

Figure 18 - DAG avec un mécanisme MAR



Sur le graphique, on voit que le mécanisme qui génère R_X *est MAR (car il ne dépend que de* Z_x*) et que le mécanisme qui génère* R_Y *est MCAR. On a* $Y \coprod R_X|X$ *et* $Y \coprod R_Y$*, ainsi la distribution* $P(Y|X)$ *peut s'écrire (cf. exemple 1(suite)) :*

$$P(Y|X) = P(Y^*|X^*, R_X = 0, R_Y = 0)$$

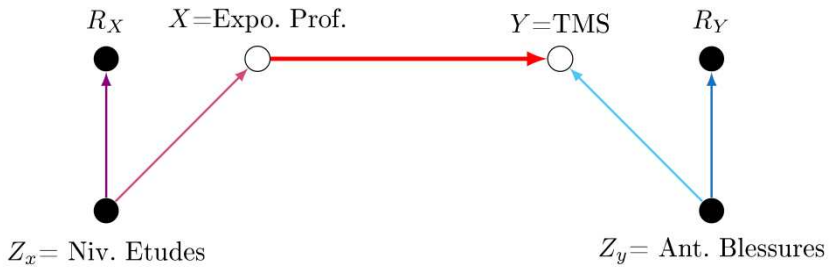
On retrouve bien que Q est recoverable et il est possible d'estimer sans biais la relation d'intérêt Q à partir des individus pour lesquels les données sont complètes.

Comme Q dépend de la distribution de Y conditionnellement à X, il n'est pas nécessaire de conditionner sur les variables qui causent X et son mécanisme de données manquantes qui génère R_X *(ici, le niveau d'études).*

Exemple 6 :

On se place dans le cas où l'exposition professionnelle X et les TMS Y sont recueillis par deux auto-questionnaires successifs et que la réponse à ceux-ci dépend respectivement du niveau d'études (Z_x) et des antécédents de blessures (Z_y) obtenus par des fichiers administratifs. On suppose qu'on n'a pas de facteur de confusion.

Figure 19 - DAG avec un mécanisme MAR



A nouveau, le mécanisme associé à R_X est MAR (car il ne dépend que de Z_x) et que le mécanisme associé à R_Y est MAR (car il ne dépend que de Z_y). On déduit du graphe que $Y \coprod R_X | X$ et $Y \coprod R_Y | Z_y$.

On a, en notant ζ l'ensemble de valeurs que prend Z_y :

$$P(Y|X) = \sum_{z \in \zeta} P\left(Y \middle| X, Z_y = z\right) P\left(Z_y = z\right)$$

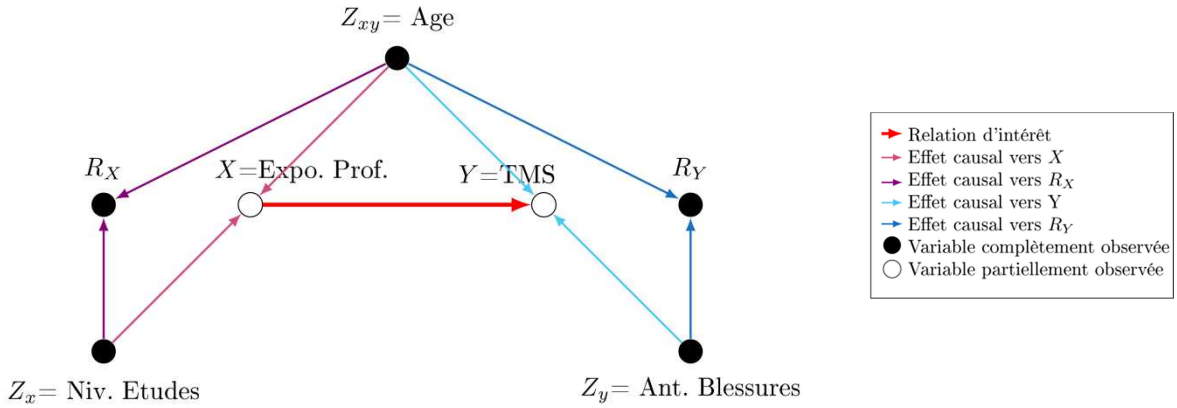
$$= \sum_{z \in \zeta} P\left(Y^* \middle| X^*, R_X = 0, R_Y = 0, Z_y = z\right) P\left(Z_y = z\right)$$

Ainsi, en prenant en compte Z_y dans les analyses (les antécédents de blessures), on pourra estimer sans biais le risque relatif (sans avoir besoin de prendre en compte le niveau d'études).

Exemple 7 :

On reprend l'exemple 6, avec le niveau d'études (Z_x) et les antécédents de blessure (Z_y). On ajoute un facteur de confusion, l'âge (Z_{xy}), qu'on suppose être une des causes de la présence de données manquantes, de l'exposition professionnelle considérée et des TMS.

Figure 20 - DAG avec un mécanisme MAR



Le mécanisme qui génère R_X est MAR (car il ne dépend que de Z_x et Z_{xy}) et le mécanisme qui génère R_Y est MAR (car il ne dépend que de Z_y et Z_{xy}). On déduit du graphe que $Y \perp\!\!\!\perp R_X | X, Z_{xy}$, $Y \perp\!\!\!\perp R_Y | Z_y, Z_{xy}$ et $Z_y \perp\!\!\!\perp Z_{xy}$.

On a, en notant ζ l'ensemble de valeurs que prend Z_y et η l'ensemble de valeurs que prend Z_{xy} :

$$P(Y|X) = \sum_{z_1 \in \zeta, z_2 \in \eta} P\left(Y \middle| X, Z_y = z_1, Z_{xy} = z_2\right) P\left(Z_y = z_1\right) P(Z_{xy} = z_2)$$

$$= \sum_{z_1 \in \zeta, z_2 \in \eta} P\left(Y^* \middle| X^*, R_X = 0, R_Y = 0, Z_y = z_1, Z_{xy} = z_2\right) P\left(Z_y = z_1\right) P(Z_{xy} = z_2)$$

Ainsi, en prenant en compte Z_y et Z_{xy} dans les analyses (les antécédents de blessures et l'âge), on pourra estimer sans biais le risque relatif.

En Annexe A , on trouve d'autres exemples de DAGs, en particulier des DAGs avec des mécanismes de données manquantes MAR séquentiels et MNAR. Ces DAGs permettent de se rapprocher du DAG qui sera utilisé dans l'étude 4.2.

Dans cette partie on a vu comment préciser les hypothèses faites sur les données manquantes et les implications qu'elles ont, d'une part sur les relations d'indépendance entre les variables considérées et d'autre part sur l'ensemble de variables qu'on peut prendre en compte afin d'estimer sans biais les données. Dans certains cas MNAR, les relations sont *recoverables (cf.* Annexe A).

2.3.4 Les différentes méthodes d'estimation

Dans le paragraphe précédent, les différents concepts introduits ont permis de caractériser les mécanismes de données manquantes et d'appréhender quelles relations d'intérêt Q sont estimables sans biais sur les cas complets (c'est-à-dire l'ensemble des observations ayant des données complètement observées pour toutes les variables). Nous présentons rapidement les quatre méthodes disponibles pour les analyses dans le cadre de données manquantes [97,98] : l'analyse sur cas complets (non détaillée dans la suite), l'analyse avec imputations simples (non détaillée dans la suite), l'analyse avec pondérations et l'analyse avec imputations multiples. On s'intéresse à une relation d'intérêt Q et selon chaque méthode, au biais et à la variance attendus pour l'estimation de Q :

- Analyse sur cas complets :
  - Principe : on exclut des analyses les observations avec des données manquantes
  - Biais : elle est sans biais pour les mécanismes MCAR et certains mécanismes spécifiques
  - Variance : cette méthode surestime la variance
  - En pratique : à utiliser si peu de données manquantes (<5%)
- Analyse avec imputations simples :
  - Principe : on complète (impute) chaque variable partiellement observée V_j^{m*} par diverses méthodes (la moyenne des observations, la moyenne conditionnelle,…) et on analyse les données complétées
  - Biais : elle peut être sans biais si les imputations sont pertinentes pour la relation d'intérêt ou pour certains mécanismes très spécifiques
  - Variance : cette méthode sous-estime la variance sans mesures correctives particulières (par exemple, si on simule les données à partir de la distribution a posteriori des paramètres du modèle d'imputation)
  - En pratique : à utiliser quand chaque variable manquante a peu de données à imputer mais que l'analyse sur cas complets supprime beaucoup d'individus
- Analyse avec pondérations : les analyses avec pondérations seront détaillées ici dans le cas particulier d'une variable partiellement observée (paragraphe 2.3.4.1, p 54, puis utilisées dans les analyses 4.3, p 114).

- Principe : on estime la propension des individus à avoir une donnée $P(R = 1)$ et on pondère les cas complets sur cette propension. Le modèle d'analyse classique est utilisé en considérant qu'on a des poids connus.
- Biais : cette méthode donne des estimations sans biais si les données manquantes sont MAR et si le modèle de pondération est correctement spécifié
- Variance : la variance supposant que les pondérations sont connues (plutôt qu'estimées) est en général surestimée. De plus, si les poids sont très dispersés et non corrélés aux variables d'intérêt, il peut y avoir un phénomène d'inflation.En pratique : à utiliser de préférence quand le mécanisme de données manquantes est modélisable facilement

- Analyse avec des imputations multiples : les analyses avec imputations multiples seront vues ici dans leur généralité (paragraphe 2.3.4.2, p 57, puis utilisées dans les analyses 4.2, p 95).
  - Principe : chaque variable partiellement observée V_j^{m*} est imputée par une liste de M variables complètes avec des valeurs plausibles pour les valeurs manquantes. Les M jeux de données imputés pour l'ensemble des V_j^{m*} sont analysés séparément et les estimations de chaque jeu de données sont recombinées.
  - Biais : cette méthode donne des estimations sans biais si le mécanisme de données manquantes est MAR (ou MAR séquentiel pour l'outil « mice » décrit dans les paragraphes suivants) et si le modèle d'imputation est correctement spécifié
  - Variance : la variance est correctement estimée sous les mêmes conditions que le point précédent
  - En pratique : à utiliser quand on est en mesure d'imputer les valeurs manquantes des variables partiellement observées.

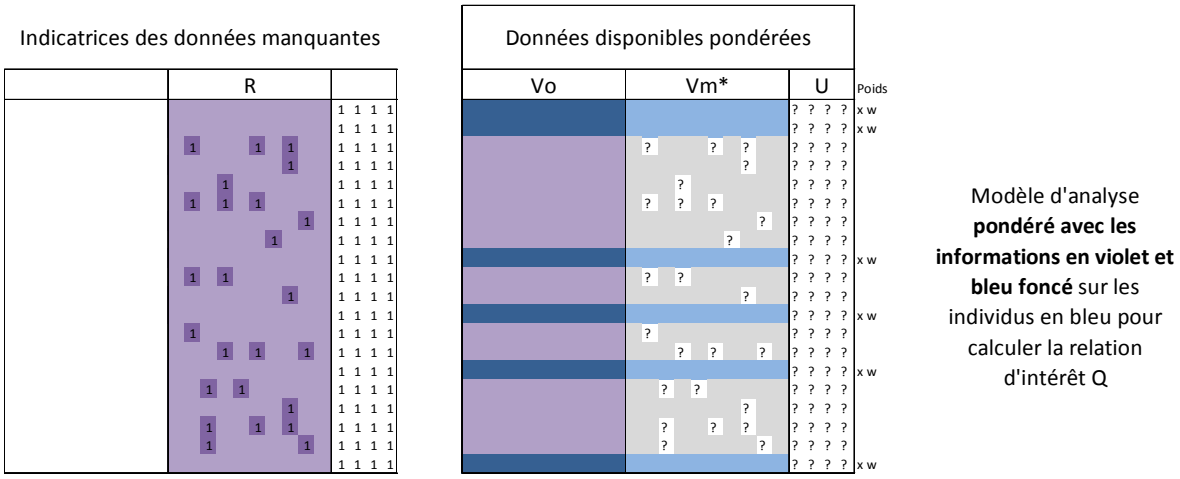
Pour terminer cette partie, un paragraphe introduira les analyses de sensibilité sur le mécanisme de données manquantes MAR dans le cas d'imputations multiples (au cas où le mécanisme est en fait MNAR). L'objectif de ces analyses de sensibilité est de savoir si les résultats obtenus sont robustes à la mauvaise spécification du mécanisme de données manquantes (MAR ou MAR séquentiel au lieu de MNAR).

2.3.4.1 Analyse sur cas complets avec pondérations

2.3.4.1.1 Principes généraux

Comme décrit précédemment, l'idée est ici de pondérer les observations complètes afin qu'elles puissent représenter l'ensemble des individus (avec ou sans données manquantes). On applique alors les méthodes classique d'analyse, mais en pondérant par la propension des individus à avoir une donnée observée (Figure 21).

Figure 21 - Représentation des données et de la méthode cas complets avec pondérations



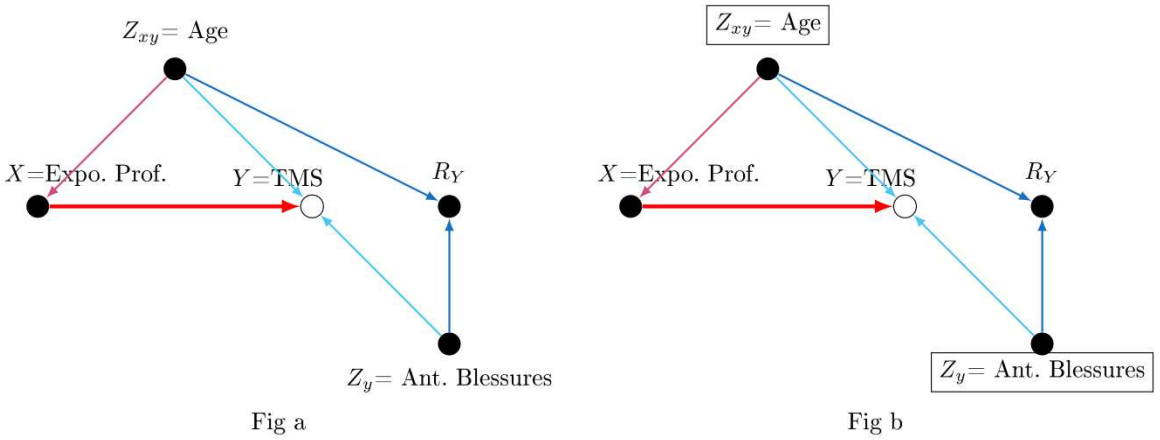
Ainsi, la modélisation s'effectue en deux étapes :

- Il faut d'abord trouver des pondérations adaptées au mécanisme de données manquantes et à la relation d'intérêt, en ajustant sur les bonnes variables (à l'aide du DAG et des propriétés précédentes).
- Il faut ensuite effectuer l'analyse pondérée sur les cas complets en ajustant sur les facteurs de confusion

2.3.4.1.2 Exemple

On reprend l'exemple 7 en simplifiant le problème avec une variable X complètement observée. On a donc une variable Y partiellement observée binaire avec un mécanisme de données manquantes MAR qui génère R_Y et une variable X complètement observée continue. On souhaite étudier l'association entre Y et X : $Q(x) = P(Y = 1|X = x)$.

Figure 22 - DAG des relations causales (Fig. a - sans conditionnement, Fig. b - avec conditionnement)



D'après les propriétés sur les mécanismes MAR décrites dans les paragraphes précédents, la relation $Q(x)$ est *recoverable* avec une décomposition faisant intervenir $P(Y^*|X^*, Z_{xy}, Z_y, R_X = 0, R_Y = 0)$. Il est suffisant de prendre en compte Z_{xy} et Z_y dans les pondérations.

Ainsi, on écrit le modèle de pondération suivant :

$$logit\ P\left(R_Y = 1 \middle| Z_{xy}, Z_y\right) = \alpha_1 + \alpha_2 Z_{xy} + \alpha_3 Z_y$$

De ce modèle, on tire les prédictions $\widehat{p_i}$, pour chaque individu i d'avoir une valeur observée pour Y et $\widehat{w_i} = {}^{1}/_{\widehat{p_i}}$. Ensuite, on effectue le modèle d'analyse pondéré sur les $\widehat{w_i}$:

$$logit\ P\left(Y = 1 \middle| X, Z_{xy}\right) = \beta_1 + \beta_2 X + \beta_3 Z_{xy}$$

Remarque : l'ajustement sur le facteur de confusion Z_{xy} est optionnel puisqu'on a déjà pris en compte ce facteur en pondérant.

2.3.4.1.3 Conclusion

Cette méthode donne des résultats non biaisés quand le modèle de pondération est bien spécifié. Cette méthode se base uniquement sur les cas observés ce qui ne permet pas de gagner en puissance par rapport à l'analyse sur cas complets. Elle a l'avantage de préserver toutes les distributions des données et peut être adaptée au calcul d'autres relations d'intérêt que celles prévues initialement.

Cette méthode peut se généraliser au cas de plusieurs variables manquantes, la probabilité sera alors modélisée par la probabilité d'avoir des données complètes pour toutes les variables contre des données complètement ou partiellement manquantes (mécanisme total des données manquantes). Les résultats sont également sans biais si le modèle pour les données manquantes est correct. Cependant, si les variables partiellement observées sont associées à des mécanismes de données

manquantes différents, il sera complexe voire impossible de modéliser le mécanisme total des données manquantes.

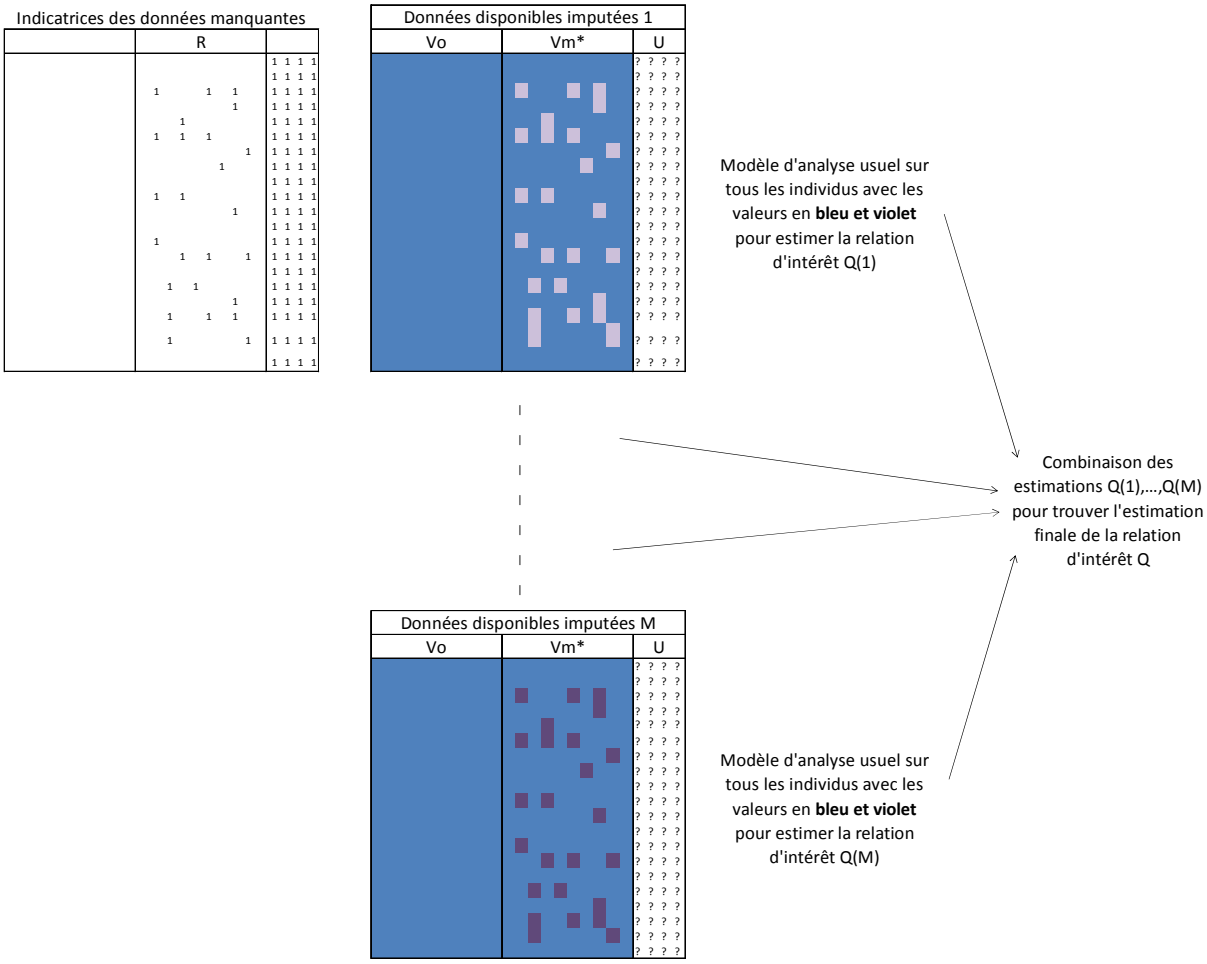
Le principal inconvénient de cette méthode vient du fait qu'elle devient difficile à appliquer lorsqu'on a beaucoup de combinaisons différentes de données manquantes, car en travaillant uniquement sur les cas complets, peu d'individus doivent alors représenter une population importante. Elle peut aussi devenir instable si les données sont trop déséquilibrées ou si on ajuste le modèle de pondération sur trop de variables (si on a trop peu de données observées dans certaines catégories) et peut nécessiter l'utilisation de pondérations plus complexes (non développées ici).

2.3.4.2 Les imputations multiples

2.3.4.2.1 Principes généraux

L'idée est d'imputer plusieurs fois les données manquantes afin de tenir compte de la variabilité des imputations effectuées. Contrairement aux imputations simples, les imputations multiples prennent en compte la variabilité des données en utilisant plusieurs valeurs simulées. Les M jeux de données imputés sont analysés séparément et les estimations sont ensuite recombinées pour obtenir un résultat final [98] (Figure 23).

Figure 23 - Représentation des données et de la méthode avec imputations multiples



Plus précisément, on procède de la façon suivante :

1. On impute les variables partiellement observées V^{m*} par $V^{m}(1)$ en complétant les valeurs non observées par des valeurs simulées $\widehat{V^{m}(1)}$ qu'on aura obtenues par les outils présentés dans la suite (cf. paragraphe 2.3.4.2.2, p 59 et paragraphe 2.3.4.2.3, p 61) :

$$V^{m}(1)=V^{m*} \text{ si } R = 0$$

$$V^{m}(1)=\widehat{V^{m}(1)} \text{ si } R = 1$$

2. On recommence ces imputations M-1 fois pour produire $V^m(2), \dots, V^m(M)$. On dispose alors de M jeux de données avec les variables complétées $(V^o, V^m(j), R)_{j=1,\dots,M}$.

3. Pour chaque jeu j de données imputées, on estime la relation d'intérêt Q sur les variables $\left(V^o, V^m(j)\right)$ par $\widehat{Q(j)}$ avec une incertitude $\widehat{E(j)}$ (en général E est la variance)

4. Pour finir, on recombine ces estimations par les règles de Rubin [91] pour avoir une estimation globale de Q et de son incertitude E :

$$\hat{Q} = \frac{1}{M}\sum_{j=1}^{M}\widehat{Q(j)}$$

$$\hat{E} = \frac{1}{M}\sum_{j=1}^{M}\widehat{E(j)} + \frac{1}{M-1}\sum_{j=1}^{M}\left(\widehat{Q(j)} - \hat{Q}\right)^2$$

Pour $\hat{E}$, la première partie correspond à l'incertitude intra-analyse et la deuxième partie correspond à l'incertitude inter-analyses.

Le point délicat de cette méthode est de fournir des imputations acceptables dans le sens où elles permettent d'obtenir une estimation ultérieure de la relation d'intérêt Q non biaisée. C'est ici qu'intervient la représentation par un DAG qui permet de déterminer les ajustements nécessaires.

L'objectif des imputations n'est pas d'obtenir la meilleure prédiction possible, mais de créer des valeurs simulées qui déformeront le moins les vraies distributions qui permettent de calculer la relation d'intérêt.

Deux outils principaux sont disponibles pour imputer les données :

- Le modèle gaussien joint : on suppose que les variables partiellement observées suivent un modèle joint gaussien
- Les équations chainées : on impute cycliquement une à une les variables partiellement observées en fonction des autres variables imputées dans les cycles précédents et des variables complètement observées. On ne spécifie pas de modèle joint mais des modèles conditionnels (d'où l'autre appellation « fully conditional specification »)

Ces deux méthodes s'interprètent naturellement avec l'approche bayésienne ; cependant, Rubin fournit des conditions fréquentistes pour que ces modèles soient valides [91].

2.3.4.2.2 Modèle d'imputation gaussien

2.3.4.2.2.1 Principe

Cette méthode a été proposée initialement par Schafer [97].

On suppose qu'on est en présence d'une distribution jointe gaussienne pour les variables V^m (qu'elles soient continues ou catégorielles). Cette hypothèse ne porte pas sur le mécanisme de données manquantes mais bien sur la distribution globale des données afin de permettre d'imputer (étapes 1. et 2. du paragraphe précédent).

L'avantage de cette hypothèse est que toutes les distributions marginales conditionnelles sont gaussiennes et qu'on connait la distribution *a posteriori* des estimateurs. On se place donc dans un cadre paramétrique avec les paramètres β, Σ correspondants à la distribution gaussienne des variables V^m conditionnellement à V^o :

$$V^m | V^o, \beta, \Sigma \sim N(\beta, \Sigma)$$

On procède itérativement sur k en deux étapes successives sur $V^m_{(k)} = \left(V^m_{j,(k)}\right)_{j=1,\dots;l}$, $\beta_{(k)}$ et $\Sigma_{(k)}$. A l'étape k, cela donne :

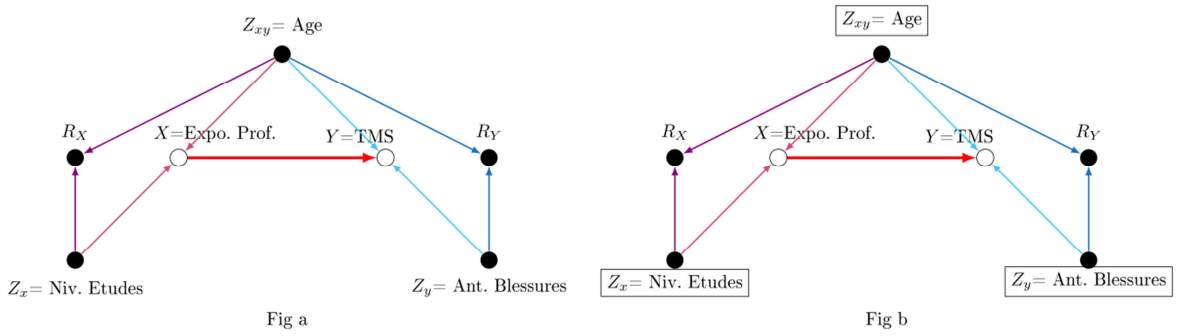
- Etape I (Imputation) : on impute le vecteur de variables $V^m_{(k+1)}$ à partir de la distribution $V^m | V^o, \beta_{(k)} \sim N(\beta_{(k)}, \Sigma_{(k)})$
- Etape P (« Posterior ») : on recalcule $\beta_{(k+1)}$ et $\Sigma_{(k+1)}$ à partir des distributions connues (respectivement gaussienne et de Wishart) conditionnellement à $V^o, V^m_{(k+1)}$

La suite $\left(V^m_{(k)}, \beta_{(k)}, \Sigma_{(k)}\right)$ converge vers la distribution de $V^m, \beta, \Sigma | V^o$. Ainsi on obtient un premier jeu d'imputations pour les variables partiellement observées V^m après un nombre de cycles k suffisamment important (souvent au bout de 1000). Ensuite, pour obtenir les jeux de données imputés suivants, on sélectionne les imputations obtenues $V^m_{(k)}$ tous les 50 cycles afin de garantir l'indépendance entre les différents jeux de données imputés. En résumé, on sélectionne les jeux de données imputés $V^m(j)$ parmi les $V^m_{(k)}$ avec $V^m(j) = V^m_{(1000+50j)}$.

2.3.4.2.2.2 Exemple

On reprend l'exemple 7 avec l'exposition professionnelle $X = V^m_1$ et les TMS $Y = V^m_2$ qui sont recueillis par deux auto-questionnaires successifs avec des données manquantes. On dispose de trois causes du mécanisme de données manquantes : l'âge (Z_{xy}), le niveau d'études (Z_x) et les antécédents de blessures (Z_y) (Figure 24). On veut estimer la relation ajustée $Q(x) = P(Y = 1 | Z_{xy}, X = x)$.

Figure 24 - DAG des relations causales (Fig. a - sans conditionnement, Fig. b - avec conditionnement)



On sait par la décomposition de la *recoverability* que pour estimer sans biais la relation d'intérêt, il suffit de prendre en compte Z_{xy} *et* Z_y. Cependant, afin d'améliorer les performances des imputations en terme de variance, on peut utiliser Z_x qui prédit X.

On suppose dans cette méthode que la distribution de (X, Y) est gaussienne avec :

$$(X, Y)|Z_x, Z_{xy}, Z_y \sim N(\beta, \Sigma)$$

Remarque 1 : On voit que dans le modèle gaussien tous les prédicteurs considérés imputent toutes les variables partiellement observées : ici, X est prédit par Z_x, Z_{xy}, Z_y et Y est prédit par Z_x, Z_{xy}, Z_y. Or d'après le DAG, il suffirait d'utiliser Z_x, Z_{xy} pour imputer X et Z_{xy}, Z_y pour imputer Y.

Une fois qu'on a effectué l'algorithme d'optimisation pour imputer X *et* Y, on obtient les M imputations $X(1), \dots X(M)$ *et* $Y(1), \dots, Y(M)$ et on peut faire les étapes 3 et 4 de la procédure d'imputations multiples, c'est-à-dire, avec un modèle paramétrique :

$$logit\ P\big(Y(j) = 1 \big| Z_{xy}, X\big) = \gamma_1(j) + \gamma_2(j) Z_{xy} + \gamma_3(j) X(j)$$

Ainsi, on a l'estimation de $Q(j)$ par $\widehat{Q(j)}(x) = \widehat{\gamma_3}(j) * x$ avec la variance associée $\widehat{E(j)}$ pour chacun des M jeux de données imputés. On en déduit l'estimation de $Q(x)$ par la moyenne des estimations $\widehat{Q(j)}(x)$ et la variance associée avec les règles de Rubin.

2.3.4.2.2.3 Conclusion sur la méthode d'imputation gaussienne

La méthode d'imputation gaussienne fournit une façon simple et plutôt robuste d'imputer les données. Elle est sans biais dans le cas gaussien avec un mécanisme de données manquantes MAR puisqu'on impute les variables partiellement observées uniquement à partir des variables complètement observées. Elle montre aussi d'assez bonnes propriétés avec des variables non gaussiennes [99].

L'inconvénient de cet outil est qu'il peut présenter une variance importante due à l'obligation d'avoir un modèle de forme identique pour toutes les variables partiellement observées (cf. remarque 1).

Si on est en présence de variables non gaussiennes (en particulier de variables catégorielles non-ordonnées), on peut tout de même craindre une distorsion des associations et des distributions par l'hypothèse gaussienne jointe. C'est pour cette raison que la méthode suivante a été développée.

2.3.4.2.3 Méthode d'imputation par « chained equation » (MICE)

2.3.4.2.3.1 Principe

Cette méthode d'imputation propose une alternative flexible à la méthode précédente dans le cadre de données manquantes de natures diverses (binaires, continues, catégorielles, ordinales) [100–104]. Dans ce cas, il est difficile voire impossible d'écrire la distribution jointe des données V^m.

La solution proposée ici est de modéliser une à une les variables $\left(V_j^m\right)_{j=1,\dots,l}$ conditionnellement aux autres variables avec un ensemble de paramètres $\left(\theta_j\right)_{j=1,\dots,l}$:

$$\forall j = 1 \dots l \quad P\left(V_j^m \middle| V^o, V_1^m, \dots, V_{j-1}^m, V_{j+1}^m \dots, V_l^m, \theta_j\right)$$

Dans un cycle k, on impute successivement toutes les variables partiellement observées. Sur le k^ième^ cycle, on a pour les variables imputées $V_{(k)}^m$ et les paramètres $\theta_{(k)}$, en notant $V_{-j,(k)}^m = \left(V_{1,(k)}^m \dots, V_{j-1,(k)}^m, V_{j+1,(k)}^m \dots, V_{l,(k)}^m\right)$:

$\theta_{1,(k)}$ est évalué à partir de la distribution de $\theta_1 | V^o, V_{-1,(k-1)}^m$

$V_{1,(k)}^m$ est imputé à partir de la distribution de $V_1^m | V^o, V_{-1,(k-1)}^m, \theta_{1,(k)}$

…

$\theta_{l,(k)}$ est évalué à partir de la distribution de $\theta_l | V^o, V_{-l,(k-1)}^m$

$V_{l,(k)}^m$ est imputé à partir de la distribution de $V_l^m | V^o, V_{-l,(k-1)}^m, \theta_{l,(k)}$

Ces cycles sont itérés jusqu'à la convergence vers la distribution conjointe des paramètres et des variables partiellement observées. Ensuite, pour obtenir les M imputations indépendantes $V^m(1), \dots, V^m(M)$ pour V^m, on récupère les imputations tous les 50 cycles (de même que pour les imputations dans le modèle gaussien), c'est-à-dire $V^m(j) = V_{(1000+50j)}^m$.

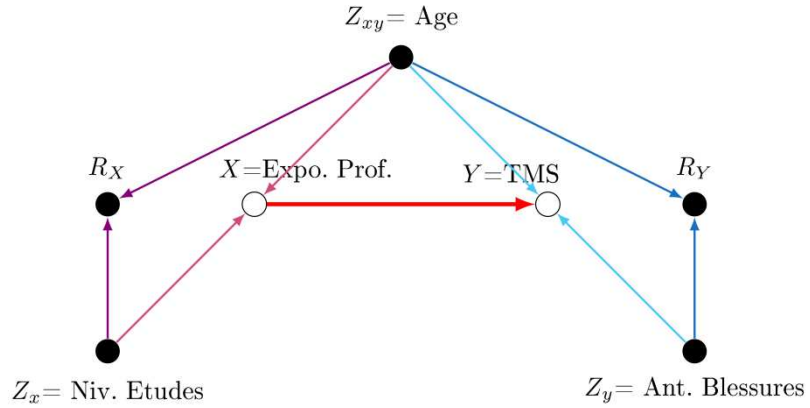
Ici, on a supposé que les variables sont toutes dépendantes les unes des autres. En pratique, on conditionnera uniquement sur les variables associées à la variable considérée dans le DAG proposé.

2.3.4.2.3.2 Exemple

2.3.4.2.3.2.1 Exemple MAR

On reprend le même exemple que pour la méthode d'imputation précédente (exemple 7, Figure 20, p 51). On a le même DAG que précédemment.

Figure 20 (duplicata) - DAG des relations causales



X est continue (supposée gaussienne) et sera modélisée par un modèle gaussien et Y est une variable binaire et sera modélisée par un modèle logistique. Les ajustements sont indiqués dans le DAG (toutes les variables prédictrices peuvent être inclues) :

$$E(X|Y, Z_{xy}, Z_x) = \beta_{11} + \beta_{12}Y + \beta_{13}Z_{xy} + \beta_{14}Z_x + \epsilon$$

$$logit\ P(Y = 1|X, Z_{xy}, Z_y) = \beta_{21} + \beta_{22}X + \beta_{23}Z_{xy} + \beta_{24}Z_y$$

Remarque : On voit qu'on impute X en fonction de Y car la relation causale va dans l'autre sens. On doit bien utiliser Y pour imputer X afin de respecter l'association qu'on suppose qu'il y a entre ces variables dans les données.

On voit que de la même façon que la méthode d'imputation précédente, on utilise Z_x pour imputer X afin de diminuer la variance alors que cette variable n'était pas nécessaire pour assurer l'estimation sans biais de la relation d'intérêt.

Une fois qu'on a appliqué l'algorithme d'optimisation pour imputer $X\ et\ Y$, on obtient les M imputations $X(1), \dots X(M)\ et\ Y(1), \dots, Y(M)$ et on peut effectuer les étapes 3 et 4 de la procédure d'imputations multiples, de même que dans la méthode gaussienne précédente (cf. paragraphe 2.3.4.2.2.2, p 59)

2.3.4.2.3.2.2 Exemple MAR séquentiel

En Annexe B , la méthode MICE est appliquée à un exemple de mécanisme MAR séquentiel, proche de celui utilisé dans les analyses ultérieures. Cette méthode fournit dans ce cas une estimation non biaisée de la relation d'intérêt puisqu'on fait intervenir des distributions dont on sait (par le DAG) qu'elles sont représentatives des données observées et non observées.

On peut généraliser cette propriété à tous les mécanismes MAR séquentiels : on montre que les estimations par imputations multiples avec équations chainées sont sans biais, à condition d'imputer les variables partiellement observées dans un ordre adéquat.

2.3.4.2.3.3 Conclusion sur la méthode d'imputation MICE

Il n'existe pas de cadre théorique satisfaisant pour justifier la convergence de la méthode d'imputation MICE, mais de nombreuses simulations ont montré de bonnes propriétés avec des résultats sans biais et des intervalles de couverture corrects [99].

Les modèles conditionnels $P\left(V_j^m \middle| V^o, V_{-j}^m, \theta_j\right)$ $\forall j = 1 \dots l$ offrent une grande flexibilité ; on peut choisir dans chaque modèle la distribution (loi binaire, loi de Poisson, loi gaussienne…), les variables explicatives et les fonctions de lien (logit, log, identité...). De plus, dans le cas d'une variable continue non gaussienne (par exemple, une variable bornée), il est possible utiliser une méthode de *predictive matching* qui, basé sur un modèle gaussien, va utiliser la valeur observée la plus proche de la prédiction obtenue par le modèle ce qui permet de simuler des valeurs imputées uniquement parmi les valeurs observées [104].

On a vu que les « chained equations » permettent d'obtenir des résultats non biaisés pour les mécanismes de données manquantes MAR séquentiels, ce qui représente un avantage très intéressant par rapport aux autres méthodes d'estimation proposées.

L'ensemble des paramètres des modèles conditionnels $\left(\theta_j\right)_{j=1\dots l}$ ne fournissent pas nécessairement une factorisation de $P(V^o, V^m|\theta)$. En réalité, il y a des chances que les θ_j soient corrélés entre eux et qu'on sur-paramétrise la distribution jointe. C'est pour cette raison qu'il faut faire attention à la concordance (ou *congeniality*) des modèles conditionnels entre eux, mais aussi des modèles avec le modèle d'analyse final [104]. Par exemple, si une variable X cause Y, il faut non seulement imputer Y avec X mais aussi imputer X avec Y.

2.3.4.3 Conclusion sur la gestion des données manquantes

Ce paragraphe a fait le tour d'horizon des méthodes de gestion des données manquantes les plus fréquemment utilisées en pratique. Elles sont valables pour des mécanismes très spécifiques (imputations simples), les mécanismes MCAR (cas complets), les mécanismes MAR (cas complets pondérés, imputations multiples) ou les mécanismes MAR séquentiels (imputations multiples par équations chainées). Leur réussite est donc conditionnée par :

- le choix et la disponibilité des variables qu'on doit prendre en compte pour avoir un mécanisme MAR (ou MAR séquentiel), et
- une modélisation appropriée (aspect peu abordé ici, mais il est essentiel de modéliser correctement les relations entre les variables : interactions, bonne fonction de lien...)

Les méthodes spécifiques aux mécanismes MNAR n'ont pas été abordées ici : il n'existe pas de méthode universelle pour traiter des données manquantes MNAR. Les paragraphes suivants vont explorer les analyses de sensibilité possibles pour évaluer les résultats dans le cas de données manquantes MNAR.

2.3.5 Analyses de sensibilité pour l'hypothèse MAR

Dans ce travail, les analyses de sensibilité pour l'hypothèse MAR vont supposer qu'on se situe en réalité sous l'hypothèse MNAR (contrairement à ce qui a été supposé dans les analyses initiale) : les analyses seront faites sous l'hypothèse MAR, puis des analyses de sensibilité seront implémentées avec des hypothèses précises sur les mécanismes MNAR plausibles.

Les paragraphes suivants précisent la nature de ces analyses dans le cadre des imputations multiples et d'un seul mécanisme supposé MNAR (les éventuels autres mécanismes étant supposés MAR).

La distribution totale des données V^o, V^m, R (données complètes et manquantes, mécanismes de données manquantes) peut être factorisée de deux façons différentes :

- Par un modèle de sélection : $P(V^o, V^m, R) = P(V^o, V^m)P(R|V^o, V^m)$
- Par un modèle « pattern-mixture » : $P(V^o, V^m, R) = P(V^o, V^m|R)P(R)$

Dans les cas MNAR, on voit que ces factorisations ne permettent pas d'« ignorer » le mécanisme qui engendre R et les analyses ultérieures nécessitent des hypothèses précises sur la nature de ce mécanisme. Les deux types de factorisations donnent lieu à des analyses différentes avec des hypothèses de natures différentes.

On dispose initialement d'imputations multiples sous l'hypothèse MAR (c'est-à-dire $V^m \coprod R \,|V^o$).

On suppose maintenant que l'on veut évaluer l'impact qu'aurait l'hypothèse MNAR sur la variable partiellement observée V_1^m et les données manquantes associées R_1.

2.3.5.1 Modèles de sélection

Dans ces modèles, on met l'accent sur le fait que le mécanisme qui génère les indicatrices R_1 dépend de V_1^m (c'est-à-dire qu'on a des données « sélectionnées ») et qu'il peut y avoir une corrélation résiduelle entre ces deux variables. En pratique, pour chaque jeu de données imputées MAR, on va créer une estimation de $P(R_1|V^o, V_1^m)$ de manière à prendre en compte cette corrélation (poids « MNAR ») puis on va pondérer les analyses de chaque jeu de données imputées MAR avec les poids « MNAR » pour se placer dans une situation MNAR.

De manière plus concrète, on procède de la façon suivante :

- On créé les imputations multiples $V_1^m(1), \dots, V_1^m(M)$ sous un DAG choisi avec l'hypothèse du mécanisme de données manquantes MAR.
- On suppose ensuite qu'on est dans le cas d'un mécanisme MNAR qui génère R_1. Ainsi, pour chaque jeu de données imputées j :
  - On modélise $logit\big(P(R_1 = 1|V^o, V_1^m(j))\big) = \alpha V^o + \delta V_1^m(j)$ avec δ fixé et α estimé sur les données imputées, on en tire une pondération « MNAR » notée $\widehat{P(j)}(\delta) = \mathrm{expit}(\alpha V^o + \delta V_1^m(j))$
  - On pondère l'analyse sur ce jeu de données imputées j par les poids « MNAR » $\widehat{P(j)}(\delta)$ et on en tire les estimations $\widehat{Q(j)}(\delta)$ de la relation d'intérêt sous l'hypothèse MNAR
- Pour finir, on combine les estimations $\widehat{Q(j)}(\delta)$ par les règles de Rubin (de la même façon que dans le cas MAR) pour obtenir une estimation $\widehat{Q(\delta)}$ sous l'hypothèse MNAR.

On remarque que si $\delta = 0$, alors on retrouve le mécanisme MAR. Si $\delta \neq 0$, alors on a un mécanisme MNAR. Le paramètre δ n'est pas estimable à partir des données disponibles. En général, plusieurs valeurs de δ sont choisies pour imputer. La valeur du paramètre δ indique à quel point le mécanisme s'éloigne d'un mécanisme MAR.

Le modèle de sélection apporte une restriction importante : les estimations α du modèle de sélection sont indépendantes des estimations faites sur $P(V^o, V^m)$ [105].

Cette méthode se concentre sur le mécanisme de données manquantes, elle serait donc plus « naturelle » dans le cadre de modèles de pondération des cas complets.

2.3.5.2 Modèles « pattern-mixture »

Dans ces modèles, on met l'accent sur le fait que la distribution des données manquantes $V^m|V^o$ est différente entre les unités observées et les unités manquantes, c'est-à-dire que c'est un mélange de « patterns » [91,97,106,107]. C'est méthode est plus facile à implémenter que la précédente dans le cadre d'imputations multiples, c'est pour cette raison qu'elle sera préférée dans ce travail. En pratique, on va donc modifier les imputations des unités manquantes avec un écart δ aux imputations initiales, puis appliquer la procédure d'analyse usuelle pour des imputations multiples.

La forme de cet écart dépend du modèle utilisé pour V_1^m. Ainsi, la procédure dépend de la nature de V_1^m. On indique la procédure à suivre dans les cas d'une variable V_1^m gaussienne et d'une variable V_1^m binaire.

2.3.5.2.1 Cas d'une variable gaussienne

On procède de la façon suivante :

- On créé les imputations multiples $V_1^m(1), \dots, V_1^m(M)$ sous un DAG choisi avec l'hypothèse du mécanisme de données manquantes MAR.
- On suppose ensuite qu'on est dans le cas d'un mécanisme MNAR qui génère R_1. Ainsi, pour chaque jeu de données imputées j :
  - On ajoute un δ à la moyenne des valeurs imputées initialement pour chaque individu :
    - On tire aléatoirement pour chaque individu un Δ tel que $\Delta \sim N(\delta, \sigma)$ où δ, σ sont fixés
    - On calcule les nouvelles imputations (pour les unités imputées) notées $V_1^m(j)(\delta) = V_1^m(j) + \Delta$
- On effectue l'analyse sur les nouveaux jeux de données imputées MNAR $(V_1^m(j)(\delta), \mathrm{V^o})_{\mathrm{j=1,\dots,M}}$ et on en tire les estimations $\widehat{Q(j)}(\delta)$ de la relation d'intérêt sous l'hypothèse MNAR
- Pour finir, on combine les estimations $\widehat{Q(j)}(\delta)$ par les règles de Rubin pour obtenir une estimation $\widehat{Q(\delta)}$ sous l'hypothèse MNAR.

2.3.5.2.2 Cas d'une variable binaire

On procède de la façon suivante :

- On créé les imputations multiples $V_1^m(1), \dots, V_1^m(M)$ sous un DAG choisit avec l'hypothèse du mécanisme de données manquantes MAR.

- On suppose ensuite qu'on est dans le cas d'un mécanisme MNAR qui génère R_1. Ainsi, pour chaque jeu de données imputées j :

  - On ajoute un δ à la probabilité d'avoir $V_1^m = 1$ pour chaque individu :

    - On modifie la probabilité MAR $P(V_1^m(j) = 1|V^o)$, en une nouvelle probabilité notée $P^\delta(V_1^m(j) = 1|V_i^o)$ de la façon suivante :
      - $logit\ P^\delta(V_1^m(j) = 1|V^o) = logit\ P(V_1^m(j) = 1|V^o) + \delta$
    - On calcule les nouvelles imputations (pour les unités imputées) notées $V_1^m(j)(\delta)$ en tirant dans la nouvelle loi de probabilité $P^\delta(V_1^m(j) = 1|V^o)$

- On effectue l'analyse sur les nouveaux jeux de données imputées MNAR $(V_1^m(j)(\delta), \mathrm{V^o})_{\mathrm{j=1,\dots,M}}$ et on en tire les estimations $\widehat{Q(j)}(\delta)$ de la relation d'intérêt sous l'hypothèse MNAR.

- Pour finir, on combine les estimations $\widehat{Q(j)}(\delta)$ par les règles de Rubin pour obtenir une estimation $\widehat{Q(\delta)}$ sous l'hypothèse MNAR.

Dans l'application de ce travail, afin de rendre le paramètre δ plus facilement interprétable, le choix a été fait de choisir δ de la façon suivante : $P^\delta(V_1^m(j) = 1|V^o) = P(V_1^m(j) = 1|V^o) + \delta$. Ainsi, on peut interpréter δ comme la différence entre la probabilité d'avoir $V_1^m = 1$ sous l'hypothèse MAR et la probabilité d'avoir $V_1^m = 1$ sous l'hypothèse MNAR. Ce choix impose cependant de tronquer les nouvelles probabilités créées entre 0 et 1.

2.3.5.2.3 Choix du paramètre δ

Comme δ est inconnu et n'est pas estimable, on choisit plusieurs valeurs plausibles pour δ afin de voir la sensibilité des analyses à ce paramètre inconnu.

On peut également prendre un paramètre δ différent selon une variable complètement observée V_1^o si on pense que les relations entre V_1^m et V_1^o sont différentes parmi ceux qui ont des données manquantes pour V_1^m ($R_1 = 1$). Par exemple, si on dispose de l'âge pour tous les individus, on peut faire dépendre δ de l'âge.

2.3.5.2.4 Cas de plusieurs variables manquantes

Cette méthode est généralisable dans le cas de plusieurs variables à imputer. Avec des imputations par « chained equations » pour vérifier la sensibilité à un mécanisme de données manquantes MNAR sur une variable donnée V_1^m, il faut en principe changer les imputations pour V_1^m dans chaque cycle itératif car les imputations de V_1^m influencent les imputations des autres variables $V_2^m, \dots, V_l^m$.

2.3.6 Conclusion générale sur la gestion des données manquantes

Cette partie a fait un tour d'horizon des hypothèses théoriques mises en œuvre dans le cadre des données manquantes, en particulier avec l'aide des DAGs, et des méthodes pratiques d'estimations en fonction des hypothèses. Tous les aspects sur la gestion des données manquantes n'ont pas été abordés ici mais ceux qui sont nécessaires à la compréhension des analyses ultérieures ont été détaillés.

La partie suivante porte sur les aspects plus concrets des analyses effectuées : les populations d'études et les analyses précisément utilisées selon les problématiques.

3 Matériel et méthodes

Pour rappel, dans ce travail, on s'intéresse à la prévalence et l'incidence des symptômes douloureux aux coudes et des épicondylites et à l'incidence des douleurs aux genoux. On cherche dans les trois études épidémiologiques à évaluer les associations avec les expositions professionnelles.

3.1 Cohorte des Salariés Ligériens (Cosali)

La population Cosali a été recrutée parmi les salariés de la région Pays de la Loire et a été suivie une fois depuis son inclusion (phase initiale et phase de suivi). La région Pays de Loire représentait 4,8% des actifs en France en 1999 (soit 1 289 937 actifs, recensement INSEE 1999) et sa structure socio-professionnelle était comparable à celle d'autres régions. Les modalités de recrutement et de suivi sont détaillées dans les paragraphes suivants.

3.1.1 Phase initiale : Réseau expérimental de surveillance épidémiologique des troubles musculo-squelettiques dans les Pays de la Loire

La première phase a eu lieu entre le 1er avril 2002 et le 30 avril 2005 [108,109]. Elle a été basée sur la constitution d'un « réseau sentinelle » de médecins du travail volontaires de la région Pays de la Loire qui ont recruté aléatoirement les participants lors des visites annuelles de médecine du travail.

Les participants étaient invités à remplir un auto-questionnaire dans la salle d'attente avant de passer un examen clinique standardisé.

3.1.1.1 Le réseau de médecins du travail

Dans les Pays de la Loire, les services de santé au travail sont présents dans les cinq départements de la région (Loire-Atlantique, Maine et Loire, Sarthe, Mayenne, Vendée) et regroupaient environ 500 médecins du travail. C'était essentiellement des services inter-entreprises qui couvraient des secteurs d'activité très diversifiés. Le nombre de salariés surveillés par un médecin du travail était compris entre 2 800 et 3 200 pour un exercice à temps plein.

Le réseau était composé de 83 médecins du travail appartenant aux principaux services de santé au travail des cinq départements de la région et ayant participé à l'étude au moins une des quatre années. Ils étaient 30 en Loire-Atlantique, 19 en Maine-et-Loire, 13 en Sarthe, 13 en Vendée et 8 en Mayenne, et représentaient 16 % des médecins du travail de la région. La participation des médecins était basée sur le volontariat, mais ils étaient comparables aux autres médecins de la région en termes de temps de travail et des secteurs économiques couverts. Un renouvellement des médecins a eu lieu au cours des trois années : 69 ont participé en 2002, 60 en 2003 et 51 en 2004.

Chaque médecin devait inclure 15 à 30 salariés par année selon qu'il travaillait à temps plein ou à temps partiel. En moyenne, 44 salariés ont été inclus par médecin sur les trois ans (avec des extrêmes allant de 1 à 112).

3.1.1.2 Les participants recrutés

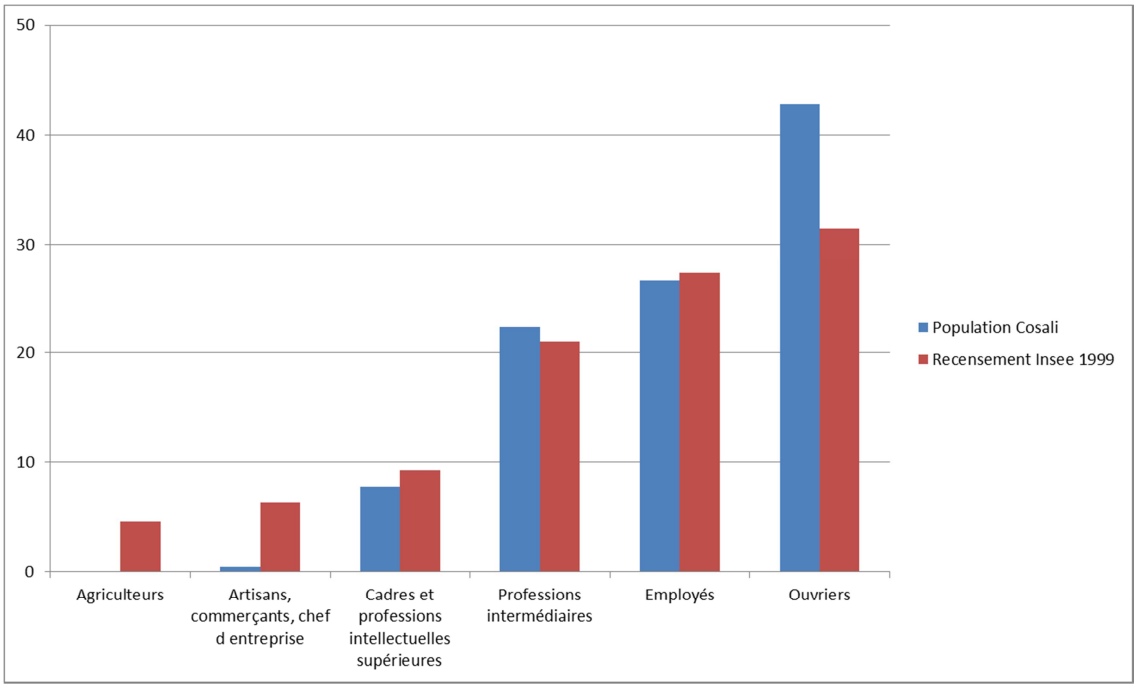
En France, tous les salariés, en particulier les intérimaires et les travailleurs à temps partiel, devaient faire annuellement une visite de médecine du travail. Les médecins volontaires avait reçu une formation pour inclure aléatoirement les travailleurs et effectuer un examen clinique standardisé spécifique aux TMS [20,108]. La population étudiée était ainsi composée de l'ensemble des salariés des entreprises publiques et privées des Pays de la Loire.

Les critères d'inclusion retenus étaient les suivants :

- salariés surveillés par l'un des médecins du réseau,
- âgés de 20 à 59 ans,
- ayant donné leur accord pour participer au réseau de surveillance épidémiologique,
- employés d'une entreprise privée ou publique localisée dans les Pays de la Loire mais dont le siège peut se situer ailleurs,
- quel que soit le statut d'emploi,
- souffrant ou non de TMS des membres ou du rachis.

Un total de 3 710 participants a été recruté sur la période considérée. Le pourcentage de participation à inclusion n'était pas connu même si il semblait qualitativement faible (peu de personnes refusaient d'être incluses d'après les médecins volontaires). La structure professionnelle des participants sélectionnés était comparable (non significativement distincte) à celle de la région Pays de la Loire si on excluait les agriculteurs qui n'étaient pas suivis par les médecins du travail (avec une sous-représentation des artisans et une surreprésentation des ouvriers) (cf. Figure 25).

Figure 25 - Comparaison de la répartition des catégories socio-professionnelles dans la cohorte Cosali avec le recensement INSEE de 1999



Tous les participants ont accepté de participer à l'étude et l'étude a été approuvée par la CNIL (Commission Nationale de l'Informatique et des Libertés).

3.1.1.3 L'auto-questionnaire initial

L'auto-questionnaire a été rempli dans la salle d'attente avant l'examen clinique, avec l'aide éventuelle du secrétaire médical ou du médecin. La durée de remplissage était de 30 à 45 minutes. Le recueil des données a porté à la fois sur l'état de santé des salariés et les expositions professionnelles [108]. Dans les paragraphes suivants, on donne le détail des données utilisées dans ce travail.

3.1.1.3.1 Facteurs de risque personnels

On disposait du sexe et de l'âge des participants. L'âge a été considéré en classes dans les analyses. L'IMC a été calculé à partir des déclarations de poids et de taille, il a ensuite été catégorisé selon les seuils proposés par l'OMS : sous-poids (inférieur strictement à 18,5), corpulence normale (18,5 à 25), surpoids (25 à 30) et obésité (supérieur à 30).

3.1.1.3.2 Les caractéristiques socio-professionnelles

On disposait des professions et catégories socioprofessionnelles (PCS) pour la plupart des individus. La PCS a été catégorisée en 4 classes : cadres et artisans (PCS 11 à 48), employés (PCS 52 à 56), ouvriers qualifiés (PCS 62 à 65) et ouvriers non qualifiés (PCS 67 à 69).

Le type de contrat a été recueilli sous les catégories suivantes : CDI, CDD, intérimaire, apprenti, stagiaire. L'ancienneté était disponible en 4 classes : moins d'un an, entre 1 et 3 ans, entre 3 et 10 ans et plus de 10 ans. Ces éléments ont été combinés pour avoir une information plus synthétique : statut temporaire (intérimaire, apprenti, stagiaire), CDD, CDI avec ancienneté de moins de 10 ans et CDI avec ancienneté de plus de 10 ans.

3.1.1.3.3 Facteurs de risque professionnels

De nombreux facteurs professionnels biomécaniques étaient disponibles dans cette étude. La plupart des questions portaient sur une journée typique de travail et étaient en 4 modalités : jamais ou presque jamais, rarement (moins de 2 heures par jour), souvent (2 à 4 heures par jour) et toujours (plus de 4 heures par jour) [20]. Certaines questions sur les gestes et les postures comportaient des images pour les illustrer. Dans ce travail, les questions suivantes ont été exploitées :

- Travail répétitif : « Votre travail nécessite-t-il habituellement de répéter les mêmes actions plus de 2 à 4 fois environ par minute? »
- S'agenouiller : « Devez-vous vous agenouiller ou vous accroupir ? »
- Port de charges : « Combien de temps passez-vous à porter une charge qui pèse : Moins de 10 kg ? 10 à 25 kg ? Plus de 25 kg ? »
- Manipulation de charges : « Manipulez-vous une charge ou un objet qui pèse : 1 à 4 kg ? Plus de 4 kg ? »
- Flexion/extension des coudes : « Combien de temps devez-vous fléchir le(s) coude(s) régulièrement ou de manière prolongée ? »
- Torsion des poignets : « Combien de temps devez-vous tordre le poignet ? »
- Utilisation d'outils vibrants : « Utilisez-vous des outils vibrants ou devez-vous poser la (les) main(s) sur des machines vibrantes ? »

Le questionnaire de Karasek dans sa version française était également posé [110,111] : il est constitué d'une série d'affirmations sur les perceptions au travail sur lesquelles le travailleur doit se prononcer du type « Les gens avec qui je travaille ont des attitudes hostiles ou conflictuelles envers moi : Tout à fait d'accord / D'accord / Pas d'accord / pas du tout d'accord ». A partir de ces questions, trois dimensions psycho-sociales ont été calculées : la demande psychologique (en terme de quantité et de complexité des tâches, et de contraintes temporelles), la latitude décisionnelle (en terme d'autonomie décisionnelle et d'utilisation des compétences) et le soutien social (en terme d'aide et de reconnaissance des collègues et des supérieurs hiérarchiques). Le « Job Strain » est défini comme une situation dans laquelle un travailleur est exposé à une forte demande psychologique et une faible latitude décisionnelle. Le « Job strain » et le faible soutien social ont été déterminés par les

seuils de l'étude SUMER 2005 qui fournit des données de cadrage sur les expositions de la population active française [108,110].

Une échelle de mesure de l'effort perçu au travail était aussi disponible : à la question, « Comment évaluez-vous l'intensité des efforts physiques de votre travail au cours d'une journée typique de travail ? », le participant devait donner un chiffre entre 6 (« Pas d'effort du tout ») et 20 (« Epuisant ») avec 13 étant « Un peu dur » et 15 étant « Dur » [112].

3.1.1.3.4 Les symptômes

Les symptômes musculo-squelettiques ont été évalués à l'aide d'une version française du questionnaire scandinave dit « Nordic » [113] (cf. Annexe E , p 163). Le questionnaire permettait d'évaluer l'existence de symptômes des membres supérieurs, du cou, et de la région lombaire, c'est-à-dire :

- l'existence au cours des 12 derniers mois et des 7 derniers jours de problèmes de type courbatures, douleurs, gêne ou engourdissement au niveau des zones anatomiques suivantes : nuque/cou, épaules/bras, coudes/avant-bras, poignets/mains, doigts, haut du dos, bas du dos, hanches/cuisses, genoux, chevilles/pieds ;
- la durée cumulée des troubles au cours des 12 derniers mois ;
- l'intensité du problème au moment de l'interrogatoire sur une échelle visuelle analogique.

3.1.1.4 L'examen clinique initial

En plus de l'examen clinique réalisé habituellement en consultation, un examen standardisé d'une durée de 2 à 15 minutes, en fonction de la présence ou non de symptômes musculo-squelettiques, a été effectué à l'aide du cahier d'examen clinique et du guide des manœuvres cliniques Saltsa, qui est une série de critères cliniques à évaluer pour détecter la présence de TMS [20].

Une formation théorique et pratique à la démarche clinique du consensus Saltsa a été proposée aux médecins du travail volontaires. Celle-ci était basée sur l'application de critères permettant de classer la sévérité des symptômes, d'orienter l'examen en fonction de la localisation des symptômes et de diagnostiquer les TMS.

La sévérité des TMS a été évaluée en trois niveaux en fonction des caractéristiques des symptômes :

- TMS latent, s'il existe des symptômes sans qu'il soit possible d'en préciser la survenue dans le temps,
- TMS symptomatique, si on retrouve des symptômes actuels, récents ou présents au moins 4 jours pendant au moins une même semaine au cours des 12 derniers mois. Par contre, il n'y a aucun signe positif à l'examen,

- TMS avéré s'il existe des signes positifs à l'examen et des symptômes actuels ou présents au moins 4 jours au cours de la semaine précédant l'examen.

Pour ce faire, le médecin devait suivre un arbre décisionnel. Il commençait par rechercher la présence d'une forme symptomatique. Si celle-ci était positive, il recherchait ensuite une forme avérée en réalisant différentes manœuvres cliniques.

Pour les épicondylites latérales, le symptôme d'intérêt était la présence de douleur intermittente, liée à l'activité manuelle, directement localisée dans la zone de l'épicondyle latéral. Les signes positifs à l'examen consistaient en une douleur localisée lors de l'extension contrariée du poignet.

Le médecin renseignait également la présence d'antécédents de TMS parmi les 6 suivants : syndrome de la coiffe des rotateurs, épicondylites latérales, syndrome du nerf ulnaire, syndrome du canal carpien, tendinite, maladie De Quervain.

3.1.2 Phase de suivi : Cosali

La mise en place du suivi longitudinal des salariés inclus dans la phase pilote du réseau a été effectuée entre le 1er janvier 2007 et le 9 octobre 2010 [114]. Cette cohorte, de type prospectif, a été baptisée COSALI, pour « COhorte des SAlariés Ligériens », et l'échantillon cible était constitué de l'ensemble des 3 710 salariés ayant participé à la phase pilote du réseau, théoriquement deux à neuf ans après leur participation initiale. Elle avait pour objectif :

- d'estimer les répercussions socio-professionnelles des TMS car elle inclut les participants actifs ou inactifs au suivi,
- d'étudier l'incidence des TMS car elle inclut les participants initialement indemnes de douleurs articulaires,
- d'étudier le pronostic des TMS puisqu'elle inclut les participants souffrant de TMS initialement.

L'inclusion dans la cohorte a été effectuée sous réserve que le salarié ait accepté d'être suivi, c'est-à-dire à condition qu'il ait renvoyé le formulaire de consentement de suivi.

3.1.2.1 L'auto-questionnaire de suivi

En 2007, un auto-questionnaire a été envoyé au domicile des 3 710 salariés participants. Un important travail d'actualisation des adresses des salariés inclus en 2002-2005 a été effectué auprès de leur ancien médecin du travail, puis par contrat auprès de La Poste (plus de 300 adresses ont ainsi été récupérées). En cas de non réponse, deux relances ont été envoyées. Au final, 2 332 (62.3%) participants ont renvoyé un questionnaire (Figure 26).

L'auto-questionnaire de suivi reprenait la plupart des questions de l'auto-questionnaire initial. Il y avait une question supplémentaire sur les changements professionnels depuis 2002 en 4 catégories : aucun changement, changement de poste uniquement, changement d'entreprise, plus en activité.

Différentes caractéristiques étaient associées à un meilleur suivi par questionnaire (Tableau 1) :

- des caractéristiques individuelles : être une femme, travailler en Vendée,
- des caractéristiques liées à l'intérêt que l'individu peut porter à l'étude : avoir eu un TMS dans le passé, avoir au moins une douleur articulaire,
- des caractéristiques associées à une certaine stabilité, ce qui améliore la capacité à pouvoir joindre ces individus : avoir plus de 45 ans, être cadre, avoir un emploi en CDI avec une ancienneté de plus de 10 ans,
- une caractéristique liée à l'organisation du suivi : avoir un syndrome de la coiffe des rotateurs (les salariés avec des problèmes d'épaule à l'inclusion ont été suivis avec plus de relances).

Tableau 1 - Caractéristiques des participants en fonction du suivi ou non par questionnaire

Variables (test du Khi-2 sur l'association avec le suivi)	Pas de questionnaire de suivi	Questionnaire de suivi	Total
Sexe (p=0.0018)			
Hommes	**848 (61,54%)**	1313 (56,30%)	2161 (58,25%)
Femmes	530 (38,46%)	**1019 (43,70%)**	1549 (41,75%)
Département (p=0.0470)			
Loire-Atlantique	437 (31,71%)	692 (29,67%)	1129 (30,43%)
Maine-et-Loire	328 (23,80%)	627 (26,89%)	955 (25,74%)
Mayenne	146 (10,60%)	229 (9,82%)	375 (10,11%)
Sarthe	**269 (19,52%)**	402 (17,24%)	671 (18,09%)
Vendée	198 (14,37%)	**382 (16,38%)**	580 (15,63%)
Age en 3 classes (p=<.0001)			
18-29 ans	**427 (30,99%)**	413 (17,71%)	840 (22,64%)
30-44 ans	593 (43,03%)	1053 (45,15%)	1646 (44,37%)
45-60 ans	358 (25,98%)	**866 (37,14%)**	1224 (32,99%)
IMC en 3 classes (p=0.3756)			
Sous-poids/Normal	886 (64,30%)	1446 (62,01%)	2332 (62,86%)
Surpoids	384 (27,87%)	694 (29,76%)	1078 (29,06%)
Obésité	108 (7,84%)	192 (8,23%)	300 (8,09%)
PCS (p=0.0076)			
Cadres/ NC	382 (27,72%)	**756 (32,42%)**	1138 (30,67%)
Employés	370 (26,85%)	616 (26,42%)	986 (26,58%)
Ouvriers qualifiés	359 (26,05%)	584 (25,04%)	943 (25,42%)
Ouvriers non qualifiés	**267 (19,38%)**	376 (16,12%)	643 (17,33%)
Contrat et ancienneté (p=<.0001)			
CDD	**78 (5,66%)**	85 (3,64%)	163 (4,39%)
Intérim, apprenti, stagiaire	**142 (10,30%)**	121 (5,19%)	263 (7,09%)
CDI et ancienneté < 10 ans	724 (52,54%)	1198 (51,37%)	1922 (51,81%)
CDI et ancienneté > 10 ans	434 (31,49%)	**928 (39,79%)**	1362 (36,71%)
Antécédents de TMS (p=<.0001)			
Non	**1162 (84,33%)**	1835 (78,69%)	2997 (80,78%)
Oui	216 (15,67%)	**497 (21,31%)**	713 (19,22%)
Douleurs à au moins une localisation (p=0.0010)			
Pas de douleur	**240 (17,42%)**	337 (14,45%)	577 (15,55%)
Douleurs ponctuelles (1-29 jours)	**704 (51,09%)**	1111 (47,64%)	1815 (48,92%)
Douleurs chroniques (>30 jours)	434 (31,49%)	**884 (37,91%)**	1318 (35,53%)
Syndrome de la coiffe des rotateurs initialement (p=<.0001)			
Non	**1311 (95,14%)**	2125 (91,12%)	3436 (92,61%)
Oui	67 (4,86%)	**207 (8,88%)**	274 (7,39%)
Total	**1378**	**2332**	**3710**

Dans l'étude sur l'incidence des épicondylites, les questionnaires de suivis utilisés étaient restreints à ceux qui étaient renvoyés au moins 3 mois avant l'examen clinique de suivi. On disposait alors de questionnaire pour 59,4% des participants (soit 2 203 participants, Figure 26). Les participants qui n'étaient plus actif lors du suivi étaient considérés comme n'étant pas exposés aux facteurs professionnels au moment du suivi (quelle que soit leur réponse effective au questionnaire).

3.1.2.2 L'examen clinique de suivi

En 2007, les salariés avec des problèmes d'épaule à l'inclusion ont été préférentiellement examinés puis entre 2008 et 2010, tous les salariés ont été suivis [114]. Dans la mesure du possible, chaque salarié a été revu par le même médecin du travail. Dans le cas où le salarié n'était plus suivi par le même médecin (pour cause de modifications de secteurs du médecin, de changement d'affectation à un service de santé au travail du médecin ou de changement d'activité du salarié), le nouveau médecin du travail concerné a été sollicité.

Certains participants n'ont pas souhaité être inclus dans le suivi de l'étude et d'autres n'ont pas été vus en consultation de médecine du travail au cours de la période de suivi. Différentes causes de non suivi sont disponibles (Tableau 2). Au final, 43,4% des participants ont eu un examen clinique de suivi (soit 1 611 participants, Figure 26).

Tableau 2 - Répartition du suivi et de causes de non suivi par sexe

Suivi et cause de non suivi pour l'examen clinique	Hommes	Femmes	Total
Suivi	921 (42.62%)	690 (44.54%)	1611 (43.42%)
Non Suivi car			
Pas de consultation sur la période 2007-2010	562 (26.01%)	391 (25.24%)	953 (25.69%)
Changement de situation*	168 (7.77%)	126 (8.13%)	294 (7.92%)
Cause inconnue	510 (23.60%)	342 (22.08%)	852 (22.96%)
Total	**2161**	**1549**	**3710**

* Décès, démission, entreprise fermée/radiée, inaptitude, invalidité, mutation, déménagement, refus de participer au suivi, congé parental, retraite

Tableau 3 - Caractéristiques des participants en fonction du passage ou non de l'examen clinique

Variables (test du Khi-2 sur l'association avec le suivi))	Pas d'examen clinique de suivi	Examen clinique de suivi	Total
Sexe (p=0.2432)			
Hommes	1240 (59.08%)	921 (57.17%)	2161 (58.25%)
Femmes	859 (40.92%)	690 (42.83%)	1549 (41.75%)
Département (p=<.0001)			
Loire-Atlantique	**680 (32.40%)**	449 (27.87%)	1129 (30.43%)
Maine-et-Loire	560 (26.68%)	395 (24.52%)	955 (25.74%)
Mayenne	181 (8.62%)	**194 (12.04%)**	375 (10.11%)
Sarthe	379 (18.06%)	292 (18.13%)	671 (18.09%)
Vendée	299 (14.24%)	**281 (17.44%)**	580 (15.63%)
Age en 3 classes (p=<.0001)			
18-29 ans	**579 (27.58%)**	261 (16.20%)	840 (22.64%)
30-44 ans	806 (38.40%)	**840 (52.14%)**	1646 (44.37%)
45-60 ans	**714 (34.02%)**	510 (31.66%)	1224 (32.99%)
IMC en 3 classes (p=0.3819)			
Sous-poids/Normal	1300 (61.93%)	1032 (64.06%)	2332 (62.86%)
Surpoids	622 (29.63%)	456 (28.31%)	1078 (29.06%)
Obésité	177 (8.43%)	123 (7.64%)	300 (8.09%)
PCS (p=0.1015)			
Cadres/ NC	638 (30.40%)	500 (31.04%)	1138 (30.67%)
Employés	531 (25.30%)	455 (28.24%)	986 (26.58%)
Ouvriers qualifiés	547 (26.06%)	396 (24.58%)	943 (25.42%)
Ouvriers non qualifiés	383 (18.25%)	260 (16.14%)	643 (17.33%)
Contrat et ancienneté (p=<.0001)			
CDD	**109 (5.19%)**	54 (3.35%)	163 (4.39%)
Intérim, apprenti, stagiaire	**224 (10.67%)**	39 (2.42%)	263 (7.09%)
CDI et ancienneté < 10 ans	1058 (50.40%)	**864 (53.63%)**	1922 (51.81%)
CDI et ancienneté > 10 ans	708 (33.73%)	**654 (40.60%)**	1362 (36.71%)
Antécédents de TMS (p=0.0832)			
Non	1675 (79.80%)	1322 (82.06%)	2997 (80.78%)
Oui	424 (20.20%)	289 (17.94%)	713 (19.22%)
Douleurs à au moins une localisation (p=0.5578)			
Pas de douleur	316 (15.05%)	261 (16.20%)	577 (15.55%)
Douleurs ponctuelles (1-29 jours)	1026 (48.88%)	789 (48.98%)	1815 (48.92%)
Douleurs chroniques (>30 jours)	757 (36.06%)	561 (34.82%)	1318 (35.53%)
Syndrome de la coiffe des rotateurs initialement (p=<.0001)			
Non	**1975 (94.09%)**	1461 (90.69%)	3436 (92.61%)
Oui	124 (5.91%)	**150 (9.31%)**	274 (7.39%)
Total	**2099**	**1611**	**3710**

Figure 26 - Suivis de la cohorte Cosali

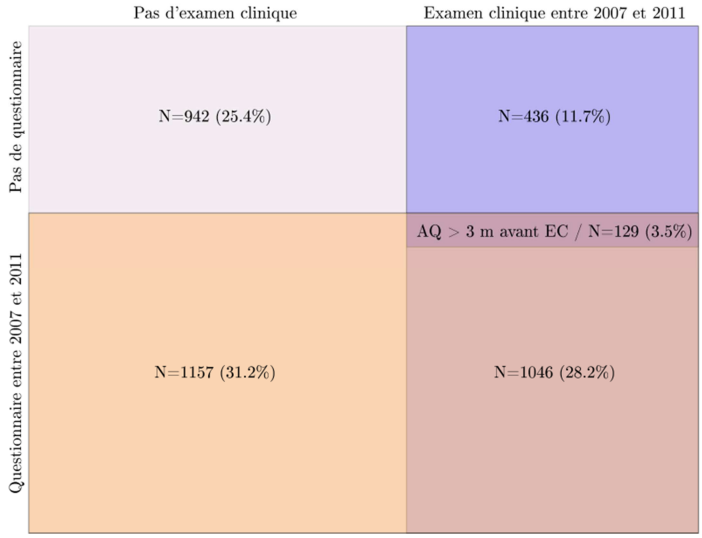
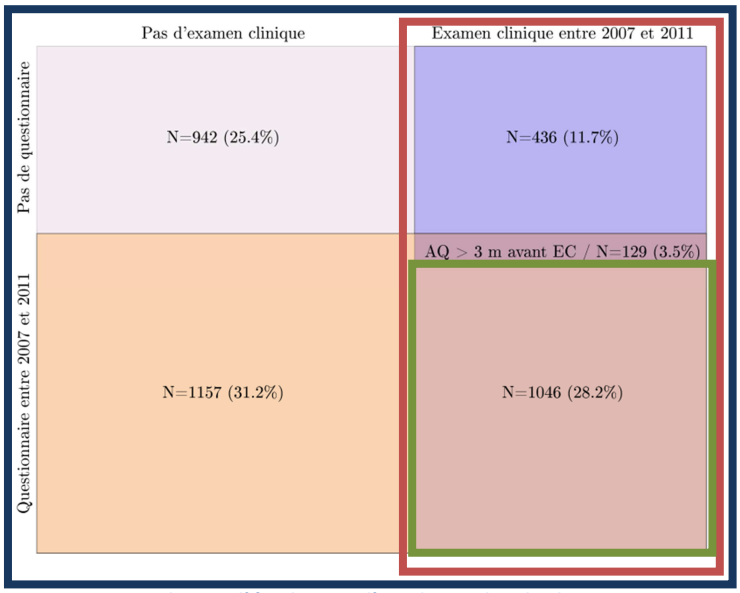


Figure 27 - Populations d'études pour l'incidence des épicondylites



Incidence des épicondylites et facteurs professionnels initiaux et au suivi – Données imputées

Incidence des épicondylites et facteurs professionnels initiaux – Cas complets

Incidence des épicondylites et facteurs professionnels initiaux et au suivi – Cas complets

Figure 28 - Population d'étude pour l'incidence des douleurs aux genoux

Incidence des douleurs aux genoux et facteurs professionnels initiaux – Cas complets

3.2 Les populations d'étude sélectionnées

Dans l'étude sur la prévalence des épicondylites, les analyses portaient sur les 3 710 participants initialement inclus (Figure 26).

Dans l'étude sur l'incidence des épicondyles (Figure 27), les analyses portaient sur :

- les 3 710 participants quel que soit leur suivi,
- les 1 611 participants avec un examen clinique de suivi,
- les 1 046 participants avec un examen clinique de suivi et un auto-questionnaire de suivi au moins 3 mois avant l'examen clinique.

Les cas prévalents de douleurs aux coudes lors de l'examen clinique initial ont été exclus des analyses sur l'incidence.

Dans l'étude sur l'incidence des douleurs aux genoux, les analyses portaient sur les 2 332 participants avec un auto-questionnaire de suivi (Figure 28).

Les cas prévalents de douleurs aux genoux déclarées dans l'auto-questionnaire initial ont été exclus des analyses sur l'incidence des douleurs aux genoux. De plus, les travailleurs en sous-poids ont été exclus des analyses puisque le sous-poids peut correspondre à des situations cliniques très hétérogènes et que l'IMC est un facteur de risque important de douleurs aux genoux [115].

3.3 Facteurs de risque utilisés dans les analyses

Tableau 4 - Facteurs étudiés dans les travaux présentés

	Analyse sur		
	La prévalence des épicondylites et des symptômes aux coudes	**L'incidence des épicondylites**	**L'incidence des douleurs aux genoux**
Premier auto-questionnaire			
Facteurs individuels	Age[a] IMC	Age[b]	Age[c] IMC
Caractéristiques cliniques initiales	Antécédent de TMS (ATCD TMS)		Douleurs à d'autres localisations[d]
Facteurs professionnels biomécaniques	Tâches répétitives** Effort perçu[e] Flexion/extension des coudes* Torsion des poignets* Utilisation d'outils vibrants**	Tâches répétitives** Effort perçu[e] Flexion/extension des coudes* Torsion des poignets*	Effort perçu[e] Port de charges*[f] Manipulation de charges de plus de 4 kg* S'agenouiller* Etre assis**
Facteurs professionnels psycho-sociaux	Soutien social[g] Job Strain[h]	Soutien social[f]	
Facteurs combinés	Effort perçu combiné avec torsion des poignets ou flexion/extension des coudes	Effort perçu combiné avec torsion des poignets ou flexion/extension des coudes	
Facteurs combinés des deux questionnaires[±]			
Facteurs professionnels biomécaniques	-	Tâches répétitives**	-
Facteurs professionnels psycho-sociaux	-	Soutien social[f]	-
Facteurs combinés	-	Effort perçu combiné avec torsion des poignets ou flexion/extension des coudes	-

[a] 18-29 ans, 30-49 ans, 50-60 ans
[b] 18-29 ans, 30-44 ans, 45-60 ans
[c] 18-44 ans, 45-60 ans
[d] en 3 classes : non = pas de douleurs, chronique = douleurs de plus de 30 jours dans les 12 derniers mois à une des 9 autres localisations (doigts, mains, coudes, épaules, nuque, haut du dos, bas du dos, hanches et chevilles) et autres = douleurs entre 1 et 29 jours dans les 12 derniers mois à une des 9 autres localisations
[e] en 2 classes sur l'échelle de Borg : léger = 6 à 13, fort = 14 à 20 (travail dur à épuisant)
[f] en 4 classes : Pas de port de charges, Port de charges de moins de 10 kg, Port de charges de 10 à 25 kg, Port de charges de plus de 25 kg
[g] en 2 classes, Faible soutien social (≤ 23), si inactif au deuxième auto-questionnaire, non exposé au deuxième questionnaire
[h] en 2 classes, Job Strain si faible latitude décisionnelle (≤ 70) et forte demande psychologique (≥ 22)
* plus de 2 heures/jour contre moins de 2 heures/jour, si inactif au deuxième auto-questionnaire, non exposé au deuxième questionnaire
** plus de 4 heures/jour contre moins de 4 heures/jour
[±] Facteurs combinés en 4 classes : jamais exposés, exposés au premier questionnaire, exposés au deuxième questionnaire, exposés au premier et deuxième questionnaire

Ces facteurs de risque ont été sélectionnés au vu des relations trouvées dans la littérature [26]. En particulier, l'effort perçu et les mouvements des coudes ont déjà été identifiés comme facteurs de risque de douleurs aux coudes et d'épicondylites [23,25,37,46], de même que les tâches répétitives [25,34,36,116] et le soutien social [37].

Ces variables ont été considérées en combinaison car certaines études transversales ont trouvé des interactions entre des facteurs similaires [25,37]. Ainsi, trois facteurs de risque ont été combinés en une variable : mouvements de flexion/extension du coude, mouvements de torsion du poignet et effort perçu. Les classes créées sont les suivantes :

- peu d'effort et pas de mouvement des coudes,
- peu d'effort et mouvements des coudes,
- effort important et pas de mouvement des coudes,
- effort important et 1 mouvement des coudes,
- effort important et 2 mouvements des coudes.

Une variable simplifiée regroupait les trois premières catégories en une dans l'analyse sur la prévalence. Dans l'analyse sur l'incidence, les deux dernières catégories ont également été regroupées (effort perçu important avec au moins un mouvement des coudes opposé à effort peu important ou pas de mouvement des coudes).

3.4 TMS et douleurs

L'évaluation des TMS par le médecin du travail et la déclaration des douleurs par le travailleur étaient disponibles pour chaque côté du corps. Quand les TMS (respectivement les douleurs) étaient présents des deux côtés du corps, on attribuait comme statut à l'individu le diagnostic le plus spécifique (respectivement la douleur la plus longue) des deux.

3.4.1 La prévalence des épicondylites et des douleurs de coudes

Dans l'étude sur la prévalence des épicondylites, la variable d'intérêt était catégorisée en trois classes : pas de symptôme aux coudes, symptômes aux coudes (cas non spécifiques, cas latents et cas symptomatiques) et épicondylites latérales (cas avérés). Les symptômes aux coudes ont été définis dans ce travail par « la présence de douleur intermittente, liée à l'activité manuelle, directement localisée dans la zone de l'épicondyle latéral » (cf. paragraphe 3.1.1.3.4, p 74). On utilisera le terme symptômes aux coudes sans préciser dans la suite la définition exacte à chaque occurrence.

3.4.2 L'incidence des épicondylites

Dans l'étude pour l'incidence des épicondylites, pour augmenter la puissance des analyses, on appelait « épicondylites » les cas avérés et les cas symptomatiques. On considérait donc les individus présentant ou ayant présenté une douleur au moins intermittente localisée dans la zone de l'épicondyle latéral au moins 4 jours sur 7 jours consécutifs durant les 12 derniers mois comme ayant une épicondylite.

L'incidence a été définie par les participants n'ayant pas déclaré de douleurs aux coudes initialement et ayant eu une épicondylite (cas avérés et cas symptomatiques) sur le suivi (cf. paragraphe 2.1.1, p 26).

3.4.3 L'incidence des douleurs aux genoux

Dans l'étude sur l'incidence des douleurs aux genoux, les douleurs aux genoux ont été considérées en deux catégories basées sur le questionnaire *Nordic*: les douleurs de plus de 30 jours dans les 12 derniers mois versus les douleurs de moins de 30 jours.

De la même manière que précédemment, l'incidence a été définie par les participants n'ayant pas déclaré de douleurs aux genoux initialement et ayant déclaré des douleurs aux genoux sur le suivi (cf. paragraphe 2.1.1, p 26).

3.5 Choix d'analyses dans toutes les études

3.5.1 Analyses séparées hommes/femmes

Les analyses ont été effectuées séparément pour les hommes et les femmes, en tenant compte des différences hommes-femmes sur les professions et les niveaux d'exposition [117,118].

Cependant, les analyses multivariées ont été choisies avec le même ajustement pour les hommes et les femmes afin de pouvoir comparer leurs risques respectifs dans les mêmes conditions.

3.5.2 Sélection des modèles multivariés

Les modèles ont été sélectionnés *a priori* sur une sélection de facteurs de risque connus ou suspectés. Les analyses se concentraient sur des facteurs de risque impliquant l'articulation considérée. Certains autres facteurs de risque biomécaniques pouvaient être significativement associés à la variable d'intérêt, mais non inclus dans les analyses multivariées si l'on a considéré qu'il s'agissait d'un signal d'un autre mouvement impliquant l'articulation. Les sélections des variables dans les modèles multivariés ont été illustrées par des DAGs représentant les hypothèses effectuées.

Dans toutes les analyses, les répartitions des facteurs de risque et des professions de la population étudiée ont été décrites. Quand il y avait des non suivis ou des données manquantes, ceux-ci ont également été décrits.

3.5.3 Logiciels utilisés

Les analyses statistiques ont été effectuées principalement avec SAS 9.3 puis 9.4 (SAS Institute Inc., Cary, NC, USA).

Les imputations multiples ont été générées avec Stata 10 et la commande mice (StataCorp. 2007. Stata Statistical Software: Release 10. College Station, TX: StataCorp LP.).

4 Etudes épidémiologiques et résultats

4.1 Prévalence des épicondylites latérales et des symptômes aux coudes

4.1.1 Analyses statistiques

Le diagramme causal de la Figure 29 représente les hypothèses retenues dans cette étude. On s'intéresse à la prévalence des épicondylites latérales et des symptômes aux coudes tels qu'ils ont été décrit dans le paragraphe 3.4.1 (p 82). Les associations avec les facteurs professionnels évalués au moment de la mesure de la variable d'intérêt sont explorées (cf. paragraphe 3.3, p 81).

Des modèles logistiques multinomiaux ont été utilisés pour étudier la présence d'épicondylites latérales contre l'absence de symptômes aux coudes, et la présence de symptômes aux coudes seuls contre l'absence de symptômes aux coudes.

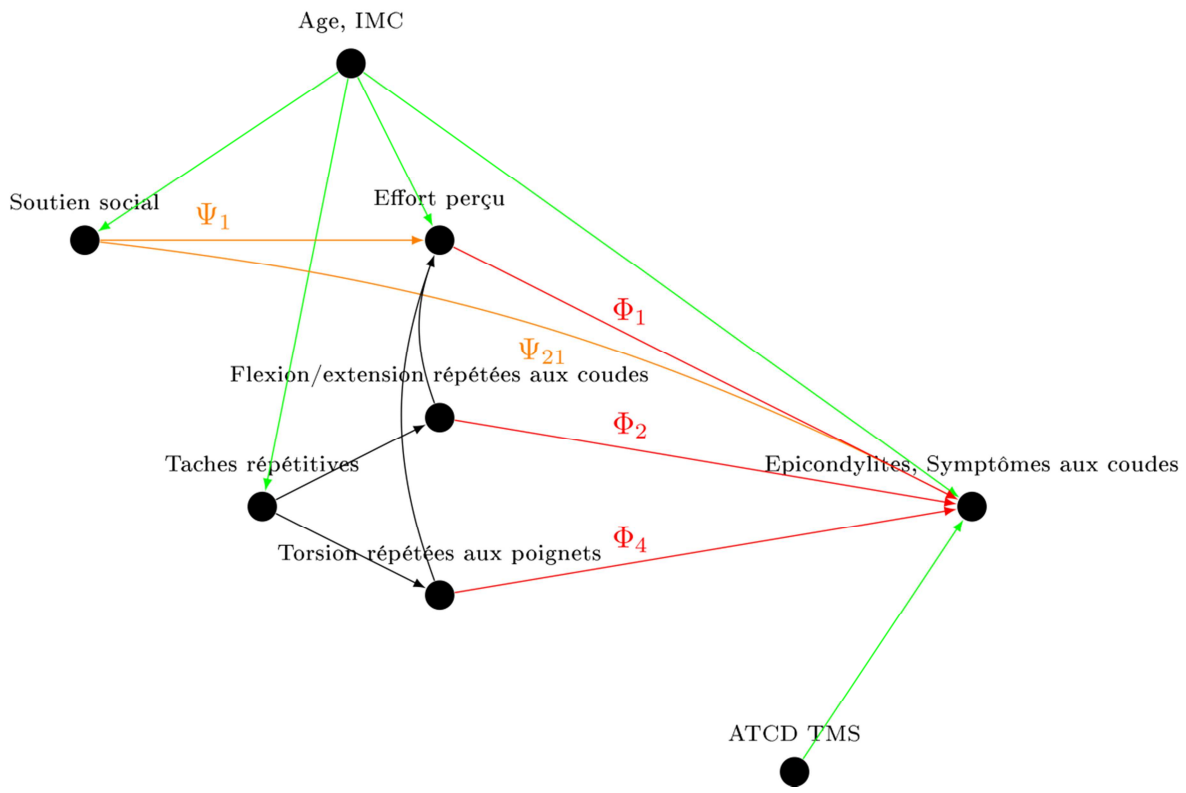
Trois modèles multivariés sont présentés :

- modèle 1 : un modèle prenant en compte l'âge, l'IMC, la répétition des tâches, l'effort physique perçu et le soutien social,
- modèle 2 : un modèle prenant en compte l'âge, l'IMC, la répétition des tâches, l'effort physique perçu combiné avec des mouvements des coudes (5 classes) et le soutien social,
- modèle 3 : un modèle prenant en compte l'âge, l'IMC, la répétition des tâches, l'effort physique perçu combiné avec des mouvements des coudes (3 classes) et le soutien social.

Un modèle supplémentaire a été utilisé mais non présenté dans les tableaux ajusté sur l'âge, l'IMC, la répétition des tâches, l'effort physique perçu, les mouvements des coudes et l'interaction entre les deux derniers facteurs.

Dans des analyses complémentaires, la population a été restreinte à ceux qui avaient plus de 10 ans d'ancienneté pour explorer les effets de l'âge et les effets des facteurs professionnels sur une population plus homogène.

Figure 29 - Hypothèses causales sur la prévalence des épicondylites et des symptômes aux coudes



4.1.2 Résultats

Parmi les 3 710 participants, 389 participants, 229 (10,6%) hommes et 160 (10,3%) femmes, présentaient des symptômes aux coudes pendant les 12 mois précédant l'examen physique. On trouvait également 90 participants, 51 (2,4%) hommes et 39 (2,5%) femmes, avec une épicondylite latérale à l'examen physique.

On observait des expositions plus importantes aux facteurs biomécaniques chez les hommes que les chez les femmes (effort physique perçu p<0,001, utilisation d'outil vibrant p<0,001) et à l'inverse une plus forte exposition des femmes à certains facteurs psychosociaux (job strain p<0,001) et au travail répétitif (p<0,001). 25,1% des ouvriers étaient exposés à un effort physique important combiné à des mouvements d'extension/flexion du coude ou de torsion du poignet contre seulement 10,5% des employés et 4,7% des cadres.

Les analyses univariées montraient que la probabilité d'avoir des symptômes aux coudes ou des épicondylites augmentait beaucoup avec l'âge, atteignant un OR de 11,0 pour les hommes de plus de 50 ans comparés aux hommes de moins de 30 ans (respectivement 8,7 pour les femmes – Tableau 5, Tableau 6). De plus, les OR des travailleurs de plus de 50 ans par rapport aux moins de 30 ans sont supérieurs pour les épicondylites que pour les symptomes aux coudes seuls (pour les hommes p=0,02 et pour les femmes p=0,02).

Les antécédents de TMS des membres supérieurs étaient fortement associés aux symptômes aux coudes (pour les hommes OR=5,5 et pour les femmes OR=4,2) et aux épicondylites (pour les hommes OR=5,8 et pour les femmes OR=8,5). L'association était plus forte avec les épicondylites qu'avec les symptômes aux coudes seuls pour les femmes (p=0,06).

On avait 44,4% des cas d'épicondylites (soit 40 cas) qui étaient associés à d'autres TMS dont 33 d'entre eux (82,5%) qui étaient associés à un syndrome de la coiffe des rotateurs ou à un syndrome du canal carpien. La probabilité d'avoir un TMS (hors épicondylites latérales) à l'examen physique (0,14, IC à 95%=[0,13 ;0,15]) était augmentée par la présence de symptômes aux coudes seuls (0,32, IC à 95% = [0,27 ;0,36]) et par la présence d'épicondylites latérales (0,44, IC à 95%=[0,34 ;0,55]).

Dans le modèle ajusté sur l'âge, l'IMC, les tâches répétitives, l'effort perçu et les mouvements aux coudes (non combinés), l'interaction entre l'effort perçu et les mouvements aux coudes était significative chez les hommes (p=0,0071) et non significative chez les femmes (p=0,7956) (modèle non présenté).

Tableau 5 - Analyses univariées des symptômes aux coudes ou des épicondylites parmi les hommes

	N	Symptômes aux coudes[±]				Epicondylites				P eq**
		N	OR	95% CI	p	N	OR	95% CI	p	
Facteurs personnels										
Age, en années					**0,0005**				**<0,0001**	
< 30	492	37	1,00			3	1,00			
30-49	1271	132	1,45	0,99-2,12		25	3,38	1,02-11,25		0,1841
≥ 50	398	60	2,33	1,51-3,59		23	11,00	3,27-36,95		**0,0166**
IMC en 3 classes					**0,0120**				0,2503	
Sous-poids, normal (<25)	1231	110	1,00			24	1,00			
Surpoids (25-30)	755	98	1,54	1,15-2,05		21	1,51	0,83-2,73		0,9564
Obésité (≥30)	175	21	1,42	0,86-2,33		6	1,85	0,75-4,61		0,6004
Antécédents de TMS					**<0,0001**				**<0,0001**	
Non	1782	124	1,00			27	1,00			
Oui	379	105	5,52	4,13-7,40		24	5,80	3,29-10,21		0,875
Facteurs biomécaniques professionnels										
Tâches répétitives (>4 h/j)					**0,0005**				0,139	
Non	1684	158	1,00			36	1,00			
Oui	477	71	1,71	1,27-2,31		15	1,59	0,86-2,93		0,8236
Effort physique perçu					**<0,0001**				**0,0022**	
Léger	993	76	1,00			13	1,00			
Fort	1168	153	1,86	1,39-2,49		38	2,70	1,43-5,11		0,1277
Mouvements de flexion/extension du coude (>2h/j)					**<0,0001**				**0,002**	
Non	1432	125	1,00			24	1,00			
Oui	729	104	1,79	1,35-2,36		27	2,41	1,38-4,22		0,331
Mouvements de torsion du poignet (>2h/j)					**0,0025**				**0,004**	
Non	1412	130	1,00			24	1,00			
Oui	749	99	1,54	1,16-2,03		27	2,27	1,30-3,97		0,2091
Mouvements des coudes					**0,0019**				**<0,0001**	
Non	1130	99	1,00			16	1,00			
Torsion du poignets ou flexion/extension du coude	584	57	1,14	0,81-1,61		16	1,99	0,99-4,00		0,1577
Torsion du poignets et flexion/extension du coude	447	73	2,11	1,52-2,92		19	3,40	1,73-6,68		0,2010
Utilisation d'outils vibrants (>2h/j)					0,1159				0,8918	
Non	1754	177	1,00			42	1,00			
Oui	407	52	1,30	0,94-1,81		9	0,95	0,46-1,97		0,4297
Mouvements des coudes* combinés à l'effort physique					**<0,0001**				**<0,0001**	
Peu d'effort physique et pas de mouvement des coudes	436	46	1,00			5	1,00			
Peu d'effort physique et 1 à 2 mouvements des coudes	299	23	0,70	0,47-1,07		2	1,34	0,46-3,90		0,2099
Effort physique important et pas de mouvement des coudes	694	53	0,70	0,42-1,19		11	0,56	0,11-2,92		0,9986
Effort physique important et 1 mouvement des coudes	382	38	0,97	0,61-1,52		15	3,51	1,26-9,76		**0,0083**
Effort physique important et 2 mouvements des coudes	350	69	2,20	1,47-3,29		18	5,27	1,93-14,37		0,0429
Facteurs psycho-sociaux										
Soutien social					0,0623				**0,0146**	
Fort	1322	128	1,00			23	1,00			
Faible	839	101	1,30	0,99-1,72		28	2,01	1,15-3,51		0,1636
Job strain					0,6860				0,1827	
Non	1726	181	1,00			37	1,00			
Oui	435	48	1,07	0,76-1,50		14	1,53	0,82-2,86		0,3144

[±] Présence de douleur intermittente à au moins un des coudes, liée à l'activité manuelle, directement localisée dans la zone de l'épicondyle latéral

* mouvements des coudes = mouvements de flexion/extension du coude ou de torsion du poignet plus de 2 heures/jour

**P eq = test d'égalité de Wald entre l'estimation pour les symptômes aux coudes seuls et l'estimation pour les épicondylites

Tableau 6 - Analyses univariées des symptômes aux coudes ou des épicondylites parmi les femmes

	N	Symptômes aux coudes[±]				Epicondylites				P eq**
		N	OR	95% CI	p	N	OR	95% CI	p	
Facteurs personnels										
Age, en années					**0,0079**				**0,0137**	
< 30	348	21	1,00			2	1,00			
30-49	909	103	2,04	1,25-3,32		24	4,99	1,17-21,22		0,2477
≥ 50	292	36	2,29	1,31-4,03		13	8,69	1,94-38,87		0,0989
IMC en 3 classes					0,1023				0,4915	
Sous-poids, normal (<25)	1101	102	1,00			29	1,00			
Surpoids (25-30)	323	41	1,43	0,97-2,10		9	1,10	0,52-2,36		0,5418
Obésité (≥30)	125	17	1,51	0,87-2,62		1	0,31	0,04-2,32		0,1336
Antécédents de TMS					**<0,0001**				**<0,0001**	
Non	1215	85	1,00			14	1,00			
Oui	334	75	4,21	2,99-5,92		25	8,52	4,36-16,63		0,0564
Facteurs biomécaniques professionnels										
Tâches répétitives (>4 h/j)					0,1948				**0,0058**	
Non	1068	104	1,00			19	1,00			
Oui	481	56	1,26	0,89-1,77		20	2,46	1,30-4,65		0,0631
Effort physique perçu					0,3155				**0,0021**	
Léger	861	84	1,00			12	1,00			
Fort	688	76	1,18	0,85-1,64		27	2,94	1,48-5,86		**0,0084**
Mouvements de flexion/extension du coude (>2h/j)					0,7396				**0,0029**	
Non	1064	109	1,00			18	1,00			
Oui	485	51	1,06	0,75-1,51		21	2,65	1,40-5,02		**0,0119**
Mouvements de torsion du poignet (>2h/j)					**0,0419**				**0,0366**	
Non	1062	99	1,00			21	1,00			
Oui	487	61	1,42	1,01-2,00		18	1,98	1,04-3,75		0,3593
Mouvements des coudes					0,1767				0,0105	
Non	859	85	1,00			13	1,00			
Torsion du poignets ou flexion/extension du coude	408	38	0,95	0,64-1,43		13	2,13	0,98-4,65		0,0661
Torsion du poignets et flexion/extension du coude	282	37	1,43	0,94-2,16		13	3,28	1,50-7,18		0,0580
Utilisation d'outils vibrants (>2h/j)					0,7469				0,2351	
Non	1487	153	1,00			36	1,00			
Oui	62	7	1,14	0,51-2,56		3	2,08	0,62-6,98		0,4010
Mouvements des coudes* combinés à l'effort physique					0,6437				0,0586	
Peu d'effort physique et pas de mouvement des coudes	597	61	1,00			7	1,00			
Peu d'effort physique et 1 à 2 mouvements des coudes	262	24	1,11	0,68-1,83		6	0,51	0,17-1,54		0,2047
Effort physique important et pas de mouvement des coudes	264	23	0,94	0,52-1,72		5	0,82	0,25-2,72		0,8115
Effort physique important et 1 mouvement des coudes	597	61	1,26	0,69-2,29		7	2,10	0,75-5,89		0,4671
Effort physique important et 2 mouvements des coudes	218	24	1,60	0,90-2,86		10	2,52	0,91-6,94		0,5692
Facteurs psycho-sociaux										
Soutien social					0,2834				0,9593	
Fort	980	95	1,00			25	1,00			
Faible	569	65	1,20	0,86-1,68		14	0,98	0,51-1,91		0,5888
Job strain					0,3263				0,0558	
Non	1120	111	1,00			23	1,00			
Oui	429	49	1,20	0,84-1,71		16	1,88	0,98-3,61		0,2164

[±] Présence de douleur intermittente à au moins un des coudes, liée à l'activité manuelle, directement localisée dans la zone de l'épicondyle latéral
* mouvements des coudes = mouvements de flexion/extension du coude ou de torsion du poignet plus de 2 heures/jour
**P eq = test d'égalité de Wald entre l'estimation pour les symptômes aux coudes seuls et l'estimation pour les épicondylites

L'effort physique combiné aux mouvements des coudes était significativement associé aux symptômes aux coudes et aux épicondylites, avec un risque supérieur pour l'effort physique important avec les deux mouvements au niveau des coudes et une relation dose-réponse significative à 10% dans les analyses multivariées sur les épicondylites (Tableau 7, Tableau 8 – modèle 3).

Les mouvements des coudes sans effort physique et l'effort physique important sans mouvement des coudes, comparés à l'absence d'effort physique sans mouvement des coudes n'étaient significativement pas associés aux symptômes aux coudes et aux épicondylites (Tableau 7, Tableau 8 – modèle 2).

Dans le modèle final (Tableau 7, Tableau 8 – modèle 3), la valeur de l'association était plus importante pour les épicondylites que pour les symptômes aux coudes seuls chez les hommes (non significatif pour les femmes).

Les associations trouvées restaient significatives pour la sous-population des hommes avec plus de 10 ans d'ancienneté (excepté pour la catégorie « moins de 30 ans » qui ne peut être estimée faute d'effectifs suffisants), avec un OR ajusté de l'effort physique important avec les deux mouvements des coudes de 2,3 [1,4 ;3,9] pour les symptômes aux coudes seuls et 4.5 [1.9 ;10.8] pour les épicondylites.

Le faible soutien social était le seul facteur psycho-social considéré (parmi les dimensions du questionnaire de Karasek) associé aux symptômes aux coudes et aux épicondylites (pour les hommes uniquement, Tableau 7).

Etre exposé à des tâches répétitives était associé aux symptômes aux coudes et aux épicondylites dans les analyses univariées, mais plus dans les analyses multivariées.

Tableau 7 - Analyses multivariées des symptômes aux coudes ou des épicondylites parmi les hommes

	Modèle 1				Modèle 2				Modèle 3				
	Symptômes aux coudes±		Epicondylites latérales		Symptômes aux coudes±		Epicondylites latérales		Symptômes aux coudes±		Epicondylites latérales		P eq**
	OR	95% CI	OR	95% CI	OR	95% CI	OR	95% CI	OR	95% CI	OR	95% CI	
Age, en années													
< 30	1,00		1,00		1,00		1,00		1,00		1,00		
30-49	1,40	0,95-2,07	**3,44**	**1,02-11,56**	1,42	0,96-2,11	**3,63**	**1,08-12,24**	1,42	0,96-2,11	**3,70**	**1,10-12,45**	0,1382
≥ 50	**2,17**	**1,38-3,41**	**11,38**	**3,30-39,25**	**2,28**	**1,44-3,61**	**13,55**	**3,90-47,03**	**2,30**	**1,46-3,64**	**13,71**	**3,96-47,51**	**0,0072**
IMC en 3 classes													
Sous-poids, normal (<25)	1,00		1,00		1,00	,	1,00		1,00		1,00		
Surpoids, obésité (>25)	**1,36**	**1,02-1,81**	1,10	0,61-1,96	**1,35**	**1,01-1,80**	1,12	0,63-2,02	**1,35**	**1,01-1,80**	1,13	0,63-2,03	0,5876
Tâches répétitives (>4 h/j)													
Non	1,00		1,00		1,00		1,00		1,00		1,00		
Oui	**1,57**	**1,15-2,13**	1,39	0,74-2,59	1,36	0,99-1,88	1,05	0,54-2,02	1,37	0,99-1,88	1,03	0,53-1,97	0,4222
Effort physique perçu													
Léger	1,00		1,00										
Fort	**1,75**	**1,30-2,36**	**2,57**	**1,34-4,90**									
Mouvements des coudes* combinés à l'effort physique													
Peu d'effort physique et pas de mvt des coudes					1,00		1,00						
Peu d'effort physique et 1 à 2 mvts des coudes					0,73	0,48-1,11	1,52	0,52-4,44	1,00		1,00		
Effort physique important et pas de mvt des coudes					0,76	0,45-1,29	0,76	0,14-3,98					
Effort physique important et 1 mvt des coudes					1,01	0,64-1,61	**4,55**	**1,60-12,88**	1,23	0,83-1,83	**3,78**	**1,85-7,70**	**0,0054**
Effort physique important et 2 mvts des coudes					**2,18**	**1,43-3,32**	**6,71**	**2,38-18,96**	**2,65**	**1,88-3,73**	**5,60**	**2,76-11,35**	**0,0517**
Soutien social													
Fort	1,00		1,00		1,00		1,00		1,00		1,00		
Faible	1,20	0,91-1,60	**1,86**	**1,05-3,28**	1,20	0,91-1,60	**1,98**	**1,11-3,52**	1,22	0,92-1,62	**1,96**	**1,11-3,48**	0,1332

± Présence de douleur intermittente à au moins un des coudes, liée à l'activité manuelle, directement localisée dans la zone de l'épicondyle latéral
* mvts des coudes = mouvements de flexion/extension du coude ou de torsion du poignet plus de 2 heures/jour
**P eq = test d'égalité de Wald entre l'estimation pour les symptômes aux coudes seuls et l'estimation pour les épicondylites

Tableau 8 - Analyses multivariées des symptômes aux coudes ou des épicondylites parmi les femmes

	Modèle 1				Modèle 2				Modèle 3				
	Symptômes aux coudes±		Epicondylites latérales		Symptômes aux coudes±		Epicondylites latérales		Symptômes aux coudes±		Epicondylites latérales		P eq**
	OR	95% CI	OR	95% CI	OR	95% CI	OR	95% CI	OR	95% CI	OR	95% CI	
Age, en années													
< 30	1,00		1,00		1,00		1,00		1,00		1,00		
30-49	**1,98**	**1,21-3,23**	**5,13**	**1,20-21,89**	**1,97**	**1,21-3,22**	**5,16**	**1,21-22,03**	**1,98**	**1,21-3,22**	**5,14**	**1,20-21,95**	0,2179
≥ 50	**2,15**	**1,21-3,79**	**9,04**	**2,00-40,78**	**2,12**	**1,20-3,76**	**9,19**	**2,03-41,53**	**2,14**	**1,21-3,78**	**9,15**	**2,03-41,30**	**0,0736**
IMC en 3 classes													
Sous-poids, normal (<25)	1,00		1,00		1,00		1,00		1,00		1,00		
Surpoids, obésité (>25)	1,35	0,96-1,92	0,69	0,33-1,46	1,34	0,95-1,91	0,67	0,32-1,41	1,34	0,95-1,90	0,67	0,32-1,41	0,091
Tâches répétitives (>4 h/j)													
Non	1,00		1,00		1,00		1,00		1,00		1,00		
Oui	1,21	0,85-1,73	**2,14**	**1,11-4,12**	1,17	0,81-1,71	1,80	0,91-3,59	1,15	0,79-1,66	1,86	0,94-3,69	0,2074
Effort physique perçu													
Léger	1,00		1,00										
Fort	1,10	0,78-1,54	**2,51**	**1,24-5,08**									
Mvts des coudes* combinés à l'effort physique													
Peu d'effort physique et pas de mvt des coudes					1,00		1,00		1,00		1,00		
Peu d'effort physique et 1 à 2 mvts des coudes					1,14	0,69-1,88	0,53	0,17-1,60					
Effort physique important et pas de mvt des coudes					0,94	0,51-1,73	0,80	0,24-2,71					
Effort physique important et 1 mvt des coudes					1,17	0,63-2,15	1,81	0,63-5,23	1,11	0,68-1,79	**2,54**	**1,12-5,76**	0,0764
Effort physique important et 2 mvts des coudes					1,47	0,80-2,68	2,06	0,72-5,93	1,39	0,87-2,23	**2,89**	**1,28-6,51**	0,1151
Soutien social													
Fort	1,00		1,00		1,00		1,00		1,00		1,00		
Faible	1,16	0,83-1,63	0,85	0,44-1,67	1,17	0,83-1,63	0,86	0,44-1,69	1,16	0,83-1,63	0,87	0,45-1,71	0,4469

± Présence de douleur intermittente à au moins un des coudes, liée à l'activité manuelle, directement localisée dans la zone de l'épicondyle latéral
* mvts des coudes = mouvements de flexion/extension du coude ou de torsion du poignet plus de 2 heures/jour
**P eq = test d'égalité de Wald entre l'estimation pour les symptômes aux coudes seuls et l'estimation pour les épicondylites

4.1.3 Discussion

Cette étude a souligné l'association des facteurs professionnels avec les épicondylites latérales et les symptômes aux coudes. Le principal résultat portait sur la forte association entre des expositions professionnelles spécifiques aux coudes combinant l'effort physique perçu et des mouvements répétés de flexion/extension des coudes ou de torsion des poignets, même après ajustement sur les facteurs de risque personnels.

Ces résultats confirment les études précédentes sur l'association entre les épicondylites et les facteurs de force et de répétition [34,37,45]. Les mouvements répétés au niveau des coudes ont déjà été mis en évidence comme la position de la main ou des mouvements de prono-supination de l'avant-bras [37,45]. Une précédente étude transversale suggérait également que la combinaison du travail en force avec des mouvements répétés est particulièrement associée aux problèmes aux coudes [37] et plusieurs études longitudinales ont mis en valeur le même genre d'exposition combinées après cette étude [119]. Dans une revue de la littérature de 2009, Van Rijn et al. ont souligné que les principaux facteurs de risque d'épicondylites latérales étaient la manipulation d'outils de plus de 1 kg (ORs de 2,1-3,0), la manipulation de charges de plus de 20 kg au moins 10 fois par jour (OR 2,6) et les mouvements répétitifs plus de 2 heures par jour (OR 2,8-4,7) [26], ce qui recoupe les facteurs trouvés ici.

L'utilisation d'outils vibrants plus de 2 heures par jour ne semblait pas être associée aux épicondylites latérales dans cette étude contrairement à ce qui avait déjà été trouvé précédemment [25,37]. Les diverses façons de définir la notion de vibration avec différents seuils d'exposition peuvent peut-être expliquer ces résultats discordants. Les résultats concernant la répétition diffèrent également d'une étude à l'autre et peuvent probablement s'expliquer par le même genre de variations dans la définition de l'exposition.

La prévalence des épicondylites dans la littérature varie entre 0,3% et 12,2% selon le degré d'exposition des populations considérées et les définitions des cas utilisées [27]. La prévalence de 2,4% dans cette étude représentative de la population au travail est concordante avec la prévalence de 1,6%-3,5% estimée dans une population finlandaise en 2006 [25].

Comme dans d'autres études sur d'autres pathologies du coude [24,29], les symptômes aux coudes et les épicondylites latérales étaient associés à d'autres troubles musculo-squelettiques, suggérant la complexité des tâches effectuées au travail qui impliquent différents mouvements et différentes articulations et ainsi peuvent provoquer des douleurs dans plusieurs zones du corps.

Les antécédents de troubles musculo-squelettiques des membres supérieurs étaient fortement associés aux épicondylites. Ce résultat a aussi été trouvé pour les symptômes aux coudes, indiquant peut-être des épisodes d'épicondylites passés pour ceux ayant des symptômes aux coudes. Ces résultats semblaient indiquer la récurrence des symptômes au cours du temps mais une étude longitudinale permettrait de mieux explorer ces aspects.

La force des associations selon une définition plus large (symptômes aux coudes) ou plus stricte (épicondylites) des pathologies aux coudes était similaire pour la plupart des expositions pour les deux définitions, comme dans d'autres études [21,23,33,34]. Cependant, l'effort physique perçu était plus fortement associé aux épicondylites qu'aux symptômes aux coudes seuls. L'âge était aussi plus fortement associé aux épicondylites, de manière concordante avec d'autres études [23], ce qui suggère que les symptômes aux coudes précèdent les épicondylites (et donc que les symptômes aux coudes sont des signes précurseurs d'apparition d'épicondylites). Ces facteurs de risque peuvent aussi jouer des rôles différents selon le stade des pathologies, mais d'autres études sont nécessaires pour apporter des réponses à cette question.

4.2 Incidence des épicondylites latérales

Dans ce paragraphe, on désigne par t_0 le temps de mesure de la phase initiale (entre 2002 et 2005) et par t_1 le temps de mesure de la phase de suivi (entre 2007 et 2009) (cf. paragraphe 3.1, p 70).

4.2.1 Analyses statistiques

Les analyses reprenaient les expositions professionnelles trouvées dans l'analyse précédente sur la prévalence des épicondylites et des symptômes aux coudes (cf. paragraphe 4.1, p 85). Les DAGs des Figure 30 et Figure 31 indiquent les relations causales qu'on suppose exactes.

Figure 30 - Hypothèses causales sur l'incidence des épicondylites

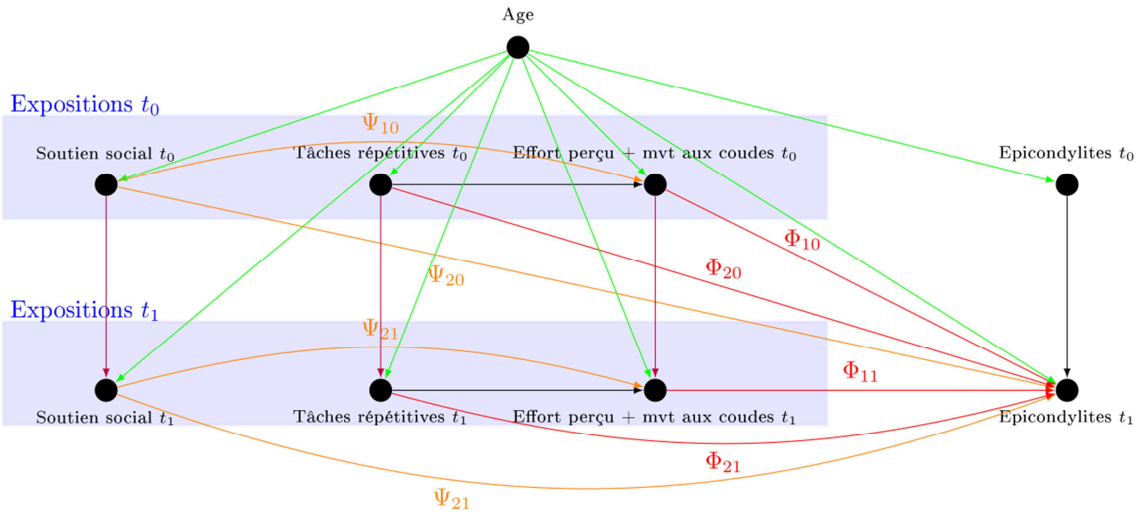
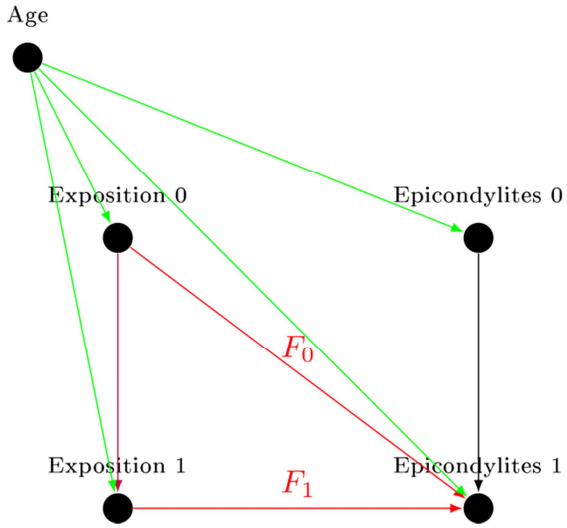


Figure 31 - Hypothèses causales sur l'incidence des épicondylites (simplifié)



Un point important dans ces analyses est qu'il y avait des données partiellement observées pour les expositions au suivi (Exposition 1) et les épicondylites au suivi (Epicondylites 1). Les analyses ont été effectuées sur les cas complets (valable si les données manquantes sont *missing completely at random*, MCAR) ou sur les données imputées avec des hypothèses *missing at random* (MAR)

séquentielles ou *missing not at random* (MNAR) (pour les définitions de ces hypothèses, cf. paragraphe 2.3.2.2, p 41).

Les taux d'incidence annuels ont été estimés en utilisant les données imputées MAR. La distribution des expositions mesurées en t_0 et en t_1 a aussi été décrite par catégorie socio-professionnelle à t_0 en utilisant les données imputées MAR séquentielles.

Les analyses ont été effectuées en trois temps : des analyses préliminaires (sur cas complets), des analyses principales (sur cas complets et sur les données imputées MAR séquentielles) et des analyses de sensibilité (sur les données imputées MNAR). Les populations utilisées pour chaque analyse ont été décrites avec la Figure 27, p 79.

Dans l'analyse préliminaire, seules les expositions déclarées lors de l'auto-questionnaire initial (à t_0) et leurs liens avec l'incidence au moment du suivi ont été étudiés (c'est-à-dire la flèche F_0 - Figure 31). Cette analyse correspondait au schéma d'analyse usuel avec ce type de données et portait sur les travailleurs ayant un suivi pour l'examen clinique.

Dans l'analyse principale, les expositions déclarées lors des auto-questionnaires initiaux et de suivi (à t_0 et t_1) ont été utilisés pour étudier les effets court terme et long terme des expositions professionnelles (c'est-à-dire les flèches F_0 et F_1 - Figure 31).

L'analyse sur données complètes sur les 3710 participants à l'inclusion, donne des résultats non biaisés dans le cas de données MCAR [98,104].

Dans l'analyse sur données multiples imputées, l'hypothèse initiale était que les deux mécanismes de données manquantes (pour l'examen clinique et pour l'auto-questionnaire) dépendaient des facteurs mesurés (les expositions professionnelles et les antécédents médicaux des participants). Autrement dit, on a supposé qu'on avait un mécanisme MAR séquentiel en fonction des variables observées, ou que les variables observées fournissaient des informations suffisantes pour imputer les données (cf. paragraphe 2.3.2.2, p 41 et paragraphe 2.3.2.3, p 42 pour plus de détail).

L'hypothèse MAR séquentielle a été élargie dans les analyses de sensibilité concernant le mécanisme de données manquantes associées à l'absence d'examen clinique à t_1. On a supposé plusieurs mécanismes MNAR spécifiques pour ce processus (cf. paragraphe 4.2.3.5, p 110 et Annexe D p 159). Le mécanisme de données manquantes pour l'auto-questionnaire est supposé MAR dans les analyses de sensibilité. En effet, une hypothèse MNAR sur les covariables du modèle aurait *a priori* moins d'impact sur les résultats qu'une hypothèse MNAR sur la variable à expliquer (c'est-à-dire les épicondylites).

Les variables suivantes de l'auto-questionnaire de suivi ont été imputées :

- les changements professionnels depuis 2002,
- les expositions professionnelles : l'effort physique important, les mouvements de flexion/extension du coude, les mouvements de torsion du poignet, le travail répétitif, le faible soutien social au travail,
- les douleurs aux coudes auto-rapportées.

En ce qui concerne l'examen clinique de suivi, les épicondylites latérales ou les symptômes aux coudes et le temps entre les examens cliniques ont été également imputés.

Les imputations ont été stratifiées par sexe et 100 jeux de données ont été générés [120].

Les associations entre les facteurs de risque et l'incidence des épicondylites latérales ont été estimées par des modèles de Poisson univariés et multivariés qui prennent en compte le temps de suivi ($t_1 - t_0$). Les résultats ont été exprimés en termes de rapport d'incidence (IRR, cf. paragraphe 2.1.1, p 26). Pour rappel, l'IRR pouvait alors s'interpréter comme le rapport entre le taux d'incidence annuel de la catégorie considérée et le taux d'incidence annuel de la catégorie de référence.

Plusieurs modèles multivariés ont été implémentés :

- Modèle 1 : modèle prenant en compte l'âge et le travail répétitif,
- Modèle 2 : modèle prenant en compte l'âge et l'effort physique combiné aux mouvements des coudes,
- Modèle 3 : modèle prenant en compte l'âge, le travail répétitif et l'effort physique combiné aux mouvements des coudes.

Le paragraphe suivant entre dans le détail des hypothèses effectuées dans les analyses sur les données imputées dont les aspects techniques sur analyses de sensibilité sont mis en annexe (Annexe D). Un lecteur qui n'est pas intéressé par les aspects méthodologiques des imputations multiples peut passer directement au paragraphe 4.2.3 (p 105) sans problème de compréhension pour la suite.

4.2.2 Gestion des données manquantes

Les différentes méthodologies pour gérer les données manquantes ont été présentées en détail dans le paragraphe méthodologique 2.3.4 (p 52). On a vu que les données manquantes de la population Cosali n'étaient pas MCAR puisque la présence d'un suivi dépend des caractéristiques initiales des individus (cf. paragraphe 3.1.2, p 75). Cela implique que l'analyse sur cas complets n'était pas la méthode la plus adaptée (cf. paragraphe 2.3.4, p 52). Dans les paragraphes suivants, on a fait des

hypothèses sur les différents mécanismes de données manquantes et on a justifié les choix méthodologiques correspondants.

4.2.2.1 Analyses sous l'hypothèse MAR séquentiel

4.2.2.1.1 Hypothèse MAR séquentiel

Dans ce paragraphe, on détaille les analyses avec imputations multiples MAR séquentielles.

Pour plus de simplicité, on désignera les épicondylites latérales initiales et les épicondylites latérales au suivi respectivement par TMS 0 et TMS 1, le temps de suivi par t et l'ensemble des expositions professionnelles considérées (gestes répétitifs, effort perçu important, torsion des poignets, flexion/extensions des coudes et faible soutien social) initialement et au suivi respectivement par Expositions 0 et Expositions 1. Pour la compréhension, on introduit deux variables non mesurées : l'état de santé au moment du suivi pour l'examen clinique, U_1, et le statut professionnel (actif/inactif) au moment du suivi pour l'examen clinique, U_2.

La Figure 32 montre un schéma simplifié des hypothèses faites sur les données manquantes ; en Annexe C se trouve le schéma complet des relations entre toutes les variables considérées.
On a supposé un mécanisme MAR pour l'absence d'auto-questionnaire de suivi conditionnellement aux variables suivantes (R_{AQ}) (Figure 32) :

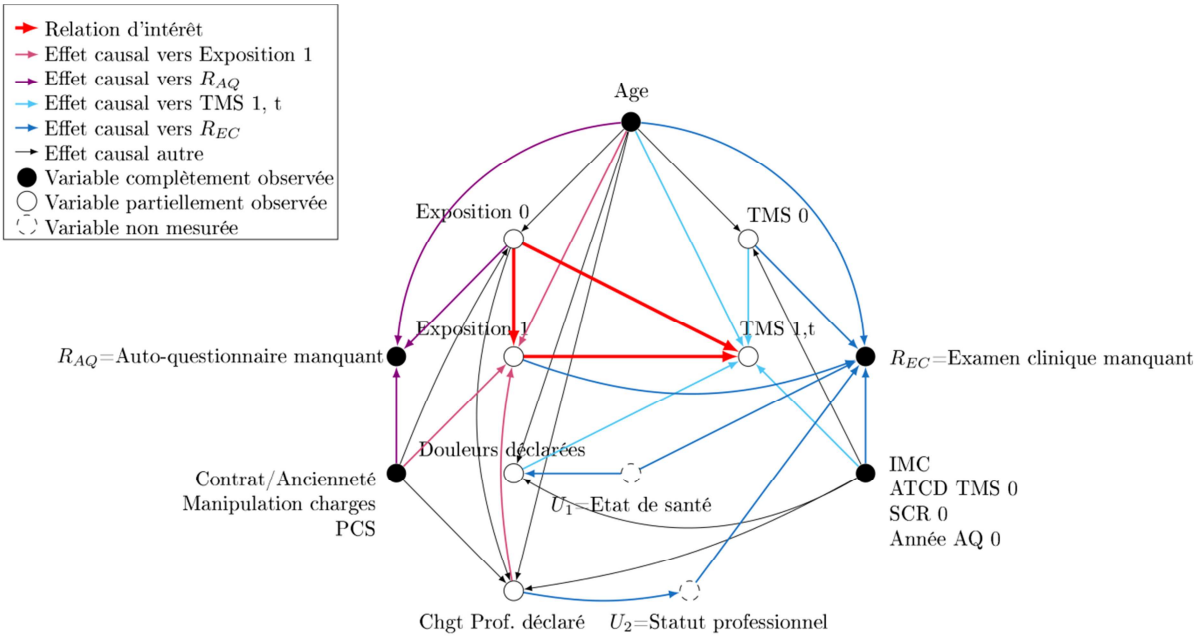
- premier auto-questionnaire : âge, PCS en 4 classes, type de contrat de travail combiné avec l'ancienneté (Contrat/ancienneté), manipulation de charges en 3 classes et expositions 0,
- premier examen clinique : syndrome de la coiffe des rotateurs (SCR 0) et TMS 0,
- deuxième examen clinique : venue à l'examen clinique en Oui/Non (Rec).

On a supposé qu'on avait un mécanisme MAR séquentiel pour l'absence d'examen clinique de suivi conditionnellement aux variables suivantes (R_{EC}) (Figure 32) :

- premier auto-questionnaire : âge, IMC, année du questionnaire (Année AQ 0), PCS en 4 classes, type de contrat de travail combiné avec l'ancienneté (Contrat/ancienneté) et expositions 0,
- premier examen clinique : antécédent de TMS (ATCD TMS 0), syndrome de la coiffe des rotateurs (SCR 0) et TMS 0,
- deuxième auto-questionnaire : changements professionnels depuis 2002 (Chgt Prof déclaré) passant par le statut professionnel à t_1 (U2), douleurs déclarées aux coudes (Douleurs déclarés) passant par l'état de santé non mesuré (U1).

On a un mécanisme de données manquantes MAR séquentiel comme on l'a montré dans la partie outils méthodologiques 2.3.3.3 (p 45) et correspond à un exemple détaillé en Annexe A (Figure 42, p 154).

Figure 32 - Hypothèses causales sur les mécanismes des données manquantes



La méthodologie choisie était la méthode de imputations multiples par « chained equation » (cf. paragraphe 2.3.4.2.3, p 61) et ce pour plusieurs raisons :

- la plupart des variables à imputer ne sont pas normales mais catégorielles,
- cette méthode d'imputation pour chaque variable successivement permet de gérer facilement les mécanismes de données manquantes MAR séquentiel, à condition d'imputer les variables manquantes dans un ordre adéquat : ici les variables partiellement observées du questionnaire à t_1 avant les variables partiellement observées de l'examen clinique à t_1 (cf. paragraphe 2.3.4.2.3.2.2, p 63 et Annexe B , p 156).

4.2.2.1.2 Les modèles d'imputations

Les imputations multiples par « chained equation » donnent des résultats non biaisés si le DAG de la Figure 32 est exact. Les ajustements qui en ont découlé sont précisés dans le Tableau 9. Dans la suite, pour plus de simplicité, on appelle Z l'ensemble de ces variables d'ajustement.

Tableau 9 - Les différentes variables d'ajustement selon les variables à imputer

		Variables complètement observées (à t_0)		**Variables partiellement observées (à t_1)**	
Variables à imputer	**Modèle**	**Variables d'ajustement du 1er auto-questionnaire**	**Variables d'ajustement du 1er examen clinique**	**Variables d'ajustement du 2ème auto-questionnaire**	**Variables d'ajustement du 2ème examen clinique**
Chgt professionnel déclaré	Logistique multinomial	Age, Contrat/ancienneté, PCS, Manipulation de charges, Expositions 0	ATCD TMS 0, TMS 0, SCR 0		Rec
Expositions 1	Logistique	Age, Contrat/ancienneté, PCS, Manipulation de charges, Expositions 0		Chgt prof déclaré, Autres Expositions 1	TMS 1, t, Rec
Douleurs déclarées	Logistique	Age, IMC, PCS, Manipulation de charges, Expositions 0	ATCD TMS 0, TMS 0, SCR 0		TMS 1, t, Rec
TMS 1	Logistique multinomial	Age, IMC, Expositions 0	ATCD TMS 0, TMS 0, SCR 0	Expositions 1, Douleurs 1	t
Temps entre les examens cliniques (t)	Normal avec *predictive matching*	Année questionnaire, Age, PCS, Expositions 0	ATCD TMS 0, TMS 0, SCR 0	Expositions 1, Douleurs 1	TMS 1

Le temps entre les examens cliniques était continu, compris entre 2 et 9 ans. Cette variable n'était pas Gaussienne, même après les transformations classiques. Ce problème a été résolu en utilisant la méthode « predictive matching » pour avoir des valeurs imputées similaires aux valeurs observées et éviter les valeurs non réalistes et extrêmes [120] (cf. 2.3.4.2.3, p 61).

La concordance des modèles entre eux devait être respectée : en particulier, si une première variable partiellement observée a été utilisée pour imputer une seconde variable partiellement observée, alors la seconde a été aussi utilisée pour imputer la première. De plus, comme on étudiait les épicondylites latérales au suivi sachant que les travailleurs n'avaient pas eu d'épicondylites latérales initialement, on a utilisé une variable combinant les épicondylites latérales initiales et au suivi dans les modèles.

Le modèle d'imputation de TMS 1 était plus général que le modèle d'analyse correspondant :

- Certaines variables auxiliaires ont été utilisées dans les imputations, mais pas dans les analyses. En effet, ces variables n'étaient pas facteurs de confusion de la relation entre les expositions professionnelles et les épicondylites latérales comme l'IMC, les antécédents de TMS (ATCD TMS 0), le syndrome de la coiffe des rotateurs (SCR 0), les douleurs auto-rapportées.
- On a pris en compte séparément les symptômes aux coudes et les épicondylites latérales car les variables citées dans le point précédent prédisaient différemment la probabilité de symptômes et d'épicondylites..

4.2.2.2 Les données imputées

On décrit les données disponibles et imputées des participants à l'étude selon leur suivi (cf. Figure 27, p 79). Le Tableau 10 et le Tableau 11 montrent que les données imputées avaient des distributions assez proches des données disponibles (données avec suivi par AQ ou par EC) ce qui était attendu. La seule exception concerne les temps de suivi pour le deuxième examen clinique : les temps de suivi imputés étaient plus longs que les temps de suivi observés, ce qui s'explique par le fait que ceux qui n'étaient pas suivis avaient des années d'inclusion plus anciennes que les suivis.

Tableau 10 - Répartition des variables imputées pour l'auto-questionnaire manquant avec les données complètes et les données avec un suivi par auto-questionnaire

	Hommes						Femmes					
	Données avec suivi par AQ et EC		Données avec suivi par AQ		Données imputées		Données avec suivi par AQ et EC		Données avec suivi par AQ		Données imputées	
	N	%	N	%	N	%	N	%	N	%	N	%
Evolution professionnelle												
Aucun changement	412	72,8	714	57,9	1208,3	55,9	342	71,3	554	57,1	851,5	55,0
Changement de poste	115	20,3	217	17,6	385,4	17,8	102	21,3	163	16,8	251,8	16,2
Changement d'entreprise	30	5,3	163	13,2	339,6	15,7	18	3,7	111	11,4	205,5	13,3
Inactivité	9	1,6	138	11,2	227,7	10,5	18	3,7	143	14,7	240,2	15,5
Mouvements des coudes et effort physique important*												
Non	408	72,1	920	74,7	1608,4	74,4	358	74,6	734	75,6	1187,0	76,6
Oui	158	27,9	312	25,3	552,6	25,6	122	25,4	237	24,4	362,0	23,4
Tâches répétitives												
Non	445	78,6	993	80,6	1728,9	80,0	328	68,3	702	72,3	1132,6	73,1
Oui	121	21,4	239	19,4	432,1	20,0	152	31,7	269	27,7	416,4	26,9
Faible soutien social												
Non	318	56,2	740	60,1	1297,1	60,0	268	55,8	593	61,1	953,5	61,6
Oui	248	43,8	492	39,9	863,9	40,0	212	44,2	378	38,9	595,5	38,4
Douleurs aux coudes déclarées												
Non	469	82,9	996	80,8	1746,2	80,8	365	76,0	755	77,7	1218,2	78,6
Oui	97	17,1	236	19,2	414,8	19,2	115	24,0	216	22,3	330,8	21,4
Total	**566**		**1232**		**2161**		**480**		**971**		**1549**	

* mouvements des coudes = flexion/extension des coudes ou torsion des poignets plus de 2 heures par jour

Tableau 11 - Comparaison de la répartition des variables imputées pour l'examen clinique manquant avec les données complètes et les données avec un suivi pour l'examen clinique

	Hommes						Femmes					
	Données avec suivi par AQ et EC		Données avec suivi par EC		Données imputées		Données avec suivi par AQ et EC		Données avec suivi par EC		Données imputées	
	N	%	N	%	N	%	N	%	N	%	N	%
Epicondylites au suivi												
Pas de symptômes aux coudes	480	84,8	784	85,1	1839,2	85,1	404	84,2	586	84,9	1316,2	85,0
Symptômes aux coudes	44	7,8	71	7,7	165,0	7,6	49	10,2	64	9,3	143,9	9,3
Epicondylites	42	7,4	66	7,2	156,8	7,3	27	5,6	40	5,8	88,9	5,7
Temps entre les examens (en années)												
2-4	66	11,7	126	13,7	224,1	10,4	52	10,8	84	12,2	148,6	9,6
4-6	295	52,1	480	52,1	1034,6	47,9	264	55,0	370	53,6	794,9	51,3
6-9	205	36,2	315	34,2	902,3	41,7	164	34,2	236	34,2	605,5	39,1
Total	**566**		**921**		**2161**		**480**		**690**		**1549**	

Les analyses statistiques ont été réalisées sur chacun des jeux de données séparément et recombinées en utilisant les règles de Rubin [91,98].

4.2.2.3 Analyses de sensibilité

Même si l'ensemble des variables pertinentes disponibles ont été prises en compte dans les analyses, certaines informations n'ont pas été mesurées pour les travailleurs perdus de vue, comme par exemple leur état de santé au suivi (qui peut être moins bon que celui des travailleurs suivis) et leur situation d'emploi au suivi (qui peut être moins favorable que les travailleurs suivis).

Comme cette étude se concentrait sur les épicondylites latérales, l'analyse de sensibilité a porté uniquement sur le mécanisme de données manquantes de cette variable (c'est-à-dire le fait d'avoir ou non l'examen clinique manquant). On disposait de trois types de raisons pour lesquelles l'examen clinique n'avait pas eu lieu (cf. paragraphe 3.1.2.2, Tableau 2, p 77) :

(1) Aucun rendez-vous n'avait été programmé pendant la période 2007-2010,

(2) Un changement de situation avait eu lieu pour le travailleur (décès, démission, entreprise fermée/radiée, inaptitude, invalidité, mutation, déménagement, refus de participer au suivi, congé parental, retraite),

(3) La raison était inconnue.

Dans les analyses de sensibilité, on suppose que la prévalence des épicondylites latérales à t_1 pour les non suivis varie en fonction de la cause d'absence d'examen clinique et l'âge. Ces analyses sont faites avec deux séries de scénarios qui sont décrites dans les parties suivantes. Les détails statistiques sur l'implémentation de ces analyses sont indiqués en Annexe D .

4.2.2.3.1 Premiers scénarios d'analyses

Dans cette partie, on détaille les hypothèses MNAR effectuées dans cette analyse et les conséquences sur les prévalences des données imputées.

4.2.2.3.1.1 Hypothèses MNAR

Le fait d'avoir un examen clinique à cause d'une absence de rendez-vous entre 2007 et 2010 laisse penser qu'on a à faire à des travailleurs qui n'étaient pas surveillés activement. Or, les travailleurs non suivis activement correspondent en principe à ceux qui ne sont pas à risque élevé de développer des TMS. On suppose donc qu'ils avaient de bonnes conditions de travail et étaient en bonne santé.

Ceux qui ont eu un changement de situation (déménagement, invalidité, arrêt maladie, licenciement...) avaient probablement des conditions de travail peu favorables et peut-être des problèmes de santé.

Ainsi, au vu de ces hypothèses, on aurait une prévalence plus importante des épicondylites latérales chez ceux qui ont eu un changement de situation que chez les travailleurs observés, et une prévalence plus faible chez ceux qui n'ont pas eu de rendez-vous entre 2007 et 2010.

Pour ceux dont on ne connaissait pas la cause d'absence d'examen, on a supposé qu'ils correspondaient un mélange des deux types de causes précédentes avec potentiellement plus de travailleurs avec changement de situation car aucune information n'était disponible à leur sujet. Ainsi, on a supposé une prévalence plus forte que chez les observés et moins forte que chez ceux avec un changement de situation.

4.2.2.3.1.2 Prévalences imputées sous les hypothèses MNAR

Différents scénario avec différents écarts entre les prévalences ont été choisis dont les détails techniques se trouvent en Annexe D . Le Tableau 12 illustre les prévalences par cause d'absence d'examen qui découlent de ces simulations. On voit par exemple que le scénario extrême a été choisi à titre d'illustration car il était hautement improbable que les prévalences des épicondylites atteignent des valeurs de 25%, même dans une sous-population à fort risque.

Tableau 12 - Prévalences par cause d'examen clinique manquant basées sur les données imputées MAR

	Pas de RDV	Inconnu	Changement de situation	Non manquant*	Total
Hommes					
Imputations MAR	7,3	6,2	7,8	7,2	7,0
Scenario bas	**0,9**	6,2	7,8	7,2	**5,3**
Scenario plausible	**2,1**	**11,1**	**17,9**	7,2	**7,6**
Scenario haut	7,3	**16,1**	**17,9**	7,2	**10,1**
Scenario extrême	**0,9**	**26,4**	**27,6**	7,2	**11,7**
Femmes					
Imputations MAR	5,7	5,0	6,2	5,8	5,6
Scenario bas	**0,8**	5,0	6,2	5,8	**4,4**
Scenario plausible	**1,7**	**9,8**	**16,3**	5,8	**6,5**
Scenario haut	5,7	**14,8**	**16,3**	5,8	**8,6**
Scenario extrême	**0,8**	**24,9**	**26,2**	5,8	**10,4**

* Cette prévalence ne change pas selon les scénarios puisqu'elle concerne ceux qui ont un examen clinique de suivi (individus non imputés)

4.2.2.3.2 Deuxièmes scénarios d'analyses

De même que la partie précédente, les hypothèses MNAR sont détaillées et les prévalences des données imputées sont décrites.

4.2.2.3.2.1 Hypothèses MNAR

La première série de scénarios ne permettait pas de prendre en compte le fait que les changements professionnels pouvaient être choisis ou subis. Dans le dernier cas, on suppose que l'état de santé de l'individu était plus mauvais que la normale.

Une façon grossière d'accéder au fait qu'un changement serait choisi ou subi était de prendre en compte l'âge : un changement professionnel jeune avait plus de chance d'être choisi qu'un changement professionnel à un âge plus avancé [121].

Parmi les salariés qui n'avaient pas de rendez-vous sur la période, un salarié plus âgé avait *a priori* un risque similaire d'avoir une épicondylite à un salarié plus jeune, autrement dit, l'écart entre les jeunes et les plus âgés pouvait être plus faible dans cette catégorie.

4.2.2.3.2.2 Prévalences imputées sous les hypothèses MNAR

En supposant une prévalence des épicondylites à t_1 et un effet de l'âge plus ou moins important selon la cause d'absence d'examen clinique, on obtient les prévalences par âge et par cause de données manquantes détaillées dans le Tableau 13 (dont les détails techniques sont précisés en Annexe D).

Tableau 13 - Prévalences par cause d'examen clinique manquant et par âge en fonction des scénarios MAR

	Pas de RDV				Inconnu				Changement de situation				Non manquant			
	< 30	30-44	≥ 45	Total	< 30	30-44	≥ 45	Total	< 30	30-44	≥ 45	Total	< 30	30-44	≥ 45	Total
Hommes																
MAR	*4,1*	*7,5*	*10,1*	*7,3*	*4,3*	*6,4*	*8,2*	*6,2*	*7,1*	*6,9*	*8,1*	*7,8*	*4,4*	*6,5*	*10,1*	*7,2*
bas	**0,1**	**0,7**	**0,8**	**0,6**	4,3	6,4	8,2	6,2	7,1	6,9	8,1	7,8	4,4	6,5	10,1	7,2
plausible	**0,4**	**1,4**	**1,5**	**1,2**	**5,4**	**9,4**	**13,9**	**9,3**	**9,1**	**14,0**	**20,2**	**18,3**	4,4	6,5	10,1	7,2
haut	4,1	**3,7**	**3,5**	**3,7**	**6,4**	**18,6**	**19,6**	**14,5**	**11,7**	**25,7**	**28,1**	**26,4**	4,4	6,5	10,1	7,2
Femmes																
MAR	*1,3*	*7,8*	*6,6*	*5,7*	*1,2*	*7,4*	*6,5*	*5,0*	*1,3*	*6,2*	*7,3*	*6,2*	*1,0*	*7,0*	*6,2*	*5,8*
bas	**0,04**	**0,7**	**0,6**	**0,5**	1,2	7,4	6,5	5,0	1,3	6,2	7,3	6,2	1,0	7,0	6,2	5,8
plausible	**0,1**	**1,5**	**1,1**	**1,0**	**2,1**	**10,2**	**12,2**	**8,0**	**3,5**	**13,7**	**19,2**	**15,7**	1,0	7,0	6,2	5,8
haut	1,3	**3,8**	**2,3**	**2,7**	**3,2**	**19,2**	**18,2**	**13,4**	**6,8**	**26,7**	**26,9**	**24,1**	1,0	7,0	6,2	5,8

De même que pour les scénarios du paragraphe 4.2.2.3.1 (p 102), le Tableau 13 montre que le scénario haut était assez improbable en pratique (avec des prévalences de l'ordre de 25%-28%).

4.2.3 Résultats

Cette partie liste l'ensemble des résultats trouvés dont le plan d'analyse est précisé dans le paragraphe 4.2.1 (p 95).

4.2.3.1 Taux d'incidence

Sur 3 231 travailleurs qui ne déclaraient pas de douleurs aux coudes, on estimait avec les données imputées à 171 (103 (5,5%) hommes et 68 (5,0%) femmes) le nombre d'épicondylites latérales incidentes. Le temps de suivi moyen entre les examens cliniques était de 5,6 ans pour les hommes et 5,5 ans pour les femmes (compris entre 2 ans et 9 ans). Les taux d'incidence annuels étaient de 1,0 [0,7;1,3] pour 100 hommes et de 0,9 [0,6;1,3] pour 100 femmes.

4.2.3.2 Distribution des expositions en fonction de la profession

Les ouvriers qualifiés et non qualifiés, les agents de service et les aides-soignants étaient fortement exposés à un effort important combiné avec des mouvements des coudes : respectivement 26% des ouvriers non qualifiés hommes, 21% des ouvriers qualifiés hommes et 17% des agents de service et des aides-soignants hommes et respectivement 26%, 29% et 23% des femmes ont déclaré être exposés à t_0 et t_1. Tandis que seulement 5% des agents de maitrise, 3% des hommes employés et 1% des cadres et des agents de la fonction publique étaient exposés à t_0 et t_1 (respectivement 5%, 6% et 2% des femmes). Les distributions de l'exposition aux tâches répétitives étaient similaires, avec des différences moins marquées entre les professions.

Parmi les hommes qui ne déclaraient pas la même exposition à t_0 et t_1 (pour le soutien social, les tâches répétitives ou l'exposition à l'effort combiné aux mouvements des coudes), 46,7% déclaraient une évolution professionnelle depuis 2002 (telle que définie en p 73) contre seulement 39,3% de ceux qui ne changeaient pas de déclaration entre les deux questionnaires, cette différence étant significative à 0,09% (pour les femmes, 47,4% contre 41,1%, significatif à 1,7%).

4.2.3.3 Analyses préliminaires

Pour rappel, l'analyse préliminaire portait sur les expositions initiales uniquement (à t_0). L'incidence des épicondylites augmentait avec tâches répétitives à t_0, non significativement pour les hommes (IRR 1,7 [0,9;3,3]) et significativement en univarié pour les femmes (IRR 2,3 [1,1;4,9]) (Tableau 14). L'exposition à l'effort combiné aux mouvements des coudes à t_0 était aussi associée avec l'incidence des épicondylites pour les femmes avec un IRR de 2,3 [1,1;5,0] en univarié et de 2,1 [1,1 ;4,6] en multivarié (Tableau 15).

Tableau 14 - Analyses préliminaires univariées d'incidence des épicondylites en fonction des facteurs de risque initiaux

	Analyses sur les cas complets, N=1394[a]											
	Hommes, N=797						Femmes, N=597					
	N	N evnt	% evnt	IRR	IC à 95%	P	N	N evnt	% evnt	IRR	IC à 95%	P
Facteurs personnels												
Age, en 3 classes												
< 30 ans	140	6	4,3	1,0	-		99	1	1,0	1,0	-	
30-44 ans	442	22	5,0	1,1	0,5- 2,9	0,76	298	17	5,7	5,6	0,7- 47,8	0,11
≥ 45 ans	215	15	7,0	1,7	0,6- 4,4	0,29	200	10	5,0	5,0	0,6- 44,7	0,15
Facteurs professionnels biomécaniques												
Tâches répétitives												
Non	636	30	4,7	1,0	-		434	15	3,5	1,0	-	
Oui	161	13	8,1	1,7	0,9- 3,3	0,11	163	13	8,0	**2,3**	**1,1- 4,9**	**0,03**
Effort perçu combiné aux mouvements des coudes[b]												
Non	562	27	4,8	1,0	-		451	16	3,5	1,0	-	
Oui	235	16	6,8	1,4	0,7- 2,6	0,33	146	12	8,2	**2,3**	**1,1- 5,0**	**0,03**
Facteurs professionnels psycho-sociaux												
Faible soutien social												
Non	504	23	4,6	1,0	-		407	18	4,4	1,0	-	
Oui	293	20	6,8	1,5	0,8- 2,8	0,20	190	10	5,3	1,2	0,5- 2,6	0,68

[a] Travailleurs avec un examen clinique de suivi
[b] Effort perçu important et flexion/extension des coudes plus de 2 heures/jour ou torsion des poignets plus de 2 heures/jour

Tableau 15 - Analyses préliminaires multivariées d'incidence des épicondylites en fonction des facteurs de risque initiaux

	Analyses sur les cas complets, N=1394[a]					
	Hommes, N=797			Femmes, N=597		
	IRR	IC à 95%	P	IRR	IC à 95%	P
Age, en 3 classes						
< 30 ans	1,0	-		1,0	-	
30-44 ans	1,2	0,5- 3,0	0,69	6,7	0,9- 49,1	0,06
≥ 45 ans	1,9	0,7- 5,1	0,20	5,4	0,7- 41,4	0,10
Tâches répétitives						
Non	1,0	-		1,0	-	
Oui	1,7	0,8- 3,5	0,15	1,9	0,9- 4,1	0,11
Effort perçu combiné aux mouvements des coudes[b]						
Non	1,0	-		1,0	-	
Oui	1,2	0,6- 2,5	0,52	2,1	1,0- 4,6	0,06

[a] Travailleurs avec un examen clinique de suivi
[b] Effort perçu important et flexion/extension des coudes plus de 2 heures/jour ou torsion des poignets plus de 2 heures/jour

4.2.3.4 Analyses principales (données complètes et données imputées MAR)

Les résultats des analyses sur données complètes étaient similaires à celles sur données imputées MAR, pour la plupart avec des intervalles de confiance plus larges. Dans ce paragraphe, seulement les résultats sur les données imputées sont discutés.

Les analyses univariées ont montré que l'incidence des épicondylites augmentait non significativement avec l'âge, atteignant un IRR de 1,8 [0,7;4,5] pour les hommes de 45-60 ans comparés aux hommes de 18-30 ans (6.4 [0,7;56,2] pour les femmes) (Tableau 16 et Tableau 17).

Les facteurs professionnels étudiés étaient fortement associés à l'incidence des épicondylites. L'effort important combiné avec des mouvements des coudes était significativement associé à l'incidence des épicondylites avec un IRR de 2,9 [1, 4;5,9] pour les travailleurs exposés à t_0 et t_1 par rapport aux travailleurs non exposés (ni au temps t_0 ni au temps t_1) parmi les hommes (IRR de 3,2 [1,4;7,5] parmi les femmes). Les hommes exposés aux tâches répétitives avaient une incidence des

épicondylites latérales significativement plus grande dans les analyses univariées. Le faible soutien social n'était pas associé à l'incidence des épicondylites dans les analyses univariées.

Tableau 16 - Analyses principales univariées d'incidence des épicondylites en fonction des facteurs de risque initiaux et au suivi pour les hommes

	Cas complets, N=491[a]					Imputations multiples, N=1881				
	N	N evnt	IRR	IC à 95%	p	N	N evnt	IRR	IC à 95%	p
Facteurs personnels										
Age, en 3 classes										
< 30 ans	81	5	1,0	-		452	19,5	1,0	-	
30-44 ans	267	14	0,9	0,3- 2,5	0,79	857	47,3	1,3	0,5- 3,4	0,54
≥ 45 ans	143	8	1,0	0,3- 3,0	0,95	572	39	1,7	0,6- 4,6	0,33
Facteurs professionnels biomécaniques										
Tâches répétitives										
Non exposé*	350	17	1,0	-		1303,3	59	1,0	-	
Exposé au temps t_0	45	2	0,9	0,2- 4,0	0,89	217,4	10,9	1,1	0,5- 2,6	0,85
Exposé au temps t_1	54	2	0,7	0,2- 3,3	0,69	186,7	14,4	1,6	0,6- 4,5	0,37
Exposé aux temps t_0 et t_1	42	6	**3,0**	**1,2- 7,7**	**0,02**	173,6	21,5	**2,7**	**1,3- 5,8**	**<,01**
Effort perçu combiné aux mouvements des coudes[b]										
Non exposé*	302	14	1,0	-		1093,6	40,7	1	-	
Exposé au temps t_0	*65*	*0*	*NE[c]*	*-*	*-*	*326,7*	*12,3*	*NE[c]*	*-*	*-*
Exposé au temps t_1	54	6	**2,5**	**1,0- 6,3**	**0,05**	195,4	21,5	**3,0**	**1,2- 7,2**	**0,02**
Exposé aux temps t_0 et t_1	70	7	2,1	0,9- 5,0	0,09	265,2	31,3	**3,1**	**1,5- 6,5**	**<,01**
Facteurs professionnels psycho-sociaux										
Faible soutien social										
Non exposé*	201	9	1,0	-		771,2	37,7	1	-	
Exposé au temps t_0	78	4	1,1	0,3- 3,9	0,82	348,4	21,3	1,2	0,5- 2,9	0,69
Exposé au temps t_1	110	5	1,0	0,3- 3,2	0,96	399,8	17,6	0,9	0,3- 2,3	0,81
Exposé aux temps t_0 et t_1	102	9	2,0	0,8- 5,2	0,15	361,6	29,2	1,7	0,8- 3,6	0,19

Tableau 17 - Analyses principales univariées d'incidence des épicondylites en fonction des facteurs de risque initiaux et au suivi pour les femmes

	Cas complets, N=412[a]					Imputations multiples, N=1350				
	N	N evnt	IRR	IC à 95%	p	N	N evnt	IRR	IC à 95%	p
Facteurs personnels										
Age, en 3 classes										
< 30 ans	62	1	1,0	-		325	3,7	1	-	
30-44 ans	210	12	3,4	0,4- 30,9	0,28	592	36,5	**7,1**	**1.0- 51,0**	**0,05**
≥ 45 ans	140	5	2,1	0,2- 21,4	0,53	433	23,9	6,4	0,9- 47,2	0,07
Facteurs professionnels biomécaniques										
Tâches répétitives										
Non exposé*	244	8	1,0	-		797,5	28,8	1	-	
Exposé au temps t_0	39	4	3,1	0,9- 11,1	0,08	197,5	12,3	1,7	0,7- 4,1	0,23
Exposé au temps t_1	61	3	1,5	0,4- 6,0	0,58	147,5	6,8	1,2	0,3- 4,4	0,73
Exposé aux temps t_0 et t_1	68	3	1,3	0,3- 5,4	0,69	207,5	16,2	2,1	0,8- 5,8	0,13
Effort perçu combiné aux mouvements des coudes[b]										
Non exposé*	262	10	1,0	-		859,1	31	1	-	
Exposé au temps t_0	45	2	1,2	0,3- 5,1	0,83	182,9	10	1,4	0,4- 4,7	0,59
Exposé au temps t_1	*50*	*0*	*NE[c]*	*-*	*-*	*137,9*	*3,5*	*NE[c]*	*-*	*-*
Exposé aux temps t_0 et t_1	55	6	**2,9**	**1,1- 7,8**	**0,03**	170,1	19,7	**3,2**	**1,43- 7,2**	**<,01**
Facteurs professionnels psycho-sociaux										
Faible soutien social										
Non exposé*	179	5	1,0	-		586	25,2	1	-	
Exposé au temps t_0	51	3	2,1	0,5- 9,4	0,35	251,1	11,2	1,0	0,3- 3,3	0,95
Exposé au temps t_1	98	5	1,8	0,5- 6,7	0,37	274	13	1,1	0,4- 3,2	0,87
Exposé aux temps t_0 et t_1	84	5	2,2	0,6- 8,1	0,24	238,9	14,6	1,4	0,5- 3,9	0,48

* Non exposé = exposé ni au temps t_0, ni au temps t_1
[a] Travailleurs avec un examen clinique de suivi et un auto-questionnaire de suivi
[b] Effort perçu important et flexion/extension des coudes plus de 2 heures/jour ou torsion des poignets plus de 2 heures/jour
[c] NE = Pas estimable car il n'y avait pas d'épicondylite dans cette catégorie pour les cas complets

Dans les analyses multivariées, l'association avec l'effort important combiné avec au moins un mouvement des coudes à t_0 et t_1 restait significative (Tableau 18 et Tableau 19). L'exposition à des tâches répétitives à t_0 et t_1 n'était pas significativement associée à l'incidence des épicondylites dans les analyses multivariées ajustées sur l'effort important combiné avec des mouvements des coudes et l'âge.

Avoir déclaré être exposé à t_0 et t_1 semblait (non significativement) plus fortement associé à l'incidence qu'avoir déclaré être exposé à t_1 pour l'effort combiné aux mouvements des coudes chez les hommes avec respectivement des IRRs de 3,2 [1,6;6,4] et 2,7 [1,2;6,1]. Quand il a été possible de l'estimer, l'association avec l'exposition était significative pour l'exposition à t_1 uniquement (par rapport à ceux qui n'étaient exposés ni à t_0 ni à t_1) et non significative pour l'exposition à t_0 uniquement.

Tableau 18 - Analyses principales multivariées d'incidence des épicondylites en fonction des facteurs de risque initiaux et au suivi pour les hommes

	Cas complets, N=491[a]								
	Modèle 1			**Modèle 2**			**Modèle 3**		
	IRR	**IC à 95%**	**p**	**IRR**	**IC à 95%**	**p**	**IRR**	**IC à 95%**	**p**
Age, en 3 classes									
< 30 ans	1,0	-		1,0	-		1,0	-	
30-44 ans	0,9	0,3- 2,6	0,86	0,8	0,3- 2,1	0,66	0,8	0,3- 2,1	0,69
≥ 45 ans	1,1	0,3- 3,4	0,91	0,9	0,3- 2,7	0,87	0,9	0,3- 2,6	0,86
Tâches répétitives									
Non exposé*	1,0	-					1,0	-	
Exposé au temps t_0	0,9	0,2- 4,2	0,91				1,1	0,3- 4,5	0,87
Exposé au temps t_1	0,7	0,2- 3,4	0,71				0,5	0,1- 2,1	0,36
Exposé aux temps t_0 et t_1	**3,0**	**1,1- 8,0**	**0,02**				**2,6**	**1,0- 6,7**	**0,05**
Effort perçu combiné aux mouvements des coudes[b]									
Non exposé*				1,0	-		1,0	-	
Exposé au temps t_0				*NE*[c]	-		*NE*[c]	-	
Exposé au temps t_1				**2,6**	**1,0- 6,5**	**0,04**	**2,5**	**1,0- 6,5**	**0,05**
Exposé aux temps t_0 et t_1				2,1	0,9- 5,1	0,10	1,7	0,6- 4,2	0,29

	Imputations multiples, N=1881								
	Modèle 1			**Modèle 2**			**Modèle 3**		
	IRR	**IC à 95%**	**p**	**IRR**	**IC à 95%**	**p**	**IRR**	**IC à 95%**	**p**
Age, en années									
< 30	1,0	-		1,0	-		1,0	-	
30-44	1,3	0,5- 3,4	0,53	1,4	0,6- 3,6	0,44	1,4	0,6- 3,5	0,46
≥ 45	1,8	0,6- 5,1	0,25	2,0	0,7- 5,4	0,19	2,0	0,7- 5,5	0,17
Tâches répétitives									
Non exposé*	1,0	-					1,0	-	
Exposé au temps t_0	1,7	0,6- 4,6	0,34				1,1	0,4- 2,6	0,85
Exposé au temps t_1	1,1	0,5- 2,6	0,84				1,2	0,4- 3,3	0,78
Exposé aux temps t_0 et t_1	**2,9**	**1,4- 6,1**	**<,01**				1,9	0,9- 4,4	0,11
Effort perçu combiné aux mouvements des coudes[b]									
Non exposé*				1,0	-		1,0	-	
Exposé au temps t_0				*NE*[c]	-		*NE*[c]	-	
Exposé au temps t_1				**3,1**	**1,3- 7,5**	**0,01**	**2,8**	**1,1- 6,9**	**0,03**
Exposé aux temps t_0 et t_1				**3,4**	**1,6- 7,0**	**<,01**	**2,8**	**1,3- 6,1**	**0,01**

* Non exposé = exposé ni au temps t_0, ni au temps t_1
[a] Travailleurs avec un examen clinique de suivi et un auto-questionnaire de suivi
[b] Effort perçu important et flexion/extension des coudes plus de 2 heures/jour ou torsion des poignets plus de 2 heures/jour
[c] NE = Pas estimable car il n'y avait pas d'épicondylite dans cette catégorie pour les cas complets

Tableau 19 - Analyses principales multivariées d'incidence des épicondylites en fonction des facteurs de risque initiaux et au suivi pour les femmes

	Cas complets, N=412[a]								
	Modèle 1			**Modèle 2**			**Modèle 3**		
	IRR	**IC à 95%**	**p**	**IRR**	**IC à 95%**	**p**	**IRR**	**IC à 95%**	**p**
Age, en 3 classes									
< 30 ans	1,0	-		1,0	-		1,0	-	
30-44 ans	3,4	0,4- 29,8	0,27	4,3	0,6- 31,3	0,15	4,3	0,6- 30,4	0,15
≥ 45 ans	2,0	0,2- 19,0	0,57	2,5	0,3- 20,2	0,38	2,1	0,3- 16,4	0,49
Tâches répétitives									
Non exposé*	1,0	-					1,0	-	
Exposé au temps t_0	3,3	0,9- 11,8	0,07				3,0	0,9- 9,9	0,08
Exposé au temps t_1	1,6	0,4- 6,4	0,53				1,2	0,3- 4,6	0,75
Exposé aux temps t_0 et t_1	1,4	0,3- 5,7	0,65				0,9	0,2- 3,4	0,85
Effort perçu combiné aux mouvements des coudes[b]									
Non exposé*				1,0	-		1,0	-	
Exposé au temps t_0				1,3	0,3- 5,6	0,74	1,1	0,2- 4,7	0,92
Exposé au temps t_1				*NE*[c]	-		*NE*[c]	-	
Exposé aux temps t_0 et t_1				**3,4**	**1,3- 9,2**	**0,01**	**3,5**	**1,2- 10,3**	**0,02**

	Imputations multiples, N=1350								
	Modèle 1			**Modèle 2**			**Modèle 3**		
	IRR	**IC à 95%**	**p**	**IRR**	**IC à 95%**	**p**	**IRR**	**IC à 95%**	**p**
Age, en 3 classes									
< 30 ans	1	-		1	-		1	-	
30-44 ans	**7,1**	**1,0- 49,3**	**0,05**	**7,3**	**1,1- 49,0**	**0,04**	**7,4**	**1,1- 49,7**	**0,04**
≥ 45 ans	6,4	0,9- 45,7	0,07	6,2	0,9- 43,6	0,06	6,3	0,9- 43,8	0,06
Tâches répétitives									
Non exposé*	1	-					1	-	
Exposé au temps t_0	1,2	0,3- 4,2	0,78				1,6	0,7- 3,7	0,3
Exposé au temps t_1	1,8	0,7- 4,3	0,19				1,0	0,3- 3,5	0,99
Exposé aux temps t_0 et t_1	2,1	0,8- 5,7	0,14				1,4	0,5- 4,1	0,57
Effort perçu combiné aux mouvements des coudes[b]	,		,						
Non exposé*				1	-		1	-	
Exposé au temps t_0				1,4	0,4- 4,7	0,58	1,3	0,4- 4,2	0,71
Exposé au temps t_1				*NE*[c]	-		*NE*[c]	-	
Exposé aux temps t_0 et t_1				**3,3**	**1,5- 7,3**	**<,01**	**2,9**	**1,2- 7,1**	**0,02**

* Non exposé = exposé ni au temps t_0, ni au temps t_1
[a] Travailleurs avec un examen clinique de suivi et un auto-questionnaire de suivi
[b] Effort perçu important et flexion/extension des coudes plus de 2 heures/jour ou torsion des poignets plus de 2 heures/jour
[c] NE = Pas estimable car il n'y avait pas d'épicondylites dans cette catégorie pour les cas complets

4.2.3.5 Analyses additionnelles (données imputées MNAR)

Les analyses de sensibilité réalisées ont montré une bonne robustesse des résultats pour les modèles univariés et multivariés. Les détails sur les résultats du modèle multivarié 3 (cf. paragraphe 4.2.1, p 95) sont présentés dans les paragraphes suivants.

4.2.3.5.1 Premiers scénarios

Pour les scénarios bas, plausibles et hauts, les résultats initiaux sont restés inchangés pour l'exposition combinée de l'effort important avec des mouvements des coudes (Tableau 20 et Tableau 21). Pour le scénario extrême, l'association avec ce facteur n'était plus significative.

On a constaté également que l'effet de l'âge chez les femmes disparaissait dans les analyses de sensibilité des trois derniers scénarios. Au vu de ces changements, l'association très forte trouvée dans les analyses principales était peut-être liée à un suivi insuffisant des femmes de moins de 30 ans, en particulier de celles qui présentaient une épicondylite au suivi.

Tableau 20 - Analyse de sensibilité de l'incidence des épicondylites sur les premiers scénarios MNARs chez les hommes

	Scénario bas		Scénario plausible		Scénario haut		Scénario extrême	
	IRR	**95% CI**	**IRR**	**95% CI**	**IRR**	**95% CI**	**IRR**	**95% CI**
Age, en 3 classes								
< 30 ans	1,0	-	1,0	-	1,0	-	1,0	-
30-44 ans	1,3	0,6- 2,9	1,0	0,5- 1,9	0,9	0,5- 1,6	0,7	0,4- 1,1
≥ 45 ans	2,2	1,0- 5,0	1,9	1,0- 3,6	1,5	0,9- 2,7	1,3	0,8- 2,0
Tâches répétitives								
Non exposé*	1,0	-	1,0	-	1,0	-	1,0	-
Exposé au temps t_0	1,1	0,4- 2,9	1,0	0,4- 2,4	1,0	0,5- 2,0	1,0	0,6- 1,8
Exposé au temps t_1	1,0	0,4- 2,6	0,9	0,4- 2,2	1,0	0,5- 2,0	0,9	0,5- 1,7
Exposé aux temps t_0 et t_1	1,9	0,9- 4,2	1,7	0,9- 3,3	1,6	0,9- 2,9	1,3	0,8- 2,3
Effort perçu combiné aux mouvements des coudes[b]								
Non exposé*	1,0	-	1,0	-	1,0	-	1,0	-
Exposé au temps t_0	*NE*[c]	-	*NE*[c]	-	*NE*[c]	-	*NE*[c]	-
Exposé au temps t_1	**2,5**	**1,1- 5,6**	1,9	1,0- 3,8	1,8	1,0- 3,2	1,3	0,8- 2,4
Exposé aux temps t_0 et t_1	**2,6**	**1,2- 5,7**	**2,0**	**1,0- 3,8**	**1,8**	**1,0- 3,2**	1,3	0,8- 2,2

[b] Effort perçu important et flexion/extension des coudes plus de 2 heures/jour ou torsion des poignets plus de 2 heures/jour
[c] NE = Pas estimable car il n'y avait pas d'épicondylites dans cette catégorie pour les cas complets

Tableau 21 - Analyse de sensibilité de l'incidence des épicondylites sur les premiers scénarios MNARs chez les femmes

	Scénario bas		Scénario plausible		Scénario haut		Scénario extrême	
	IRR	**95% CI**	**IRR**	**95% CI**	**IRR**	**95% CI**	**IRR**	**95% CI**
Age, en 3 classes								
< 30 ans	1,0	-	1,0	-	1,0	-	1,0	-
30-44 ans	**7,3**	**1,2- 43,9**	2,2	0,9- 5,7	1,8	0,8- 3,8	1,1	0,6- 2,0
≥ 45 ans	**7,1**	**1,2- 42,8**	2,5	0,9- 6,7	1,8	0,8- 4,0	1,3	0,7- 2,5
Tâches répétitives								
Non exposé*	1,0	-	1,0	-	1,0	-	1,0	-
Exposé au temps t_0	2,0	0,8- 5,1	1,7	0,8- 3,6	1,5	0,7- 3,1	1,4	0,8- 2,5
Exposé au temps t_1	1,2	0,4- 3,7	1,0	0,3- 2,7	0,8	0,3- 2,2	0,8	0,4- 1,8
Exposé aux temps t_0 et t_1	1,5	0,5- 4,1	1,2	0,5- 3,0	1,2	0,6- 2,4	1,0	0,5- 1,9
Effort perçu combiné aux mouvements des coudes[b]								
Non exposé*	1,0	-	1,0	-	1,0	-	1,0	-
Exposé au temps t_0	1,3	0,5- 3,8	1,4	0,6- 3,1	1,3	0,6- 2,7	1,3	0,7- 2,4
Exposé au temps t_1	*NE*[c]	-	*NE*[c]	-	*NE*[c]	-	*NE*[c]	-
Exposé aux temps t_0 et t_1	**2,8**	**1,2- 6,4**	**2,2**	**1,1- 4,6**	**2,0**	**1,0- 3,9**	1,5	0,8- 2,8

[b] Effort perçu important et flexion/extension des coudes plus de 2 heures/jour ou torsion des poignets plus de 2 heures/jour
[c] NE = Pas estimable car il n'y avait pas d'épicondylites dans cette catégorie pour les cas complets

4.2.3.5.2 Deuxièmes scénarios

L'effet de l'âge a été mécaniquement changé dans les analyses de sensibilité (de par la nature des hypothèses faites, Tableau 22 et Tableau 23). Les analyses de sensibilité ont montré une bonne robustesse des résultats sur les expositions professionnelles : même si la valeur de l'effet diminuait avec l'importance des changements effectués, elle restait significative.

Tableau 22 - Analyse de sensibilité de l'incidence des épicondylites sur les deuxièmes scénarios MNARs chez les hommes

	Scénario bas		Scénario plausible		Scénario haut	
	IRR	**95% CI**	**IRR**	**95% CI**	**IRR**	**95% CI**
Age, en 3 classes						
< 30 ans	1,0	-	1,0	-	1,0	-
30-44 ans	1,3	0,6- 2,8	1,3	0,6- 2,7	1,4	0,8- 2,6
≥ 45 ans	2,2	1,0- 4,9	**3,0**	**1,5- 6,0**	**2,7**	**1,5- 5,0**
Tâches répétitives						
Non exposé*	1,0	-	1,0	-	1,0	-
Exposé au temps t_0	1,0	0,4- 2,9	1,1	0,5- 2,5	1,0	0,5- 2,1
Exposé au temps t_1	1,0	0,4- 2,6	0,9	0,4- 2,2	0,9	0,5- 2,0
Exposé aux temps t_0 et t_1	1,9	0,9- 4,2	1,8	0,9- 3,5	1,6	0,9- 3,0
Effort perçu combiné aux mouvements des coudes[b]						
Non exposé*	1,0	-	1,0	-	1,0	-
Exposé au temps t_0	*NE*[c]	-	*NE*[c]	-	*NE*[c]	-
Exposé au temps t_1	2,5	1,1- 5,6	1,9	0,9- 3,9	1,7	0,9- 3,2
Exposé aux temps t_0 et t_1	**2,6**	**1,2- 5,6**	**2,0**	**1,0- 4,0**	**1,8**	**1,0- 3,2**

[b] Effort perçu important et flexion/extension des coudes plus de 2 heures/jour ou torsion des poignets plus de 2 heures/jour
[c] NE = Pas estimable car il n'y avait pas d'épicondylites dans cette catégorie pour les cas complets

Tableau 23 - Analyse de sensibilité de l'incidence des épicondylites sur les deuxièmes scénarios MNARs chez les femmes

	Scénario bas		Scénario plausible		Scénario haut	
	IRR	**95% CI**	**IRR**	**95% CI**	**IRR**	**95% CI**
Age, en 3 classes						
< 30 ans	1,0	-	1,0	-	1,0	-
30-44 ans	**7,2**	**1,2- 43,6**	**5,4**	**1,1- 25,7**	**4,5**	**1,3- 15,3**
≥ 45 ans	**7,0**	**1,2- 42,4**	**7,2**	**1,5- 34,9**	**5,3**	**1,6- 18,0**
Tâches répétitives						
Non exposé*	1,0	-	1,0	-	1,0	-
Exposé au temps t_0	2,0	0,8- 5,1	1,7	0,8- 3,9	1,5	0,8- 2,9
Exposé au temps t_1	1,2	0,4- 3,8	1,0	0,3- 2,8	0,8	0,3- 2,1
Exposé aux temps t_0 et t_1	1,4	0,5- 4,0	1,3	0,5- 3,1	1,1	0,5- 2,4
Effort perçu combiné aux mouvements des coudes[b]						
Non exposé*	1,0	-	1,0	-	1,0	-
Exposé au temps t_0	1,3	0,5- 3,8	1,4	0,6- 3,1	1,4	0,7- 2,8
Exposé au temps t_1	*NE*[c]	-	*NE*[c]	-	*NE*[c]	-
Exposé aux temps t_0 et t_1	**2,8**	**1,2- 6,3**	**2,3**	**1,1- 4,8**	**2,1**	**1,0- 4,1**

[b] Effort perçu important et flexion/extension des coudes plus de 2 heures/jour ou torsion des poignets plus de 2 heures/jour
[c] NE = Pas estimable car il n'y avait pas d'épicondylites dans cette catégorie pour les cas complets

4.2.4 Discussion

Cette étude a estimé l'incidence des épicondylites latérales dans une large population de travailleurs et a souligné le rôle des facteurs physiques professionnels pour l'incidence des épicondylites latérales, particulièrement l'effort physique important combiné avec des mouvements des coudes. Ces résultats ont déjà été trouvés en transversal. Les différentes analyses ont montré qu'une exposition élevée répétée était fortement associée avec l'incidence des épicondylites.

Le taux annuel d'incidence des épicondylites latérales a été estimé à 1,0 [0,7 ;1,3] pour 100 travailleurs parmi les hommes et à 0,9 [0,7 ;1,3] pour 100 travailleurs parmi les femmes. Feleus et al. ont trouvé un taux d'incidence annuel de 1,1 pour 100 patients en médecine générale aux Pays-Bas et Silverstein et al. ont trouvé un taux d'incidence annuel de demande de compensation de 0,1 pour 100 travailleurs à plein temps dans l'état de Washington [41,43]. En France, Roquelaure et al. ont trouvé un taux d'incidence annuel de 4,0 pour 100 travailleurs fortement exposés aux contraintes physiques [40,42]. Le fait que le métier soit un déterminant important d'apparition d'épicondylites peut expliquer les différences entre les incidences selon la population d'étude [47]. En effet, la population de cette étude était une population de travailleurs avec des tâches très variées incluant des travailleurs peu exposés.

Le résultat principal de cette étude était l'effet important de l'effort physique combiné avec des mouvements des coudes sur l'incidence des épicondylites. Ce résultat est concordant avec d'autres études sur l'excès d'incidence des épicondylites parmi les travailleurs manuels [38,41,119]. Dans l'article de Leclerc et al., les mouvements répétés des poignets et des coudes "tourner, visser" ont aussi été identifiés comme associés à l'incidence des épicondylites [40]. Ces mouvements présentent de plus une certaine plausibilité biologique, comme montré dans de précédentes études [122]. Les mouvements répétitifs étaient aussi associés dans les analyses univariées avec les épicondylites, ce résultat étant en accord avec la forte incidence des épicondylites pour les travailleurs exposés aux gestes répétitifs [40] et les associations déjà trouvées sur la prévalence [25,34,36,116]. Cependant, cette étude a suggéré que la répétition globale était plutôt un indicateur de mouvements répétitifs spécifiques qu'un facteur de risque indépendant. Les interactions trouvées dans d'autres études entre la répétition et l'effort perçu n'ont pas été explorées ici par manque d'effectifs dans les catégories croisées. Ces interactions n'ont pu être exclues ici, même si elles n'avaient pas été trouvées en transversal sur les mêmes données [46].

L'analyse préliminaire sur le lien entre l'exposition initiale et l'incidence des épicondylites latérales est l'approche traditionnelle avec ce genre de données : elle limite la mesure de l'exposition à la mesure initiale (à t_0). Cette analyse a mis en évidence des résultats comparables avec l'analyse

principale basée sur les mesures des expositions initiales et au suivi (à t_0 et t_1), mais elle n'exploitait qu'une partie de l'information disponible. En effet, l'analyse préliminaire revenait à comparer le sous-groupe des travailleurs exposés aux temps t_0 et $t_1$et ceux exposés au temps t_0 avec le sous-groupe des travailleurs exposés au temps t_1 et ceux qui ne sont exposés ni au temps t_0, ni au temps t_1. L'analyse principale avec les deux mesures a exploré indirectement l'effet de la durée d'exposition sur l'incidence des épicondylites, puisque le nombre de déclarations d'expositions pouvait être considéré comme une approximation grossière de la durée d'exposition (ceux qui étaient exposés seulement à t_1 ont été exposés pour une durée plus courte que ceux qui étaient exposés à t_0 et t_1). Les résultats trouvés dans cette étude sont compatibles avec l'hypothèse que l'incidence des épicondylites augmente avec la durée d'exposition. En particulier, être exposé à t_0 et t_1 semblait être plus fortement associé aux épicondylites (même si non significativement) qu'être exposé à t_1 pour l'exposition à l'effort physique combiné à des mouvements des coudes parmi les hommes. Les résultats sur l'exposition à t_0 seulement ou à t_1 seulement suggéraient également plutôt un effet à court terme qu'un effet avec temps de latence.

En ce qui concerne la mesure répétée des expositions professionnelles, on peut se demander si les évolutions entre les déclarations traduisent un réel changement de tâches à exécuter au travail, de matériel de travail, de poste, ou traduisent simplement les limites de l'exposition auto-rapportée. Sur le dernier point, il n'est pas impossible que même si les expositions professionnelles d'un individu n'ont pas changé, la réponse à une même question portant sur l'exposition soit différente à quelques années d'écart en fonction de son ressenti au moment de la passation du questionnaire. Quoiqu'il en soit, ceux qui déclaraient deux fois une exposition représentaient vraisemblablement une population homogène très exposée tandis que ceux qui déclaraient une fois une exposition pourrait représenter une population hétérogène incluant des individus peu exposés. On constatait qu'une grande partie de ceux qui déclaraient deux expositions différentes aux deux questionnaires déclaraient aussi une évolution professionnelle, mais aucun élément n'était disponible sur les évolutions qui auraient pu avoir eu lieu au sein du même poste. D'autres études seraient nécessaires pour explorer la signification de ces changements de déclarations.

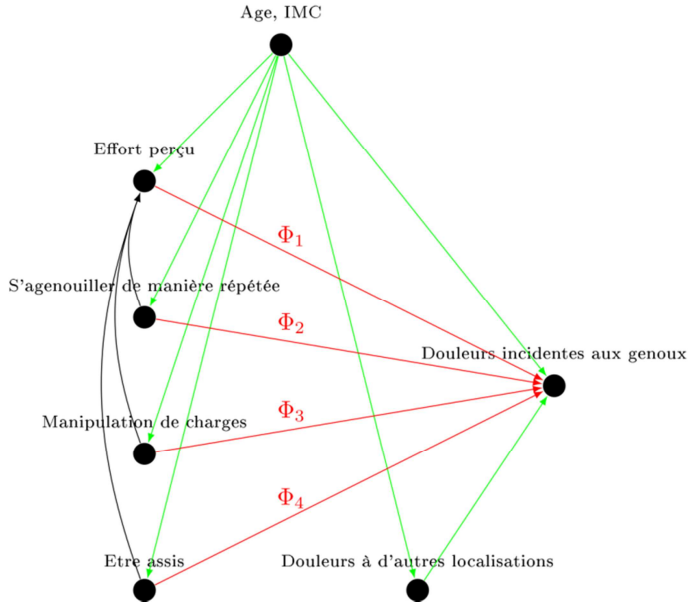
Le paragraphe suivant porte sur la dernière étude de ce travail de thèse et sur un autre problème articulaire : les douleurs de genoux.

4.3 Incidence des douleurs aux genoux

4.3.1 Analyses statistiques

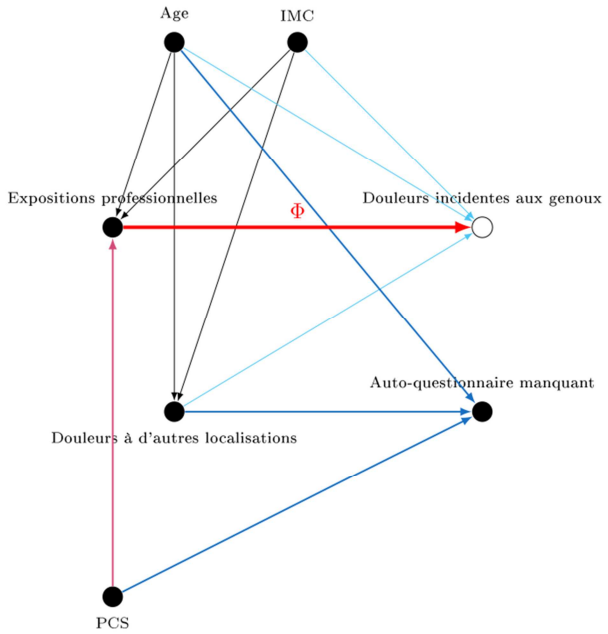
Le modèle causal supposé dans cette analyse est représenté dans la Figure 33.

Figure 33 - DAG des hypothèses de travail pour l'incidence des douleurs aux genoux



Les analyses ont été effectuées sur les travailleurs ayant un suivi par auto-questionnaire. Les travailleurs n'ayant pas été suivis ont été comparés à ceux avec suivi sur leurs caractéristiques initiales. Cette comparaison a donné lieu à la création d'un score de propension au suivi qui a été utilisé pour effectuer des analyses pondérées (avec les hypothèses représentées dans la Figure 34).

Figure 34 - DAG des hypothèses de travail pour l'incidence des douleurs aux genoux avec les données manquantes



Le taux d'incidence des douleurs aux genoux a été estimé par le nombre de nouveaux cas divisé par le temps entre les deux questionnaires cumulé sur tous les individus (pour plus de détail, cf. paragraphe 2.1.1, p 26).

L'analyse principale a porté sur l'association entre les facteurs de risque mesurés initialement et l'incidence des douleurs aux genoux. Les associations avec l'incidence de douleurs aux genoux de longue et de courte durées ont été étudiées à l'aide de modèles logistiques multinomiaux univariés et multivariés. Certaines égalités entre les coefficients du modèle ont été testées par des tests de Wald.

Deux modèles multivariés ont été présentés dans ce travail :

- un modèle ajusté sur l'âge, l'IMC, le fait de s'agenouiller, la manipulation de charges et le fait de se tenir debout,
- un modèle ajusté sur les facteurs précédents et les douleurs à d'autres localisations.

Les fractions attribuables parmi les exposés ont été calculées à partir du premier ajustement.

Les interactions multiplicatives entre le fait de s'agenouiller, la manipulation de charges et l'IMC ont été testées par des tests de Wald.

Deux analyses supplémentaires ont été effectuées sur des sous-populations :

- sur les travailleurs sans changement professionnel pendant le suivi,
- sur les travailleurs sans douleurs (toutes localisations confondues) initialement.

Les résultats ont été présentés pour les analyses non pondérées. Les analyses pondérées ont été présentées uniquement pour les modèles multivariés.

4.3.2 Résultats

La population d'étude était composée des 2 576 participants qui n'avaient pas de douleur aux genoux et qui n'étaient pas en sous-poids en 2002-2005 (cf. paragraphe 3.2, p 80 et Figure 35).

4.3.2.1 Données manquantes et score de propension à répondre

Sur les 2 576 participants, 62,7% ont rempli un auto-questionnaire de suivi (c'est-à-dire 1616 participants). Ceux qui ont répondu étaient plus âgés (plus de 45 ans), plus souvent des cadres et plus exposés à la station assise et ayant plus de douleurs à d'autres localisations initialement. Le suivi n'était pas associé à l'IMC ou à la plupart des expositions professionnelles qui ont été étudiées dans ce travail (s'agenouiller, la manipulation de charges et l'effort physique perçu).

Figure 35 - Population d'étude pour l'analyse des douleurs aux genoux

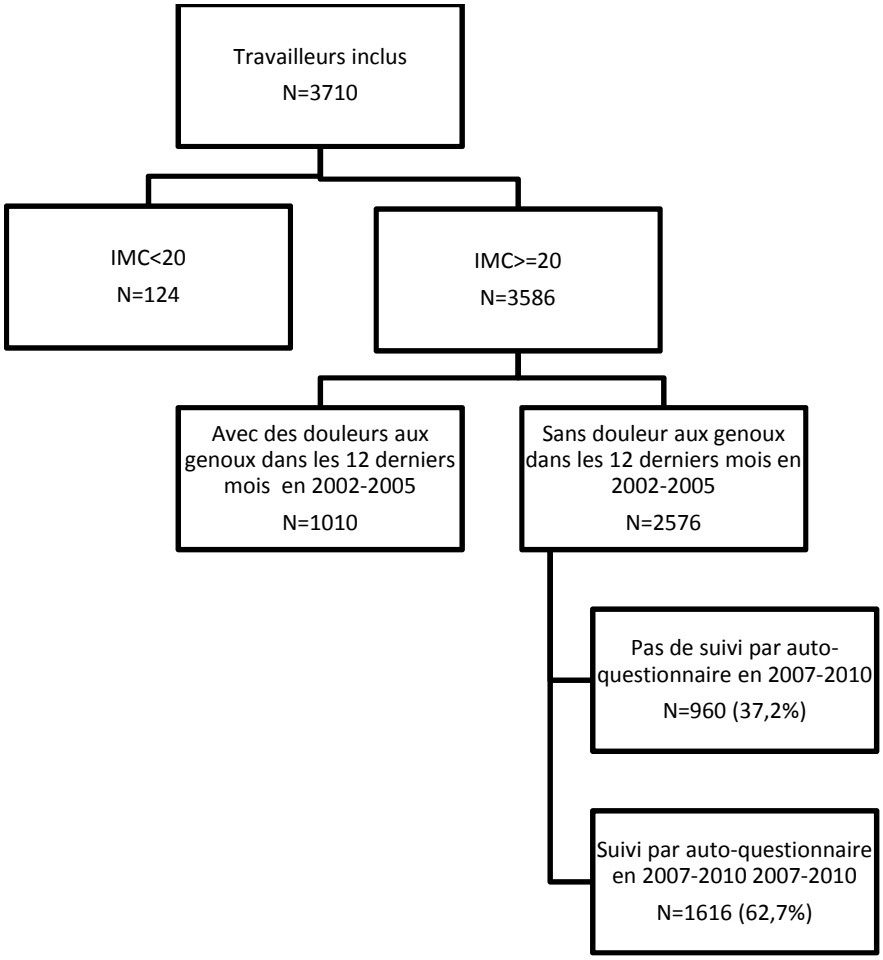


Tableau 24 - Comparaison entre les suivis et les non suivis par auto-questionnaires sur les principales caractéristiques initiales

	Hommes					Femmes				
Variable	**Non suivi**		**Suivi**			**Non suivi**		**Suivi**		
	N	**(%)**	**N**	**(%)**	**p**	**N**	**(%)**	**N**	**(%)**	**p**
Age en années*	35	(28-44)	40	(32-48)		36	(28-44)	40	(32-48)	
Age en 2 classes					**<,0001**					**0,0001**
<45 ans	447	**75,9%**	587	63,9%		279	75,2%	445	63,8%	
≥45 ans	142	24,1%	331	**36,1%**		92	24,8%	253	**36,3%**	
Indice de Masse Corporelle					0,1818					0,3872
Corpulence normale	349	59,3%	517	56,3%		266	71,7%	500	71,6%	
Surpoids	190	32,3%	336	36,6%		81	21,8%	138	19,8%	
Obésité	50	8,5%	65	7,1%		24	6,5%	60	8,6%	
Catégorie Socio-Professionnelle					**0,0008**					0,2890
Manquant	2	0,3%	0	0,0%		0	0,0%	3	0,4%	
Cadres	180	30,6%	364	**39,7%**		82	22,1%	179	25,6%	
Employés	61	**10,4%**	78	8,5%		207	55,8%	349	50,0%	
Ouvriers qualifiés	223	37,3%	334	36,4%		24	6,5%	54	7,7%	
Ouvriers non qualifiés	123	**20,9%**	142	15,5%		58	15,6%	113	16,2%	
Effort physique important					0,7890					0,7694
Non	282	47,9%	446	48,6%		215	58,0%	398	57,0%	
Oui	307	52,1%	472	51,4%		156	42,1%	300	43,0%	
Port de charges					0,2988					0,3153
Non ou < 2 heures/jours	421	71,5%	692	75,4%		309	83,3%	597	85,5%	
0 à 10 kg	72	12,2%	103	11,2%		41	11,0%	61	8,7%	
10 à 25 kg	57	9,7%	67	7,3%		12	3,2%	30	4,3%	
Plus de 25 kg	39	6,6%	56	6,1%		9	2,4%	10	1,4%	
Manipulation de charges de plus de 4 kg					0,1658					0,7992
Non	454	77,1%	735	80,1%		333	89,8%	623	89,3%	
Oui	135	22,9%	183	19,9%		38	10,2%	75	10,7%	
S'agenouiller					0,1628					0,1112
Non	429	72,8%	698	76,0%		305	82,2%	545	78,1%	
Oui	160	27,2%	220	24,0%		66	17,8%	153	21,9%	
Se tenir debout**					**0,0241**					0,2461
Non	332	56,4%	571	**62,2%**		240	64,7%	476	68,2%	
Oui	257	**43,6%**	347	37,8%		131	35,3%	222	31,8%	
Douleurs à d'autres localisations					**0,0087**					**0,0394**
Pas de douleur	153	**26,0%**	179	19,5%		81	21,8%	141	20,2%	
Douleurs de courte durée	294	49,9%	480	52,3%		193	**52,0%**	322	46,1%	
Douleurs de longue durée	142	24,1%	259	**28,2%**		97	26,2%	235	**33,7%**	
Total	**589**		**918**			**371**		**698**		

*Age médian et interquartiles
**S'asseoir moins de 4h/j

Le score de propension a été calculé à partir d'un modèle logistique ajusté sur l'âge, la catégorie socio-professionnelle et les douleurs à d'autres localisations (séparément pour les hommes et les femmes). Même si les variables choisies étaient significativement associées au suivi (Tableau 25), les Figure 36 et Figure 37 montraient des capacités de prédiction très moyennes (aires sous la courbe de 0.6 pour les hommes et pour les femmes).

Tableau 25 - Modèle de suivi par auto-questionnaire pour les hommes et les femmes

	N	% réponse	OR	IC à 95%	P
Hommes					
Age en 2 classes					**<,0001**
<45 ans	1034	56,8	1,0		
≥45 ans	473	70,0	**1,7**	**1,3-2,1**	
Catégorie Socio-Professionnelle					**0,0097**
Cadres / NC	546	66,7	1,0		
Employés	139	56,1	**0,6**	**0,5-1,0**	
Ouvriers qualifiés	557	60,0	**0,8**	**0,6-1,0**	
Ouvriers non qualifiés	265	53,6	**0,6**	**0,4-0,8**	
Douleurs à d'autres localisations					**0,0112**
Pas de douleur	332	53,9	1,0		
Douleurs de courte durée	774	62,0	**1,4**	**1,1-1,9**	
Douleurs de longue durée	401	64,6	**1,5**	**1,1-2,0**	
Femmes					
Age en 2 classes					**0,0008**
<45 ans	724	61,5	1,0		
≥45 ans	345	73,3	**1,6**	**1,2-2,2**	
Catégorie Socio-Professionnelle					0,3814
Cadres / NC	264	68,9	1,0		
Employés	556	62,8	0,8	06-1,1	
Ouvriers qualifiés	78	69,2	1,0	0,6-1,8	
Ouvriers non qualifiés	171	66,1	0,9	0,6-1,3	
Douleurs à d'autres localisations					0,1450
Pas de douleur	222	63,5	1,0		
Douleurs de courte durée	515	62,5	0,9	0,7-1,3	
Douleurs de longue durée	332	70,8	1,3	0,9-1,8	

Figure 36 - Probabilité prédite de répondre aux questionnaires en fonction du suivi par sexe

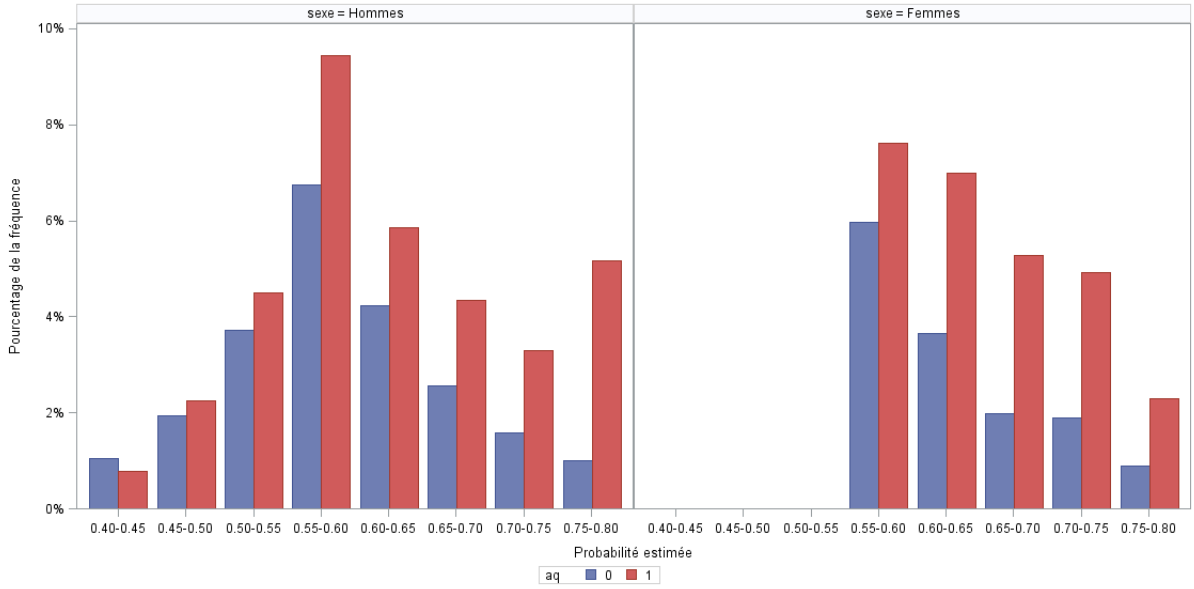
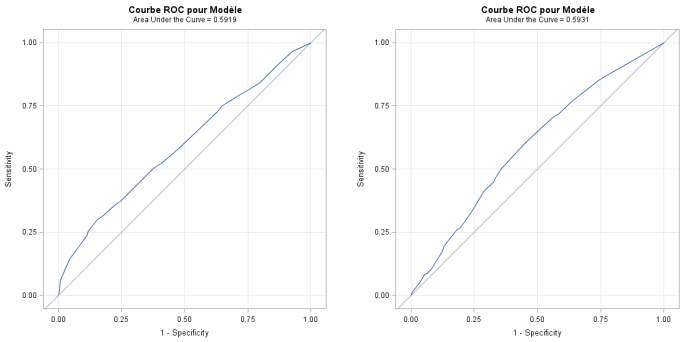


Figure 37 - Courbes de ROC du modèle sur le suivi pour les hommes (à gauche) et pour les femmes (à droite)



On obtenait alors des poids entre 1,2 et 2,4 pour les individus suivis avec une médiane de 1,6.

4.3.2.2 Incidence des douleurs aux genoux

Sur les 1 616 participants sans douleur initialement, 122 (7,5%) ont déclaré des douleurs de longue durée au suivi et 243 (15,0%) des douleurs de courte durée au suivi. La durée de suivi étant de 3,8 ans (3,9 pour les hommes, 3,7 pour les femmes), le taux d'incidence des douleurs aux genoux (toutes durées confondues) a été estimé à 5,9 (IC à 95% : 5,3-6,5) pour 100 travailleurs-années ;5,7 (5,0-6,6) pour les hommes et 6,1 (5,2-7,1) pour les femmes. Le taux d'incidence des douleurs aux genoux de longue durée a été estimé à 2,0 (1,6-2,3) pour 100 travailleurs-années ; 2,0 (1,5-2,5) pour les hommes et 2,0 (1,5-2,6) pour les femmes.

4.3.2.3 Les expositions en fonction de la catégorie socio-professionnelle

Parmi les hommes, 30,3% des ouvriers non qualifiés et 32,6% des ouvriers qualifiés étaient exposés à la manipulation de charges de plus de 4 kg, comparé à 5,5% d'exposés chez les cadres. De manière similaire pour les femmes, 27,4% des ouvrières non qualifiées et 20,4% des ouvrières qualifiées étaient exposées à la manipulation de charges de plus de 4 kg, comparées à 3,3% d'exposées chez les cadres. Pour l'exposition au fait de s'agenouiller de manière répétée, des pourcentages similaires ont été trouvés, avec plus d'exposés parmi les cadres : 25,3% des ouvriers non qualifiés et 37,1% des ouvriers qualifiés chez les hommes comparés à 14,8% des cadres chez les hommes, 25,7% des ouvrières non qualifiées et 29,6% des ouvrières qualifiées chez les femmes comparé à 15,9% des cadres chez les femmes.

La répartition des IMCs était similaire entre les catégories socio-professionnelles chez les hommes (environ 36,6% des travailleurs en surpoids et 7,1% des travailleurs obèses). La prévalence de l'obésité chez les femmes était plus forte pour les ouvrières non qualifiées (14,2%) et pour les ouvrières qualifiées (16,7%) que pour les employées (8,0%) et les cadres (3,8%).

4.3.2.4 Les facteurs de risque d'incidence des douleurs aux genoux

Les facteurs de risque individuels étudiés (l'âge et l'IMC) étaient associés à l'incidence des douleurs aux genoux de longue durée pour les femmes, mais pas pour les hommes (Tableau 26, Tableau 27). L'âge et l'IMC étaient plus fortement associés aux douleurs de longue durée qu'aux douleurs de courte durée pour les femmes.

Les douleurs aux genoux de longue durée étaient significativement plus souvent incidentes chez les ouvriers que chez les cadres pour les femmes. Un effort physique perçu important et la manipulation de charges de plus de 4 kg augmentaient significativement le risque de douleurs aux genoux de longue durée pour les deux sexes. L'association avec la manipulation de charges était

significativement plus forte pour les douleurs incidentes de longue durée que pour celles de courte durée chez les femmes. S'agenouiller de manière répétée était significativement associé aux douleurs aux genoux de longue durée et de courte durée pour les hommes, mais pas pour les femmes. S'asseoir moins de 4 heures par jour était significativement associé aux douleurs aux genoux de courte durée, mais pas avec les douleurs aux genoux de longue durée pour les hommes.

Après ajustement sur l'âge et l'IMC, la manipulation de charges de plus de 4 kg et le fait de s'agenouiller étaient significativement associés aux douleurs aux genoux de longue durée pour les hommes (seule la manipulation de charges était significative pour les femmes) (Tableau 28,Tableau 29). Il n'y avait pas d'interactions multiplicatives entre s'agenouiller et manipuler des charges (p=0,73 pour les hommes et p=0,37 pour les femmes) ou entre manipuler des charges et l'IMC (p=0,98 pour les hommes, p=0,79 pour les femmes) ou entre s'agenouiller et l'IMC (p=0,69 pour les hommes, p=0,18 pour les femmes). Le port de charges plus de 2 heures par jour n'était pas significativement associé aux douleurs aux genoux.

Pour les hommes, les douleurs initiales à d'autres localisations étaient fortement associées aux douleurs incidentes aux genoux (de courte ou de longue durée). Elles n'étaient plus associées dans les analyses multivariées.

Tableau 26 - Analyse univariée de l'incidence des douleurs aux genoux de courte et de longue durées pour les hommes

Variable	N	Douleurs aux genoux de longue durée vs pas de douleur			Douleurs aux genoux de courte durée vs pas de douleur			P eq*
		N evnt	OR	IC à 95%	N evnt	OR	IC à 95%	
Facteurs personnels								
Age en 2 classes								0,0868
<45 ans	587	37	1,0		87	1,0		
≥45 ans	331	34	**1,7**	**1,0-2,8**	48	1,0	0,7-1,5	
Indice de Masse Corporelle								0,3764
Corpulence normale	517	35	1,0		74	1,0		
Surpoids	336	31	1,5	0,9-2,4	57	1,3	0,9-1,8	
Obésité	65	5	1,0	0,4-2,8	4	0,4	0,1-1,1	
Douleurs à d'autres localisations								0,1263
Pas de douleurs	179	10	1,0		24	1,0		
Douleurs de courte durée	259	30	**2,3**	**1,1-4,8**	38	0,9	0,5-1,7	
Douleurs de longue durée	480	31	1,2	0,6-2,5	73	1,2	0,7-1,9	
Catégorie Socio-Professionnelle								0,6057
Cadres / Manquant	364	23	1,0		49	1,0		
Employés	78	6	1,2	0,5-3,0	8	0,8	0,3-1,6	
Ouvriers qualifiés	334	33	1,7	1,0-3,0	54	1,3	0,8-2,0	
Ouvriers non qualifiés	142	9	1,0	0,5-2,3	24	1,3	0,8-2,2	
Facteurs professionnels biomécaniques								
Effort physique important								0,3763
Non	446	27	1,0		60	1,0		
Oui	472	44	**1,7**	**1,0-2,7**	75	1,3	0,9-1,8	
Port de charges								0,3753
Non	692	49	1,0		98	1,0		
0 à 10 kg	103	7	1,0	0,4-2,3	18	1,3	0,7-2,2	
10 à 25 kg	67	9	2,0	0,9-4,3	8	0,9	0,4-1,9	
Plus de 25 kg	56	6	1,7	0,7-4,2	11	1,6	0,8-3,2	
Manipulation de charges de plus de 4 kg								**0,0156**
Non	735	47	1,0		110	1,0		
Oui	183	24	**2,2**	**1,3-3,7**	25	1,0	0,6-1,6	
S'agenouiller								0,6262
Non	698	46	1,0		92	1,0		
Oui	220	25	**2,0**	**1,2-3,4**	43	**1,7**	**1,1-2,6**	
Se tenir debout								0,2746
Non	571	43	1,0		71	1,0		
Oui	347	28	1,2	0,7-1,9	64	**1,6**	**1,1-2,3**	

*P-value pour le test du Khi-2 d'égalité des ORs pour les douleurs de longue durée et pour les douleurs de courte durée.

Tableau 27 - Analyse univariée de l'incidence des douleurs aux genoux de courte et de longue durées pour les femmes

Variable	N	Douleurs aux genoux de longue durée vs pas de douleur			Douleurs aux genoux de courte durée vs pas de douleur			P eq*
		N evnt	OR	IC à 95%	N evnt	OR	IC à 95%	
Facteurs personnels								
Age en 2 classes								**0,0138**
<45 ans	445	25	1,0		75	1,0		
≥45 ans	253	26	**1,8**	**1,0-3,3**	33	0,8	0,5-1,2	
Indice de Masse Corporelle								**0,0012**
Corpulence normale	500	27	1,0		83	1,0		
Surpoids	138	11	1,5	0,7-3,1	20	0,9	0,5-1,5	
Obésité	60	13	**4,5**	**2,1-9,3**	5	0,6	0,2-1,5	
Douleurs à d'autres localisations								0,3078
Pas de douleurs	141	11	1,0		20	1,0		
Douleurs de courte durée	235	22	1,2	0,6-2,7	36	1,1	0,6-2,0	
Douleurs de longue durée	322	18	0,7	0,3-1,6	52	1,1	0,6-2,0	
Catégorie Socio-Professionnelle								0,1379
Cadres / Manquant	182	10	1,0		34	1,0		
Employés	349	21	1,0	0,5-2,2	45	0,6	0,4-1,0	
Ouvriers qualifiés	54	5	1,9	0,6-6,0	13	1,5	0,7-3,1	
Ouvriers non qualifiés	113	15	**2,5**	**1,1-5,9**	16	0,8	0,4-1,5	
Facteurs professionnels biomécaniques								
Effort physique important								0,5637
Non	398	23	1,0		54	1,0		
Oui	300	28	**1,8**	**1,0-3,2**	54	1,5	1,0-2,2	
Port de charges								0,6732
Non	597	42	1,0		91	1,0		
0 à 10 kg	61	5	1,1	0,4-3,0	8	0,8	0,4-1,9	
10 à 25 kg	30	3	1,7	0,5-6,1	8	2,1	0,9-5,0	
Plus de 25 kg	10	1	1,4	0,2-11,3	1	0,6	0,1-5,2	
Manipulation de charges de plus de 4 kg								**0,0370**
Non	623	40	1,0		98	1,0		
Oui	75	11	**2,5**	**1,2-5,1**	10	0,9	0,4-1,9	
S'agenouiller								0,2233
Non	476	33	1,0		67	1,0		
Oui	222	18	1,3	0,7-2,3	41	1,4	0,9-2,2	
Se tenir debout								0,7450
Non	545	41	1,0		77	1,0		
Oui	153	10	0,9	0,4-1,9	31	1,5	1,0-2,4	

*P-value pour le test du Khi-2 d'égalité des ORs pour les douleurs de longue durée et pour les douleurs de courte durée.

Tableau 28 - Analyse multivariée de l'incidence des douleurs aux genoux de courte et de longue durées pour les hommes

Variable	Douleurs aux genoux de longue durée vs pas de douleur		Douleurs aux genoux de courte durée vs pas de douleur		P eq*
	OR	IC à 95%	OR	IC à 95%	
Age en 2 classes					0,0893
<45 ans	1,0		1,0		
≥45 ans	**1,8**	**1,1-3,1**	1,1	0,7-1,6	
Indice de Masse Corporelle					0,5131
Corpulence normale	1,0		1,0		
Surpoids	1,3	0,8-2,2	1,2	0,8-1,8	
Obésité	0,9	0,3-2,4	0,4	0,1-1,1	
Manipulation de charges de plus de 4 kg					**0,0048**
Non	1,0		1,0		
Oui	**2,2**	**1,2-3,9**	0,8	0,5-1,3	
S'agenouiller					0,6499
Non	1,0		1,0		
Oui	**1,8**	**1,0-3,2**	**1,6**	**1,0-2,4**	
Se tenir debout					0,0726
Non	1,0		1,0		
Oui	0,9	0,5-1,5	**1,5**	**1,0-2,3**	

*P-value pour le test du Khi-2 d'égalité des ORs pour les douleurs de longue durée et pour les douleurs de courte durée.

Tableau 29 - Analyse multivariée de l'incidence des douleurs aux genoux de courte et de longue durées pour les femmes

Variable	Douleurs aux genoux de longue durée vs pas de douleur		Douleurs aux genoux de courte durée vs pas de douleur		P eq*
	OR	IC à 95%	OR	IC à 95%	
Age en 2 classes					**0,0150**
<45 ans	1,0		1,0		
≥45 ans	**1,9**	**1,0-3,4**	0,8	0,5-1,2	
Indice de Masse Corporelle					**0,0016**
Corpulence normale	1,0		1,0		
Surpoids	1,4	0,6-2,9	0,9	0,5-1,5	
Obésité	**4,2**	**2,0-8,9**	0,5	0,2-1,4	
Manipulation de charges de plus de 4 kg					**0,0230**
Non	1,0		1,0		
Oui	**2,4**	**1,1-5,4**	0,7	0,3-1,5	
S'agenouiller					0,1505
Non	1,0		1,0		
Oui	0,8	0,4-1,7	1,5	0,9-2,4	
Se tenir debout					0,3332
Non	1,0		1,0		
Oui	0,9	0,5-1,9	1,4	0,9-2,2	

*P-value pour le test du Khi-2 d'égalité des ORs pour les douleurs de longue durée et pour les douleurs de courte durée.

4.3.2.5 Les fractions attribuables parmi les exposés

La fraction de douleur de longue durée attribuable à la manipulation de charges parmi les exposés était de 54,1% parmi les hommes et 58,5% parmi les femmes. La fraction de douleurs de longue durée attribuable à l'exposition au fait de s'agenouiller plus de 2 heures par jour parmi les exposés était de 45,5% parmi les hommes (36,3% pour les douleurs de courte durée). Enfin, la fraction attribuable à l'exposition au fait d'être debout (en fait, de ne pas être assis) était de 35,0% parmi les hommes.

4.3.2.6 Les analyses sur des sous-populations

4.3.2.6.1 Participants sans douleurs initialement

Il y avait 320 travailleurs sans douleurs (toutes localisations confondues) initialement, les résultats suivants portent sur cette sous-population.

L'âge n'était pas associé aux douleurs de longue durée pour les hommes, mais était significativement associé aux douleurs de longue durée pour les femmes, avec un OR de 5,6 (IC à 95% : [1,3 ;23,5]) pour les travailleurs âgés de plus de 45 ans comparés à ceux de moins de 45 ans.

Sur les 14 hommes obèses, 3 ont développé des douleurs aux genoux, avec un OR ajusté de 7,6 [1,3 ;46,4] pour les travailleurs obèses comparés aux travailleurs avec une corpulence normale. Sur les 12 femmes obèses, 4 ont développé des douleurs aux genoux, avec un OR ajusté de 14,0 [2,1 ;94,5]. S'agenouiller était associé à la limite de la significativité aux douleurs de longue durée chez les hommes (p=0,0808), avec un OR ajusté de 4,4 [1,0 ;20,2]. La manipulation de charges n'était pas significativement associée avec les douleurs de longue durée chez les hommes, avec un OR de 0,2 [0,0 ;2,5]. La manipulation de charges et le fait de s'agenouiller n'étaient pas associés à l'incidence des douleurs chez les femmes. Se tenir debout n'était pas significativement associé aux douleurs incidences parmi les femmes et les hommes sans douleur initialement.

4.3.2.6.2 Participants sans changement de poste

On a trouvé 952 travailleurs qui déclaraient ne pas avoir eu de changement professionnel depuis 2002 (ni changement de poste, ni changement d'entreprise, ni invalidité). Les résultats pour cette sous-population étaient quasiment identiques à ceux trouvés dans la population générale.

4.3.2.7 Analyses pondérées

Les analyses non pondérées donnaient des résultats similaires aux analyses pondérées présentées dans les paragraphes précédents. Le Tableau 30 et le Tableau 31 montraient des résultats quasi-identiques pour le modèle multivarié final sans pondération.

Tableau 30 - Analyse multivariée pondérée de l'incidence des douleurs aux genoux de courte et de longue durée pour les hommes

Variables	Douleurs aux genoux de longue durée vs pas de douleur		Douleurs aux genoux de courte durée vs pas de douleur		P eq*
	OR	IC à 95%	OR	IC à 95%	
Age en 2 classes					**0,0416**
<45 ans	1,0		1,0		
≥45 ans	**1,8**	**1,2-2,7**	1,1	0,8-1,5	
Indice de Masse Corporelle					0,2011
Corpulence normale	1,0		1,0		
Surpoids	1,2	0,8-1,9	1,2	0,9-1,6	
Obésité	1,0	0,5-2,2	**0,4**	**0,2-0,9**	
Manipulation de charges de plus de 4 kg					**0,0041**
Non	1,0		1,0		
Oui	**1,9**	**1,2-3,0**	0,8	0,6-1,2	
S'agenouiller					0,3691
Non	1,0		1,0		
Oui	**2,0**	**1,3-3,1**	**1,6**	**1,1-2,2**	
Se tenir debout					
Non	1,0		1,0		0,0179
Oui	0,9	**0,6-1,3**	1,6	**1,1-2,1**	

*P-value pour le test du Khi-2 d'égalité des ORs pour les douleurs de longue durée et pour les douleurs de courte durée.

Tableau 31 - Analyse multivariée pondérée de l'incidence des douleurs aux genoux de courte et de longue durée pour les femmes

Variables	Douleurs aux genoux de longue durée vs pas de douleur		Douleurs aux genoux de courte durée vs pas de douleur		P eq*
	OR	IC à 95%	OR	IC à 95%	
Age en 2 classes					**0,0416**
<45 ans	1,0		1,0		
≥45 ans	**1,8**	**1,1-3,0**	0,8	0,5-1,2	
Indice de Masse Corporelle					0,2011
Corpulence normale	1,0		1,0		
Surpoids	1,3	0,7-2,4	0,9	0,5-1,3	
Obésité	**4,2**	**2,2-7,7**	0,5	0,2-1,1	
Manipulation de charges de plus de 4 kg					**0,0041**
Non	1,0		1,0		
Oui	**2,4**	**1,3-4,7**	0,7	0,4-1,3	
S'agenouiller					0,3691
Non	1,0		1,0		
Oui	0,8	0,4-1,4	1,5	1,0-2,2	
Se tenir debout					**0,0179**
Non	1,0		1,0		
Oui	0,9	0,5-1,6	1,4	1,0-2,0	

*P-value pour le test du Khi-2 d'égalité des ORs pour les douleurs de longue durée et pour les douleurs de courte durée.

4.3.3 Discussion

Dans une population active, l'incidence des douleurs aux genoux a été estimée à 5,7 pour 100 personnes-années pour les hommes et 6,1 pour 100 personnes-années pour les femmes, avec la définition d'au moins un jour de douleurs dans les 12 derniers mois. L'incidence des douleurs chroniques (au moins 30 jours dans les 12 derniers mois) a été estimée à 2,0 pour 100 personnes-années pour les hommes et 1,9 pour 100 personnes-années pour les femmes. Au vu du temps de suivi moyen de 3,8 années, il n'est pas exclu que la définition choisie pour les cas incidents, c'est-à-dire la déclaration de nouveaux cas de douleurs au moment du suivi, ne reflétait pas l'incidence exacte. En effet, le protocole de l'étude en deux phases ne permettait pas d'évaluer les possibles fluctuations entre les états douloureux et non douloureux : par exemple, les douleurs pouvaient avoir eu lieu avant le premier questionnaire ou entre les deux questionnaires. On pouvait aussi constater moins de douleurs pour les participants suivis que pour les participants non suivis. Cependant, les caractéristiques principales des suivis et non suivis étaient similaires initialement (excepté pour l'âge et les douleurs à d'autres localisations). D'autres études sur l'incidence des douleurs aux genoux ont été effectuées dans des contextes très différents : des déclarations rétrospectives sur 12 ans [57], des déclarations prospectives chaque année pendant deux ans [56] ou un an [55], deux déclarations prospectives séparées de 10 ans [54]. Par exemple, Miranda et al. ont trouvé un taux d'incidence de 10 cas pour 100 travailleurs-années avec plus de 30 jours de douleurs dans l'année qui précède parmi des travailleurs forestiers [55]. L'incidence trouvée dans cette étude était inférieure à l'étude citée, ce qui est attendu étant donné le large éventail d'expositions ou non expositions représentées (pour rappel, 50% des individus déclarent un travail avec un effort physique peu important). Enfin, l'incidence des douleurs aux genoux étaient similaires entre les hommes et les femmes, contrairement à ce qui a été trouvé dans d'autres études.

Cette étude se concentrait sur les douleurs aux genoux en général et pas sur un diagnostic précis (blessure ou arthrose par exemple). Bien qu'un diagnostic spécifique soit nécessaire au niveau individuel pour donner un traitement adapté à la douleur, les interventions primaires en santé au travail se concentrent sur la prévention de la douleur et des limitations fonctionnelles ultérieures. Or, les douleurs aux genoux constituent un déterminant important de l'invalidité, meilleur que le diagnostic plus spécifique d'arthrose aux genoux [50,52]. Puisque la principale cause de douleurs aux genoux après 55 ans est l'arthrose (de manière croissante avec l'âge), la majorité des cas de douleurs aux genoux était ici probablement causés par une arthrose au vu des âges d'inclusion (âge médian : 40 ans) bien que d'autres diagnostics n'étaient pas exclus (46). Les résultats de cette étude confirmaient d'ailleurs ce qui a été trouvé dans les études sur les arthroses aux genoux [66,68]. Comme les études précédentes sur l'arthrose ou les douleurs aux genoux, un âge plus avancé était

associé à une plus forte incidence de douleurs. L'obésité et le surpoids étaient fortement associés à l'incidence des douleurs chroniques pour les femmes mais pas pour les hommes. D'autres études ont trouvé un effet sur l'arthrose pour les hommes et les femmes [59]. La prévalence de l'obésité est plus faible en France que dans d'autres pays, avec une prévalence de 12% en France et de 8% dans cette étude [123]. A titre de comparaison, aux Etats-Unis, la prévalence de l'obésité en 2003-2004 était de 32% [124]. Les associations entre l'obésité et les douleurs aux genoux étaient en conséquence plus difficiles à examiner dans cette étude. Cependant, dans la sous-population sans aucune douleur initialement, l'obésité était associée avec les douleurs chroniques incidentes aux genoux chez les hommes. De plus, l'obésité et le surpoids ont un effet sur toute la vie et pas seulement pendant la carrière professionnelle : l'âge auquel l'IMC est mesuré ici (autour de 40 ans) peut être remis en question, sachant que certaines études montrent qu'il est plus pertinent de considérer l'IMC à 18 ans pour prédire l'arthrose aux genoux [125]. Les différences trouvées entre les hommes et les femmes pour l'effet de l'IMC pouvaient être expliquées par un facteur de confusion non mesuré, les antécédents de blessures aux genoux. Les hommes avec une corpulence normale sont supposés faire plus de sport faisant intervenir les genoux (football, basket...) que les hommes obèses, menant à plus de blessures aux genoux. Pour les femmes, cet effet peut être moins important.

Dans cette étude, se tenir debout au travail n'était pas significativement associé à l'incidence des douleurs aux genoux de longue durée parmi les hommes et les femmes, contrairement à ce qui a été trouvé dans d'autres études [63,126]. Se tenir debout était associé à l'incidence des douleurs aux genoux de courte durée significativement chez les hommes (avec une fraction attribuable de 35%) et significativement à 10% chez les femmes. Certaines études sur l'arthrose aux genoux n'ont également pas trouvé d'association avec les douleurs fortes et une étude a même trouvé un effet protecteur [62,127]. La définition de « se tenir debout » était discutable dans cette étude et ne permettait pas de trancher sur l'effet de « se tenir debout » au travail sur les douleurs aux genoux.

Cette étude confirmait que le fait de s'agenouiller fréquemment pour les hommes et le fait de manipuler des charges lourdes étaient associés à l'incidence des douleurs aux genoux chroniques [55,128]. Le port de charges n'était pas significativement associé aux douleurs aux genoux, contrairement à de précédentes études [62,71,126]. Cependant, porter des charges de plus de 10 kg entrainait un sur-risque non significatif pour les douleurs de longues durées pour les hommes, ce qui est cohérent avec la littérature. S'agenouiller fréquemment n'était pas un facteur de risque significatif chez les femmes. Une explication possible est que s'agenouiller au travail peut avoir une signification différente pour les hommes et pour les femmes [129]. Le seuil de 2 heures par jour n'était vraisemblablement pas à même d'identifier de manière adéquate l'intensité et le type de posture spécifiques aux femmes. De plus, la charge qui portait sur les genoux a peut-être été sous-

estimée chez les femmes car on ne disposait pas de données sur les tâches ménagères qui sont susceptibles d'être plus importantes pour les femmes [70]. Dans cette étude, aucun interaction n'a été mise en évidence entre les facteurs professionnels et l'IMC, contrairement à certaines études transversales sur l'arthrose aux genoux qui avait trouvé des interactions au moins additives et presque multiplicatives [62,130]. En plus de la puissance limitée, les différences trouvées avec les études précédentes pouvaient être expliquées par un biais de sélection : les populations au travail n'incluent pas les cas les plus graves de douleurs aux genoux qui peuvent être exclus du marché du travail (biais du travailleur sain). Le seul facteur professionnel associé significativement avec l'incidence des douleurs aux genoux de courte durée dans cette étude était le fait de s'agenouiller chez les hommes et les femmes. Des facteurs de risque personnels importants sur la survenue de douleurs n'étaient pas disponible dans cette étude, en particulier les antécédents de blessures aux genoux et les activités sportives ou de loisirs impliquant des mouvements des genoux (comme le jardinage ou le vélo). Ces facteurs ont pu plus particulièrement jouer sur les douleurs de courte durée, ce qui expliquerait l'absence de facteurs pertinents associés à ces douleurs dans ces analyses.

5 Discussion générale

5.1 Synthèse des résultats

5.1.1 Epicondylites latérales et symptômes aux coudes

Les deux études réalisées sur les pathologies aux coudes ont souligné l'importance des facteurs professionnels dans l'incidence et la prévalence des épicondylites (Tableau 32). En particulier, un travailleur exposé à un effort physique important cumulé avec des mouvements répétés des coudes (flexion/extension des coudes ou torsion des poignets) avait un risque plus important d'avoir des symptômes douloureux aux coudes ou de développer des épicondylites latérales. Globalement, les tâches répétitives étaient associées à la prévalence et à l'incidence d'épicondylites latérales.

Tableau 32 - Résumé des associations univariées entre les épicondylites et les expositions considérées

	Prévalence des épicondylites et des symptômes aux coudes		***Incidence des épicondylites***
	Symptômes	**Epicondylites**	
Premier auto-questionnaire			
Age	++	++	NS
IMC	♂ :+ / ♀ :NS	NS	
Antécédent de TMS	++	++	
Tâches répétitives	♂ :+ / ♀ :NS	♂ :NS /♀ :++	♂ :NS / ♀ :++
Effort perçu	♂ :+ /♀ :NS	++	
Flexion/extension des coudes	♂ :+ / ♀ :NS	++	
Torsion des poignets	+	♂ :++ / ♀ :+	
Utilisation d'outils vibrants	NS	NS	
Soutien social	NS	♂ :++ / ♀ :NS	NS
Job Strain	NS	NS	
Effort perçu combiné avec torsion des poignets ou flexion/extension des coudes	♂ :++ / ♀ :NS	++	♂ :NS / ♀ :++
Facteurs combinés des deux questionnaires			
Tâches répétitives			++
Soutien social			NS
Effort perçu combiné avec torsion des poignets ou flexion/extension des coudes			++

++ : Risque significatif supérieur à 2
+ : Risque significatif compris entre 1 et 2
NS : Risque non significatif

Les dimensions psychosociales du questionnaire de Karasek étaient peu corrélées à la présence ou à l'occurrence d'épicondylites ; on a observé seulement un lien chez les hommes avec le faible soutien social. Cette absence d'association n'est pas toujours vérifiée dans la littérature ; certaines études trouvent un lien entre les symptômes aux coudes et certaines contraintes psycho-sociales comme le faible soutien social et les exigences élevées au travail pour les métiers avec une forte composante répétitive [37,38]. On pourrait expliquer ces différences par un effet des facteurs psycho-sociaux spécifiquement parmi les métiers répétitifs. Cependant, il n'y avait pas d'interaction multiplicative significative entre les facteurs biomécaniques et psycho-sociaux (testée sur l'étude de prévalence). Une étude précédente sur les mêmes données avait trouvé un effet de certains facteurs psychosociaux (forte demande psychologique et faible latitude décisionnelle) sur les TMS des

membres supérieurs conditionnellement aux facteurs professionnels biomécaniques [108]. Ces résultats laissent penser que l'absence d'effet trouvé est spécifique aux localisations étudiées. La quasi-absence de lien entre les facteurs psycho-sociaux dans cette population et les symptômes aux coudes a empêché d'aller plus avant sur leur mécanisme potentiel d'action dans les études présentées [9,131] (cf. Figure 5, p 17).

L'âge et les antécédents de TMS étaient des facteurs un peu à part dans ces analyses puisqu'ils reflètent en partie les expositions passées. En effet, un jeune salarié ne peut avoir été exposé longtemps aux facteurs de risque de symptômes aux coudes puisque en général, il est depuis peu dans la vie active. A l'inverse, des antécédents de TMS peuvent être un indice d'expositions professionnelles passées qui auraient causé ces TMS.

L'âge avait un effet très important sur la prévalence mais n'a pas d'effet significatif sur l'incidence. L'âge est un facteur de risque connu pour la prévalence [25,45,46] et l'incidence des épicondylites [40]. L'âge restait associé à la prévalence parmi les travailleurs ayant plus de 10 ans d'ancienneté, ce qui laisse penser que le vieillissement des tissus articulaires seul peut avoir un lien avec les symptômes aux coudes [132]. Si cette hypothèse était exacte, on devrait en principe observer un effet de l'âge également sur l'incidence, ce qui n'était pas le cas. L'absence d'effet de l'âge sur l'incidence peut s'expliquer de deux façons :

- La proportion importante de perdus de vue parmi les travailleurs de plus de 45 ans et de moins de 30 ans impliquait un manque de puissance pour estimer l'effet de l'âge (il y a un sur-risque non significatif pour les analyses sur données imputées) ;
- Les travailleurs de plus de 45 ans sans douleur aux coudes initialement constituaient probablement un sous-groupe spécifique, puisqu'ils étaient sans douleur avec pour certains d'entre eux des expositions de longue durée (biais du travailleur sain, cf. paragraphe 1.1.2, p 14).

Il se peut également que l'âge ait peu d'effet sur l'incidence mais ait un effet sur la durée de la maladie, avec une guérison plus lente à un âge plus avancé. Dans ce cas, on aurait bien un effet de l'âge plus fort sur la prévalence que sur l'incidence. Il serait intéressant d'explorer cet aspect.

Dans l'étude sur la prévalence, certains facteurs semblaient associés plus fortement avec les épicondylites qu'avec les symptômes aux coudes, notamment l'âge et l'effort perçu. Ces différences suggèrent une progression des symptômes aux coudes vers les épicondylites en fonction du niveau d'exposition.

L'étude sur l'incidence des épicondylites soulignait l'importance de la dimension temporelle des facteurs de risque professionnels sur l'incidence des épicondylites, en particulier pour les tâches répétitives et l'effort physique important combiné aux mouvements répétitifs au niveau des coudes. Elle indiquait que l'exposition répétée à ces facteurs professionnels augmentait considérablement le risque de développer des épicondylites ultérieurement.

Cette dernière étude était particulièrement concordante avec deux autres études prospectives sur les épicondylites latérales et les expositions professionnelles [119,133] et a fait l'objet d'un éditorial qui détaille en quoi ces études peuvent ou non répondre aux critères de Hill sur la causalité [134].

5.1.2 Incidence des douleurs aux genoux

L'étude sur les douleurs aux genoux soulignait l'incidence importante des douleurs aux genoux dans une population au travail et le rôle des facteurs professionnels, en particulier le fait de devoir s'agenouiller fréquemment et la manipulation de charges (Tableau 33). Ces facteurs étaient surtout associés à l'incidence d'épisodes de douleurs longs.

Tableau 33 - Résumé des associations univariées entre les douleurs aux genoux et les expositions considérées

	Incidence des douleurs aux genoux	
	Douleurs courtes	**Douleurs longues**
Premier auto-questionnaire		
Age	NS	+
IMC	NS	♂ :NS / ♀ :++
Douleurs à d'autres localisations	NS	♂ :++ / ♀ :NS
Effort perçu	NS	+
Manipulation de charges de plus de 4 kg	NS	++
S'agenouiller	♂ :+ / ♀ :NS	♂ :++ / ♀ :NS
Ne pas être assis	♂ :+ / ♀ :NS	NS

++ : Risque significatif supérieur à 2
+ : Risque significatif compris entre 1 et 2
NS : Risque non significatif

5.2 Limites

La population considérée était une population sélectionnée potentiellement sur l'état de santé puisqu'elle incluait des individus qui étaient actifs au premier suivi (en moyenne à 38 ans). Ce problème a été évoqué plusieurs fois dans ce travail sous l'appellation « biais du travailleur sain » et pouvait expliquer que certaines associations attendues n'ont pas été retrouvées. En ce qui concerne la phase de suivi, il n'y avait pas de critère de sélection sur le statut d'activité des individus, ce qui a pu limiter cet effet de sélection.

La qualité des données auto-rapportées en tant que mesure subjective des conditions de travail est toujours une préoccupation pour la santé au travail [7,135]. Une étude de 2005 a montré que les réponses auto-rapportées ont une bonne reproductibilité, en particulier sur l'échelle d'effort de Borg et les contraintes physiques [136]. Cependant, les travailleurs avec des douleurs articulaires étaient plus susceptibles de se rappeler leurs expositions au niveau des coudes et de décrire leur travail

comme ayant beaucoup de contraintes. Ainsi, une influence de l'état de santé du travailleur sur l'exposition professionnelle déclarée n'était pas exclue, en particulier pour la première étude transversale. Pour les études longitudinales sur les travailleurs sans douleurs au niveau de l'articulation d'intérêt initialement, on pouvait penser que cet effet allait être modéré puisque l'état de santé initial était *a priori* plutôt bon. Pour l'analyse de l'incidence des épicondylites qui prenait en compte une mesure d'exposition au suivi (antérieure d'au moins 3 mois à la mesure clinique), on ne pouvait pas complètement éliminer l'hypothèse d'une sur-déclaration de l'exposition due à la présence de symptômes aux coudes puisque ceux-ci ont pu être présents au moment de la déclaration de l'exposition.

Les cas ont été définis dans les deux premières études par le diagnostic posé par le médecin du travail et dans la troisième étude par la déclaration du travailleur. Pour les diagnostics, bien que les médecins du travail étaient au courant des expositions professionnelles des participants, les erreurs de mesure sont supposées être marginales car ils avaient une formation spécifique aux troubles musculo-squelettiques avec des procédures de détection standardisées [20,108,109]. Les symptômes déclarés peuvent présenter les mêmes limites que toute déclaration, sujette aux interprétations individuelles.

Une limitation importante dans ces études était le grand nombre de données manquantes pour le suivi, avec les problèmes potentiels liés au manque de précision. De plus, les suivis étant dépendants des caractéristiques professionnelles et de l'état de santé des individus, les analyses sur cas complets étaient en principe biaisées. Pour pallier ce problème, des analyses adaptées aux données manquantes ont été effectuées : des analyses sur données multiples imputées pour l'incidence des pathologies des coudes et des analyses pondérées pour les pathologies des genoux. Celles-ci sont valables si le mécanisme de données manquantes est MAR (ou MAR séquentiel pour les imputations multiples), ce qui n'a pas été testé ici (cf. paragraphe 2.3.2.1, p 38). Dans ces deux études, les analyses sur cas complets et avec imputations multiples MAR ont donnés des résultats similaires. Les analyses de sensibilité sur le mécanisme de données manquantes pour l'examen clinique (données « missing not at random ») ont été effectuées pour l'analyse de l'incidence des épicondylites et ont montrées une bonne robustesse des résultats à des modifications du mécanisme de données manquantes lié aux conditions de travail et à l'état de santé pendant la deuxième phase de l'étude.

Une autre hypothèse forte dans ce travail concernant les calculs d'incidence était qu'aucun épisode d'épicondylites n'a eu lieu entre les deux mesures sans être encore présente au suivi (cf. paragraphe 2.1.1, p 26). Cette hypothèse peut impliquer une sous-estimation des taux d'incidences. Elle a pu également causer une légère sous-estimation des associations puisqu'il est établi que les travailleurs

manuels sont plus susceptibles d'avoir des épisodes récurrents que les travailleurs non manuels (par exemple pour les épicondylites) [27]. De plus, dans les études d'incidence, on a également supposé que tous les cas avec des antécédents d'épicondylites étaient exclus des analyses. Or, même si des questions rétrospectives sur les antécédents de douleurs avaient été disponibles, il est de manière générale très difficile d'être sûr qu'un individu n'a jamais eu de symptômes articulaires. Ce problème est lié à la difficulté de définir une date de début et de fin pour les TMS sachant que ces pathologies sont progressives et récurrentes [137].

5.3 Applications de ce travail

5.3.1 Pour le chercheur

5.3.1.1 En santé publique

Ces études portaient sur une population représentative de la population active française : la comparaison avec les catégories socio-professionnelles dans l'échantillon par rapport au recensement en population française de 1999 [138] ne montre pas de différence significative, excepté pour les professions qui ne sont pas soumises à la visite obligatoire de médecine du travail (les indépendants et les artisans). De plus, les proportions de travailleurs exposés à différentes expositions professionnelles étaient similaires à celles observées dans l'étude SUMER 2003 [139]. Cette représentativité permet d'avoir des chiffres de référence sur les prévalences des symptômes musculo-squelettiques au niveau des coudes et des genoux. En effet, dans un objectif de santé publique, il est utile de connaitre la prévalence en population au travail des douleurs aux genoux (environ 10%) et des symptômes aux coudes (environ 15%).

Ce travail peut également avoir des conséquences sur l'utilisation des tableaux de reconnaissance des maladies professionnelles, en particulier pour les individus qui ne rentrent pas dans le cadre de ces tableaux : en effet, les travaux présentés augmentent le faisceau de preuves disponibles pour indiquer si l'exposition d'un travailleur est bien la cause ou non de sa maladie. Ce travail pourra également faire évoluer les critères pris en compte dans ces tableaux.

5.3.1.2 Sur les hypothèses de recherche épidémiologique

Les nouveaux outils méthodologiques pour illustrer les hypothèses formulées en épidémiologie sont une façon attrayante de représenter les problématiques posées. Les DAGs sont un outil illustratif intéressant permettant de poser les hypothèses et figurer les conclusions. Les Figure 38, Figure 39 et Figure 40 ci-après illustrent des DAGs compatibles avec les analyses effectuées et les résultats obtenus. Il faut tout de même utiliser cet outil avec prudence : d'une part, plusieurs DAGs peuvent être compatibles avec les mêmes résultats, d'autre part, les données dont on dispose ne sont en général pas en mesure d'établir des liens causaux (facteurs de confusion non mesurés, données

manquantes, données transversales). Il faut donc veiller à ne pas sur-interpréter les DAGs et bien penser qu'il s'agit de figuration d'hypothèses. Dans les recherches à venir, nous pensons que la force des DAGs sera de proposer plusieurs alternatives ou hypothèses de recherche pour une même question et d'évaluer la robustesse des résultats à des variations sur un DAG initial. Par exemple, on pourrait introduire des simulations sur un facteur de confusion non mesuré et voir dans quelle mesure celui-ci peut changer les résultats obtenus [140].

Ces outils doivent être utilisés en complément des critères de causalité introduit par Hill en 1965 [19] : l'exposition doit précéder l'effet, les associations trouvées doivent être comparables à la littérature, les effets causaux doivent avoir une explication biologique sous-jacente.

Figure 38 - DAG compatible avec les résultats obtenus pour la prévalence des symptômes aux coudes

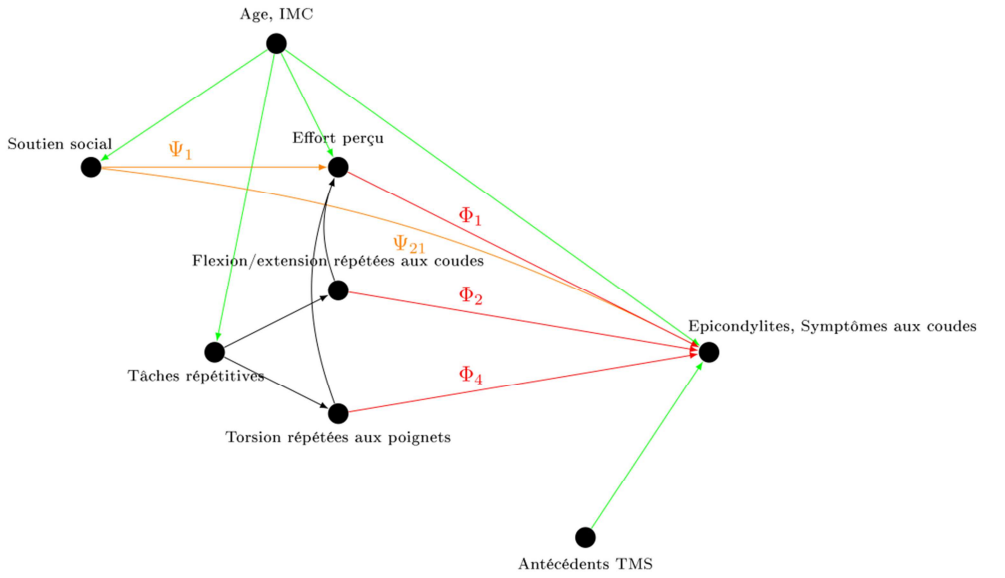


Figure 39 - DAG compatible avec les résultats obtenus pour l'incidence des épicondylites

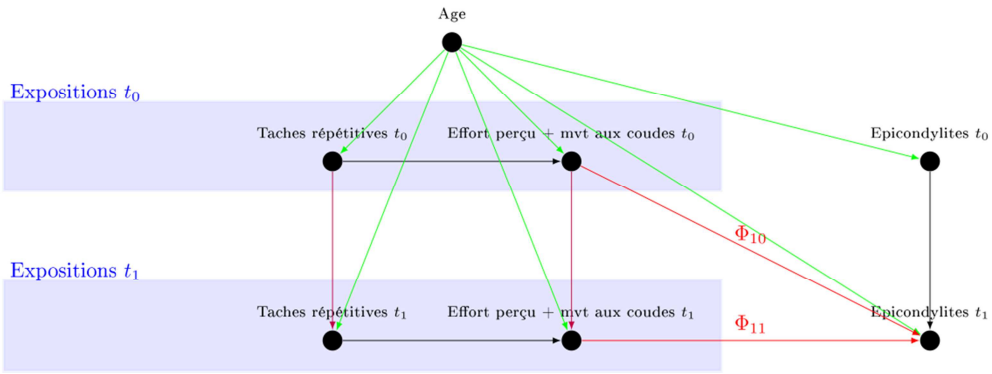
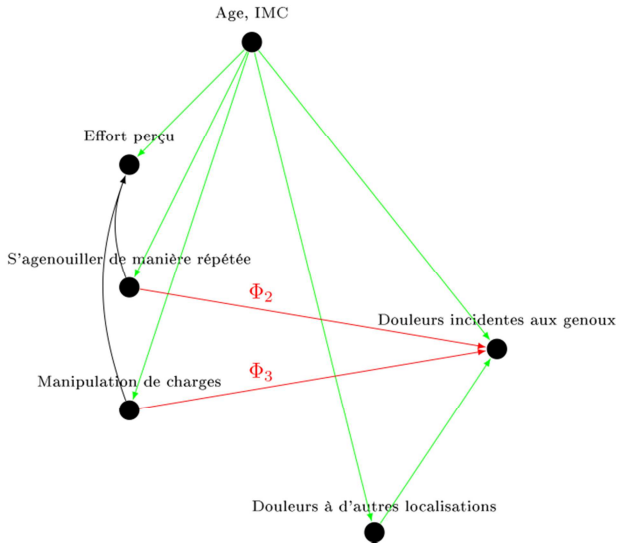


Figure 40 - DAG compatible avec les résultats obtenus pour l'incidence des douleurs aux genoux



5.3.1.3 Sur les études avec des données manquantes

La multiplication des cohortes à grande échelle amène à s'intéresser de plus en plus aux problèmes liés à la participation aux études, et aux réponses que l'on peut y apporter, d'où un intérêt accru pour le traitement des données manquantes. Ce travail décrit de nombreuses méthodes pour traiter les données manquantes et en applique deux différentes : les pondérations et les imputations multiples. Dans les deux cas, les analyses sur les cas complets donnent des résultats similaires aux analyses qui prennent en compte les données manquantes. Cependant, dans le cas des épicondylites, le gain en puissance est non négligeable et certaines estimations varient légèrement. Bien que l'analyse sur cas complets puisse donner dans certains cas précis des résultats non biaisés, cette approche très souvent utilisée est au mieux sous-optimale puisqu'elle n'utilise qu'une partie des informations disponibles.

Diverses barrières existent pour l'utilisation de méthodes plus sophistiquées que les analyses sur cas complets :

- L'implémentation dans les logiciels (SAS, Stata) : La méthode de pondération est facilement réalisable dans les logiciels habituels. Les imputations multiples sont faisables dans SAS, mais ne permettent pas pour l'instant la flexibilité disponible dans Stata (choix de covariables différentes pour chaque modèle d'imputation, imputations dans le cas de données manquante non monotones...). Afin de démocratiser l'usage de ces méthodologies, un point important serait d'établir en routine de ces méthodes d'imputations dans chaque type d'analyse (comme ce qui est disponible pour les pondérations). Certaines méthodes sur les modèles joints entre la variable d'intérêt et le mécanisme de données manquantes ne sont pas implémentées dans les logiciels comme SAS ou Stata. Leur implémentation amènerait un usage plus fréquent de ces méthodes.

- Certains éléments théoriques qui restent à développer : dans le cadre des imputations multiples par « chained equation », il n'y a pas de cadre théorique pour justifier cette méthode qui présente de bonnes performances en pratique. Cela pose en particulier problème pour l'hypothèse d'existence d'une loi jointe des variables imputées, avec des modèles conditionnels qui peuvent être rapidement incompatibles entre eux si on ajoute des interactions.
- La conviction qu'une analyse qui prend en compte les données manquantes demandera beaucoup d'effort et de temps pour un rendement limité, dans le sens où les analyses sur cas complets donneront les mêmes résultats que les analyses plus sophistiquées. La première partie de cette conviction est tout à fait vraie : ces procédures demandent beaucoup plus de temps et de réflexion sur la nature du recueil des données que les analyses « classiques ». Cependant, ces réflexions devront être menées par l'investigateur quelle que soit la nature des analyses qu'il effectuera pour des questions évidentes de validation externe des résultats, et pour savoir quelle est réellement la population cible de l'étude. Pour ce qui concerne les résultats identiques entre les différentes méthodologies, un point souvent oublié est que selon la nature des hypothèses retenues (ajout de variables auxiliaires ou non, variables d'intérêts et/ou variables explicatives manquantes), il est parfois attendu que les différentes méthodes donnent des résultats identiques. La notion d'ajout d'information supplémentaire est cruciale dans cette question. Par exemple, dans le cas où seule la variable d'intérêt a des valeurs manquantes, si on impute celle-ci en utilisant un modèle identique à celui de l'analyse, on s'attend à avoir des résultats identiques, aucune information supplémentaire n'ayant été ajoutée.
- La conviction qu'on peut imputer les variables explicatives mais pas la variable d'intérêt est également erronée. Ici se pose la compréhension des hypothèses retenues pour les données MAR. En effet, cette hypothèse est en réalité très forte puisqu'elle suppose que les données non observées dépendent uniquement des données complètement observées. Si cette hypothèse est vérifiée, il n'y a aucun problème à imputer les variables explicatives et les variables d'intérêt puisque les relations observées sur les cas complets sont valables sur les cas partiellement observés. La stratégie qui consiste à imputer uniquement les variables explicatives et à exclure les individus avec une variable d'intérêt manquante est erronée car si on suppose que le mécanisme de données manquantes pour la variable d'intérêt est MNAR (c'est-à-dire qu'il dépend de variables partiellement observées), on n'a d'autant moins de chance d'avoir un résultat sans biais en appliquant une stratégie « cas complets » sur la variable d'intérêt. Cette réflexion porte bien sur la validité des hypothèses MAR qui est d'autant plus discutable qu'un nombre important de variables sont manquantes et donc que

les associations sur les cas complets sont représentatives de celles sur les cas partiellement observés. Dans le cas où la proportion de données manquantes est importante, il est crucial de discuter de la pertinence des hypothèses MAR et de faire des analyses de sensibilité sur cette hypothèse afin d'évaluer la robustesse des résultats.

- La conviction que le modèle d'imputation doit être exact (avec les bonnes interactions, fonction de lien...) et/ou avoir une très bonne valeur prédictive pour que les analyses finales soient non biaisées. L'aspect de la modélisation a été peu abordé dans ce travail. Cependant, certains travaux montrent que les imputations multiples sont assez robustes vis-à-vis de certaines mauvaises spécifications des modèles d'imputation [99]. Des travaux proposent d'utiliser des stratégies doubles robustes pour se prémunir de mauvaises spécifications des modèles d'imputation ou des modèles du mécanisme de données manquantes [141] mais ces stratégies n'ont pas été appliquées dans ces travaux par manque de temps.

Pour conclure, la prise en compte des données manquantes implique couramment des hypothèses fortes sur les données (mécanismes MAR). Les récentes discussions sur les tests des données manquantes et la *recoverability* des données laissent espérer le développement de méthodes avec des hypothèses plus flexibles qui permettraient une prise en compte optimale des données observées (cf. paragraphe 2.3.2, p 38) [90,92].

5.3.2 Pour le praticien

Les facteurs personnels sont essentiels à prendre en compte pour le repérage et la prise en charge des troubles musculo-squelettiques. Dans ce travail, les expositions professionnelles s'ajoutent aux facteurs personnels sur l'incidence de symptômes aux coudes et aux genoux. Ce travail pousse à une plus grande considération des facteurs professionnels vis à vis des pathologies articulaires. En effet, non seulement les troubles musculo-squelettiques peuvent être dus aux expositions professionnelles passées, mais leur intensité peut être exacerbée par les expositions professionnelles actuelles. A titre d'exemple, ce travail suggère fortement un sur-risque associé à l'exposition prolongée aux facteurs professionnels de mouvements répétés des coudes pour les symptômes aux coudes. Pour conclure, la prise en compte de ces dimensions en dehors du cadre strict de la médecine du travail pourrait permettre une meilleure prise en charge des patients.

5.4 Perspectives

5.4.1 Evaluation des expositions professionnelles

L'évaluation des expositions professionnelles pose le problème du choix de la mesure. Dans ce travail, le parti pris est de choisir des mesures auto-rapportées sur la durée « typique » d'exposition journalière. Ces mesures ne sont pas objectives puisqu'elles font appel à la perception que le travailleur a de sa journée de travail. De plus, elles n'incluent pas de notion d'intensité et de fréquence des expositions qui pourraient modifier fortement l'effet de ces expositions sur les TMS. En ce qui concerne les mesures auto-rapportées, il y a toujours un équilibre à trouver entre le temps de passation de questionnaire et la précision des informations souhaitées. En effet, si le questionnaire est trop long, le risque est d'avoir des questionnaires mal ou peu complétés. Une évaluation ergonomique des expositions par un intervenant extérieur peut être souhaitable pour avoir une mesure plus précise et objective [137]. Mais cette évaluation peut être très couteuse et longue, en particulier si la population à évaluer est grande. Quand la seule information disponible est l'intitulé d'emploi, une autre solution est l'utilisation de matrices emploi-expositions, comme ce qui se fait déjà pour les expositions professionnelles chimiques ou dans le cadre des cancers professionnels par exemple. Initialement, les matrices emploi-expositions ne portaient pas sur les expositions biomécaniques car, selon les expositions biomécaniques considérées, la variabilité pouvait être très grande entre les individus ayant un même intitulé d'emploi [7]. Cependant, des études ont montré qu'il était possible d'évaluer également les expositions biomécaniques avec les intitulés d'emploi et les résultats pouvaient être en accord avec les résultats connus [142,143]. Cependant, la concordance entre les 3 types de mesures restant assez modérée, la question de l'existence d'une mesure de référence pour les expositions biomécaniques reste ouverte.

En plus de la question du type de mesure utilisée, le choix du ou des temps de mesure est fondamental. La question de savoir quand les expositions doivent être mesurée dépend des hypothèses que l'on veut tester sur les liens entre les expositions et les TMS. La plupart des études longitudinales en santé au travail se basent sur les expositions mesurées à l'inclusion. Cette mesure est ponctuelle et dépend de l'état de santé et des expositions qui ont précédé l'inclusion, mais est un bon point de départ pour évaluer les liens avec les TMS. Ce travail a souligné l'importance de l'évaluation répétée de l'exposition professionnelle, puisque l'exposition répétée dans le temps semble plus prédictive de survenue de TMS que l'exposition à l'inclusion seule. Les nouvelles cohortes en population générale en France laissent espérer des recueils d'expositions répétés régulièrement qui permettraient d'affiner les relations trouvées entre les facteurs d'expositions en fonction de leur durée mais aussi de leur répétition dans le temps. Selon les localisations des TMS et les diagnostics, les hypothèses sur les relations temporelles entre les expositions et l'incidence

peuvent différer. Il peut aussi exister des effets seuils en durée ou en intensité, ceci restant pour l'instant largement à explorer. A cela s'ajoute le caractère au moins partiellement choisi de l'exposition professionnelle qui dépend fortement des expositions précédentes et de l'état de santé. Ainsi, si on cherche l'effet causal de l'exposition totale au cours de la carrière, il faut également prendre en compte l'état de santé au cours du temps et ceci ne peut se faire que par des méthodes complexes de g-estimation [144]. De futures recherches permettront une utilisation plus large de ces concepts et de ces méthodes.

Les récentes discussions sur la pénibilité au travail dans le cadre de la réforme des retraites au travail exhibent les mêmes facteurs de risque, comme par exemple porter des charges lourdes plus de 600 heures par an, avoir des postures pénibles plus de 900 heures par an ou avoir un travail répétitif plus de 900 heures par an. Il est intéressant de constater que les facteurs de risque mis en évidence dans ce travail de thèse l'ont été à partir d'une exposition de plus de 2 heures par jour pour les postures contraignantes et la manipulation de charges, soit à partir d'une exposition d'environ 500 heures par an pour un travailleur à temps plein, ce qui est une exposition inférieure à ce qui est prévu pour la définition de pénibilité. Cette discussion souligne l'importance et les implications que peuvent avoir la définition des expositions sur la santé publique en général.

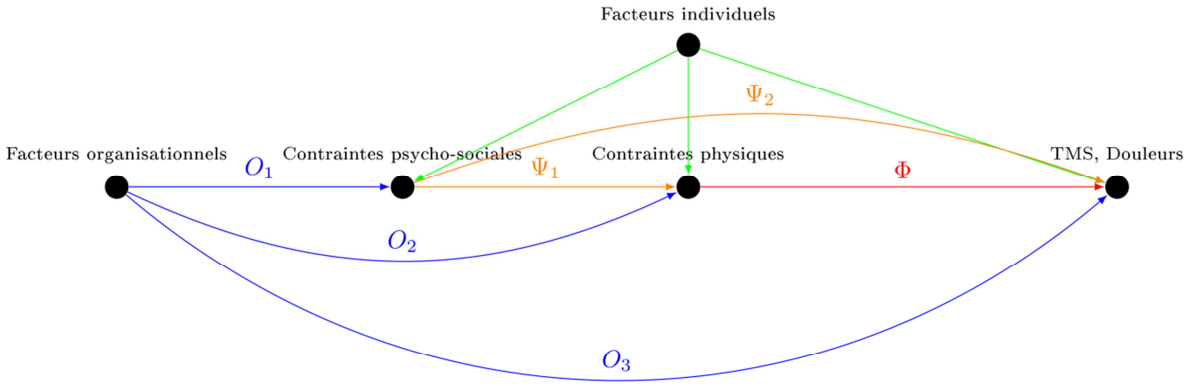
5.4.2 Importance de la définition des cas

Il est difficile de savoir précisément comment définir l'existence de symptômes articulaires : en effet, il existe divers degrés de douleurs (en fréquence, en intensité ou en durée) et diverses natures de symptômes expérimentés (douleurs localisées ou irradiante par exemple). En fonction des objectifs des études, les seuils sont choisis en fonction soit des limitations fonctionnelles qui peuvent en découler ou soit de critère diagnostics sur certaines pathologies. Un critère diagnostic spécifique sera préféré pour une étude de facteurs de risque tandis qu'un critère symptomatique subjectif sera préféré pour une étude dans un objectif de prévention. Les variations dans la définition ou la mesure des cas sont susceptibles de mettre ou non en évidence le rôle de certains facteurs de risque. Dans deux des études de ce travail (l'étude sur la prévalence des symptômes aux coudes et l'étude sur l'incidence des douleurs aux genoux), deux définitions des cas sont prises et les résultats sont comparés en fonction de la définition prise. Les mêmes facteurs de risque professionnels ont été trouvés pour les symptômes aux coudes et les épicondylites (avec une intensité plus grande pour les épicondylites) tandis que les facteurs de risque professionnels de douleurs aux genoux de longue durée étaient différents de ceux des douleurs aux genoux de courte durée. Ce travail encourage à considérer plusieurs définitions des cas dans les études de santé au travail afin d'évaluer la sensibilité des facteurs de risque à ces définitions.

5.4.3 Etude des liens entre les facteurs psychosociaux, les facteurs biomécaniques et les TMS

Dans cette population et au niveau des deux articulations considérées, les facteurs mis en évidence correspondent à des contraintes physiques (mouvements répétitifs, effort physique perçu) et à des contraintes intermédiaires entre les contraintes physiques et psycho-sociales (tâches répétitives). Les contraintes psycho-sociales, ici les facteurs associés au questionnaire de Karasek, n'ont pas montré d'associations significatives avec les symptômes considérés. Bien qu'il s'inscrive dans le schéma causal présenté en introduction (Figure 4) et rappelé en bas de cette page, ce travail n'a pas permis d'éclaircir les relations entre les différentes contraintes professionnelles entre elles et avec les troubles musculo-squelettiques. D'une part, les facteurs organisationnels au niveau des entreprises n'ont pas pu être étudiés car les données auto-rapportées par les individus sont peu adaptées à la mesure de ces facteurs. D'autre part l'absence d'effet des contraintes psycho-sociales étudiées a limité l'analyse de celles-ci. De futurs travaux sur les effets directs et indirects de ces facteurs professionnels seraient bénéfiques à l'éclaircissement de la part causale de chaque type de contrainte professionnelle et les interventions qui pourraient en découler, en particulier, avec l'utilisation de modèles multiniveaux pour prendre en compte les facteurs qui jouent au niveau de l'entreprise et au niveau individuel. En terme de prévention, il serait très intéressant de savoir quel type d'intervention serait le plus efficace pour diminuer la prévalence des troubles musculo-squelettiques, c'est-à-dire des interventions sur les facteurs organisationnels, psycho-sociaux ou biomécaniques (ou encore des interventions combinées).

Figure 4 (duplicata) - Schéma causal des liens entre les différentes contraintes professionnelles et les troubles musculo-squelettiques



Cette thèse avait pour objectif d'explorer les expositions professionnelles associées aux troubles musculo-squelettiques aux coudes et aux genoux. Les facteurs professionnels mis en évidence sont globalement des facteurs de mouvements répétitifs impliquant les articulations étudiées. Le travail en force (manipulation de charges, efforts physiques importants) et les tâches répétitives ont également été mis en évidence dans les études. L'approfondissement des associations trouvées passera par l'utilisation de données de cohortes avec des mesures plus fines des expositions et de nouvelles méthodologies d'analyse des effets longitudinaux et des effets de sélection.

6 Bibliographie

1 Ha C, Roquelaure Y. Troubles musculo-squelettiques d'origine professionnelle en France. Où en est-on aujourd'hui ? TMS d'origine professionnelle. Une préoccupation majeure. BEH n°5-6. 2010.

2 Diricq N. Rapport de la commission instituée par l'article L. 176-2 du code de la Sécurité sociale. 2011. http://www.securite-sociale.fr/IMG/pdf/11_diricq.pdf

3 Breivik H, Collett B, Ventafridda V, et al. Survey of chronic pain in Europe: Prevalence, impact on daily life, and treatment. Eur J Pain 2006;**10**:287–287. doi:10.1016/j.ejpain.2005.06.009

4 Picavet HSJ, Schouten JSAG. Musculoskeletal pain in the Netherlands: prevalences, consequences and risk groups, the DMC3-study. Pain 2003;**102**:167–78. doi:10.1016/s0304-3959(02)00372-x

5 Carayontt P, Smith MJ, Haimst MC. Work organization and psychosocial work factors: Definitions. Int Encycl Ergon Hum Factors 2001;**3**:1612.

6 Huang GD, Feuerstein M, Sauter SL. Occupational stress and work-related upper extremity disorders: Concepts and models. Am J Ind Med 2002;**41**:298–314. doi:10.1002/ajim.10045

7 Leclerc A. Exposure assessment in ergonomic epidemiology: is there something specific to the assessment of biomechanical exposures? Occup Environ Med 2005;**62**:143–4. doi:10.1136/oem.2004.017889

8 Bernard BP, Putz-Anderson V, others. Musculoskeletal disorders and workplace factors: a critical review of epidemiologic evidence for work-related musculoskeletal disorders of the neck, upper extremity, and low back. 1997.

9 Macdonald LA, Härenstam A, Warren ND, et al. Incorporating work organisation into occupational health research: an invitation for dialogue. Occup Environ Med 2008;**65**:1–3. doi:10.1136/oem.2007.033860

10 Marras WS. State-of-the-art research perspectives on musculoskeletal disorder causation and control: the need for an intergraded understanding of risk. J Electromyogr Kinesiol Off J Int Soc Electrophysiol Kinesiol 2004;**14**:1–5. doi:10.1016/j.jelekin.2003.09.005

11 Bongers PM, de Winter CR, Kompier MA, et al. Psychosocial factors at work and musculoskeletal disease. Scand J Work Environ Health 1993;**19**:297–312. doi:10.5271/sjweh.1470

12 Sauter SL, Swanson NG. An ecological model of musculoskeletal disorders in office work. Biomech Psychosoc Asp Musculoskelet Disord Off Work 1996;:3–21.

13 Kompier MA, van der Beek AJ. Psychosocial factors at work and musculoskeletal disorders. Scand J Work Environ Health 2008;**34**:323–5. doi:10.5271/sjweh.1281

14 Punnett L, Cherniack M, Henning R, et al. A Conceptual Framework for Integrating Workplace Health Promotion and Occupational Ergonomics Programs. Public Health Rep 2009;**124**:16.

15 Marras WS, Cutlip RG, Burt SE, et al. National occupational research agenda (NORA) future directions in occupational musculoskeletal disorder health research. Appl Ergon 2009;**40**:15–22. doi:10.1016/j.apergo.2008.01.018

16 Elfering A, Grebner S, Gerber H, et al. Workplace observation of work stressors, catecholamines and musculoskeletal pain among male employees. Scand J Work Environ Health 2008;**34**:337–44. doi:10.5271/sjweh.1280

17 Burdorf A A, van der Beek A. Exposure assessment strategies for work-related risk factors for musculoskeletal disorders. Scand J Work Environ Health 1999;**25**:25–30.

18 Da Costa BR, Vieira ER. Risk factors for work-related musculoskeletal disorders: a systematic review of recent longitudinal studies. Am J Ind Med 2010;**53**:285–323.

19 Grimes DA, Schulz KF. Bias and causal associations in observational research. Lancet 2002;**359**:248–52. doi:10.1016/S0140-6736(02)07451-2

20 Sluiter JK, Rest KM, Frings-Dresen MH. Criteria document for evaluating the work-relatedness of upper-extremity musculoskeletal disorders. Scand J Work Environ Health 2001;**27 Suppl 1**:1–102.

21 Palmer KT, Harris EC, Linaker C, et al. Optimising case definitions of upper limb disorder for aetiological research and prevention: a review. Occup Environ Med Published Online First: 17 October 2011. doi:10.1136/oemed-2011-100086

22 Nandi S, Maschke S, Evans PJ, et al. The Stiff Elbow. HAND 2009;**4**:368–79. doi:10.1007/s11552-009-9181-z

23 Walker-Bone K, Palmer KT, Reading I, et al. Occupation and epicondylitis: a population-based study. Rheumatol Oxf Engl 2012;**51**:305–10. doi:10.1093/rheumatology/ker228

24 Descatha A, Leclerc A, Chastang J-F, et al. Incidence of ulnar nerve entrapment at the elbow in repetitive work. Scand J Work Environ Health 2004;**30**:234–40.

25 Shiri R, Viikari-Juntura E, Varonen H, et al. Prevalence and determinants of lateral and medial epicondylitis: a population study. Am J Epidemiol 2006;**164**:1065–74.

26 Van Rijn RM, Huisstede BMA, Koes BW, et al. Associations between work-related factors and specific disorders at the elbow: a systematic literature review. Rheumatol Oxf Engl 2009;**48**:528–36. doi:10.1093/rheumatology/kep013

27 Shiri R, Viikari-Juntura E. Lateral and medial epicondylitis: Role of occupational factors. Best Pract Res Clin Rheumatol 2011;**25**:43.

28 Walz DM, Newman JS, Konin GP, et al. Epicondylitis: Pathogenesis, Imaging, and Treatment1. Radiographics 2010;**30**:167–84.

29 Descatha A, Leclerc A, Chastang JF, et al. Medial epicondylitis in occupational settings: prevalence, incidence and associated risk factors. J Occup Environ Med Am Coll Occup Environ Med 2003;**45**:993–1001. doi:10.1097/01.jom.0000085888.37273.d9

30 Hamilton PG. The prevalence of humeral epicondylitis: a survey in general practice. J R Coll Gen Pract 1986;**36**:464.

31 Bot SDM, van der Waal JM, Terwee CB, et al. Incidence and prevalence of complaints of the neck and upper extremity in general practice. Ann Rheum Dis 2005;**64**:118–23. doi:10.1136/ard.2003.019349

32 Luopajärvi T, Kuorinka I, Virolainen M, et al. Prevalence of tenosynovitis and other injuries of the upper extremities in repetitive work. Scand J Work Environ Health 1979;**5 suppl 3**:48–55.

33 Viikari-Juntura E, Kurppa K, Kuosma E, et al. Prevalence of epicondylitis and elbow pain in the meat-processing industry. Scand J Work Environ Health 1991;**17**:38–45.

34 Chiang HC, Ko YC, Chen SS, et al. Prevalence of shoulder and upper-limb disorders among workers in the fish-processing industry. Scand J Work Environ Health 1993;**19**:126–31.

35 Ritz BR. Humeral epicondylitis among gas- and waterworks employees. Scand J Work Environ Health 1995;**21**:478–86.

36 Ono Y, Nakamura R, Shimaoka M, et al. Epicondylitis among cooks in nursery schools. Occup Environ Med 1998;**55**:172–9.

37 Haahr JP, Andersen JH. Physical and psychosocial risk factors for lateral epicondylitis: a population based case-referent study. Occup Environ Med 2003;**60**:322–9.

38 Kurppa K, Viikari-Juntura E, Kuosma E, et al. Incidence of tenosynovitis or peritendinitis and epicondylitis in a meat-processing factory. Scand J Work Environ Health 1991;**17**:32–7.

39 Verhaar JA. Tennis elbow. Anatomical, epidemiological and therapeutic aspects. Int Orthop 1994;**18**:263–7.

40 Leclerc A, Landre MF, Chastang JF, et al. Upper-limb disorders in repetitive work. Scand J Work Environ Health 2001;**27**:268–78.

41 Silverstein B, Viikari-Juntura E, Kalat J. Use of a prevention index to identify industries at high risk for work-related musculoskeletal disorders of the neck, back, and upper extremity in Washington state, 1990-1998. Am J Ind Med 2002;**41**:149–69.

42 Roquelaure Y, Mariel J, Fanello S, et al. Active epidemiological surveillance of musculoskeletal disorders in a shoe factory. Occup Environ Med 2002;**59**:452–8.

43 Feleus A, Bierma-Zeinstra SMA, Miedema HS, et al. Prognostic indicators for non-recovery of non-traumatic complaints at arm, neck and shoulder in general practice—6 months follow-up. Rheumatology 2007;**46**:169–76.

44 Walker-Bone K, Reading I, Coggon D, et al. Risk factors for specific upper limb disorders as compared with non-specific upper limb pain: assessing the utility of a structured examination schedule. Occup Med 2006;**56**:243–50.

45 Fan ZJ, Silverstein BA, Bao S, et al. Quantitative exposure-response relations between physical workload and prevalence of lateral epicondylitis in a working population. Am J Ind Med 2009;**52**:479–90. doi:10.1002/ajim.20700

46 Herquelot E, Bodin J, Roquelaure Y, et al. Work-related risk factors for lateral epicondylitis and other cause of elbow pain in the working population. Am J Ind Med 2013;**56**:400–9. doi:10.1002/ajim.22140

47 Palmer KT, Harris EC, Coggon D. Compensating occupationally related tenosynovitis and epicondylitis: a literature review. Occup Med Oxf Engl 2007;**57**:67–74. doi:10.1093/occmed/kql127

48 Felson DT, Hannan MT, Naimark A, et al. Occupational physical demands, knee bending, and knee osteoarthritis: results from the Framingham Study. J Rheumatol 1991;**18**:1587–92.

49 Bedson J, Croft PR. The discordance between clinical and radiographic knee osteoarthritis: A systematic search and summary of the literature. BMC Musculoskelet Disord 2008;**9**:116. doi:10.1186/1471-2474-9-116

50 Peat G, McCarney R, Croft P. Knee pain and osteoarthritis in older adults: a review of community burden and current use of primary health care. Ann Rheum Dis 2001;**60**:91–7. doi:10.1136/ard.60.2.91

51 Agaliotis M, Mackey MG, Jan S, et al. Burden of reduced work productivity among people with chronic knee pain: a systematic review. Occup Environ Med 2014;:oemed – 2013–101997. doi:10.1136/oemed-2013-101997

52 McAlindon TE, Cooper C, Kirwan JR, et al. Determinants of disability in osteoarthritis of the knee. Ann Rheum Dis 1993;**52**:258–62. doi:10.1136/ard.52.4.258

53 Andersen LL, Clausen T, Carneiro IG, et al. Spreading of chronic pain between body regions: prospective cohort study among health care workers. Eur J Pain Lond Engl 2012;**16**:1437–43. doi:10.1002/j.1532-2149.2012.00143.x

54 Descatha A, Teysseyre D, Cyr D, et al. Long-term effects of biomechanical exposure on severe shoulder pain in the Gazel cohort. Scand J Work Environ Health 2012;**38**:568–76.

55 Miranda H, Viikari-Juntura E, Martikainen R, et al. A prospective study on knee pain and its risk factors. Osteoarthritis Cartilage 2002;**10**:623–30.

56 Jones GT, Harkness EF, Nahit ES, et al. Predicting the onset of knee pain: results from a 2-year prospective study of new workers. Ann Rheum Dis 2007;**66**:400–6. doi:10.1136/ard.2006.057570

57 Ingham SL, Zhang W, Doherty SA, et al. Incident knee pain in the Nottingham community: a 12-year retrospective cohort study. Osteoarthritis Cartilage 2011;**19**:847–52. doi:10.1016/j.joca.2011.03.012

58 Andersen LL, Clausen T, Persson R, et al. Perceived physical exertion during healthcare work and risk of chronic pain in different body regions: prospective cohort study. Int Arch Occup Environ Health 2013;**86**:681–7. doi:10.1007/s00420-012-0808-y

59 Jiang L, Tian W, Wang Y, et al. Body mass index and susceptibility to knee osteoarthritis: A systematic review and meta-analysis. Joint Bone Spine 2012;**79**:291–7. doi:10.1016/j.jbspin.2011.05.015

60 Heidari B. Knee osteoarthritis prevalence, risk factors, pathogenesis and features: Part I. Casp J Intern Med 2011;**2**:205.

61 Järvholm B, From C, Lewold S, et al. Incidence of surgically treated osteoarthritis in the hip and knee in male construction workers. Occup Environ Med 2008;**65**:275–8.

62 Coggon D, Croft P, Kellingray S, et al. Occupational physical activities and osteoarthritis of the knee. Arthritis Rheum 2000;**43**:1443–9.

63 Allen KD, Chen J-C, Callahan LF, et al. Associations of Occupational Tasks with Knee and Hip Osteoarthritis: The Johnston County Osteoarthritis Project. J Rheumatol 2010;**37**:842–50. doi:10.3899/jrheum.090302

64 Jensen LK, Eenberg W. Occupation as a risk factor for knee disorders. Scand J Work Environ Health 1996;:165–75.

65 Vignon E, Valat J-P, Rossignol M, et al. Osteoarthritis of the knee and hip and activity: a systematic international review and synthesis (OASIS). Jt Bone Spine Rev Rhum 2006;**73**:442–55. doi:10.1016/j.jbspin.2006.03.001

66 Jensen LK. Knee osteoarthritis: influence of work involving heavy lifting, kneeling, climbing stairs or ladders, or kneeling/squatting combined with heavy lifting. Occup Environ Med 2008;**65**:72–89.

67 Fransen M, Agaliotis M, Bridgett L, et al. Hip and knee pain: Role of occupational factors. Best Pract Res Clin Rheumatol 2011;**25**:81–101. doi:10.1016/j.berh.2011.01.012

68 McWilliams DF, Leeb BF, Muthuri SG, et al. Occupational risk factors for osteoarthritis of the knee: a meta-analysis. Osteoarthritis Cartilage 2011;**19**:829–39. doi:10.1016/j.joca.2011.02.016

69 Palmer KT. Occupational activities and osteoarthritis of the knee. Br Med Bull 2012;**102**:147–70. doi:10.1093/bmb/lds012

70 Ezzat AM, Li LC. Occupational physical loading tasks and knee osteoarthritis: a review of the evidence. Physiother Can Physiothérapie Can 2014;**66**:91–107. doi:10.3138/ptc.2012-45BC

71 Lau EC, Cooper C, Lam D, et al. Factors Associated with Osteoarthritis of the Hip and Knee in Hong Kong Chinese: Obesity, Joint Injury, and Occupational Activities. Am J Epidemiol 2000;**152**:855–62. doi:10.1093/aje/152.9.855

72 Seidler A, Bolm-Audorff U, Abolmaali N, et al. The role of cumulative physical work load in symptomatic knee osteoarthritis - a case-control study in Germany. J Occup Med Toxicol 2008;**3**:14. doi:10.1186/1745-6673-3-14

73 Vingård E, Alfredsson L, Goldie I, et al. Occupation and Osteoarthrosis of the Hip and Knee: A Register-Based Cohort Study. Int J Epidemiol 1991;**20**:1025–31. doi:10.1093/ije/20.4.1025

74 Klussmann A, Gebhardt H, Nübling M, et al. Research article Individual and occupational risk factors for knee osteoarthritis: results of a case-control study in Germany. Arthritis Res Ther 2010;**12**:R88.

75 Agresti A. Categorical Data Analysis. John Wiley & Sons 2003.

76 Menvielle G, Leclerc A. Glossaire statistique et épidémiologique. Pathol. Prof. Environ. EMC Encycl. Méd.-Chir.

77 Kleinbaum DG, Sullivan KM, Barker ND. A Pocket Guide to Epidemiology. Springer Science & Business Media 2007.

78 Steenland K, Armstrong B. An overview of methods for calculating the burden of disease due to specific risk factors. Epidemiol Camb Mass 2006;**17**:512–9. doi:10.1097/01.ede.0000229155.05644.43

79 Hanley JA. A heuristic approach to the formulas for population attributable fraction. J Epidemiol Community Health 2001;**55**:508–14.

80 Greenland S, Robins JM. Conceptual problems in the definition and interpretation of attributable fractions. Am J Epidemiol 1988;**128**:1185–97.

81 Greenland S, Pearl J, Robins JM. Causal diagrams for epidemiologic research. Epidemiology 1999;:37–48.

82 Shrier I, Platt RW. Reducing bias through directed acyclic graphs. BMC Med Res Methodol 2008;**8**:70. doi:10.1186/1471-2288-8-70

83 Tsai C-L, Camargo Jr. CA. Methodological considerations, such as directed acyclic graphs, for studying "acute on chronic" disease epidemiology: Chronic obstructive pulmonary disease example. J Clin Epidemiol 2009;**62**:982–90. doi:10.1016/j.jclinepi.2008.10.005

84 VanderWeele TJ, Hernán MA, Robins JM. Causal directed acyclic graphs and the direction of unmeasured confounding bias. Epidemiol Camb Mass 2008;**19**:720–8. doi:10.1097/EDE.0b013e3181810e29

85 Hernan MA, Hernandez-Diaz S, Robins JM. A Structural Approach to Selection Bias. Epidemiol Sept 2004 2004;**15**:615–25. doi:10.1097/01.ede.0000135174.63482.43

86 VanderWeele TJ, Hernán MA. Results on Differential and Dependent Measurement Error of the Exposure and the Outcome Using Signed Directed Acyclic Graphs. Am J Epidemiol 2012;**175**:1303–10. doi:10.1093/aje/kwr458

87 VanderWeele TJ, Robins JM. Four Types of Effect Modification: A Classification Based on Directed Acyclic Graphs. Epidemiology 2007;**18**:561–8. doi:10.1097/EDE.0b013e318127181b

88 Potthoff RF, Tudor GE, Pieper KS, et al. Can one assess whether missing data are missing at random in medical studies? Stat Methods Med Res 2006;**15**:213–34.

89 Mohan K, Pearl J, Jin T. Missing data as a causal inference problem. In: Proceedings of the Neural Information Processing Systems Conference (NIPS). 2013.

90 Pearl J, Mohan K. Recoverability and testability of missing data: Introduction and summary of results. 2013.

91 Little RJA, Rubin DB. Statistical analysis with missing data. Wiley New York 1987.

92 Mohan K, Pearl J. On the testability of models with missing data. In: Proceedings of the Seventeenth International Conference on Artificial Intelligence and Statistics. 2014. 643–50.

93 Pearl J. The Mathematics of Causal Inference. Rochester, NY: : Social Science Research Network 2013.

94 Gill RD, Robins JM. Sequential models for coarsening and missingness. In: Proceedings of the First Seattle Symposium in Biostatistics. Springer 1997. 295–305.

95 Pearl J. The Deductive Approach to Causal Inference. 2014.

96 Daniel RM, Kenward MG, Cousens SN, et al. Using causal diagrams to guide analysis in missing data problems. Stat Methods Med Res 2012;**21**:243–56.

97 Schafer JL. Analysis of incomplete multivariate data. CRC press 1997.

98 Schafer JL, Graham JW. Missing data: our view of the state of the art. Psychol Methods 2002;**7**:147.

99 Lee KJ, Carlin JB. Multiple imputation for missing data: fully conditional specification versus multivariate normal imputation. Am J Epidemiol 2010;**171**:624–32.

100 Van Buuren S, Oudshoorn K. Flexible multivariate imputation by MICE. Leiden Neth TNO Prev Cent 1999.

101 Raghunathan TE, Lepkowski JM, Van Hoewyk J, et al. A multivariate technique for multiply imputing missing values using a sequence of regression models. Surv Methodol 2001;**27**:85–96.

102 Buuren S van. Multiple imputation of discrete and continuous data by fully conditional specification. Stat Methods Med Res 2007;**16**:219–42. doi:10.1177/0962280206074463

103 Buuren S van, Brand JPL, Groothuis-Oudshoorn K, et al. Fully conditional specification in multivariate imputation. J Stat Comput Simul 2006;**76**:1049–64.

104 White IR, Royston P, Wood AM. Multiple imputation using chained equations: Issues and guidance for practice. Stat Med 2011;**30**:377–99.

105 Carpenter JR, Kenward MG, White IR. Sensitivity analysis after multiple imputation under missing at random: a weighting approach. Stat Methods Med Res 2007;**16**:259–75. doi:10.1177/0962280206075303

106 Michiels B, Molenberghs G, Lipsitz SR. Selection Models and Pattern-Mixture Models for Incomplete Data with Covariates. Biometrics 1999;**55**:978–83. doi:10.1111/j.0006-341X.1999.00978.x

107 Enders CK. Applied missing data analysis. Guilford Press 2010.

108 Roquelaure Y, Ha C, Leclerc A, et al. Epidemiologic surveillance of upper-extremity musculoskeletal disorders in the working population. Arthritis Rheum 2006;**55**:765–78. doi:10.1002/art.22222

109 Ha C, Roquelaure Y, Leclerc A, et al. The French Musculoskeletal Disorders Surveillance Program: Pays de la Loire network. Occup Environ Med 2009;**66**:471–9. doi:10.1136/oem.2008.042812

110 Niedhammer I, Chastang JF, Gendrey L, et al. [Psychometric properties of the French version of Karasek's "Job Content Questionnaire" and its scales measuring psychological pressures, decisional latitude and social support: the results of the SUMER]. Santé Publique Vandoeuvre-Lès-Nancy Fr 2006;**18**:413–27.

111 Karasek RA. Job Demands, Job Decision Latitude, and Mental Strain: Implications for Job Redesign. Adm Sci Q 1979;**24**:285–308.

112 Borg G. Borg's perceived exertion and pain scales. Human kinetics 1998.

113 Kuorinka I, Jonsson B, Kilbom A, et al. Standardised Nordic questionnaires for the analysis of musculoskeletal symptoms. Appl Ergon 1987;**18**:233–7. doi:10.1016/0003-6870(87)90010-X

114 Bodin J, Ha C, Petit Le Manac'h A, et al. Risk factors for incidence of rotator cuff syndrome in a large working population. Scand J Work Environ Health 2012;**38**:436–46. doi:10.5271/sjweh.3285

115 Flegal KM, Graubard BI, Williamson DF, et al. Cause-specific excess deaths associated with underweight, overweight, and obesity. J Am Med Assoc 2007;**298**:2028–37. doi:10.1001/jama.298.17.2028

116 Hansson GA, Balogh I, Ohlsson K, et al. Impact of physical exposure on neck and upper limb disorders in female workers. Appl Ergon 2000;**31**:301–10.

117 Messing K, Stock SR, Tissot F. Should studies of risk factors for musculoskeletal disorders be stratified by gender? Lessons from the 1998 Québec Health and Social Survey. Scand J Work Environ Health 2009;**35**:96–112.

118 Silverstein B, Fan ZJ, Smith CK, et al. Gender adjustment or stratification in discerning upper extremity musculoskeletal disorder risk? Scand J Work Environ Health 2009;**35**:113–26.

119 Fan ZJ, Silverstein BA, Bao S, et al. The association between combination of hand force and forearm posture and incidence of lateral epicondylitis in a working population. Hum Factors 2014;**56**:151–65.

120 Royston P. Multiple imputation of missing values: update of ice. Stata J 2005;**5**:527.

121 Dupray A, Recotillet I. Mobilités professionnelles et cycle de vie. 2009.

122 Eygendaal D, Rahussen FTG, Diercks RL. Biomechanics of the elbow joint in tennis players and relation to pathology. Br J Sports Med 2007;**41**:820–3. doi:10.1136/bjsm.2007.038307

123 Diouf I, Charles MA, Ducimetière P, et al. Evolution of Obesity Prevalence in France: An Age-Period-Cohort Analysis. Epidemiology 2010;**21**:360–5. doi:10.1097/EDE.0b013e3181d5bff5

124 Ogden CL, Carroll MD, Curtin LR, et al. Prevalence of overweight and obesity in the united states, 1999-2004. JAMA 2006;**295**:1549–55. doi:10.1001/jama.295.13.1549

125 Jiang L, Rong J, Wang Y, et al. The relationship between body mass index and hip osteoarthritis: a systematic review and meta-analysis. Joint Bone Spine 2011;**78**:150–5.

126 Sandmark H, Hogstedt C, Vingård E. Primary osteoarthrosis of the knee in men and women as a result of lifelong physical load from work. Scand J Work Environ Health 2000;**26**:20–5. doi:10.5271/sjweh.505

127 Manninen P, Heliövaara M, Riihimäki H, et al. Physical workload and the risk of severe knee osteoarthritis. Scand J Work Environ Health 2002;:25–32.

128 Jensen LK, Rytter S, Marott JL, et al. Relationship between years in the trade and the development of radiographic knee osteoarthritis and MRI-detected meniscal tears and bursitis in floor layers. A cross-sectional study of a historical cohort. BMJ Open 2012;**2**.

129 Messing K, Stock SR, Tissot F. Should studies of risk factors for musculoskeletal disorders be stratified by gender? Lessons from the 1998 Québec Health and Social Survey. Scand J Work Environ Health 2009;**35**:96–112.

130 Vrezas I, Elsner G, Bolm-Audorff U, et al. Case–control study of knee osteoarthritis and lifestyle factors considering their interaction with physical workload. Int Arch Occup Environ Health 2010;**83**:291–300. doi:10.1007/s00420-009-0486-6

131 Macfarlane GJ, Pallewatte N, Paudyal P, et al. Evaluation of work-related psychosocial factors and regional musculoskeletal pain: results from a EULAR Task Force. Ann Rheum Dis 2009;**68**:885–91. doi:10.1136/ard.2008.090829

132 Hagberg M. Clinical assessment of musculoskeletal disorders in workers exposed to hand-arm vibration. Int Arch Occup Environ Health 2002;**75**:97–105.

133 Descatha A, Dale AM, Jaegers L, et al. Self-reported physical exposure association with medial and lateral epicondylitis incidence in a large longitudinal study. Occup Environ Med 2013;**70**:670–3. doi:10.1136/oemed-2012-101341

134 Descatha A, Dale AM, Silverstein BA, et al. Lateral epicondylitis: New evidence for work relatedness. Jt Bone Spine Rev Rhum Published Online First: 29 December 2014. doi:10.1016/j.jbspin.2014.10.013

135 Barrero LH, Katz JN, Dennerlein JT. Validity of self-reported mechanical demands for occupational epidemiologic research of musculoskeletal disorders. Scand J Work Environ Health 2009;**35**:245–60.

136 Stock SR, Fernandes R, Delisle A, et al. Reproducibility and validity of workers' self-reports of physical work demands. Scand J Work Environ Health 2005;**31**:409–37.

137 Punnett L, Wegman DH. Work-related musculoskeletal disorders: the epidemiologic evidence and the debate. J Electromyogr Kinesiol 2004;**14**:13–23. doi:10.1016/j.jelekin.2003.09.015

138 INSEE des Pays-de-la-Loire. INSEE Pays-de-la-Loire. Tableaux économiques des Pays de la Loire: 2000–2001. 2001.

139 Niedhammer I, Chastang J-F, David S, et al. The contribution of occupational factors to social inequalities in health: Findings from the national French SUMER survey. Soc Sci Med 2008;**67**:1870–81. doi:10.1016/j.socscimed.2008.09.007

140 Richiardi L, Bellocco R, Zugna D. Mediation analysis in epidemiology: methods, interpretation and bias. Int J Epidemiol 2013;**42**:1511–9.

141 McIsaac M, Cook RJ. Statistical methods for incomplete data: Some results on model misspecification. Stat Methods Med Res 2014;:0962280214544251. doi:10.1177/0962280214544251

142 Rubak TS, Svendsen SW, Andersen JH, et al. An expert-based job exposure matrix for large scale epidemiologic studies of primary hip and knee osteoarthritis: The Lower Body JEM. BMC Musculoskelet Disord 2014;**15**:204.

143 Evanoff B, Zeringue A, Franzblau A, et al. Using job-title-based physical exposures from O*NET in an epidemiological study of carpal tunnel syndrome. Hum Factors 2014;**56**:166–77.

144 Chevrier J, Picciotto S, Eisen EA. A Comparison of Standard Methods With G-estimation of Accelerated Failure-time Models to Address The Healthy-worker Survivor Effect: Application in a Cohort of Autoworkers Exposed to Metalworking Fluids. Epidemiology 2012;**23**:212–9. doi:10.1097/EDE.0b013e318245fc06

Annexe A Exemples de DAG avec plus de deux variables manquantes

Dans le paragraphe 2.3.3.3 (p 48), différents DAGs sont présentés avec des données manquantes MCAR ou MAR. L'objectif de cette partie est de montrer différents exemples plus complexes de données manquantes MAR séquentiels et MNAR.

On repart de l'exemple 7 (cf. paragraphe 2.3.3.3, p 48 et Figure 20, p 51). On a une exposition professionnelle binaire X, une mesure de la douleur W et les changements professionnels T recueillis par un auto-questionnaire. La présence de TMS, Y, est évaluée par un examen clinique. Ces données sont partiellement observées, avec deux mécanismes de données manquantes qui génèrent $R_X = R_W = R_T$ et $\mathrm{R_Y}$.

On souhaite étudier l'association entre l'exposition professionnelle et les TMS :

$$Q = \frac{P(Y = 1|X = 1)}{P(Y = 1|X = 0)}$$

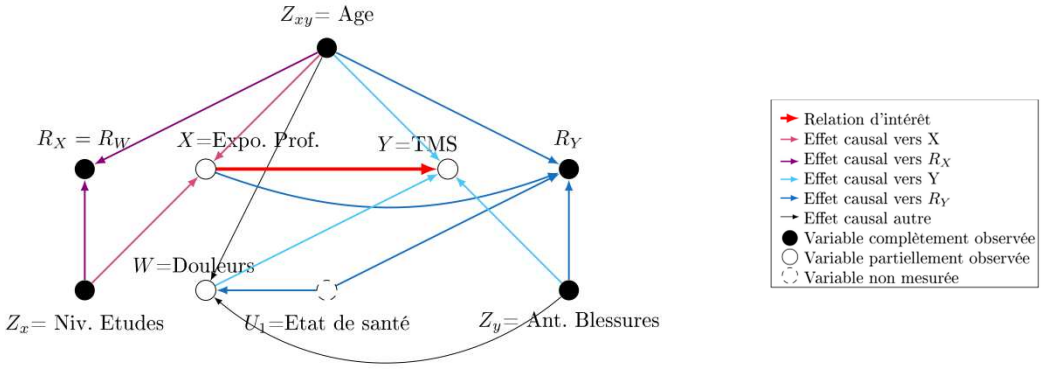
Trois hypothèses sur les mécanismes des données manquantes sont explorées :

- Exemple A1 : le mécanisme de données manquantes de Y dépend de l'état de santé général non mesuré (U_1) et de X
- Exemple A2 : le mécanisme de données manquantes de Y dépend de l'état de santé général non mesuré (U_1), de X et du statut d'emploi au suivi pour les TMS non mesuré (actif/inactif, U_2).
- Exemple A3 : le mécanisme de données manquantes de Y dépend de l'état de santé général non mesuré (U_1) et du statut d'emploi au suivi pour les TMS non mesuré (actif/inactif, U_2). De plus, le mécanisme de données manquantes de X est causé par la valeur de X.

Exemple A1 :

On suppose que la mesure de douleur (W) partiellement mesurée est un facteur intermédiaire entre l'état de santé global non mesuré (U_1) et les TMS Y (Figure 41).

Figure 41 - DAG avec un mécanisme MAR séquentiel (exemple A1)



Le mécanisme R_X est MAR (car il ne dépend que de Z_x et Z_{xy}) et le mécanisme R_Y est MNAR puisqu'il dépend de W, lui-même associé au mécanisme de données manquantes R_X qui est MAR. Le mécanisme global de données manquantes est MAR séquentiel puisqu'en ordonnant les variables partiellement observées en (Y, X, W), on voit sur le DAG que $W \coprod R_X \,|Z_{xy}$; $X \coprod R_X \,|Z_{xy}, Z_x$; $Y \coprod R_Y \,|Z_{xy}, Z_y, X, W$ et $Y \coprod R_X \,|Z_{xy}, X$

Dans les équations suivantes, on utilise également que $Z_y \coprod Z_{xy} \coprod W$.

On a, en notant ζ l'ensemble de valeurs que prend Z_y, η l'ensemble de valeurs que prend Z_{xy} et ξ les valeurs que prend W :

$$P(Y|X) = \sum_{z_1 \in \zeta, z_2 \in \eta, w \in \xi} P\big(Y\big|X, Z_y = z_1, Z_{xy} = z_2, W = w\big)\, P\big(Z_y = z_1\big) P\big(Z_{xy} = z_2\big) P(W = w)$$

$$= \sum_{z_1 \in \zeta, z_2 \in \eta, w \in \xi} P\big(Y^*\big|X^*, R_X = 0, R_Y = 0, Z_y = z_1, Z_{xy} = z_2, W^* = w\big)\, P\big(Z_y = z_1\big) P\big(Z_{xy} = z_2\big) P(W = w)$$

W n'est pas totalement observé, mais par $W \coprod R_X|Z_{xy}$, on a :

$$P(W = w) = \sum_{z_2 \in \eta} P\big(W\big|Z_{xy} = z_2\big)\, P\big(Z_{xy} = z_2\big)$$

$$= \sum_{z_2 \in \eta} P\big(W^*\big|Z_{xy} = z_2, R_X = 0\big)\, P\big(Z_{xy} = z_2\big)$$

Ainsi, on voit que la décomposition de la distribution conditionnelle $P(Y|X)$ en plusieurs facteurs donne une idée de la procédure à suivre pour estimer le risque relatif Q :

(1) il faut estimer la probabilité d'avoir des douleurs $P(W = w)$ en prenant en compte Z_{xy} (les antécédents de blessures)

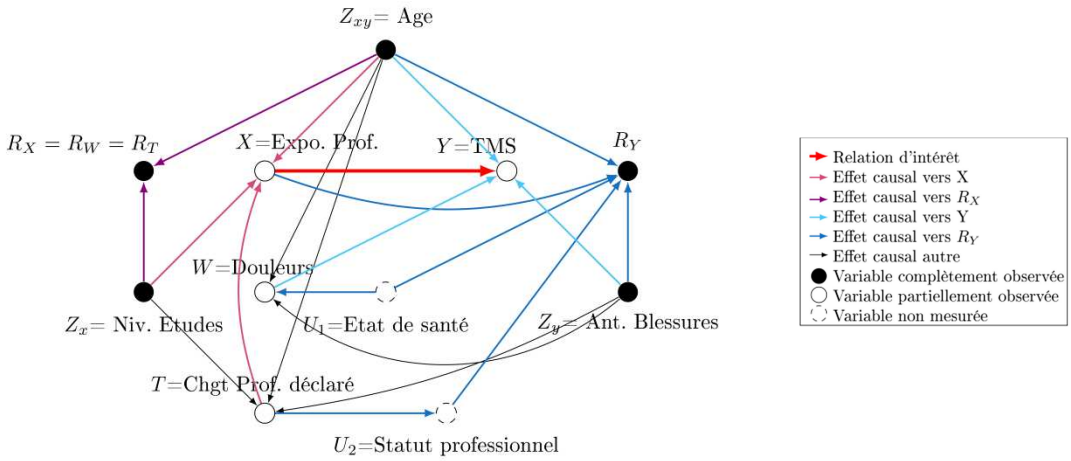
(2) *il faut estimer ensuite la probabilité d'avoir un TMS conditionnellement aux expositions X et aux douleurs W, $P(Y|X, W = w)$, en prenant en compte Z_y et Z_{xy} dans les analyses (les antécédents de blessures, l'âge et les douleurs déclarées)*

(3) *on peut alors estimer sans biais $P(Y|X)$ puis le risque relatif Q.*

<u>Exemple A2 :</u>

On suppose que les changements professionnels (T) partiellement mesurés causent les données manquantes pour les TMS Y (R_Y) passant par le statut professionnel au suivi (U_2). (Figure 42).

Figure 42 - DAG avec un mécanisme MAR séquentiel (exemple A2)



Le mécanisme R_X est MAR (car il ne dépend que de Z_x et Z_{xy}) et le mécanisme R_Y est MNAR puisqu'il dépend de X, W et T, eux-mêmes associés au mécanisme de données manquantes R_X qui est MAR. Le mécanisme global de données manquantes est MAR séquentiel puisqu'en ordonnant variables partiellement observées en (Y, X, W, T), on voit sur le DAG que $W \perp\!\!\!\perp R_X | Z_{xy}$; $X \perp\!\!\!\perp R_X | Z_{xy}, Z_x$; $T \perp\!\!\!\perp R_X | Z_x, Z_{xy}$; $Y \perp\!\!\!\perp R_X | X, Z_x, Z_{xy}$ et $Y \perp\!\!\!\perp R_Y | Z_y, Z_{xy}, W$.

De la même façon que pour l'exemple A1, on peut décomposer la distribution $P(Y|X)$ sur Z_y, Z_{xy} et W. Ainsi, par la même procédure que précédemment, on peut estimer sans biais le risque relatif Q recherché.

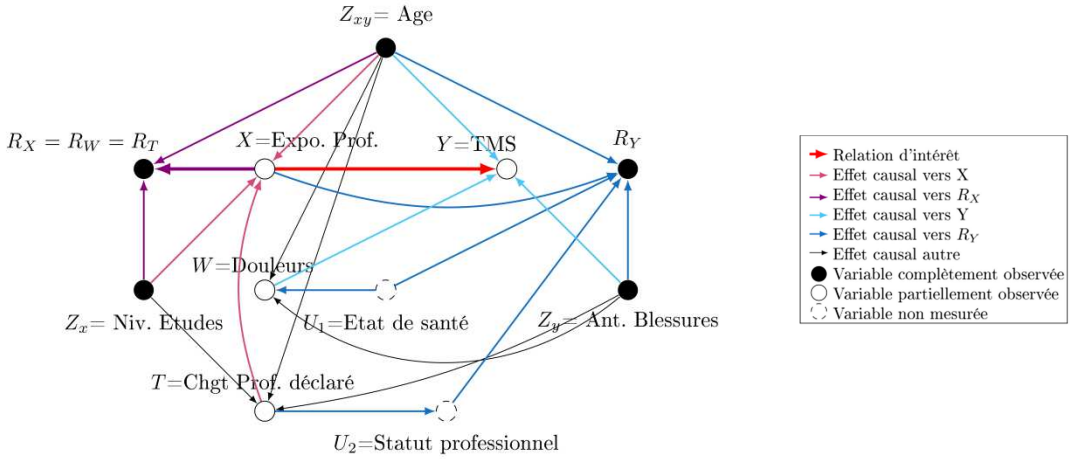
Il n'est pas nécessaire de prendre en compte les changements professionnels T puisqu'on conditionne les analyses sur les expositions professionnelles X.

Remarque : ces calculs indiquent la procédure et les ajustements minimaux pour obtenir un résultat non biaisé. Cependant, selon les méthodes d'analyses choisies, il pourra être nécessaire de prendre en compte des variables supplémentaires, comme les changements professionnels T.

Exemple A3 :

On ajoute aux hypothèses des exemples précédents (A1 et A2) une relation causale entre le fait d'être manquant pour les expositions professionnelles R_X et la valeur des expositions professionnelles X (flèche rouge, Figure 43).

Figure 43 - DAG avec un mécanisme MNAR recoverable (exemple A3)



Le mécanisme R_X est MNAR puisqu'il dépend de X et le mécanisme R_Y est MNAR puisqu'il dépend de W et T. Le mécanisme global de données manquantes n'est pas MAR séquentiel puisque que quelle que soit la séquence choisie R_X sera dépendant de X.

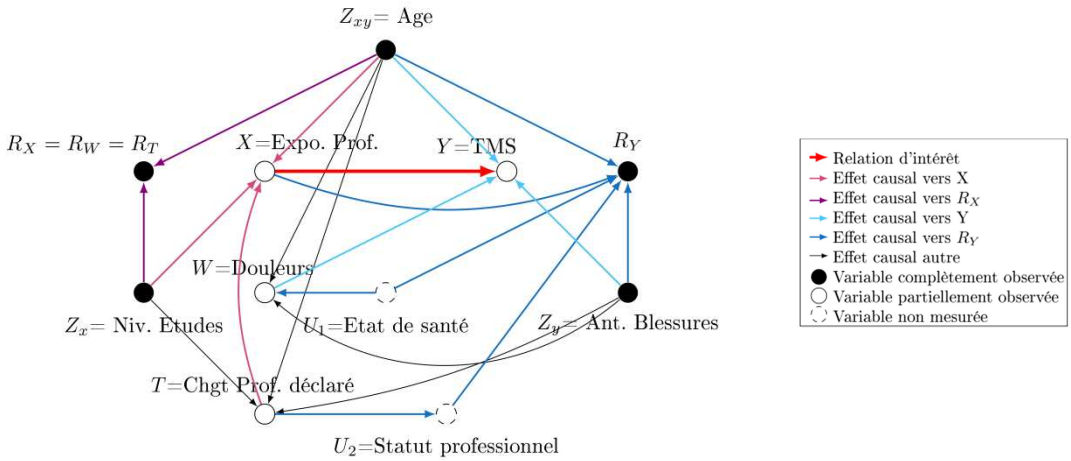
Pourtant on est dans un cas recoverable pour la relation d'intérêt. En effet, les relations d'indépendance précédentes ($W \coprod R_X | Z_{xy}$; $X \coprod R_X | Z_{xy}, Z_x$; $T \coprod R_X | Z_x, Z_{xy}$; $Y \coprod R_X | X, Z_x, Z_{xy}$ et $Y \coprod R_Y | Z_y, Z_{xy}, W$) restent vérifiées, on peut donc écrire la relation d'intérêt à partir des données observées.

Le scénario A2 correspond aux types d'hypothèses qui seront effectuées dans les analyses sur l'incidence des épicondylites (cf. Figure 32, p 99 et Figure 44 de l'Annexe C , p 158). Il présente les difficultés techniques sur les données manquantes même l'étude en question portera sur des variables complètement observées supplémentaires et une relation d'intérêt légèrement différente.

Annexe B Procédure d'imputation pour un exemple de mécanisme MAR séquentiel

On reprend l'exemple A2 de l'Annexe A (cf. Figure 42, p 154) et on présente la procédure pour les imputations multiples avec les *chained equations*. On est alors dans un cas MAR séquentiel avec l'ordre (Y, X, W). On en conclut alors qu'il faut imputer dans l'ordre W, X et Y puisque les premières variables dépendent des suivantes.

Figure 42 (duplicata) - DAG des relations causales



Dans la décomposition de *recoverability* de $P(Y = 1|X = x)$ (cf. Annexe A , p 152), on a vu que la probabilité d'avoir des douleurs $P(W = w)$ en prenant en compte Z_{xy} (les antécédents de blessures) est estimée sans biais à partir des données disponibles (c'est-à-dire si $R_X = 0$). Ainsi, on sait qu'on imputer correctement les non suivis par les suivis pour les douleurs déclarées W en ajustant sur Z_{xy}.

De même, $P(Y = 1|X = x, W = w)$ est estimable sans biais en prenant en compte Z_y, Z_{xy} sur les données disponibles (c'est-à-dire si $R_X = 0$ et $R_Y = 0$). Donc, on peut imputer correctement les TMS Y en ajustant sur Z_y, Z_{xy} et sur X et W imputés.

Ainsi, on estime dans l'ordre les modèles suivants à l'étape d'itération k :

$logit\ P\left(W = 1\middle|Y_{(k-1)}, Z_{xy}, Z_y, R_Y\right) = \beta_{11} + \beta_{12}Y_{(k-1)} + \beta_{13}Z_{xy} + \beta_{14}Z_y + \beta_{15}R_Y \rightarrow$ imputation $W_{(k)}$

$logit\ P\left(X = 1\middle|Y_{(k-1)}, Z_{xy}, Z_x, R_Y\right) = \beta_{11} + \beta_{12}Y_{(k-1)} + \beta_{13}Z_{xy} + \beta_{14}Z_x + \beta_{15}R_Y \rightarrow$ imputation $X_{(k)}$

$logit\ P\left(Y = 1\middle|X_{(k)}, W_{(k)}, Z_{xy}, Z_y\right) = \beta_{21} + \beta_{22}X_{(k)} + \beta_{23}W_{(k)} + \beta_{24}Z_{xy} + \beta_{25}Z_y \rightarrow$ imputation $Y_{(k)}$

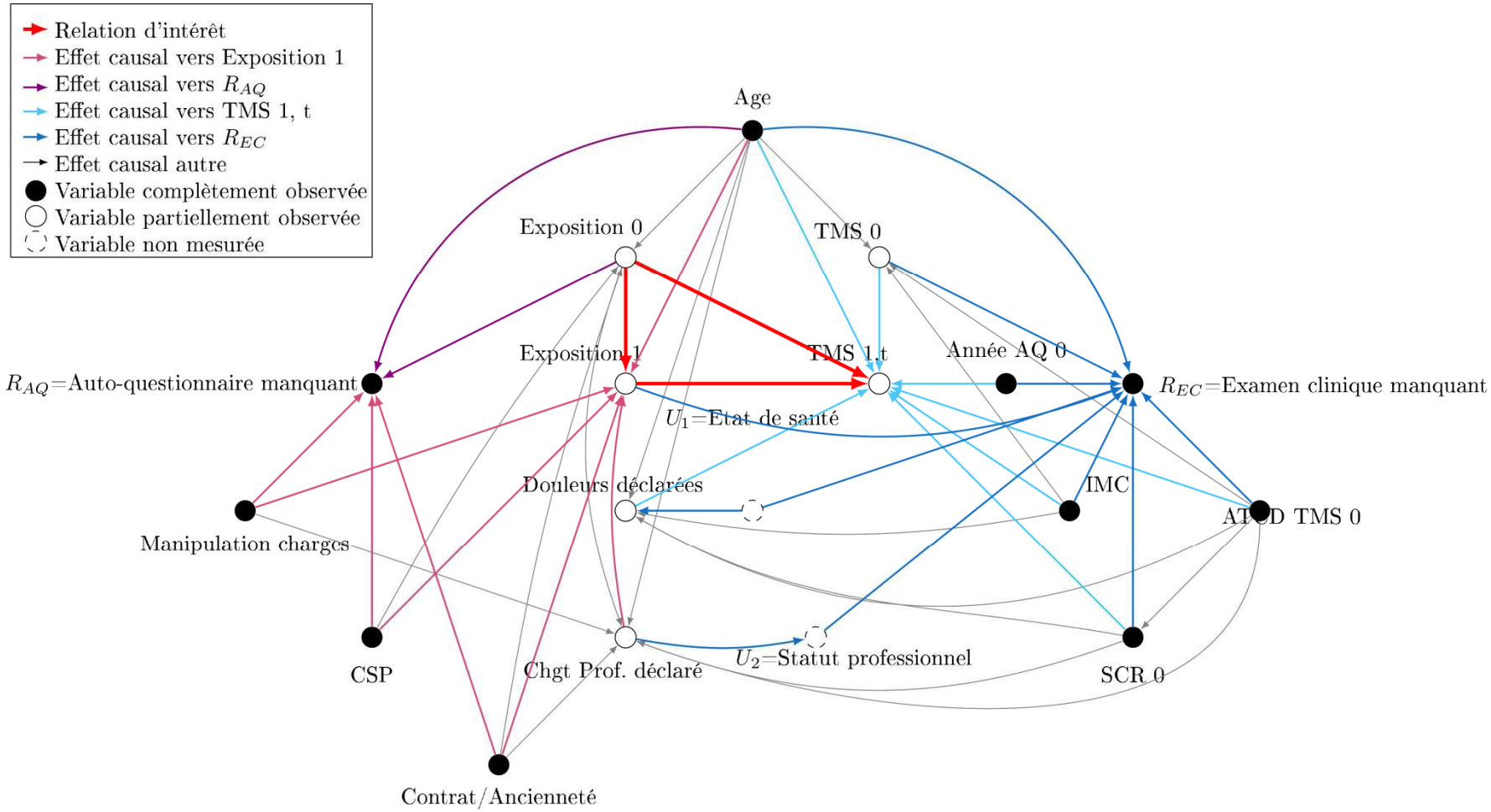
On itère ce cycle pour obtenir les $X(1), \dots X(M),\ W(1), \dots, W(M)\ et\ Y(1), \dots, Y(M)$.

Ainsi, cette procédure donnera une estimation non biaisée de la relation d'intérêt puisqu'on fait intervenir des distributions dont on sait (par le DAG) qu'elles sont représentatives des données observées et non observées.

On peut généraliser cette propriété à tous les mécanismes MAR séquentiel : on montre les estimations par imputations multiples avec équations chainées sont sans biais, à condition d'imputer les variables partiellement observées dans un ordre adéquat (cf. paragraphe 2.3.4.2.3.2.2, p 63).

Annexe C Schéma causal complet pour l'analyse des épicondylites incidentes

Figure 44 - Schéma causal complet pour l'incidence des épicondylites et les données manquantes associées



Annexe D Analyses de sensibilité sur l'incidence des épicondylites – détails techniques

Dans cette annexe, les aspects techniques des analyses de sensibilité pour le mécanisme de données manquantes pour l'examen clinique de suivi sont détaillés.

Afin de faciliter la compréhension dans cette partie, on pose les notations suivantes :

- Y_1 la présence d'épicondylites à t_1 et R_Y l'indicatrice des données manquantes correspondante
- X_0 les facteurs professionnels à t_0 complètement observées
- A l'âge à t_0 en 2 classes
- W_1 les douleurs déclarées à $t_1$et X_1 les expositions professionnelles à t_1
- C la cause d'absence d'examen clinique qui prend les valeurs "$Pas\ RDV$" pour l'absence de rendez-vous sur la période, "$Chgt$" pour les changements de situation ou "Inc" pour une cause inconnue.

Dans l'hypothèse MAR séquentiel pour R_Y, on suppose que la distribution de $Y_1|A, X_0, X_1, W_1$ est identique parmi les suivis et les non suivis (équivalent à $Y_1 \coprod R_Y\ |A, X_0, X_1, W_1$) :

$$P(Y_1 = 1|R_Y = 1, A, X_0, X_1, W_1) = P(Y_1 = 1|R_Y = 0, A, X_0, X_1, W_1)$$

Dans l'hypothèse MNAR, on va supposer que cette prévalence conditionnelle est différente pour les suivis et les non suivis et plus particulièrement différente selon les causes de non suivi pour l'examen clinique et l'âge. Les justifications conceptuelles se trouvent dans le paragraphe 4.2.2.3 (p 102).

Dans les premiers scénarios d'analyse, on suppose que les prévalences dépendent des causes de non suivi pour l'examen clinique. Dans les deuxièmes scénarios d'analyse, on suppose qu'en plus, l'effet de l'âge dépend en plus des causes de non suivi pour l'examen clinique.

Premiers scénarios d'analyses

Comme décrit dans le paragraphe 2.3.5.2 (p 67), on va définir un écart sur les prévalences de Y_1 entre les suivis ($R_Y = 0$) et les non suivis selon leur cause de données manquantes (C). On exprime ces écarts en termes de différences de prévalence pour plus de lisibilité.

On note δ_c les écarts fixes introduits pour faire varier les prévalences de Y_1 selon C :

$$\forall\, c \in \{"Pas\ RDV", "Inc", "Chgt"\}\ \delta_c = P(Y_1 = 1|A, X_0, X_1, W_1, C = c) - P(Y_1 = 1|R_Y = 0, A, X_0, X_1, W_1)$$

Les hypothèses présentées dans le texte sur les prévalences des TMS Y_1 selon les causes d'absence de suivi C correspondent à l'équation suivante :

$$P(Y_1 = 1|A, X_0, X_1, W_1, C = "Pas\ RDV") \leq P(Y_1 = 1|R_Y = 0, A, X_0, X_1, W_1) \leq P(Y_1 = 1|A, X_0, X_1, W_1, C = "Inc") \leq P(Y_1 = 1|A, X_0, X_1, W_1, C = "Chgt")$$

Ce qui se traduit en terme de différences de prévalence par $\delta_{Pas\ RDV} \leq 0 \leq \delta_{Inc} \leq \delta_{Chgt}$.

Pour avoir une idée de l'ordre de grandeur de ce qui serait intéressant d'utiliser pour les valeurs des δ_c, on regarde les ordres de grandeur des probabilités prédites d'avoir un TMS en fonction de la cause d'absence de suivi pour l'examen clinique basés sur les imputations MAR (Tableau 34). On voyait un écart interquartile d'environs 0,05 et une amplitude d'environs 0,8. Dans les modèles d'imputation, on a choisi de faire varier les δ_c entre -0,2 et 0,2, afin d'avoir une modification suffisante des probabilités obtenues tout en évitant d'avoir des prévalences globales trop extrêmes. Les probabilités ainsi modifiées ont été tronquées entre 0 et 1.

Tableau 34- Répartition des probabilités d'avoir un TMS en fonction de la cause d'absence d'examen clinique à partir des données imputées MAR

Suivi et cause de non suivi (R, C)	Minimum	Quartile inférieur	Médiane	Quartile supérieur	Maximum	Ecart interquartile
Pas de RDV	0,001	0,020	0,034	0,073	0,802	0,053
Inconnu	0,001	0,019	0,031	0,065	0,805	0,046
Changement de situation	0,001	0,025	0,044	0,091	0,581	0,066
Suivi	0,002	0,022	0,037	0,081	0,731	0,059

Quatre scénarios ont été choisis pour tester la robustesse des résultats :

- Scénario bas (les prévalences les plus basses possibles) :

$$\delta_{Pas\ RDV} = -0{,}2\ \ \delta_{Inc} = 0{,}0\ \ \delta_{Chgt} = 0{,}0$$

- Scénario plausible :

$$\delta_{Pas\ RDV} = -0{,}1\ \ \delta_{Inc} = 0{,}05\ \ \delta_{Chgt} = 0{,}1$$

- Scénario haut (les prévalences les plus hautes possibles) :

$$\delta_{Pas\ RDV} = 0{,}0\ \ \delta_{Inc} = 0{,}1\ \ \delta_{Chgt} = 0{,}1$$

- Scénario extrême (les prévalences les plus éloignées des prévalences initiales) :

$$\delta_{Pas\ RDV} = -0{,}2 \;\; \delta_{Inc} = 0{,}2 \;\; \delta_{Chgt} = 0{,}2$$

Les prévalences résultantes se trouvent dans le paragraphe 4.2.2.3.1 (p 102) et les résultats des analyses sous ces hypothèses MNAR sont dans le paragraphe 4.2.3.5.1 (p 110).

Deuxièmes scénarios d'analyse

Dans cette deuxième série de scénarios, on suppose qu'en plus de la prévalence des épicondylites, l'effet de l'âge sur les épicondylites varie en fonction de la cause d'absence d'examen clinique. On exprime à nouveau ces effets en termes de différences de prévalence δ_a^c :

$$\begin{aligned}\delta_c^a &= (Effet\ de\ A = a\ sur\ Y_1\ si\ C = c) - (Effet\ de\ A = a\ sur\ Y_1\ si\ R = 0)\\ &= \big(P(Y_1 = 1|C = c, X_0, X_1, W_1, A = a) - P(Y_1 = 1|C = c, X_0, X_1, W_1, A = ref)\big)\\ &\qquad - \big(P(Y_1 = 1|R = 0, X_0, X_1, W_1, A = a) - P(Y_1 = 1|R = 0, X_0, X_1, W_1, A = ref)\big)\end{aligned}$$

D'après les hypothèses formulées dans le paragraphe 4.2.2.3.2 (p 104), on a, pour chaque classe d'âge a (30-44 ans et ≥45 ans), $\delta_{Pas\ RDV}^a \leq \delta_{Inc}^a \leq \delta_{Chgt}^a$.

De façon similaire à ce qui a été présenté pour les premiers scénarios, on a calculé les ordres de grandeur de l'effet de l'âge pour chaque cause de suivi manquant (Tableau 35). On voit que l'effet de l'âge variait entre 0,019 et 0,024 selon les catégories d'âge et la cause. On a donc fait varier les effets de l'âge δ_c^a entre -0,1 et 0,1.

Tableau 35 - Répartition des différences de probabilités d'avoir un TMS au suivi par classe d'âge en fonction de la cause d'absence de suivi ($P(Y_1 = 1|C, X_0, X_1, W_1, Age = a) - P(Y_1 = 1|C, X_0, X_1, W_1, Age = ref)$) basée sur les données imputées MAR

Suivi et cause de non suivi (R, C)	Age (a)	Minimum	Quartile inférieur	Médiane	Quartile supérieur	Maximum
Pas de RDV	30-44 vs <30	0,002	0,010	0,019	0,033	0,363
	≥45 vs <30	0,003	0,014	0,024	0,048	0,218
Inconnu	30-44 vs <30	0,002	0,009	0,019	0,033	0,352
	≥45 vs <30	0,004	0,014	0,021	0,044	0,232
Changement de situation	30-44 vs <30	0,002	0,014	0,023	0,043	0,331
	≥45 vs <30	0,004	0,015	0,024	0,048	0,217
Non manquant	30-44 vs <30	0,002	0,010	0,019	0,032	0,358
	≥45 vs <30	0,003	0,014	0,024	0,049	0,244

On a fait varier les effets de l'âge et les prévalences initiales de cette façon :

- Scénario bas : un effet de l'âge affaibli et une prévalence globale faible pour les sans RDV:

$$\delta_{Pas\ RDV} = -0{,}20 \;\; \delta_{Inc} = 0{,}0 \;\; \delta_{Chgt} = 0{,}0$$

$$\delta_{Pas\ RDV}^1 = -0{,}05 \;\; \delta_{Inc}^1 = 0{,}0 \;\; \delta_{Chgt}^1 = 0{,}0$$

$$\delta_{Pas\ RDV}^2 = -0{,}10 \;\; \delta_{Inc}^2 = 0{,}0 \;\; \delta_{Chgt}^2 = 0{,}0$$

- Scénario plausible : un effet de l'âge plausible et une prévalence globale plausible :

$$\delta_{Pas\ RDV} = -0{,}10 \;\; \delta_{Inc} = 0{,}01 \;\; \delta_{Chgt} = 0{,}02$$

$$\delta^1_{Pas\ RDV} = -0{,}05 \;\; \delta^1_{Inc} = 0{,}02 \;\; \delta^1_{Chgt} = 0{,}05$$

$$\delta^2_{Pas\ RDV} = -0{,}10 \;\; \delta^2_{Inc} = 0{,}05 \;\; \delta^2_{Chgt} = 0{,}10$$

- Scénario haut : un effet de l'âge fort et une prévalence globale forte :

$$\delta_{Pas\ RDV} = 0{,}00 \;\; \delta_{Inc} = 0{,}02 \;\; \delta_{Chgt} = 0{,}05$$

$$\delta^1_{Pas\ RDV} = -0{,}05 \;\; \delta^1_{Inc} = 0{,}10 \;\; \delta^1_{Chgt} = 0{,}15$$

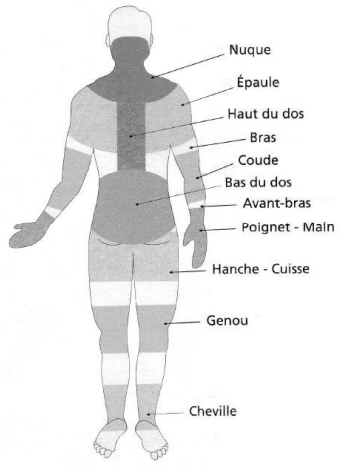
$$\delta^2_{Pas\ RDV} = -0{,}10 \;\; \delta^2_{Inc} = 0{,}10 \;\; \delta^2_{Chgt} = 0{,}15$$

Les prévalences résultantes se trouvent dans le paragraphe 4.2.2.3.2 (p 104) et les résultats des analyses sous ces hypothèses MNAR sont dans le paragraphe 4.2.3.5.2 (p 111).

Annexe E Questionnaire Nordique

Voici un extrait des questions posées sur les douleurs dans l'étude COSALI:

8. Avez-vous eu, au cours des **12 derniers mois**, des problèmes (courbatures, douleurs, gêne, engourdissement) au niveau des zones du corps suivantes ? *Pour chacune des zones du corps, cochez la case correspondant*



	Oui	Non		Du côté droit	Du côté gauche	Des deux côtés
Nuque / cou……	☐	☐				
Epaule / bras…….	☐	☐	Si oui :	☐	☐	☐
Coude/ avant-bras	☐	☐	Si oui :	☐	☐	☐
Main / poignet.....	☐	☐	Si oui :	☐	☐	☐
Doigts.....	☐	☐	Si oui :	☐	☐	☐
Haut du dos……..	☐	☐				
Bas du dos……	☐	☐				
Hanche / cuisse...	☐	☐	Si oui :	☐	☐	☐
Genou / jambe…..	☐	☐	Si oui :	☐	☐	☐
Cheville / pied……	☐	☐	Si oui :	☐	☐	☐

Si vous avez répondu NON à toutes ces questions, passez directement à la question 12

9. Durant ces **12 derniers mois**, combien de temps, au total, avez-vous souffert ? *Pour chacune des zones du corps où vous avez souffert, cochez la case qui correspond le mieux à votre réponse*

	Moins de 24 heures	1 à 7 jours	8 à 30 jours	Plus de 30 jours	En permanence
Nuque / cou……..	☐	☐	☐	☐	☐
Epaule / bras…….	☐	☐	☐	☐	☐
Coude/ avant-bras…	☐	☐	☐	☐	☐
Main / poignet……	☐	☐	☐	☐	☐
Doigts……	☐	☐	☐	☐	☐
Haut du dos……..	☐	☐	☐	☐	☐
Bas du dos……..	☐	☐	☐	☐	☐
Hanche / cuisse……	☐	☐	☐	☐	☐
Genou / jambe……..	☐	☐	☐	☐	☐
Cheville / pied……	☐	☐	☐	☐	☐