Haut Conseil de la Santé Publique

Pour une **meilleure utilisation** des **bases de données nationales** pour la **santé publique** et la **recherche**

Collection
Documents

Mars 2012

Haut Conseil de la Santé Publique

POUR UNE MEILLEURE UTILISATION DES BASES DE DONNEES ADMINISTRATIVES ET MEDICO-ADMINISTRATIVES NATIONALES POUR LA SANTE PUBLIQUE ET LA RECHERCHE

Mars 2012

SOMMAIRE

La saisine

Liberté • Égalité • Fraternité
RÉPUBLIQUE FRANÇAISE

MINISTERE DU TRAVAIL, DE L'EMPLOI ET DE LA SANTE

SECRETARIAT D'ETAT A LA SANTE

Le Directeur général de la santé

Dossier suivi par Alain Fontaine
Mission stratégie et recherche
alain.fontaine@sante.gouv.fr

Paris, le 08 DEC. 2011

Monsieur le Président,

Le rapport publié en novembre 2009 par le Haut Conseil de la santé publique sur les systèmes d'information pour la santé publique attirait l'attention sur les difficultés rencontrées pour tirer pleinement parti des possibilités d'appariement des grandes bases de données nationales, notamment des bases de l'assurance maladie et de celles de la caisse nationale d'assurance vieillesse.

Ces difficultés limitent notamment les possibilités de tirer parti des données ainsi recueillies, pour la mise en œuvre notamment d'études de pharmaco épidémiologie, de grandes études de cohortes, d'évaluations économiques dans le domaine de la santé ou pour les systèmes de surveillance multi-sources.

Ces difficultés sont pour une part liées aux contraintes règlementaires qui imposent actuellement un décret en Conseil d'Etat pour l'utilisation du Numéro d'Inscription au Répertoire national d'identification des personnes physiques (NIR) à des fins de recherche et d'études en santé, sauf dans le cas où la recherche a son origine dans des données recueillies par des organismes déjà autorisés à collecter le NIR.

Sensibilisée à ces difficultés, la Commission Nationale de l'informatique et des libertés (CNIL) a approuvé en séance plénière le 22 juillet 2010 le principe de l'élaboration d'un décret cadre qui autoriserait les chercheurs et autorités sanitaires à utiliser le NIR dans des conditions à définir avec la CNIL et l'ensemble des acteurs concernés.

L'assemblée plénière de la Conférence Nationale de Santé a également adopté le 19 octobre 2010 un avis favorable à la résolution de ces difficultés dans le but d'optimiser l'utilisation des données à des fins de recherche permettant d'éclairer les politiques de santé publique.

La secrétaire générale des ministères sociaux a mandaté la Délégation à la stratégie des systèmes d'information en santé pour identifier une solution juridique et règlementaire appropriée, dans le contexte de la mise en œuvre de l'identifiant national de santé (INS) distinct du NIR. Elle a également mandaté récemment l'Agence des systèmes d'information partagée de santé (ASIP Santé) pour réaliser un état des lieux des différentes questions relatives à la sécurité des données de santé.

./...

Monsieur le Professeur Roger Salamon
Président
Haut Conseil de la santé publique

14, avenue Duquesne – 75 350 Paris 07 SP
Tél. : 01 40 56 60 00 – Télécopie : 01 40 56 40 56 – www.sante.gouv.fr – www.sante.fr

En complément de ces démarches, je vous serais reconnaissant de charger le groupe de travail du Haut Conseil sur les systèmes d'information pour la santé publique de nous apporter une information complète et cohérente sur les besoins identifiés par les principaux acteurs de la recherche et de la surveillance en santé vis-à-vis de l'appariement de données enregistrées dans les principales bases de données nationales.

Cette analyse devrait notamment permettre d'éclairer les points suivants :

- décrire la nature et l'intérêt des différents types d'informations que peut apporter l'appariement de données issues des principales grandes bases de données nationales dans le cadre de systèmes de surveillance, d'études et de travaux de recherche en santé, en identifiant les types de travaux de recherche ou de surveillance et et les bases de données concernées dans chaque cas ;
- rappeler les obstacles réglementaires limitant actuellement les possibilités de réaliser ces appariements ;
- identifier les difficultés techniques à résoudre pour assurer d'une part la faisabilité de ces appariements, d'autre part la protection des personnes et la sécurité des données ;
- décrire le cas échéant les solutions déjà mises en œuvre ou envisagées pour répondre à ces difficultés techniques.

Elle devrait nous être communiquée avant la fin du mois de janvier 2012.

Croyez, Monsieur le Président, à l'assurance de ma considération distinguée.

Le Directeur Général de la Santé,

Dr Jean-Yves GRALL

2.

Composition du groupe de travail

Président

Marcel Goldberg

Membres du HCSP

Claudine Berr, CS Maladies chroniques

Chantal Cases, CS Evaluation, stratégie et prospective

François Dabis, CS Maladies transmissibles

Jean-Pierre Hugot, CS Risques liés à l'environnement

Eric Jougla, CS Evaluation, stratégie et prospective

Catherine Sermet, CS Maladies chroniques

Représentants des membres de droit du HCSP

Sandrine Danet, Drees

Isabelle Grémy, InVS

Yves Charpak, EFS

Experts extérieurs

Jean-Claude Desenclos, InVS - Suppléante : Anne Doussin

Vincent Poubelle, Cnav

Philippe Ricordeau, Cnam-TS

Alain Trugeon, Fnors

Alain Weill, Cnam-TS

Secrétariat général du HCSP

Gérard Badéyan, coordonnateur

Résumé et synthèse des principales propositions

Le contexte

La France dispose de bases de données médico-sociales et économiques nationales centralisées, constituées et gérées par des organismes publics, couvrant de façon exhaustive et permanente l'ensemble de la population dans divers domaines stratégiques pour la santé publique et la recherche : recours aux soins, hospitalisation, handicaps, prestations et situation professionnelle, sociale et économique. De plus, un identifiant individuel unique (le NIR : numéro d'identification au répertoire) est actuellement utilisé par pratiquement toutes les bases de données nationales. Malgré certaines limites en termes de couverture, de qualité et de validité des données, ces bases de données, concernant plus de 60 millions de personnes, constituent un patrimoine considérable, vraisemblablement sans équivalent au monde.

Cependant, l'utilisation à des fins de recherche et de surveillance de ces bases de données nationales se heurte actuellement à des obstacles divers, dont les plus importants sont de nature juridique et opérationnelle.

Les bases de données et leur utilisation pour la recherche et la surveillance

Les principales bases de données nationales mobilisables pour la recherche pour la surveillance et la santé publique sont brièvement décrites dans ce rapport. Elles concernent :

- **Des données de santé** : mortalité, données d'hospitalisation *via* le PMSI (Programme de médicalisation du système d'information), données de consommations de soins et de prise en charge de l'assurance maladie, ces deux dernières bases de données étant réunies au sein du Système national d'information inter-régimes de l'assurance maladie - SNIIR-AM, dont un échantillon aléatoire (l'EGB **:** échantillon généraliste des bénéficiaires) peut être utilisé à distance. Les principales limites des bases de données du PMSI et de l'assurance maladie sont qu'elles ne contiennent pratiquement pas de données concernant la situation socioprofessionnelle des personnes et que les informations sur le domicile des patients ne sont pas suffisamment précises pour permettre une exploitation territorialisée à une échelle fine ; de plus, la validité des données de santé de ces bases est de qualité variable.
- **Situation socioprofessionnelle** : les bases de données de la Caisse nationale d'assurance vieillesse (Cnav) permettent de retracer, pour chaque personne ayant appartenu durant sa vie au moins une fois au régime général de sécurité sociale, ses différentes périodes d'activité : périodes d'activité professionnelle ou assimilées (chômage, maladie, maternité ou congés parentaux, …). De plus, la Cnav gère le système national de gestion des identités qui contient l'ensemble des données (NIR, état-civil, statut vital) de la population française, ainsi que le RNIAM (Répertoire national inter-régimes des bénéficiaires de l'assurance maladie), qui contient les informations de rattachement des personnes aux organismes servant les prestations d'assurance maladie.
- **Autres sources** : outre les trois dispositifs cités, il existe d'autres bases de données nationales concernant des problèmes de santé ou des populations spécifiques.

Des exemples d'utilisation des bases nationales, récents ou en cours, illustrent la diversité de leurs apports, qu'il s'agisse de l'utilisation de chaque base de données indépendamment des autres, de l'appariement de bases différentes ou de l'enrichissement par des sources

administratives d'enquêtes avec recueil de données auprès des personnes, comme les études de cohorte.

Les principales difficultés pour l'utilisation des bases de données nationales à des fins de recherche et de surveillance

- Le **cadre juridique** permettant d'accéder aux données à caractère personnel des bases nationales est complexe et dépend notamment de la nature de la base de données concernée et de celle de l'organisme utilisateur : organismes de recherche et agences de santé, service statistique public, organismes privés à but lucratif. Dans l'ensemble, on peut considérer que les textes actuels ne constituent pas un obstacle insurmontable à l'accès aux données à caractère personnel des bases nationales, mais ils rendent complexes les circuits de transmission de données. Une difficulté majeure concerne l'identifiant pour l'accès aux données à caractère personnel des bases nationales. L'identifiant individuel actuellement utilisé dans les bases nationales est le NIR, pour la collecte duquel un décret en conseil d'État est nécessaire, ce qui constitue un obstacle infranchissable pour la plupart des équipes concernées, dès lors qu'elles n'appartiennent pas à un organisme habilité à disposer de cet identifiant ou qu'il n'est pas possible d'établir des flux de données reposant sur un tel organisme. Un autre problème risque de se poser prochainement avec le déploiement de l'INS (Identifiant national de santé), destiné à être le seul identifiant utilisé pour les applications en santé, ce qui rendrait impossible l'appariement de données de santé avec d'autres sources utilisant le NIR comme identifiant. Enfin, dans nombre de situations, il n'est en pratique pas possible de recueillir un consentement explicite des personnes concernées avant chaque étude, ce qui présente une difficulté vis-à-vis de la loi informatique et libertés.
- **Obstacles organisationnels et techniques** : l'accessibilité effective des données reste aujourd'hui un problème majeur, qu'il s'agisse d'identifier les personnes pour lesquelles on veut extraire des données dans les bases, d'extraire les données et de mettre les données en forme pour les analyses. L'utilisation du NIR pour l'identification des personnes nécessite une participation active d'un organisme détenteur du NIR, mais il n'existe pas de dispositif formalisé mis en place par un ou plusieurs organismes détenteurs du NIR pour prendre en charge ce type d'activité ; l'extraction proprement dite des données des bases nécessite également une participation active des organismes gestionnaires des bases de données, car cette activité implique diverses étapes techniques lourdes et, malgré les très importants efforts de la Cnam-TS, de l'Insee et de la plupart des autres organismes gestionnaires de bases de données, l'accessibilité effective des données reste cependant encore fortement contingentée ; la complexité de la base de données du SNIIR-AM rend l'utilisation des données individuelles particulièrement difficile et nécessite des moyens spécialisés importants dont peu d'équipes disposent.

Finalement, l'ensemble des aspects juridiques, organisationnels et techniques à prendre en compte pour pouvoir utiliser les bases de données nationales de façon respectueuse des textes et efficace en termes de résultats constitue un **véritable maquis juridico-institutionnel** dans lequel les équipes de recherche et de surveillance ont souvent du mal à se repérer.

Principales propositions

- **Règles d'ouverture des bases de données nationales** : les projets pour lesquels des données sont demandées doivent avoir fait l'objet d'évaluations scientifiques préalables par des organismes légitimes ; la finalité d'intérêt public de la demande doit être établie ; l'utilisation est réservée à des organismes publics ou parapublics, ou à des structures

privées à but lucratif (directement ou par un intermédiaire académique) dans le cadre de la réalisation de certaines études (post-autorisation de mise sur le marché) commanditées par les pouvoirs publics ; un contrôle préalable par les organismes gestionnaires des bases de données doit être possible, le cas échéant suivi d'un avis négatif argumenté, de même que la mise en œuvre d'analyses à titre de contre-expertise.

- **Gouvernance** : il s'agit de l'organisation de l'accès aux bases, de la supervision de la réalisation des demandes et du contrôle de l'utilisation des données. Actuellement, chaque organisme gestionnaire de base « traite » directement avec les demandeurs ; cet état de choses n'est pas satisfaisant. Deux modèles de gouvernance ont été envisagés : décentralisée (chaque organisme public gestionnaire de bases de données fixe des règles explicites d'accès et met en place un « guichet » destiné à traiter les demandes) ; centralisée (une structure centrale gère un guichet unique et fait office d'interface entre les demandeurs et les organismes gestionnaires de bases de données). Les avantages et inconvénients de chaque modèle ont été analysés, notamment en termes de simplicité, d'autonomie et de moyens.
- **L'identifiant pour l'accès aux bases de données lorsqu'on ne dispose pas du NIR** : deux possibilités peuvent être envisagées : (*i*) **recueil du NIR en clair** auprès des personnes ou des professionnels et transfert à un « centre d'appariement sécurisé » (CAS), structure jouant un rôle de tiers de confiance habilité à recevoir les NIR des sujets concernés, qui applique ensuite des procédures d'anonymisation ; dans l'état actuel des textes, cette solution ne permet pas d'éviter la prise d'un décret en conseil d'État spécifique de chaque étude ; (*ii*) **recueil par appariement indirect (probabiliste) et consultation d'une base contenant les NIR de toute la population** (RNIPP, RNIAM ou SNGI) ; cette méthode est employée pour l'accès au statut vital et implique que le demandeur recueille uniquement les informations suivantes : nom, prénom, date et lieu de naissance des personnes concernées et les transmettent (accompagnées d'un numéro d'étude destiné aux transferts des données) *via* un tiers de confiance à l'organisme détenteur de la base des NIR, qui peut ainsi retrouver les NIR par une méthode d'appariement indirect, et les transférer à qui de droit, sans que l'investigateur n'en ait connaissance. Sa mise en œuvre pratique (dont les aspects opérationnels sont décrits dans le rapport) implique la création d'un centre d'appariement sécurisé (CAS), tiers de confiance qui serait l'opérateur central des procédures à mettre en œuvre. Cette procédure a déjà été mise en œuvre, elle est éprouvée et fonctionne avec des performances satisfaisantes lorsque les informations initiales (nom, prénom, date et lieu de naissance) sont de bonne qualité ; elle implique cependant qu'au moins un organisme détenteur des NIR (Insee ou Cnav) mette en place un « service » *ad hoc*. Séduisante, la méthode d'appariement indirect n'est pas adaptée à toutes les situations : par exemple les études avec des données en provenance des professionnels de santé ne pourraient pas obtenir le lieu de naissance, ou encore les études en situation d'urgence impliquent une rapidité incompatible avec les délais nécessaires. C'est pour répondre à ces situations que **la Cnil propose l'élaboration d'un décret-cadre en conseil d'État** « *permettant l'utilisation encadrée du NIR à des fins de recherche médicale et d'études en santé publique* », solution qui s'impose dans le cadre juridique actuel, dans la mesure où il n'existe pas de procédures sans utilisation du NIR permettant de répondre à toutes les situations.

  Au total, les deux approches (appariement indirect et recueil en clair du NIR autorisé par un décret-cadre) sont complémentaires : la méthode d'appariement indirect doit être privilégiée pour toutes les situations où elle est possible et suffisamment efficace et un principe de parcimonie doit être appliqué pour le recueil en clair du NIR ; en pratique, c'est à la Cnil que revient d'examiner dans chaque cas les solutions possibles.

- **Correspondance INS – NIR** : le déploiement de l'INS rendra indispensables l'établissement et la maintenance d'une table de correspondance entre INS et NIR, seule à permettre les appariements entre bases de données de santé, identifiées à l'avenir avec l'INS, et les autres bases identifiées avec le NIR ; il semble prévu que la Cnav gèrera cette table de correspondance, ce qui reste à confirmer.

Pour la création d'une plateforme d'interface entre les utilisateurs et les bases de données nationales

Une solution proposée pour améliorer la situation actuelle est la création d'une plateforme spécialisée qui jouerait le rôle d'interface entre les chercheurs et les bases de données nationales. Cette plateforme aurait comme missions principales : le conseil aux utilisateurs, la préparation des requêtes, la transmission des requêtes vers les bases de données et la récupération des fichiers extraits, la restitution aux utilisateurs de données synthétisées, après sélection préalable des variables d'intérêt à partir des données brutes.

Deux modèles organisationnels peuvent être envisagés, qui ont chacun leurs avantages et inconvénients : chaque organisme gestionnaire de base développe un « guichet » ouvert aux utilisateurs remplissant ces fonctions ; création d'une plateforme centrale.

Propositions diverses

- **Le consentement des personnes** : la recommandation de la Conférence nationale de santé « *d'unifier le régime de consentement à la collecte, au traitement, à l'échange et à l'hébergement des données de façon à ce qu'il soit aisément compréhensible par les usagers et commode à exprimer »* doit être soutenue et mise en œuvre.
- **Identification indirecte dans le SNIIR-AM** : bien que le SNIIR-AM soit une base anonymisée, son développement, avec des données de plus en plus nombreuses et diversifiées, des extractions répétées sur les mêmes personnes, peut rendre possible l'identification indirecte des personnes par croisement de données ; il est indispensable de développer des mesures de précaution renforcées, faisant appel à des techniques diverses.
- **La propriété des fichiers appariés** : les fichiers constitués à partir de l'appariement de données provenant de bases gérées par des organismes différents posent des problèmes de propriété. Chaque organisme contributeur de données doit avoir le moyen de s'opposer à des utilisations des données qu'il fournit.
- **La politique tarifaire** : si l'accès aux données fait l'objet d'un paiement, il est indispensable que les tarifs pratiqués pour les demandeurs relevant d'un organisme public de recherche ou de surveillance ne soient pas incompatibles avec les budgets que les équipes publiques demandeuses sont susceptibles d'obtenir pour leurs travaux ; par contre, pour des demandes provenant de structures à but lucratif, il est légitime que les tarifs pratiqués soient établis de façon à correspondre au moins au coût véritable des données.
- **La localisation spatiale des personnes** : on recommande que les organismes collecteurs de données de premier niveau (l'hôpital, la CPAM, etc.) mettent en place une procédure interne automatisée par laquelle l'adresse des personnes serait géocodée, convertie en code Iris et transférée sous cette forme dans les bases nationales.

1 NATURE ET INTERET DES DIFFERENTS TYPES D'INFORMATIONS DANS LE CADRE DE SYSTEMES DE SURVEILLANCE, D'ETUDES ET DE TRAVAUX DE RECHERCHE EN SANTE

1.1 LES BASES DE DONNEES PUBLIQUES ADMINISTRATIVES ET MEDICO-ADMINISTRATIVES NATIONALES : UNE RICHESSE INSUFFISAMMENT EXPLOITEE

La France est un des rares pays qui dispose de bases de données médico-sociales et économiques nationales centralisées, constituées et gérées par des organismes publics, couvrant de façon exhaustive et permanente l'ensemble de la population dans divers domaines stratégiques : recours aux soins, hospitalisation, handicaps, prestations et situation professionnelle et sociale. De plus, un identifiant individuel unique (le numéro d'identification au répertoire ou NIR) est actuellement utilisé (directement ou sous forme anonymisée) par pratiquement toutes les bases de données nationales, dont la constitution repose sur des activités liées aux missions de l'administration et d'organismes publics. Leur collecte est très encadrée, notamment par le code de la santé publique pour les bases médico-administratives. Un intérêt majeur est qu'elles sont le plus souvent exhaustives (et donc localisées) et produites régulièrement, le plus souvent sur une base annuelle.

Ces bases de données présentent évidemment des limites diverses en termes de couverture, de qualité et de validité des données, variables selon les types d'utilisation qu'on peut envisager. Ces bases de données, concernant plus de 60 millions de personnes, constituent néanmoins un patrimoine immatériel considérable, vraisemblablement sans équivalent au monde. D'autres pays ont su depuis longtemps mettre au service de la santé publique et de la recherche leurs systèmes d'information médico-sociaux, notamment les pays scandinaves ou le Canada, en créant de véritables « *Population Data Centers* », largement ouverts à la communauté scientifique qui permettent de très nombreuses études de grande qualité dans des domaines divers (voir par exemple le centre mis en place à la *British Columbia University*[1]). On peut, dans ce contexte, signaler l'élaboration en cours par l'OCDE d'un document sur l'utilisation secondaire des données de santé. Cette demande de l'OCDE fait partie des priorités de travail retenues par les ministres de la santé lors de leur dernière réunion d'octobre 2010. Il s'agit de faire un état des lieux sur les bases de données contenant des informations individuelles pour divers champs identifiés et pertinents pour l'OCDE (données d'hospitalisation, de soins de longue durée, de mortalité…), de faire le point sur les possibilités d'appariements de ces bases et les pratiques en cours dans les différents pays pour les aspects légaux concernant l'utilisation de ces données. Un document de synthèse devrait être produit par l'OCDE en mai 2012.

Dans notre pays, les bases de données administratives et médico-administratives nationales sont cependant encore insuffisamment exploitées en dehors des organismes qui les constituent et les gèrent, même si plus d'une centaine de publications référencées, se rapportant à des travaux réalisés sur les données de remboursement de l'assurance maladie, avaient été recensées en juin 2009[2]. Les équipes de recherche les

1 http://www.popdata.bc.ca/.

2 Martin-Latry K, Bégaud B. Pharmacoepidemiological research using French reimbursement databases: yes we can ! Pharmacoepidemiology and drug safety 2010; 19: 256–265.

utilisent encore trop peu souvent, comme l'illustre le fait que l'Institut des données de santé (IDS) avait enregistré fin 2010 au total moins de vingt demandes d'accès au SNIIR-AM provenant d'organismes publics et privés divers[3]. En revanche, les agences (InVS, HAS, Afssaps…) travaillent désormais régulièrement sur les données de l'assurance maladie, tant dans une approche de surveillance sanitaire que de suivi du médicament (en post inscription, notamment) ou médico-économique.

Les bases de données administratives et médico-administratives nationales ont un potentiel considérable. Elles répondent à des besoins d'information très diversifiés de surveillance, d'études et de recherches, dépassant largement les préoccupations de ces organismes, et peuvent rendre de grands services à la communauté de santé publique et de recherche.

L'utilisation à des fins de recherche et de surveillance de ces bases de données nationales se heurte actuellement à des obstacles divers, dont les plus importants sont de nature juridique et opérationnelle :

- L'identifiant utilisé de façon directe ou cryptée par les bases de données nationales étant le NIR, l'interdiction de fait de le recueillir auprès des personnes ou des organismes qui en disposent limite très fortement les possibilités d'accès aux bases de données, qui n'est possible que dans certaines circonstances et qui, de plus, nécessite une participation active des gestionnaires de ces bases.
- la mise à disposition des données à la communauté de santé publique et de recherche dans des conditions en permettant l'exploitation, nécessite des ressources scientifiques, techniques et organisationnelles complexes et de haut niveau de compétence. Ces ressources dépassent largement les moyens actuellement disponibles au sein des équipes françaises, quel que soit leur organisme d'appartenance.

1.2 LES PRINCIPALES BASES DE DONNEES NATIONALES POUR LA RECHERCHE ET LA SANTE PUBLIQUE

1.2.1 Données de santé

1.2.1.1 Mortalité

Le statut vital et les causes de décès peuvent actuellement être obtenus selon la procédure décrite dans le décret n° 98-37 autorisant l'accès au Répertoire national d'identification des personnes physiques (RNIPP) et à la base de données du Centre d'épidémiologie des causes de décès de l'Inserm (CépiDc). Dans le cadre de cette procédure, le plateau technique informatique du Centre de recherche en épidémiologie et santé des populations (CESP, ex-CRI IFR 69) joue un rôle central, du type « plateforme » pour cette prestation spécifique d'accès au statut vital et aux causes de décès.

3 Source: http://www.institut-des-donnees-de-sante.fr/.

1.2.1.2 Données d'hospitalisation : le PMSI

Le PMSI (Programme de médicalisation du système d'information) a pour objectif de produire des informations à contenu médical sur les fonctions hospitalières et de permettre une allocation de ressources dépendante de l'activité hospitalière. Il consiste en un recueil exhaustif systématique et un traitement automatisé d'informations administratives et médicales. Chaque séjour est ensuite classé dans l'un des 560 GHM (Groupes homogènes de malades), économiquement et médicalement considérés comme « homogènes ».

Au sein des établissements hospitaliers, les départements d'information médicale (DIM) jouent un rôle central. Le médecin responsable de l'information médicale a un rôle de conseil pour la production des informations et il veille à leur qualité. Les données recueillies sont soumises au secret médical et sont sous la responsabilité du médecin responsable du DIM. Les établissements transmettent trimestriellement les fichiers anonymisés à l'Agence régionale de santé (ARS), et celles-ci les transmettent à l'Agence technique de l'information sur l'hospitalisation (ATIH) en vue de la constitution des bases de données nationales.

Cette transmission se fait sous la forme de résumés de sortie anonymisés (RSA), qui contiennent les informations suivantes :

Identification du séjour : modes d'entrée et de sortie de l'établissement – nombre d'unités médicales fréquentées – mois et année de sortie - durée de séjour de la totalité de l'hospitalisation– numéro Finess de l'établissement.

Identification du patient : sexe - âge en année ou en jours pour les enfants de moins d'un an - numéro d'anonymat, construit par l'anonymisation irréversible du NIR, de la date de naissance et du sexe du patient.

Données médicales : poids de naissance - diagnostic principal et ensemble des diagnostics associés et des actes pratiqués. Les diagnostics sont codés selon la CIM 10. Depuis 2005, la Classification commune des actes médicaux (CCAM), qui harmonise la codification des actes entre médecine de ville et médecine hospitalière, est utilisée pour le PMSI.

1.2.1.3 Les données de l'assurance maladie

Échelon loco-régional du régime général de sécurité sociale

Parmi les données enregistrées par les systèmes d'information de l'assurance maladie du régime général de sécurité sociale (RGSS), on distingue les données dites « de production », portant principalement sur les consommations de soins, et dont l'objectif premier est la liquidation des prestations d'assurance maladie, et les données « de référentiels », qui concernent les informations sur les assurés, les établissements de santé et les professionnels de santé. Par ailleurs, les services médicaux des caisses primaires d'assurance maladie (CPAM) disposent de leurs propres fichiers comportant des informations médicales sur les affections de longue durée (ALD), les accidents du travail (AT) et les maladies professionnelles (MP), et dont l'objectif initial est le contrôle, par les médecins conseil, des pathologies ouvrant droit à une prestation.

Toutes ces données sont rassemblées au niveau des Centres de traitement informatique régionaux (CTI), qui jouent ainsi un rôle central dans la gestion des données de l'assurance maladie. Il existe plusieurs CTI en France, chacun rassemblant les données d'un groupe de CPAM. Une validation des données est faite à ce niveau. Ces données sont regroupées au sein de deux bases principales : ERASME (données de production essentiellement) et HIPPOCRATE (données médicales).

- **La base ERASME** (Extraction recherches analyses pour un suivi médico-économique) enregistre les consommations de soins et consommables pharmaceutiques de façon précise (médicaments, actes de biologie) des personnes affiliées au régime général et aux sections locales mutualistes (SLM), incluant l'identification des professionnels de santé (prescripteurs et exécutants) et des établissements sanitaires et sociaux prestataires de soins. ERASME est gérée au niveau des CTI. Il ne s'agit pas d'une base anonyme, les bénéficiaires étant identifiés par le NIR de l'assuré, leur date et leur rang de naissance ; elle contient par ailleurs les nom, prénom, date de naissance, sexe, adresse et qualité des bénéficiaires (assuré, ayant droit conjoint ou enfant).
- **La base HIPPOCRATE** constitue le système d'information du service médical de l'assurance maladie ; elle est hébergée et administrée par les CTI. Elle enregistre les données médicales (diagnostics codés en CIM-10) des patients en ALD, AT et MP. Les ALD, qui concernent les affections susceptibles d'ouvrir droit à une exonération du ticket modérateur, sont d'un intérêt particulier pour l'épidémiologie. Il s'agit des affections de la liste ALD 30 (affections comportant un traitement prolongé et une thérapeutique particulièrement coûteuse, inscrites sur une liste établie par décret) ; des affections dites « hors liste » (maladies graves de forme évolutive ou invalidante, non inscrites sur la liste des ALD 30, comportant un traitement prolongé d'une durée prévisible supérieure à six mois et une thérapeutique particulièrement coûteuse) ; des polypathologies (patient atteint de plusieurs affections caractérisées entraînant un état pathologique invalidant et nécessitant des soins continus d'une durée prévisible supérieure à six mois). Les médecins de l'échelon local ont accès à l'identité des patients (nom, prénom, adresse) ; les médecins de l'échelon régional n'ont qu'une version de la base où les patients sont identifiés par un numéro d'anonymat.

Les données de l'assurance maladie – Autres régimes

Les autres régimes d'assurance maladie ont des bases de données qui contiennent pour l'essentiel des données de même nature que le RGSS.

Le Système national d'information inter-régimes de l'assurance maladie

L'ensemble des bases de données concernant les événements de santé est réuni au sein du *Système national d'information inter-régimes de l'assurance maladie* (SNIIR-AM). Depuis sa création et la constitution en 2003 d'un entrepôt, limité aux remboursements du régime général, le SNIIR-AM s'est considérablement enrichi et l'accès à ses données a été facilité avec la mise en place, en 2005, d'un échantillon au 1/100 dédié aux institutions publiques, aux agences et au monde de la recherche

(EGB : *cf*. ci-dessous). L'enrichissement du SNIIR-AM s'est traduit par la mise à disposition des utilisateurs :

- du chaînage en routine des données ambulatoires et du PMSI (en 2007),
- de l'exhaustivité de l'activité médicale (actes techniques) codée (en 2007),
- de l'activité externe des hôpitaux publics (en 2009),
- des dates de décès provenant de l'Insee (en 2009),
- des médicaments et dispositifs médicaux facturés en sus des forfaits d'hospitalisation (en 2009),
- des données provenant de l'ensemble des régimes d'assurance maladie (en 2009),
- des dates exactes d'hospitalisation (en 2010).

Les données du SNIIR-AM incluent donc depuis 2009 tous les régimes de l'assurance maladie : Cnam-TS, MSA, RSI et les seize autres régimes spéciaux (y compris les sections locales mutualistes-SLM) et concernent aussi bien la médecine de ville que les hospitalisations. Les objectifs du SNIIR-AM sont la connaissance des dépenses de l'ensemble des régimes de l'assurance maladie, le retour de ces informations auprès des professionnels de santé (informations pertinentes relatives à leur activité, leurs recettes, et s'il y a lieu, à leurs prescriptions), la définition, le suivi et l'évaluation des politiques de santé publique (loi de santé publique du 13 août 2004).

Le SNIIR-AM constitue une solution particulièrement efficace pour l'accès à des données individuelles, en évitant le passage par les échelons locaux et régionaux des différents régimes qui rendent complexes et lourdes les procédures d'accès. La base SNIIR-AM est en effet alimentée par les fichiers des bases de données citées ci-dessus ; elle est gérée par le Centre national de traitement informatique (Centi) de la Cnam-TS.

Le SNIIR-AM est une base de **données individuelles mais anonymes** qui rassemble les données décrites plus haut : les données de remboursement avec le détail du codage des actes et des médicaments ; les identifiants des professionnels de santé et des établissements de santé qui ont participé aux soins du patient ; les informations sur la pathologie traitée pour les patients en ALD et en AT-MP ; les données issues du PMSI. Pour les prestations concernées, les codes affinés sont disponibles (médicament, LPP[4], CCAM[5], GHS[6], biologie). L'anonymisation des variables identifiantes est réalisée par le module FOIN (Fonction d'occultation *des informations nominatives*). Cette fonction repose sur le NIR de l'ouvrant droit, la date de naissance et le sexe du bénéficiaire ; pour les ayants droit (enfants, conjoints ne travaillant pas…), la Cnam-TS gère pour ses affiliés un fichier intitulé Référentiel individus (RFI) qui permet de retrouver le NIR de l'ouvrant droit et ainsi de calculer leurs propres identifiants FOIN en introduisant la date de naissance et le sexe du bénéficiaire[7]. Les données sont

4 Liste des produits et prestations.
5 Classification commune des actes médicaux.
6 Groupe homogène de séjours.
7 Des systèmes équivalents existent pour les autres régimes obligatoires et les SLM (Sections locales mutualistes).

anonymisées en deux étapes : au niveau locorégional (FOIN-1), et au niveau national (FOIN-2). L'application des algorithmes FOIN construit un identifiant anonyme non réversible : à partir d'un identifiant, on ne peut pas retrouver les données nominatives qui ont servi à son calcul.

L'architecture actuelle du SNIIR-AM est composée de plusieurs sous-ensembles :

- **L'entrepôt SNIIR-AM,** contenant toutes les données de remboursement issues des différents régimes d'assurance maladie obligatoire, sert à construire les autres sous-ensembles. Les données de cet entrepôt ne sont pas accessibles. La durée de conservation de ces données initialement de deux ans plus l'année en cours, a été portée récemment à trois ans plus l'année en cours.
- **La base DCIR** (données de consommation inter-régimes) est une extraction de l'entrepôt SNIIR-AM des données du régime général, de la MSA, du RSI et de quelques régimes spéciaux. L'historique des données est le même que l'entrepôt SNIIR-AM (deux ans plus l'année en cours), mais tous les organismes qui alimentent l'entrepôt SNIIR-AM n'utilisant pas le même niveau de norme, le DCIR ne couvre pas le même périmètre que l'entrepôt SNIIR-AM. Les données du DCIR sont détaillées par bénéficiaire (identifiants anonymes) et par offreur de soins (prescripteur et exécutant, identifiant en clair pour l'assurance maladie obligatoire). Par construction, le DCIR concerne les bénéficiaires ayant « consommé » des soins au cours de la période couverte.
- **L'EGB** (échantillon généraliste des bénéficiaires) permet de suivre l'évolution de la consommation de soins de 600 000 personnes (identifiants anonymes) sur vingt ans sélectionnés à partir d'une clef de tirage. Il est à ce jour alimenté par les données du régime général (hors SLM), de la MSA et du RSI. Il devrait prochainement contenir des données des autres régimes d'assurance maladie et des SLM. Il contient les mêmes données que celles du DCIR, mais il s'agit à la différence de ce dernier d'un échantillon de bénéficiaires.
- **Les Datamarts** (magasins de données) contiennent, à la différence du DCIR et de l'EGB, des données agrégées répondant à des « besoins métiers », par exemple : cliniques privées, offre de soins….
- **Les données du PMSI** (détaillées, exhaustives et anonymisées) sont accessibles aux utilisateurs ayant accès à la base DCIR. L'identifiant anonyme (NIR anonymisé) des patients est commun à la base DCIR et au PMSI, ce qui autorise le chaînage par les utilisateurs au cas par cas des données issues de ces deux bases. Le chaînage des données issues de l'EGB et du PMSI pourra être réalisé et les utilisateurs accéderont donc à des données déjà chaînées.

Le SNIIR-AM, qui est un dispositif relativement récent (constitué en 2003), est complété et amélioré régulièrement : le chaînage ville-hôpital a été réalisé en 2006/2007, les dates de décès ont été intégrées en 2009, les dates précises d'hospitalisation en 2010. Concernant l'enrichissement par les causes de décès, une étude expérimentale nationale (Projet Amphi : analyse de la mortalité post-hospitalisation) est en cours (*cf*. plus loin).

1.2.1.4 Intérêt et limites des bases de données de l'assurance maladie

Bien qu'elles n'informent évidemment pas sur de nombreuses données personnelles et environnementales pouvant être indispensables pour la recherche et la surveillance (comportements, expositions à des facteurs de risque de nature diverse, etc.), les bases de données du PMSI et de l'assurance maladie ont à l'évidence un intérêt majeur. Elles présentent cependant certaines limites :

- Elles ne contiennent aucune donnée concernant la situation socioprofessionnelle des personnes (en dehors de la notion de CMUC), seuls le sexe, l'âge et l'organisme de sécurité sociale de rattachement étant enregistrés.
- Les informations sur le domicile des patients ne sont pas suffisamment précises pour permettre une exploitation territorialisée à une échelle fine. Le code postal parfois utilisé pose problème, le code commune étant toujours préférable. De plus, disposer d'une géolocalisation plus fine permettant de produire l'Iris ou des données « carroyées » (c'est-à-dire selon un quadrillage fin, par exemple de 200 mètres de côté), serait particulièrement utiles pour les études sur les inégalités territoriales de santé et sur la santé environnementale.
- Elles ne contiennent pas de résultat d'examen clinique ou paraclinique.
- Elles ne contiennent pas d'information sur l'hébergement en structure médico-sociale des personnes âgées ou sur les hospitalisations en long séjour (situation en voie d'amélioration).
- Elles ne contiennent pas d'information médicale sur les séjours en centre hospitalier spécialisé (psychiatrie).
- Le PMSI ne contient pas d'information sur les passages aux urgences ; cependant, la DGOS envisage de généraliser la remontée d'informations sur les passages aux urgences (Résumés de passage aux urgences-RPU) déjà mise en place pour environ 350 établissements et de les intégrer au PMSI en utilisant le même numéro d'anonymat individuel.
- La validité des données de santé de ces bases est de qualité variable. Bien que l'ensemble des bases de données citées n'ait pas fait l'objet d'analyses systématiques de validité, quelques études plus ou moins ponctuelles ont porté sur les données issues des différents fichiers. L'utilisation du PMSI comme source d'information sur les pathologies s'avère délicate et les diagnostics ne sont pas toujours fiables[8,9]. Les ALD ont des limites connues[10] et la qualité du codage des pathologies est encore mal connue. Des travaux récents concernant les cancers menés avec la collaboration des registres du cancer montrent que, utilisés isolément, ni le PMSI ni les ALD ne permettent d'avoir une bonne

8 Couris CM, Forêt Dodelin C, Rabilloud M, Colin C, Bobin JY, Dargent D, Raudran D, Schott AM. Sensibilité et spécificité de deux méthodes d'identification des cancers du sein incidents dans les services spécialisés à partir des données médico-administratives. Rev Epidemiol Sante Publique 2004, 52, 151-60.
9 Couris CM et al. Method of correction to assess the number of hospitalized incident breast cancer cases based on claims databases. Journal of Clinical Epidemiology, 2002, 55 : 386-391.
10 Incidence médico-sociale des ALD30 en 1999. CNAMTS-DSM-Mission des Soins de ville-Mission Statistique. Avril 2004. Disponible sur le site www.ameli.fr/245/doc/1391/article_pdf.html.

estimation de l'incidence et de la prévalence, mais que leur utilisation conjointe peut être efficace[11]. La base de données de remboursements de l'assurance maladie est adaptée aux objectifs d'analyse des pratiques de prescription[12], d'évaluation de l'impact de campagne d'information[13]. Par contre, elle ne comporte pas d'information directe sur la nature des maladies traitées et exclut par définition l'automédication et les prestations non présentées au remboursement.

L'intérêt potentiel des bases de données du SNIIR-AM apparaît clairement dans la mesure où elles fournissent des données individuelles médicalisées, structurées et codées de manière standardisée[14]. Leur utilisation, notamment dans une optique épidémiologique, nécessite cependant un important travail méthodologique, de contrôle et de validation.

1.2.2 Situation socioprofessionnelle

1.2.2.1 Les bases de données

Les bases de données de la Caisse nationale d'assurance vieillesse (Cnav) sont un élément essentiel, à la fois pour l'accès aux données socioprofessionnelles et pour le traçage des sujets. Le rôle de cet organisme est notamment d'assurer le droit au paiement de la retraite pour toute personne ayant appartenu durant sa vie au moins une fois au régime général de sécurité sociale. Pour cela, la Cnav a mis un place un système permettant de collecter et traiter les données sociales issues de différents organismes tels que les régimes de retraite, Pôle Emploi… (aux niveaux national, régional et local). La Cnav exerce la mission de collecte, de contrôle et de traitement de données sociales utiles pour les droits à la retraite pour certains de ces partenaires, chacun d'entre eux étant ensuite rendu destinataire des informations qui le concernent.

Pour remplir son rôle, la Cnav a mis en place et gère plusieurs bases de données, qu'on présente succinctement, ainsi que l'origine des données qui les alimentent.

- **Le SNGI (Système national de gestion des identités)** qui contient l'ensemble des données (NIR, état-civil, statut vital) pour toute personne née en France métropolitaine ou dans les DOM, ainsi que les données d'identification des personnes nées à l'étranger ou dans les TOM et résidant ou ayant résidé sur le territoire français ; il a pour finalité de certifier l'identité d'une personne. L'Insee a en charge l'immatriculation de toute personne née en France métropolitaine ou dans les DOM ; ces informations sont contenues dans le RNIPP. Il incombe à la Cnav depuis 1988, dans le cadre de sa mission déléguée par l'Insee, de procéder à l'immatriculation des ayants droit nés à l'étranger ou dans les TOM et résidant sur le

11 Grosclaude et al et Lauzeille et al, BEH 2012 (à paraître).
12 Deprez Ph-H, Chinaud F, Clech S, Germanaud J, Weill A, Cornille JL, Fender P. La population traitée par médicaments de la classe des antihistaminiques en France métropolitaine : données du régime général de l'assurance maladie, 2000. Revue médicale de l'assurance maladie Avril-juin 2004, 35 (1), 3-11.
13 Lecadet J, Vialaret K, Vidal P, Baris B, Fender P. Mesure à l'échelle d'une région des effets d'un programme national d'information sur le bon usage des antibiotiques. Revue médicale de l'assurance maladie Avril-Juin 2004, 35 (2) ,81-91.
14 Fender P, Weill A. Epidémiologie, santé publique et bases de données médico-tarifaire. (Éditorial) Rev Epidemiol Santé Publique, 2004, 52,113-117.

territoire français. Le SNGI contient l'ensemble de des éléments d'identification des personnes (NIR, nom de famille[15], nom d'usage, nom marital, prénom(s), sexe, date et lieu de naissance, date et lieu de décès, éventuellement numéros d'acte de naissance et d'actes de décès et, pour les personnes nées à l'étranger ou dans les COM, les éléments de filiation), soit reçus de l'Insee, soit intégrés par la Cnav elle-même. L'Insee et la Cnav se transmettent mutuellement les notifications (immatriculations et mises à jour) apportées à leur champ respectif, ceci afin que le SNGI et le RNIPP soient synchrones.[16]

- **Le SNGC (Système national de gestion des carrières)** qui permet de retracer pour chaque individu dès la première validation d'un droit (premier salaire, etc.) et jusqu'à la liquidation de ses droits à la retraite, ses différentes périodes d'activité : périodes d'activité professionnelle ou assimilées (chômage, maladie, maternité ou congés parentaux, …). Le SNGC contient donc l'ensemble des données inhérentes à la carrière des assurés du régime Général, y compris les données concernant d'éventuelles périodes effectuées dans les autres régimes de base (MSA, Cancava, Organic), ainsi que dans les régimes particuliers ou spéciaux (SNCF, EDF-GDF, CNRACL, Mines, etc.).
- **Le SNGD** (**Système national de gestion des dossiers)** de retraites en cours d'instruction ou de paiement).
- **Le RNIAM (Répertoire national inter-régimes des bénéficiaires de l'assurance maladie)** est également mis en œuvre par la Cnav pour le compte et sous le contrôle des organismes d'assurance maladie ; il est constitué pour chaque bénéficiaire, en plus de son NIR et de son état-civil, des informations de rattachement à l'organisme lui servant les prestations d'assurance maladie.

Pour la constitution et l'enrichissement de ces bases de données, la Cnav reçoit régulièrement des données en provenance de différentes sources. Les *Déclarations annuelles des données sociales* (DADS), sont transmises chaque année par les employeurs ayant un numéro Siret. Les *Données nominatives trimestrielles* (DNT) sont transmises par les employeurs de personnel de maison. Les informations de périodes d'activité/non-activité des individus relevant de l'Unedic (chômage), de la Cnam-TS (maladie), de la Cnaf (maternité, …), des régimes particuliers ou spéciaux (SNCF, EDF, RATP,…) sont également transmises à la Cnav. Il en est de même pour certains autres régimes, et il est prévu qu'à terme la Cnav reçoive les données de tous les régimes. Cependant, les données des autres régimes ne sont actuellement pas toutes connues « en temps réel » : certaines le sont au moment de la régularisation de la carrière ; pour d'autres régimes (RSI, par exemple), l'alimentation est annuelle pour tous les actifs quel que soit l'âge.

Après avoir reçu des différents organismes gestionnaires des prestations sociales les informations relatives à l'activité des individus, la Cnav procède à des opérations de consolidation : validation des données ; envoi à chaque partenaire (Insee, services fiscaux,…) des données le concernant ; recodage et intégration dans le SNGC de la

15 Sur l'application SNGI, apparaît, à la place du terme « nom de famille », le terme « nom officiel ». Il semble qu'il s'agisse de la même chose.
16 Le SNGI est plus large que le RNIPP, car il donne des informations de rattachement (mutation vieillesse, RNIAM).

partie des données nécessaires pour le traitement des retraites ; destruction des données initialement transmises par les différents organismes gestionnaires des prestations sociales.

1.2.2.2 Intérêt et limites

Les principales caractéristiques des données issues de la Cnav sont leur exhaustivité et leur qualité : pour des raisons évidentes (elles servent de base au calcul des retraites), ces données sont complètes et particulièrement bien validées, notamment pour les périodes les plus récentes, et leur qualité (complétude et exactitude) s'améliore régulièrement au fil des années avec l'informatisation du recueil à la source. Les bases de données de la Cnav peuvent grandement faciliter des opérations particulièrement lourdes et complexes, dont les résultats sont souvent médiocres, et qui sont courantes dans de nombreuses études. Pour l'essentiel, ces opérations concernent :

- ***Le suivi et le traçage des sujets*** : tous les épisodes socioprofessionnels de la quasi-totalité des personnes vivant en France sont enregistrés de façon régulière et détaillée ; seules les personnes très désocialisées et ne bénéficiant d'aucun salaire et d'aucune prestation sociale échappent à cet enregistrement. Il est donc théoriquement possible de suivre les personnes incluses dans un protocole longitudinal de cohorte tout au long de leur vie et de minimiser ainsi les perdus de vue.
- ***L'accès aux données socioprofessionnelles*** : certains domaines de la recherche et de la surveillance, notamment l'épidémiologie sociale et l'épidémiologie des risques professionnels, s'intéressent particulièrement au statut social et professionnel et à son évolution dans le temps. Les données enregistrées dans les bases de la Cnav sont particulièrement riches de ce point de vue, d'une très bonne qualité, et susceptibles d'intéresser différentes équipes, aussi bien pour sélectionner des sujets à inclure dans des études sur des critères socioprofessionnels, que pour avoir accès aux données socioprofessionnelles les concernant tout au long de suivis de longue durée.

Les principales limites des bases de la Cnav dans l'optique d'une utilisation pour la recherche et la surveillance en santé publique sont l'imprécision des données concernant la catégorie socioprofessionnelle (souvent donnée sur les deux premiers chiffres seulement) et le fait qu'elles sont parfois manquantes.

1.2.3Autres bases de données pertinentes

Outre les trois dispositifs cités, d'autres bases de données nationales et d'un grand intérêt potentiel pour la recherche et la surveillance sont actuellement en construction et viendront prochainement enrichir le dispositif national, comme celle des mutuelles (Système national des données mutualistes). Ces sources peuvent être qualifiées de « généralistes ».

Il existe également des bases de données plus spécifiques, enregistrant de façon exhaustive les personnes présentant une caractéristique de santé particulière. Certaines couvrent la totalité de la population française, d'autres un territoire plus restreint (département ou région).

Sans chercher à être exhaustif, on peut citer parmi les bases de données ayant un intérêt pour la santé publique :

- Les registres de maladie : registres des cancers, des maladies rares, des malformations congénitales, des cardiopathies ischémiques, des accidents vasculaires cérébraux.
- REIN (Réseau épidémiologie et information en néphrologie), qui enregistre les patients en insuffisance rénale chronique traitée par un traitement de suppléance.
- Les certificats de santé de l'enfant, en particulier du 8^{e} jour ; celui du 24^{e} mois est aussi intéressant, car il permet de disposer des informations concernant la vaccination.
- Le système d'information de l'Établissement français du sang (EFS) qui enregistre les donneurs et les produits reçus par des patients.
- Cancer : des fiches standardisées de compte-rendu anatomopathologique (CRFS) ont été élaborées par l'INCa et la Société française de pathologie (SFP) pour la plupart des localisations de cancer. Ces CRFS devraient alimenter le dossier communicant de cancérologie (DCC), service du dossier médical personnel (DMP), expérimenté dans sept régions ; la mise en place du DCC sur l'ensemble du territoire est prévue avant la fin 2013. Le schéma général de ce système prévoit la constitution d'un entrepôt de données à visée épidémiologique, mais les conditions de mise en œuvre de celui-ci ne sont pas arrêtées.
- Reproduction : l'Agence de la biomédecine gère le registre national des tentatives de FIV (fécondation *in vitro*), ainsi que le registre national des IMG (interruptions médicales de grossesse).
- Handicap : la Caisse nationale de solidarité pour l'autonomie des personnes âgées et des personnes handicapées (CNSA) met actuellement en place le SipaPH (Système d'Information partagé pour l'autonomie des personnes handicapées) destiné à permettre la mise à disposition de données aidant au pilotage des politiques en faveur des personnes en situation de handicap. Il est actuellement en phase de mise en place dans les maisons départementales du handicap, et l'informatisation des dossiers est aussi en train de débuter.
- L'échantillon démographique permanent (EDP), créé en 1967. Il correspond à peu près à un sondage au 100^{e} de la population (4 jours de naissance) ; depuis 2006, son effectif a été quadruplé (16 jours de naissance). Pour chaque personne incluse, il contient des informations issues des bulletins d'état-civil de naissance, de mariage, de reconnaissance et de décès depuis 1968, ainsi que des recensements de 1968, 1975, 1982, 1990 et 1999. Les sujets décédés pour lesquels un bulletin de décès a été reçu par l'Insee sont conservés dans l'EDP. Par sa taille, l'échantillon démographique permanent permet des analyses fines qui peuvent notamment prendre en compte les effets de génération et des différenciations selon les qualifications, l'origine...

- La base des prestations sociales de la Caisse nationale d'allocations familiales pour la population générale et la Caisse centrale de la mutualité sociale agricole avec des données identiques.
- Données de l'administration fiscale concernant les revenus, qui peuvent beaucoup apporter aux travaux sur les déterminants sociaux de la santé et l'étude des inégalités sociales de santé.

1.3 QUELQUES EXEMPLES D'UTILISATION POSSIBLE DES BASES DE DONNEES ADMINISTRATIVES ET MEDICO-ADMINISTRATIVES NATIONALES POUR LA RECHERCHE ET LA SURVEILLANCE

Il est clair que les bases de données administratives et médico-administratives nationales ne peuvent être une panacée qui résoudrait toutes les difficultés rencontrées par les investigateurs, mais elles pourront apporter une aide importante. Ceci est particulièrement vrai pour les études et recherches de grande dimension et de longue durée, comme les études de cohorte qui vont continuer de se développer dans beaucoup de domaines, notamment en épidémiologie, où l'effectif envisagé de certaines cohortes ne se compte plus en dizaines, mais en centaines de milliers de sujets, une envergure comparable à ce qui existe dans plusieurs pays[17]. Les très grandes études cas-témoins en population générale, les systèmes de surveillance épidémiologique, les études concernant le recours aux soins, etc. peuvent bénéficier de ces bases de données. De plus, dans le domaine de la surveillance, la rapidité de la remontée des informations dans le SNIIR-AM, qui semble d'ailleurs en constante amélioration, en fait un outil spécifique pouvant contribuer, outre la surveillance « au long cours », à la surveillance et les investigations à moyen et court termes, comme le suivi d'épidémies, l'investigation de clusters, la surveillance « autour » d'une catastrophe environnementale. Sa disponibilité devrait être un atout pour répondre aux fortes contraintes de temps inhérentes aux missions de surveillance des agences sanitaires.

Il est évidemment impossible d'imaginer toutes les utilisations possibles de bases de données aussi riches en informations et couvrant des domaines aussi différents que les grandes bases nationales médicales et socioéconomiques. Cependant, afin d'illustrer l'apport potentiel d'une large ouverture de ces bases à des utilisateurs diversifiés, on peut évoquer quelques utilisations typiques, ayant déjà fait l'objet d'expérimentations ou qui sont en préparation.

1.3.1 Utilisation de chaque base de données indépendamment des autres

Les bases de données médico-sociales existantes couvrent de façon large des domaines spécifiques. Même si, comme on l'a vu, elles présentent certaines limites de nature diverse, de nombreuses utilisations relevant de la recherche, de la surveillance ou des études peuvent bénéficier de l'accès indépendant à l'une ou l'autre de ces bases.

17 Voir par exemple le site : http://www.p3gobservatory.org/studylist.htm.

1.3.1.1 Analyse de données contenues dans chaque base

Depuis longtemps, les données de l'assurance maladie sont utilisées à des fins descriptives, pour **estimer la fréquence de divers paramètres d'intérêt concernant les consommations de soins et la santé**, malgré diverses limites concernant la qualité des données.

La mise en place du SNIIR-AM, qui permet de combiner, pour les mêmes sujets, données de l'assurance maladie et données d'hospitalisation, et l'introduction du chaînage des données individuelles dans le PMSI ont permis d'améliorer très nettement la qualité des estimations, notamment grâce à des travaux développant des algorithmes destinés à identifier avec une bonne validité des pathologies spécifiques. On voit ainsi depuis peu des travaux présentant des estimations de la prévalence et/ou de l'incidence de certains cancers à partir des données du PMSI[18] au moyen d'algorithmes combinant diagnostics et actes techniques, ou des résultats concernant la maladie de Parkinson ou l'asthme à partir des données d'ALD et de consommations de certains médicaments[19].

Dans le **domaine de la pharmacoépidémiologie**, il est possible de réunir des échantillons d'effectif important de sujets correspondants à un ou plusieurs critères d'intérêt ; qui plus est, il est possible de suivre les sujets sélectionnés de façon longitudinale. Un exemple récent et largement médiatisé est celui de l'étude du risque de valvulopathies cardiaques chez les patients diabétiques utilisateurs de benfluorex (Médiator®). Une cohorte exhaustive des diabétiques affiliés au régime général, âgés de 40 à 69 ans et ayant présenté au moins trois remboursements d'antidiabétiques oraux et/ou d'insuline à des dates différentes a été constituée à partir du SNIIR-AM ; plus d'un million de sujets a ainsi été inclus, et des comparaisons entre exposés (consommation de benfluorex en 2006) et non-exposés (aucune consommation de benfluorex en 2006, 2007 ou 2008) ont porté sur les hospitalisations pour insuffisance mitrale ou aortique ou chirurgie de remplacement valvulaire pour insuffisance valvulaire survenus en 2006 ou 2007 recherchées dans le PMSI[20].

Au-delà de ce travail particulièrement démonstratif de l'intérêt du SNIIR-AM, il est clair que de très nombreuses études de pharmacoépidémiologie et **de suivi post-AMM** peuvent être réalisées uniquement à partir de cette base. Ceci est particulièrement vrai dans le cas de l'étude de situations peu fréquentes, comme des maladies rares, ou des traitements très spécifiques, qui peuvent nécessiter l'étude de la totalité des sujets concernés : dans de tels cas, le recours au SNIIR-AM est la seule méthode possible. D'une façon plus générale, face aux difficultés opérationnelles d'un suivi détaillé des patients bénéficiaires de traitements spécifiques en termes de consommations de soins et d'événements de santé, il faut souligner qu'il existe une forte demande de l'Afssaps, de la HAS et du ministère chargé de la santé pour ce type d'utilisation du SNIIR-AM, notamment pour les études de suivi post-AMM.

---

18 Couris et al. Breast cancer incidence using administrative data: correction with sensitivity and specificity. J Clin Epidemiol. 2009;62:660-6.

19 Moisan F et al. Prediction model of Parkinson's disease based on antiparkinsonian drug claims. Am J Epidemiol 2011; 174:354-363. Iwatsubo Y et al. Prediction model of asthma using antiasthma drug claims for epidemiological surveillance of asthma in self-employed workers in France. EPICOH Conference, Oxford, 7-9 September 2011.

20 Weill A, Païta M, Tuppin P, Fagot JP, Neumann A, Simon D, Ricordeau P, Montastruc JL, Allemand H.Benfluorex and valvular heart disease: a cohort study of a million people with diabetes mellitus. Pharmacoepidemiol Drug Saf. 2010;19:1256-62.

Un autre domaine d'utilisation des données du SNIIR-AM est l'étude de **phénomènes territoriaux**, notamment ce qui concerne les inégalités territoriales de soins de santé, mais aussi dans le domaine environnemental. Les études à l'échelle d'un territoire limité peuvent en effet réunir la totalité des personnes qui y habitent. La couverture nationale exhaustive de la population permet des études de comparaison entre zones géographiques, même de petite taille ou de faible population. L'analyse du recours aux soins peut bénéficier de données concernant des échantillons ou la totalité des patients qui sont traités pour une pathologie donnée, qui consultent tel type de professionnel ou qui utilisent tel médicament ou dispositif médical ; de plus, ces analyses peuvent être transversales ou longitudinales, permettant ainsi d'étudier des filières et des parcours de soins ou l'impact d'expositions environnementales.

À titre d'exemple dans le domaine environnemental, une étude de l'InVS sur les gastro-entérites d'origine hydrique montre que les données du SNIIR-AM peuvent contribuer au repérage de secteurs vulnérables quant à la qualité de leurs ressources en eau[21]. Dans le domaine de la surveillance autour des catastrophes, les données locales de l'assurance maladie ont été mobilisées pour suivre l'impact de la catastrophe d'AZF ; une étude est actuellement en cours sur les données du SNIIR-AM afin d'évaluer l'impact de la tempête Xinthia sur la consommation de psychotropes. Il faut cependant souligner que si l'adresse des personnes figure dans les bases de l'assurance maladie à l'échelle locorégionale, l'information disponible dans le SNIIR-AM se limite à la commune de résidence du bénéficiaire ; on peut également disposer de la commune d'exercice des professionnels de santé (prescripteur, pharmacie…).

Dans un domaine différent, les bases de données de la Cnav peuvent apporter une aide majeure pour l'étude de phénomènes socioéconomiques intéressant au premier chef **l'épidémiologie sociale ou l'épidémiologie des risques professionnels**. Ainsi, le Département santé travail de l'InVS discute actuellement avec la Cnav de la reconstitution de cohortes industrielles constituées de salariés ayant travaillé dans certaines entreprises à des fins d'analyse de clusters et, plus généralement, de mise en place de systèmes de surveillance des risques professionnels.

1.3.1.2 Enrichissement d'enquêtes avec recueil de données auprès des personnes

Dès 2006, le Cnis a souligné la complémentarité entre sources administratives et données d'enquêtes[22]. L'appariement de données administratives avec données d'enquête s'est développé ces dix dernières années[23] mais reste encore relativement rare.

Depuis plusieurs années, l'Insee et l'ensemble du système statistique public se sont engagés dans un mouvement croissant de recours aux sources administratives et médico-administratives. Dans cette perspective, il a été décidé de compléter le dispositif des enquêtes Handicap-Santé (HSM) 2008 et 2009 en les appariant pour la première fois avec les données de remboursements de l'assurance maladie (SNIIR-AM). Parmi

21 Utilisation des données de l'assurance maladie pour évaluer l'impact sanitaire d'une épidémie de gastro-entérites d'origine hydrique, Bourg-Saint-Maurice (Arc 1800), 2006. BEH 31, 6 septembre 2011.
22 Cnis (2006), Chroniques n°5, Enquêtes statistiques et sources administratives : une complémentarité à exploiter.
23 Gensbittel M.-H., Riandey B., Appariements sécurisés et statistiques (2000-2011) : Une décennie d'expériences. Courrier des statistiques n°131 Septembre 2011.

les avantages, on peut citer l'intérêt pratique, l'appariement permettant notamment de réduire le temps d'enquête et d'alléger la charge de réponse pour les enquêtés, ainsi que l'amélioration de la qualité des études par le recueil de données sans « biais de mémoire » ou de « biais de désirabilité sociale » comme cela peut être le cas pour les enquêtes en face-à-face, par téléphone ou par auto-questionnaire. Cette base de données issue de l'appariement HSM-SNIIR-AM a permis de développer des travaux originaux, en particulier un chiffrage des dépenses de santé des personnes âgées dépendantes dans le cadre de la préparation de la réforme de la dépendance. Le parcours de l'appariement s'est cependant avéré très lourd et complexe, car si les sources administratives et médico-administratives présentent l'incomparable avantage d'être déjà disponibles, les utiliser peut poser d'importantes difficultés[24].

Le suivi de cohortes épidémiologiques longitudinales peut également bénéficier **de l'enrichissement des données recueillies directement auprès des sujets par des données de consommation de soins et d'hospitalisation dans les bases de données de santé**. Un exemple est celui de l'étude Entred (*Échantillon national témoin représentatif des personnes diabétiques*) coordonnée par l'InVS, qui s'intéresse à la qualité de prise en charge médicale des diabétiques, à l'évolution du contrôle des facteurs de risque vasculaire et de la fréquence des complications chez les personnes diabétiques. Deux cohortes représentatives d'environ 10 000 diabétiques (respectivement constituées en 2001 et 2007), tirées au sort dans les bases de l'assurance maladie parmi les personnes ayant bénéficié d'au moins trois remboursements de médicaments antidiabétiques oraux et/ou d'insuline au cours des douze derniers mois. Entred 2007 a reposé sur cinq sources de données : 1) un suivi de la consommation de soins (médicaments, actes médicaux et biologiques, montants remboursés) pour l'ensemble de l'échantillon tiré au sort ; 2) un suivi des hospitalisations de 2006 à 2009 pour les personnes n'ayant pas refusé de participer à l'enquête ; 3) une enquête téléphonique réalisée auprès des personnes ; 4) une enquête postale détaillée auprès des personnes ; 5) une enquête postale auprès des médecins des personnes diabétiques[25].

D'autres projets complétant les recueils auprès de personnes par les bases de données nationales sont menés en collaboration avec la Cnam-TS. Sans prétendre être exhaustif, on peut citer l'enquête sur la santé et la protection sociale (ESPS) de l'Irdes[26], la cohorte Record[27] (étude des effets du contexte physique et social de résidence sur la santé cardiovasculaire de l'Unité Inserm 707), le projet Cesir[28] (étude de l'influence de la consommation de médicaments et de l'état de santé sur l'insécurité routière, réalisée par l'Unité 897 de l'Inserm) ou les programmes de suivi post-professionnel Amiante Spirale[29] (Inserm Unité 1018 et Cnam-TS) et Espri[30] (DST-InVS et RSI).

24 L'appariement Handicap-Santé et données de l'assurance maladie : Une source de données originale, mais un parcours semé d'embûches. Alexis Montaut, Lucie Calvet, Gérard Bouvier, Lucie Gonzalez. Journées de méthodologie statistique, INSEE. http://jms.insee.fr.
25 http://www.invs.sante.fr/entred/.
26 http://www.irdes.fr/EspaceRecherche/Enquetes/ESPS/index.html.
27 http://www.u707.jussieu.fr/ds3/Recherche.htm.
28 Orriols et al. (2010) Prescription Medicines and the Risk of Road Traffic Crashes: A French Registry-Based Study. PLoS Med 7(11): e1000366. doi:10.1371/journal.pmed.1000366.
29 http://www.spirale.rppc.fr/amiante.html.
30 http://www.invs.sante.fr/surveillance/espri/default.htm.

Concernant les **aspects socioprofessionnels**, certains importants projets de recherche et de surveillance épidémiologique récents s'appuient également largement sur des appariements avec les bases de données nationales de la Cnav, comme la cohorte Constances[31], qui complète les données recueillies auprès des individus et dans le SNIIR-AM par des données de situation sociale et de trajectoire professionnelle de façon prospective, ce qui procure un avantage considérable.

Enfin, il faut souligner qu'un des avantages du recours aux bases nationales est qu'il permet d'avoir au minimum pour tous les sujets un suivi passif et **d'éviter, ou du moins de limiter fortement, les « perdus de vue »** dans les études longitudinales, qui constituent des sources de biais potentiel qui peuvent être majeurs.

1.3.2 Appariement de bases de données entre elles

Dans le cadre du projet AMPHI (2010-2013), le CépiDc, en partenariat avec la Drees et la Cnam-TS, réalise l'appariement entre causes médicales de décès et données du SNIIR-AM. Il s'agit d'une étude expérimentale nationale, portant sur les séjours hospitaliers en court séjour (PMSI-MCO) de 2008-2009 appariés, *via* le SNIIR-AM, aux données médicales de décès pour les personnes décédées dans l'année suivant la sortie (décès 2008-2010). L'objectif est de décrire la mortalité hospitalière et post-hospitalière, pour, à terme, évaluer la faisabilité d'indicateurs de mortalité par établissements représentatifs de la qualité des soins.

Rappelons que le SNIIR-AM, à l'instar de pratiquement toutes les sources médicales, ne contient pas de données sur la situation socioprofessionnelle des personnes ; et que, de leur côté, les bases de la Cnav ne contiennent pas de données sur la santé (en dehors de données concernant certaines prestations sociales occasionnées pour raisons de santé). Dans un contexte de fort développement des recherches en **épidémiologie sociale**, et où les études concernant les **inégalités sociales et territoriales de santé**, les **risques professionnels ou la pénibilité du travail** sont particulièrement nécessaires pour venir en appui aux politiques publiques en matière de santé et d'emploi, des appariements à l'échelle des individus permettant de combiner des données en provenance du SNIIR-AM et de la Cnav (voire d'autres bases nationales) sont indispensables. Ceci a notamment un intérêt particulier pour constituer un « système permanent de surveillance des inégalités de santé » comme l'ont recommandé les rapports des groupes de travail sur les inégalités de santé[32] et sur les systèmes d'information pour la santé publique[33] du HCSP.

Quelques projets d'appariement de bases de données ont déjà été réalisés ou sont en cours. Ainsi, la base HYGIE gérée par l'Irdes provient de l'échantillon au 1/20^{e} de la population de personnes âgées de 22 à 70 ans en 2005 et ayant cotisé au moins une fois au régime général de retraite au cours de leur vie. Il s'agit d'un panel (cohorte) constitué de données issues du SNGC et du système national statistiques prestataires (SNSP) de la Cnav, et de données du SNIIR-AM obtenus par appariement (le taux

31 http://www.constances.fr/.

32 Haut Conseil de la santé publique. Les inégalités sociales de santé : sortir de la fatalité. Rapport HCSP, décembre 2009.

33 Haut Conseil de la santé publique. Les systèmes d'information pour la santé publique. Rapport HCSP, décembre 2009.

d'appariement est de 96,8 %). La base contient l'historique des salaires et des trimestres validés, consommations médicales, les ALD, les AT-MP et les arrêts de travail (maladie et AT/MP). La population d'analyse est d'environ 500 000 bénéficiaires (actifs et retraités). Parmi les travaux en cours, on peut citer : (i) l'analyse des mécanismes d'arrêts de travail des salariés du privé, en lien avec la nature et les spécificités des établissements qui les emploient ; (ii) la connaissance de l'impact des maladies chroniques psychiatriques sur les parcours professionnels[34].

Un premier appariement entre les causes médicales de décès et l'EDP d'une part, le panel DADS d'autre part a été réalisé dans le cadre de la surveillance systématique de la mortalité par profession et par secteur d'activité en population générale (InVS, projet COSMOP). Le projet EDISC (Inserm Unité 1018) a également réalisé diverses analyses sur les inégalités sociales de mortalité à partir de l'EDP. Le renouvellement à intervalle régulier de ces appariements présente un grand intérêt, et le CépiDc, l'InVS et l'Inserm travaillent actuellement avec l'Insee dans ce sens.

Le CépiDc[35] souhaite pouvoir apparier les causes individuelles de décès avec les données sociales détenues par la Cnav ; il en est de même de la Cnam-TS qui souhaite réaliser une opération d'appariement des sujets inclus dans l'échantillon généraliste des bénéficiaires (EGB) avec celles du CépiDc et de la Cnav. Dans un autre domaine, il serait possible de (re)constituer des cohortes de personnes travaillant dans des entreprises d'intérêt particulier et de les apparier avec des bases de données de santé à des fins d'analyse de clusters de maladies, de surveillance des risques professionnels en général ou pour d'autres besoins.

Ces quelques exemples montrent tout l'intérêt des appariements entre bases de données nationales dont on n'a pas encore exploré le très riche potentiel. Ainsi, le système d'information de l'Établissement français du sang (EFS) qui est centré sur les donneurs et les produits reçus par des patients pourrait être apparié avec les systèmes d'information hospitaliers. En effet, l'EFS dispose de données sur les produits administrés aux patients et les hôpitaux ont l'information sur les patients *via* le PMSI : l'appariement de ces deux sources permettrait de décrire les utilisateurs de produits sanguins, les contextes médicaux de la prescription ou leur devenir. L'assurance maladie de son côté n'a pas non plus de données sur ce sujet, puisque les produits sanguins ne sont pas dispensés en ambulatoire. Pourtant, des exemples locaux d'analyse conjointe des différentes bases existent, qui montrent l'intérêt potentiel d'appariements à des échelles plus larges (projet en cours au CHU de NICE en collaboration avec l'EFS Alpes-Méditerranée).

34 http://www.irdes.fr/EspaceRecherche/Partenariats/Hygie/Presentation.html
35 Service de l'Inserm qui gère la base de donnés nationale des causes de décès.

2 LES PRINCIPALES DIFFICULTES POUR L'UTILISATION DES BASES DE DONNEES NATIONALES A DES FINS DE RECHERCHE ET DE SURVEILLANCE

2.1 OBSTACLES REGLEMENTAIRES ET LEGAUX

Le cadre juridique permettant d'accéder aux données à caractère personnel des bases nationales est le suivant. Le dispositif légal et règlementaire concernant la protection des données à caractère personnel dans le domaine de la santé est encadré notamment par la loi n° 78-17 du 6 janvier 1978 modifiée par la loi du 6 août 2004 relative à l'informatique, aux fichiers et aux libertés, le décret n° 2005-1309 du 20 octobre 2005 modifié par le décret n° 2007-451 du 25 mars 2007 relatif à l'informatique, aux fichiers et aux libertés, ainsi que par les articles 226-16 à 226-21 du code pénal et les articles L.4113-7 et L.4163-9 du code de la santé publique.

2.1.1 Le SNIIR-AM

Concernant le SNIIR-AM, l'arrêté du 20 juin 2005 fixe la liste des organismes habilités à accéder aux données. Les organismes de recherche et les agences de santé font partie de cette liste ; par contre, les organismes privés à but lucratif en sont exclus.

2.1.1.1 Cas où on veut extraire des données du SNIIR-AM pour des personnes sélectionnées uniquement sur leurs caractéristiques

Il s'agit de situations où le demandeur veut uniquement extraire des données pour des personnes sélectionnées selon des critères spécifiques correspondant à des variables enregistrées dans le SNIIR-AM (âge, sexe, période, consommation de certains médicaments, etc.), sans croisement avec d'autres sources individuelles de données.

Dans tous les cas, s'agissant de données à caractère personnel, l'autorisation doit être donnée par la Cnil, conformément au régime d'autorisation réglementaire fixé, en fonction des objectifs, soit par la section II du chapitre IV (article 25), soit par le chapitre IX (art. 53 à 61) ou le chapitre X de la loi du 6 janvier 1978 relative à l'informatique, aux fichiers et aux libertés, modifié par la loi n° 2004-801 du 6 août 2004.

Des règles complémentaires s'appliquent selon les caractéristiques des données demandées, la nature du demandeur et le contexte de l'étude :

- Lorsqu'il s'agit de disposer de données provenant du SNIIR-AM, un avis de l'Institut des données de santé (IDS) est requis avant soumission à la Cnil (arrêté mentionné à l'article L.161-28-1 du code de la sécurité sociale).
- Lorsqu'il s'agit de demandes du service statistique public (Insee et services statistiques ministériels), c'est le Conseil national de l'information statistique (Cnis) qui doit donner un avis préalable[36].
- Prochainement, un nouveau GIP - dont la création est prévue par la loi relative au renforcement de la sécurité sanitaire du médicament et des produits de santé adoptée le 19 décembre 2011 - pourra également autoriser l'accès au SNIIR-AM : « *l'accès ou l'extraction peuvent être autorisés par un groupement d'intérêt*

36 Article 8 § II 7° de la loi du 6 janvier 1978 relative à l'informatique, aux fichiers et aux libertés.

public constitué à cette fin entre l'État, la Haute Autorité de santé, l'Agence nationale de sécurité du médicament et des produits de santé, l'Institut de veille sanitaire et la Caisse nationale d'assurance maladie des travailleurs salariés. » (article L. 5121-28 du code de la santé publique).

Cas particulier de l'Institut des données de santé (IDS) : l'arrêté mentionné à l'article L.161-28-1 du code de la sécurité sociale, qui approuve le protocole définissant les modalités de gestion et de renseignement du SNIIR-AM, autorise les membres de l'IDS[37] à accéder aux datamarts et à l'EGB. Les « membres de membres »[38] de l'IDS n'ont accès qu'aux datamarts, sauf les fédérations d'assureurs maladie complémentaires qui ont aussi accès à l'EGB. Ce même arrêté n'autorise l'accès à la base DCIR qu'aux régimes d'assurance maladie obligatoire. La réalisation de requêtes se fait dans le respect de contraintes fixées par la Cnil (limites dans le croisement de certaines données sensibles, pas d'affichage des résultats d'une requête lorsque le nombre de bénéficiaires est inférieur à dix) ; la réalisation de requêtes dans l'EGB se fait dans le respect de contraintes fixées par la Cnil (limites dans le croisement de certaines données sensibles).

2.1.1.2 Cas où l'on veut apparier les données individuelles du SNIIR-AM avec celles d'autres systèmes d'information disposant du NIR

Dans la mesure où les organismes gestionnaires des bases concernées sont habilités à utiliser le NIR, seule l'autorisation de la Cnil est nécessaire.

2.1.1.3 Cas où on veut extraire des données pour des personnes préalablement identifiées

Typiquement, il s'agit de situations où le demandeur veut enrichir des données individuelles recueillies auprès des sujets d'une étude, de professionnels de santé ou d'autres entités (cohorte, cas-témoins ou transversale) par des données de consommation de soins et d'hospitalisation.

S'il s'agit de données à caractère personnel ayant pour fin la recherche dans le domaine de la santé, l'avis du CCTIRS est nécessaire préalablement à la soumission d'un dossier à la Cnil (chapitre IX de la loi). Pour les études ayant obtenu un avis positif du CCTIRS et qui veulent compléter le recueil de données *via* le SNIIR-AM, l'avis de l'IDS doit être demandé, mais l'autorisation est automatique, grâce à un accord entre le CCTIRS, la Cnil et l'IDS.

Mais pour réaliser l'appariement entre ces données d'enquêtes et les données du SNIIR-AM, il faut, soit réaliser un appariement indirect (aussi appelé « probabiliste » ; *cf. infra*), soit détenir ou se procurer le NIR et l'anonymiser selon la même procédure que celle du SNIIR-AM (FOIN 1 et 2). Or « *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 numéro d'inscription des personnes au répertoire*

37 État, Cnam-TS, CCMSA, RSI, Unocam, Union nationale des professionnels de santé, Union nationale des régimes spéciaux, CNSA, FHF, Fédération des établissements hospitaliers et d'assistance privés, Fédération de l'hospitalisation privée, Collectif interassociatif sur la santé, Fédération nationale des centres de lutte contre le cancer.
38 Il s'agit des organismes qui sont des membres d'organismes participant directement à l'IDS.

national d'identification des personnes physiques » … sont autorisés par décret en conseil d'État après avis motivé et publié de la Cnil (article 27 de la loi relative à l'informatique, aux fichiers et aux libertés). Il faut donc un décret en conseil d'État pour pouvoir recueillir le NIR à des fins d'étude épidémiologique ou de surveillance sanitaire.

2.1.2 Les autres bases nationales

2.1.2.1 Base des causes de décès du CépiDc

L'accès au statut vital et aux causes individuelles de décès est réglementé par le décret n° 98-37 du 16 janvier 1998. Techniquement, l'accès au statut vital s'obtient *via* le Répertoire national d'identification des personnes physiques (RNIPP, géré par l'Insee) à partir de l'identité du sujet, son sexe et sa date et lieu de naissance. L'accès aux causes de décès repose ensuite sur un appariement indirect (probabiliste) basé sur le sexe, les dates et lieux de naissance et de décès.

2.1.2.2 Bases de la Cnav

Il n'existe pas de textes spécifiques concernant l'accès aux bases de la Cnav.

2.1.2.3 Bases de l'Insee

Jusqu'à une date récente, seul l'Insee était habilité à avoir accès aux données sur les personnes qu'il recueillait ou dont il était destinataire (en plus des archives nationales). La loi sur les archives de juillet 2008 (ouverture d'une possibilité d'accès aux données individuelles sur les personnes *via* le Comité du secret statistique) en rendent l'accès possible, à des fins de statistique publique ou de recherche scientifique ou historique, sur décision de l'administration des archives prise après avis du Comité du secret statistique, et sous réserve de l'accord de l'Insee. Celui-ci a fait savoir qu'il pourrait donner son accord sous réserve du respect de certains protocoles[39] :

- la règle générale est que des données indirectement nominatives (c'est-à-dire ne comportant ni le nom, ni le NIR - mais pouvant comporter un NIR haché) sont accessibles sur un centre d'accès sécurisé distant (CASD) (*cf. 2.2.2) ;*
- cependant, les services statistiques ministériels[40] peuvent avoir accès à ces mêmes données sur leur propre système d'information, à condition d'avoir produit un document prouvant un degré de protection suffisant et une étanchéité complète avec les autres systèmes d'information du ministère ; cette note doit être approuvée par le responsable des systèmes d'information de l'Insee ;
- des informations directement nominatives peuvent être fournies, uniquement pour la réalisation d'enquêtes statistiques, ayant reçu le visa du ministre chargé de l'économie, et après accord formel et explicite du comité de direction de l'Insee.

39 Voir le « Guide du secret statistique » à l'adresse : http://insee.fr/fr/insee-statistique-publique/statistique-publique/guide-secret-18-10-2010.pdf.
40 La liste de ces services est annexée au décret n° 2009-205 du 3 mars 2009 relatif à l'Autorité de la statistique publique.

Concernant l'EDP, base particulièrement utile pour l'étude des inégalités sociales de mortalité, elle est également couverte par le secret statistique et son accès se fait dans les mêmes conditions que les autres données couvertes par le secret statistique, c'est-à-dire, en général, *via* le CASD. Alors qu'il avait été possible de l'utiliser dans le cadre des projets COSMOP et EDISC déjà cités, les demandes d'accès à l'EDP doivent désormais passer par le Comité du secret statistique et l'accès et le traitement des données doivent le plus souvent être réalisés *via* le CASD. Le CépiDc a ainsi obtenu en 2011 un accord auprès du Comité du secret statistique pour réaliser un appariement entre l'EDP et la base des causes de décès. Ce recours quasi systématique au CASD implique la vérification des résultats produits par des statisticiens de l'Insee. Cette méthode de travail s'avère assez contraignante à l'usage.

2.1.2.4 Autres bases

À notre connaissance, il n'existe pas de textes spécifiques concernant l'accès aux diverses autres bases de données citées, à l'exception d'une recommandation de la Cnil concernant l'information des patients et les règles de sécurité à observer pour protéger la confidentialité des données[41].

2.1.3 Le problème de l'identifiant pour l'accès aux données à caractère personnel des bases nationales

Dans l'ensemble, on peut considérer que les textes actuels ne constituent pas un obstacle insurmontable à l'accès aux données à caractère personnel des bases nationales. Mais ils rendent complexes les circuits de transmission de données, ce qui induit des lourdeurs techniques et des délais parfois longs pour obtenir les autorisations nécessaires.

Globalement, la Cnil, chargée d'appliquer ces textes et de contrôler leur mise en œuvre, ne s'oppose pas à l'accès aux données des bases médico-administratives pour la santé publique. L'obtention des autorisations nécessite inévitablement des délais indispensables à la vérification du respect de la confidentialité des données à caractère personnel. Mais la complexité et le nombre croissant des projets et des demandes d'autorisation associées induisent une charge de travail pour la Cnil qui rend ces délais souvent imprévisibles et la gestion des projets complexes parfois difficile.

2.1.3.1 Le NIR

L'identifiant individuel actuellement utilisé dans les bases nationales étant le NIR[42], le problème majeur est l'impossibilité de pouvoir l'utiliser dans certaines situations. Il ne s'agit pas d'une interdiction stricte, mais la procédure d'autorisation pour la collecte du NIR (décret en conseil d'État) constitue dans les faits un obstacle infranchissable pour la plupart des équipes concernées.

En pratique, lorsqu'on souhaite pouvoir extraire des données des bases nationales, les situations qui se rencontrent sont les suivantes :

41 Cnil. Délibération n°03-053 du 27 novembre 2003 portant adoption d'une recommandation relative aux traitements de données à caractère personnel mis en œuvre par les registres du cancer.
42 On évoquera au paragraphe suivant les problèmes posés par l'Identifiant national de santé (INS).

- Études qui ne nécessitent pas d'appariement avec d'autres données recueillies auprès des personnes (cas 3.1.1.1 ci-dessus), comme dans l'exemple de l'étude Médiator® : il suffit alors d'obtenir les autorisations nécessaires, ce qui ne pose pas de problème particulier de nature juridique (uniquement pour les organismes habilités, dans le cas du SNIIR-AM).
- Appariements entre bases de données qui disposent chacune du NIR (cas de l'appariement de l'EGB avec les données de la Cnav, par exemple) : il suffit là aussi simplement d'obtenir les autorisations nécessaires.
- Appariements entre enquêtes individuelles et bases nationales : selon les modalités d'inclusion des sujets de l'enquête, deux situations peuvent se présenter :
    - les modalités d'inclusion impliquent un organisme qui dispose du NIR (organismes de protection sociale, employeur…) : un exemple est celui de la cohorte Constances où l'inclusion des participants est faite à partir d'un tirage au sort dans les bases de la Cnav, qui transmet le NIR des sujets sélectionnés à la Cnam-TS, qui peut donc appliquer les algorithmes FOIN pour extraire les données individuelles du SNIIR-AM et les transférer *via* des flux de données où le NIR a été supprimé.
    - l'inclusion se fait par contact avec les sujets de l'étude ou en recueillant les informations auprès de professionnels de santé, d'entreprises, ou d'autres sources : dans ce cas, l'investigateur ne peut pas collecter le NIR et l'extraction de données des bases nationales pour les sujets inclus est impossible de façon directe. Il est parfois possible malgré tout d'accéder indirectement au NIR en utilisant une méthode d'appariement probabiliste, à condition d'avoir pu collecter auprès des personnes des variables d'appariement pertinentes et de qualité suffisante. Un exemple maintenant bien établi est celui de l'accès au statut vital du RNIPP qui permet de retrouver les causes de décès par appariement indirect avec la base de données du CépiDc (*cf*. 2.1.2.1). Cette procédure n'est toutefois pas applicable systématiquement et le taux d'appariement est variable selon la qualité des données d'appariement.

2.1.3.2 L'INS

L'identifiant national de santé est institué par l'article L.1111-8-1 du code de la santé publique (loi nº 2007-127 du 30 janvier 2007) : « *Un identifiant de santé des bénéficiaires de l'Assurance maladie pris en charge par un professionnel de santé ou un établissement de santé ou dans le cadre d'un réseau de santé défini à l'article L. 6321-1 est utilisé, dans l'intérêt des personnes concernées et à des fins de coordination et de qualité des soins, pour la conservation, l'hébergement et la transmission des informations de santé. Il est également utilisé pour l'ouverture et la tenue du dossier médical personnel institué par l'article L. 161-36-1 du code de la sécurité sociale et du dossier pharmaceutique institué par l'article L. 161-36-4-2 du même code. Un décret, pris après avis de la Commission nationale de l'informatique et des libertés, fixe le choix de cet identifiant ainsi que ses modalités d'utilisation* ».

L'INS, dont le but est de permettre de sécuriser le contenu des échanges, doit être un identifiant pérenne généré aléatoirement et attribué par un système central, en garantissant les qualités suivantes : sans collision, sans doublon, non signifiant et non prévisible. L'INS est destiné à être le seul identifiant utilisé pour les applications en santé.

Son déploiement rendrait donc impossible à l'avenir l'appariement de données de santé avec d'autres sources utilisant le NIR comme identifiant. Afin d'éviter cette situation, il est envisagé que la Cnav maintienne une table de correspondance NIR-INS.

2.2 OBSTACLES ORGANISATIONNELS ET TECHNIQUES

En pratique, même si on a pu obtenir toutes les autorisations nécessaires, l'accessibilité effective des données reste aujourd'hui un problème majeur (voire insurmontable pour nombre d'équipes), du fait de difficultés pratiques. Celles-ci concernent les trois étapes indispensables pour être en mesure de disposer de données extraites des bases nationales « prêtes à analyser » : (1) identifier les personnes pour lesquelles on veut extraire des données dans les bases ; (2) extraire les données ; (3) mettre les données en forme pour les analyses.

2.2.1 Identification des personnes dans les bases de données

En dehors de la situation d'études qui ne nécessitent pas d'appariement avec d'autres données recueillies auprès des personnes (*cf*. 1.3.1.1 ci-dessus), la disponibilité du NIR dans le circuit des données est nécessaire pour retrouver dans les bases de données les sujets pour lesquels on veut extraire des données. Comme on l'a vu ci-dessus, cela est parfois possible directement *via* un organisme qui dispose du NIR, ou en utilisant des méthodes d'appariement probabiliste pour retrouver le NIR.

Mais dans tous les cas, l'utilisation du NIR pour l'identification des personnes **nécessite une participation active d'un organisme détenteur du NIR** (Insee, Cnav, Cnam-TS…).

Il n'existe pas à notre connaissance de dispositif formalisé mis en place par un ou plusieurs organismes détenteur du NIR pour prendre en charge ce type d'activité, comme par exemple le dispositif prévu dans le cadre du décret n° 98-37 autorisant l'accès au Répertoire national d'identification des personnes physiques (RNIPP) pour l'extraction des causes de décès. Les demandes sont donc discutées au cas par cas entre un organisme et un demandeur. Cette absence de règles explicites et de dispositif *ad hoc* constitue un obstacle opérationnel important à l'accès aux bases nationales qui les rend de fait inaccessible à la plupart des équipes potentiellement demandeuses.

2.2.2 Extraction des données

L'extraction proprement dite des données des bases nécessite également une participation active des organismes gestionnaires des bases de données, car cette activité implique diverses étapes techniques qui peuvent être lourdes et qui nécessitent des moyens humains, et une organisation rigoureuse : formulation des requêtes en fonction des besoins des utilisateurs, tests, production, vérifications, transferts sécurisés…

Il faut souligner les très importants efforts de la Cnam-TS qui a déployé un dispositif facilitant l'accès à l'échantillon généraliste des bénéficiaires (EGB) et qui cherche à faciliter l'accès aux données individuelles du SNIIR-AM et du DCIR, et de l'Insee qui a mis en place un centre d'accès sécurisé distant (CASD), ainsi que la bonne volonté de la plupart des autres organismes gestionnaires de bases de données nationales. L'accessibilité effective des données reste cependant encore fortement contingentée. Ceci s'explique en partie par le manque de moyens humains suffisants, tant au sein des organismes gestionnaires des bases, que du côté des organismes utilisateurs potentiels. En effet, une connaissance suffisante des bases de données par nature très complexes, de leur structure, des données enregistrées, des nomenclatures utilisées, et des évolutions constantes est indispensable pour les utiliser correctement. Ceci implique un investissement intellectuel important, le plus souvent sans commune mesure avec les ressources des équipes potentiellement concernées, particulièrement quand il s'agit d'une étude ponctuelle de durée limitée.

Cas particulier du CASD mis en place par l'Insee

L'Insee a mis en place un centre d'accès sécurisés distant (CASD), maintenant géré par le Groupe des écoles nationales d'économie et de statistique (Genes) ; le serveur du CASD est accessible au moyen de terminaux dédiés (les SD-Box) qui sont installés dans les locaux de divers organismes et permettent de procéder à des extractions de données et à des analyses statistiques.

Contrairement au dispositif mis en place par la Cnam-TS concernant l'accès à l'EGB (qui peut également se faire à distance), cette organisation ne permet pas d'exporter les données extraites des bases, et les analyses doivent être faites sur un serveur CASD situé dans les locaux du Genes ; seuls les résultats des analyses statistiques peuvent être transmis à l'utilisateur, qui ne peut rien exporter de façon autonome, ni sur une imprimante, ni sur un fichier externe. Si l'utilisateur extérieur souhaite apparier les données extraites à des données individuelles qu'il a lui-même recueillies, il doit importer le fichier correspondant sur le serveur CASD et procéder aux appariements sur ce serveur ; il ne pourra donc récupérer le fichier apparié que sur ce serveur.

Ce dispositif constitue un important progrès pour l'accès aux données individuelles détenues par l'Insee. Il ne satisfait cependant qu'une partie des besoins potentiels, notamment quand il s'agit de l'enrichissement des données recueillies directement auprès des sujets par des données provenant de l'Insee, comme dans le cas du suivi longitudinal de cohortes. **L'impossibilité d'exporter les fichiers appariés impliquerait en effet que le Genes prendrait entièrement en charge la gestion des bases de données de ces cohortes, y compris de leur utilisation, ce qui n'est évidemment pas réaliste, ni même concevable**.

2.2.3 Mise en forme des données pour les analyses

La complexité de la base de données du SNIIR-AM rend l'utilisation des données individuelles particulièrement difficile.

Dans cette base de données toutes les informations sont centrées sur les prestations remboursées. À une table centrale « Prestations » est donc rattaché un ensemble de tables caractérisant de façon plus précise les éléments de ces prestations

(bénéficiaires, établissements, biologie, pharmacie…). Une fois qu'un sujet est identifié, pour avoir l'ensemble des informations sur ses consommations de soins, il faut d'abord extraire les prestations dont il a bénéficié à partir de la table « Prestations ». Ceci permet de savoir de quel type de prestation l'individu a bénéficié (consultation, acte de biologie, remboursement de médicament…). Ensuite, pour avoir une description plus précise de la prestation, il faut chercher dans les tables annexes : table « Prestation affinée Biologie », table « Prestation affinée Pharmacie », table « Prestation affinée Transport »… Par exemple, la prestation permet de savoir que l'individu a eu une prescription, mais pour avoir le code CIP (code inter pharmaceutique) du médicament, il faut extraire l'information dans une table « Prestation affinée Pharmacie » à partir du numéro de la prestation.

Les données extraites par la Cnam-TS sont fournies à l'utilisateur sous forme « brute », et c'est à lui de procéder à la synthèse de ces informations pour disposer des variables qui l'intéressent réellement. L'exemple présenté dans l'encadré ci-dessous illustre le travail de synthèse de données nécessaire avant exploitation statistique.

Exemple de demande de données synthétisées

Un demandeur souhaite avoir une seule table, contenant les informations suivantes pour chaque individu de son étude : nombre de boîtes prescrites de médicaments antidiabétiques oraux ; nombre de consultations chez un cardiologue.

L'extraction brute du SNIIR-AM permet d'avoir, pour l'ensemble des sujets de l'étude, une table « Prestation », ainsi que l'ensemble des tables affinées pour la pharmacie, la biologie… Il faut donc à partir de ces tables construire une table contenant pour chaque individu l'ensemble des informations demandées. Chaque information nécessite un travail sur la base de données. Ainsi, pour avoir l'information sur le nombre de boîtes prescrites de médicaments antidiabétiques oraux et sur le nombre de consultations chez un cardiologue :

- nombre de boîtes prescrites de médicaments antidiabétiques oraux : il faut faire la jointure entre la table « Prestation » et la table affinée « Pharmacie » et sélectionner les codes CIP correspondant à des médicaments antidiabétiques oraux. Ensuite, il faut compter le nombre de boîtes par individu.

- nombre de consultations chez un cardiologue : il faut dans un premier temps sélectionner dans la table « Prestation » le code « spécialité médicale », correspondant au cardiologue, et compter le nombre de consultations par individu. Cependant, il ne faut pas faire la somme des lignes par individu : une consultation chez un cardiologue peut correspondre dans la base de données du SNIIR-AM à plusieurs lignes, et pour une prestation, il peut y avoir une ligne correspondant à l'acte de base et plusieurs lignes correspondant à des majorations ou à des compléments d'acte. Dans ce cas, il faut compter une seule consultation.

Chaque demande d'extraction est évidemment spécifique, générant donc chaque fois un travail long et fastidieux de gestion de données, qui nécessite des moyens spécialisés importants dont peu d'équipes disposent.

2.3 DIFFICULTES DIVERSES

2.3.1 Le consentement des personnes

Dans la plupart des situations d'utilisation des données à caractère personnel des bases nationales envisagées ici, il n'est en pratique pas possible de recueillir un consentement explicite des personnes concernées avant chaque étude ou appariement systématique. Ceci présente donc une difficulté vis-à-vis de la loi informatique et libertés, et nécessite des dispositifs d'information acceptables par la Cnil.

2.3.2 Un problème potentiel : l'identification indirecte dans le SNIIR-AM

Le SNIIR-AM a été conçu comme une base de données anonymes où les personnes sont enregistrées sous un numéro non signifiant et irréversible (FOIN-2).

Il faut cependant souligner que la richesse et la diversité des données, qui s'accroît régulièrement grâce aux appariements avec d'autres bases et à l'augmentation récente de la profondeur de durée de conservation du SNIIR-AM, rendra de plus en plus techniquement possible l'identification indirecte des personnes par croisement des données les concernant.

Un problème voisin est celui de l'analyse de données à caractère personnel à une échelle territoriale fine : ceci peut poser des problèmes de confidentialité, car il devient alors parfois possible d'identifier des personnes.

2.3.3 La propriété des fichiers appariés

Les fichiers issus de l'appariement de bases de données, ou de l'enrichissement d'une étude par des données en provenance des bases de données nationales, sont *de facto* en « multipropriété », car composés de données dont chaque source est propriétaire, selon des règles propres à chacun des organismes concernés. L'expérience montre que ceci peut poser des problèmes en termes de propriété intellectuelle et de droit d'usage.

2.3.4 Le maquis juridico-institutionnel

L'ensemble des aspects juridiques, organisationnels et techniques à prendre en compte pour pouvoir utiliser les bases de données nationales de façon respectueuse des textes et efficace en termes de résultats constitue un véritable maquis dans lequel les équipes de recherche et de surveillance ont souvent du mal à se repérer. Il faut connaître le contenu précis des bases de données, savoir quels textes s'appliquent selon sa situation institutionnelle et la nature de sa demande, trouver les interlocuteurs compétents dans les organismes concernés, établir plusieurs dossiers successifs de demande d'autorisation… L'absence, déjà évoquée, d'un ou plusieurs « guichets » destinés à accompagner les demandeurs dans leurs démarches est un obstacle pratique qui explique en partie la sous-utilisation des bases de données nationales.

3 PROPOSITIONS

3.1 L'UTILISATION DES BASES DE DONNEES NATIONALES : POUR QUI ? POUR QUOI ?

Avant d'envisager les propositions pour améliorer l'utilisation des bases de données nationales à des fins de recherche et de surveillance, il est important d'examiner qui sont les utilisateurs potentiels de ces bases et quelles règles il faudrait définir pour leur ouverture à des organismes extérieurs à ceux qui les gèrent.

3.1.1 Les utilisateurs potentiels des bases de données nationales

Les exemples cités montrent que les utilisateurs des bases de données sont potentiellement nombreux, pour des travaux de nature diversifiée. Il convient de distinguer les utilisateurs qui appartiennent à des structures publiques ou parapubliques, et les sociétés privées à but lucratif.

Structures publiques ou parapubliques : divers organismes de recherche publics et les établissements d'enseignement supérieur ont des équipes de différentes disciplines (épidémiologie, économie de la santé, sociologie, démographie…) qui sont demandeuses d'accès aux bases de données nationales dans le cadre de leurs recherches :

- Le Service statistique public, composé de l'Insee et des services statistiques ministériels.
- Les agences de santé, notamment l'Afssaps, la HAS, l'Agence de la biomédecine, les ARS réalisent ou font réaliser des études qui font appel à des données provenant des bases nationales. Une place à part doit être faite à l'InVS, dont plusieurs systèmes de surveillance, qui par nature ont vocation à couvrir l'ensemble de la population et à être pérennes, reposent largement sur les données des bases nationales.
- D'autres structures de santé publique, comme les Observatoires régionaux de la santé, certaines collectivités territoriales, ou des organismes d'études (comme l'Irdes) sont également demandeuses de données, tout comme certains services ministériels.
- Les organismes d'assurance maladie sont demandeurs d'appariements entre leurs propres bases et celles d'autres organismes.

Enfin, un des constats qui a été fait est que nombre des équipes qui, de façon évidente, pourraient bénéficier pour leurs travaux des bases de données nationales ignorent jusqu'à leur existence et connaissent mal les données disponibles et ne savent pas dans quelles conditions elles pourraient y accéder.

Structures privées à but lucratif : des structures privées à but lucratif sont des utilisateurs potentiels des bases de données nationales. Il s'agit notamment des industriels de la santé (produits pharmaceutiques, dispositifs médicaux…), mais aussi des assureurs (santé, retraite) et de bureaux d'études travaillant pour l'industrie. Actuellement, les textes n'autorisent pas les structures privées à but lucratif à accéder aux données du SNIIR-AM (du moins directement), même lorsqu'il s'agit de réaliser une

étude commandée par les pouvoirs publics. La Cnav ne fournit pas non plus de données à des organismes privés à but lucratif.

3.1.2 Règles d'ouverture des bases de données nationales

En préalable, il faut rappeler qu'il existe un ensemble de textes qui encadrent l'accès aux données des bases nationales et qu'en tout état de cause l'obtention des autorisations réglementaires est indispensable avant tout accès aux données ; les principaux textes concernés sont cités plus loin. Mais au-delà des textes, il est indispensable de se poser la **question de l'opportunité de l'ouverture en fonction de la nature de la demande et du demandeur**, car les bases de données concernées par cette note sont établies à des fins d'intérêt public, gérées par des organismes publics financés par des fonds publics ; de plus, le recueil des données individuelles est une obligation imposée par les pouvoirs publics à des fins de gestion et de contrôle et toute utilisation à d'autres fins doit être spécifiquement justifiée.

- **Nature de la demande** : il est évidemment impossible de définir à l'avance toutes les utilisations possibles des données contenues dans les bases nationales en dehors de celles qui ont été définies pour leur recueil initial. On peut simplement rappeler une règle consensuelle dans le monde de la recherche et de la surveillance : les projets pour lesquels des données sont demandées doivent avoir fait l'objet d'évaluations scientifiques préalables par des organismes légitimes (CCTIRS, Cnis, comités scientifiques, organismes subventionnaires…).
- **Nature du demandeur** : en raison de la finalité d'intérêt public qui préside à la constitution des bases de données nationales rappelée ci-dessus, il est légitime que leur utilisation soient réservée à des organismes publics ou parapublics. Il faut souligner que ce n'est pas seulement la nature juridique de l'organisme demandeur qui doit être prise en compte, mais la finalité d'intérêt public de la demande, des organismes privés à but lucratif confiant parfois des études à des structures de recherche publiques. Il faut cependant tenir compte du fait que les pouvoirs publics imposent aux industriels de santé la réalisation de certaines études (post-AMM) : il semble que, dans de tels cas, ceux-ci sont fondés à demander l'accès aux données publiques qui peuvent être nécessaires pour la bonne conduite de ces études, qu'elles soient réalisées directement par une structure privée à but lucratif ou par un intermédiaire « académique ».
- **Contrôle préalable et contre-expertise** : il semble donc légitime que les organismes gestionnaires des bases de données puissent exercer un contrôle préalable sur les demandes qui leur sont adressées et le cas échéant donner un avis négatif de façon argumentée, même si les demandes se situent dans le cadre d'études commanditées par les autorités de santé et même si elles ont obtenu les autorisations réglementaires nécessaires. Ils sont évidemment également fondés à mettre en œuvre ou commanditer des analyses s'ils le jugent utile.

3.1.3 La gouvernance

3.1.3.1 Définition de la gouvernance

Dans le contexte de l'utilisation des bases de données nationales pour la recherche et la surveillance, la « gouvernance » concerne l'organisation de l'accès aux bases, la supervision de la réalisation des demandes et le contrôle de l'utilisation des données ainsi transférées à des demandeurs extérieurs aux organismes gestionnaires. Il faut aussi prendre en compte l'attribution et le contrôle des ressources nécessaires et, le cas échéant, la priorisation des demandes.

Actuellement, comme on l'a souligné, il n'existe pas à proprement parler de gouvernance pour l'utilisation des bases de données nationales pour la recherche et la surveillance, chaque organisme gestionnaire de base « traitant » directement avec les demandeurs. Cet état de choses n'est pas satisfaisant et il est utile d'envisager une meilleure formalisation des modalités d'accès et d'utilisation des bases de données nationales.

Remarque : *on met ici à part l'aspect des autorisations réglementaires, qui relèvent comme on l'a rappelé de la Cnil et selon les cas, du CCTIRS, de l'IDS, du Cnis ou du GIP qui doit prochainement être créé dans le cadre de la loi relative au renforcement de la sécurité sanitaire du médicament et des produits de santé : ces autorisations sont nécessaires, en tout état de cause, dans l'état actuel des textes*.

3.1.3.2 Organisation de la gouvernance

On peut *a priori* envisager deux modèles de gouvernance :

- *Gouvernance décentralisée* : chaque organisme public gestionnaire de bases de données fixe des règles explicites d'accès (incluant une politique tarifaire et la possibilité de refuser l'accès à ses données) et met en place un « guichet » destiné à traiter les demandes et accompagner les demandeurs.
- *Gouvernance centralisée* : une structure centrale gère un guichet unique et fait office d'interface entre les demandeurs et les organismes gestionnaires de bases de données, selon des règles homogènes et sous le contrôle d'une instance de gouvernance unique.

La première solution est plus facile à mettre en œuvre et laisse une pleine autonomie à chaque organisme gestionnaire de bases de données. Elle implique que les organismes se dotent des moyens de mettre en place un guichet d'accueil. Son inconvénient principal est la multiplication des points d'accès (régimes obligatoires et sections locales mutualistes - SLM) et la possibilité que les organismes définissent des règles d'accès trop différentes. Il serait donc certainement judicieux qu'un minimum de règles communes inter-organismes soit établi et qu'une information d'ensemble homogène soit disponible pour les demandeurs potentiels.

La seconde solution est certainement plus commode pour les demandeurs potentiels ; elle faciliterait également l'homogénéisation des règles d'accès. Elle est cependant plus complexe à mettre en œuvre, car elle implique la création d'une structure *ad hoc* dotée de moyens. Cette structure devrait idéalement associer, sous l'égide des pouvoirs publics, les organismes concernés par l'utilisation des bases de données nationales

pour la recherche et la surveillance : organismes publics gestionnaires de bases de données et organismes utilisateurs de données. Ceci implique également d'élaborer des règles de gouvernance communes et de définir une politique budgétaire, voire tarifaire. Sur le plan pratique, il faut considérer que les organismes potentiellement concernés sont nombreux et qu'il n'est certainement pas réaliste de les associer tous individuellement dans une structure de gouvernance qui serait alors difficilement gérable. Il est de la responsabilité des pouvoirs publics de proposer des solutions réalistes, efficaces et acceptables par l'ensemble des organismes concernés.

3.2 PROPOSITIONS CONCERNANT L'IDENTIFIANT POUR L'ACCES AUX BASES DE DONNEES

Dans l'état actuel, l'accès aux bases de données nationales nécessite l'utilisation du NIR (on envisagera plus loin la situation où l'INS sera déployé). On a vu que ceci ne présente pas de problèmes dans certaines situations et les procédures pour accéder aux données individuelles fonctionnent sans difficulté particulière.

On traite donc ici des situations où il faudrait recueillir le NIR des personnes pour lesquelles on souhaite extraire des données (cas de l'appariement entre enquête individuelle et bases nationales : *cf*. 1.3.1.2), ce qui, en l'état actuel des textes, nécessite un décret en conseil d'État après avis de la Cnil.

3.2.1 Recueil du NIR en clair

Deux possibilités peuvent théoriquement être envisagées :

- Recueil en clair auprès des personnes ou des professionnels (auto-questionnaires ou enquêteurs) et transfert immédiat à un centre d'appariement sécurisé (CAS), structure jouant un rôle de tiers de confiance habilité à recevoir, dans le cadre d'études autorisées, les NIR des sujets concernés. L'application par le CAS de procédures d'anonymisation et de cryptage et des flux sécurisés pourrait alors permettre des opérations d'appariement dans des conditions garantissant le respect des contraintes de confidentialité des données à caractère personnel. Un problème majeur est le maintien de la confidentialité pendant les transferts jusqu'au CAS : sous réserve d'omission, la seule possibilité serait que les personnes sollicitées lui envoient elles-mêmes directement (par courrier ou Internet) leur NIR ; cette solution est cependant impossible à mettre en œuvre en pratique dans de nombreuses situations.
- Techniquement, le recueil peut aussi être fait sur ordinateur (y compris la clé pour minimiser les erreurs) auprès des personnes ou d'autres sources, puis immédiatement crypté, et transféré de façon sécurisée à un organisme habilité à disposer du NIR. On pourrait argumenter que la procédure est alors anonyme et dispensée du décret (ce qui reste à vérifier).

Dans l'état actuel des textes, **il semble que ces solutions ne puissent pas permettre d'éviter la prise d'un décret en conseil d'État spécifique de chaque étude**, procédure impraticable de fait. De plus, la Cnil pourrait considérer ces procédures comme un détournement de la loi, car pouvant être assimilées à un « traitement » du NIR au sens de la loi informatique et libertés. Enfin, ces procédures

ne sont pas applicables dans de nombreuses situations concrètes d'étude (enquêtes en situation d'urgence, par exemple).

3.2.2 Recueil par appariement indirect (probabiliste)

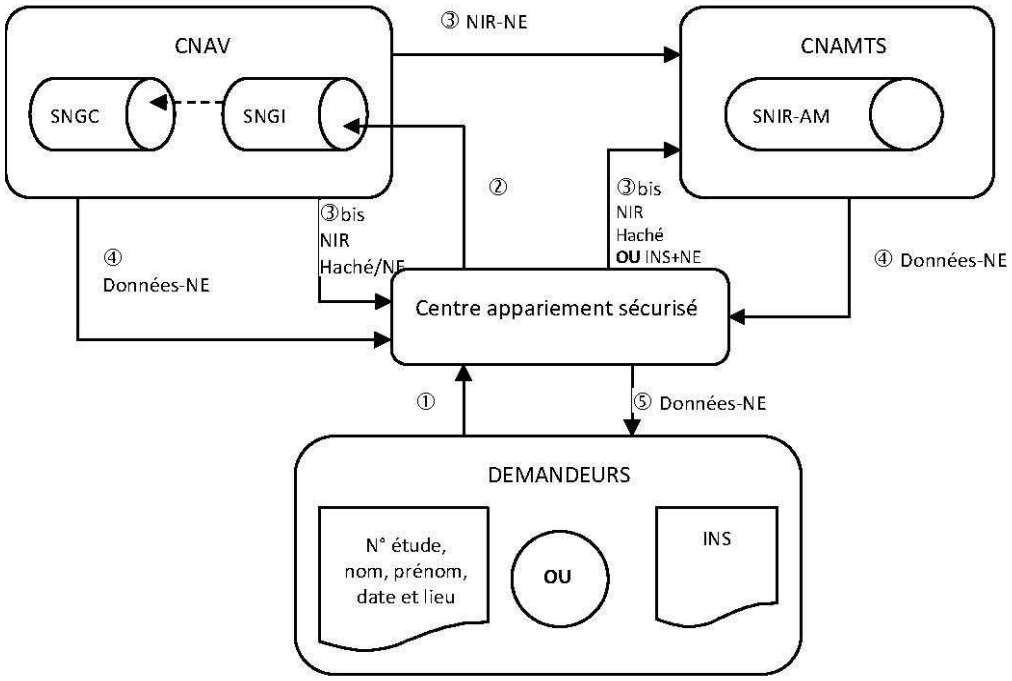
L'accès au NIR peut se faire par appariement indirect et consultation d'une base contenant les NIR de toute la population : RNIPP géré par l'Insee, RNIAM ou SNGI gérés par la Cnav. Cette méthode est employée depuis longtemps pour l'accès au statut vital selon la procédure définie par le décret n° 9 8-37 déjà cité et elle est également utilisée par la Cnav. Elle implique que le demandeur recueille uniquement les informations suivantes : nom, prénom, date et lieu de naissance des personnes concernées, et les transmettent (accompagnées d'un numéro d'étude destiné aux transferts des données) *via* un tiers de confiance à l'organisme détenteur de la base des NIR, qui peut ainsi retrouver les NIR par une méthode d'appariement indirect et les transférer à qui de droit, sans que l'investigateur n'en ait connaissance.

3.2.3 Aspects opérationnels

3.2.3.1 Appariement indirect

Principe : la mise en œuvre pratique implique la création d'un centre d'appariement sécurisé (CAS), tiers de confiance qui serait l'opérateur central des procédures résumées par le schéma ci-dessous.

Procédures d'accès aux bases de données nationales sans le NIR



Remarque : *on se restreint au cas de l'accès aux bases du SNIIR-AM et de la Cnav, mais le schéma peut être étendu à l'accès à d'autres bases identifiées par le NIR.*

Le point de départ est le demandeur qui a déjà recueilli les informations nécessaires à l'appariement avec les NIR (ou l'INS dans le futur). Les opérations sont les suivantes :

- le demandeur transmet au CAS un fichier comprenant pour chaque sujet le nom, prénom, lieu et date de naissance des personnes (ou l'INS), accompagné d'un numéro d'étude (NE) : flèche 1 ;
- le CAS transmet ces données à la Cnav : flèche 2 ;
- la Cnav apparie, par l'intermédiaire du SNGI, le NIR avec le numéro d'étude NE de chaque sujet, et transmet le fichier apparié NIR-NE à la Cnam-TS : flèche 3 ;
- la Cnam-TS extrait du SNIIR-AM, grâce au NIR, les données pour chaque sujet, et les transmet au CAS associées au NE : flèche 4 ;
- le CAS renvoie au demandeur le fichier de données identifiées par le NE : flèche 5.

On peut envisager des variantes à ce schéma de base :

- si le demandeur souhaite des données socioprofessionnelles de la Cnav, celle-ci les extrait du SNGC grâce au NIR et les transmet au CAS associées au NE : flèche 4 venant de la Cnav ;
- on peut imaginer aussi que la Cnav transmette des NIR hachés+NE au CAS au lieu d'envoyer des NIR directement à la Cnam-TS ; le CAS envoie alors les NIR hachés avec les NE à la Cnam-TS : flèches 3bis entre la Cnav et le CAS et entre le CAS et la Cnam-TS. Cependant, il faudrait que l'algorithme de hachage utilisé par la Cnav soit celui qui génère FOIN-1 pour que la Cnam-TS puisse appliquer le hachage FOIN-2, ce qui n'est actuellement pas possible ;
- au lieu du SNGI, on peut utiliser le RNIAM géré également par la Cnav ou le RNIPP de l'Insee (comme c'est le cas pour l'accès aux causes de décès) en tant que base contenant les NIR de la population : les procédures mises en œuvre seraient voisines.

Dans le cas de demandes répétées pour les mêmes sujets (suivi régulier de cohortes notamment), on peut envisager que la Cnav conserve de façon sécurisée le lien entre le NE et le NIR, ou la Cnam-TS le lien entre le NE et FOIN-1, afin d'alléger les opérations d'appariement ; ceci implique une organisation *ad hoc* au niveau de ces organismes.

Avantages : cette procédure a déjà été mise en œuvre, elle est éprouvée et fonctionne avec des performances satisfaisantes lorsque les informations initiales (nom, prénom, date et lieu de naissance) sont de bonne qualité. Elle présente le très grand avantage d'être parfaitement compatible avec la loi informatique et libertés. Elle implique cependant qu'au moins un organisme détenteur des NIR (Insee ou Cnav) mette en place un « service » *ad hoc*, avec les moyens afférents pour réaliser les opérations d'appariement et de transfert et qu'un décret autorise la généralisation des procédures prévues par le décret n° 98-37 (qui réglemente l'accès au statut vital et à la base des causes de décès) à l'ensemble des bases de données concernées.

Limites : tout séduisant qu'il paraisse, l'appariement indirect n'est pas adapté à toutes les situations. Les principales limites sont les suivantes :

- L'efficacité de la procédure dépend de la qualité des données pour les variables d'appariement : elle peut être excellente (proche de 100 %) lorsque les données sont sans erreur (cas de certains fichiers d'entreprise par exemple), mais beaucoup moins satisfaisante dans de nombreuses situations. Ainsi, l'expérience de l'InVS montre qu'environ 15 % des sujets sont retrouvés sans équivoque, environ 70 % des appariements nécessitent une validation (une variable avec une discordance) et environ 15 % des sujets ne sont pas retrouvés. Selon l'Insee, le taux de réussite est globalement de l'ordre de 80 %, ce qui est souvent très insuffisant par rapport aux objectifs du demandeur.
- Les études avec des données en provenance des professionnels de santé ne pourraient pas obtenir le lieu de naissance. Cette procédure devrait donc se baser sur l'INS, et aujourd'hui il n'est accessible que pour les titulaires de la carte vitale (ouvrant droit ; environ six millions de personnes ne possèdent pas d'INS) ; ce problème devrait être résolu théoriquement avec le passage à l'INS anonyme.
- Les procédures proposées impliquent des mesures d'ordre organisationnel et institutionnel : il faut créer un CAS, doté des moyens nécessaires et d'une gouvernance adéquate, ainsi que d'un encadrement juridique compatible avec les textes.

3.2.3.2 Décret-cadre en conseil d'État

C'est une solution envisagée par la Cnil qui propose l'élaboration d'un décret-cadre en conseil d'État « *permettant l'utilisation encadrée du NIR à des fins de recherche médicale et d'études en santé publique* »[43].

Cette solution s'impose dans le cadre juridique actuel, dans la mesure où il n'existe pas de procédures sans utilisation du NIR permettant de répondre à toutes les situations, comme on l'a vu. En effet, la Cnil, en proposant un décret cadre qui autoriserait après une procédure d'évaluation et d'autorisation définie, le recueil, la détention et l'utilisation du NIR par un organisme conduisant une recherche ou une surveillance en santé publique, reconnaît les besoins nouveaux de la recherche et de la surveillance au nom de l'intérêt général au regard des possibilités offertes par l'appariement des bases de données. Elle propose ainsi un cadre qui assouplit les démarches et procédures, tout en respectant la loi avec un contrôle renforcé en cas d'utilisation du NIR.

Un tel décret-cadre devrait reprendre certaines des contraintes ci-dessus et imposer évidemment pour chaque étude une autorisation de la Cnil. Des procédures opérationnelles doivent être élaborées, qui devraient *a priori* être proches de celles proposées ci-dessus pour les appariements probabilistes.

43 Courrier du 12 août 2010 du président de la Cnil au Premier ministre.

3.2.4 Problèmes particuliers

3.2.4.1 Ayants droit

Actuellement, l'identifiant individuel du SNIIR-AM est obtenu à partir du NIR de l'ouvrant droit. Comme on l'a indiqué plus haut, le fichier RFI permet d'attribuer leurs propres identifiants FOIN aux ayants droits du régime général par ajout du sexe et de la date de naissance. Pour les affiliés des autres régimes obligatoires et des SLM, des procédures équivalentes existent ; dans le cas où les sujets pour lesquels on veut extraire des données du SNIIR-AM n'appartiennent pas au régime général, il faut donc mettre en place des circuits spécifiques :

- en cas de recueil du NIR en clair directement auprès des personnes, on peut demander son propre NIR ainsi que celui de son ouvrant droit ;
- en cas d'appariement indirect (probabiliste), le schéma ci-dessus devrait être complété par une première étape consistant à consulter le RNIAM (géré par la Cnav) pour identifier le régime d'affiliation ; une deuxième étape interne à chaque régime permet d'obtenir le NIR souhaité.

Ces procédures alourdissent beaucoup les circuits et les flux de données et impliquent une participation active des organismes concernés. De plus, elles présentent des difficultés dans certaines situations : cas des enfants couverts par les deux parents, jumeaux de même sexe, changement de statut ayant droit/ouvreur de droits…

Une note complémentaire décrira plus en détail ce point spécifique, qui peut s'avérer problématique pour certaines utilisations du SNIIR-AM et du PMSI.

3.2.4.2 Le futur déploiement de l'INS

Une table de correspondance entre INS (ou INS haché) et NIR (ou NIR haché) est indispensable pour les appariements entre bases de données de santé, identifiées à l'avenir avec l'INS, et les autres bases identifiées avec le NIR. Il semble prévu que la Cnav gèrera cette table de correspondance, ce qui reste à confirmer.

On peut de plus envisager un problème potentiel, car il est envisagé par certains de constituer des bases de données avec des INS hachés de façon spécifique à chaque base : on ne pourrait alors pas utiliser l'INS pour apparier des bases protégées par leur propre hachage. Une solution serait que les INS de toutes les bases soient hachés à l'aide d'une clé commune, puis cryptés avec une clé spécifique à chaque base, et que ces clés soient confinées dans une structure de coordination des identifiants, qui pourrait être un CAS (centre d'appariement sécurisé, évoqué plus haut) habilité à recevoir les INS des sujets inclus dans les études impliquant un appariement avec des sources utilisant le NIR.

3.2.5 Conclusion

Il ressort de l'analyse présentée ci-dessus que les deux approches (appariement indirect ou probabiliste et recueil en clair du NIR autorisé par un décret-cadre) sont complémentaires. **Notre point de vue est qu'il faut appliquer un principe de parcimonie pour le recueil en clair du NIR et que la méthode d'appariement**

indirect décrite ci-dessus doit être privilégiée pour toutes les situations où elle est possible et suffisamment efficace, le recueil en clair du NIR n'étant alors pas indispensable. En pratique, c'est à la Cnil que revient d'examiner dans chaque cas la solution proposée par les équipes et à autoriser ou non sa mise en œuvre.

3.3 PROPOSITIONS CONCERNANT L'EXTRACTION ET LA TRANSMISSION DE DONNEES DES BASES NATIONALES

Actuellement, les extractions sont réalisées au sein des organismes gestionnaires des bases de données, après concertation avec les demandeurs quant au choix de la population concernée, des variables d'intérêt, etc. La transmission des fichiers issus des requêtes se fait par des voies diverses : DVD, télétransmission cryptée sécurisée… Les expériences sont maintenant suffisamment nombreuses et diversifiées pour qu'on puisse constater que l'extraction et la transmission des données ne présente pas de difficulté particulière.

Il reste cependant des obstacles à une ouverture plus large des bases nationales :

- **Difficultés pour définir les demandes d'extraction** : la complexité des bases de données (particulièrement du SNIIR-AM) et leur évolutivité les rendent difficilement compréhensibles aux demandeurs, comme on l'a souligné plus haut (3.2.2).
- **Délais d'extraction et de transmission** : les phases techniques d'extraction sont lourdes (formulation de requêtes, contrôles, mise en forme, transmission…). Lorsqu'il s'agit de demandes impliquant des extractions régulières planifiées, l'expérience montre que les organismes gestionnaires des bases se sont organisés pour répondre dans des conditions satisfaisantes. Lorsqu'il s'agit de demandes qui ne s'insèrent pas dans un projet planifié longtemps à l'avance, ils n'ont habituellement pas les ressources internes suffisantes pour répondre toujours dans les délais souhaités par les chercheurs (hors situations de « crise »), malgré les importants efforts récents de la Cnam-TS dans ce sens concernant le SNIIR-AM. Il existe donc une priorisation de fait du traitement des demandes, dont les critères ne semblent pas formalisés, et des délais difficilement prévisibles pour les demandeurs et qui peuvent être longs.

On peut imaginer en théorie qu'une solution serait de permettre un accès direct aux bases de données (entrepôt SNIIR-AM ou DCIR, SNGC…), à l'instar du dispositif que la Cnam-TS a mis en place pour l'EGB. Cette solution semble exclue pour diverses raisons : importante logistique informatique à mettre en place, complexité des bases impliquant une connaissance approfondie et constamment actualisée des utilisateurs (donc un important investissement et une formation permanente assurée par les gestionnaires des bases), sécurisation des accès et des transmissions, ainsi que vraisemblablement des modifications réglementaires, etc. En tout état de cause, une telle solution écarterait les « petits » demandeurs, comme c'est le cas pour l'EGB, en raison de l'investissement considérable nécessaire pour les utilisateurs.

Il semble donc réaliste que les organismes gestionnaires des bases nationales continuent d'assurer la charge de l'extraction et de la transmission des données, et on ne peut que souhaiter qu'ils soient en mesure de disposer des moyens nécessaires pour répondre aux demandes dans des conditions optimales.

Cas des bases de données de l'Insee

Comme on l'a vu, les bases de l'Insee ne sont accessibles que *via* le centre d'accès sécurisé distant (CASD) du Genes et les données extraites ne peuvent pas être exportées, pas plus que les fichiers appariés éventuellement générés. Il serait souhaitable que l'Insee reconsidère cette politique et puisse, notamment dans le cas du suivi longitudinal de cohortes, faire en sorte que sur dérogation justifiée, l'enrichissement des données recueillies directement auprès des sujets par des données provenant de l'Insee ne soit pas impossible.

3.4 PROPOSITIONS CONCERNANT L'UTILISABILITE DES DONNEES PROVENANT DES BASES NATIONALES

Comme on l'a souligné plus haut, c'est surtout pour la base de données du SNIIR-AM que l'utilisation des données individuelles est particulièrement difficile, du fait de la complexité des données fournies à l'utilisateur sous forme brute. C'est à lui de procéder à la synthèse de ces informations pour disposer des variables qui l'intéressent réellement, générant donc pour chaque extraction un travail long et fastidieux de gestion de données nécessitant des moyens humains spécialisés importants.

Dans l'état actuel des choses, si on met à part le cas de certains « gros » organismes dont l'usage récurrent des bases de données le justifie et qui disposent des ressources adéquates, la plupart des équipes potentiellement concernées n'ont pas les moyens de faire ce travail, ce qui constitue un des obstacles majeurs à une utilisation plus large des bases de données nationales pour la recherche et la surveillance. Il faut donc mettre à disposition des milieux concernés un support technique et scientifique adapté aux besoins. C'est un des objets de la proposition suivante.

3.5 SYNTHESE DES PROPOSITIONS : POUR LA CREATION D'UNE PLATEFORME D'INTERFACE ENTRE LES CHERCHEURS ET LES BASES DE DONNEES NATIONALES

La solution proposée pour améliorer la situation actuelle et résoudre au moins en partie la plupart des difficultés évoquées est la création d'une plateforme spécialisée, réunissant des moyens et des compétences adéquates, qui jouerait le rôle d'interface entre les chercheurs et les bases de données nationales. Cette plateforme aurait comme missions principales :

- le conseil aux utilisateurs concernant, en fonction de leurs besoins, le contenu et la structure des bases, les populations et périodes couvertes, la signification des variables, etc. Elle prendrait en charge l'historisation des évolutions, la mise en ligne de catalogues de données... ;
- la préparation des requêtes, permettant de transformer une demande de nature scientifique en une requête de type informatique ;
- la transmission des requêtes vers les bases de données et la récupération des fichiers extraits ;
- la restitution aux utilisateurs de données synthétisées, après sélection préalable des variables d'intérêt à partir des données brutes extraites du SNIIR-AM.

Il faut souligner que la création d'une plateforme centrale n'est évidemment pas exclusive d'un partenariat direct entre certaines équipes et les organismes gérant les bases, lorsque cela leur semble opportun.

Deux modèles organisationnels peuvent être envisagés :

- chaque organisme gestionnaire de base développe un « guichet » ouvert aux utilisateurs remplissant ces fonctions ;
- création d'une plateforme centrale, qui pourrait être celle qui est envisagée ci-dessus sous forme d'un centre d'appariement sécurisé pour gérer l'utilisation du NIR (*cf*. 3.2.3.1).

La première solution présente l'avantage pour le chercheur d'être en contact direct avec l'organisme, ce qui permet une bonne réactivité. Étant producteur des données, l'organisme est le mieux placé pour aider à leur utilisation. Cette solution a également l'avantage de laisser toute autonomie à chaque organisme gestionnaire de bases de données et n'implique pas la mise en place d'une structure inter-organismes. Cette solution nécessite néanmoins une bonne coordination entre les organismes pour proposer une offre de service homogène aux utilisateurs et nécessite de la part des demandeurs une connaissance *a priori* des possibilités offertes par les différentes bases de données disponibles.

La seconde solution offre un « guichet unique » pour des chercheurs souvent peu informés sur les bases disponibles et leurs spécificités et leur évite des démarches multiples ; elle garantit *a priori* une réponse homogène aux demandeurs ; elle permet de mutualiser les compétences et peut étendre rapidement ou à moyen terme ses prestations à diverses bases de données autres que celles du SNIIR-AM et de la Cnav ; elle permet un meilleur contrôle de l'ensemble des demandes et une meilleure connaissance des besoins du monde de la recherche et de la surveillance. Elle a l'inconvénient de nécessiter la création d'une structure impliquant plusieurs organismes, donc une certaine lourdeur institutionnelle et des moyens budgétaires.

3.6 PROPOSITIONS DIVERSES

3.6.1 Le consentement des personnes

Devant la multiplication et l'intrication de systèmes d'information distincts, l'ouverture de plus en plus large de ces systèmes à des utilisateurs toujours plus nombreux et l'impossibilité pratique de recueillir un consentement explicite individuel des personnes avant chaque étude ou appariement systématique, et dans la mesure où l'ensemble de la population est concerné, il conviendrait certainement d'envisager, comme le recommande la Conférence nationale de santé, « *d'unifier le régime de consentement à la collecte, au traitement, à l'échange et à l'hébergement des données de façon à ce qu'il soit aisément compréhensible par les usagers et commode à exprimer. […] La Conférence nationale de santé estime judicieux que soit menée une campagne publique articulant l'intérêt de la collecte des données de santé et le régime de protection en vigueur. La question du consentement pourrait être utilement mise en exergue pour faire progresser l'acceptabilité de l'informatisation des données de santé. Au-delà de la seule question du consentement, une telle campagne pourrait présenter les droits des*

usagers du système de santé : droit d'accès aux traces des échanges, droit de rectification, droit au masquage, voies de recours. »[44].

On ne peut que soutenir ces propositions, en n'omettant pas de l'étendre à tous types de données susceptibles d'être utilisées dans les études de santé.

3.6.2 Identification indirecte dans le SNIIR-AM

Le SNIIR-AM est une base anonymisée et les dispositions prises par la Cnam-TS en accord avec la Cnil pour l'accès à l'EGB (limites dans le croisement de certaines données sensibles, pas d'affichage des résultats d'une requête lorsque le nombre de bénéficiaires est inférieur à 10…) sont efficaces pour éviter l'identification des personnes.

Le développement du SNIIR-AM, qui contient des données individuelles de plus en plus nombreuses et diversifiées, les projets faisant appel à des extractions répétées sur les mêmes personnes, l'ouverture souhaitée à de nouvelles catégories d'utilisateurs peuvent cependant modifier cette situation. Même si les tentatives d'identification indirecte des personnes par croisement de données restent largement théoriques, il est indispensable de développer des mesures de précaution renforcées, faisant notamment appel à des techniques diverses de masquage de données sensibles.

Les problèmes de confidentialité posés par l'analyse de données individuelles à une échelle territoriale fine rendant possible l'identification des personnes, n'ont pas de solutions simples. On peut imaginer que la Cnil peut autoriser de telles analyses au cas par cas, au vu de demandes suffisamment justifiées.

3.6.3 La propriété des fichiers appariés

Les fichiers constitués à partir de l'appariement de données provenant de bases gérées par des organismes différents posent des problèmes de propriété. Bien entendu, les organismes concernés doivent se concerter et décider au cas par cas des conditions de leur coopération, des modalités d'utilisation des fichiers appariés, des possibilités de cession de données, de confidentialité, etc.

Il semble que le principe que chaque organisme contributeur de données doit avoir le moyen de s'opposer à des utilisations des données qui ne lui conviennent pas devrait être respecté en tout état de cause.

3.6.4 La politique tarifaire

La fourniture de données a un coût, même s'il est marginal par rapport à celui de la collecte et la gestion des bases de données. Actuellement, il n'existe pas de politique tarifaire homogène inter-organismes. Certains font payer la fourniture de données (CépiDc, Cnav, par exemple), et d'autres fournissent gracieusement les données, comme la Cnam-TS. Il semble cependant que le fait de faire payer la fourniture de données présente des avantages : contribution au budget des organismes et surtout sélection des demandes « sérieuses » du fait de l'effort demandé.

44 Conférence nationale de santé. Avis sur les données de santé informatisées adopté par l'assemblée plénière le 19 octobre 2010.

Il revient certes à chaque organisme gestionnaire de bases de données de définir sa propre politique en matière tarifaire. Cependant, pour les demandeurs relevant d'un organisme public de recherche ou de surveillance, il est indispensable que les tarifs pratiqués ne soient pas incompatibles avec les budgets que les équipes publiques demandeuses sont susceptibles d'obtenir de façon réaliste pour leurs travaux. Ceci signifie que les organismes gestionnaires de bases de données subventionnent de fait actuellement l'accès à leurs données, car leur coût véritable les rendrait inaccessibles aux équipes publiques. Il semble par contre légitime que si des demandes provenant de structures à but lucratif sont prises en compte, les tarifs pratiqués soient établis de façon correspondant au moins au coût véritable des données.

Dans cet esprit, la plateforme, qui jouerait le rôle d'interface entre les chercheurs et les bases de données nationales dont la création est proposée ci-dessus, doit disposer d'un financement provenant de deux sources : subventions de fonctionnement récurrentes de la part d'organismes publics (utilisateurs et fournisseurs de données, pouvoirs publics) ; rémunération des services en veillant à ce que le coût des prestations de la plateforme ne soit pas un obstacle pour les équipes publiques qui pourraient en bénéficier.

3.6.5 La localisation spatiale des personnes

On a insisté à plusieurs reprises sur les limites imposées pour divers usages par l'imprécision de la domiciliation des personnes dans les bases de données nationales, la résolution spatiale la plus fine enregistrée étant la commune. Pour remédier à ce problème, on peut recommander que les organismes collecteurs de données de premier niveau (l'hôpital, la CPAM, etc.), ou les Centres de Traitement Informatique (CTI) qui centralisent à l'échelle interrégionale l'ensemble des données à intégrer dans le SNIIR-AM, mettent en place une procédure interne automatisée par laquelle l'adresse des personnes serait géocodée, convertie en code Iris et transférée sous cette forme dans les bases nationales ; ce code pourrait être alors accessible et utilisable par les chercheurs sans rompre l'anonymat, à l'instar des autres données à caractère personnel contenues dans ces bases. Une étape supplémentaire serait que le géocodage des adresses au niveau de l'organisme de collecte soit réalisé au niveau le plus précis (coordonnées X/Y), ce qui autorise ultérieurement des agrégations à des niveaux plus ou moins agrégés en fonction des besoins.

Liste des sigles

ABM	Agence de la biomédecine
Afssaps	Agence française de sécurité sanitaire des produits de santé
ALD	Affection de longue durée
AMM	Autorisation de mise sur le marché
ARS	Agence régionale de santé
AT	Accident du travail
Cas	Centre d'appariement sécurisé
CASD	Centre d'accès sécurisé distant
CCAM	Classification commune des actes médicaux
CCTIRS	Comité consultatif sur le traitement de l'information en matière de recherche dans le domaine de la santé
Centi	Centre national de traitement informatique de la Cnam-TS
CépiDc	Centre d'épidémiologie des causes de décès
CIM 10	Classification internationale des maladies et causes de décès (dixième révision)
CMUC	Couverture maladie universelle complémentaire
Cnaf	Caisse nationale d'allocations familiales
Cnam-TS	Caisse nationale de l'assurance maladie des travailleurs salariés
Cnav	Caisse nationale d'assurance vieillesse
Cnil	Commission nationale de l'informatique et des libertés
Cnis	Conseil national de l'information statistique
CNSA	Caisse nationale de solidarité pour l'autonomie
CPAM	Caisse primaire d'assurance maladie
CRFS	Fiche standardisée de compte-rendu anatomopathologique
DADS	Déclaration annuelle des données sociales
DCC	Dossier communicant de cancérologie
DCIR	Données de consommation inter-régimes
DGS	Direction générale de la santé
DMP	Dossier médical personnel
DNT	Donnée nominative trimestrielle
Drees	Direction de la recherche, des études, de l'évaluation et des statistiques
EDP	Échantillon démographique permanent
EFS	Établissement français du sang
EGB	Échantillon généraliste des bénéficiaires
ESPS	Enquête santé et protection sociale
Fnors	Fédération nationale des observatoires régionaux de santé
FOIN	Fonction d'occultation des informations nominatives
Genes	Groupe des écoles nationales d'économie et de statistique
GHM	Groupe homogène de malades
GHS	Groupe homogène de séjour

Gip	Groupement d'intérêt public
HAS	Haute Autorité de santé
HSM	Enquête handicap-santé
HCSP	Haut Conseil de la santé publique
IDS	Institut des données de santé
INS	Identifiant national de santé
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
LPP	Liste des produits et prestations
MP	Maladie professionnelle
MSA	Mutualité sociale agricole
NIR	Numéro d'inscription au répertoire
PMSI	Programme de médicalisation des systèmes d'information
Rein	Réseau épidémiologie et information en néphrologie
RNIAM	Répertoire national d'inter-régimes des bénéficiaires de l'assurance maladie
RNIPP	Répertoire national d'identification des personnes physiques
RPU	Résumé de passage aux urgences
RSI	Régime social des indépendants
SipaPH	Système d'information partagé pour l'autonomie des personnes handicapées
SLM	Section locale mutualiste
SNGC	Système national de gestion des carrières
SNGD	Système national de gestion des dossiers de retraites en cours d'instruction ou de paiement
SNGI	Système national de gestion des identités
SNIIR-AM	Système national d'information inter-régimes de l'assurance maladie

Table des matières

Figure

Haut Conseil de la Santé Publique

Pour une **meilleure utilisation** des **bases de données nationales** pour la **santé publique** et la **recherche**

La France dispose de bases de données médico-sociales et économiques nationales centralisées, constituées et gérées par des organismes publics, couvrant de façon exhaustive et permanente l'ensemble de la population dans divers domaines stratégiques pour la santé publique et la recherche : recours aux soins, hospitalisation, handicaps, prestations et situation professionnelle, sociale et économique. De plus, un identifiant individuel unique (le NIR : Numéro d'identification au répertoire) est actuellement utilisé par pratiquement toutes les bases de données nationales. Malgré certaines limites en termes de couverture, de qualité et de validité des données, ces bases de données, concernant plus de 60 millions de personnes, constituent un patrimoine considérable, vraisemblablement sans équivalent au monde. Cependant, l'utilisation à des fins de recherche et de surveillance de ces bases de données nationales se heurte actuellement à des obstacles divers, dont les plus importants sont de nature juridique et opérationnelle.

En réponse à une saisine de la DGS ce document fait le point sur les principales bases existantes, expose les difficultés pour leur utilisation et présente de façon détaillée une série de propositions pour les surmonter.