RÉPUBLIQUE FRANÇAISE
Liberté
Égalité
Fraternité

Drees
DIRECTION DE LA RECHERCHE, DES ÉTUDES, DE L'ÉVALUATION ET DES STATISTIQUES

LES DOSSIERS DE LA DREES

n° 66 • septembre 2020

L'EDP-Santé

un appariement des données socio-économiques de l'échantillon démographique permanent au Système national des données de santé

Claire-Lise Dubost, Aude Leduc (DREES)

STATISTIQUE PUBLIQUE
La DREES fait partie du Service statistique public piloté par l'Insee.

L'EDP-Santé

un appariement des données socio-économiques de l'échantillon démographique permanent au Système national des données de santé

Claire-Lise Dubost, Aude Leduc (DREES)

Remerciements : Samuel Allain, Muriel Barlet, Christine de Peretti, Matthieu Doutreligne, Sébastien Durier, Natacha Gualbert, Javier Nicolau, Philippe Raynaud

Sommaire

SYNTHÈSE

Si le système national des données de santé apporte une grande quantité d'informations sur la consommation et les parcours de soins des bénéficiaires de l'Assurance maladie, son contenu est très limité pour caractériser la situation socio-économique de ces personnes. Dans l'optique d'évaluer la Stratégie nationale de santé 2018-2022, et en particulier son objectif de réduire les inégalités sociales de santé, la DREES a donc piloté un projet visant à enrichir le SNDS de ce type de données en appariant l'échantillon démographique permanent de l'Insee au SNDS. La base de données résultant de cet appariement, appelée l'EDP-Santé, doit permettre de réaliser des études longitudinales sur les parcours de soins entre 2008 et 2022, à partir d'un échantillon d'un peu plus de 3 millions de personnes, qu'il est possible de croiser avec des données issues des fichiers fiscaux, du recensement (exhaustif, puis par enquête), du panel d'actifs tous salariés de l'Insee, du fichier électoral et de l'État civil.

Ce document présente les contours de ce projet d'appariement, sa finalité et sa place dans le paysage des données de santé (partie 1), ses enjeux juridiques et les étapes de sa conception (partie 2), et enfin sa mise en œuvre ainsi que quelques premiers résultats méthodologiques sur sa qualité et ses perspectives d'utilisation (partie 3).

INTRODUCTION

La DREES a pour mission de fournir aux décideurs publics, aux citoyens, et aux responsables économiques et sociaux des informations fiables et des analyses sur les populations et les politiques sanitaires et sociales. Cette mission porte notamment sur le domaine de la santé et de l'assurance maladie. Or, une des problématiques qui traverse les études dans ce domaine est celle de la nette persistance d'inégalités sociales et territoriales de santé : sur la période 2012-2016, l'écart d'espérance de vie à 35 ans entre les 5 % les plus aisés et les 5 % les plus pauvres est de 13 ans pour les hommes, et de 8 ans pour les femmes. Cet écart important est connu des pouvoirs publics (OMS, 2009, HCSP, 2009) mais se présente comme la conséquence de nombreux mécanismes à l'œuvre tout au long de la vie. Comprendre leurs rouages est donc nécessaire pour la mise en place et le ciblage de politiques publiques efficaces.

Il existe déjà une littérature riche sur les inégalités sociales de santé et la façon dont celles-ci se développent, depuis la grossesse de la mère (Vilain et al., 2013) et la petite enfance (Chardon et al., 2015) jusqu'à la prise en charge des problèmes de santé et l'accès aux soins (HCAAM, 2011). Ces études sont révélatrices de l'existence d'un gradient social, autrement dit une dégradation continue de la plupart des indicateurs de santé en allant des catégories les plus favorisées aux catégories les plus défavorisées. Mais la recherche doit continuer à se développer dans ce domaine pour combler les besoins de connaissance notamment sur la question des déterminants de ces inégalités sociales de santé permettant ainsi de définir des priorités d'action pour les politiques publiques (DREES, 2017).

En matière de politiques publiques, la Stratégie nationale de santé (SNS) 2018-2022, adoptée officiellement par le gouvernement à la fin de l'année 2017 constitue aujourd'hui le cadre de la politique de santé. Porteuse d'objectifs multiples, elle vise à répondre aux défis rencontrés par notre système de santé, notamment les risques sanitaires liés à l'augmentation prévisible de l'exposition aux polluants et aux toxiques, les risques d'exposition de la population aux infections, les maladies chroniques et leurs conséquences, ainsi que l'adaptation du système de santé aux enjeux démographiques, épidémiologiques, et sociétaux. Chacun de ces objectifs doit aussi contribuer à lutter contre l'ensemble des inégalités sociales et territoriales, qu'elles se traduisent par des écarts d'espérance de vie entre niveaux de vie par exemple, ou plus en amont par des écarts en matière d'états de santé, de recours à la prévention, ou aux soins, lorsqu'une pathologie est dépistée. En lien avec sa mission et les études déjà engagées sur ces thématiques, le suivi pluriannuel et le pilotage des évaluations de la SNS ont été confiés à la DREES par l'arrêté du 1er février 2018.

Pour contribuer à cette évaluation, la DREES a mis en œuvre un rapprochement inédit de sources de données administratives et ainsi constitué l'EDP-Santé, une base de données offrant un vaste potentiel pour étudier les inégalités sociales et territoriales de santé, dont les contours et enjeux sont présentés dans ce dossier.

La première partie présente les deux sources, l'EDP et le SNDS, et revient sur l'intérêt de les rapprocher pour répondre à de nombreuses questions et approfondir la recherche en économie et sociologie de la santé, bien que les utilisations de cette source ne se limitent pas à ces champs (l'épidémiologie, la santé publique mais également la recherche clinique pourront également bénéficier de ces nouvelles données). La seconde partie concerne les aspects techniques et juridiques du projet depuis son commencement et jusqu'à sa mise en œuvre, un processus complexe sur lequel il est intéressant de revenir, afin de faire bénéficier les futurs projets du même type de cette expérience. Enfin, la dernière partie présente les premiers aspects méthodologiques liés à l'appariement et à son exploitation, dans l'optique d'harmoniser les traitements futurs prévus sur les données.

APPARIER LES DONNÉES MÉDICO-ADMINISTRATIVES ET LES DONNÉES SOCIO-ÉCONOMIQUES DE L'EDP : UNE OPPORTUNITÉ POUR L'ÉTUDE DES INÉGALITÉS SOCIALES DE SANTÉ

L'échantillon démographique permanent : un panel phare de l'Insee

Construction de l'EDP

L'échantillon démographique permanent (EDP) est un panel socio-démographique de grande taille, créé en 1968 par l'Insee, afin d'étudier les parcours professionnels, résidentiels ou familiaux des personnes résidant en France métropolitaine.

Il rapproche les informations de différentes sources administratives et enquêtes pour des individus nés certains jours de l'année (4 jours par an jusqu'au début des années 2000, puis 16 jours par an, soit 4,4 % de la population), et constitue un échantillon représentatif. Enrichi chaque année par les nouvelles données à disposition dans chacune des sources, il a aussi vu sa forme évoluer, qu'il s'agisse de la taille de l'échantillon, qui a donc quadruplé en 2004, ou des sources de données collectées pour ces individus.

Stable depuis 2011, l'EDP est désormais le produit de l'appariement de données en provenance des cinq sources suivantes (Jugnot, 2014)[1] :

- les bulletins d'État civil de naissance, de mariage, de décès depuis 1968 ;
- les données issues des cinq recensements exhaustifs (1968, 1975, 1982, 1990, 1999) et des enquêtes annuelles de recensement (EAR) depuis 2004 ;
- le fichier électoral donnant les inscriptions électorales actuelles et passées depuis 1990 ;
- les informations issues du panel d'actifs « tous salariés » depuis 1967 : salaires, durée de paie, catégorie socioprofessionnelle, secteur d'activité…
- les données socio-fiscales issues de la base Fideli (Fichier démographique des logements et des individus) et de la base FiLoSoFi (Fichier localisé social et fsiscal) depuis 2011.

Dès lors qu'un individu né un des jours concernés par l'EDP apparaît dans l'une de ces sources, il est intégré au fichier. En 2017, la base études de l'EDP contient 3,7 millions d'individus. Le champ couvert concerne donc les personnes résidant, travaillant, ou connaissant un événement d'état-civil sur le territoire national. Les départements d'outre-mer (DOM) sont inclus dans les sources à partir des années 2000.

Le NIR (numéro d'inscription au répertoire national d'identification des personnes physiques) n'est pas restitué dans la base études de l'EDP, mais il est présent dans les bases de production de l'Insee. Il rend possible l'enrichissement des données avec le panel « tous salariés », les autres appariements étant faits sur l'état civil de la personne.

Informations contenues dans l'EDP

La base de l'EDP contient une table centrale sur les individus (tableau 1), avec des caractéristiques synthétiques, à savoir les informations démographiques issues de l'état civil et une variable pour chaque source de données permettant notamment de savoir si l'individu apparaît dans les tables provenant des différentes sources.

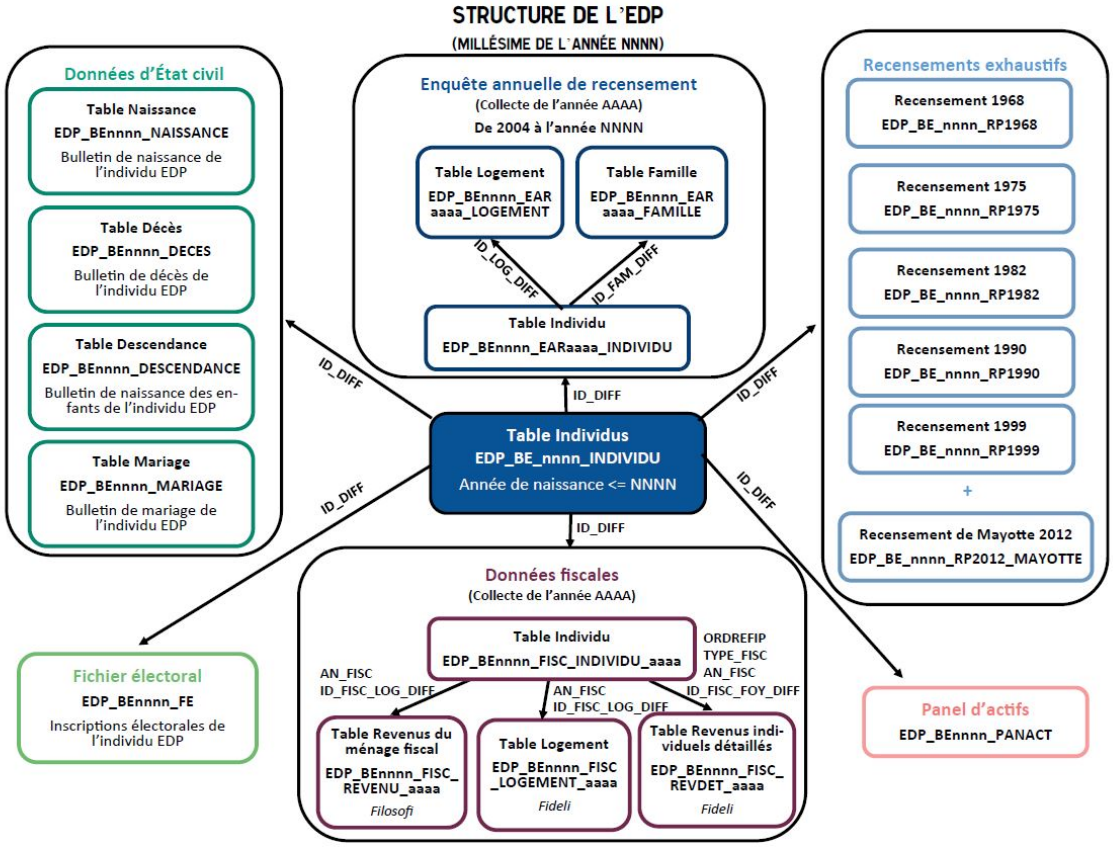
[1] Le document de travail de Stéphane Jugnot (INSEE Document de Travail N°F1406) donne plus de précisions sur les informations disponibles dans l'EDP. L'appariement avec les données fiscales et le panel « tous salariés » est cependant postérieur à la rédaction de ce document. Des compléments rapides sur ces sources récentes sont disponibles à partir de ce lien : https://www.insee.fr/fr/metadonnees/source/s1166.

Tableau 1 •Contenu de la table Individus de l'EDP

Source	Variable
Données d'état civil	Naissance Décès Nombre d'enfants repérés dans les actes d'état civil Nombre de mariages repérés dans les actes d'état civil
Recensements exhaustifs	Présence ou non à chacun des recensements exhaustifs
Enquêtes annuelles de recensement	Présence ou non dans chacune des enquêtes annuelles de recense-ment
Fichier électoral	Nombre d'inscriptions électorales
Panel d'actifs	Nombre d'années d'activité salariée
Données fiscales	Nombre de déclarations fiscales

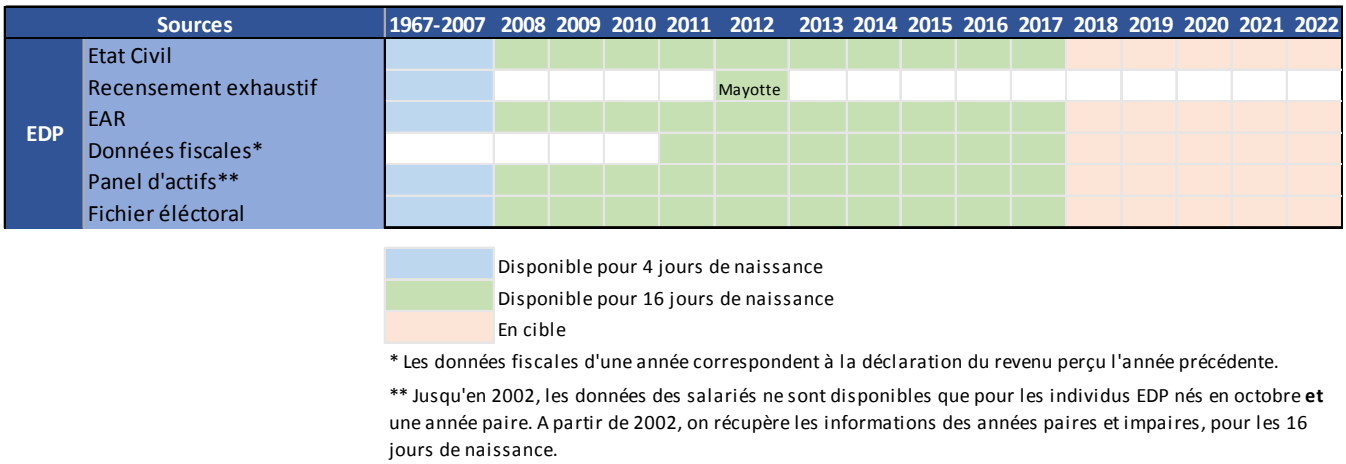
Chacun de ces événements, lorsqu'il existe, renvoie ensuite à une table plus détaillée à partir de la source concernée.

Figure 1 • Structure des tables de l'EDP



Le champ couvert par l'EDP peut néanmoins varier selon les sources et les années (l'EDP va par exemple évoluer avec le prélèvement à la source).

Figure 2 • Synthèse du champ couvert par chacune des sources de l'EDP



* Les données fiscales d'une année correspondent à la déclaration du revenu perçu l'année précédente.

** Jusqu'en 2002, les données des salariés ne sont disponibles que pour les individus EDP nés en octobre **et** une année paire. A partir de 2002, on récupère les informations des années paires et impaires, pour les 16 jours de naissance.

Les données de l'EDP permettent ainsi de faire des études sur la fécondité, la mortalité, les parcours familiaux, les migrations géographiques au sein du territoire national, la mobilité sociale et la mobilité professionnelle, les carrières salariales et les niveaux de vie ainsi que les interactions possibles entre ces différents aspects.

Une bibliographie recense les exploitations de ces données et figure dans le manuel décrivant la base étude. Un exemple de travaux notables menés par l'Insee à partir de cette source porte sur les inégalités sociales d'espérance de vie. D'abord calculées selon la catégorie sociale et le diplôme, elles ont récemment été évaluées sous l'angle du niveau de vie (Blanpain, 2018) et constituent un résultat de référence dans le champ de la santé publique.

Mise à disposition de l'EDP

La demande d'accès est adressée sur projet au Comité du secret statistique. Lorsque l'avis est favorable et après accord des Archives de France, la base de données est accessible au Centre d'accès sécurisé aux données (CASD) pour les chercheurs ou organismes extérieurs au service statistique public (SSP). En effet, l'article R135 D-1 du livre des données fiscales dispose que ces dernières, présentes dans l'EDP, ne peuvent être mises à disposition que via le CASD. Les services statistiques ministériels (SSM) et l'Insee peuvent néanmoins y accéder sans passer par cette infrastructure, après avis du Comité du secret.

Le Système national des données de santé : précision et richesse d'information sur les parcours de soins

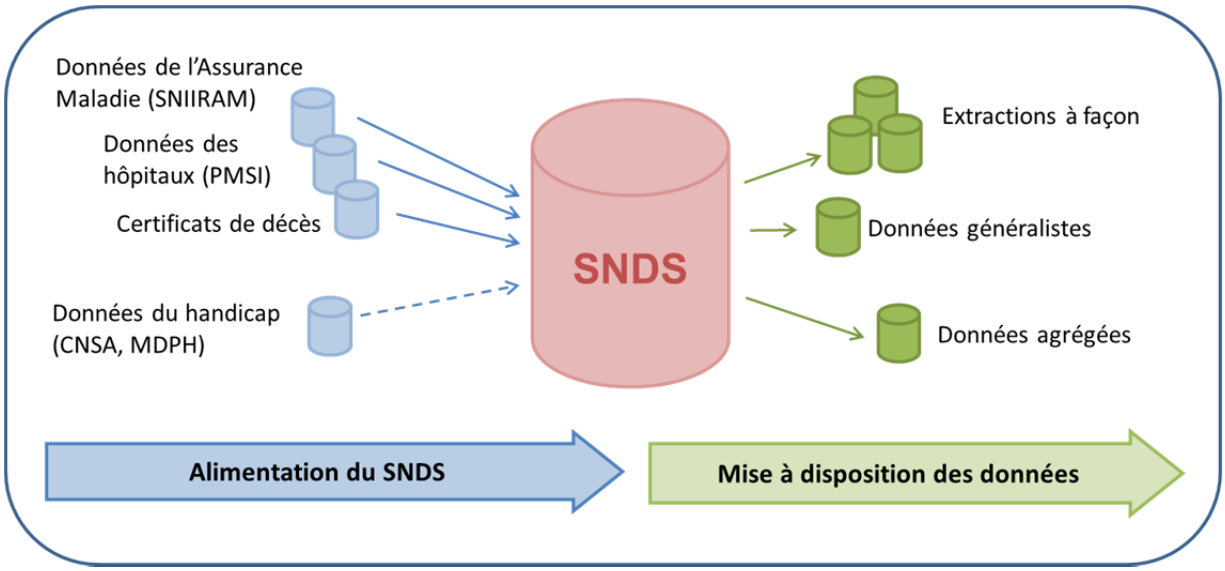
Construction du SNDS

Le Système national des données de santé (SNDS) est un entrepôt de données médico-administratives. Géré initialement par la Caisse nationale d'Assurance maladie (CNAM), il a été créé par l'article 193 de la loi de modernisation de notre système de santé en 2016. La loi du 24 juillet 2019 relative à l'organisation et à la transformation du système de santé a élargi le périmètre du SNDS et confié à la plateforme des données de santé (Health Data Hub) sa mise à disposition. Le SNDS contient désormais une base principale, seule source concernée par l'appariement EDP-Santé, et un catalogue de bases de données. La base principale contient:

- les données de l'Assurance maladie (consommations de soins en ville et en établissement remontées dans le SNIIRAM) ;
- les données hospitalières du PMSI ;
- les données sur les causes médicales de décès du CépiDC-Inserm.

Il est également prévu que la base principale du SNDS contienne, lorsqu'elles seront constituées, les données relatives au handicap en provenance des maisons départementales des personnes handicapées (MDPH).

Figure 3 • Construction de la base principale du SNDS



Le SNDS couvre donc l'ensemble des personnes ayant eu recours au système de soins français ou étant décédées sur le territoire. Il permet de reconstituer, de manière quasi-exhaustive, l'intégralité des parcours de soins, notamment grâce au chaînage des soins en ville et à l'hôpital (Tuppin et al., 2017).

Le NIR, numéro d'inscription au répertoire national d'identification des personnes physiques, est utilisé comme identifiant des personnes et donne la clé pour chaîner entre elles ces différentes sources de données. Comme dans l'EDP, le NIR n'est pas restitué dans les bases de données d'étude.

Informations contenues dans le SNDS

Le SNDS contient depuis 2006 des informations sur le recours aux soins et le montant des dépenses associées, les informations concernant les séjours hospitaliers. Les causes de décès des individus sont également chaînées avec ces données, mais elles ne sont pour l'instant intégrées que pour les années 2013 à 2015.

Plus précisément, on peut distinguer différentes catégories de données :

- 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, causes médicales de décès) ;
- Prestations, dépenses et remboursements (soins de ville, en établissements de santé, et montants associés) :
  - Prestations de soins de ville (consultations, visites…),
  - Prescriptions de médicaments,
  - Actes techniques,
  - Prélèvements biologiques,
  - Dispositifs médicaux (aides techniques),
  - Autres prestations (cures, transports…),
  - Séjours hospitaliers,
  - Soins hospitaliers (hors séances),
  - Indemnités journalières (maladie, accidents du travail et maladies professionnelles, maternité).
- Informations sur les professionnels de santé (spécialité) et les établissements de santé fréquentés par le bénéficiaire[2] ;

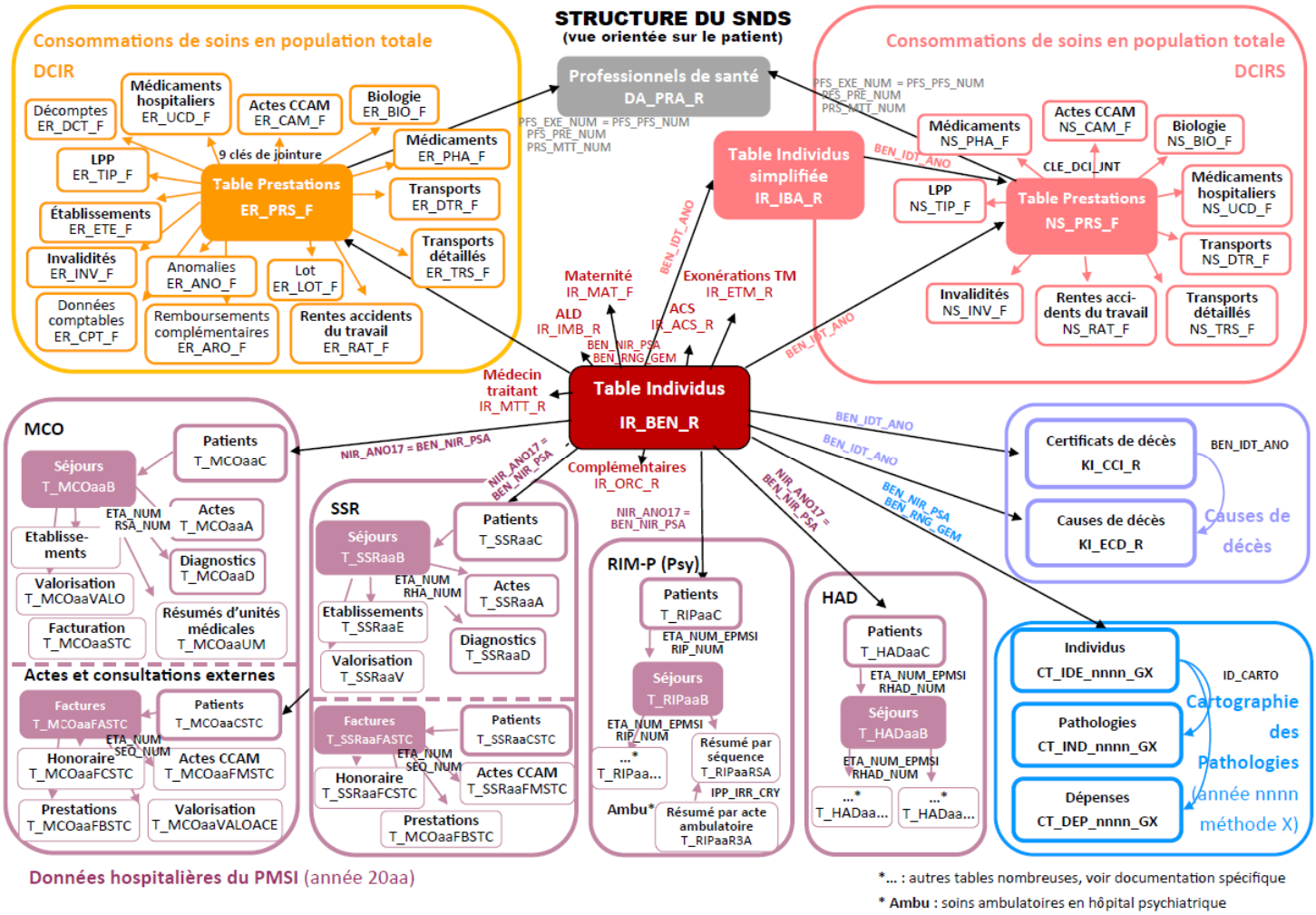
[2] Dans la première version de l'extraction de l'EDP-Santé, les données sur les professionnels de santé ne sont pas intégralement disponibles : seules les données présentes dans les tables de prestations sont restituées, mais celles issues du référentiel des professionnels de santé sont absentes.

- Informations sur l'état de santé du bénéficiaire : les affections de longue durée renseignent sur certaines pathologies chroniques des bénéficiaires, ainsi que les diagnostics hospitaliers, ou la consommation de médicaments ou d'actes très spécifiques à des pathologies.

L'approche de la santé à travers le SNDS est néanmoins cantonnée à l'étude du recours aux soins. Ainsi, les parcours de santé ne peuvent être analysés qu'à partir du moment où une prise en charge par le système de soins commence. En conséquence, le SNDS ne contient pas d'informations sur les personnes exclues du système de soins, ou les personnes dont la pathologie n'a pas encore été repérée. Il s'agit également de soins pris en charge par l'Assurance maladie obligatoire. Ils n'incluent donc pas les consommations sans prescription ou qui ne sont pas remboursées par l'Assurance maladie (automédication, consultations d'un psychologue ou d'un ostéopathe par exemple).

La structure du SNDS est similaire à celle de l'EDP : une table centrale contient les informations principales sur les bénéficiaires de l'Assurance maladie, puis elle peut être jointe à d'autres tables liées aux différentes sources présentes dans le SNDS (figure 4).

Figure 4 • Schéma de la structure du SNDS



Mise à disposition du SNDS

Depuis la loi de 2016, toute personne ou structure, publique ou privée, à but lucratif ou non lucratif, peut accéder aux données (ou à une partie des données) du SNDS pour un projet d'intérêt public à des fins de recherche, d'étude et d'évaluation dans le domaine de la santé. Cette finalité de recherche, étude ou évaluation devait être précise, ce qui empêchait la constitution d'entrepôts de données pour un ensemble de sujets de recherche à partir du SNDS. La loi de 2019 a donc supprimé la finalité de recherche, étude ou évaluation, notamment pour permettre la constitution d'entrepôts de données de santé comme l'EDP Santé. L'accès sur projet nécessite une

autorisation de la CNIL sauf dans des cas particuliers (méthodologies de référence ou procédures simplifiées, pour l'accès à l'échantillon généraliste de bénéficiaires par exemple[3]). Pour obtenir cette autorisation, il existe une procédure standard dont la plateforme des données de santé[4] est le guichet central.

Par ailleurs, un décret prévoit, pour certains organismes chargés d'une mission de service public, un accès permanent aux données du SNDS. Cet accès est défini pour un périmètre spécifique des données, dont les niveaux d'agrégation (données individuelles ou indicateurs agrégés, …) et la profondeur historique peuvent varier.

L'intérêt scientifique de rapprocher ces deux sources

Le projet de constitution de l'EDP-Santé mené par la DREES consiste à extraire les données de santé du SNDS sur la période 2008-2022 pour les individus présents dans l'EDP, et à mettre en regard l'ensemble de ces données. Le NIR, identifiant commun aux deux sources[5], constitue la clé pour rapprocher les données des deux systèmes d'information.

L'objectif poursuivi par cet appariement est d'obtenir une source inédite permettant de croiser des données socio-économiques et des données médico-administratives sur la santé et le recours aux soins, et ainsi de fournir un outil d'une grande richesse pour permettre l'évaluation de la stratégie nationale de santé 2018-2022.

Plus précisément, l'appariement vise à répondre à un certain nombre de besoins :

- Le besoin d'un échantillon de taille suffisante pour analyser des situations épidémiologiques relativement rares, et pour pouvoir notamment faire des recoupements de plusieurs variables sans dégrader la puissance statistique de l'analyse (par exemple pour une étude sur le suicide dans certaines professions) ;
- Le besoin d'un échantillon de taille suffisante également pour décliner les analyses à une échelle territoriale assez fine et pouvoir tenir compte de la situation en matière d'offre de soins ;
- Le besoin d'une profondeur historique pour être en mesure d'évaluer les tendances sur les dix dernières années, et d'éventuelles ruptures ou évolutions notables sur les années récentes, mais aussi pour disposer d'un certain recul dans les parcours de soins, afin notamment de reconstruire de manière fiable les pathologies chroniques et d'étudier leur évolution[6] ;
- Le besoin de données fiables et objectives sur des informations sur le recours aux soins, souvent entachées de biais de mémoire (notamment au vu de la profondeur historique souhaitée) ou de biais de désirabilité[7] lorsqu'elles sont collectées directement auprès des individus via des enquêtes.

Pour répondre à ces différents besoins, les données requises pour l'appariement concernent les trois bases du SNDS : SNIIRAM, PMSI et CépiDC. En effet, l'étude des parcours de soins au regard des caractéristiques sociales et économiques est au cœur des enjeux ciblés par la stratégie nationale de santé et requiert de disposer de données sur les soins réalisés en ville ainsi qu'à l'hôpital, de même que sur les causes du décès lorsque ce dernier ponctue le parcours.

---

[3] L'EGB est un échantillon permanent représentatif de la population protégée par l'Assurance maladie française. Il contient des informations anonymes sur les caractéristiques socio-démographiques (les mêmes que dans le SNDS, à savoir âge, sexe, lieu de résidence et repérage des bénéficiaires de droits pris en charge par l'Assurance maladie tels que la CMU-C) et médicales des bénéficiaires et les prestations qu'ils ont perçues.

[4] La Plateforme des Données de Santé (ou Health Data Hub) a remplacé l'Institut National des données de santé (INDS) le 1er décembre 2019. Ce Groupement d'Intérêt Public vise à construire une infrastructure qui facilite le croisement des bases de données de santé entre elles ou avec d'autres données. Les démarches juridiques nécessaires à la création de l'EDP-Santé ont néanmoins été menées avant la création de la PDS, et donc auprès de l'INDS.

[5] Plus exactement, le NIR dans le SNDS est un NIR pseudonymisé par un algorithme de hachage appelé FOIN. Voir partie sur le NIR.

[6] La CNAM met à disposition des algorithmes issus de la Cartographie des pathologies qui permettent d'identifier une cinquantaine de pathologies à partir du recours aux soins sur les cinq années précédentes. Dix ans de profondeur permettraient donc de reconstruire les pathologies sur cinq ans.

[7] Document de travail n°39, L'appariement handicap-santé et données de l'Assurance maladie, Alexis MONTAUT, Lucie CALVET, Gérard BOUVIER, Lucie GONZALEZ.

Les données demandées concernent la période 2008-2022 : la profondeur historique permettra de dresser un état des lieux de la situation et de repérer d'éventuelles tendances, de façon à mieux mesurer les évolutions amenées par le déploiement de la stratégie. Pour cela, les données sur la période 2018-2022 seront indispensables pour mesurer les effets des mesures prises. La première version de l'appariement contient les données 2008-2018, et la base sera actualisée chaque année jusqu'à ce que les données des soins réalisés en 2022 soient consolidées, a priori en 2023.

L'EDP étant lui aussi enrichi chaque année des dernières informations disponibles, la base étude sera actualisée lors de la diffusion du dernier millésime de l'EDP, avec un enrichissement des données les plus récentes sur le SNDS.

Ainsi en juillet 2021 par exemple serait mis à disposition l'EDP-Santé 2021 comportant les données du SNDS de 2008 à 2020 (au bout de 6 mois, il est considéré que les données de l'année sont complètes) et l'EDP 2019 avec, pour chaque source de l'EDP, des données incluant l'année 2019 (pour les données fiscales, cela correspond aux revenus de l'année d'avant, donc 2018, avant mise en place du prélèvement à la source).

Ces données seront conservées cinq ans après l'établissement de la version complète de la base 2008-2022 (sur l'hypothèse d'une base complète en 2023, les données seront conservées jusqu'en 2028) pour les besoins d'études et de statistiques en matière d'évaluation de la stratégie nationale de santé. À la fin de ce délai, les données devraient être définitivement supprimées.

Les sources déjà existantes sur ces thématiques

L'EDP-Santé s'inscrit dans le prolongement de systèmes d'informations existants, qui apportent déjà des réponses utiles à la problématique des inégalités sociales et territoriales de santé. L'appariement doit être considéré comme une maille supplémentaire conçue pour dépasser différentes limites rencontrées dans l'exploitation de ces bases de données.

Les études socio-économiques à partir du SNDS

Les données du SNDS seules contiennent un petit nombre d'informations socio-démographiques élémentaires sur les bénéficiaires, à savoir l'âge, le sexe, le lieu de résidence au niveau communal et la date de décès lorsqu'elle existe. Elles permettent également de repérer des populations précaires grâce à l'information sur l'accès à certains droits pris en charge par l'Assurance maladie[8] :

- la couverture maladie universelle complémentaire (CMU-C)[9], mais qui implique de restreindre les analyses sur les personnes âgées de moins de 65 ans, puisqu'à partir de cet âge, l'allocation de solidarité aux personnes âgées, non repérable dans le SNDS, peut se substituer à la CMU-C. Cela exclut également les analyses sur les personnes résidant à Mayotte, où la CMU-C ne s'applique pas ;
- l'aide à l'acquisition d'une complémentaire santé (ACS) ;
- l'aide médicale de l'État (AME).

Enfin, un indice de défavorisation sociale a également été constitué, permettant de connaître le niveau social de la commune dans laquelle un individu réside (Rey et al, 2009). Cet indicateur, intégré dans les bases du SNDS, est un agrégat de plusieurs grandes variables socio-économiques propres au territoire : la part des ouvriers dans la population active, la part des chômeurs dans la population active, la part des diplômés de niveau baccalauréat au moins et le revenu fiscal médian des ménages. Cet indice n'est cependant pas disponible pour les résidents dans les DROM et les COM. Il ne décrit par ailleurs pas directement la situation sociale de l'individu et manque notamment de pertinence pour les résidents de grandes villes, au sein desquelles l'hétérogénéité des situations sociales est plus grande (Ducros, 2015). Bien qu'approximatif, cet indice a suffi à révéler dans certains travaux des inégalités sociales de santé, par exemple dans le suivi par les patients diabétiques des recommandations de la Haute Autorité de Santé (Fosse, Mandereau-Bruno, 2015).

[8] Pour plus d'informations, voir la documentation du SNDS : https://documentation-snds.health-data-hub.fr/

[9] À partir du 1er novembre 2019, l'ACS et la CMU-C ont fusionné pour former la Complémentaire santé solidaire, également repérable dans le SNDS.

Les enquêtes appariées au SNDS

Pour compléter ces informations individuelles limitées, des enquêtes déclaratives ou d'autres dispositifs (données de registres médicaux par exemple) sont enrichis avec le SNDS. Ce fut par exemple le cas des enquêtes Handicap Santé en ménages 2008-2009 (Drees, Montaut et al, 2013), des enquêtes Santé et Protection Sociale de l'Irdes, et de l'enquête Care (Drees). Cela permet de croiser les informations collectées dans l'enquête, pouvant porter sur la situation sociale et économique de l'enquêté, ses habitudes de vie, ou son état de santé perçu avec les informations sur le recours aux soins. Ce type de projets permet donc aussi d'éclairer un certain nombre de questions liées aux inégalités sociales de santé, et a par exemple pu donner lieu à des études sur les inégalités d'accès à la prévention (Guthmann, 2016) ou de reste à charge. Si elles permettent de collecter des informations qui ne seront pas disponibles dans l'EDP-Santé (consommation de tabac, d'alcool, perception de sa santé, limitations d'activités), il s'agit néanmoins d'enquêtes le plus souvent ponctuelles et qui ne permettent donc pas d'étudier des évolutions concernant la partie déclarative des informations (un changement de profession ou dans la structure du ménage par exemple). En outre, le coût des enquêtes contraint nécessairement la taille des échantillons et donc les possibilités de représentativité au niveau régional ou sur des sous-populations.

Certaines cohortes, comme la cohorte Constances, ont des effectifs plus importants permettant un suivi longitudinal sur les données d'enquête également et des analyses plus fines. Leur recueil peut être enrichi de données d'examens de santé, non disponibles dans le SNDS à l'heure actuelle et permettant par exemple une évaluation de la gravité de la maladie. La participation y est cependant fondée sur le volontariat et elle peut, pour un dispositif d'interrogation lourd, souffrir de biais de représentativité, ou encore de problèmes d'attrition.

Le programme Cosmop, les prémices de l'EDP-Santé

Au début des années 2000, le projet de la cohorte pour la surveillance de la mortalité par profession, COSMOP, piloté par l'Institut de veille sanitaire (InVS) reposait sur l'appariement de l'échantillon démographique permanent aux données des causes médicales de décès du CépiDC (Inserm). Ce programme de surveillance de la mortalité avait pour objectif de décrire de façon régulière la fréquence des différentes causes de décès par profession.

Cette opération du programme COSMOP n'a finalement été produite qu'une fois (avec l'EDP dans sa version disponible en 2002) mais a déjà permis de produire des études sur les causes de mortalité par secteur d'activité et profession (Geoffroy-Perez, 2006).

Un rapport très détaillé sur la faisabilité et la mise en place du programme de surveillance COSMOP soulève les différentes limites liées à ce projet : taille d'échantillon encore trop restreinte (l'EDP ne contenait que les individus nés sur 4 jours de l'année), profession pas toujours bien identifiée via le recensement. L'EDP-Santé doit donc permettre de dépasser ces limites, notamment grâce au quadruplement du champ EDP depuis 2004, ainsi qu'à la plus grande profondeur de données disponibles, le projet étant réalisé treize ans plus tard. Enfin, l'intégration du panel DADS à l'EDP permettra de pallier le manque de finesse dans la description des parcours professionnels pour les salariés.

Ainsi, dans la continuité de ces sources, l'appariement de l'EDP au SNDS, basé sur des informations issues de sources administratives ou déjà collectées (enquêtes annuelles de recensement) répond à l'objectif de concilier la profondeur historique, la fiabilité des données, la grande taille – et donc la puissance statistique d'études à un niveau fin – et la représentativité de l'échantillon, la régularité et le suivi des informations, pour un coût très limité.

Enrichir l'EDP-Santé pour permettre des études plus ciblées

Si l'EDP-Santé viendra donc dépasser un certain nombre de limites posées par ces précédents projets, il ne permettra pas lui non plus de répondre à toutes les questions sur les inégalités sociales et territoriales de santé, et des pistes sont déjà à l'étude pour enrichir l'appariement. Ainsi, au-delà des limites inhérentes au contenu des données appariées, telles que l'absence de données cliniques décrivant l'état de santé des personnes, le périmètre des personnes concernées par l'appariement peut également constituer une limite.

Le critère d'inclusion dans l'EDP-Santé est un critère individuel : les données de l'EDP seul portent parfois sur les personnes du foyer (informations sur le ménage et la personne de référence dans les données fiscales ou les données du recensement par exemple), mais les données du SNDS ne porteront que sur l'individu EDP. Cela

limite donc les études à l'échelle du ménage, pourtant intéressantes pour la recherche sur les déterminants de la santé et du parcours de soins. Des travaux sur le rôle des pairs dans les comportements de santé, ou sur les déterminants du parcours de soins des uns sur la santé des autres pourraient être très utiles. Pour les études en périnatalité par exemple, il est primordial de pouvoir relier le suivi de la grossesse et de l'accouchement de la mère à l'état de santé et au parcours du nouveau-né. De telles études ne sont pour l'instant pas possibles avec l'EDP-Santé.

Pour dépasser ces limites, il serait envisageable d'enrichir l'appariement de données sur le ménage. C'est notamment ce qui est fait pour l'enquête Santé et Protection Sociale, collectée et appariée en 2010, 2012 et 2014. L'échantillon de l'enquête est constitué de bénéficiaires tirés au sort dans les bases administratives de l'Assurance Maladie, et de leur ménage. Les informations collectées par le biais des questionnaires sont ensuite appariées aux données du SNIIRAM-PMSI pour le bénéficiaire tiré dans l'échantillon, ainsi que toutes les personnes qui lui sont rattachées pour la prise en charge de leurs frais de santé, appelées ayants-droit, classiquement un parent et les enfants qui lui sont affiliés.

Ce type d'appariement appliqué à l'EDP-Santé serait assez simple à mettre en œuvre et permettrait donc d'obtenir l'information sur le recours aux soins des personnes appartenant à la même « grappe[10] » de l'Assurance Maladie que les individus EDP. Pour autant, cela ne garantit pas de disposer exhaustivement des informations sur l'ensemble du ménage, dont les membres ne sont pas nécessairement tous au sein d'une même grappe. Une autre possibilité, pour l'étude de la périnatalité en particulier, serait d'enrichir les données avec celles sur les nouveau-nés des mères présentes dans l'EDP, qui eux ne sont pas forcément nés des jours EDP, cela à l'aide du chaînage mère-enfant disponible dans le SNDS.

[10] Une « grappe » Assurance maladie est composée d'un ouvrant droit et de l'ensemble de ses ayants droit, classiquement un parent et les enfants qui lui sont affiliés.

LES ÉTAPES POUR BÂTIR UN PROJET D'ENVERGURE

Initié peu de temps après la création du SNDS, ce projet a permis de mettre en application l'expertise de la DREES sur les aspects juridiques liés aux données de santé, et plus particulièrement aux appariements avec le SNDS. Retracer les différentes étapes de la réflexion en amont du projet pourra permettre de capitaliser sur cette expérience pour les projets futurs.

Genèse du projet

Le projet a démarré en 2015, à la suite de réflexions au sein du Service statistique public : dans le cadre de la stratégie « Insee 2025 », un groupe de travail sur les nouvelles données intégrant des personnes de la DREES et de l'Insee avait souligné dans son rapport final l'opportunité d'intégrer des données socio-démographiques au SNIIRAM. Deux projets avaient été envisagés, l'appariement avec le dispositif SRCV (Statistiques sur les ressources et conditions de vie) et celui avec l'EDP. L'appariement du dispositif SRCV au SNIIRAM pouvait « permettre d'enrichir notablement cette source dans le contexte d'un règlement européen qui impose désormais d'intégrer tous les trois ans à l'enquête SRCV une vingtaine de questions sur les aspects santé. »[11] L'appariement avec l'EDP pouvait quant à lui permettre « d'envisager de nombreux nouveaux travaux […] et combler en partie le manque d'information socio-démographique qui est l'une des limites les plus importantes du SNIIRAM actuellement. »

La DREES et la Direction des statistiques démographiques et sociales (DSDS) de l'Insee ont alors rencontré la CNAM pour développer les réflexions sur les contours de l'appariement en juillet 2015. À partir de septembre 2016, fort de son expérience sur de précédents appariements d'enquêtes (Handicap-Santé, CARE) au SNIIRAM, la DREES prend en charge le pilotage du projet.

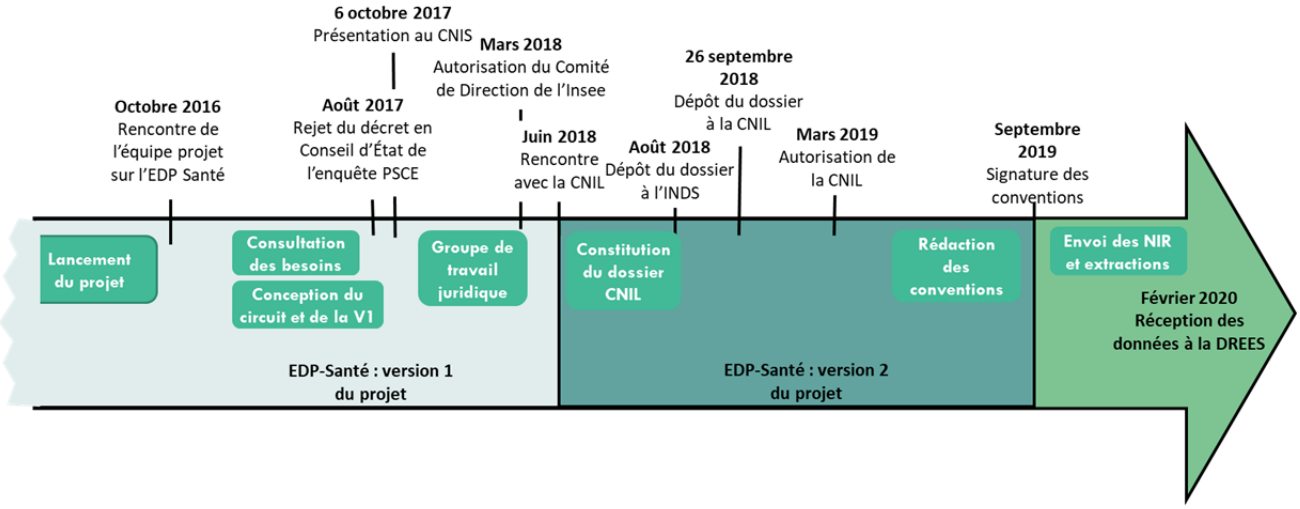
Les trois années qui ont suivi ont permis de concevoir les différentes étapes du projet, qu'il s'agisse des démarches juridiques pour le mettre en œuvre, du contenu final de la base de données, ou des conditions d'accès à ces données. Ce travail a été le fruit d'une coordination entre de nombreux acteurs et d'une adaptation aux évolutions permanentes de l'écosystème, notamment sur le plan juridique.

Résumé chronologique des étapes parcourues

- Septembre 2016 : Création d'un poste dans le bureau État de santé de la population à la DREES pour le pilotage de ce projet.
- Octobre 2016 : Rencontre avec l'Insee pour présentation de la nouvelle équipe projet de la DREES.
- Novembre 2016 – Février 2017 : Allers-retours Insee/DREES pour discussion sur la faisabilité juridique du projet avec notamment des interrogations sur la conciliation de la loi santé encadrant les données du SNDS et la loi sur les données de la statistique publique. La démarche envisagée pour donner une base légale au projet est celle d'un décret en Conseil d'État (voir supra).
- Avril 2017 – Juin 2017 : Consultation d'organismes utilisant l'EDP ou/et le SNDS pour réfléchir au format et au niveau de finesse des données (Ined, Santé publique France, Irdes, DARES, Inserm).
- 5 juillet 2017 : Validation de la note à destination de l'Insee, la Cnamts et le CASD (GENES) pour soumission du circuit de données.
- 29 août 2017 : Passage de la DREES au Conseil d'État pour le décret permettant l'appariement de l'enquête Protection sociale et Complémentaire d'entreprise (PSCE) au SNDS. Le Conseil d'État a jugé qu'un décret n'était pas requis et que ce traitement relevait du chapitre IX de la loi informatique et libertés (LIL). Son retour est détaillé dans une note et conduit la DREES à repenser les démarches juridiques pour l'EDP-Santé en considérant que le projet ne s'inscrit pas dans le cadre d'un décret en Conseil d'État mais dans le cadre d'une autorisation de la CNIL.
- 4 octobre 2017 : Présentation du projet au CNIS.

[11] Extrait du rapport final du groupe Nouvelles Données – Insee 2025.

- Octobre 2017 – Janvier 2018 : Lancement d'un groupe de travail juridique (voir composition en annexe 1) dont l'objectif est de bien cerner les enjeux autour des référentiels juridiques encadrant l'hébergement des deux bases de données.
  - La 1e réunion (18 octobre) porte sur la possibilité d'héberger les données de santé à l'Insee et au CASD.
  - La 2e réunion (12 décembre) porte sur la possibilité d'héberger les données de statistique publique de l'EDP sur le portail de la Cnamts.
  - La 3e réunion (25 janvier) vise à établir une proposition de procédure d'accès à l'appariement, basée sur le croisement des procédures d'accès à l'EDP et au SNDS.
- 16 mars 2018 : Rencontre avec le cabinet d'avocat Beslay consulté sur le choix de la procédure (Procédure INDS dans le cadre du Chapitre IX de la loi informatique et libertés), du responsable de traitement, la validation du circuit de données et la procédure d'accès aux données.
- 19 mars 2018 : Présentation du projet au comité de direction de l'Insee et validation.
- 6 juin 2018 : Rencontre avec la CNIL, objection au projet de création d'une base de données ouverte de manière pérenne à la recherche, perçue comme un entrepôt de données. Réorientation vers un projet en deux temps : mise en place de l'appariement pour une finalité précise et une temporalité limitée (qui fera de plus office de prototype) et poursuite des réflexions pour une pérennisation ultérieure de la base de données.
- Juin 2018 : La DREES revoit le contour du projet et propose à l'Insee et à la CNAM un projet en deux étapes.
- 19 juillet 2018 : Le directeur de la DREES, Jean-Marc Aubert, initie une réunion avec la directrice de la DSDS, Chantal Cases, au cours de laquelle le nouveau projet est discuté avec l'Insee. Il est proposé de leur soumettre le dossier INDS pour relecture et association en co-responsabilité. Finalement, l'absence de l'historique des données du CépiDC dans le SNDS central représente une limite importante pour l'usage que l'Insee souhaitait faire de la base de données. La DREES sera donc le seul responsable de traitement.
- Lancement de l'étape 1 du projet : mettre en œuvre le prototype.
- 13 août 2018 : Le dossier est validé par l'INDS et transmis au CEREES.
- 12 septembre 2018 : Réception de l'avis favorable du CEREES.
- 26 septembre 2018 : Transmission du dossier à la CNIL.
- 16 octobre 2018 : Rencontre de l'équipe DEMEX à la CNAM pour préparer l'extraction et finaliser le document d'expression des besoins.
- 24 octobre 2018 : Démarrage des travaux sur les conventions.
- 22 janvier– 1er février 2019 : Questions-réponses avec la CNIL.
- 28 février 2019 : Séance de délibération de la CNIL. Le dossier n'est pas examiné lors de la séance et ne sera jamais reprogrammé.
- 28 mars 2019 : Avec deux mois de retard sur la procédure, conformément à l'article 54 du chapitre IX de la loi informatique et libertés en vigueur à cette date, et en accord avec la commissaire du gouvernement auprès de la CNIL, la demande d'autorisation de l'appariement EDP-SNDS est réputée tacitement acceptée.
- 4 juillet 2019 : Le serveur Big Data de la DREES est ré-homologué pour tenir compte des évolutions sur les données hébergées et des risques que cela peut entraîner en matière de confidentialité.
- Juillet-août 2019 : Allers-retours sur les conventions DREES-INSEE et DREES-CNAM.
- Septembre 2019 : Signature des deux conventions.
- Octobre 2019 : deux tentatives infructueuses puis une troisième tentative réussie de transfert des NIR via la procédure SAFE de l'Insee à la CNAM, le 22 octobre 2019.
- Février 2020 : réception des données de l'appariement.



Le contexte juridique sur le SNDS : s'inscrire dans un cadre réglementaire en chantier

Le NIR est présent dans l'EDP, et sous forme pseudonymisée dans le SNDS. Il permet donc un appariement des sources sur cet identifiant. L'enjeu du projet n'a donc pas été technique, mais juridique. Il s'agissait de savoir dans quel cadre situer le projet, à la frontière des données de la statistique publique et de santé et quelle procédure suivre pour le rendre possible. Un groupe de travail juridique s'est penché sur cette question lors de trois réunions tenues entre octobre 2017 et janvier 2018.

Quelle démarche pour constituer la base de données ?

La DREES s'est appuyée sur l'expérience de précédents appariements d'enquêtes menées par la Drees au SNDS : celui de l'enquête Handicap-Santé collectée en 2008 (Montaut et al., 2013), et celui de l'enquête CARE collectée en 2015-2016 (Carrère, 2016) dont l'appariement avait été achevé en 2017, à partir du NIR. Jusqu'alors, la loi informatique et libertés imposait, pour tout traitement sollicitant l'utilisation du numéro d'identification au répertoire (NIR), un décret en conseil d'État. Le décret pour l'enquête CARE fut ainsi publié au journal officiel en mars 2015, après 22 mois de travail.[12]

La difficulté propre à ce nouvel appariement reposait sur le fait que la loi de modernisation de notre système de santé, promulguée en janvier 2016, fixait un nouveau cadre de mise à disposition des données de santé, redéfini dans l'article 193. Le cadre réglementaire était donc en cours de construction au moment du développement du projet et tous les décrets faisant suite à la loi n'étaient pas encore publiés en 2017. L'article contient notamment la création du SNDS, prévue pour avril 2017, dont les finalités sont définies à l'article L 1461-1 III 6 du code de la santé publique comme la recherche, les études et l'évaluation.

Une difficulté supplémentaire concernait la distinction, présente dans la loi, entre les traitements statistiques et les traitements nécessaires à la recherche, les études et les évaluations dans le domaine de la santé. Ces traitements figurent en effet tous deux dans l'article 8 de la loi informatique et libertés, en exception à l'interdiction des traitements sur les données à caractère personnel relatives à la santé.

- Une première exception était permise s'il s'agissait d'un traitement statistique. Dans ce cas, il entrait dans le cadre de l'article 25. Celui-ci autorisait les traitements sur les données à caractère personnel après autorisation de la CNIL. Dans le cas d'une opération utilisant le NIR en clair[13], le traitement devait être autorisé par décret en Conseil d'État, pris après avis de la CNIL.
- Une seconde exception était permise s'il s'agissait d'un traitement sur les données de santé. Dans ce cas, il entrait dans le cadre du chapitre IX. L'article 54 de ce chapitre prévoyait que les traitements sur les données à

[12] Document de travail n°56, Les enrichissements prévus pour l'enquête CARE-Ménages, Amélie Carrère.

[13] Le NIR est dit « en clair » lorsqu'il n'est pas pseudonymisé ou chiffré par une méthode ne permettant pas de l'identifier.

caractère personnel ayant une finalité d'intérêt public de recherche, d'étude ou d'évaluation dans le domaine de la santé sont soumis à l'autorisation de la CNIL. Celle-ci prend sa décision après avis d'un comité, le Comité de protection des personnes (CPP) lorsque la personne humaine est impliquée[14], le Comité d'expertise pour les recherches, les études et les évaluations scientifiques (CEREES) lorsque ce n'est pas le cas.

L'interprétation de la loi par la DREES s'est appuyée sur l'idée que l'exception liée au statut de statistique publique primait sur celle des données de santé parce qu'elle apparaissait avant dans la liste des exceptions à l'interdiction de traiter des données de santé. Ainsi, les démarches juridiques suivantes avaient été envisagées :

- Un avis d'opportunité du CNIS sur le projet d'appariement entre les données de l'EDP et les données du SNDS ;
- Une demande d'accès aux données auprès du CNIS, au titre de l'article 7 bis de la loi de 1951 autorisant l'INSEE et les SSM à accéder à des données administratives, dont les données du SNDS ;
- Un décret en Conseil d'État pris après avis de la CNIL pour autoriser l'appariement de l'EDP aux données du SNDS à partir du NIR.

En août 2017, un autre projet de la DREES, l'appariement de l'enquête Protection sociale et complémentaire d'entreprise (PSCE) aux données du SNDS, est soumis au Conseil d'État. Ce dernier a finalement considéré que ce traitement relevait du chapitre IX de la loi informatique et libertés (recherches, études et évaluations en santé) et que ces dispositions devaient s'appliquer à tout projet d'élaboration de statistiques à partir du SNDS.

Encadré 1 • Extrait de la Note de rejet sur le projet de décret autorisant la mise en œuvre d'un traitement de données à caractère personnel relatif à un dispositif d'enquête portant sur la complémentaire d'entreprise

« Or, si en vertu du 1° du I de l'article 27 de la loi du 6 janvier 1978, « les traitements de données à caractère personnel mis en œuvre pour le compte de l'État, d'une personne morale de droit public ou d'une personne morale de droit privé gérant un service public » qui portent sur des données parmi lesquelles figure le NIR sont en principe autorisés par décret en Conseil d'État, une exception est désormais prévue, au IV de ce même article 27, pour les « traitements à des fins de recherche, d'étude ou d'évaluation dans le domaine de la santé » qui sont soumis au chapitre IX de la loi du 6 janvier 1978. Ce chapitre IX prévoit, à l'article 54, que ces traitements sont soumis à une simple autorisation de la Commission nationale de l'informatique et des libertés délivrée, pour les demandes d'autorisation relatives à des études ou à des évaluations ainsi qu'à des recherches n'impliquant pas la personne humaine, après avis du comité d'expertise pour les recherches, les études et les évaluations dans le domaine de la santé.

Le Conseil d'État a jugé que le projet de traitement de données décrit au deuxième alinéa devait être regardé comme un « traitement à des fins de recherche, d'étude ou d'évaluation dans le domaine de la santé », au sens des articles 27 et 54 de la loi du 6 janvier 1978. Il a en particulier estimé :

- D'une part, que la circonstance – mise en avant par le Gouvernement pour justifier la nécessité de prendre un décret en Conseil d'État – que le II de l'article 8 de loi du 6 janvier 1978 semble distinguer les traitements statistiques réalisés par l'un des services statistiques ministériels soumis à la loi n°51-711 du 7 juin 1951 sur l'obligation, la coordination et le secret en matière de statistiques, mentionnés à son 7°, des traitements nécessaires à la recherche, aux études et évaluations dans le domaine de la santé étant, eux, mentionnés à son 8°, était sans incidence sur l'applicabilité des dispositions du chapitre IX de la loi du 6 janvier 1978 au projet de traitement. Le II de l'article 8 a en effet seulement pour objet de dresser la liste des traitements de données soumis à l'interdiction énoncée au I, et non de déterminer le champ d'application du chapitre IX ;

- D'autre part, que la circonstance que certaines données objet du projet de traitement soient étrangères au domaine de la santé ne faisait pas obstacle à ce que ce traitement soit regardé comme relevant dans son ensemble du « domaine de la santé » au sens de la loi du 6 janvier 1978

Dans ces conditions, le Conseil d'État a considéré que le projet qui lui était soumis ne relevait pas d'un décret en Conseil d'État et devait, en application des dispositions de l'article 54 de la loi du 6 janvier 1978, être autorisé par la Commission nationale de l'informatique et des libertés, après avis du comité d'expertise pour les recherches, les études et les évaluations dans le domaine de la santé. »

[14] La loi n° 2012-300 du 5 mars 2012 relative aux recherches impliquant la personne humaine, communément appelée loi Jardé, définit les procédures de soumission de projet de recherche s'inscrivant dans ce cadre. Sont considérées comme recherche impliquant la personne humaine les recherches « organisées et pratiquées sur des personnes volontaires saines ou malades qui visent à évaluer les mécanismes de fonctionnement de l'organisme humain normal ou pathologique, l'efficacité et la sécurité de réalisation d'actes ou de l'utilisation ou de l'administration de produits dans un but de diagnostic, de traitement ou de prévention d'états pathologiques » (décret 2017-884, JO du 10 mai 2017)

La procédure retenue à l'été 2017 est donc la suivante :

- Un avis d'opportunité du CNIS sur le projet d'appariement entre les données de l'EDP et les données du SNDS ;
- Une demande d'accès aux données devant le CNIS, au titre de l'article 7 bis de la loi de 1951 autorisant l'INSEE et les SSM à accéder à des données administratives, dont les données du SNDS ;
- Une demande d'autorisation à la CNIL après avis du CEREES, via la procédure auprès de l'INDS.

Quel hébergement? Concilier les enjeux juridiques de différents domaines statistiques

Outre les interrogations sur la démarche juridique à retenir, le projet a impliqué de concilier des référentiels en matière de protection des données personnelles et d'accès à ces données, relativement différents, à savoir les exigences en matière de secret statistique et le référentiel de sécurité du SNDS.

Le principe du secret statistique est décrit succinctement dans le code des bonnes pratiques de la statistique européenne (2011), au cinquième principe, et de manière plus détaillée dans le cadre d'assurance qualité du système statistique européen. Il impose un contrôle strict des exploitations des données et une surveillance de ce qui est extrait à partir de ces données.

Le secret statistique est décrit et cadré, en France, dans la loi 1951 sur l'obligation, la coordination et le secret en matière de statistiques. Il s'appuie notamment sur un engagement de confidentialité des personnels concernés, pour tous les agents du service statistique public. Il est également étendu (via le comité du secret) à toute personne ayant accès à un échantillon ou une base de données.

Des sanctions sont prévues par le code pénal en cas de violation du secret statistique.

En matière de confidentialité, l'Insee informe les producteurs via un guide de bonnes pratiques rendu public (diffusé également au personnel) et contrôle pour sa part la diffusion de ses données. La sécurité et l'intégrité des bases de données sont garanties par un certain nombre de dispositions et une unité spécifique est en charge de ces problématiques à l'Insee.

Le sixième principe du code des bonnes pratiques de la statistique européenne est celui qui concerne plus spécifiquement l'accès des chercheurs aux données individuelles, et correspond donc au cas de l'accès à l'EDP sur un portail extérieur au service statistique public. Dans ce cas, les conditions d'accès sont soumises au secret (autorisation via le comité du secret), l'accès doit être sécurisé et des contrôles doivent être prévus pour interdire toute copie des données. C'est ce qui justifie le contrôle a priori des sorties mis en place au CASD, seul emplacement extérieur au système statistique public où l'EDP est actuellement accessible.

Le second référentiel, celui du SNDS, s'est inscrit par anticipation dans l'esprit du règlement européen sur la protection des données (RGPD, 2018). Sa logique est de responsabiliser les acteurs, de conserver des traces de ce qui est fait et d'auditer pour repérer d'éventuelles mauvaises pratiques ou utilisations. Le périmètre de ce référentiel de sécurité comprend toutes les données à caractère personnel du SNDS. Tout responsable de traitement sur ces données doit être en conformité au référentiel avant l'accès aux données.

Il repose sur cinq principes :

- La pseudonymisation ;
- L'authentification forte (soit deux étapes indépendantes ou supports pour s'authentifier lors de l'accès aux données) ;
- Une traçabilité élevée (tracer tous les traitements, notamment les requêtes et les sorties) ;
- La sensibilisation et formation des utilisateurs ;
- Le contrôle.

Une fois tenu compte de ces cinq principes, la conformité au référentiel de sécurité repose sur l'homologation d'un système-fils[15]. Celle-ci passe par une analyse de risque sur le traitement, compte tenu des mesures de sécurité proposées par le gestionnaire du système, et éventuellement la mise en évidence d'un risque résiduel.

[15] Un système-fils est un système d'informations issu du SNDS élargi hébergeant ou mettant à disposition des données relatives au SNDS, cédées par le SNDS central, un système source ou un autre système fils.

En matière d'hébergement des données envisagé pour l'EDP-Santé, le portail SNDS géré par la CNAM, ainsi que le CASD sont deux plateformes conformes au référentiel du SNDS et offraient donc la possibilité de s'inscrire dans ce second cadre. Le CASD, déjà hébergeur des données de l'EDP, offrait également les garanties du respect des conditions du secret statistique et la CNAM proposait d'instaurer le contrôle a priori de l'ensemble des sorties produites à partir de l'EDP-Santé sur son portail pour s'inscrire également dans ce référentiel.

Un problème subsistait néanmoins sur la spécificité des données fiscales, pour lesquelles un décret précise (article R135-D du livre des procédures fiscales) que leur accès s'effectue au moyen du CASD, ce qui restreignait règlementairement l'hébergement de l'EDP (pour sa partie sur les données fiscales) sur une autre plate-forme que le CASD. Une solution envisagée était de proposer une version synthétisée et plus limitée des données de l'EDP sur le portail de la CNAM, avec par exemple une variable renseignant le niveau de vie qui n'est pas considérée comme une donnée fiscale directe.

La conception initiale du projet comme une base ouverte à la recherche

Quelles données ? Optimiser la base pour l'évaluation des inégalités sociales de santé

Dès son démarrage, l'objectif du projet est de mettre à disposition les données de l'appariement à d'autres organismes et au monde de la recherche. En effet, si la DREES souhaite utiliser ces données dans le cadre de ses études sur les inégalités sociales de santé et de l'évaluation de la SNS, elle souhaite également capitaliser sur la mise en œuvre du projet pour faciliter l'exploitation des données par d'autres équipes, sans reproduire les mêmes étapes à chaque fois.

Les différents échanges avec des organismes susceptibles d'être intéressés (Ined, Irdes, Santé publique France, INSERM, DARES), ainsi que la présentation au Conseil national de l'information statistique, ont permis de finaliser un projet qui pourrait répondre au maximum de besoins. Le principal enjeu était de proposer un outil adapté à la fois aux chercheurs utilisateurs des données de l'EDP, familiers de l'environnement de travail au CASD, et aux chercheurs utilisateurs des données du SNDS, plus habitués à travailler sur le portail de la CNAM. Une difficulté portait spécifiquement sur les données fiscales, composante de l'EDP (*cf.* infra).

Il est alors prévu de mettre à disposition le projet sous deux formes :

- Un EDP-Santé détaillé (EDP allégé – SNDS détaillé) disponible sur le portail SNDS géré par la CNAM, dans lequel seraient appariées les données de l'EDP (sans les données fiscales, ou avec des indicateurs synthétiques) aux données brutes du SNDS ;
- Un EDP-Santé agrégé (EDP complet – SNDS synthétisé) disponible au Centre d'accès sécurisé aux données (CASD) dans lequel seraient appariées les données de l'EDP (y compris les données fiscales) à des indicateurs de santé individuels agrégés par types de soins issus du SNDS, plus simples d'utilisation.

Quel circuit de mise en œuvre ? Minimiser le risque

L'EDP et le SNDS sont deux bases adossées au NIR, qui constitue donc la clé de jointure pour l'appariement. Pour limiter le risque de ré-identification, le NIR est cependant pseudonymisé dans le SNDS par le biais d'un algorithme de hachage (initialement deux étapes, appelées FOIN 1 et FOIN 2, pour fonction d'occultation d'informations nominatives[16]) qui n'est pas réversible (DREES, 2015). La base études de l'EDP qui est diffusée aux utilisateurs externes à l'INSEE *via* le CASD ne contient quant à elle pas le NIR, mais l'INSEE l'utilise pour réaliser l'appariement avec les DADS et il est donc présent dans les bases de travail de l'EDP à l'INSEE.

L'enjeu de la réflexion était donc de concevoir un circuit permettant d'établir la table de passage entre l'identifiant des individus EDP et leurs identifiants SNDS et de déterminer les flux de données. Ce circuit doit respecter les référentiels juridiques qui encadrent ces deux bases de données et notamment ne pas amener à ce que le NIR en clair soit associé à des données de santé dans un même fichier, ni à ce qu'une table de passage entre le NIR

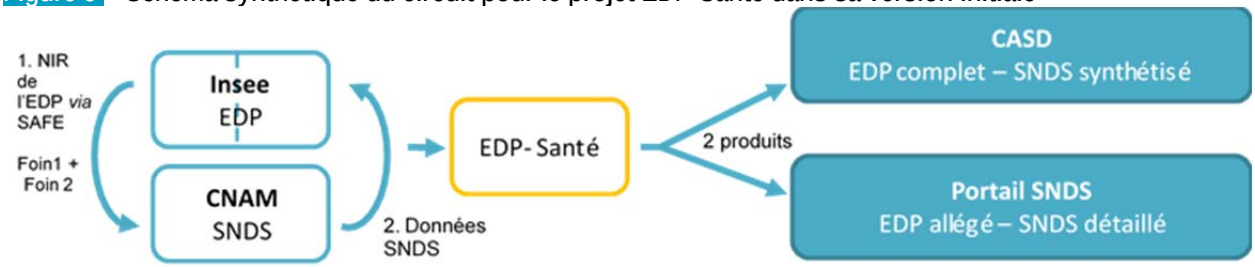
[16] Depuis le 1er septembre 2019, la CNAM chiffre les identifiants par le biais d'une troisième étape FOIN 3 afin de renforcer la sécurité de la base de données, mais cela n'a pas eu d'impact sur le circuit.

en clair et le NIR pseudonymisé soit créée. Cela implique donc le recours à un tiers de confiance pour assurer la transmission et le cryptage intermédiaire de ce NIR. Pour s'appuyer sur l'expérience acquise lors de l'appariement de l'enquête CARE au SNDS, le CASD a été envisagé pour jouer ce rôle de tiers de confiance : réception de la part de l'INSEE de la table des identifiants du champ EDP (NIR) avec l'identifiant anonyme correspondant ; cryptage de ces NIR via le premier algorithme de hachage FOIN 1, qui permet d'obtenir l'identifiant dans le SNDS, et envoi ensuite de cette table à la CNAMTS.

Cependant, le CASD ayant également été envisagé pour la mise à disposition du produit EDP-Santé agrégé, il ne pouvait pas à la fois recevoir les données et prendre en charge le hachage des NIR. En effet, cela aurait impliqué d'héberger au même endroit des NIR en clair, des données de santé et leur clé d'appariement. Il était donc prévu qu'une autre cellule de l'Insee joue le rôle de récepteur des données du SNDS pour les apparier à celles de l'EDP et transmettre au CASD l'ensemble des données avec un nouvel identifiant généré aléatoirement.

Finalement, à l'été 2018, la CNAM a fait homologuer par la CNIL une procédure automatisée de chiffrement FOIN qui permet de ne plus solliciter de tiers de confiance. Cette procédure, appelée SAFE, permet une transmission sécurisée des NIR à la CNAM pour les appariements directs avec le SNDS. Les informations concernant les individus de l'extraction (NIR, date de naissance et sexe) sont déposées sur la plateforme « SAFE » accompagnées d'un identifiant propre au projet, nommé ID-FLUX. À l'issue de cette transmission, les algorithmes de hachage du NIR, FOIN1 et FOIN2, sont appliqués de manière automatisée afin d'obtenir l'identifiant SNDS. Une autre cellule de la CNAM peut alors extraire les données de correspondance entre le NIR FOIN 2 qui correspond à l'identifiant dans le SNDS et l'ID-FLUX et conserver cette table de passage uniquement le temps des échanges visant à valider l'appariement.

Figure 5 • Schéma synthétique du circuit pour le projet EDP-Santé dans sa version initiale



La redéfinition du projet : limiter le projet pour limiter les risques

L'incompatibilité du projet avec la loi de 2016

En mai 2018, le projet ainsi conçu a été présenté à la CNIL, qui peut accompagner les utilisateurs de données de santé en amont de leur demande d'autorisation. Les experts de la CNIL ont alors fait remarquer que le périmètre du projet dépassait celui d'une autorisation dans le cadre du chapitre IX (section II) de la loi informatique et libertés alors en vigueur. Celui-ci peut porter sur des traitements à partir de données du SNDS pour une recherche bien précise, et délimitée dans le temps. L'EDP-Santé supposait cependant la constitution d'une base pérenne, à laquelle pourraient ensuite accéder les organismes ou chercheurs pour leurs traitements, mais dont les sujets de recherche ne pourraient être tous précisés dès la création de la base. Le projet d'appariement consistait donc en la création d'un entrepôt de données, tel que défini par la CNIL, mis à disposition dans deux bulles sécurisées, le CASD et le portail de la CNAM.

La section I du chapitre IX de la LIL, habituellement utilisée pour la création d'entrepôts de données, ne pouvant pas s'appliquer aux données du SNDS du fait des dispositions de l'article L 1461-3 du code de la santé publique alors en vigueur, le traitement tel qu'envisagé n'était donc pas possible légalement.

Toutefois, en matière de sécurité sur les données à caractère personnel, il est plus judicieux de réaliser une seule fois l'appariement et de l'héberger sur un serveur sécurisé tel que ceux envisagés dans le cadre du projet plutôt que de multiplier les appariements et les flux de données à chaque demande de traitement d'un nouvel organisme.

Sans renoncer au projet de constitution de la base pérenne, il a donc été décidé de réaliser le projet en deux étapes : une première demande, qui s'inscrirait dans le cadre du chapitre IX de la loi informatique et libertés, sur le projet de recherche de la DREES, et en parallèle, des échanges et réflexions sur la loi pour parvenir dans un second temps à pérenniser l'EDP-SNDS et à le mettre à disposition sur une plateforme sécurisée, comme prévu.

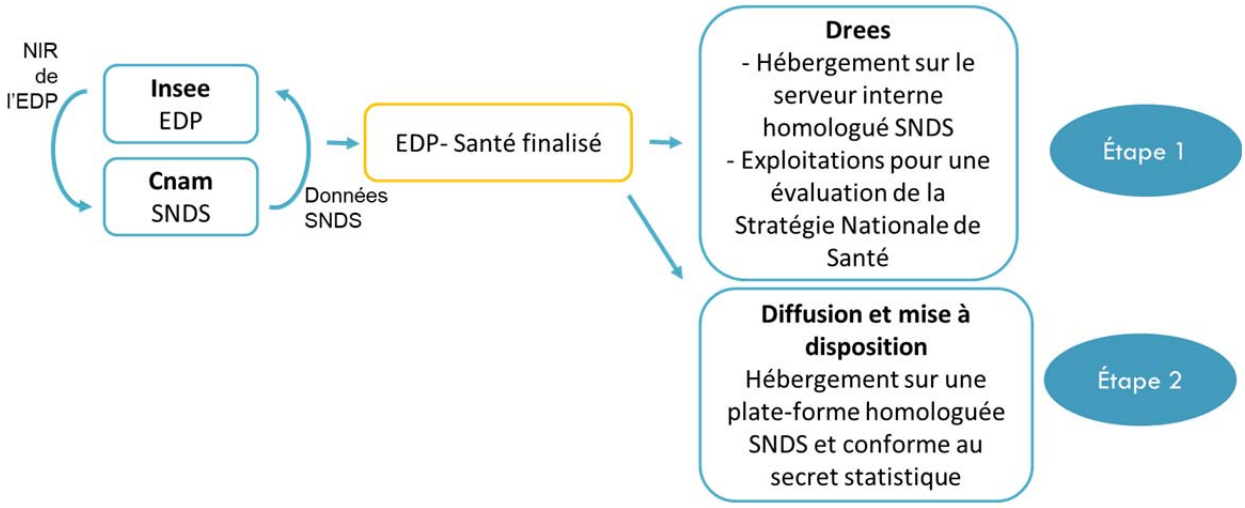
Le choix du projet final

Pour la première étape, il est proposé de constituer une première version de l'appariement à la DREES, dont la finalité serait toujours de permettre l'évaluation de la stratégie nationale de santé en matière de réduction des inégalités sociales et territoriales de santé. À cette finalité s'ajoute désormais un second objectif, celui de disposer d'un retour d'expérience sur la mise en œuvre de cet appariement et d'une expertise de la qualité des données obtenues. Cet objectif s'inscrit alors en amont d'une seconde étape du projet visant à mettre à disposition les données à d'autres organismes ou équipes de recherche.

Cette deuxième étape est désormais légalement réalisable : la possibilité de créer des entrepôts de données contenant des données du SNDS a été introduite dans le code de la santé publique, lors des modifications portées par la loi sur l'organisation et la transformation du système de santé de juillet 2019. Cette nouvelle possibilité n'est cependant pas encore mise en œuvre, la DREES s'étant concentrée sur la réalisation de la première phase du projet.

À l'été 2018, le projet est donc redéfini pour simplifier le nombre d'acteurs impliqués dans la construction du prototype de l'EDP-Santé (figure 6).

Figure 6 • Schéma synthétique des deux étapes prévues pour le projet EDP-Santé



La DREES, en tant que SSM, héberge l'EDP dans son intégralité et dispose déjà d'un serveur dont les caractéristiques répondent aux exigences du référentiel de sécurité du SNDS. À ce titre, elle est en mesure d'héberger l'EDP-Santé.

Plus précisément, cela signifie que la DREES réceptionne l'envoi sécurisé par l'Insee des données de l'EDP accompagnées d'un identifiant individuel propre au projet, ainsi que l'envoi sécurisé par la CNAM des données du SNDS accompagnées du même identifiant. Ces données sont pseudonymisées et aucune base ne contient de traces du NIR. Après réception, la DREES génère un nouvel identifiant aléatoire pour supprimer dans sa base études l'identifiant du projet et éviter tout lien avec les tables ayant circulé pour la constitution de l'EDP-Santé (voir circuit détaillé en Annexe 2).

La DREES initie donc en août 2018 sa démarche auprès de l'INDS dans l'optique d'obtenir une autorisation de la CNIL pour ce traitement. Le dossier pour un « Enrichissement de l'échantillon démographique permanent (EDP) avec le système national de données de santé pour une évaluation de la Stratégie Nationale de Santé 2018-2022 » est déposé le 13 août 2018.

Conformément à la procédure, ce dossier contient :

– Un protocole scientifique décrivant l'intégralité du projet, ses objectifs, ses conditions de mise en œuvre et d'hébergement des données finales ;

– Un résumé du traitement à destination du CEREES, comité d'experts qui examine la cohérence entre la finalité de l'étude proposée, la méthodologie présentée et le périmètre des données auxquelles il est demandé accès[17] ;
– Un formulaire de demande d'autorisation d'un traitement de recherche, étude ou évaluation dans le domaine de la santé, qui précise la nature des données concernées par le traitement, les moyens utilisés pour informer les personnes concernées, et la sécurité et l'architecture informatique liées au traitement.

Déroulé de la procédure d'autorisation par la CNIL

Le 6 septembre 2018, le CEREES a rendu un avis favorable, ce dans le respect du délai prévu d'un mois. Le dossier a ensuite été transmis à la CNIL le 26 septembre.

D'après l'article 54 du chapitre IX de la loi informatique et libertés alors en vigueur, la CNIL dispose d'un délai de deux mois à compter de la réception de la demande. Ce délai peut toutefois « être prolongé une fois pour la même durée sur décision motivée de son président ou lorsque l'Institut national des données de santé est saisi en application du second alinéa de l'article 61 ».

L'article prévoit que « lorsque la Commission nationale de l'informatique et des libertés ne s'est pas prononcée dans ces délais, la demande d'autorisation est réputée acceptée. Cette disposition n'est toutefois pas applicable si l'autorisation fait l'objet d'un avis préalable en application de la section 2 du présent chapitre et que l'avis ou les avis rendus ne sont pas expressément favorables ».

Un courrier de la CNIL a d'abord été adressé à la DREES le 29 octobre pour informer du prolongement du délai de deux mois, « compte tenu des caractéristiques [du] dossier ».

Des questions ont ensuite été transmises par la CNIL une semaine avant la fin du délai légal de quatre mois. Ces questions ont permis d'améliorer le protocole en matière d'information faite aux personnes et d'envisager un plan d'action pour sécuriser davantage la plateforme d'hébergement des données à la DREES. Les échanges se sont achevés au bout de deux semaines et n'étaient pas bloquants pour le projet. Le dossier fut programmé à la séance de la Commission du 28 février, mais ne fut pas examiné. La CNIL ne s'est pas prononcée jusqu'à la fin du mois de mars.

Conformément à l'article 54 du chapitre IX de la loi informatique et libertés, en accord avec la commissaire du gouvernement auprès de la CNIL, la demande d'autorisation de l'appariement des données du SNDS aux données de l'EDP, afin de constituer l'EDP-Santé, est réputée acceptée.

Cette autorisation implique que la DREES peut mettre en œuvre le traitement demandé et s'en porte responsable.

Des contraintes de sécurité à respecter : analyser les risques sur la plateforme d'hébergement retenue

Lors des échanges avec la CNIL, le dossier a également été complété par une analyse d'impact relative à la protection des données (AIPD), qui a « pour objectif de construire et de démontrer la mise en œuvre des principes de protection de la vie privée afin que les personnes concernées conservent la maîtrise de leurs données à caractère personnel. » (Guide rédigé par la CNIL sur la méthode pour conduire une AIPD, édition février 2018). Cette analyse n'est pas obligatoire pour tous les traitements, mais la CNIL a établi une liste des types d'opérations de traitement pour lesquelles elle estime obligatoire de la réaliser. Les traitements des données de santé nécessaires à la constitution d'un entrepôt de données ou d'un registre font partie de cette liste, et, dans la mesure où l'EDP-Santé doit être alimenté pendant quatre ans jusqu'à ce que la base soit complète, il est apparu nécessaire de faire une AIPD à son sujet[18].

Cette analyse s'est appuyée en grande partie sur une analyse de risques menée avec un prestataire extérieur sur la plate-forme d'hébergement des données de la DREES, mais elle présente également les mesures adop-

[17] https://www.snds.gouv.fr/SNDS/Processus-d-acces-aux-donnees

[18] Plus largement, à la DREES, tout traitement sur des données de santé est associé à une AIPD.

tées aux différentes étapes du processus de mise en œuvre de l'EDP-Santé pour minimiser les risques d'atteintes à la vie privée.

En effet, les données sont pseudonymisées dans chacune des bases initiales et font l'objet d'attribution d'un nouvel identifiant aléatoire non signifiant dans la base finale. Néanmoins, étant donnée la quantité d'informations contenue dans l'EDP-Santé, et la finesse de ces informations, elles sont indirectement identifiantes : cela signifie qu'il est possible, en croisant quelques informations, de réidentifier une personne ou un groupe de personnes (DREES, 2015). Il était donc primordial dans ce projet de minimiser les risques. Le serveur sur lequel sont hébergées les données respecte donc un certain nombre de principes permettant de limiter l'accès aux données. Conformément au référentiel de sécurité du SNDS, il prévoit une double authentification pour les utilisateurs, leur sensibilisation aux notions de confidentialité, ainsi que la traçabilité de toutes les opérations effectuées sur les données.

Ouverture des données à la recherche : un objectif de moyen terme

La seconde étape du projet consiste donc à mettre en place les fondements juridiques et organisationnels pour permettre la création de l'EDP-Santé comme un entrepôt de données ouvert à l'extérieur de la Drees.

En matière d'accès aux données de ce futur entrepôt, les deux procédures existantes, nécessaires l'une à l'EDP et l'autre au SNDS, sont distinctes. Il semble donc indispensable pour les futures équipes de recherche d'obtenir un avis favorable *via* les deux procédures (autorisation des archives après avis du comité du secret pour l'EDP, autorisation CNIL après avis du CESREES[19] pour le SNDS), mais une mutualisation des étapes pouvait être envisagée :

- Pour la constitution du dossier, il faudrait faciliter les échanges avec d'une part le service producteur de l'EDP à l'Insee et d'autre part le service producteur du SNDS, la CNAM, qui pourrait éventuellement être représentée par la plateforme des données de santé (ou Health Data Hub). La DREES propose aussi de centraliser ces échanges, dans la mesure où elle pourrait devenir rapidement familière des deux sources et s'adresser aux services producteurs (Insee ou CNAM) une fois le dossier pré-validé.
- Concernant la double procédure, il a été envisagé d'insérer la procédure de demande via le Comité du secret au sein de la procédure d'autorisation de la CNIL, une fois l'avis émis favorable par le CESREES. Le Comité du secret statistique pourrait émettre un avis général via une procédure accélérée (par exemple par consultation électronique) s'appuyant sur l'avis préalablement émis par le CESREES et visant plus spécifiquement la pertinence du recours aux données de l'EDP. La CNIL se prononcerait ensuite.

Ces réflexions sont en cours pour la concrétisation de cette seconde étape du projet.

[19] Suite à la loi sur l'organisation et la transformation du système de santé de juillet 2019, le CEREES a été remplacé par le CESREES (comité éthique et scientifique pour les recherches, études et évaluations dans le domaine de la santé).

MISE EN ŒUVRE ET EXPERTISE DE LA BASE ISSUE DE L'APPARIEMENT

Le suivi des procédures : un calendrier au temps long

Une fois l'autorisation officialisée en avril 2019 auprès des producteurs de données, la mise en œuvre du projet était conditionnée à la signature de conventions entre le responsable de traitement et les gestionnaires des bases de données.

L'option retenue fut la rédaction de deux conventions bi-partites, principalement pour faciliter les échanges. La CNAM, familière de la constitution de systèmes-fils, dispose d'un modèle pré-établi de convention faisant intervenir le responsable de traitement du SNDS central (la CNAM), le responsable de traitement, et le gestionnaire du système-fils (la DREES dans les deux cas). Dans le cas de l'EDP-Santé, il n'y avait pas de sous-traitant à inclure dans la convention, mais le rôle de l'Insee en tant qu'expéditeur des NIR via la procédure SAFE a tout de même été consigné dans la convention. Cette convention CNAM-DREES précise les engagements du responsable de traitement ainsi que celles du gestionnaire du système-fils relatifs au traitement des données, les modalités de l'appariement direct, les données constituant le système-fils, et les modalités de transmission des données. La convention court sur une durée de trois ans et devra donc être renouvelée en septembre 2022 pour maintenir le partenariat jusqu'à atteindre la complétude de l'EDP-Santé.

La convention INSEE-DREES a fait l'objet d'une co-rédaction et précise le double rôle de l'Insee en tant qu'expéditeur des NIR à la CNAM et expéditeur des données de l'EDP ainsi que ses engagements en matière de protection des données lors de ces différentes opérations, de même que pour la DREES en tant que responsable de traitement.

Les échanges sur ces conventions se sont étendus de mai à juillet pour une signature finale des deux conventions par l'ensemble des parties au début du mois de septembre 2019.

Les dernières étapes pour la mise en œuvre du circuit ont pu être ralenties par la nécessité de valider l'envoi des NIR par l'Insee via la plateforme SAFE proposée par la CNAM, et qui constituait une nouvelle procédure pour l'Insee. Cette procédure demande par ailleurs la mise au format spécifique de la table en entrée qui doit être préparée à l'Insee et sera à reproduire à chaque nouvel envoi. La CNAM a finalement réceptionné les NIR après passage de FOIN 1 et FOIN 2 le 22 octobre 2019.

L'extraction et l'expédition des données par l'Insee et la CNAM ont ensuite pris trois mois : la CNAM est sollicitée par de nombreuses demandes d'extractions et le projet a donc été ajouté à une file d'attente qui a pu le retarder. L'Insee devait de son côté remettre en place la procédure d'envoi des archives de l'EDP.

Les données ont finalement été reçues par la DREES le 21 février 2020.

De la livraison des données à l'expertise

La DREES a développé un outil de validation des extractions, qui permet notamment de parcourir l'ensemble des tables réceptionnées, et de construire quelques statistiques élémentaires pour s'assurer de la complétude du produit livré. Cela permet de rapidement faire un retour à la CNAM sur les corrections nécessaires, et ce, sans avoir besoin d'expertiser finement l'ensemble des tables et variables. La convention avec la CNAM prévoit en effet trois mois au gestionnaire du système-fils pour vérifier que la qualité des données est conforme à ses attentes.

Pour l'EDP par exemple, ce script a permis d'identifier rapidement que deux années de soins étaient incomplètes dans l'extraction et de demander à la CNAM une correction des données.

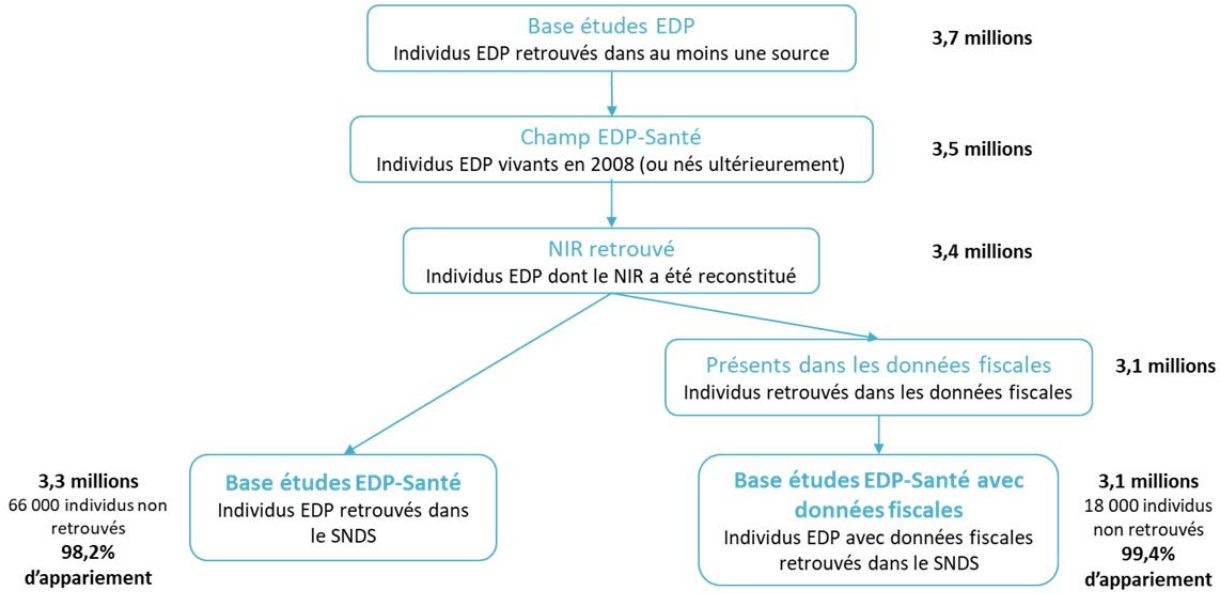
Qualité de l'appariement

L'appariement repose sur l'envoi des NIR des individus inclus dans le champ de l'EDP de l'Insee à la CNAM, qui reconstitue le NIR tel qu'il est crypté dans le SNDS et extrait les informations des personnes ainsi retrouvées dans le SNDS.

La qualité de l'appariement se mesure donc à la part des personnes du champ EDP retrouvées dans le champ du SNDS. Cependant, le champ de l'EDP est par construction plus large que celui du SNDS puisque certaines données sont collectées depuis 1967, et il faut mesurer la qualité de l'appariement sur le champ réduit aux individus que l'on pouvait effectivement retrouver dans le SNDS, soit les individus en vie sur la période 2008-2018 et dont le NIR a pu être identifié. Pour cet échantillon restreint à 3,4 millions d'individus, le taux de personnes retrouvées dans le SNDS est de 98,2 %.

Les étapes de cet appariement sont résumées dans la figure 7.

Figure 7 • Synthèse des étapes de l'appariement de l'EDP au SNDS



Du champ théorique de l'EDP à la base études EDP

Le champ de l'EDP est défini théoriquement sur la base de jours de naissances (4 jours puis 16 jours à partir de 2004). En pratique, les individus concernés peuvent être inclus dans ce champ de deux façons :

- D'une part les personnes identifiées au RNIPP (répertoire national d'identification des personnes physiques, construit à partir de l'État civil) sur ces jours de naissances sont constitutives du champ de l'EDP. Les informations à leur sujet sont leurs NIR, et leur identité.
- D'autre part, toute personne qui apparaît dans une des sources de l'EDP avec une date de naissance comprise dans le champ, qu'elle soit ou non identifiée au RNIPP, rejoint également le champ de l'EDP.

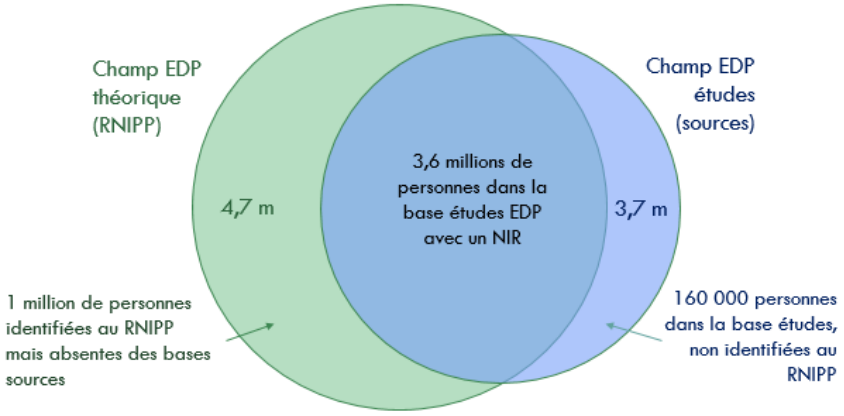
Ces champs ne se recoupent pas totalement et c'est finalement la deuxième définition qui correspond au champ de la base études mise à disposition des utilisateurs (figure 8)[20].

1 million d'individus qui sont inscrits au RNIPP sur les bons jours de naissance et devraient être dans le champ de l'EDP ne sont pas retrouvés dans la base études, car n'apparaissent dans aucune source. Cela tient au fait que l'historique des sources n'a pas été récupéré lorsque le champ a été quadruplé en 2004. Il s'agit donc principalement de personnes ne résidant plus sur le territoire depuis l'élargissement du champ (un certain nombre est probablement décédé, mais le bulletin de décès n'est pas remonté à l'état civil), ou de personnes

[20] La « population EDP » et les « individus EDP » nommés dans la suite du dossier font donc référence aux personnes présentes dans la base études de l'EDP.

nées avant 2004, mais trop jeunes pour apparaître dans une des sources[21]. Une partie de ces individus a d'ailleurs été retrouvée dans le SNDS, ce qui nous permet de mieux caractériser ces personnes (voir encadré 2).

Figure 8 • Recoupement des deux définitions du champ de l'EDP



À l'inverse, 160 000 individus présents dans la base études n'ont pas été retrouvés au RNIPP et ne sont donc pas associés à un NIR (par exemple une personne vivant en France et recensée mais non inscrite à l'État civil). Ces personnes étant présentes dans la base études, on peut connaître leurs caractéristiques grâce à l'EDP.

Encadré 2 • Caractéristiques des personnes entrant dans le champ théorique de l'EDP, absentes des sources EDP et pourtant retrouvées dans le SNDS

Pour l'extraction des données du SNDS, les informations liées à l'ensemble des personnes présentes dans le champ théorique de l'EDP ont été transmises à la CNAM. Cela correspond donc à 4,7 millions de NIR qui ne sont pas tous associés à un individu de la base études de l'EDP. Parmi ceux qui ne sont pas associés à un individu de l'EDP, une grande partie n'a pas été retrouvée dans le SNDS[22] : ces personnes n'apparaissant dans aucune source de l'EDP, elles ont pu quitter le territoire ou être décédées avant la construction du SNIIRAM-PMSI. En revanche, pour 226 171 individus, des informations ont été retrouvées dans le SNDS, ce qui permet de mieux comprendre le décalage entre le champ théorique des jours de naissance de l'EDP et le champ pratique de la base études.

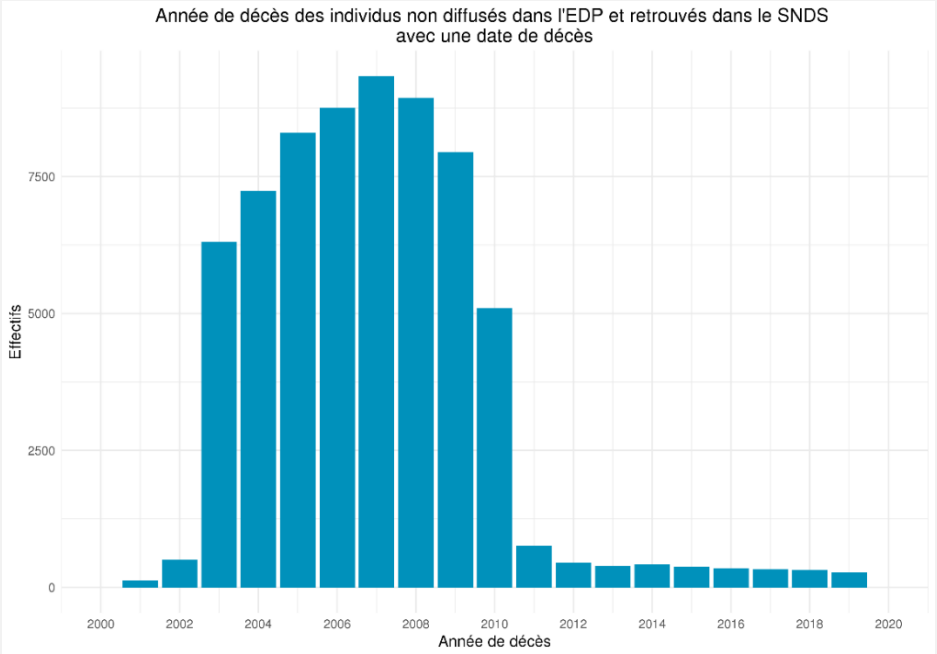
La moitié de ces 226 171 individus est née avant 1942 (78 ans ou plus en 2020), et un quart est né avant 1924 (96 ans en 2020). Pour les individus présents dans la base études de l'EDP, l'année de naissance médiane est 1973 (47 ans en 2020), ce qui confirme la surreprésentation de personnes âgées et potentiellement décédées parmi ces personnes non retrouvées dans les sources de l'EDP.

Parmi les individus retrouvés dans le SNDS et absents de l'EDP, seuls 36 % ont eu recours à au moins une prestation de soins entre 2013 et 2020, contre 95 % des individus présents dans l'EDP.

Enfin, si la date de décès n'est pas une information remontée exhaustivement dans le SNDS, elle est tout de même renseignée pour 29 % de ces individus, et 22 % sont décédés avant 2008.

[21] Depuis 2004, la seule source quasi-exhaustive sur l'échantillon est le fichier de déclaration fiscale, mais il ne capte pas bien les mineurs. Ainsi, une personne née en 2002 sur un des trois mois de naissances ajoutés en 2004 a rejoint le champ de l'EDP à ce moment, mais son bulletin de naissance n'a pas été récupéré dans l'EDP. Si cette personne n'a pas été recensée depuis (depuis 2004, 8 % de la population est recensée chaque année), et si elle n'a pas été identifiée dans les données fiscales (l'identification des mineurs est plus compliquée que celle des personnes majeures), il n'y a pas d'informations sur elle dans l'EDP.

[22] Pour retrouver les individus dans le SNDS, il suffit qu'ils apparaissent dans le référentiel des bénéficiaires, dont le champ temporel est plus large que celui de l'appariement : les données de l'appariement portent sur les prestations de soins ayant eu lieu à partir de 2008, mais le référentiel des bénéficiaires et son historique portent sur l'ensemble des bénéficiaires couverts dans le SNDS. Il est donc possible de retrouver des individus décédés avant 2008 s'ils ont consommé au moins une fois sur la période couverte par le SNDS, mais leurs données de consommations de soins ne seront pas disponibles dans l'appariement.



Champ > Individus présents dans le champ de l'EDP car retrouvés au RNIPP mais absents de la base études de l'EDP, et retrouvés dans le SNDS.

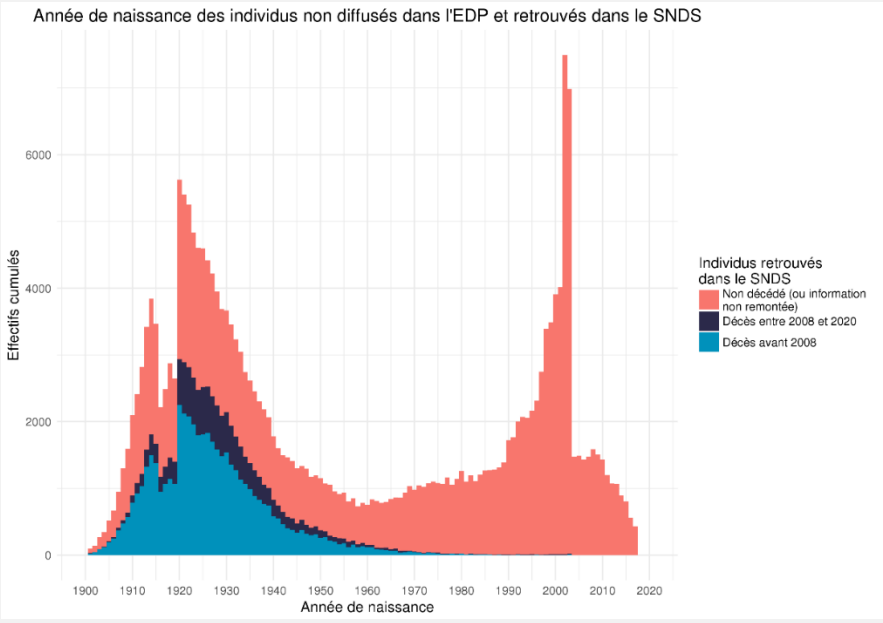
Source > EDP-Santé.

La pyramide des âges de ces personnes retrouvées dans le SNDS semble corroborer l'hypothèse de deux phénomènes à l'origine de l'absence des personnes du champ théorique de l'EDP dans la base études :

- la surreprésentation des personnes âgées, et en particulier de personnes âgées décédées avant 2008, liée à des personnes probablement hors du champ de l'EDP-Santé ;
- la surreprésentation des personnes mineures nées avant 2004, liée à l'ajout des 12 jours de naissance supplémentaires, sans rapatriement des données historiques de l'EDP.

Dans ce second cas, l'absence de ces personnes dans l'EDP peut affecter la représentativité de l'échantillon, mais le calage des poids des individus dans les différentes sources doit permettre de pallier ce biais ; de plus dans les prochaines années, il est fort probable que ces personnes apparaissent dans une des bases sources de l'EDP et soient donc réintégrées à la base étude.

Sans information disponible sur ces personnes dans l'EDP, on choisit donc de les retirer de la base études de l'EDP-Santé.



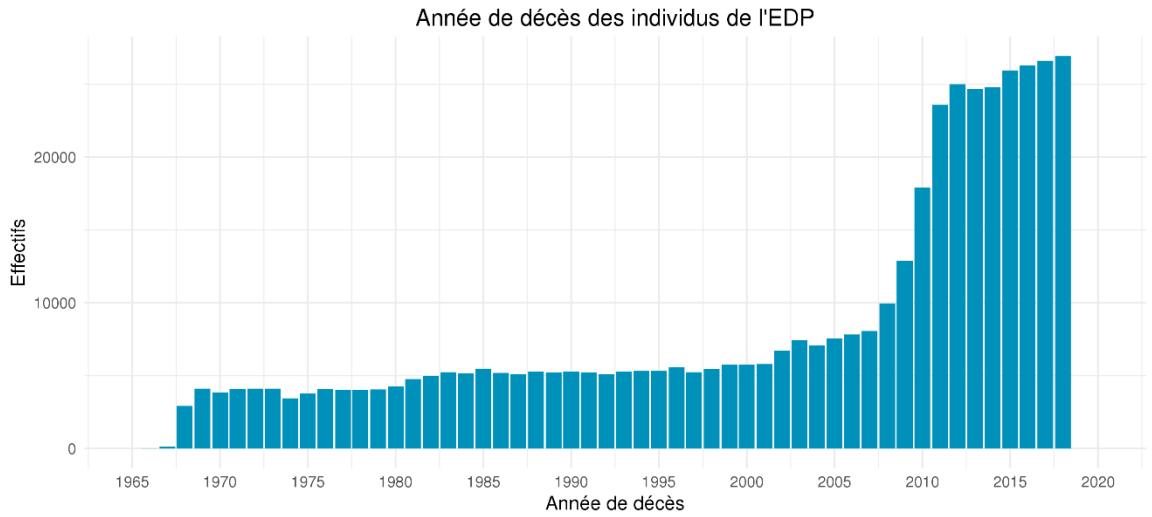
Champ > Individus présents dans le champ de l'EDP car retrouvés au RNIPP mais absents de la base études de l'EDP, et retrouvés dans le SNDS.

Source > EDP-Santé.

De la base études de l'EDP au champ de l'EDP-Santé : les personnes susceptibles d'avoir consommé des soins depuis 2008

Les personnes décédées avant 2008 n'étant pas susceptibles d'avoir consommé sur la période de soins extraite pour l'EDP-Santé, elles sont exclues du champ de l'EDP-Santé. Cela correspond à 208 000 personnes. Après cette exclusion, l'échantillon est composé de 3,5 millions d'individus EDP, en vie après 2008, dont 245 000 sont décédés au cours de la période d'étude (figure 9).

Figure 9 • Répartition des individus EDP décédés selon l'année de décès

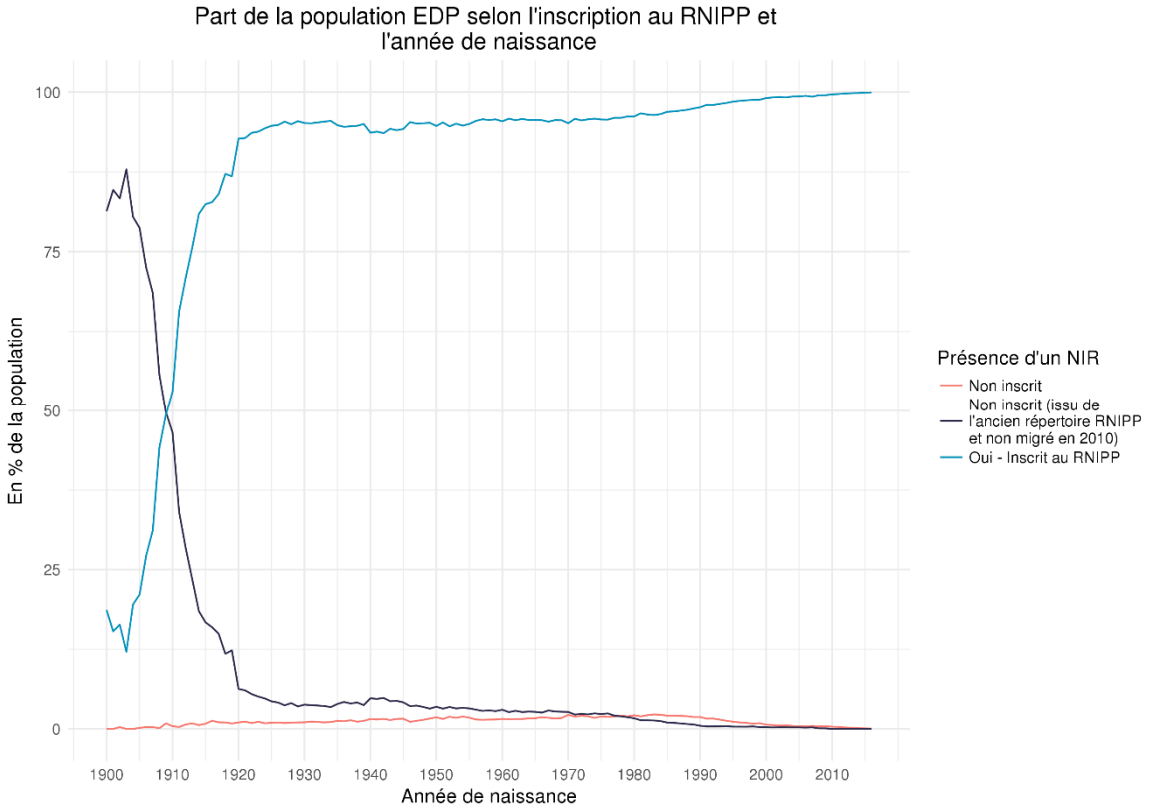


Champ > Individus présents dans la base études de l'EDP.
Source > EDP-Santé.

Parmi les individus EDP en vie après 2008, 140 000 personnes, soit 4 %, ne sont pas associées à un NIR et ne peuvent pas être identifiées dans le SNDS. Il s'agit de personnes repérées dans une des sources de l'EDP, mais qui ne sont pas inscrites au RNIPP, comme expliqué plus haut, ou d'individus qui n'ont pas été retrouvés dans le nouveau répertoire lors du basculement en 2011 (voir figure 10)[23]. La distribution par âge de ces individus révèle qu'il s'agit d'une population plus âgée dont une partie a pu décéder sans que l'information ne soit remontée dans l'EDP (21 300 sont nés avant 1920). Pour un grand nombre cependant, il s'agit de personnes nées à l'étranger dont la reconstruction du NIR n'a pas été possible, en particulier lors de la migration vers le nouveau répertoire (Jugnot, 2012) : 92 % d'entre elles sont nées à l'étranger et si 99,5 % des individus EDP nés en France ont bien un NIR, ce n'est le cas que de 79 % de ceux nés à l'étranger. Pour 40 % de ces personnes nées à l'étranger, l'information ne provient que d'un recensement exhaustif, 28 % sont apparues dans une enquête annuelle de recensement, et 13 % apparaissent uniquement à l'État civil. Seules 7 % d'entre elles apparaissent au moins une fois dans les sources Fideli ou Filosofi (soit 11 000 individus). Il paraît donc très probable que ces personnes ne soient pas restées longtemps sur le territoire français et aient été identifiées uniquement lors d'un recensement.

[23] Jusqu'en 2009, l'EDP gérait son propre répertoire (avec des rapprochements épisodiques avec le RNIPP). À partir de 2010, l'Insee a basculé sur une gestion mutualisée, qui assure une meilleure qualité à l'identification. Mais, cela a posé des difficultés pour basculer une partie des individus de l'ancien répertoire, principalement pour les personnes nées à l'étranger. (Jugnot, 2014).

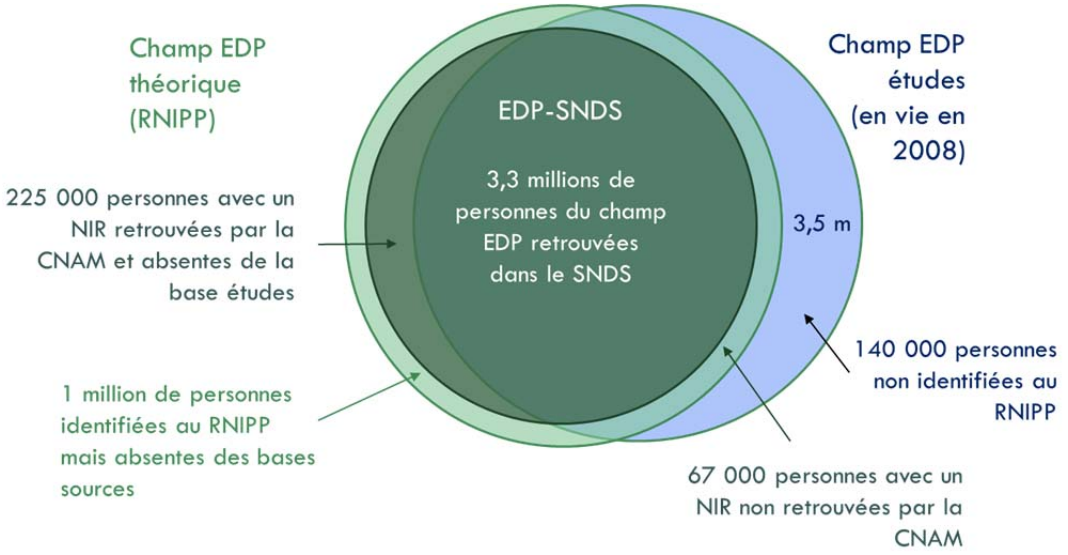
Figure 10 • Répartition par année de naissance de la population EDP avec et sans NIR



Champ > Individus présents dans la base études de l'EDP.
Source > EDP-Santé.

Parmi les 3,4 millions d'individus présents dans l'EDP avec un NIR, on retrouve finalement dans le SNDS 3,3 millions d'individus, soit 98 % des NIR envoyés pour des personnes dans le champ. 67 000 individus n'ont donc pas été repérés dans le SNDS malgré l'envoi de leur NIR (figure 11).

Figure 11 • Recoupement des champs EDP et SNDS pour les personnes en vie après 2008

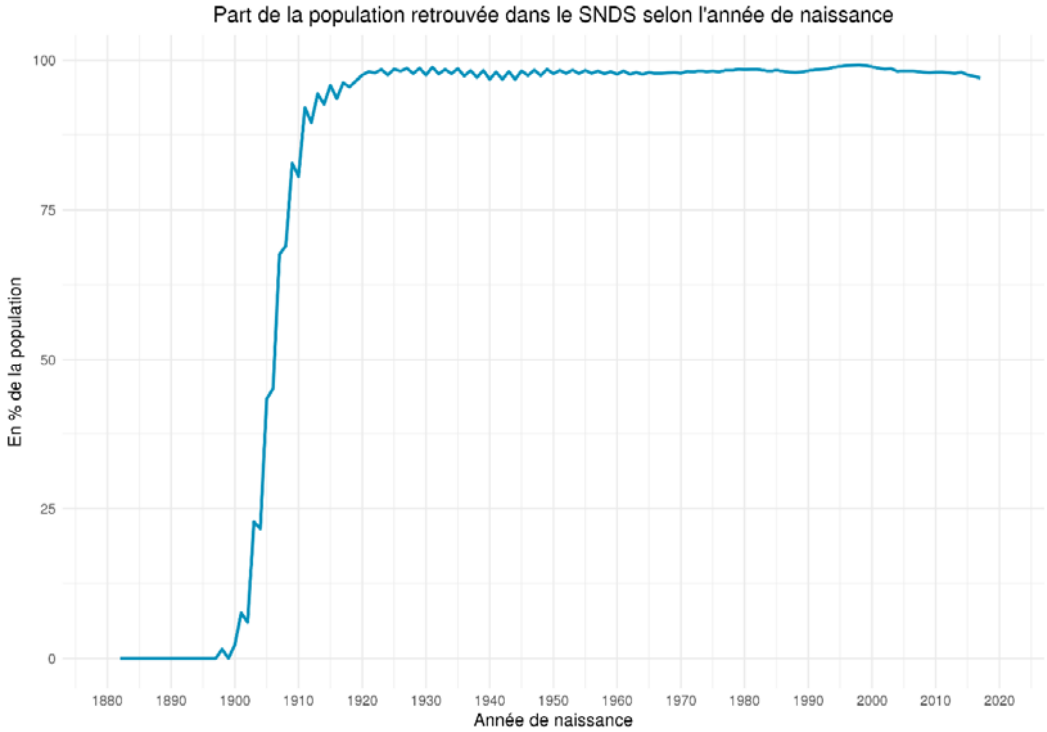


Ces 67 000 personnes peuvent être des personnes dont le NIR n'a pas été bien retrouvé dans le SNDS. C'est notamment possible dans le cas où un individu consomme des soins en étant ayant-droit d'une autre personne, et que son propre NIR n'est pas connu de la CNAM. Pour la plupart, les caisses de remboursement de l'Assurance maladie transmettent les NIR propres aux ayants-droit (ou NIR individuel) ce qui permet de les retrouver, mais elles ne le font pas toutes. Ainsi, un enfant qui consomme des soins au titre des droits ouverts par

un de ses parents et dont le NIR individuel n'est pas remonté administrativement par la caisse d'Assurance maladie de ce parent ne pourra pas être retrouvé[24]. Ceci explique la surreprésentation des jeunes mineurs (principale composante de la population ayant-droit) parmi les personnes non retrouvées. Néanmoins, l'amélioration de la qualité de remontée du NIR individuel dans le SNDS ces dernières années limite l'ampleur de ce problème. L'absence du NIR individuel ne concerne que 3 % de l'ensemble des bénéficiaires, et environ 5 % des moins de 18 ans en 2019, ce qui signifie qu'il doit être possible de retrouver 95 % des enfants sur la base de leur NIR individuel.

Parmi ces personnes non retrouvées, il peut également y avoir des personnes qui n'ont pas consommé de soins sur la période. Cela s'explique soit parce qu'elles n'ont eu aucun recours au système de soins français sur la période couverte par le SNDS, mais cette hypothèse paraît peu réaliste, soit parce qu'elles ont quitté le territoire[25]. Ainsi, parmi les 67 000 individus non appariés, 20 % apparaissent uniquement dans les DADS. On constate d'ailleurs un moins bon taux d'appariement des personnes nées les années paires entre 1920 et 1970, que l'on peut directement relier au champ des DADS : jusqu'en 2002, il intégrait uniquement les personnes nées en octobre des années paires. N'apparaissant dans aucune autre source, il y a de bonnes raisons de penser que ces individus ont pu décéder ou quitter le territoire et ne devraient donc dans tous les cas pas être inclus dans le champ. En particulier, l'absence du fichier fiscal sur la période couverte par la source (déclarations collectées à partir de 2011, pour un revenu perçu ou une taxe d'habitation versée en 2010) semble peu vraisemblable pour un résident français majeur.

Figure 12 • Part de la population retrouvée dans le SNDS selon l'année de naissance



Champ > Individus EDP en vie après 2008 et associés à un NIR.
Source > EDP-Santé.

[24] Lors d'appariements sur des enquêtes dont l'échantillon inclut des mineurs, il est possible de collecter le NIR de l'ouvrant-droit en plus du NIR individuel, et la CNAM est en mesure de retrouver les individus ayant-droit par ce biais. Dans le cas de l'EDP ce n'était pas possible puisque seuls les NIR des individus EDP sont disponibles.

[25] La CNAM garde une archive des bénéficiaires ayant consommé des soins au moins une fois depuis la création de sa base de données. Les individus qui n'ont pas été retrouvés seraient donc des personnes qui n'ont pas consommé de soins sur les quinze dernières années (même si le système d'information est monté en charge et a gagné en qualité sur les années les plus récentes).

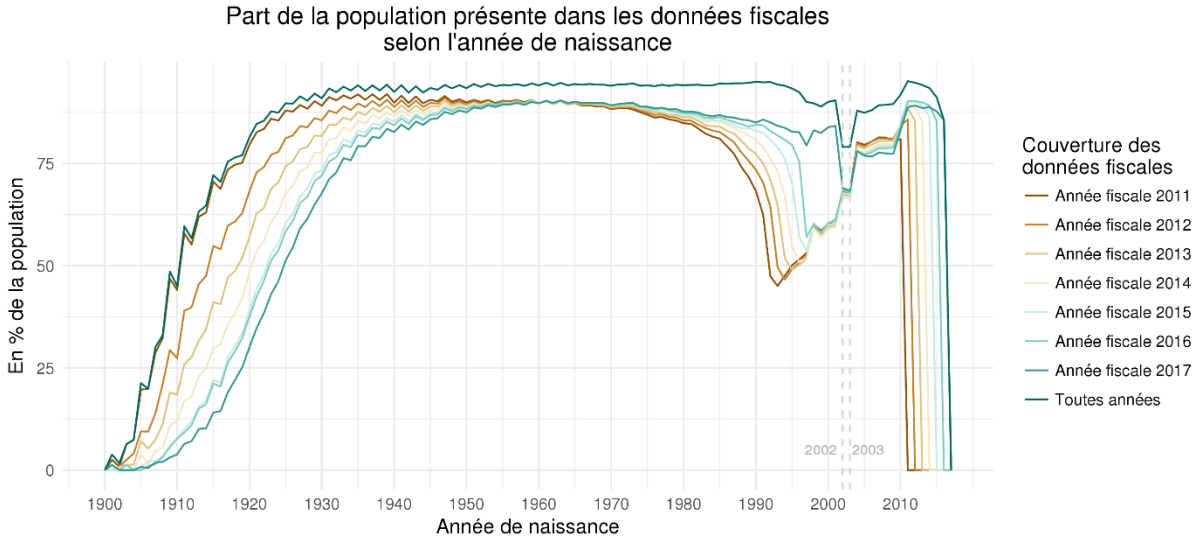
Le champ de l'EDP-Santé restreint aux personnes présentes dans les données fiscales

Pour s'assurer de la présence des personnes de l'EDP-Santé sur le territoire au cours de la période couverte, on restreint donc dans un dernier temps le champ de l'EDP aux personnes apparues au moins une fois dans les sources fiscales (Fideli ou Filosofi) depuis 2011, ce qui représente 3,1 millions de personnes. Les personnes qui ne sont pas dans les données fiscales sont celles qui n'ont pas de taxe d'habitation ni de déclaration de revenu en France depuis 2011. Si elles sont majeures, cela devrait concerner une minorité de personnes dont on peut plus vraisemblablement penser qu'elles ne résident plus en France. Pour information, les données de Fideli constituent dorénavant le nouvel échantillon maître de l'Insee (projet Nautile) pour le tirage d'échantillons d'enquêtes.

Parmi les personnes présentes dans les données fiscales, 99,6 % sont identifiées au RNIPP et sont donc associées à un NIR, ce qui introduit beaucoup moins de biais liés à l'absence de NIR. En outre, cela garantit une information disponible sur le niveau de vie et la situation du ménage fiscal de l'individu EDP.

Enfin, parmi les 3,1 millions d'individus EDP présents dans les données fiscales, 99,5 % sont retrouvés dans le SNDS, offrant donc une très bonne couverture en matière de données de santé. Cette bonne couverture se retrouve sur l'ensemble des années de naissance des individus EDP.

Figure 13 • Part de la population de l'EDP présente dans les données fiscales selon l'année de naissance



Champ > Individus EDP en vie après 2008 et associés à un NIR.
Source > EDP.

La restriction du champ aux données fiscales implique tout de même quelques limites :

- D'une part, les mineurs ne sont pas bien identifiés dans les données fiscales car l'information est donnée indirectement par les parents sur les déclarations, et ne permet pas toujours le chaînage. Les bulletins de naissance des enfants collectés à l'État civil permettent plus facilement de repérer leurs parents dans les données fiscales et d'associer les bulletins de leurs enfants lorsque la date de naissance coïncide. Cependant pour les enfants nés avant 2004 sur un des trois mois de naissance de l'élargissement, les bulletins de naissance n'ont pas été collectés rétrospectivement, ce qui explique la dégradation de la qualité de la couverture pour les personnes nées avant 2003. Après 18 ans, la couverture s'améliore sensiblement du fait du basculement sur la déclaration directement nominative (figure 13). Depuis 2016, le seuil de passage de la déclaration indirectement nominative à la déclaration directement nominative pour les mineurs a été abaissé de 18 ans à 14 ans, améliorant la couverture des 14-18 ans. Si le taux de couverture de la population EDP reste moins bon pour les mineurs, à hauteur de 80 % environ, il devrait donc continuer à s'améliorer dans les années à venir. Dans tous les cas, un poids est calculé sur les marges de Fideli séparément pour les adultes et pour les mineurs, afin de tenir compte de cette moins bonne représentativité. Il permet donc de corriger ce biais.
- D'autre part, le champ temporel couvert par les données fiscales limite celui de l'EDP : les personnes décédées avant 2011, année de la première déclaration dans Fideli-Filosofi, en sont absentes, de même que les

nouveau-nés de la dernière année du millésime, puisqu'ils ne sont pas inclus dans la déclaration fiscale portant sur l'année précédente. En l'occurrence dans le millésime 2017, aucune information n'est disponible sur les enfants nés en 2017. Ainsi les études peuvent porter sur les personnes en vie entre 2011 et 2016.

Deux périmètres de l'EDP-Santé à exploiter selon les besoins

En définitive, l'EDP-Santé peut être exploité sur deux périmètres, à adapter selon les besoins de l'étude :

- Si le champ des personnes en vie entre 2011 et 2016 convient pour l'étude, il est préférable de se restreindre aux personnes présentes dans les données fiscales, dont le taux d'appariement avec le SNDS est très élevé, et dont la représentativité est déjà permise par les données et le calcul d'une pondération spécifique dans l'EDP ;
- Si l'étude doit porter sur une population aux limites des données fiscales, ou si un échantillon de plus grande taille est nécessaire, il est préférable d'élargir l'échantillon et de réfléchir à une pondération spécifique.

Les caractéristiques de ces deux extractions sont résumées dans le tableau 2.

Figure 14 • Part de la population de l'EDP présente dans les données fiscales (2011-2017) selon l'année de naissance

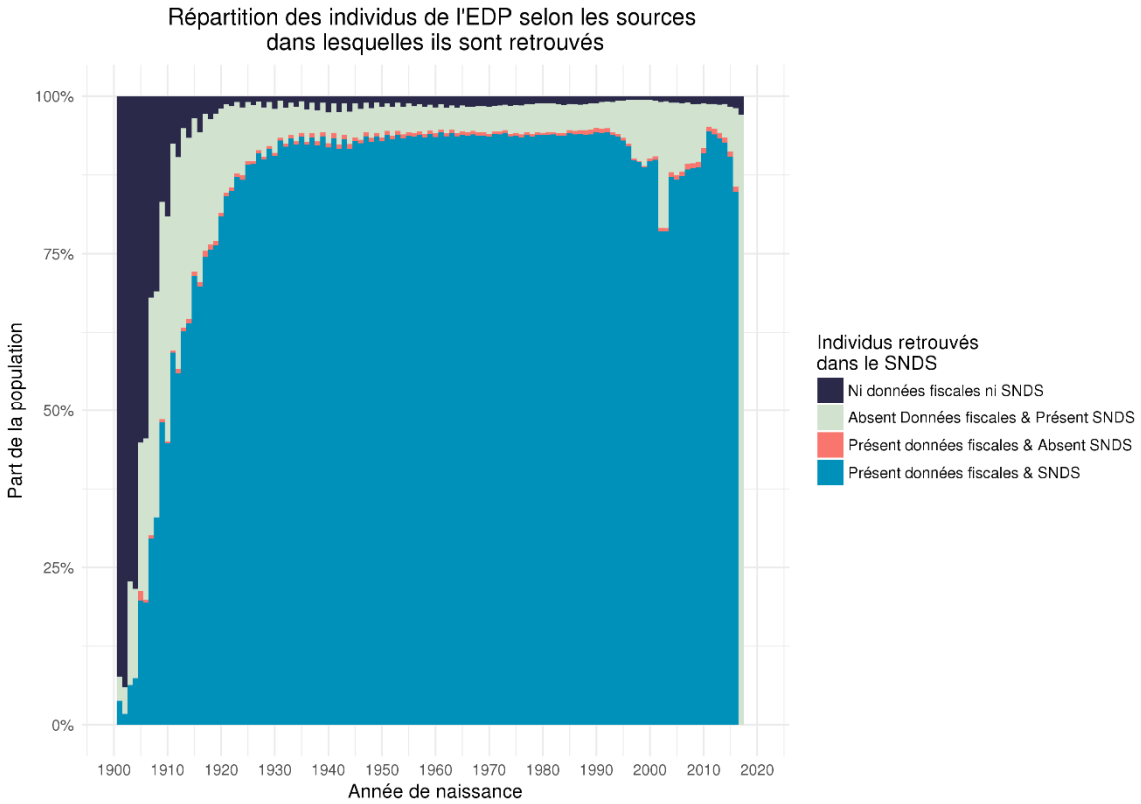


Tableau 2 • Caractéristiques des deux périmètres d'étude recommandés pour l'EDP Santé

Individus EDP vivants en 2008 (ou nés ultérieurement)	Individus EDP vivants en 2008 (ou nés ultérieurement) et présents dans les données fiscales
96% dispose d'un NIR -> 3,4 millions de personnes	99,6% dispose d'un NIR -> 3,1 millions de personnes
98% d'appariement au SNDS	99,5% d'appariement au SNDS
3,3 millions d'individus dans l'appariement final	3,1 millions d'individus dans l'appariement final
Avantages : - effectifs plus élevés - toute la population ayant pu consommer sur 2008-2017 présente	Avantages : - Quasiment tous les individus couverts par les données fiscales ont un NIR - Poids annuel déjà calculé pour représentativité de l'échantillon par rapport à la population Fideli

	- Au moins une information sur le niveau de vie pour tous ces individus
Inconvénients : - Potentiellement des personnes qui ne résident plus sur le territoire (personnes majeures non décédées et sans données fiscales) - Biais possible dû à l'absence de NIR des personnes nées à l'étranger	Inconvénients : - Absence des personnes décédées avant 2011 - Absence des nouveau-nés de 2017 - Difficultés d'identification des mineurs, mais une nette amélioration sur les dernières années

Mise à disposition des données

Documentation pour les utilisateurs

Chaque étude mobilisant l'EDP-Santé nécessitera une bonne compréhension et utilisation d'une information qui est complexe notamment car elle est multi-sources. Un enjeu du projet d'appariement est donc aussi de faciliter ces exploitations en apportant une documentation suffisante.

- Pour l'utilisation de l'EDP, différentes ressources sont disponibles sur le site internet des utilisateurs de l'EDP, en particulier un document détaillé mis à disposition par l'Insee pour chaque millésime de l'EDP et décrivant les spécificités de chaque source.

  Par ailleurs, la DREES a élaboré un dictionnaire interactif des variables et des tables de l'EDP, disponible en accès restreint, sur demande à la DREES.
- Pour l'utilisation du SNDS, un certain nombre de documents sont accessibles sur le portail des données du SNDS hébergé par la CNAM. Il existe de plus un site internet de documentation du SNDS accessible en ligne. Un forum public d'entraide a également été mis en place pour partager des questions, qu'elles émergent en amont de l'accès aux données ou pendant l'exploitation.

CONCLUSION

L'appariement de l'échantillon démographique permanent au système national des données de santé est apparu comme la réponse la plus adaptée et la plus riche pour mesurer l'ampleur des inégalités sociales de santé, comprendre leurs mécanismes et étudier leur évolution sur les années récentes. La constitution de cette base, l'EDP-Santé, vient ainsi compléter le paysage des données existantes sur la question telles que l'enquête santé européenne (EHIS), les cohortes épidémiologiques ou les dispositifs d'enquêtes plus ciblés sur des thématiques (santé au travail, ou critères pathologiques). Elle doit permettre l'évaluation de la stratégie nationale de santé 2018-2022.

Les appariements de données sont des opérations qui se sont multipliées sur les dernières années. Ils sont encouragés dans la statistique publique car ils allègent les coûts des enquêtes, permettent de diminuer la charge des questionnaires auprès des ménages ou des entreprises, et de contourner des biais de déclaration ou de mémoire souvent présents dans les données d'enquêtes. La DREES, pleinement inscrite dans la démarche de favoriser les appariements, bénéficie aujourd'hui de l'expérience de plusieurs enrichissements de ses enquêtes, notamment à partir des données médico-administratives du SNDS (Montaut, 2013, Carrère, 2016). L'évolution du contexte juridique autour de ces données (création du SNDS en 2016, loi relative à l'organisation et la transformation du système de santé en 2019) et de la protection des données personnelles de manière plus générale (règlement général sur la protection des données en 2016), a pour autant imposé de nombreuses adaptations du projet et une vigilance particulière sur le respect des procédures liées à la confidentialité de ces données sensibles. Le projet a donc été conçu de façon à garantir une sécurité maximale dans le circuit de transmission des données, ainsi que dans l'hébergement et l'accès à la base finale, dont la seule finalité est celle de permettre la recherche et l'évaluation sur les inégalités sociales de santé. Bien que moindre par rapport à celui d'une enquête, le coût d'un appariement n'est donc pas à négliger et ne se résume pas à l'envoi de données déjà existantes. Il mobilise des ressources juridiques et techniques dans différentes institutions, en amont de la mise en œuvre pour concevoir le traitement dans le respect de la protection des données personnelles, et en aval pour expertiser la qualité des données réceptionnées. Pour l'EDP-Santé, l'ensemble de cette chaîne de production s'est étiré sur quatre années, de septembre 2017 à juillet 2020. On peut néanmoins espérer que la multiplication de ces opérations permettra de réduire le temps et l'investissement nécessaires à un appariement au cours des prochaines années.

L'EDP-Santé, aujourd'hui disponible sur un serveur sécurisé à la DREES, ouvre désormais la perspective de nombreuses études permettant d'évaluer la stratégie nationale de santé 2018-2022. L'expertise des données reçues a en effet permis de mettre en évidence un très bon taux d'appariement, de 98 % pour l'ensemble des personnes vivantes en 2008 ou nées après 2008 et disposant d'un NIR dans l'EDP, et de 99,5 % pour l'ensemble de ces personnes par ailleurs présentes dans la source fiscale depuis 2011. L'utilisation des poids existant dans cette source, ou la reconstruction de poids sur un périmètre adapté, garantiront des résultats représentatifs de la population résidant sur le territoire français et la disponibilité d'une grande quantité d'informations sur la situation socio-économique des individus dans l'EDP.

Au-delà des études à mener sur l'EDP-Santé, les perspectives de prolongement du projet ne manquent pas. La priorité doit maintenant porter sur la mise à disposition de ces données à la recherche. Celle-ci passera nécessairement par une réflexion sur la mise en œuvre de cette ouverture, en particulier sur la procédure d'accès aux données, la responsabilité des traitements et la solution d'hébergement de ces données. Une autre piste d'ouverture consiste à enrichir ces données afin de notamment permettre des études sur la santé périnatale en ajoutant les données liées aux nouveau-nés des mères EDP.

BIBLIOGRAPHIE

- Blanpain, N. (2011) L'espérance de vie s'accroît, les inégalités sociales face à la mort demeurent. *Insee première,* n°1372, Insee, octobre.
- Blanpain, N. (2018) L'espérance de vie par niveau de vie : chez les hommes, 13 ans d'écart entre les plus aisés et les plus modestes. *Insee première,* n°1687, Insee, février.
- Carrère, A. (2016) Les enrichissements prévus pour l'enquête CARE-Ménages, *Drees Document de travail,* n°56, janvier.
- Chardon, O., Guignon, N., de Saint Pol, T., Guthmann, J.-P., Ragot, M., Delmas, M.-C., Paget, L., Perrine, A.-L., Thélot, B. (2015) La santé des élèves de grande section de maternelle en 2013 : des inégalités sociales dès le plus jeune âge*, Études et Résultats*, n°920, Drees, juin.
- DREES (2015), Données de santé : anonymat et risque de ré-identification, *Dossiers solidarité et santé*, n°64, Drees, juillet.
- DREES (2017), Les inégalités sociales de santé, Actes du séminaire de recherche de la Drees 2015-2016.
- Ducros, D., Nicoules, V., Chehoud, H., Bayle, A., Souche, A., Tanguy, M. & Grosclaude, P. (2015) Les bases médico-administratives pour mesurer les inégalités sociales de santé. *Santé Publique*, pp. 383-394.
- Fosse-Edorh S, Mandereau-Bruno L. (2015) Suivi des examens recommandés dans la surveillance du diabète en France en 2013. *Bull Epidémiol Hebd.* ;(34-35):645-54. 62(8).
- Geoffreoy-Perez, B. (2006) Analyse de la mortalité et des causes de décès par secteur d'activité de 1968 à 1999 à partir de l'Échantillon démographique permanent, Rapport « Cosmop », InVS, septembre.
- Guthmann, J.-P., Pelat, C., Célant, N., Parent du Chatelet, I., Duport, N., Rochereau, T. et Lévy-Bruhl, D. (2016) Inégalités socioéconomiques d'accès à la vaccination contre les infections à papillomavirus humains en France : résultats de l'Enquête santé et protection sociale (ESPS), *Bulletin Epidémiologique Hebdomadaire (BEH),* pp. 288-297, juin.
- HCAAM (2011), « L'accessibilité financière des soins : comment la mesurer ? », Avis du Haut Conseil de l'avenir de l'Assurance maladie, novembre.
- HCSP (2009), Les inégalités sociales de santé : sortir de la fatalité, Rapport, décembre.
- Jugnot S. (2014) La constitution de l'échantillon démographique permanent de 1968 à 2012, *Insee Document de travail*, n°F1406, septembre.
- Montaut, A., Calvet, L., Bouvier, G., Gonzalez, L. (2013) L'appariement handicap-santé et données de l'Assurance-maladie, *Drees Document de travail,* n°39, janvier.
- OMS (2009), Combler le fossé en une génération : instaurer l'équité en santé en agissant sur les déterminants sociaux de la santé. Rapport final de la Commission des déterminants sociaux de la santé, 246 p.
- Tuppin P., Rudant, J., Constantinou, P., Gastaldi-Ménager, C., Rachas, A., Roquefeuil, L., Maura, G., Caillol, H., Tajahmady, A., Coste, J., Gissot, C., Weill, A., Fagot-Campagna, A.. (2017) L'utilité d'une base médico-administrative nationale pour guider la décision publique : du système national d'information interrégimes de l'Assurance Maladie (SNIIRAM) vers le système national des données de santé (SNDS) en France. *Revue d'épidémiologie et de santé publique.*
- Rey, G., Jougla, E., Fouillet, A.,, et Hémon D. (2009). "Ecological Association between a Deprivation Index and Mortality in France over the Period 1997 - 2001: Variations with Spatial Scale, Degree of Urbanicity, Age, Gender and Cause of Death." *BMC Public Health* 9: 33.
- Vilain, A., Gonzalez, L., Rey, S., Matet, N., Blondel, B. (2013), Surveillance de la grossesse en 2010 : des inégalités socio-démographiques*, Études et Résultats*, n°848, Drees, juillet.

Annexe 1. Composition du groupe de travail juridique

Les personnes ayant participé aux trois séances du groupe de travail juridique sur les aspects liés à l'appariement EDP-SNDS sont les suivantes (par institutions) :

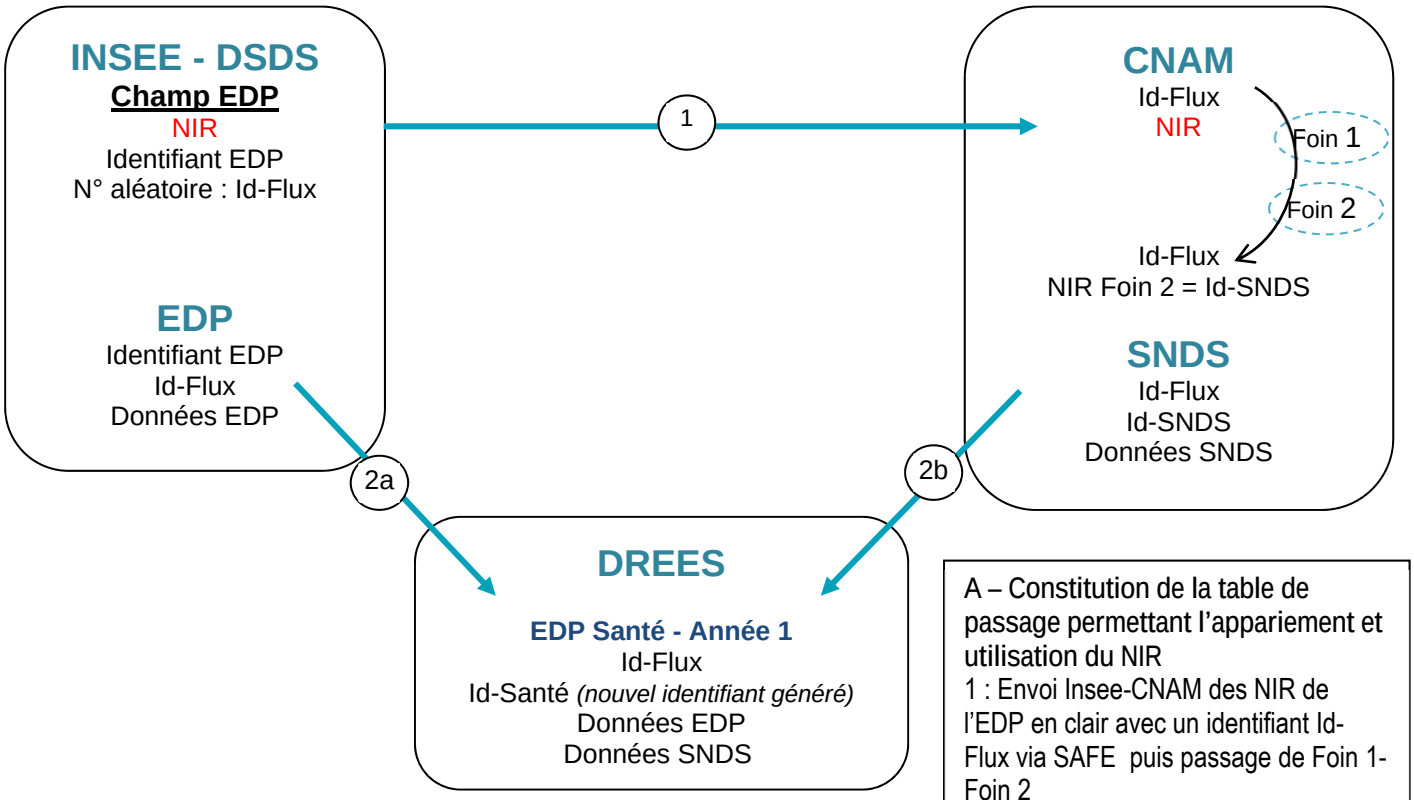
CASD : Anaïs Bréchard, Eric Debonnel, Kamel Gadouche

CNAM : Hélène Caillol, Stéphanie Naux, Anne Poisson

DREES : Claire-Lise Dubost, Pierre Fleutiaux, Aude Leduc, Javier Nicolau, Matthias Pigneur, Philippe Raynaud,

INSEE : Jean-Luc Bottet, Mylène Chaleix, Frédéric Comte, Sébastien Durier, Patrick Redor, Marie Reynaud, Isabelle Robert-Bobée

Annexe 2. Circuit des flux de données pour l'appariement



A – Constitution de la table de passage permettant l'appariement et utilisation du NIR
1 : Envoi Insee-CNAM des NIR de l'EDP en clair avec un identifiant Id-Flux via SAFE puis passage de Foin 1-Foin 2
B - Rapprochement des données à l'aide de l'identifiant commun
2a : Envoi Insee-Drees des données de l'EDP avec Id-Flux
2b : Envoi CNAM-Drees des données SNDS avec Id-Flux

Annexe 3. Glossaire des sigles

ACS	Aide à l'acquisition d'une complémentaire santé
AIPD	Analyse d'impact relative à la protection des données
AME	Aide médicale de l'État
BRPP	Base des répertoires des personnes physiques (Insee)
CARE	Enquête Capacités, Aides et Ressources des seniors
CASD	Centre d'accès sécurisé aux données
CépiDC	Centre d'épidémiologie sur les causes médicales de Décès
CEREES	Comité d'expertise pour les recherches, les études et les évaluations dans le domaine de la santé
CESREES	Comité éthique et scientifique pour les recherches, études et évaluations dans le domaine de la santé
CMU-C	Couverture maladie universelle complémentaire
CNAM	Caisse nationale de l'Assurance maladie
CNIL	Commission nationale de l'informatique et des libertés
CNIS	Conseil national de l'information statistique
CNSA	Caisse nationale de solidarité pour l'autonomie
COSMOP	Cohorte pour la surveillance de la mortalité par profession
CPP	Comités de protection des personnes
DADS	Déclaration annuelle des données sociales
DARES	Direction de l'animation de la recherche, des études et des statistiques
DEMEX	Cellule de la CNAM en charge de l'accompagnement des demandes d'extraction
DREES	Direction de la recherche, des études, de l'évaluation et des statistiques
DSDS	Direction des statistiques démographiques et sociales (Insee)
EAR	Enquête annuelle de recensement
EDP	Échantillon démographique permanent
EGB	Échantillon généraliste des bénéficiaires
EHIS	European Health interview survey
ESPS	Enquête Santé et Protection Sociale
Fideli	Fichier Démographique des logements et des individus
Filosofi	Fichier Localisé social et fiscal
FOIN	Fonction d'occultation d'informations nominatives
GENES	Groupe des écoles nationales d'économie et statistique
INDS	Institut National des données de santé
Ined	Institut national des études démographiques
INSEE	Institut national de la statistique et des études économiques
Inserm	Institut national de la santé et de la recherche médicale
INVS	Institut de veille sanitaire
Irdes	Institut de recherche et documentation en économie de la santé
LIL	Loi informatique et libertés

MDPH	Maisons départementales des personnes handicapées
NIR	Numéro d'inscription au répertoire national d'identification des personnes physiques
PMSI	Programme de médicalisation des systèmes d'information
PSCE	Enquête Protection sociale complémentaire d'entreprise
RGPD	Règlement général sur la protection des données
RNIPP	Répertoire national d'identification des personnes physiques
SNDS	Système national des données de santé
SNIIRAM	Système national d'information inter-régimes de l'Assurance maladie
SNS	Stratégie nationale de santé
SRCV	Statistiques sur les ressources et conditions de vie
SSM	Service statistique ministériel
SSP	Service statistique public

Les dossiers de la DREES

N° 66 • septembre 2020

L'EDP-Santé : Un appariement des données socio-économiques de l'échantillon démographique permanent du Système national des données de santé

Directeur de la publication
Fabrice LENGLART

Responsable d'édition
Souphaphone Douangdara

ISSN
2495-120X

Ministère des Solidarités et de la Santé
Direction de la recherche, des études, de l'évaluation et des statistiques (DREES)
14 avenue Duquesne - 75 350 paris 07 SP
Retrouvez toutes nos publications sur drees.solidarites-sante.gouv.fr et nos données sur www.data.drees.sante.fr