Santé publique France

SANTÉ TRAVAIL

MARS 2025

MÉTHODES ET REPÈRES

SURVEILLANCE ÉPIDÉMIOLOGIQUE DE L'ÉTAT DE SANTÉ DES TRAVAILLEURS EN FRANCE SELON L'ACTIVITÉ PROFESSIONNELLE - SEESTA

Protocole à partir de l'EDP et de l'EDP-Santé

Résumé

Surveillance épidémiologique de l'état de santé des travailleurs en France selon l'activité professionnelle - SEESTA

Protocole à partir de l'EDP et de l'EDP-Santé

Une approche classique de la surveillance épidémiologique des populations au travail consiste à analyser systématiquement la fréquence de survenue de maladies et des causes médicales de décès en fonction de l'activité professionnelle.

Le secteur d'activité représente un niveau pertinent pour cette surveillance permettant à la fois d'approcher indirectement certaines spécificités professionnelles en matière de conditions de travail, et l'appropriation aisée des données épidémiologiques par les acteurs de la prévention des risques professionnels. Par ailleurs, à côté des déterminants professionnels, la description de l'état de santé par secteur d'activité peut contribuer à identifier des groupes professionnels particuliers où des actions de prévention spécifiques seraient à mener vis-à-vis de comportements de santé non professionnels.

Les bases de données de l'échantillon démographique permanent (EDP) et de l'EDP-Santé (version enrichie des données de l'assurance maladie et des données d'hospitalisation) constituent de volumineux échantillons longitudinaux représentatifs de la population, comportant à la fois des trajectoires individuelles socioprofessionnelles et des informations relatives à la santé. Ces sources de données présentent ainsi les qualités nécessaires pour répondre aux objectifs de description et de surveillance de la santé en lien avec l'activité professionnelle.

Ce document décrit les sources de données utilisées et les méthodes d'analyse prévues pour :

- Identifier si certains secteurs d'activité sont caractérisés par des risques d'événements de santé spécifiques (en termes d'espérance de vie, de fréquence de pathologies et d'évolution dans le temps) ;
- Identifier si des typologies de carrières professionnelles sont caractérisées par des risques d'événements de santé spécifiques ;
- Évaluer le rôle des secteurs d'activité dans les inégalités de santé observées entre groupes de travailleurs particuliers, par exemple les travailleurs séniors (≥ 55 ans).

De plus, le document présente le rôle des différentes institutions impliquées dans cette étude, les démarches réglementaires réalisées et les modalités de respect et d'application des droits des personnes quant à leurs données. Le document décrit également les limites scientifiques attendues et un calendrier prévisionnel de travail.

MOTS-CLÉS : SURVEILLANCE ÉPIDÉMIOLOGIQUE, SECTEURS D'ACTIVITÉ, TRAVAILLEURS, PROTOCOLE, ÉCHANTILLON DÉMOGRAPHIQUE PERMANENT, SNDS

Citation suggérée : Geoffroy-Perez B, Moisan F. Surveillance épidémiologique de l'état de santé des travailleurs en France selon l'activité professionnelle - SEESTA. Protocole à partir de l'EDP et de l'EDP-Santé. Saint-Maurice : Santé publique France, 2025. 40 p. Disponible à partir de l'URL : https://www.santepubliquefrance.fr

ISSN : 2647-4816 ; ISBN-NET : 979-10-289-0987-1 ; RÉALISÉ PAR LA DIRECTION DE LA COMMUNICATION, SANTÉ PUBLIQUE FRANCE - DÉPÔT LÉGAL : MARS 2025

Abstract

Epidemiological Surveillance of Workers' Health Status in France by Occupational Activity - SEESTA

Protocol based on the EDP and the EDP-Santé samples

A classic approach to epidemiological surveillance of working populations involves systematically analyzing the frequency of disease occurrence and the medical causes of death according to occupational activity.

The economic sector represents a relevant level for this surveillance, allowing both an indirect approach to certain occupational specificities in terms of working conditions and the easy appropriation of epidemiological data by those involved in occupational risk prevention.
Furthermore, alongside occupational determinants, describing health status by economic sector can help identify particular professional groups where specific preventive actions are needed regarding non- occupational health behaviors.

The permanent demographic sample (EDP) and EDP-Santé databases (an enriched version with health insurance and hospitalization data) constitute large, representative longitudinal samples of the population, containing both individual socio-professional trajectories and health-related information.
These data sources thus have the necessary qualities to meet the objectives of describing and monitoring health in relation to occupational activity.

This document describes the data sources used and the analysis methods planned with the aim of:
• Identifying whether certain economic sectors are characterized by specific health risks (in terms of life expectancy, disease frequency, and trends over time);
• Identifying whether certain type of professional careers are characterized by specific health risks;
• Evaluating the role of economic sectors in the health inequalities observed among particular groups of workers, such as senior workers (≥55 years old).

Additionally, the document presents the role of the different institutions involved in this study, the regulatory procedures carried out, and the methods of respecting and applying individuals' rights regarding their data. The document also describes the expected scientific limitations and a provisional work schedule.

KEY WORDS: EPIDEMIOLOGICAL SURVEILLANCE, ECONOMIC SECTORS, WORKERS, PROTOCOL, PERMANENT DEMOGRAPHIC SAMPLE, SNDS

Équipe projet

Responsable de traitement

Santé publique France (Agence nationale de santé publique, 12 rue du Val d'Osne, 94415 Saint-Maurice Cedex), représentée par sa directrice générale Caroline Semaille

Responsable de la mise en œuvre du traitement

Direction Santé Travail Environnement (Dset) de Santé publique France

Déléguée à la protection des données (DPO)

Clothilde Hachin, Santé publique France, dpo@santepubliquefrance.fr

Responsables du projet

Béatrice Geoffroy-Perez et Frédéric Moisan, chargés de projets scientifiques à la Direction Santé Environnement Travail (Dset) de Santé publique France, chargés de la rédaction du protocole, de la définition des objectifs, du pilotage des analyses et de la valorisation des résultats.

Les statisticiens de la Direction Appui, Traitements et Analyse de données (Data) de Santé publique France chargés du traitement des données.

Abréviations

ALD	Affection longue durée
ANSSI	Agence nationale de la sécurité des systèmes d'information
CASD	Centre d'accès sécurisé aux données
CépiDc	Centre d'épidémiologie sur les causes médicales de décès
Cesrees	Comité éthique et scientifique pour les recherches, les études et les évaluations dans le domaine de la santé
Cnam	Caisse nationale de l'assurance maladie
Cnil	Commission nationale de l'informatique et des libertés
CSS	Comité du secret statistique
DADS	Déclaration annuelle des données sociales
Dares	Direction de l'animation de la recherche, des études et des statistiques (ministère chargé du travail)
Dcir	Datamart de consommation interrégimes
DPO	Délégué à la protection des données
Drees	Direction de la recherche, des études, de l'évaluation et des statistiques (ministère chargé de la santé)
EAR	Enquêtes annuelles de recensement
EDP	Échantillon démographique permanent
Ined	Institut national des études démographiques
Insee	Institut national de la statistique et des études économiques
NAF	Nomenclature d'activités française
NIR	Numéro d'inscription au répertoire des personnes physiques
PMSI	Programme de médicalisation des systèmes d'information
RGPD	Règlement général sur la protection des données
SEESTA	Surveillance épidémiologique de l'état de santé des travailleurs en France selon l'activité professionnelle
SIR	*Standardized Incidence Ratio*
SNDS	Système national des données de santé

Sommaire

1. CONTEXTE ET JUSTIFICATION DE L'ÉTUDE

1.1 Importance des problèmes de santé d'origine professionnelle

Il est largement établi que les facteurs professionnels sont une source majeure d'inégalité sociale en matière de santé et pèsent largement sur la santé de la population. Ainsi, on estime qu'environ le tiers des différences sociales de mortalité par cancer dans les pays industrialisés (différences qui sont très fortes, en Europe et en France en particulier [1]), est expliqué par l'exposition à des facteurs d'origine professionnelle, et que cette fraction s'élève chez les hommes jusqu'à 40 % pour les cancers de la cavité nasale et 20 % pour le cancer du poumon [2]. À côté des cancers, qui font l'objet de recherches nombreuses, il existe de très importants problèmes de santé qui ont tout ou partie de leur origine dans l'environnement professionnel [3, 4] : troubles musculo-squelettiques, troubles de l'audition (liés au bruit d'origine professionnelle), pathologies respiratoires (asthme), maladies cardiovasculaires (infarctus du myocarde, cardiopathie), chutes, accidents de la route, etc. À côté des nuisances de nature physico-chimique et biologique, on connaît aujourd'hui l'influence considérable des facteurs psychosociaux associés à l'organisation du travail, au stress au travail, à la perte de l'emploi, au temps de travail, dont les conséquences pour la santé concernent aussi bien la sphère somatique (maladies cardiovasculaires, par exemple) que mentale (dépression, épuisement, suicide, par exemple) [5].

1.2 Surveillance de la santé selon l'activité professionnelle

Une approche classique de la surveillance épidémiologique des risques professionnels consiste à analyser systématiquement la fréquence de survenue de maladies et les causes médicales de décès en fonction de l'activité professionnelle.

L'appariement de données longitudinales, à l'échelle de la population nationale, relatives aux facteurs professionnels et aux événements de santé, issues de sources de données collectées de manière systématique et indépendante présente de nombreux atouts dans le domaine de la surveillance des risques professionnels, notamment en termes de représentativité et de puissance statistique. Le caractère longitudinal permet en outre déterminer la temporalité dans la survenue des événements de santé et des changements professionnels, de prendre en compte l'ensemble de la carrière des personnes dans l'analyse des risques pour la santé, et d'étudier des problèmes de santé dont la survenue est parfois très longtemps différée par rapport à l'exposition, ce qui est souvent le cas des cancers.

Après des premiers travaux commencés en 2002 sur la mortalité par causes (programme Cosmop [11]), l'accès plus récent aux données de l'assurance maladie et aux données d'hospitalisation, disponibles pour l'ensemble de la population française, et la possibilité de les apparier à des données professionnelles permettent désormais d'étudier, outre la mortalité par cause, la fréquence de survenue des maladies selon le type d'activité professionnelle. Le projet de Surveillance épidémiologique de l'état de Santé des travailleurs en France selon l'activité professionnelle, baptisé SEESTA, s'inscrit dans ce cadre et dans la continuité des premiers travaux sur la mortalité, englobant ainsi le projet Cosmop.

1.3 Utilité et pertinence d'une description épidémiologique par secteur d'activité

Un secteur d'activité est un regroupement d'entreprises/établissements de fabrication, de commerce ou de service qui ont la même activité principale. Chaque établissement est affecté à un unique secteur par l'Institut national de la statistique et des études économiques (Insee) selon la

nomenclature d'activités française (NAF) [6]. La finalité de la NAF est essentiellement statistique. Elle facilite l'organisation de l'information économique et sociale et permet une description exhaustive, et sans double compte ou recoupement, de l'ensemble de l'activité économique productive en France. De nombreuses productions de l'Insee sont diffusées au niveau du secteur d'activité (plus ou moins agrégé).

D'un point de vue scientifique, la description épidémiologique au niveau du secteur d'activité permet de regrouper des travailleurs ayant plus vraisemblablement des conditions de travail et des expositions professionnelles similaires car exerçant dans des structures avec la même activité principale. Cette approche, utilisée dès les années quatre-vingt, est classique dans les études épidémiologiques en santé au travail [7]. Elle permet d'approcher indirectement certaines spécificités professionnelles. L'utilisation du secteur d'activité par d'autres acteurs de la statistique publique ou de la protection sociale (Insee, Direction de l'animation de la recherche, des études et des statistiques (Dares) de la Direction générale du travail, Caisse nationale de l'assurance maladie (Cnam)) permet également d'établir des liens entre les résultats épidémiologiques produits et d'autres données contextuelles.

De plus, l'approche par secteur d'activité est pertinente pour permettre l'appropriation des données épidémiologiques par les acteurs de la prévention des risques professionnels. En effet, un employeur connaît le secteur d'activité auquel il appartient. Par ailleurs, c'est sur la base du secteur d'activité que les structures employant des salariés dans le secteur privé se regroupent pour établir des conventions collectives. Ces dernières sont des accords signés et négociés entre les organisations syndicales représentatives et des employeurs (ou groupements d'employeurs) définissant les conditions de travail, les salaires, les congés payés, les droits et obligations des employeurs et des salariés ainsi que les adaptations des règles du Code du travail aux situations particulières des secteurs concernés [8].

Enfin, en considérant comme le propose Dejours [9], que la santé n'est pas un état naturel mais une construction intentionnelle par rapport à des normes et des valeurs, on comprend que des dynamiques collectives dans l'environnement professionnel, notamment via la socialisation professionnelle, peuvent venir moduler, dans un sens ou dans un autre, les comportements de santé ou le rapport au système de soins pour des groupes de travailleurs appartement à un même secteur d'activité. Des différences de consommation de tabac et de l'alcool entre les secteurs d'activité sont d'ailleurs observées [10] et, comme indiqué précédemment, il est admis que le travail est un des déterminants sociaux des inégalités sociales de santé [1]. Ainsi, en plus des déterminants professionnels, la description de l'état de santé par secteur d'activité contribue à identifier des groupes de travailleurs particuliers permettant de cibler des actions de prévention et de promotion de la santé en milieu de travail.

1.4 Justification du respect de l'éthique et de l'intérêt public

Parmi ses missions, Santé publique France a notamment l'amélioration de la connaissance de l'état de santé de la population, des comportements et des risques pour la santé. Pour cela, Santé publique France développe des activités de surveillance et des enquêtes en population. Ces activités lui permettent d'apprécier l'état de santé de la population et de formuler des recommandations en fonction des caractéristiques des populations et des priorités qui se dessinent. La surveillance épidémiologique des risques professionnels entre dans le cadre de cette surveillance sanitaire. L'identification de catégories de personnes plus à risque de décès pour certaines causes a pour objectif de contribuer à guider l'action des pouvoirs publics et prioriser les actions de prévention et de promotion de la santé dans les milieux professionnels identifiés à risque particulier.

Dans le cadre du présent dossier, les données seront traitées globalement sans intention de réidentifier un individu particulier. À ce titre, le traitement des données dans le Centre d'accès sécurisé aux données (CASD) empêche toute sortie de résultats contrevenant au secret statistique.

Les finalités et les modalités de mise en œuvre de l'étude, qui reposent uniquement sur l'analyse de données existantes issues de bases de données médico-administratives, ne correspondent pas à celles d'une recherche impliquant la personne humaine et ne relèvent donc pas de l'avis d'un comité de protection des personnes.

1.5 Publication des résultats et valorisations attendues

Tous les résultats produits dans le cadre de ce projet seront rendus publics via la mise en ligne de rapports d'études et la communication lors de colloques. Des restitutions pourront aussi être réalisées auprès des instances paritaires regroupant organisations syndicales représentatives et employeurs ou groupements d'employeurs. Des articles scientifiques portant sur les principales thématiques de l'étude seront rédigés et soumis à des revues à comité de lecture. Du fait de la quantité d'indicateurs produits, il est envisagé que les résultats soient publiés et consultables à travers un outil en ligne accessible via le site Internet de Santé publique France permettant de filtrer les résultats selon des critères de type d'indicateur, de genre, de secteur ou de cause.

2. OBJECTIFS DE L'ÉTUDE

Les objectifs principaux de l'exploitation, objet du présent protocole, sont les suivants :

- Identifier si certains secteurs d'activité sont caractérisés par des risques d'événements de santé spécifiques
  - Décrire l'espérance de vie selon le secteur d'activité le plus longtemps exercé et la comparer à la valeur nationale
  - Décrire l'incidence de plusieurs événements de santé (liste détaillée après) et les comparer aux valeurs nationales
    - Selon le fait d'avoir travaillé dans un secteur donné
    - Selon le secteur d'activité le plus longtemps exercé
  - Décrire, selon le secteur d'activité le plus longtemps exercé, l'évolution temporelle de l'incidence de plusieurs événements de santé et la comparer aux évolutions nationales
  - Décrire la survenue de plusieurs événements de santé à cinq ans et dix ans après avoir travaillé dans un secteur d'activité et la comparer à des personnes ayant travaillé dans un autre secteur
- Identifier si des typologies de carrières professionnelles sont caractérisées par des risques d'événement de santé spécifiques
  - Définir des typologies de carrières (successions de secteurs d'activité)
  - Décrire l'incidence de plusieurs événements de santé selon la typologie de carrière et les comparer aux valeurs nationales
- Étudier les liens entre les secteurs d'activité occupés et les inégalités de l'état de santé observées dans des groupes de travailleurs spécifiques
  - Définir a priori des groupes de travailleurs spécifiques (par exemple, travailleurs séniors, ≥ 55 ans)
  - Décrire la part que représentent les secteurs d'activité exercés pour décrire la survenue de plusieurs événements de santé parmi ces travailleurs spécifiques.

Les résultats obtenus et l'évolution dans le temps des indicateurs épidémiologiques doivent contribuer à repérer et surveiller des situations à risque, à alerter sur l'apparition de nouveaux facteurs de risque potentiels d'origine professionnelle et/ou partagés par des groupes professionnels afin de prioriser les actions de prévention et de promotion de la santé. À plus long terme, cette surveillance suppose de reproduire les analyses à intervalles de temps réguliers sur des sources actualisées.

Ponctuellement, la base d'analyse constituée pourra également être mobilisée afin de documenter, à l'échelle du secteur d'activité, des observations et signalements de terrain, à l'échelle d'une entreprise par exemple, laissant suspecter des problèmes de santé en fréquence anormale au sein d'un groupe professionnel donné.

3. MÉTHODOLOGIE

3.1 Conception générale

Pour répondre aux objectifs de description et de surveillance de la santé en lien avec l'activité professionnelle, il est nécessaire de disposer d'informations socioprofessionnelles et de données de santé sur de larges échantillons représentatifs de la population cible de l'étude, ici constituée de l'ensemble de la population active en France.

Les informations socioprofessionnelles doivent idéalement décrire les trajectoires socio-professionnelles dans leur intégralité.

La représentativité des sources de données à l'échelon national est également un atout essentiel afin de limiter les biais dans les estimations. Enfin, la taille des échantillons doit être suffisante pour envisager la surveillance de facteurs professionnels ou d'événements de santé peu fréquents.

Or, la constitution ad hoc de telles bases de données est très longue et extrêmement coûteuse. Il est donc important, lorsque cela est possible, de s'appuyer sur des sources de données déjà constituées et qui permettent, parfois au prix d'un appariement, de disposer de bases de données répondant aux besoins d'analyse.

L'échantillon démographique permanent (EDP) géré par l'Insee, ainsi que sa version enrichie par les données du Système national des données de santé (SNDS), baptisée EDP-Santé, gérée par la Direction de la recherche, des études, de l'évaluation et des statistiques (Drees), présentent les principales qualités nécessaires pour répondre aux objectifs de description et de surveillance de la santé de l'étude SEESTA.

Ces bases de données comportent à la fois des données permettant de reconstituer des trajectoires professionnelles et des informations relatives à la santé. De plus, ces données sont collectées de manière systématique et non sur la base du volontariat.

Elles couvrent l'ensemble des classes d'âges à partir de 16 ans, les deux sexes et toutes les situations de travail en France quels que soient le régime de protection sociale, la profession et le secteur d'activité.

Elles constituent des grands échantillons, autorisant l'étude de nombreux facteurs professionnels ou d'événements de santé.

Enfin, déjà constituées et alimentées en routine, elles permettent non seulement de faire l'économie d'un recueil supplémentaire, mais aussi d'envisager à plus longue échéance le suivi dans le temps des indicateurs épidémiologiques produits grâce à la mise à jour régulière des sources de données.

3.2 Description des sources de données

Les analyses prévues reposent donc sur l'exploitation d'une part de l'EDP géré par l'Insee, d'autre part de l'extraction de données du SNDS, produit de l'appariement déterministe de l'EDP avec le SNDS. Cet appariement étant réalisé par la Drees pour ses propres objectifs [12], Santé publique France a souhaité avoir accès à la dernière version de l'EDP-Santé disponible.

L'EDP représente une source de données très complète, combinant les atouts des deux sources de données : EDP (version 2002) et Panel DADS (Déclaration annuelle des données sociales) (version 2006) déjà exploitées par Santé publique France à titre de faisabilité [11, 13]. De par le quadruplement

du champ de l'échantillon en 2004, l'EDP dispose d'une puissance statistique permettant la surveillance épidémiologique de la santé en lien avec les facteurs professionnels.

L'enrichissement de l'EDP, lors de la création de l'EDP-Santé, par les informations de Santé du SNDS, notamment la mortalité et les causes médicales de décès, ainsi que les données de remboursements de l'Assurance maladie, les hospitalisations et les affections de longue durée (ALD) permet d'étudier la mortalité et la morbidité à partir d'une base déjà constituée sans nécessiter d'appariement supplémentaire.

La demande concerne ainsi l'autorisation d'exploiter :

- L'EDP géré par l'Insee en lui-même pour les analyses concernant la mortalité globale ;
- Sa version enrichie par les données du SNDS, baptisé EDP-Santé, déjà constitué et géré par la Drees grâce à un appariement déterministe de l'EDP aux données du SNDS, pour les analyses sur les causes de décès et, pour l'étude de la morbidité.

3.2.1 Échantillon démographique permanent - EDP

L'EDP est un panel sociodémographique géré par l'unité « Études démographiques et sociales » de l'Insee avec pour objectif de constituer une vaste base de données longitudinale représentative des personnes vivant en France. Mise en place au recensement de 1968, la base étude accumule et relie au niveau individuel des données de différentes sources administratives et d'enquête pour des individus sélectionnés sur leur jour de naissance. Le champ de l'EDP, de 4 jours de naissance en 1968, a été élargi à 16 jours en 2004 et l'EDP couvre désormais 4,4 % de la population (3,7 millions d'individus).

Depuis sa constitution, l'EDP a continué de s'enrichir au fil du temps par de nouvelles sources de données. Stable depuis 2011, il résulte actuellement de l'appariement de données des sources suivantes :

- Les données des bulletins d'état civil (naissance, décès, naissance d'enfants, mariage) depuis 1968 ;
- Les données issues des cinq recensements exhaustifs (1968, 1975, 1982, 1990, 1999) puis des enquêtes annuelles de recensement à partir de 2004 ;
- Les données du fichier électoral donnant les inscriptions électorales depuis 1990 ;
- Les informations issues du Panel d'actifs « tous salariés » (1/2 des 4 jours de naissance depuis 1967 puis complet à partir de 2004) ;
- Les données socio-fiscales issues du Fichier démographique des logements et des individus (base Fidéli) et du Fichier localisé social et fiscal (base FiLoSoFi) depuis 2011.

Par ailleurs, les informations professionnelles détaillées issues du Panel DADS et des bulletins d'état civil sont rassemblées dans la base de données « Panel tous salariés » disponible dans sa version appariée à l'EDP. Concernant les activités non salariées, l'Insee dispose également d'un panel non-salarié constitué à partir de 2006 et dont l'intégration dans un prochain millésime de l'EDP a été annoncée. Si disponibles, ces données sont également nécessaires pour la bonne identification des travailleurs non-salariés dans l'échantillon.

3.2.2 EDP-Santé

L'EDP-Santé a été constitué par l'appariement de l'EDP dans sa version élargie avec les données du SNDS (remboursements des soins en ville et en établissements de santé, actes et diagnostics des séjours d'hospitalisation du Programme de médicalisation des systèmes d'information (PMSI) et causes médicales de décès).

Cet appariement à l'échelon individuel a été réalisé une première fois en 2019 par la Drees en collaboration avec l'Insee, grâce au Numéro d'inscription au répertoire des personnes physiques (NIR) conservé dans la base étude de l'Insee [12]. L'autorisation de mise en œuvre comportait un

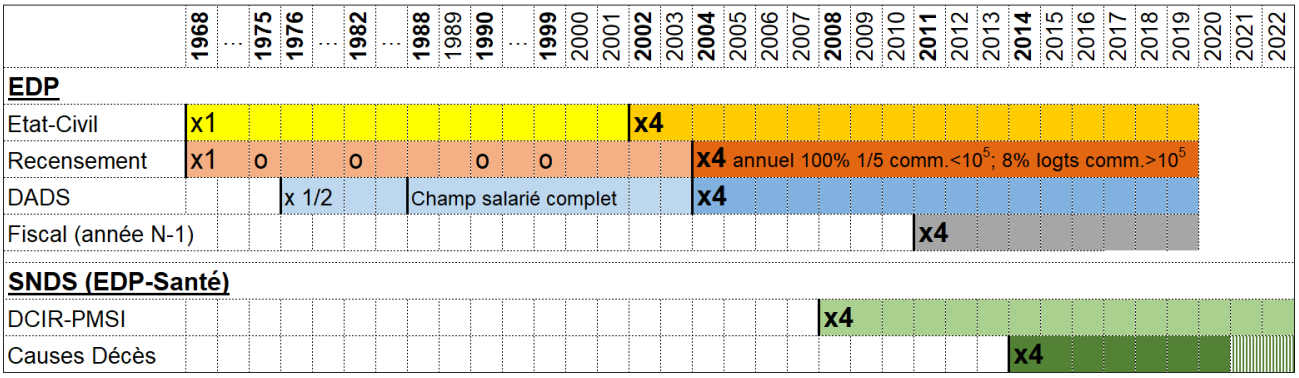
appariement et exploitation pour cinq années par la Drees. Cet échantillon est exploité par la Drees pour la production des indicateurs de la Loi de santé.

L'échantillon constitué, baptisé EDP-Santé, a pour champ les individus du champ EDP encore vivants en 2008 ou nés ultérieurement. Ainsi, 3,3 millions d'individus EDP ont été appariés au SNDS (dont 3,1 millions présents dans les sources fiscales).

Du fait de l'appariement déjà réalisé par la Drees, Santé publique France souhaite exploiter le produit de l'appariement dans la dernière version qui sera disponible à la date de mise à disposition des données.

La profondeur des données disponibles selon la source et la période de collecte est schématisée sur la Figure 1.

Figure 1. Profondeur des données disponibles dans l'EDP et l'EDP-Santé selon la source



x1 : échantillon à 4 jours ; x4 : échantillon à 16 jours.

3.3 Description de la population d'étude

Le ciblage de la population correspond en premier lieu aux individus entrant dans le champ des échantillons de population concernés. Pour l'EDP, ce sont les individus nés un des 4 jours de naissance clés initiaux depuis 1968 ou un des 16 jours de naissance clés depuis 2004. Pour l'EDP-Santé, il s'agit des personnes de l'EDP vivantes en 2008 ou nées après.

3.3.1 Critères de ciblage

Dans l'optique de l'analyse de la mortalité globale selon l'activité professionnelle, la population d'intérêt est représentée par les personnes appartenant à l'EDP : âgées de 0 à 65 ans et vivantes au 1er janvier 2004 (donc âgées de 18 ans ou plus à la fin de la période d'étude) ET ayant exercé une activité professionnelle, salariée ou non, antérieure au 1er janvier 2004 ou au cours de la période d'étude.

Pour l'étude des événements de santé spécifiques, il s'agira des mêmes critères de ciblage, mais avec le critère structurel complémentaire d'être en vie en 2008 (critère lié à la constitution de l'EDP-Santé).

3.3.2 Période de ciblage

Pour l'échantillon démographique permanent et l'étude de la mortalité générale, l'analyse portera sur la période à partir de 2004 et jusqu'à 2022 (dernière année disponible dans la version la plus récente de l'échantillon apparié).

Pour l'EDP-Santé et l'étude des événements de santé plus spécifiques, la période d'étude portera sur la période à partir de 2008 et jusqu'à 2022 (dernière année disponible dans la version la plus récente de l'échantillon apparié).

3.4 Nature des données extraites et période d'extraction

3.4.1 Données sociales et professionnelles

Les données socioprofessionnelles d'intérêt portent sur l'intégralité de l'historique professionnel des individus concernés, rétrospectivement jusqu'en 1968, année du plus ancien recensement exhaustif de population disponible dans l'EDP. Elles permettent de décrire et caractériser le parcours socioprofessionnel des personnes. Ces données proviennent de multiples sources disponibles au sein de l'EDP :

- Les données des recensements exhaustifs depuis 1968 pour l'échantillon à 4 jours ;
- Les données des Enquêtes annuelles de recensement (EAR) collectées en continu pour tout l'échantillon à 16 jours de naissance depuis 2004 ;
- Les données du Panel salarié, dont le champ est considéré comme constant depuis 1992 ; disponibles à partir de 2002 pour l'ensemble de l'échantillon à 16 jours et pour la moitié de l'échantillon à 4 jours avant 2002 (1/8 de l'échantillon à 16 jours). Les informations sur les périodes de chômage indemnisé sont disponibles depuis 2009 dans le Panel tous salariés ;
- Les données fiscales disponibles à partir de 2011. Elles permettent de documenter les épisodes sans activité professionnelle salariée connue.

Les principales informations sociales et professionnelles disponibles qui seront utilisées dans l'analyse sont listées en Annexe 1.

3.4.1.1 Données de recensements et bulletins d'état civil

Différentes données permettant de caractériser la population au travail sont présentes dans les recensements et dans les bulletins d'état civil.

L'origine géographique du travailleur permettra par ailleurs de décrire la population des travailleurs d'origine étrangère.

3.4.1.2 Données du Panel tous salariés

Le Panel « tous salariés » compile des données issues des DADS, des fichiers de paie de l'état et des déclarations des particuliers-employeurs. L'EDP récupère, via le NIR, les informations des individus EDP en agrégeant au niveau annuel et en conservant les caractéristiques du poste principal.

Le champ du Panel DADS et de l'EDP ayant évolué depuis la mise en place des échantillons, les informations issues des DADS sont disponibles pour les individus de l'échantillon initial nés en octobre des années paires de 1967 à 2001 et pour l'ensemble des individus de l'échantillon à 16 jours à partir de 2002.

Les données remontent à 1967 mais portent sur un champ sectoriel et géographique évolutif :

- Le secteur privé depuis 1967 (hors secteur agricole) ;
- La fonction publique à partir de 1988 ;
- Les DOM (départements d'outre-mer) à partir de 2002 ;
- Le secteur agricole (pour les salariés non couverts par le Régime agricole) à partir de 2002 ;
- Les salariés des particuliers-employeurs à partir de 2009.

Les informations détaillées reprenant l'ensemble des informations professionnelles disponibles du Panel DADS et des bulletins d'état civil sont dans la base de données « Panel tous salariés » disponible dans sa version appariée à l'EDP.

L'accès pour exploitation de cette base de données est également demandé.

3.4.1.3 Données sociales et fiscales

Les informations sur les logements proviennent du Fichier démographique des logements et des individus (Fidéli) qui contient lui-même des informations sur l'ensemble des locaux retrouvés dans le fichier des propriétés bâties (FPB) et de la taxe d'habitation.

Les informations concernant les données de revenu sont issues du Fichier localisé social et fiscal (Filosofi) pour le ménage fiscal rattaché au logement. Il s'agit des données fiscales issues de la Direction générale des Finances publiques (déclarations de revenus des personnes physiques, taxe d'habitation et fichier d'imposition des personnes physiques) et des données sur les prestations sociales en provenance de la Caisse nationale des allocations familiales, de la Caisse nationale de l'assurance vieillesse et de la Caisse centrale de la Mutualité sociale agricole.

3.4.2 Données de santé

Les données relatives aux recours aux soins et aux événements de santé d'intérêt portent sur l'ensemble de la période d'étude. Ces données proviennent de multiples sources disponibles au sein du SNDS : le Datamart de consommation interrégimes (Dcir), le PMSI et les causes médicales de décès [14].

Les données de santé du PMSI et du Dcir sont disponibles dans l'EDP-Santé à partir de l'année 2008. L'analyse concernera donc la période à partir de 2008 et jusqu'à 2022 (dernière année disponible dans la version la plus actuelle de l'échantillon apparié).

Les causes médicales de décès sont intégrées dans le SNDS à partir de 2006 mais leur exhaustivité est médiocre en 2008 et avant (moins de 50 %) [15] et ne devient correcte qu'à partir de 2012.

L'étude des profils d'hospitalisation et de la morbidité portera donc sur la période 2008-2022, elle s'appuie principalement sur les données du PMSI. L'étude de la mortalité par cause sera quant à elle possible sur la période à partir de 2012 et jusqu'à 2022.

La notion de décès et sa date de survenue sont disponibles dans l'EDP via les bulletins de décès et l'information issue du Répertoire national d'identification des personnes physiques (RNIPP).

Les autres données de santé sont disponibles dans l'EDP-Santé et sont issues du SNDS qui couvre l'ensemble des personnes ayant eu recours au système de soins français ou étant décédées sur le territoire. Il s'agit :

- Des données de l'Assurance maladie (remboursements de soins en ville et en établissement, motifs d'exonération du ticket modérateur remontés dans le SNDS (Dcir principalement) ;
- Des données hospitalières du PMSI ;
- Des causes médicales de décès du Centre d'épidémiologie sur les causes médicales de décès (CépiDc) de l'Institut national de la santé et de la recherche médicale (Inserm).

Les informations remontées correspondent à l'ensemble des données du SNDS. Elles portent sur les années 2008 à 2022 et sont de quatre ordres :

- Informations sur le bénéficiaire (sexe, mois et année de naissance, rang de naissance, lieu de résidence, régime, couverture maladie universelle complémentaire, aide à la complémentaire santé) ;
- Pathologies, notamment les affections de longue durée ;
- Dépenses et remboursements (prestations en soins de ville, en établissements de santé, et montants associés) ;
    - Consommations de soins de ville (consultations, actes techniques…) ;
    - Dispositifs médicaux (aides techniques) ;
    - Prescriptions (médicaments) ;
    - Autres prestations (cures, transports…) ;

  - Soins hospitaliers (hors séances) ;
  - Séjours hospitaliers dont séances ; diagnostics selon la Classification internationale des maladies (10e révision), actes réalisés ;
  - Indemnités journalières (maladies, accidents du travail et maladies professionnelles, maternité) ;
- Causes médicales de décès.

L'enrichissement de l'EDP par les données du SNDS, mis en œuvre par la Drees, permet de disposer, sur une fenêtre de dix ans, d'informations sur les recours aux soins, des informations concernant les séjours hospitaliers, ainsi que les causes de décès des individus.

Les événements de santé retenus pour cette étude seront ceux :

- qu'il est possible de définir avec qualité à partir des données disponibles du SNDS (par exemple, algorithme défini et validé) et à interpréter d'un point de vue épidémiologique ;
- dont l'incidence est assez élevée pour disposer des effectifs permettant de calculer une incidence dans chaque secteur d'activité avec une précision suffisante (rapport des bornes supérieure et inférieure de l'intervalle de confiance à 95 % inférieur ou égal à 2) ;
- dont la période de latence documentée dans la littérature est compatible entre la période d'observation de l'événement de santé et les périodes de carrières professionnelles ;
- qui sont les plus diverses en termes de nature (accident *vs*. maladie) ou d'appareil atteint (respiratoire, santé mentale, etc.).

Les événements de santé analysés en premier sont :

Le décès toutes causes ;
Le décès par tumeur maligne trachée, bronches, poumon ;
Le décès par tumeur maligne du sein ;
Le décès par tumeur maligne de la prostate ;
Le décès par accident ;
Le décès par suicide ou lésion auto-infligée ;
L'indemnisation d'un arrêt de travail pour maladie ;
La perception d'une pension d'invalidité ;
Une prise en charge médicale pour diabète ;
Le traitement chirurgical d'un syndrome du canal carpien ;
Le traitement chirurgical d'une hernie discale lombaire ; Un accouchement prématuré ;
Une prise en charge médicale pour un syndrome coronaire aigu dont infarctus du myocarde ;
Une prise en charge médicale pour accident vasculaire cérébral.

Plusieurs conditions de travail sont connues pour être associées à la survenue des événements ci-dessus, par exemple : les fumées d'échappement de moteur diesel et le cancer broncho-pulmonaire [16], le travail de nuit et le cancer du sein chez les femmes [17] et aussi les accidents et le diabète [18], les pesticides organochlorés et le cancer de la prostate [19], une charge de travail importante et le suicide [20], les tâches répétitives ou le travail sous contrainte de temps et les troubles musculosquelettiques [21], le nombre d'heures travaillées et l'accouchement prématuré [22] ou l'accident vasculaire cérébral [23], le bruit et le stress au travail et les cardiopathies ischémiques [24].

Santé publique France exploite le produit de l'appariement déjà réalisé par la Drees dans sa dernière version, permettant ainsi d'éviter une nouvelle circulation de données identifiantes, la mise en œuvre d'un appariement particulièrement lourd et complexe ainsi qu'une nouvelle extraction.

3.4.3 Autres données d'intérêt

Afin de comparer la survenue des événements de santé entre les personnes ayant travaillé dans un secteur par rapport à celle n'y ayant jamais travaillé, plusieurs variables complémentaires seront utilisées pour réaliser des ajustements dans les modèles de régression ou des analyses de médiation afin d'analyser les éventuelles augmentations de risque observées. Les variables, envisagées selon l'événement de santé analysé, incluent la région de résidence, le diplôme obtenu, les revenus de la personne ou du foyer fiscal, la perception d'une pension d'invalidité ou d'une rente d'incapacité, le fait de bénéficier de la couverture médicale universelle complémentaire, le statut marital, l'indicateur de

qualité d'immigré et l'ancienneté d'arrivée en France. En effet, les facteurs liés à la qualité d'immigré sont importants à prendre en compte, d'une part car ils sont associés à un moindre recours aux soins [25] ou à des remontées d'informations médico-administratives différentes, d'autre part la proportion de travailleurs immigrés par secteur peut être non négligeable (par exemple, 21 % dans le secteur de l'hébergement et restauration en 2018) [26].

3.5 Taille de l'échantillon

Depuis son élargissement en 2004, l'EDP couvre 4,4 % de la population et compte actuellement environ 3,7 millions d'individus dont environ 80 % sont âgés de plus de 15 ans au premier janvier 2004 soit environ 3 millions de personnes selon les critères de ciblage, cette population ayant pour l'essentiel eu au moins un épisode d'activité. Un quart des individus du champ actuel sont théoriquement suivis depuis leur naissance ou depuis le recensement de 1968 si nés avant. Et pour la moitié de ces individus, ceux nés les années paires, on dispose des données de salaires du Panel en remontant au mieux jusqu'en 1976 pour les salariés déjà intégrés au Panel à cette date.

Le champ de l'EDP-Santé couvre les personnes de l'EDP vivantes en 2008 ou nées après.

Compte tenu de la répartition des travailleurs par secteur d'activité en France en 2017 (données du recensement de l'Insee), des changements de secteur d'activité dans une carrière (données de l'échantillon d'histoires professionnelles de Santé publique France), de la proportion d'individus sélectionnés pour l'EDP-Santé et de l'incidence des événements de santé envisagés dans un premier temps, le nombre d'événements de santé attendus par secteur d'activité a été calculé ainsi que la précision relative associée (Annexe 2) et la plus petite augmentation d'incidence pouvant être mise en évidence de façon statistiquement significative dans chaque secteur (Annexe 3). Le nombre de secteurs où les effectifs permettent de répondre à nos objectifs et la part d'emploi que cela représente sont présentés dans le Tableau 1.

Concernant les analyses descriptives des incidences, les données de EDP-Santé seront donc suffisantes pour les événements dont la fréquence est supérieure ou égale à 100 pour 100 000 personnes-années. Pour les événements moins fréquents, seule une minorité de secteurs d'activité pourra être considérée mais ces analyses incluront quand même la majorité des travailleurs (> 60 %).

Concernant les analyses de comparaison d'incidences, les données de l'EDP-Santé permettront de mettre en évidence des excès d'ampleur faible (RR=1,3) seulement pour les événements de santé dont la fréquence est supérieure ou égale à 100 pour 100 000 personnes-années. Pour les événements de santé moins fréquents, seules des augmentations d'incidence de plus grande ampleur pourront être mises en évidence.

Tableau°1. Nombre de secteurs d'activité (parmi 88 secteurs) et proportion d'emplois pouvant être analysés avec des effectifs considérés comme suffisants pour répondre aux objectifs

		Description des incidences[a]		Comparaison des incidences[b]	
		(A) Secteurs avec précision satisfaisante[e]	Emplois appartenant à (A)	(B) Secteurs avec précision satisfaisante[f]	Emplois appartenant à (B)
Événement de santé	Sexe	N (%[c])	%[d]	N (%[c])	%[d]
Arrêts maladie indemnisés	Hommes	86 (98 %)	100 %	86 (98 %)	100 %
	Femmes	82 (93 %)	100 %	81 (92 %)	100 %
Mortalité toutes causes	Hommes	82 (93 %)	100 %	72 (82 %)	99 %
	Femmes	76 (86 %)	100 %	62 (70 %)	98 %
Diabète de type 2	Hommes	74 (84 %)	99 %	65 (74 %)	98 %
	Femmes	66 (75 %)	99 %	54 (61 %)	97 %
Accouchement prématuré	Femmes	37 (42 %)	89 %	15 (17 %)	70 %
Accident vasculaire cérébral	Hommes	48 (55 %)	91 %	21 (24 %)	68 %
	Femmes	37 (42 %)	89 %	15 (17 %)	70 %
Pension d'invalidité	Hommes	48 (55 %)	91 %	21 (24 %)	68 %
	Femmes	37 (42 %)	89 %	15 (17 %)	70 %
Syndromes coronaires aigus dont infarctus du myocarde	Hommes	48 (55 %)	91 %	21 (24 %)	68 %
	Femmes	37 (42 %)	89 %	15 (17 %)	70 %
Mortalité par cancer trachée, bronches, poumon	Hommes	25 (28 %)	73 %	9 (10 %)	46 %
	Femmes	17 (19 %)	73 %	8 (9 %)	54 %
Mortalité par cancer du sein	Hommes	25 (28 %)	73 %	9 (10 %)	46 %
	Femmes	17 (19 %)	73 %	8 (9 %)	54 %
Mortalité par accidents	Hommes	25 (28 %)	73 %	9 (10 %)	46 %
	Femmes	17 (19 %)	73 %	8 (9 %)	54 %
Mortalité par cancer de la prostate	Hommes	16 (18 %)	61 %	5 (6 %)	32 %
Mortalité par suicide	Hommes	16 (18 %)	61 %	5 (6 %)	43 %
	Femmes	13 (15 %)	67 %	5 (6 %)	32 %

N Nombre.
a Détail par secteur en Annexe 2.
b Détail par secteur en Annexe 3.
c Parmi les 88 secteurs couvrant l'ensemble des activités françaises.
d Nombre d'emplois dans les secteurs avec une précision satisfaisante, rapporté au nombre total d'emplois.
e C'est-à-dire dont le rapport des bornes de l'intervalle de confiance à 95 % est inférieur à 2.
f C'est-à-dire dont une augmentation de 30 % de l'incidence (rapport d'incidence = 1,3) peut être mise en évidence de façon statistiquement significative.

3.6 Traitements et analyses des données

Avant de procéder aux analyses de données, un travail préalable sera effectué pour la préparation et la mise en qualité des données. L'évaluation de la disponibilité et de l'exhaustivité des variables, y compris les données manquantes sera effectuée.

S'agissant de bases de données exploitées et maintenues, la suppression des doublons sur les individus a d'ores et déjà été effectuée. La cohérence entre les dates de naissance, de décès, les périodes d'activité et informations professionnelles sera cependant vérifiée et les observations discordantes seront documentées et exclues de l'analyse le cas échéant.

Les données seront traitées dans le CASD à l'aide des logiciels SAS ou R selon sa disponibilité dans l'espace sécurisé de mise à disposition des données et l'expérience des personnes intervenant dans le traitement des données.

Toutes les analyses principales seront réalisées au niveau national, et séparément chez les hommes et les femmes.

3.6.1 Construction des indicateurs professionnels

À partir de l'ensemble des informations professionnelles disponibles dans les différentes sources de données, pour chaque individu, deux variables caractérisant le fait d'avoir travaillé dans un secteur d'activité donné seront construites de façon dynamique dans le temps selon deux définitions :

- avoir travaillé au moins une fois dans le secteur d'activité X (oui/non/manquant) ;
- avoir travaillé le plus longtemps dans le secteur d'activité X au cours de sa carrière (oui/non/manquant).

En effet, afin de définir ces variables, une attention particulière sera portée à la prise en compte de la nature des données dont la quantité d'information peut varier d'un individu à l'autre (par exemple, deux sources d'information pour les salariés (recensement et DADS) mais une seule source pour les indépendants), disponibles ponctuellement au cours de la carrière de l'individu (minimum de neuf ans entre deux années de recensement avant 2004) ou manquantes. Pour cela, des méthodes d'imputation pourront être appliquées pour prendre en compte les données manquantes et reconstituer les historiques professionnels ainsi que des méthodes d'extrapolation ou de recalcul/combinaison de poids de sondage similaires à celles réalisées par les équipes de l'Institut national des études démographiques (Ined) dans leurs exploitations de l'EDP pour décrire les trajectoires résidentielles et sociales [27].

En plus de ces définitions longitudinales d'appartenance à un secteur à un instant t, nous tâcherons de définir des types de carrière correspondant à des individus ayant des successions de secteurs d'activité (appelées carrière ou séquence) similaires. La similarité des carrières entre deux individus sera calculée en termes de nombre minimum de modifications (substitution, suppression ou insertion) nécessaire à l'une des carrières pour obtenir l'autre (optimal matching) en considérant soit toutes les modifications de secteur d'activité comme identiques, soit en leur attribuant des poids différents en fonction la proximité des secteurs d'activité dans la nomenclature NAF de l'Insee ou en termes d'expositions professionnelles. En utilisant ces similarités pour quantifier les différences/distances entre les carrières des différents individus, plusieurs algorithmes de *clustering* (méthode de regroupement hiérarchique ou partitionnement en un nombre prédéfini de groupes) seront utilisés pour définir les groupes de carrière (appelés *clusters* ou types). Le choix de l'algorithme et du nombre final de groupes sera déterminé en maximisant différents critères de qualité [28]. Ces analyses seront réalisées, séparément chez les hommes et les femmes, en utilisant l'âge comme axe du temps, à l'aide des packages TraMineR [29] et WeightedCluster [30] de R afin de prendre en compte la pondération liée aux poids de sondage des individus aux différentes étapes d'analyse.

3.6.2 Analyse des événements de santé

Compte tenu de la disponibilité des données de santé, il est proposé de décrire les indicateurs de santé de la manière suivante :

- Pour la mortalité globale, on étudiera l'espérance de vie à 35 ans afin de disposer d'un recul suffisant tout en négligeant le moins possible le début de la carrière des individus ;
- La mortalité par cause sera étudiée sur la période 2013-2017 ;
- La morbidité et son évolution seront étudiées sur les périodes 2008-2012, 2013-2017 et 2018-2022.

Une analyse descriptive des données sera préalablement réalisée en décrivant les caractéristiques des travailleurs (décennie de naissance, sexe, région, diplôme) sur l'ensemble de la période et par période de 5 ans.

Les taux d'incidence des différents événements de santé étudiés seront décrits dans l'ensemble, selon les caractéristiques des travailleurs et au cours du temps entre 2008 jusqu'à la fin du suivi dans l'échantillon.

L'espérance de vie sera calculée selon une méthode développée par l'Ined [31] par catégorie socio-professionnelle à partir de la mortalité observée sur la période la plus récente. Elle sera décrite selon le secteur d'activité le plus longtemps exercé.

Les ratios standardisés d'incidences (*Standardized Incidence Ratio* - SIR) seront calculés par secteur d'activité (selon les deux définitions) par standardisation indirecte sur l'âge et le sexe de l'ensemble des travailleurs sur la même période.

Les évolutions temporelles des taux d'incidence de chaque événement de santé, dans chaque secteur d'activité le plus longtemps exercé, seront représentées graphiquement. Puis leur évolution temporelle sera modélisée à l'aide d'une régression binomiale négative incluant la période, le secteur le plus longtemps exercé et leur interaction, afin de tester si l'effet période pour un secteur donné est statistiquement différent de l'effet période dans l'ensemble.

Pour comparer la survenue des événements de santé à cinq ou dix ans selon le secteur d'activité exercé, pour chaque secteur d'activité, les travailleurs exerçant dans ce secteur d'activité en 2009 (groupe exposé) seront appariés aléatoirement à 5 référents aux caractéristiques socio-démographiques similaires (groupe non exposé) mais travaillant dans un autre secteur.

La survenue des événements de santé non récurrents dans les deux groupes (exposés vs non exposés) sera comparée, pour chaque secteur d'activité, à l'aide d'un modèle de survie. Selon l'événement de santé étudié, son histoire naturelle et ses associations possibles avec l'environnement professionnel, le suivi sera censuré entre zéro et cinq ans après le moment où la personne quitte le secteur d'activité. Les phénomènes d'exclusion au travail et du secteur d'activité étudié seront pris en compte à l'aide de la formule g paramétrique [32].

Afin d'évaluer l'impact potentiel de facteurs de confusion non disponibles dans l'étude (par exemple du tabagisme), des analyses de sensibilité de type E-value [33] ou analyse de biais probabiliste [34] seront réalisées pour tenir compte des différences de comportement de santé, entre secteur d'activité observées dans des sources de données externes [10].

Les associations entre les types de carrière (cf. Construction des indicateurs professionnels) et les événements de santé seront mesurées en utilisant la même méthode que celle utilisée pour comparer les secteurs entre eux.

Pour évaluer la contribution des secteurs d'activité sur les inégalités d'état de santé observées dans un groupe de travailleurs spécifiques déterminé a priori – par exemple les travailleurs seniors – l'importance des secteurs d'activité dans la survenue des événements de santé sera évaluée en calculant le coefficient de corrélation intra-classes dans un modèle multiniveaux utilisant le niveau secteur comme variable aléatoire [35].

Plus de détail sur les méthodes d'analyse est présenté en Annexe 4.

3.7 Limites de l'étude

La représentativité des échantillons est un atout essentiel pour répondre aux objectifs descriptifs de l'étude. Les travaux de l'Insee et de la Drees analysant les croisements entre les sources d'information ont permis d'identifier et documenter les écarts observés dans le repérage des personnes au sein des différentes sources de données en fonction du champ théorique de l'EDP. Certaines personnes entrées dans le champ théorique de l'EDP en 2004 n'ont pas d'événement collecté dans l'EDP. D'autres personnes identifiées par un événement dans l'EDP ne sont pas retrouvées au Répertoire des personnes physiques. Enfin certains individus EDP n'ont pas de correspondance dans le SNDS. Selon la situation, il s'agit le plus souvent de personnes sorties du territoire, voire décédées à l'étranger sans collecte du bulletin de décès donc plutôt âgées, ou au contraire de personnes trop jeunes pour apparaître dans les événements collectés, et pas encore immatriculées sous leur propre NIR dans le SNDS [12].

Il est donc peu probable que l'absence de ces personnes ait un impact important sur la représentation des individus ayant eu une activité professionnelle sur la période d'étude, sauf pour ce qui concerne les travailleurs étrangers. Une analyse comparative des situations professionnelles sera donc réalisée pour documenter cette limite.

D'autre part, s'appuyer sur des données existantes et recueillies de manière systématique présente une contrepartie non négligeable. Ces données, collectées par un autre organisme et pour un autre objectif, ne comportent pas toutes les informations souhaitées pour répondre aux objectifs de la surveillance épidémiologique. En particulier, l'information individuelle sur les facteurs extra-professionnels est constamment absente, notamment la consommation de tabac, d'alcool, les comportements en matière d'activité physique qui peuvent avoir un rôle de confusion très important car potentiellement lié à la fois à la situation socioprofessionnelle et à la santé. Pour les objectifs descriptifs de l'étude, cette absence d'information ne limite pas l'utilité des données produites et ne remet pas en cause les éventuels excès d'événements de santé observés ; ils sont suffisants à eux seuls pour orienter la mise en place d'actions de prévention. En revanche, pour les objectifs cherchant à mieux caractériser le rôle des secteurs d'activité sur la santé, des stratégies sont envisagées pour compenser en partie ce manque d'information : notamment en utilisant des fréquences de comportement de santé (tabagisme, etc.) décrites à partir d'autres sources de données [10] permettant d'ajuster les résultats sur des comportements moyens [34, 36].

Enfin malgré la taille importante de l'échantillon, l'étude sera limitée pour décrire avec précision la fréquence d'événements rares dans les secteurs d'activité ayant les plus petits effectifs et pour mettre en évidence des augmentations d'incidence faibles pour des événements rares. De plus, en raison de la profondeur de disponibilité des données de santé (onze ans), nos analyses sont limitées pour évaluer les effets à long terme du fait de travailler dans un secteur d'activité.

4. RÔLE DES DIFFÉRENTS ACTEURS ET COMITOLOGIE

Les analyses sont conduites en accord avec la stratégie de la Direction Santé Environnement Travail de Santé publique France dans le cadre de ses missions de surveillance de la santé en lien avec les risques professionnels.

Santé publique France a la charge de :

- La conception de l'étude ;
- Le financement de l'étude ;
- La réalisation des démarches réglementaires ;
- La coordination des différents acteurs ;
- L'application de l'information collective ;
- L'analyse statistique des données ;
- L'interprétation des résultats ;
- L'écriture de rapport d'étude pour chaque période de suivi ;
- La diffusion des résultats et leur communication.

L'Insee met à disposition les données utiles au projet et leur documentation. La mise à disposition de ces données est encadrée par le comité du secret statistique (cf. Démarches réglementaires). Les résultats de l'exploitation feront par ailleurs l'objet d'échanges avec les chercheurs exploitant également ces données dans le cadre du groupe utilisateurs animé par la direction de la démographie de l'Insee.

La Drees a réalisé l'appariement entre l'EDP et les données du SNDS. La Drees a été sollicitée dès l'initiation du projet d'exploitation de l'EDP-Santé, afin de définir les modalités d'accès aux données de l'échantillon apparié, à la fois sur le plan réglementaire et technique. Cette mise à disposition des données est encadrée par une convention non financière conclue entre la Drees et Santé publique France.

Le CASD est un groupement d'intérêt public proposant, sous forme de prestations techniques, une technologie de bulles sécurisées destinée à sécuriser le traitement sécurisé des données sensibles pour la recherche scientifique. Un espace de travail mis à disposition de l'utilisateur par le CASD lui permet de travailler à distance sur les données tout en garantissant au producteur que les données ne puissent être extraites (cf. Support de données et sécurité).

Aucun comité n'est associé à ce projet. Des experts internes à Santé publique France ou extérieurs pourront être mobilisés sur des thématiques ou des problématiques spécifiques, avec le cas échéant, si cela est nécessaire, la constitution d'un comité ad hoc.

5. DÉMARCHES RÉGLEMENTAIRES

Un certain nombre de démarches réglementaires a été nécessaire pour obtenir les autorisations d'exploiter les données nécessaires à ce projet, lequel impliquait de concilier les exigences en matière de secret statistique mais aussi de traitement de données de santé.

Le principe du secret statistique (décrit et cadré, en France, dans la loi 1951 sur l'obligation, la coordination et le secret en matière de statistiques) impose un contrôle strict des exploitations des données. Pour les personnes souhaitant exploiter des données individuelles, les conditions d'accès sont soumises au secret (autorisation via le comité du secret) et exigent des conditions de sécurité et des contrôles afin d'interdire toute copie des données et vérifier que les sorties de résultats ne contreviennent pas au principe du secret statistique (en particulier grâce à l'accès et l'exploitation dans le CASD).

Après engagement de confidentialité de toute personne ayant accès aux données et information des producteurs de données, le protocole a été soumis au Comité du secret qui a autorisé l'accès aux données pour le présent projet le 19 mars 2024. Cette autorisation a été complétée par une demande auprès des Archives de France pour accéder par dérogation à des archives publiques non librement communicables. Cette dérogation a été accordée le 10 avril 2024.

Le projet prévoyant également le traitement de données de santé, des formalités préalables complémentaires ont été nécessaires auprès du *Health Data Hub* (HDH)[1]. S'agissant d'une recherche n'impliquant pas la personne humaine et ne répondant pas à une méthodologie de référence, une demande d'avis a été déposée auprès du Comité éthique et scientifique pour les recherches, les études et les évaluations dans le domaine de la santé (Cesrees) chargé de se prononcer notamment sur la pertinence éthique et la qualité scientifique du projet et sur son caractère d'intérêt public au regard de sa finalité et de la méthodologie utilisée. Un avis favorable a été rendu par ce comité le 14 décembre 2023.

Le dossier accompagné des avis des deux comités (Comité du secret et Cesrees) a enfin été soumis à l'examen de la Commission nationale de l'informatique et des libertés (Cnil) (dossier 924113), en se plaçant sous le régime de l'autorisation, ce traitement présentant une finalité d'intérêt public. Le traitement a été autorisé par la décision DR 2024-125 du 2 août 2024, complétée par la décision DR 2024-264 du 25 octobre 2024.

Une analyse d'impact relative à la protection des données (AIPD) a également été effectuée, rendue nécessaire pour l'instruction de ce type de demande d'autorisation et également par la mobilisation de sources d'origine administrative, par l'usage de données « sensibles », et par le rapprochement entre différentes sources de données. Cette analyse n'a pas révélé de risque élevé pour les droits et libertés des personnes physiques.

[1] Également appelée Plateforme des données de santé.

6. PROTECTION DE LA VIE PRIVÉE, SÉCURITÉ ET CONFIDENTIALITÉ DES DONNÉES

6.1 Respect des droits des personnes concernées

6.1.1 Information collective des personnes

Les données EDP mises à disposition par l'Insee et les données appariées de l'EDP-Santé constituées par la Drees sont des données pseudonymisées. En conséquence, il est impossible pour Santé publique France de contacter et d'informer individuellement les personnes concernées, dans le cadre de ce projet d'exploitation.

S'agissant des organismes en charge des bases de données exploitées dans le cadre de ce projet :

- La Drees dispose elle aussi de données pseudonymisées ne permettant pas la ré-identification des personnes concernées ;
- L'Insee peut retrouver l'identité des personnes mais ne dispose pas des coordonnées permettant une information individuelle. De plus, les données mises à disposition de Santé publique France comportent un identifiant pseudonymisé ne permettant pas de refaire le lien avec l'identité.

Ainsi, une dérogation à l'obligation d'information individuelle a été sollicitée auprès de la Cnil, dans le cadre de l'article l'art. 14.5 b) du Règlement général sur la protection des données (RGPD).

Une information collective publique est mise en place sur le site web de l'agence. Cette page internet présente les informations suivantes :

- Les objectifs du projet et le contexte du projet ;
- La description des données utilisées ;
- Les références bibliographiques sur les sources de données ;
- Le calendrier de réalisation de l'étude, notamment la durée de conservation des données ;
- Les résultats obtenus et les publications ;
- Le dispositif visant à garantir la sécurisation et la confidentialité des données ;
- Les dispositions visant à garantir l'exercice des droits des personnes concernées ;
- Les coordonnées du responsable de traitement et du délégué à la protection des données (DPD) : adresse électronique, adresse postale, numéro de téléphone ;
- Les coordonnées de la Cnil en cas de réclamation ;
- Et un lien « contact » pour toute question sur le projet : email, adresse postale.

Une note d'information collective sera également mise en ligne sur le site de la Drees dans la rubrique « autres travaux » : https://drees.solidarites-sante.gouv.fr/sources-outils-et-enquetes/ledp-sante-enrichissement-de-lechantillon-demographique-permanent-par

6.1.2 Droits d'accès, de rectification et d'opposition

Les données étant pseudonymisées, la réidentification des personnes est impossible et les droits d'accès, de rectification et d'opposition ne pourront s'exercer qu'à travers les producteurs de données selon leurs modalités : la Drees, la Cnam et l'Insee.

Concernant les sources EDP et EDP-Santé : les personnes concernées souhaitant faire valoir leurs droits devront en faire la demande auprès de Santé publique France en s'adressant à l'adresse

dpo@santepubliquefrance.fr en indiquant le code EDP-Santé dans leur demande.

Cette procédure figure également dans la note d'information collective publiée sur le site de Santé publique France (cf. Annexe 5).

Après enregistrement et vérification de la demande, la délégué à la protection des données (DPO) de Santé publique France pourra :

- Concernant les données de l'EDP-Santé, informer la Drees de l'exercice de droit d'une personne en envoyant un courriel à l'adresse drees-rgpd@sante.gouv.fr avec le code EDP-Santé.
- Concernant les données provenant de la base de données du SNDS, réorienter vers la procédure détaillée sur le site https://www.snds.gouv.fr/SNDS/Protection-de-la-donnee précisant que les personnes concernées souhaitant faire valoir leurs droits devront en faire la demande auprès du directeur de l'organisme gestionnaire du régime d'assurance maladie obligatoire auquel la personne est rattachée, conformément aux articles 92 à 95 du décret n° 2005-1309 du 20 octobre 2005.
- Concernant les données provenant de la base de données de l'EDP, réorienter vers l'adresse contact-rgpd@insee.fr

Le droit d'opposition prévu aux premier et troisième alinéas de l'article 56 de la loi n° 78-17 du 6 janvier 1978 porte sur l'utilisation des données dans les traitements mentionnés au 1° du I de l'article L. 1461-3 du code de la santé publique. À noter que ce droit ne s'applique pas aux traitements nécessaires à l'accomplissement des missions des services de l'État, des établissements publics ou des organismes chargés d'une mission de service public compétents.

6.2 Confidentialité et sécurité des données

6.2.1 Gestion du risque de réidentification

Toutes les données issues de l'EDP ou de l'EDP-Santé sont pseudonymisées à l'aide d'un numéro unique propre à la présente étude.

Aucune donnée directement identifiante n'est disponible dans les bases mises à disposition.

Le risque de réidentification est éventuellement possible via la date de naissance, la date de l'éventuel décès et les données de cursus professionnels ou d'événements de santé. Le risque est cependant limité par la généralisation des informations géographiques agrégées à l'échelle du département.

Aucune donnée liée à l'affiliation ethnique, politique, religieuse, philosophique ou syndicale, ni donnée génétique, ni donnée biométrique, ni donnée sur la vie ou l'orientation sexuelle n'est traitée dans le cadre de cette étude.

6.2.1 Support de données et sécurité

6.2.1.1 Mesures techniques et organisationnelles visant à assurer la sécurité des données

Concernant l'EDP ou l'EDP-Santé, les données pseudonymisées seront traitées au sein d'un espace projet sécurisé du CASD avec interdiction d'extraire des données individuelles (Certification ISO 27001).

Pour l'exploitation de l'EDP-Santé, une convention signée entre la Drees et Santé publique France, encadrera la transmission des données en précisant les conditions de sécurité des données, ainsi que les objectifs et le calendrier.

Les données seront chiffrées avec GnuPG par la Drees et protégées par un mot de passe. Cette procédure garantit la sécurisation de la transmission des données et la protection des données. Les données chiffrées protégées par un mot de passe seront ensuite transmises par la Drees au CASD, qui placera les données dans un espace sécurisé du CASD. Les données seront stockées et traitées sur cet espace sécurisé via une SD-Box et seront accessibles uniquement aux membres de l'équipe projet identifiée précédemment.

La SD-Box, mise à disposition de l'utilisateur, lui permet de travailler à distance sur les données tout en garantissant au producteur de données :

- Qu'aucun fichier ne puisse être récupéré par l'utilisateur (pas de copier/coller, d'impressions, d'insertion de clé USB…) ;
- Qu'il s'agit bien de l'utilisateur habilité qui se connecte sur la SD-Box (authentification forte biométrique : le contrôle d'accès de l'utilisateur est réalisé à l'aide d'une carte à puce contenant un certificat de sécurité et un lecteur biométrique d'empreintes digitales. Conformément à la loi, ce traitement a fait l'objet d'une demande d'autorisation à la Commission nationale de l'informatique et des libertés qui lui a été accordée : Cnil – délibération n° 2014-369) ;
- Que toutes les communications entre la SD-Box et les serveurs sont chiffrées.

6.2.1.2 Minimisation des risques pour la vie privée

L'espace dédié au CASD est certifié ISO 27701, Hébergeur de données de Santé, et homologué au Référentiel de sécurité des données de santé. Il permet la traçabilité des accès.

Les mesures de sécurité s'appliquant à cet espace projet ont été mises en place à la fois par le *Health Data Hub*, conformément à la démarche de sécurité validée par l'ANSSI (Agence nationale de la sécurité des systèmes d'information) et la Cnil. https://www.casd.eu/technologie/securite-certifications/

Le dispositif de sécurité repose sur les mesures suivantes :

- Une authentification forte des utilisateurs à base de certificats et de biométrie (empreinte digitale) ;
- Des points d'accès sécurisés et authentifiés (SD-Box) ;
- L'accès d'un utilisateur ou d'un point d'accès peut être arrêté à distance ;
- L'utilisateur n'a accès qu'aux applications autorisées (accès distant) ;
- Chaque projet dispose d'un espace dédié dont l'accès est réservé aux seuls utilisateurs habilités ;
- Un contrôle par les gestionnaires du CASD des informations exportées de manière à garantir le respect du secret statistique, et la conservation de toutes les opérations d'export d'informations ;
- La traçabilité de tous les traitements sur les données sensibles ;
- L'activité (en débit) d'un point d'accès est surveillée ;
- Les données stockées sont chiffrées.

Les données sont stockées et traitées dans un espace projet sécurisé de la plateforme cible, accessible aux personnes habilitées seulement. Elles seront conservées pendant le nombre d'années autorisées pour la réalisation de l'étude et la publication des résultats, puis elles seront définitivement supprimées.

6.2.2 Circuit des données

Le circuit des données pour l'appariement entre la source EDP et la source SNDS comporte 3 étapes.

Étape 1 - Constitution de la cohorte initiale et transmission du NIR à la Cnam

Dans le cadre de la base source EDP, le producteur dispose pour chaque personne de son NIR, de sa date de naissance et de son sexe.

À partir de la base source, le producteur constitue la cohorte initiale, selon les critères de ciblage.

Un numéro de cohorte unique non signifiant (ID_PROJSPF), différent de l'identifiant communément présent dans les données sources, est attribué de façon aléatoire à chaque personne concernée.

Le producteur transmet, à une structure déconcentrée de la Cnam, la table de pseudonymisation [ID_PROJSPF ; NIR] constituée, pour chaque personne, de son NIR propre et/ou celui de son ouvrant droit, de son sexe, de sa date de naissance ainsi que du numéro de cohorte non signifiant (ID_PROJSPF) qui lui a été attribué. Ce fichier est transmis selon la procédure SAFE suivant les modalités en vigueur.

Étape 2 - Extraction des données du SNDS à partir du NIR

La structure déconcentrée de la Cnam applique aux NIR en clair figurant dans le fichier normé (généré à partir de la table de pseudonymisation [ID_ PROJSPF ; NIR]), les mêmes algorithmes de hachage utilisés lors de la consolidation de la base principale du SNDS. À partir de l'identifiant SNDS ainsi obtenu, la Cnam cible les personnes concernées de l'étude dans le SNDS et extrait les données requises pour le projet. Elle attribue un identifiant non signifiant spécifique au projet (ID_SNDS_PROJSPF) et consolide en parallèle une table de correspondance des identifiants pseudonymisés [ID_PROJSPF ; ID_SNDS_PROJSPF] entre l'identifiant temporaire transmis par le producteur (ID_PROJSPF) et l'identifiant présent dans l'extraction SNDS qui sera fournie (ID_SNDS_PROJSPF).

La correspondance entre les pseudonymes pourra être conservée par la Cnam pendant la durée d'accès du projet aux données SNDS, notamment pour des corrections ou des mises à jour des données SNDS.

Étape 3 – Appariement des jeux de données et traitement dans la plateforme cible

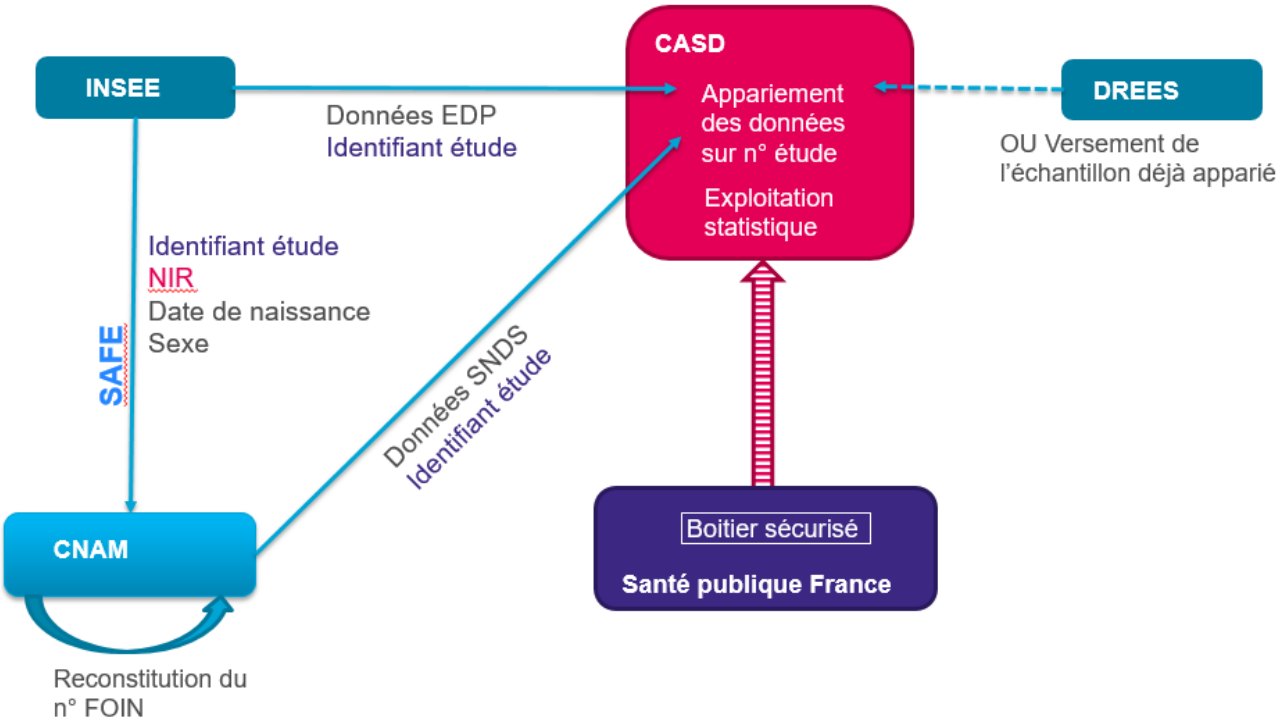
La cohorte initiale pseudonymisée (données EDP) et l'extraction du SNDS avec la table de correspondance des identifiants pseudonymisés [ID_PROJSPF; ID_SNDS_PROJSPF] sont transmises respectivement par le producteur et la Cnam au gestionnaire de la plateforme cible dédiée pour le projet via des réseaux dédiés, permettant un transfert privé (hors internet) et sécurisé des données.

À partir de la correspondance [ID_PROJSPF; ID_SNDS_PROJSPF], les jeux de données sont ensuite appariés par l'opérateur de la plateforme et un nouvel identifiant est créé de façon aléatoire pour chaque personne concernée (ID_EDPSNDS_SPF).

Les données sont stockées et traitées dans un espace projet sécurisé de la plateforme cible, accessible aux personnes habilitées seulement.

La figure 2 ci-dessous schématise ce circuit d'appariement entre l'EDP et le SNDS pour leur mise à disposition au sein du CASD accessible via un boîtier sécurisé.

Figure 2. Schéma de circulation des données



Cet appariement, lourd et complexe et nécessitant la circulation du NIR, ayant déjà été réalisé par la Drees, Santé publique France a sollicité l'autorisation de réutiliser le produit de cet appariement qui est mis à disposition par la Drees par versement dans l'espace sécurisé distant du CASD.

6.3 Durée de conservation des données.

Les données seront conservées en base active pendant le nombre d'années autorisé pour la réalisation du projet, c'est-à-dire jusqu'au 31 décembre 2028. Puis elles seront archivées temporairement et resteront consultables durant deux années après la fin de l'exploitation, le temps nécessaire à la valorisation complète des travaux.

À l'issue de cette période d'archivage temporaire, l'accès à l'espace projet sera interrompu et les données seront supprimées définitivement de l'espace projet.

7. CALENDRIER PRÉVISIONNEL ET FAISABILITÉ DU PROJET

Santé publique France dispose d'une expérience dans l'analyse de données issues de cohorte de travailleurs et de grandes bases de données. En particulier, l'agence a déjà exploité les données de l'EDP et du Panel DADS dans leur version 2002 appariées aux causes médicales de décès afin de produire des indicateurs de mortalité par cause selon l'activité professionnelle [11, 13, 36, 37]. Elle a donc acquis une bonne connaissance des données, de leur exploitabilité et de leurs limites. L'ouverture de ces données à la communauté scientifique et leurs exploitations ont contribué par ailleurs à améliorer la qualité des données au fil du temps (meilleure identification, dédoublonnage des individus).

Concernant les données du SNDS, Santé publique France dispose d'un accès permanent au portail et exploite ces données en routine. Ainsi, l'Agence possède une longue expérience dans l'exploitation de ces données. Par ailleurs, les données de mortalité par cause issues directement du CépiDc ont été exploitées à de nombreuses reprises par Santé publique France dans le champ de la surveillance de la santé au travail.

Toutes les personnes impliquées dans le projet SEESTA ont suivi la formation dispensée par la Cnam sur l'architecture des données du SNDS complétée par la formation pour l'exploitation de données à travers le portail ou sur extraction spécifique.

Le calendrier prévisionnel de l'étude est présenté dans le tableau 2.

Tableau 2. Calendrier prévisionnel

Étapes	Période prévisionnelle
Démarches réglementaires	
Dépôt du dossier au Cesrees	Novembre 2023
Examen au Comité du secret statistique	Mars 2024
Accord Cnil	Mai-Novembre 2024
Information collective	Juin-Juillet 2024
Démarche administrative	
Convention Santé publique France/Drees	Décembre 2024
Exploitation des données	
Mise à disposition des données	Janvier 2025
Mise en œuvre des analyses	Janvier 2025-Juin 2028
Valorisation	
Publications de rapports/articles	Fin 2025-Fin 2028
Clôture du projet	
Finalisation du projet	Fin 2028
Suppression de l'accès aux données	Fin 2030

Durée totale d'accès = Six ans et demi (de septembre 2024 à fin 2030)

8. RÉFÉRENCES BIBLIOGRAPHIQUES

[1] Haut Comité de la santé publique. Les inégalités sociales de santé : sortir de la fatalité. Paris : HCSP; 2010. 103 p. [consulté le 29/08/2024]. Disponible: https://www.hcsp.fr/explore.cgi/avisrapportsdomaine?clefr=113

[2] Centre international de recherche sur le cancer. Les cancers attribuables au mode de vie et à l'environnement en France métropolitaine. 2018. 271 p. [consulté le 29/09/2024]. Disponible: https://gco.iarc.fr/includes/PAF/PAF_FR_report.pdf

[3] Institut national de la santé et de la recherche médicale. Santé et conditions de travail, une recherche à développer. Paris : Documentation Française; 1985.

[4] World Health Organization, International Labour Organization. WHO/ILO joint estimates of the work-related burden of disease and injury, 2000-2016: global monitoring report.; 2021. 92 p. [consulté le 29/08/2024]. Disponible: https://iris.who.int/bitstream/handle/10665/345242/9789240034945-eng.pdf?sequence=1

[5] Niedhammer I, Bertrais S, Witt K. Psychosocial work exposures and health outcomes: a meta-review of 72 literature reviews with meta-analysis. Scand J Work Environ Health. 2021;47(7):489-508.

[6] Institut national de la statistique et des recherches économiques. Nomenclature d'activités française (2e révision) [En ligne]. 2008. [consulté le 29/08/2024]. Disponible: https://www.insee.fr/fr/information/3281579

[7] Mannetje At, Kromhout H. The use of occupation and industry classifications in general population studies. International Journal of Epidemiology. 2003;32(3):419-28.

[8] Service-public. Convention collective [En ligne]. 2003. [consulté le 29/08/2024]. Disponible: https://www.service-public.fr/particuliers/vosdroits/F78

[9] Dejours C. Chapitre VIII. Comment formuler une problématique de la santé en ergonomie et en médecine du travail (1995). Situations du travail. Paris : Presses universitaires de France; 2016. p. 195-217. [consulté le 29/08/2024]. Disponible: https://shs.cairn.info/situations-du-travail--9782130735380-page-195?lang=fr

[10] Andler R, Rabet G, Guignard R, Pasquereau A, Quatremère G, Richard J-B, *et al.* Consommation de substances psychoactives et milieu professionnel. Résultats du Baromètre de Santé publique France 2017. Saint-Maurice : Santé publique France; 2021. 17 p. [consulté le 29/08/2024]. Disponible: https://www.santepubliquefrance.fr/determinants-de-sante/alcool/documents/enquetes-etudes/consommation-de-substances-psychoactives-et-milieu-professionnel.-resultats-du-barometre-de-sante-publique-france-2017

[11] Geoffroy-Perez B. Analyse de la mortalité et des causes de décès par secteur d'activité de 1968 à 1999 à partir de l'Échantillon démographique permanent. Étude pour la mise en place du programme Cosmop : cohorte pour la surveillance de la mortalité par profession. Saint-Maurice : Institut de veille sanitaire; 2006. 162 p. [consulté le 29/08/2024]. Disponible: https://www.santepubliquefrance.fr/docs/analyse-de-la-mortalite-et-des-causes-de-deces-par-secteur-d-activite-de-1968-a-1999-a-partir-de-l-echantillon-demographique-permanent.-etude-pour

[12] Dubost C-L, Leduc A. L'EDP-Santé, un appariement des données socio-économiques de l'échantillon démographique permanent au Système national des données de santé. Les dossiers de la Drees. 2020;66.

[13] Geoffroy-Perez B, Fouquet A, Rabet G, Julliard S. Programme Cosmop : surveillance de la mortalité par cause selon l'activité professionnelle : Analyse de la mortalité et des causes de décès par secteur d'activité de 1976 à 2005. Saint-Maurice : Santé publique France; 2018. 49 p. p. [consulté le 29/08/2024]. Disponible: https://www.santepubliquefrance.fr/docs/programme-cosmop-surveillance-de-la-mortalite-par-cause-selon-l-activite-professionnelle-analyse-de-la-mortalite-et-des-causes-de-deces-par-sec

[14] Tuppin P, Rudant J, Constantinou P, Gastaldi-Menager C, Rachas A, de Roquefeuil L, *et al*. Value of a national administrative database to guide public decisions: From the Système national d'information interregimes de l'assurance maladie (Sniiram) to the Système national des données de santé (SNDS) in France. Revue d'épidemiologie et de santé publique. 2017;65 Suppl 4:S149-s67.

[15] De Roquefeuil L, G. R, Bounebache K, Imbaud C. Guide causes médicales de décès [En ligne]. 2020. [consulté le 29/08/2024]. Disponible: https://documentation-snds.health-data-hub.fr/snds/formation_snds/documents_cnam/guide_cepidc/2-Chap2CausesDeces.html

[16] Delva F, Andujar P, Lacourt A, Brochard P, Pairon JC. [Occupational risk factors for lung cancer]. Revue des maladies respiratoires. 2016;33(6):444-59.

[17] Cordina-Duverger E, Menegaux F, Popa A, Rabstein S, Harth V, Pesch B, *et al.* Night shift work and breast cancer: a pooled analysis of population-based case-control studies with complete work history. European journal of epidemiology. 2018;33(4):369-79.

[18] Kecklund G, Axelsson J. Health consequences of shift work and insufficient sleep. Bmj. 2016;355:i5210.

[19] Gandaglia G, Leni R, Bray F, Fleshner N, Freedland SJ, Kibel A, *et al.* Epidemiology and Prevention of Prostate Cancer. Eur Urol Oncol. 2021;4(6):877-92.

[20] Greiner BA, Arensman E. The role of work in suicidal behavior - uncovering priorities for research and prevention. Scandinavian journal of work, environment & health. 2022;48(6):419-24.

[21] Roquelaure Y. Troubles musculo-squelettiques et facteurs psychosociaux au travail [Internet]. Bruxelles : Etui ; 2018 [cité 18 mai 2022] p. 84. Report N°.: 142. Bruxelles2018.

[22] Cai C, Vandermeer B, Khurana R, Nerenberg K, Featherstone R, Sebastianski M, *et al.* The impact of occupational shift work and working hours during pregnancy on health outcomes: a systematic review and meta-analysis. Am J Obstet Gynecol. 2019;221(6):563-76.

[23] Descatha A, Sembajwe G, Pega F, Ujita Y, Baer M, Boccuni F, *et al.* The effect of exposure to long working hours on stroke: A systematic review and meta-analysis from the WHO/ILO Joint Estimates of the Work-related Burden of Disease and Injury. Environ Int. 2020;142:105746.

[24] Moretti Anfossi C, Ahumada Munoz M, Tobar Fredes C, Perez Rojas F, Ross J, Head J, *et al.* Work Exposures and Development of Cardiovascular Diseases: A Systematic Review. Ann Work Expo Health. 2022;66(6):698-713.

[25] Berchet C, Jusot F. État de santé et recours aux soins des immigrés en France : une revue de la littérature. Bulletin épidémiologique hebdomadaire. 2012(2-3-4):17-21.

[26] Immigrés selon le secteur d'activité et le domaine professionnel - Recensement de la population 2018 [En ligne]. 2018 [consulté le 29/08/2024]. Disponible: https://www.insee.fr/fr/statistiques/5367268?sommaire=5363676

[27] Le Roux G, Bonvalet C, Bringé A. Apports et limites de l'échantillon démographique permanent à l'analyse des trajectoires résidentielles et des inégalités spatiales (1968-2014). Institut national d'études démographiques (Ined); 2021. Disponible: https://www.ined.fr/fr/publications/editions/document-travail/apports-et-limites-echantillon-demographique-permanent-analyse-des-trajectoires-residentielles-et-des-inegalites-spatiales-1968-2014/

[28] Studer M. Le manuel de la librairie WeightedCluster - Un guide pratique pour la création de typologies de trajectoires en sciences sociales avec R. 2013. [consulté le 29/08/2024]. Disponible: https://cran.r-project.org/web/packages/WeightedCluster/vignettes/WeightedClusterFR.pdf

[29] Gabadinho A, Ritschard G, Müller NS, Studer M. Analyzing and Visualizing State Sequences in R with TraMineR. Journal of Statistical Software. 2011;40(4):1 - 37.

[30] Studer M. WeightedCluster Library Manual: A practical guide to creating typologies of trajectories in the social sciences with R. Suisse : 2013. [cité le 29/08/2024]. Disponible: https://www.centre-lives.ch/sites/default/files/2020-11/24_lives_wp_studer_weightedcluster.pdf

[31] Currie ID, Durban M, Eilers PH. Smoothing and forecasting mortality rates. Statistical Modelling. 2004;4(4):279-98.

[32] Keil AP, Edwards JK, Richardson DB, Naimi AI, Cole SR. The parametric g-formula for time-to-event data: intuition and a worked example. Epidemiology (Cambridge, Mass). 2014;25(6):889-97.

[33] VanderWeele TJ, Mathur MB. Commentary: Developing best-practice guidelines for the reporting of E-values. Int J Epidemiol. 2020;49(5):1495-7.

[34] Lash TL, Fox MP, Fink AK. Applying Quantitative Bias Analysis to Epidemiologic Data. New York : Springer; 2009. Disponible: https://link.springer.com/book/10.1007/978-0-387-87959-8

[35] Nakagawa S, Johnson PCD, Schielzeth H. The coefficient of determination R(2) and intra-class correlation coefficient from generalized linear mixed-effects models revisited and expanded. J R Soc Interface. 2017;14(134).

[36] Lauzeille D, Marchand JL, Ferrand M. Consommation de tabac par catégorie socio-professionnelle et secteur d'activité. Outil méthodologique pour l'épidémiologie. Saint-Maurice : Institut de veille sanitaire; 2009. 208 p. [consulté le 29/08/2024]. Disponible: https://www.santepubliquefrance.fr/determinants-de-sante/tabac/documents/rapport-synthese/consommation-de-tabac-par-categorie-socioprofessionnelle-et-secteur-d-activite.-outil-methodologique-pour-l-epidemiologie

[37] Cohidon C, Geoffroy Perez B, Fouquet A, Le Naour C, Goldberg M, Imbernon E. Suicide et activité professionnelle en France : premières exploitations de données disponibles. Saint-Maurice : Institut de veille sanitaire; 2010. 8 p. [consulté le 29/08/2024]. Disponible: https://www.santepubliquefrance.fr/maladies-et-traumatismes/sante-mentale/suicides-et-tentatives-de-suicide/documents/rapport-synthese/suicide-et-activite-professionnelle-en-france-premieres-exploitations-de-donnees-disponibles

[38] Glynn RJ, Buring JE. Ways of measuring rates of recurrent events. Bmj. 1996;312(7027):364-7.

[39] Mulder PGH. An exact method for calculating a confidnce interval of a poisson parameter. American Journal of Epidemiology. 1983;117(3):377.

[40] Glynn RJ, Stukel TA, Sharp SM, Bubolz TA, Freeman JL, Fisher ES. Estimating the variance of standardized rates of recurrent events, with application to hospitalizations among the elderly in New England. Am J Epidemiol. 1993;137(7):776-86.

[41] Amorim LD, Cai J. Modelling recurrent events: a tutorial for analysis in epidemiology. Int J Epidemiol. 2015;44(1):324-33.

[42] Steen J, Loeys T, Moerkerke B, Vansteelandt S. medflex: An R Package for Flexible Mediation Analysis using Natural Effect Models. Journal of Statistical Software. 2017;76(11):1 - 46.

[43] Makhzoum S. Les seniors sur le marché du travail en 2021 - Un taux d'emploi toujours en progression. Dares Résultats. 2023;2.

9. ANNEXES

Annexe 1. Liste des principales informations de l'EDP utiles à l'analyse selon la source

L'EDP se compose d'une table centrale contenant les individus, avec leurs principales caractéristiques démographiques, ainsi que des variables indicatrices renseignant sur les événements susceptibles d'apparaître dans les différentes sources. Chacun de ces événements renvoie ensuite à une table plus détaillée à partir de la source concernée.

Libellé	Modalités
Identifiant de diffusion de l'individu EDP	
Date de naissance de l'individu EDP	
Commune et département de naissance de l'individu EDP	Code officiel géographique
Sexe de l'individu EDP	F-Femme ; H-Homme ; I-Indéterminé, Non renseigné
Date du décès de l'individu EDP	
Commune et département du décès de l'individu EDP	Code officiel géographique
Nombre d'enfants de l'individu EDP	Selon l'état civil, y compris les enfants adoptés
Indicateur du lieu de naissance des parents de l'individu EDP	France/Étranger ; département si France
Nombre d'inscriptions électorales de l'individu EDP	
Nombre de bulletins individuels aux différents recensements de l'individu EDP	1968 ; 1975 ; 1982 ; 1990 ; 1999 ; année aaaa (EAR)
Existence d'un acte de décès, d'un jugement de disparition ou d'absence	
Existence d'un acte de naissance ou d'un acte d'adoption de l'individu EDP	
Nombre d'actes de mariage de l'individu EDP	
Nombre d'années d'activité de l'individu EDP dans le panel d'actif	
Inscription au Répertoire National d'Identification des Personnes Physiques	Inscrit / non inscrit / non migré (refonte 2010)
Nombre de déclarations fiscales de l'individu EDP pour l'année aaaa	aaaa = 2011 à 2017

Informations issues des bulletins des différents recensements

	BI RP68	BI RP 75	BI RP82 **	BI RP90	BI RP99	EAR
Identifiant de diffusion de l'individu EDP						
Sexe			r			
Date de naissance						
Département de naissance						
Indicateur du lieu de naissance (France / Étranger)						
Département de naissance ou pays si naissance à l'étranger						
Nationalité				r		
Situation matrimoniale						
Lieu de résidence				r		
Catégorie de population (logement ordinaire ou autre précisé)						
Diplôme				r		
Catégorie socioprofessionnelle (détail profession ≥ 1990)				r		
Activité économique employeur*				r		
Statut						
Recherche de travail						
Ancienneté de recherche de travail						
Lieu de travail				r		
Temps de travail						
Situations professionnelles et professions antérieures						

BI RP : Bulletin individuel du recensement de population ; EAR : enquête annuelle de recensement (à partir de 2004)
** NAE59 en 1968; NAP73 de 1975 à 1982 + 1990 ; NAF90 en 1990 et 1999 ; NAF2003 EAR de 2004 à 2008 et NAF2008 EAR à partir de 2009.*
*** Informations disponibles pour seulement le ¼ des individus de l'échantillon initial à 4 jours, soit 1/16 des individus de l'échantillon actuel.*
(r) disponible uniquement en information redressée

Informations issues du Panel salariés

Données sur l'activité salariée issues des panels de salariés de l'Insee, constitués à partir des déclarations annuelles de données sociales (DADS) et des fichiers de paye des agents de l'état (FPE).

Libellé	Modalités
Identifiant de diffusion de l'individu EDP	
Année d'activité de l'individu	
Année d'entrée de l'individu dans le panel	
Nombre d'entreprises dans lesquelles l'individu a travaillé durant l'année	
Département et commune de résidence du salarié	Code officiel géographique
Département et commune de travail du salarié	
Type de contrat de travail depuis 2005	01 CDI ; 02 CDD ; 03 Contrat travail temporaire…
Catégorie sociale et profession de l'individu	Nomenclature PCS (2 digits) et Code PCS (4 digits) depuis 1993
Domaine d'emploi depuis 1986	1 - Fonction publique d'État 2 – FP territoriale 3
Catégorie, statut, grade et qualification des agents de la FP d'État	
Salaire net en euros constants de l'individu EDP durant l'année	
Durée de paie, entre 1 et 360, de l'individu EDP durant l'année	En jours + conversion équivalent « temps complet/plein salaire »
Nombre d'heures rémunérées de l'individu durant l'année depuis 1993	
Temps de travail et condition d'emploi de l'individu	Temps complet/partiel, Intérim, travail à domicile, Indemnités chômage
Activité économique agrégée jusqu'en 1992 et à partir de 1994	38 niveaux
Activité économique de l'établissement sauf en 1993	NES 5
Activité économique de l'entreprise (jusqu'en 2008)	NAP73 (jusqu'en 1993) ; NAF (de 1993 à 2003) ; NAF rév. 1 (> 2003)
Activité économique de l'entreprise et de l'établissement (depuis 2008)	NAF rév. 2 (>= 2008)
Nombre de salariés dans l'entreprise et dans l'établissement au 31/12	
Salaire net et brut	
Date de début et de fin de rémunération	

Informations issues des bulletins d'état civil

Libellé	Modalités
Bulletin naissance d'enfant (descendance de l'individu EDP)	
Rôle de l'individu EDP : mère ou père	MÈRE - Mère de l'enfant ; PÈRE - Père de l'enfant
Date de naissance de l'enfant	
Date et lieu de naissance des parents	Département-Commune, Indication du lieu de naissance (France/Étranger)
Nationalité des parents	
Situation matrimoniale des parents	
Catégorie socioprofessionnelle du père et de la mère	Nomenclature des professions et catégories socioprofessionnelles
Bulletin de décès	
Date et lieu de naissance	Département – Commune
Lieu de décès de l'individu EDP	Département – Commune
Nationalité	
Catégorie socioprofessionnelle de l'individu	Au moment du décès ; Nomenclature des professions et catégories socioprofessionnelles
Bulletin de Mariage	
Date et lieu de naissance des époux	Département – Commune
Nationalité des époux	
Catégorie socioprofessionnelle des époux	Nomenclature des professions et catégories socioprofessionnelles

Informations issues des fichiers fiscaux

Informations liées au ménage fiscal de l'individu	Informations liées à l'individu
Année de déclaration	Situation conjugale
Année de perception des revenus	Type de revenu principal : Inconnu ; inférieur 2 500 € annuels ; Salaire ; Chômage/préretraite ; Retraite ; pensions alimentaires ; Bénéfices agricoles ; BIC ; BNC
Niveau de vie du ménage et centile de niveau de vie	Pensions alimentaires perçues
Revenu disponible	Type et montants des différentes sources de revenus : Bénéfices industriels et commerciaux (BIC) ; Bénéfices non commerciaux (BNC) ; Bénéfices agricoles ; Allocation-chômage ; Retraite et pensions ; Salaires

Annexe 2. Nombre d'événements attendus et précision par secteur d'activité selon différentes incidences (Période 2013-2017)

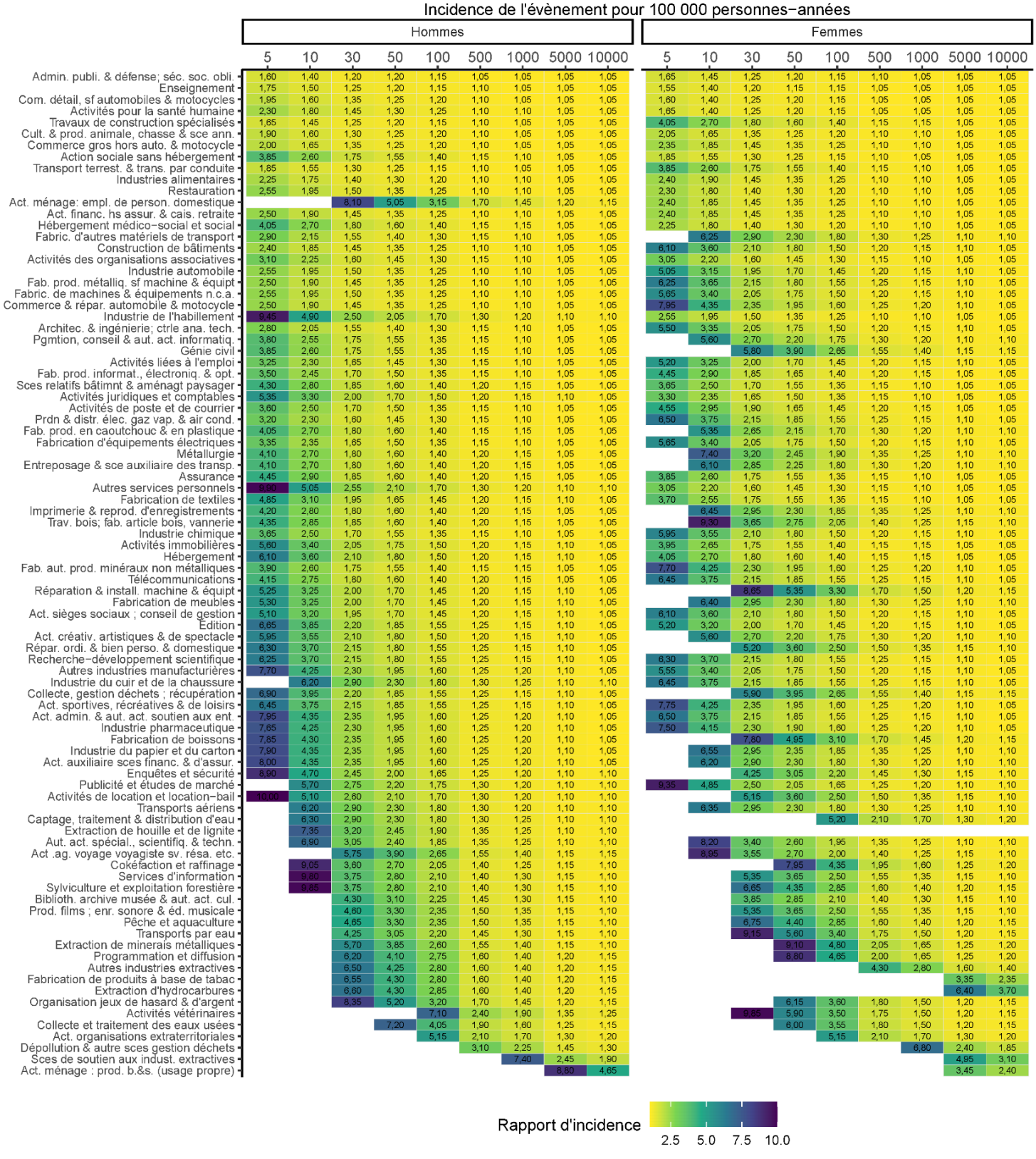
Incidence de l'évènement pour 100 000 personnes−années

	Hommes									Femmes								
	5	10	30	50	100	500	1 000	5 000	10 000	5	10	30	50	100	500	1 000	5 000	10 000
Admin. publi. & défense; séc. soc. obli.	30	60	179	298	595	2977	5953	29766	59532	25	50	150	250	500	2498	4996	24981	49962
Enseignement	20	40	121	201	402	2012	4025	20123	40246	31	63	188	313	627	3133	6266	31331	62662
Com. détail, sf automobiles & motocycles	14	28	83	138	276	1380	2760	13802	27604	29	57	172	287	573	2866	5733	28663	57326
Activités pour la santé humaine	9	17	52	87	174	869	1738	8692	17385	26	53	159	265	530	2648	5297	26483	52965
Travaux de construction spécialisés	25	49	147	245	491	2453	4905	24527	49054	3	6	19	31	62	310	621	3105	6209
Cult. & prod. animale, chasse & sce ann.	15	29	88	146	292	1460	2919	14596	29193	12	24	73	121	242	1211	2422	12111	24222
Commerce gros hors auto. & motocycle	13	26	79	131	263	1313	2627	13134	26267	9	17	51	86	171	857	1714	8568	17136
Action sociale sans hébergement	3	7	20	33	66	331	662	3308	6616	16	33	98	164	328	1640	3280	16402	32804
Transport terrest. & trans. par conduite	16	32	95	159	317	1587	3173	15867	31735	3	7	20	33	66	328	656	3278	6555
Industries alimentaires	9	19	57	95	190	948	1896	9482	18964	8	16	47	78	156	780	1560	7802	15604
Restauration	7	14	42	69	139	693	1385	6927	13855	9	18	54	90	180	902	1805	9024	18048
Act. financ. hs assur. & cais. retraite	7	15	44	73	146	730	1459	7296	14593	8	16	48	80	160	800	1601	8003	16006
Hébergement médico−social et social	3	6	18	31	61	306	611	3055	6110	9	18	55	92	185	923	1847	9233	18466
Construction de bâtiments	8	16	48	80	160	800	1600	8001	16002	2	4	11	18	36	181	362	1808	3616
Activités des organisations associatives	5	9	28	47	94	469	939	4693	9386	5	10	29	48	97	485	970	4849	9698
Fab. prod. métalliq. sf machine & équipt	7	15	44	73	146	732	1464	7320	14640	2	4	11	18	36	179	358	1789	3578
Industrie automobile	7	14	41	68	136	681	1362	6812	13623	2	5	14	23	45	226	452	2261	4522
Fabric. de machines & équipements n.c.a.	7	14	41	69	137	687	1373	6866	13732	2	4	12	20	40	198	397	1984	3967
Commerce & répar. automobile & motocycle	7	14	43	72	145	725	1449	7246	14492	1	3	8	14	28	141	282	1409	2819
Act. ménage: empl. de person. domestique	0	0	1	2	4	22	45	224	448	8	16	47	78	157	783	1567	7833	15665
Industrie de l'habillement	1	2	7	12	23	116	232	1162	2324	7	14	41	69	138	688	1376	6882	13764
Architec. & ingénierie; ctrle ana. tech.	6	12	35	58	116	580	1160	5801	11602	2	4	12	21	42	208	416	2082	4163
Activités liées à l'emploi	4	9	27	44	89	443	887	4435	8870	2	4	13	22	45	224	448	2239	4477
Fab. prod. informat., électroniq. & opt.	4	8	23	38	76	378	756	3782	7565	3	5	16	27	53	267	534	2672	5344
Sces relatifs bâtimnt & aménagt paysager	3	6	17	28	57	283	566	2831	5662	4	7	22	36	72	359	718	3592	7184
Activités juridiques et comptables	2	4	13	21	43	213	427	2135	4269	4	9	26	43	85	426	852	4258	8516
Activités de poste et de courrier	4	7	22	36	73	363	726	3630	7260	3	5	16	26	52	261	523	2614	5228
Prdn & distr. élec. gaz vap. & air cond.	4	9	27	44	89	444	889	4444	8888	2	3	10	17	34	171	343	1714	3428
Fabric. d'autres matériels de transport	5	10	31	52	104	520	1041	5203	10407	1	2	5	9	18	89	178	890	1780
Autres services personnels	1	2	7	11	23	115	229	1145	2290	5	10	29	49	98	488	977	4884	9768
Assurance	3	5	16	27	53	267	535	2673	5347	3	7	20	33	67	333	667	3335	6669
Fabrication d'équipements électriques	4	8	24	40	80	401	803	4015	8030	2	4	12	20	40	198	397	1984	3969
Fabrication de textiles	2	5	14	23	47	234	469	2343	4685	3	7	21	35	69	347	693	3465	6930
Industrie chimique	4	7	21	35	71	353	705	3527	7055	2	4	11	19	37	187	375	1874	3748
Activités immobilières	2	4	12	20	40	202	404	2020	4040	3	6	19	32	64	321	641	3206	6411
Hébergement	2	4	11	18	37	183	366	1829	3657	3	6	19	31	62	309	619	3093	6185
Télécommunications	3	6	18	30	59	296	591	2955	5910	2	3	10	17	34	172	343	1717	3434
Fab. aut. prod. minéraux non métalliques	3	6	19	32	65	323	647	3234	6467	1	3	9	14	28	142	284	1420	2839
Pgmtion, conseil & aut. act. informatiq.	3	7	21	34	69	344	688	3439	6879	1	2	6	10	21	103	207	1033	2065
Act. sièges sociaux ; conseil de gestion	2	5	14	23	46	228	456	2280	4560	2	4	11	19	37	187	373	1866	3732
Fab. prod. en caoutchouc & en plastique	3	6	18	31	61	305	611	3054	6109	1	2	6	11	21	107	214	1072	2144
Entreposage & sce auxiliaire des transp.	3	6	18	30	61	305	610	3048	6096	1	2	6	9	19	94	187	936	1872
Edition	2	3	10	17	33	166	333	1663	3326	2	4	13	22	44	220	440	2199	4398
Métallurgie	3	6	18	30	60	299	599	2993	5987	1	1	4	7	15	74	148	739	1477
Imprimerie & reprod. d'enregistrements	3	6	17	28	57	284	569	2844	5689	1	2	5	9	17	86	171	856	1712
Génie civil	3	7	20	33	66	329	659	3293	6585	0	1	2	3	7	33	65	326	652
Recherche−développement scientifique	2	4	11	18	35	177	354	1768	3536	2	4	11	18	35	177	354	1769	3538
Autres industries manufacturières	1	3	9	14	29	143	286	1429	2859	2	4	12	20	40	201	402	2012	4023
Trav. bois; fab. article bois, vannerie	3	6	17	28	55	276	553	2763	5526	1	1	4	6	12	60	120	600	1199
Act. sportives, récréatives & de loisirs	2	3	10	17	34	172	344	1718	3436	1	3	9	14	29	144	288	1439	2879
Act. admin. & aut. act. soutien aux ent.	1	3	8	14	28	140	279	1397	2794	2	3	10	17	35	174	348	1738	3477
Fabrication de meubles	2	4	13	21	43	214	428	2138	4275	1	2	5	9	17	86	172	860	1719
Act. créativ. artistiques & de spectacle	2	4	11	19	38	188	375	1876	3751	1	2	6	10	20	102	204	1020	2041
Industrie pharmaceutique	1	3	9	14	28	142	285	1423	2845	1	3	9	15	29	147	294	1471	2942
Industrie du cuir et de la chaussure	1	2	5	9	18	89	178	889	1777	2	3	10	17	34	170	341	1704	3408
Réparation & install. machine & équipt	2	4	13	22	44	220	441	2203	4406	0	0	1	2	4	22	44	218	437
Act. auxiliaire sces financ. & d'assur.	1	3	8	14	27	137	274	1372	2743	1	2	6	9	18	92	185	924	1849
Industrie du papier et du carton	1	3	8	14	28	139	277	1385	2770	1	2	5	8	17	84	169	845	1689
Publicité et études de marché	1	2	6	10	20	99	197	986	1973	1	2	7	12	24	120	241	1205	2410
Répar. ordi. & bien perso. & domestique	2	3	10	17	35	175	349	1746	3491	0	1	2	4	7	37	73	365	730
Collecte, gestion déchets ; récupération	2	3	10	16	32	159	319	1594	3187	0	1	2	3	6	32	64	320	640
Transports aériens	1	2	5	9	18	90	180	898	1795	1	2	5	9	17	87	175	874	1749
Enquêtes et sécurité	1	3	8	13	25	127	253	1266	2532	0	1	3	5	10	48	96	481	963
Fabrication de boissons	1	3	8	14	28	138	276	1379	2758	0	0	1	2	5	24	47	236	473
Activités de location et location−bail	1	2	7	11	22	112	224	1119	2239	0	1	2	4	8	38	76	380	759
Aut. act. spécial., scientifiq. & techn.	1	2	5	8	16	80	161	803	1606	1	1	4	7	14	68	136	682	1365
Biblioth. archive musée & aut. act. cul.	0	1	3	5	9	47	93	466	932	1	1	3	6	11	56	111	556	1113
Captage, traitement & distribution d'eau	1	2	5	9	17	87	174	872	1744	0	0	1	1	2	11	23	113	226
Act .ag. voyage voyagiste sv. résa. etc.	0	1	2	3	6	32	65	324	649	1	1	4	6	13	63	125	627	1253
Services d'information	1	1	3	6	11	57	114	570	1141	0	1	2	4	7	36	72	359	718
Sylviculture et exploitation forestière	1	1	3	6	11	57	115	575	1150	0	1	2	3	5	27	55	274	548
Prod. films ; enr. sonore & éd. musicale	0	1	3	4	9	44	87	436	871	0	1	2	4	7	37	73	366	731
Cokéfaction et raffinage	1	1	4	6	12	61	121	607	1215	0	0	1	1	3	14	27	137	274
Extraction de houille et de lignite	1	1	4	7	15	73	146	732	1464	0	0	0	0	0	0	0	0	0
Pêche et aquaculture	0	1	3	4	8	42	84	418	837	0	1	2	3	5	27	54	270	540
Transports par eau	0	1	3	5	10	48	95	477	955	0	0	1	2	4	20	41	203	405
Extraction de minerais métalliques	0	1	2	3	6	32	64	321	643	0	0	1	1	2	12	24	120	241
Programmation et diffusion	0	1	2	3	6	30	60	298	597	0	0	1	1	3	13	26	129	259
Organisation jeux de hasard & d'argent	0	0	1	2	4	22	44	218	436	0	0	1	2	4	18	36	181	362
Collecte et traitement des eaux usées	0	0	1	2	3	16	31	156	312	0	0	1	2	4	19	37	186	372
Autres industries extractives	0	1	2	3	6	28	57	283	565	0	0	0	0	1	3	6	29	58
Fabrication de produits à base de tabac	0	1	2	3	5	27	55	273	546	0	0	0	0	0	0	1	4	8
Activités vétérinaires	0	0	0	1	2	8	16	79	158	0	0	1	2	4	20	39	195	391
Extraction d'hydrocarbures	0	1	2	3	5	27	54	272	544	0	0	0	0	0	0	0	2	4
Act. organisations extraterritoriales	0	0	1	1	2	11	23	113	226	0	0	1	1	2	11	23	114	228
Dépollution & autre sces gestion déchets	0	0	0	0	1	5	10	49	97	0	0	0	0	0	1	2	8	17
Sces de soutien aux indust. extractives	0	0	0	0	0	1	2	8	15	0	0	0	0	0	0	0	2	5
Act. ménage : prod. b.&s. (usage propre)	0	0	0	0	0	0	0	1	3	0	0	0	0	0	0	1	4	8

Précision : Limitée / Acceptable

Précision considérée comme limitée quand il est attendu que le rapport des bornes de l'intervalle de confiance à 95% soit supérieur à 2

Annexe 3. Plus petite augmentation d'incidence pouvant être mise en évidence par secteur d'activité selon différentes incidences (période 2008-2017)



Les secteurs d'activité sans valeur sont ceux où la plus petite augmentation pouvant être mis en évidence est supérieure à 10.

Annexe 4 : Précisions sur la méthodologie

Définition des incidences

Pour les différents événements de santé étudiés, le nombre survenu sur l'ensemble de la période de l'étude sera calculé. Une période de suivi sera calculée à partir du 01/01/2008 ou de la date de la première date d'intégration dans l'EDP si postérieure à 2008 et en prenant comme date de fin de suivi, la date de l'événement si non récurrent ou la date de décès, la date de sortie de l'échantillon ou le 31/12/2022 pour les personnes encore vivantes à cette date. Les taux d'incidence des différents événements de santé seront décrits globalement, selon les caractéristiques des travailleurs et au cours du temps. Pour les événements récurrents les taux d'incidence seront calculés en divisant le nombre total d'événements observés par la durée de suivi [38].

Calcul des espérances de vie

Pour le calcul de l'espérance de vie une collaboration sera réalisée avec l'Ined (Emmanuelle Cambois) pour utiliser leur méthode développée pour estimer des espérances de vie par catégorie socio-professionnelle. Cette méthode repose sur le lissage des données de mortalité observées à l'aide de P-splines [31]. Cette approche adaptée pour les petites populations permet de calculer des intervalles de confiance pour la mortalité et l'espérance de vie, sans avoir recours aux techniques de *bootstrap* ou de simulation. Cette méthode sera appliquée à partir de la mortalité observée sur la période la plus récente des données selon le secteur d'activité le plus longtemps exercé.

Comparaison des incidences par secteur

Les taux d'incidence des différents événements de santé seront décrits selon les deux variables relatives au secteur d'activité de travail (avoir travaillé dans le secteur ; le secteur le plus longtemps) pour la période la plus récente disponible. Des indices comparatifs d'incidence (SIR) seront calculés par secteur d'activité (selon les deux définitions) par standardisation indirecte sur l'âge et le sexe en divisant les effectifs d'événement de santé observés sur la période la plus récente aux nombres attendus calculés en appliquant les taux d'incidence observés de chaque événement de santé sur la même période parmi l'ensemble des travailleurs au nombre de personnes-années observées dans chaque secteur. L'intervalle de confiance à 95 % du SIR sera calculé en supposant pour les effectifs observés d'événements de santé une distribution de Poisson ou une distribution binomiale négative en cas de surdispersion des données (notamment pour les événements récurrents) [39, 40].

Comparaison temporelle des incidences.

Dans chaque secteur d'activité le plus longtemps exercé, les taux d'incidence par période de cinq ans seront tout d'abord représentés graphiquement. Puis l'évolution temporelle de chaque événement de santé sera quantifiée à l'aide d'une régression de Poisson (ou binomiale négative) modélisant les taux incidences standardisés sur l'âge et incluant une variable période, une variable secteur le plus longtemps exercé et leur interaction. Ce modèle sera utilisé pour tester si l'effet période pour un secteur donné est statistiquement différent de l'effet période dans son ensemble.

Estimation du risque d'événement de santé par secteur

Pour comparer la survenue des événements de santé à cinq ou dix ans selon le secteur d'activité exercé, pour chaque secteur d'activité, les travailleurs exerçant dans ce secteur d'activité en 2009 (groupe exposé) seront appariés aléatoirement à 5 référents chacun (groupe non exposé) travaillant dans un autre secteur mais similaires en termes de sexe, d'âge (+/- 2 ans), de région de résidence, de diplôme, de niveau de revenus et de remboursements de soins l'année précédente. Selon l'événement de santé étudié, son histoire naturelle et ses associations possibles avec les conditions de travail dans le secteur étudié, le suivi sera censuré chez les travailleurs exposés soit au moment où la personne quitte le secteur d'activité (par exemple pour les accidents), soit un, trois ou cinq ans après (une durée plus longue ne pouvant pas être envisagée compte tenu de la profondeur disponible des données de santé). La fréquence de survenue des événements de santé dans les deux groupes (exposés vs non exposés) sera comparée, pour chaque secteur d'activité, à l'aide d'un modèle de survie pour les événements de santé non récurrents en utilisant l'âge comme axe du temps. Pour les événements récurrents, un modèle des taux marginaux (*marginal rates model*) prenant en compte, le cas échéant, un événement terminal, sera utilisé [41]. Les modèles seront ajustés à l'aide de variables

dépendantes du temps sur la période calendaire et la région de résidence. Les phénomènes d'exclusion au travail et du secteur d'activité étudié seront pris en compte à l'aide de la formule g paramétrique [32]. Pour tenir compte des différences de comportement de santé, notamment tabagisme et consommation d'alcool, observées entre secteurs d'activité [10], mais pour lesquels nous ne disposons pas de données dans l'étude, des analyses de biais pour ces facteurs de confusion non mesurées seront réalisées [34]. Dans le cas d'événement de santé où un risque augmenté sera estimé, des analyses de médiation utilisant l'approche contrefactuelle avec imputation [42] seront conduites afin d'identifier des caractéristiques (revenus, historique de carrière, accès au soin, etc.) médiant l'association observée.

Quantification de l'importance des secteurs

Pour évaluer le rôle des secteurs d'activité dans les inégalités d'état de santé observées dans un groupe de travailleurs spécifiques, une caractéristique des travailleurs sera utilisée comme critère ***a priori*** pour définir le groupe de travailleurs spécifiques ; par exemple le fait d'être un travailleur sénior c'est-à-dire d'avoir 55 ans ou plus [43]. L'importance des secteurs d'activité pour comprendre la survenue des événements de santé dans cette population spécifique sera évaluée en calculant le coefficient de corrélation intra-classes dans un modèle multiniveaux utilisant le niveau secteur comme variable aléatoire [35].

Annexe 5 : Notice d'information collective sur l'utilisation des données pour le projet SEESTA - Suivi épidémiologique de l'état de santé des travailleurs en France selon l'activité professionnelle

Responsable du traitement : Santé publique France

Contexte

On estime qu'environ le tiers des différences sociales de mortalité par cancer dans les pays industrialisés (différences qui sont très fortes, en Europe et en France en particulier), est expliqué par l'exposition à des facteurs d'origine professionnelle. Ainsi, de très nombreux problèmes de santé (cancers, pathologies musculo-squelettiques, respiratoires, cardio-vasculaires, neurologiques, troubles mentaux) trouvent tout ou partie de leur origine dans l'environnement professionnel, à travers l'exposition à des nuisances physiques, chimiques mais également aux facteurs psychosociaux en lien avec l'organisation du travail.

Afin de définir les politiques de prévention adaptées, les appliquer et les évaluer, il est important de disposer d'un outil de surveillance épidémiologique permettant de décrire dans le temps la fréquence de survenue de maladie et les causes médicales de décès en fonction des caractéristiques professionnelles.

L'utilisation et le croisement de sources de données historiques existantes, relatives aux facteurs professionnels et aux événements de santé, régulièrement mises à jour et disponibles à l'échelle nationale présente de nombreux atouts dans le domaine de la surveillance des risques professionnels.

Objectifs

Notre projet d'exploitation s'inscrit dans la suite de nos premiers travaux sur la mortalité par cause (Programme Cosmop). Ses objectifs principaux sont d'identifier si certains secteurs d'activité ou typologies de carrières professionnelles sont caractérisés par des risques plus élevés d'événements de santé, et d'évaluer le rôle des secteurs d'activité exercés dans les inégalités de l'état de santé observées dans des groupes de travailleurs particuliers.

Les résultats obtenus et l'évolution dans le temps des indicateurs épidémiologiques doivent contribuer à repérer et surveiller des situations à risque, à alerter sur l'apparition de nouveaux facteurs de risque potentiels d'origine professionnelle et/ou partagés par des groupes professionnels afin de prioriser les actions de prévention et de promotion de la santé pertinentes.

L'intérêt public de cette étude, sa qualité scientifique et sa pertinence éthique ont été confirmés par un comité éthique et scientifique pour les recherches, les études et les évaluations dans le domaine de la santé (Cesrees), indépendant du responsable de traitement. Il a également reçu un avis favorable du Comité du secret statistique.

Finalité et base juridique du traitement

Dans le cadre de cette recherche, un traitement informatique de vos données personnelles va être effectué pour répondre à ces objectifs. Le responsable du traitement des données, qui est le gestionnaire de l'étude, est Santé publique France dont les coordonnées figurent en dernière page de ce document.

Le traitement de ces données, qui a pour finalité d'évaluer le rôle des secteurs d'activité exercés dans les inégalités de l'état de santé observées dans des groupes de travailleurs particuliers, est conforme au Règlement général européen sur la protection des données 2016/679 (RGPD) et à la loi du 6 janvier 1978 modifiée relative à l'informatique, aux fichiers et aux libertés (Loi Informatique et liberté).

Le traitement de vos données personnelles est fondé sur la mission d'intérêt public dont est investie, Santé publique France, responsable de traitement (article 6.1.e du RGPD – licéité du traitement) et la dérogation de traiter des données de santé à des fins de recherche scientifique (article 9.2.j du RGPD – exception permettant de traiter des données de santé).

Utilisation des données

Les données seront accessibles exclusivement par les membres de l'équipe projet de Santé publique France habilités, ayant signé un engagement de confidentialité. Aucune information individuelle ne pourra être communiquée en dehors de l'équipe. Par ailleurs les résultats produits auront un caractère collectif ne permettant pas d'identifier les individus.

Les données seront conservées pendant le nombre d'années autorisé pour la réalisation du projet, soit quatre ans à compter de la mise à disposition des données, puis elles feront l'objet d'un archivage intermédiaire en étant consultables jusqu'au 19 mars 2030, le temps nécessaire à la valorisation scientifique complète des travaux.

Personnes concernées

Le projet s'appuie sur un appariement entre les données socio-économiques des individus présents dans l'Échantillon démographique permanent (EDP) de l'Insee et leurs données de soins et d'hospitalisation issues Système national des données de santé (SNDS). L'échantillon obtenu appelé EDP-Santé a été constitué par la Direction de la recherche, des études, de l'évaluation et des statistiques (Drees).

L'EDP-Santé ne contient pas d'information permettant l'identification des personnes. Les données sont mises à disposition et utilisées par les personnels habilités de Santé publique France dans un espace sécurisé. Les données seront conservées pendant la durée du projet, jusqu'en 2028, puis archivées jusqu'à publication des travaux.

Conformément à l'article 66 de la loi du 6 janvier 1978 relative à l'informatique, aux fichiers et aux libertés, ce projet de recherche a fait l'objet d'une autorisation de la Commission nationale de l'informatique et des libertés (Cnil) en date du 2 août 2024 par Décision DR 2024-125 autorisant Santé publique France à réaliser un traitement de données ayant pour finalité une étude portant sur la surveillance épidémiologique de l'état de santé des travailleurs et travailleuses en France selon l'activité professionnelle, intitulée « SEESTA » (demande d'autorisation n° 924113) et complétée par la décision DR 2024-264 en date du 25 octobre 2024 (demande d'autorisation n° 924113v1). La Cnil est l'autorité de contrôle chargée de surveiller l'application des règles relatives à la protection des données, afin de protéger les libertés et droits fondamentaux des personnes physiques à l'égard d'un traitement de données.

Vos droits si vous êtes concerné(e) par ce traitement de données

En application des dispositions de la loi informatique et libertés modifiée, ainsi que du RGPD, vous disposez à tout moment d'un droit d'accès et de rectification des données vous concernant et d'un droit d'opposition au traitement de ces données.

Pour exercer ces droits ou en savoir plus sur les modalités d'exercice de ces droits, vous pouvez vous adresser au délégué à la protection des données de Santé publique France :

Par courriel : dpo@santepubliquefrance.fr (en indiquant le code EDP-Santé dans votre demande) ou par courrier à Santé publique France, déléguée à la protection des données, 12 rue du Val d'Osne, 94415 Saint Maurice Cedex

Par ailleurs, vous disposez d'un droit d'introduire une réclamation auprès d'une autorité de contrôle, en particulier auprès de la Cnil si vous considérez que le traitement de données à caractère personnel vous concernant constitue une violation du règlement général sur la protection des données et de la loi informatique et libertés https://www.cnil.fr/fr/adresser-une-plainte .

À l'issue de l'étude, et si vous le souhaitez, vous pourrez être informé(e) des résultats globaux qui seront référencés sur la documentation ouverte du *Health Data Hub* (HDH) et sur le site de Santé publique France.

Pour toute autre question concernant l'utilisation des données de l'EDP-Santé dans le cadre de ce projet, merci de les adresser à seesta@santepubliquefrance.fr.