

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(**)*

() Insee, Département des Études Économiques d'Ensemble*

*(**) Direction de la Recherche, des Études, de l'Évaluation et des Statistiques*

*(***) Ministère de l'Immigration et de l'Intégration*

Introduction

Depuis plusieurs années, l'Insee s'est engagé dans un mouvement croissant de recours aux sources administratives. Dans cette perspective, il a été décidé de compléter le dispositif des enquêtes Handicap-Santé 2008 et 2009 en les appariant pour la première fois avec les données de remboursements de l'assurance-maladie.

Dès 2006, le Cnis souligne la complémentarité entre sources administratives et données d'enquêtes [1], mais si les sources administratives présentent l'incomparable avantage d'être déjà disponibles, les utiliser peut poser d'importantes difficultés. Il faut d'abord s'assurer que les données administratives mesurent vraiment les phénomènes que l'on cherche à étudier. De plus, le cadre juridique pour rapprocher des fichiers est très strict, et, in fine, intégrer ces données dans les bases statistiques exige souvent de coûteux retraitements.

Dans le cas de l'appariement de l'enquête Handicap-santé avec les données de l'assurance-maladie, la première difficulté est facilement levée car si les données de l'assurance-maladie ne mesurent que les soins remboursés, et non les soins réellement consommés, ces définitions coïncident la plupart du temps, et les données de remboursements présentent l'avantage d'être beaucoup plus détaillées et fiables que les données d'enquêtes. C'est l'objet de la première partie de l'article.

La deuxième partie vise à décrire le processus de l'appariement, qui s'est révélé truffé d'embûches. Tout d'abord, le décret en Conseil d'État autorisant les enquêteurs à collecter le Numéro d'Identification au Répertoire (Nir) des personnes, nécessaire à l'appariement, a été publié trop tard, à la toute fin de la collecte. De plus, le rapprochement de fichiers est soumis à des contraintes de sécurité très strictes et met en jeu de nombreux acteurs avec des priorités et des calendriers différents. Tout ceci n'a pas été sans conséquences sur la qualité des données puisque sur 29 000 enquêtés dans le volet « ménage », seuls 20 700 sont appariés *in fine* avec les données de l'assurance-maladie.

Enfin, une fois les fichiers rapprochés, les traitements statistiques sont nombreux avant de pouvoir livrer une base de données exploitable par les statisticiens. En particulier, l'échantillon apparié doit rester représentatif, et un important travail de repondération est nécessaire pour compenser les échecs d'appariements. La troisième partie de l'article décrit ces traitements statistiques.

1. Les enquêtes Handicap-Santé et les données de l'assurance-maladie se complètent

1.1. Un dispositif d'enquêtes Handicap-Santé riche et complet

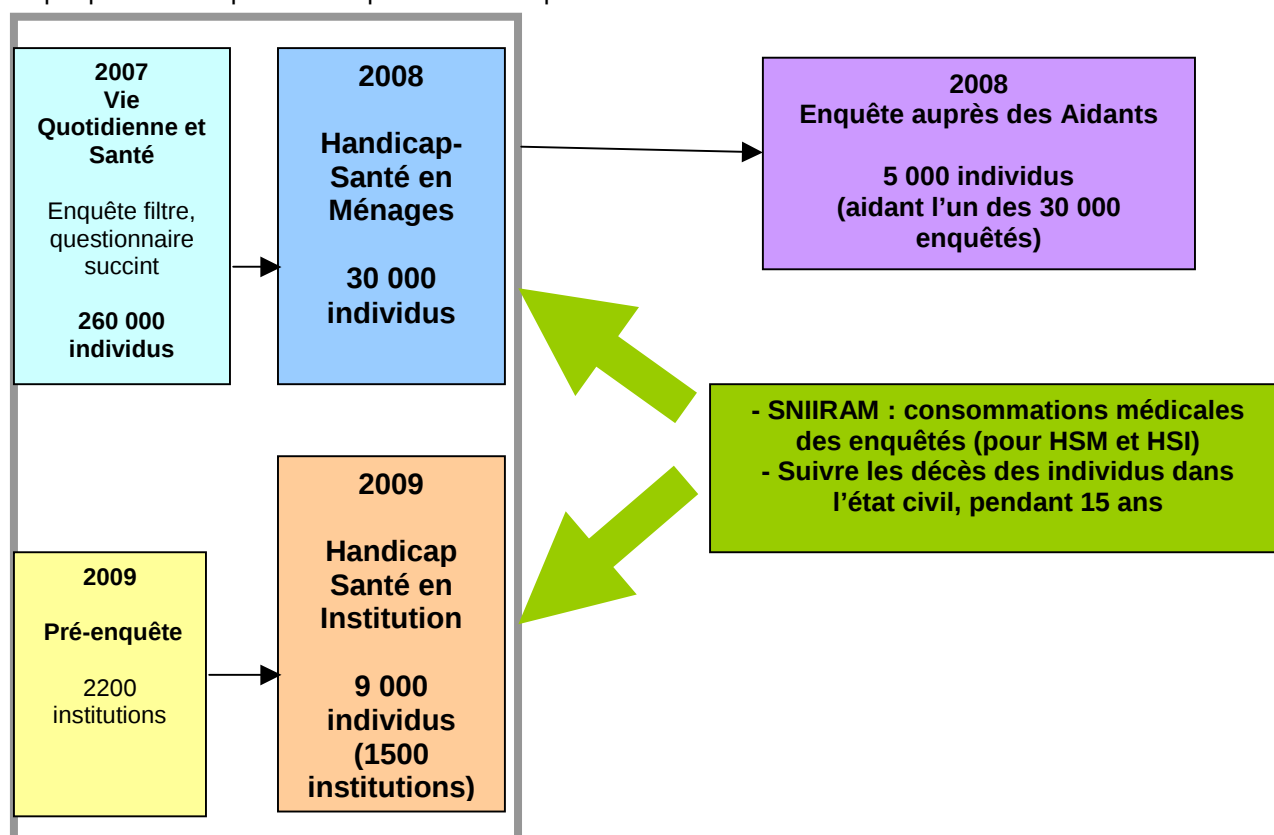
L'enquête Handicap Santé 2008-2009 (HS) est un dispositif complexe (Bouvier, 2011, [2]). L'échantillon de l'enquête auprès des ménages ordinaires (Handicap-Santé en Ménages, HSM, 2008) est sélectionné à partir d'une enquête-filtre, l'enquête Vie Quotidienne et Santé 2007, adossée au recensement. Cette dernière permet de repérer et de sur-échantillonner les personnes en situation de handicap.

Pour couvrir tout le champ de la population générale, on enquête également les personnes hébergées en institutions¹ (enquête Handicap-Santé en institution, HSI, 2009) : établissements pour personnes âgées dépendantes, unités de soins de longue durée, établissements pour personnes handicapées, établissements psychiatriques et centres d'hébergement et de réinsertion sociale. Ensemble, ces deux volets en ménage (HSM) et en institution (HSI) forment le cœur de l'enquête.

Enfin, on interroge aussi 5 000 individus qui ont été déclarés « aidants informels », *i.e.* ils apportent une aide à une personne enquêtée dans le volet ménage, mais sans être des professionnels de l'aide : c'est l'enquête auprès des aidants (Handicap-santé aidants, HSA, 2008).

Pour compléter ce dispositif, les données de l'enquête sont appariées avec le Système National d'Information Inter-Régime de l'Assurance Maladie (Sniir-am). Cette source administrative (Lenormand, 2005, [3]) permet de connaître l'ensemble des remboursements effectués par l'assurance-maladie, et par là, les soins de santé consommés par les personnes. L'enquête est aussi appariée avec les fichiers d'état civil, afin de suivre la mortalité des enquêtés pendant 15 ans. L'articulation est résumée dans le schéma ci-dessous.

Graphique 1 : le dispositif d'enquêtes Handicap-Santé



¹ Personnes séjournant en institution au moins trois semaines.

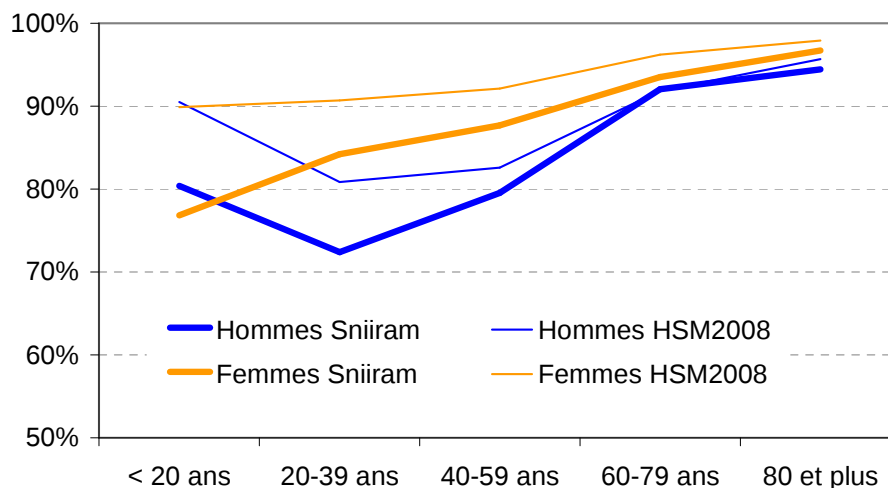
1.2. Le Sniir-am : plus détaillé et plus objectif que les données d'enquêtes

Apparier l'enquête avec les données de l'assurance-maladie, cela présente plusieurs intérêts :

1) **Économiser 25 minutes de questionnaire** sur le recours aux soins et alléger ainsi la charge de réponse pour les enquêtés. De plus, le Sniir-am fournit un niveau de détail bien supérieur à celui d'une enquête et permet par exemple de reconstituer des parcours de soins : en permettant d'identifier les médicaments prescrits par tel spécialiste, qui lui-même avait été consulté suite à la recommandation de tel généraliste, etc.

2) Dans une enquête par entretien, les **biais de mémoire** sont importants [4], car les soins consommés au cours des 12 derniers mois sont traditionnellement demandés. Graphique 2 et Tableau 1.

Graphique 2 : Proportion d'hommes et de femmes qui ont eu recours à un médecin généraliste au cours des 12 derniers mois, déclaré dans l'enquête HSM, ou selon les données administratives Sniir-am



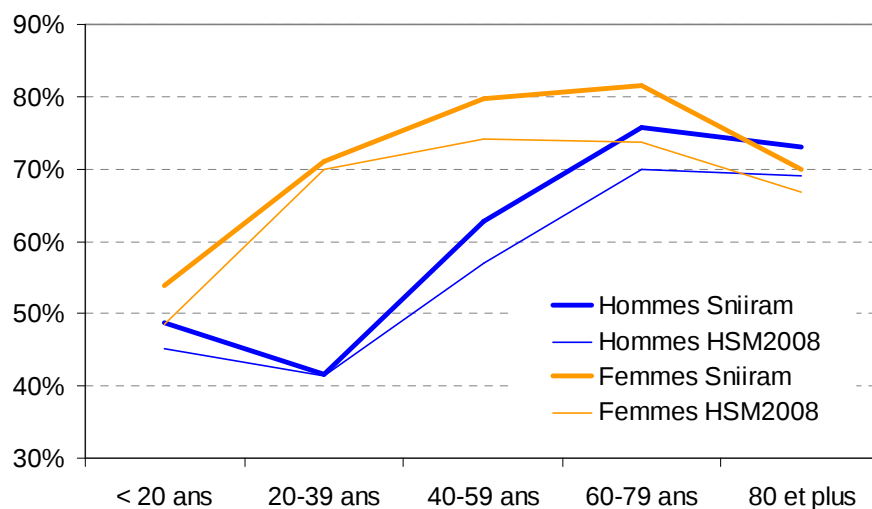
Lecture : Selon les données de l'assurance-maladie, 80% des hommes de moins de 20 ans ont consulté au moins un médecin généraliste au cours des 12 derniers mois, alors qu'ils sont 90% à déclarer avoir consulté dans l'enquête.

Source : Handicap-santé 2008, Insee.

Champ : France métropolitaine et Dom.

Par ailleurs, le questionnaire peut aussi être **mal compris** par certains : la question « Au cours des 12 derniers mois, avez-vous vu au moins une fois un médecin spécialiste pour vous-même ? » suppose de l'interviewé une parfaite connaissance des diverses spécialités, ce qui n'est pas nécessairement le cas (l'ophtalmologiste est un spécialiste, mais pas le dentiste, ni le kiné). Graphique 3. De plus ces questions peuvent servir de filtre à d'autres questions qui seront donc d'emblée entachées de ces biais éventuels.

Graphique 3 : Proportion de personnes qui ont eu recours à un médecin spécialiste au cours des 12 derniers mois, déclaré dans l'enquête HSM, ou selon les données administratives Sniir-am



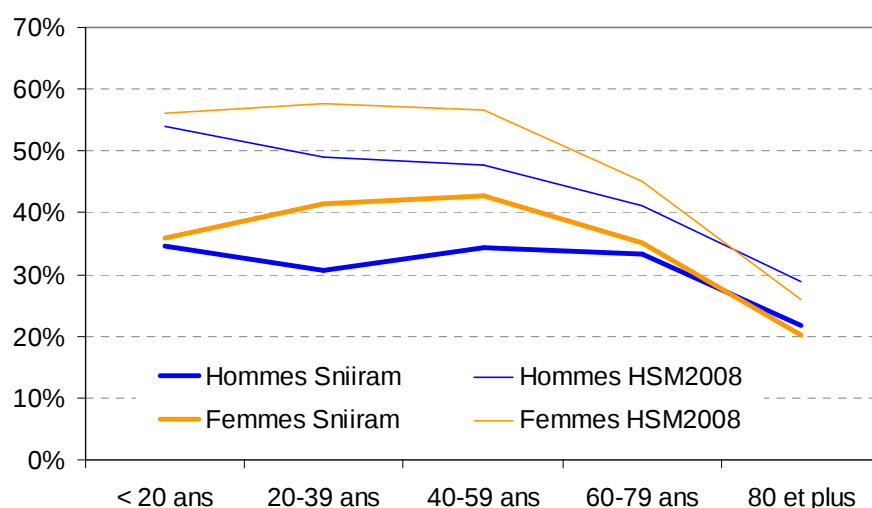
Lecture : Selon les données de l'assurance-maladie, 55% des femmes de moins de 20 ans ont consulté au moins un médecin spécialiste au cours des 12 derniers mois, alors qu'elles sont 50% à déclarer avoir consulté dans l'enquête.

Source : Handicap-santé 2008, Insee.

Champ : France métropolitaine et Dom.

3) Enfin, le **biais de « désirabilité sociale »** désigne le fait de chercher à se présenter à l'enquêteur sous son meilleur jour, et de donner la réponse que l'on pense être attendue par la société. Comme il est par exemple recommandé de se rendre au moins une fois par an chez le dentiste², ce biais pourrait expliquer une partie de l'écart entre le recours déclaré et le recours relevé à l'Assurance Maladie, plus faible. Graphique 4. Ce mécanisme peut être inconscient.

Graphique 4 : Proportion de personnes qui ont eu recours à un dentiste au cours des 12 derniers mois, déclaré dans l'enquête HSM, ou selon les données administratives Sniir-am



Lecture : Selon les données de l'Assurance Maladie, 40% des femmes de moins de 20 à 39 ans ont consulté un dentiste au cours des 12 derniers mois, alors qu'elles sont 60% à déclarer l'avoir fait dans l'enquête.

Source : Handicap-santé 2008, Insee.

Champ : France métropolitaine et Dom.

² <http://www.ameli-sante.fr/carie-dentaire/consulter-regulierement-son-chirurgien-dentiste.html>

Tableau 1 : Taux de recours selon la source de donnée (déclarée ou administrative) et cohérence des deux sources

	Proportion de personnes qui ont eu recours		
	Selon les données déclaratives de l'enquête HSM	Selon les données administratives du Sniir-am	Conjointement dans les deux sources
Généraliste	89%	83%	80%
Spécialiste	55%	63%	46%
Infirmier	15%	21%	9%
Hospitalisation	12%	12%	7%
Dentiste	54%	39%	35%
Biologie	52%	49%	37%
Kinésithérapeute	17%	14%	11%

Lecture : 17% des personnes ont consulté un kiné au cours des 12 derniers mois selon les données de l'assurance-maladie et 14% déclarent l'avoir fait selon l'enquête, mais 11% ont consulté à la fois selon l'enquête et les données administratives.

Source : Handicap-santé 2008, Insee.

Champ : France métropolitaine et Dom.

1.3. Mais les soins remboursables et non les soins consommés

Néanmoins, compte tenu des besoins des utilisateurs, le Sniir-am n'est pas sans limites. En particulier, il ne couvre pas l'ensemble des soins de santé consommés par les individus, mais uniquement la partie qui est remboursable.

- 1) Ainsi, il exclut les dépenses de santé qui ne sont pas prises en charge par l'assurance-maladie obligatoire (les psychologues, par exemple). Il exclut aussi les médicaments consommés sans prescription (automédication), lorsqu'ils sont achetés en pharmacie sans ordonnance ou bien lorsque les personnes consomment les médicaments qu'il leur reste d'une prescription antérieure. Il exclut aussi certaines consultations gratuites. Enfin, si un patient oublie de transférer sa feuille de soin à l'assurance maladie, l'information est perdue.
- 2) Il existe des trous dans le champ du Sniir-am : les données de certaines sections locales mutualistes ne remonteraient pas. Pour certains fonctionnaires et étudiants, et aussi quelques régimes spéciaux, les données ne seraient donc pas disponibles³.
- 3) Enfin, les deux sources n'ont pas exactement la même période d'observation. Les données du Sniir-am portent sur la période fixe du 1^{er} juillet 2007 au 30 juin 2008 pour tous les individus interrogés dans l'enquête handicap santé volet ménage, alors que les données déclarées portent sur les douze mois précédant l'interrogation, selon la date de visite de l'enquêteur, comprise d'avril à juillet 2008⁴. Ceci explique une partie des divergences entre données déclarées et données administratives.

Le champ du Sniiram ne coïncide donc pas parfaitement avec ce que l'on cherche à mesurer, mais la richesse et l'objectivité de cette source administrative compensent largement ces inconvénients.

2. L'appariement, un parcours semé d'embûches

³ Néanmoins, les individus affiliés au régime de sécurité sociale des fonctionnaires ou des étudiants ont un taux d'appariement dans la moyenne de l'ensemble de l'échantillon (Annexe 2).

⁴ Pour le volet institution de l'enquête, réalisé en 2009, le problème est identique : les données de l'assurance-maladie portent sur la période fixe du 1^{er} janvier au 31 décembre 2009, tandis que le questionnaire porte sur les douze mois précédant le jour de l'entretien.

Ce chapitre reprend en grande partie la note interne de Gérard Bouvier et de Pascale Breuil de l'unité des études démographiques et sociales de l'INSEE, du 23 mars 2011 [5]. Il présente les difficultés rencontrées pour réaliser l'appariement du volet ménage de l'enquête (HSM 2008) et quelques recommandations. Le volet institution de l'enquête (HSI 2009) a lui aussi été apparié, mais son cas, peu différent du volet ménage, n'est pas présenté ici.

2.1. Pour la première fois un appariement au niveau individuel, rendu possible depuis la nouvelle loi « Informatique et Libertés » de 2004

Dès l'expression des besoins de l'enquête, en 2006, il est précisé que l'« on souhaite pouvoir appairer les données individuelles de l'enquête aux données individuelles du fichier Sniir-am de la Caisse Nationale d'Assurance Maladie des Travailleurs Salariés (Cnam-TS) afin de connaître les consommations médicales des répondants à l'enquête sans avoir à passer un questionnaire lourd sur le sujet. Cette demande sera faite à la Cnil ».

Dans l'enquête décennale Santé 2003 déjà, un appariement avait été réalisé avec les données du Sniir-am, mais par catégories sociodémographiques agrégées, et non au niveau individuel, car la loi Informatique et Libertés ne l'autorisait pas à l'époque. C'est donc la première fois qu'une opération de telle envergure est lancée autour d'une enquête Santé⁵.

Pour rapprocher les fichiers de l'enquête et de l'assurance-maladie, il faut disposer du Numéro d'Identification au Répertoire (Nir) des personnes, qui ne peut être collecté qu'après avis de la Cnil et publication d'un décret en Conseil d'État. Au regard des objectifs des lois de 2004 sur la santé publique et de 2005 sur le handicap, la Cnil et le Conseil d'État autorisent l'appariement, au niveau individuel, des données d'enquête et des consommations de soins. En effet, la loi informatique et liberté a été modifiée en 2004 : « Sont autorisés par décret en Conseil d'État, pris après avis motivé et publié de la Commission nationale de l'informatique et des libertés, 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 national d'identification des personnes physiques. »

L'ensemble du dispositif d'enquête bénéficie du label d'intérêt général et de qualité statistique, assortie de l'obligation de réponse.

2.2. Une convention pour définir les rôles et les outils

L'Insee et la Cnam-TS rédigent conjointement et signent une convention permettant de réaliser cet appariement. Le Centre National Informatique d'Orléans (Insee - CNIO) est le tiers de confiance : la procédure « Fonction d'Occultation des Informations Nominatives » (FOIN) est utilisée afin de crypter les identifiants. La procédure d'échange des données est dite en double aveugle, de sorte qu'aucun des acteurs ne peut avoir en sa possession l'ensemble des fichiers, ce qui permettrait de rompre l'anonymat [6], [7].

Dans un premier temps, le schéma proposé par l'Insee pose problème à la Cnam-TS, car le tiers de confiance, le CNIO, est un organe de l'Insee. L'Insee indique que le schéma a été soumis à la Cnil, consciente de la proximité « institutionnelle », mais au vu de l'éloignement géographique des unités en charge des traitements, la Cnil n'a pas émis de réticences. La Cnam-TS indique donc qu'elle suivra l'avis de la Cnil, avis qui a été donné favorable lors de la séance du 10 janvier 2008.

2.3. Marche à suivre et rôle de chacun : la théorie

Les concepteurs proposent une démarche pour réaliser cet appariement et définissent les rôles de chacun.

⁵ La Cnam est co-financeur de l'enquête Handicap-Santé

- Les informations identifiantes sont demandées à la toute fin de l'enquête :
 - o Dans tous les cas, on recueille le Nir et les éléments d'état civil (nom, prénom, sexe, date et lieu de naissance) de l'enquêté
 - o Si l'enquêté n'a jamais travaillé, ce qui concerne de nombreuses personnes de l'enquête (inactifs n'ayant jamais travaillé, enfants, personnes handicapées) il bénéficie de l'assurance maladie via les droits d'un proche (parent, conjoint, ...). L'enquêté est appelé « l'ayant droit » et le proche est « l'ouvrant-droit ». Dans ce cas, on recueille aussi le Nir et les éléments d'état civil de l'ouvrant droit. Sans cette information, l'appariement est impossible pour les ayant droit.
- Le Responsable Informatique de l'Enquête (RIE) prépare un fichier Nir-états civils, et l'envoie au Centre National Informatique de Nantes (CNIN). Ce dernier certifie les Nir en les retrouvant dans la base-répertoire des personnes physiques et livre quelques éléments de qualité (nombre de Nir directement identifiés, identifiés après corrections, en échec).
- Sous contrôle du RIE et du CNIO, le CNIN crypte les données et les transmet à la Cnam
- La Cnam extrait, pour ces individus, les consommations de soins enregistrées pendant la période du 1^{er} juillet 2007 au 30 juin 2008, et remplace le Nir par l'identifiant anonyme de l'enquête.
- Sous contrôle du RIE, la Cnam-TS retourne ces bases de données au CNIO. Le CNIO réalise l'appariement et met à disposition de la maîtrise d'ouvrage les données rapprochées.
- À partir des données administratives, la Drees est chargée des retraitements statistiques, afin de constituer un fichier à usage des utilisateurs statisticiens, notamment en construisant des variables synthétiques faciles à manier, et une pondération.

2.4. Le décret autorisant la collecte du Nir est publié trop tard

Dans le cadre d'un appariement, les enquêtés doivent être prévenus que leurs réponses seront rapprochées des données d'autres fichiers. Les enquêteurs doivent leur demander s'ils acceptent de communiquer leur Nir. Même si la Cnil émet un avis favorable le 10 janvier 2008 dans le cas de l'enquête Handicap-Santé, la collecte démarre, en avril 2008, alors que le décret en Conseil d'État n'est toujours pas officiellement publié et la collecte du Nir est donc illégale. Les concepteurs doivent adapter leur stratégie dans l'urgence.

Quelques jours après le début de la collecte, on décide donc de ne plus demander le Nir, et de ne recueillir que l'état civil. Les données d'état civil de l'enquêté sont déjà présentes dans le questionnaire, dispersées en divers endroits. Le questionnaire CAPI doit aussi permettre de recueillir l'état civil d'un éventuel ouvrant-droit. Les nouvelles consignes sont transmises aux enquêteurs. Comme la collecte a déjà débuté, le questionnaire CAPI est adapté sans être testé. Le décret ne sera finalement publié que début juillet, tout à la fin de la collecte.

2.5. L'état civil à défaut du Nir : une réorientation de dernière minute qui pèse sur la qualité

Une fois la collecte totalement terminée, fin 2008, 23 300 états civils ont été collectés et 19 200 Nir sont reconstitués et certifiés par le pôle de Nantes. Le nombre d'individus correspondants à ces Nir, et pour lesquels on peut donc espérer un appariement, est forcément inférieur à 19 200, puisque à de nombreux individus sont associés deux Nir (leur propre Nir et celui de leur ouvrant-droit). Ces résultats, présentés en janvier 2009 en comité de pilotage, sont décevants⁶. Cette étape aurait pu être améliorée si l'on avait programmé une reprise manuelle des échecs, mais dans l'urgence et par inexpérience, rien n'a été prévu.

2.6. Le projet avance peu jusqu'au printemps 2010.

La signature de la convention entre l'Insee et la Cnam-TS est un processus long qui aboutit en mai 2009. Étant la plus intéressée par l'exploitation de l'appariement, la Drees est chargée de piloter le dispositif. Elle est cependant dans une position difficile pour mobiliser les différents acteurs (Cnam-TS

⁶ Compte tenu de l'appariement définitif, on estime à environ 17 000 le nombre d'individus qui auraient pu être appariés sur les 29 000 enquêtés au départ, si l'on s'était contenté des Nir retrouvés à cette étape.

et informaticiens Insee de Lille, du CNIO et du CNIN) et assurer leur coordination, notamment parce que la convention a été établie entre la Cnam-TS et l'Insee et que la Drees n'a autorité ni sur l'un ni sur l'autre de ces deux organismes. Par ailleurs, la Cnam-TS dispose de peu de temps à consacrer à l'opération : le projet avance peu pendant une année et ne sera relancé qu'en 2010.

2.7. L'Insee et la Drees mettent à profit ce délai pour améliorer la reconstitution des Nir

L'Insee et la Drees décident de mettre à profit ce délai pour améliorer la reconstitution des Nir. Plusieurs pistes d'améliorations sont proposées, en particulier sur la commune de naissance.

- En raison de la réforme du 1^{er} janvier 1968, les codes des communes de la périphérie parisienne ont changé. Par exemple, Neuilly-sur-Seine est dans le département 92 depuis 1968, mais était dans le département 75, « Seine », avant 1968. Ainsi, le Nir des individus nés avant 1968 à Neuilly comprend donc le numéro 75 et non 92.
- Les noms de communes fréquents (Saint Martin, Saint Rémy, Chatillon) peuvent être dans plusieurs départements. On décide alors de multiplier les individus en déclinant les départements : Saint Martin 97, Saint Martin 66, Saint Martin 83, en espérant que l'une de ces lignes sera certifiée.
- On décide d'envoyer aussi en recherche les Nir des personnes dont on ne sait pas si elles sont leur propre ouvrant-droit. En effet, si ces individus sont retrouvés dans les fichiers de la Cnam-TS, c'est qu'ils étaient bien leur propre ouvrant-droit.
- Les questionnaires de l'enquête en direction régionale sont sollicités : en revenant aux fiches-adresses papier, on récupère les noms patronymiques de 500 femmes pour lesquelles on ne disposait que du nom d'épouse.

2.8. Un échantillon final apparié et représentatif de 20 700 individus

Ces travaux de reprise portent leurs fruits, puisque l'on passe de 23 300 à 28 100 états civils envoyés à Nantes. *In fine*, on dispose dorénavant de 22 400 Nirs reconstitués et certifiés, contre 19 200 initialement.

Les contacts avec la Cnam-TS sont repris au printemps 2010 et on instaure un suivi régulier entre l'Insee, la Drees, la Cnam-TS et le CNIO pour faire avancer les travaux. Au cours de l'été 2010, la Cnam-TS parvient à libérer du temps à consacrer à l'opération et l'occasion est saisie pour réaliser l'appariement. Les phases de test et de cryptage sont alors réalisées très rapidement et la Cnam-TS met à disposition de la maîtrise d'ouvrage (Insee-Drees) les fichiers de données le 19 août 2010 pour le volet ménage⁷.

À partir des 22 400 Nir transmis, la Cnam-TS extrait des consommations de soins pour 19 250 individus. Comme 1 450 individus non retrouvés sont considérés comme n'ayant pas consommé de soins (voir plus loin, paragraphes 3.2 et 3.3), on dispose à l'issue de l'ensemble du processus d'un fichier de données apparié et représentatif comportant 20 700 personnes, soit un taux de réussite global de 70% par rapport aux 29 000 enquêtés de départ.

2.9. Des contraintes de sécurité scrupuleusement respectées

Le dossier Cnil a été instruit par la maîtrise d'ouvrage Insee en collaboration avec les experts juridiques de l'institut (division environnement juridique). Les fichiers de données sont cloisonnés et l'accès y est réglementé. Le dossier Cnil mentionne ces règles pour les fichiers finaux, mais elles sont aussi appliquées aux fichiers en cours de construction. Les contraintes sont rigoureusement

⁷ L'appariement du volet institution est réalisé dans la foulée de celui du volet ménage.

respectées, mais elles sont peut-être un peu trop strictes, si bien qu'aucun des acteurs ne possède une vue détaillée du processus.

Le principal problème est que les opérations se succèdent en aveugle pour la maîtrise d'ouvrage : en particulier, on ne peut pas distinguer les Nir reconstitués par le pôle de Nantes et ceux qui finissent en échec.

À la fin du processus, parmi les individus sans consommation de soins, on est donc incapable de distinguer les individus en échec de ceux qu'on n'a pas retrouvés dans les fichiers de l'Assurance Maladie car n'ayant pas consommé de soins sur la période de l'enquête (non-consommant).

La pondération qui doit permettre de conserver un échantillon apparié représentatif est donc beaucoup plus difficile à construire en l'absence de ces informations essentielles. Le souci de satisfaire les exigences de la Cnil en matière de protection des enquêtés est bien sûr légitime mais ces contraintes peuvent pénaliser la qualité des données.

2.10. Recommandations

Au final, l'Insee et la Drees tirent les sept recommandations suivantes de l'expérience HSM-Sniiram 2008 :

Recommandation 1 : Anticiper des délais suffisants pour l'autorisation réglementaire.

Le dossier de l'enquête (HS) est d'abord validé par le Cnis puis présenté à la Cnil. Enfin, le Conseil d'état doit *in fine* approuver l'utilisation du Nir. En pratique, il faut prévoir plus d'un an pour réaliser ce cheminement.

À titre d'exemple, le processus d'appariement pour le volet de l'enquête en institution a largement repris celui du premier volet, ce qui devait alléger les temps de constitution et d'examen. Il a pourtant fallu encore près d'un an entre le passage au label (Cnis) et la parution du décret, notamment parce que des personnes (à la Cnil puis au Conseil d'État) avaient changé entre-temps.

Recommandation 2 : Prévoir un recueil alternatif de l'état civil.

Même s'il est prévu de collecter le Nir des enquêtés, il faut concevoir un module autonome de recueil des états civils en clair. Ce module pourra ainsi être testé correctement, et les enquêteurs pourront être formés au protocole, en cas d'imprévu. Prévoir un questionnement précis, notamment pour disposer du nom patronymique et du nom d'usage des femmes.

Recommandation 3 : Ne pas anticiper *a priori* des réticences des enquêtés à communiquer leurs éléments d'état civil ou leur Nir.

Contrairement à ce que l'on pourrait peut-être croire, les enquêtés acceptent bien, de façon générale, de communiquer leur Nir ou des éléments d'état civil, dès lors que le thème de l'enquête le justifie, comme nous l'apprennent les bilans de l'enquête HSM. Quoique très identifiant, le Nir semble toutefois peu explicite, certains enquêtés ressentent même parfois moins de gêne à le donner que des éléments d'état civil plus explicites. Les formations aux enquêteurs doivent insister sur l'importance pour l'enquête de récupérer ces informations nécessaires à l'appariement.

Recommandation 4 : Limiter autant que possible le nombre des acteurs impliqués.

De plus, l'unité qui pilote le projet doit disposer de suffisamment d'autorité auprès des acteurs afin de les coordonner correctement.

Recommandation 5 : Permettre à un des acteurs de disposer d'un accès suffisant aux données.

Il convient de distinguer les bases de travail et les bases finales. Si le dossier Cnil mentionne un accès restreint aux données identifiantes dans les bases finales, il doit pouvoir permettre à un des acteurs d'avoir accès, si besoin pour une durée limitée, à suffisamment de données pour assurer un appariement efficace et fournir des éléments sur la qualité.

Recommandation 6 : Le parcours du fichier (CNIO-CNIN-DG Insee-Cnam) doit être défini de façon précise.

Ainsi, chacun peut faire part de ses contraintes suffisamment en amont de la réalisation des travaux. Le processus doit être testé dans son ensemble.

Recommandation 7 : Prévoir une reprise manuelle pour les états civils en échec.

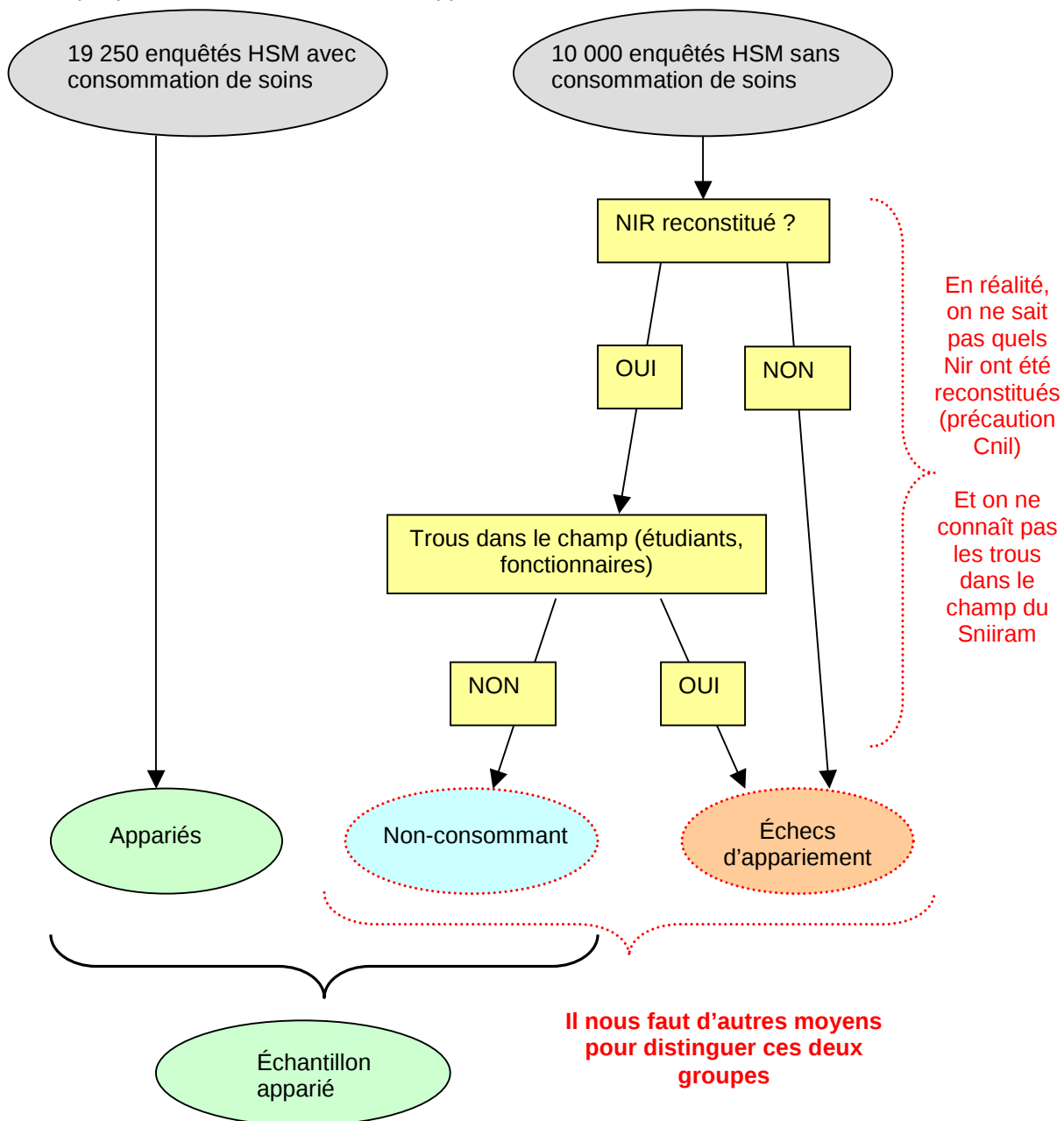
En combinant des procédures automatisées et une reprise manuelle des cas en échec, on peut améliorer sensiblement le nombre de Nir reconstitués.

3. Un échantillon apparié et représentatif

Sur les 29 000 individus enquêtés, la Cnam retourne *in fine* des consommations de soins pour 19 250 d'entre eux. Il reste donc presque 10 000 individus pour lesquels aucune information n'est disponible. La suite de l'article présente la façon dont on a traité ces 10 000 individus afin que l'échantillon apparié reste représentatif.

3.1. Non-consommant ou échec d'appariement ?

Graphique 5 : constituer l'échantillon apparié



Sur les 10 000 individus pour lesquels nous n'avons pas de consommation de soin, deux cas de figure sont possibles.

- Certains individus n'ont tout simplement pas consommé de soins remboursables sur la période de l'enquête. Ainsi, il est tout à fait normal que la Cnam ne les retrouve pas dans ses fichiers. Graphique 5. On parle d'individus « **non-consommant** ». Leur consommation de soins est donc en réalité connue : elle est égale à 0.

- Les autres sont en échec dans le processus de l'appariement. Pour certains d'entre eux, on dispose du NIR de l'individu, mais ce dernier est hors du champ du Sniiram. Pour les autres, le NIR n'a tout simplement pas pu être reconstitué. Graphique 5. Dans tous les cas, la consommation de soins de ces individus est inconnue, et on parle d' « **échec d'appariement** ».

Les individus pour lesquels on dispose des consommations de soins et les non-consommant forment l'échantillon apparié et cet échantillon doit être repondéré afin de compenser les échecs. La principale difficulté est la distinction entre les non-consommant et les échecs d'appariement parmi les 10 000 absents du Sniir-am. On va s'appuyer sur les informations socio-économiques et sur le recours aux soins déclaré dans le questionnaire pour faire cette distinction.

3.2. D'après le questionnaire, 650 non-consommant déclarés

L'enquête Handicap Santé interroge les individus sur leur recours à quelques soins : généraliste, spécialiste, dentiste, etc. Ainsi, un individu qui déclare n'avoir consommé aucun soin dans le questionnaire est un bon candidat pour être un véritable non-consommant selon l'assurance maladie. Tableau 2.

Tableau 2 : définitions d'un non-consommant à l'assurance maladie et selon le questionnaire HSM

Définition d'un consommant selon	
L'Assurance Maladie	L'enquête Handicap Santé
Avoir au moins une prestation entre le 1 ^{er} juillet 2007 et le 30 juin 2008 pour un soin remboursable, à savoir :	Déclarer avoir consulté au moins une fois au cours des 12 derniers mois ⁸ un
généraliste	généraliste
spécialiste	spécialiste
dentiste	dentiste
sage femme	-
kiné	kiné
infirmière	infirmière
orthophoniste	orthophoniste
orthoptiste	-
pédicure	pédicure
biologie	laboratoire médical
pharmacie ⁹	-
Matériel médical (pansements, prothèses, ...)	-
transports et cures thermales	-

D'après le questionnaire, 650 individus (représentant 4,2% de la population après redressement) ne déclarent aucune consommation de soin au cours des douze derniers mois et seront considérés comme non-consommant. Toutefois, d'après l'Enquête Santé et Protection Sociale 2006 (ESPS 2006), les non-consommant représentent 8,2% de la population en ménage. Même si ces 650 individus sont probablement de vrai non-consommant, cette première méthode, un peu trop restrictive, peine à repérer l'ensemble des non-consommant.

Par ailleurs, d'après ESPS 2006, les non-consommant ont tendance à se concentrer dans certaines tranches de population : ce sont souvent des hommes, des jeunes, des personnes qui ont un faible niveau de vie ou sans couverture complémentaire de santé (mutuelle, ...). Ces informations socio-économiques peuvent nous permettre d'améliorer notre méthodologie.

⁸ Le volet ménage a été collecté entre le 31 mars et le 19 juillet 2008.

⁹ Pour qu'un médicament ou matériel médical soit remboursé, il faut qu'il soit prescrit par un professionnel de santé. Ainsi, même si on ne dispose pas de données déclarées sur les médicaments, les personnes qui ont consommé sont tout de même repérées par le questionnaire, via le fait qu'elles ont consulté un professionnel de santé au préalable.

3.3. Un modèle logistique pour compléter la recherche des non-consommant

Dans ESPS 2006, les non-consommant sont connus et on retrouve un certain nombre de questions identiques à celles posées dans HSM 2008 : outre les variables socio-économiques (sexe, âge, catégorie socioprofessionnelle, couverture santé complémentaire, nombre d'adultes et nombre d'enfants du ménage), les enquêtés déclarent aussi s'ils ont eu recours à un médecin généraliste et à un médecin spécialiste au cours des douze derniers mois¹⁰.

Dans ESPS, on modélise par une régression logistique la probabilité d'être non-consommant, en fonction des caractéristiques socio-économiques de l'individu et du recours aux soins qu'il déclare. Le modèle est ensuite appliqué aux données de l'enquête handicap-santé pour repérer les individus susceptibles d'être non-consommant. Les hommes et les femmes ayant des consommations de soins assez différentes, deux modèles distincts sont établis. Annexe 1. Cette modélisation est de qualité correcte pour les hommes, un peu moins bonne pour les femmes. Ainsi, 800 individus supplémentaires sont prédits comme non-consommant dans HSM 2008.

Au final, sur les 10 000 individus absents du SNIIR-AM, 650 individus sont non-consommant directement d'après le questionnaire, et 800 autres individus prédits comme non-consommant d'après le modèle, ce qui fait 1 450 non-consommant. L'échantillon final apparié « Handicap-Santé + Sniiram » comprend donc $19\,250 + 1\,450 = 20\,700$ individus et les autres sont considérés comme des échecs d'appariement.

3.4. Traitement de la non-réponse et calage

Les échecs d'appariement ne sont pas distribués au hasard, mais ont tendance à se concentrer sur certains groupes de populations, notamment les femmes, les jeunes et les inactifs, les personnes qui sont plus souvent « ayant droit » à l'assurance maladie. Annexe 2. En effet, pour retrouver les consommations de soins de ces enquêtés, il faut deux Nir, celui de l'ayant-droit et celui de son ouvrant-droit, ce qui augmente les risques d'échecs.

Les poids de sondage doivent être redressés afin que l'échantillon apparié reste représentatif de la population vivant à domicile, en France métropolitaine et DOM. Le redressement est réalisé en deux étapes : la première étape « corrige la non-réponse » [8], c'est à dire le fait que l'échantillon des seuls individus appariés est réduit et déformé. La méthode choisie est une repondération par classes de non-réponse : chaque individu en échec d'appariement est compensé en augmentant le poids des individus appariés de la même classe. Ici, les classes résultent du croisement de six variables¹¹ : sexe, âge, statut d'activité, santé perçue, recours déclaré au généraliste et recours déclaré à un infirmier libéral.

À cette première étape les poids sont au plus multipliés par 2,3.

La seconde étape est un calage sur marges qui permet d'intégrer des variables dites externes, c'est à dire pour lesquelles on ne dispose que des marges sur la population (proportion de non-consommant par sexe et âge issue d'ESPS 2006). Il permet aussi de prendre en compte les autres variables de la base de sondage explicatives de la non-réponse qui n'ont pas pu être utilisées à la première étape. Le calage vise à réduire la variance d'échantillonnage. La macro CALMAR 2 et la méthode linéaire tronquée [9] sont utilisées afin de ne pas trop augmenter la dispersion des poids qui est déjà importante dans l'enquête.

Un premier calage est effectué sur les marges de non-consommant par sexe et âge calculées à partir d'ESPS 2006¹². Tableau 4. Les poids sont au plus multipliés par 2 à cette étape. Ensuite, le calage

¹⁰ Les variables de recours aux soins déclaré communes aux deux enquêtes sont toutefois moins nombreuses que dans la première méthode, c'est pourquoi cette deuxième méthodologie ne remplace pas la première, mais la complète.

¹¹ Moyennant quelques regroupements de modalités judicieux pour ne pas trop multiplier les croisements, on construit 48 classes

¹² Les proportions de non-consommant par sexe et âge semblent plutôt stables dans le temps donc les taux de non consommant en 2006 sont proches de ceux qu'on aurait observé en 2008

final est réalisé sur toutes les autres variables significativement associées à la non-réponse qui n'avaient pas pu être introduite en première étape : nationalité et pays de naissance, revenu et diplôme, recours au spécialiste, et déclarer au moins une maladie chronique. On réintègre aussi les marges globales de chaque variable sociodémographique, santé perçue, généraliste, infirmier, non-consommant, pour « déformer » le moins possible le redressement effectué en amont. Les poids sont multipliés au plus par 1,8 à cette étape.

3.5. Analyse des résultats

Les tableaux suivants donnent une idée de l'ampleur et de la pertinence des redressements effectués.

Tableau 3 : répartition de quelques variables socio-démographiques avant et après redressement.

Répartition (en %)			
	de référence (HSM)	HSM-SNIIRAM (poids initiaux)	HSM-SNIIRAM (poids finaux)
Sexe			
Hommes	48,4	49,9	48,4
Femmes	51,6	50,1	51,6
Classe d'âge			
Moins de 16 ans	19,9	13,7	19,9
16-29	17,8	16,8	17,8
30-44	20,6	22,4	20,6
45-59	20,6	23,5	20,6
60-74	13,5	15,0	13,5
75 ans et plus	7,6	8,7	7,6
Catégorie Socioprofessionnelle			
Agriculteur	2,6	3,2	2,6
Artisan, Commerçant, Chef d'entreprise	4,2	4,6	4,2
Cadre et profession intellectuelle supérieure	9,5	10,8	9,5
Profession intermédiaire	14,4	16,4	14,4
Employé	19,6	21,6	19,6
Ouvrier	15,9	18,4	15,9
Autres	33,8	25,0	33,8
Recours aux soins			
Hospitalisé	11,5	11,6	11,5
Dentiste	53,4	52,8	53,8
Kiné	16,3	16,8	16,5
Psychologue	4,3	3,6	4,2
Complémentaire Santé			
CMU-C	7,2	6,5	7,1
Mutuelle assurance	85,5	86,4	85,9
Sans couverture complémentaire	7,3	7,1	7,0

Source et champ : HSM seule, et HSM appariée avec Sniiram. Calculs Drees. France métropolitaine et Dom.

Note de lecture : D'après l'enquête Handicap santé 2008, 7,2% des individus sont bénéficiaires de la CMU-C. Si l'on se restreint au sous-échantillon apparié avec le Sniiram et en utilisant les pondérations initiales, cette part n'est plus que de 6,5%. Après traitements statistiques 7,1% des individus sont bénéficiaires de la CMU-C dans le sous-échantillon apparié.

Tableau 4 : Proportion de non-consommant par sexe et âge avant et après redressement.

Proportion de non-consommant par sexe et âge			
	de référence (ESPS 2006)	HSM-SNIIRAM (poids initiaux)	HSM-SNIIRAM (poids finaux)
H 00-15	6,9%	3,0%	6,6%
H 16-29	17,0%	17,2%	20,7%
H 30-44	12,6%	9,6%	11,1%
H 45-59	9,1%	7,6%	7,8%
H 60-74	4,9%	3,4%	4,1%
H 75-E+	2,6%	0,9%	1,8%
F 00-15	5,9%	2,6%	5,9%
F 16-29	8,0%	7,4%	10,2%
F 30-44	7,1%	5,6%	6,9%
F 45-59	7,0%	5,3%	7,1%
F 60-74	5,9%	3,9%	5,6%
F 75-E+	2,6%	1,2%	2,1%
Ensemble	8,2%	7,8%	8,2%

Source et champ : ESPS 2006, Irdes, France métropolitaine.

HSM appariée avec Sniiram, Insee, France métropolitaine et Dom.

Calculs Drees.

Note de lecture : D'après l'enquête ESPS 2006, 12,6% des hommes âgés de 33 à 44 ans n'ont pas consommé de soins dans l'année. Si l'on se restreint au sous-échantillon apparié avec le Sniiram et en utilisant les pondérations initiales, cette part n'est plus que de 9,6%. Après traitements statistiques 11,1% des hommes âgés de 33 à 44 ans n'ont pas consommé de soins dans l'année.

L'appariement a été utilisé pour un chiffrage des dépenses de santé des personnes âgées dépendantes dans le cadre de la réforme de la dépendance. Ces travaux ont également été l'occasion de valider les choix méthodologiques retenus pour redresser les données.

Les personnes de 60 ans ou plus ont dépense 1,33 milliard d'euros en 2008 en soins de kiné selon l'enquête Handicap-santé appariée avec l'assurance maladie, soit 101€ par personne. Tableau 5. D'après les données exhaustives de la Cnam, la dépense totale s'élève à 1,16 milliard, soit 102€ par personne. La cohérence des sources nous rassure sur la validité du redressement.

Tableau 5 : estimations des dépenses en soins de kiné et d'infirmiers, selon l'enquête, et selon les données exhaustives de l'assurance maladie.

	Soins de kiné		Soins d'infirmiers	
	Dépense totale (milliards d'euros)	Dépense par personne ayant eu recours (euros)	Dépense totale (milliards d'euros)	Dépense par personne ayant eu recours (euros)
Appariement HS Sniiram données 2007-2008	1,33	101	1,27	96
Données assurance maladie exhaustives 2009	1,16	102	1,19	104

Source : Handicap-Santé ménages 2008 et données Assurance-Maladie, calculs Drees.

3.6. Pour conclure

L'appariement de données administratives avec données d'enquête s'est développé ces dix dernières années (Gensbittel et Riandey, [6]), mais reste encore relativement rare. Une des difficultés majeures est la gestion de la confidentialité des données et des contraintes qui en découlent. Cet article présente les solutions aux problèmes rencontrés lors de l'appariement HS-Sniiram et propose des recommandations pour anticiper les difficultés.

Les données de l'Assurance Maladie permettent de connaître les consommations en soin de ville par type de soin (médecins, pharmacie, matériel médical, ...), ainsi que leur coût total et le reste à charge pour le patient. Les données de l'Assurance Maladie sont ainsi rapprochées de données sur la santé et le handicap et sur les caractéristiques socio-économiques des individus, ouvrant de larges débouchés et des possibilités d'exploitations inédites en économie de la santé.

Bibliographie

[1] Cnis (2006), Chroniques n° 5, « Enquêtes statistiques et sources administratives : une complémentarité à exploiter ».

[2] Bouvier G., « L'enquête Handicap-Santé, présentation générale », Document de travail n°F1109 INSEE, Octobre 2011

[3] Lenormand F., 2005, « Le système d'information de l'assurance maladie », Courrier des statistiques n°113-114.

[4] Waleckx P., « L'hospitalisation, lieu motifs et coûts - France 1980 », CREDES 1985

[5] Bouvier G., « Appariement entre données de l'enquête Handicap-Santé (HS) et données de l'assurance maladie (SNIIRAM) : retour sur les difficultés et recommandations », note interne Insee n°1283/DG75/F170

[6] Gensbittel M.-H., Riandey B., « Appariements sécurisés et statistiques (2000-2011) : Une décennie d'expériences » Courrier des statistiques n°131 Septembre 2011

[7] M. Goldberg, C. Quantin, A. Guéguen et M. Zins, « Bases de données médico-administratives et épidémiologie : intérêts et limites »

[8] N. Caron, « Les principales techniques de correction de la non-réponse, et les modèles associés », document de travail n° 9604 de l'Insee

[9] O. Sautory, « La macro CALMAR, redressement d'un échantillon par calage sur marges », document de travail de l'Insee

Annexe 1 : Modélisation de la probabilité d'être non consommant dans les 12 mois.

	Risque relatif	
	Hommes	Femmes
Age		
Moins de 16 ans	Ref	Ref
16-29	2,8*	2,4*
30-44	2,6*	2,9*
45-59	1,9*	3,0*
60-74	1,2	2,7
75-E+	0,8	1,4
Catégorie socioprofessionnelle		
Agriculteur	0,5	0,6
Artisan, commerçant	0,9	1,6
Cadre	Ref	Ref
Profession intermédiaire	1,0	1,1*
Employé	1,5*	0,6
Ouvrier	0,8	0,9
Autres	1,3*	1,7*
Assurance santé		
CMUC	0,9	-
Mutuelle	Ref	Ref
Sans mutuelle	2,2*	-
Nombre d'adulte dans le ménage		
1	Ref	Ref
2	-	1,5*
3	-	1,7*
Nombre d'enfant		
0	Ref	Ref
1	0,7	0,8
2	0,7	0,7
3	1,0	1,3
Nombre de visite chez le médecin généraliste		
0	4,6*	6,4*
1-6	1,1	2,5*
7-10	Ref	Ref
11 et plus	0,5	1,3
Ne sais pas	2,6*	4,0*
Recours à un médecin spécialiste		
Non	Ref	Ref
Oui	0,4	0,6
Ne sais pas	0,7	0,8
C (aire sous la courbe Roc, qualité du modèle)	0,75	0,69
Sensibilité (non consommant : prédit / réel)	46%	23%
Spécificité (consommant : prédit / réel)	86%	91%

Source, ESPS 2006, calculs Drees. * : modalités significatives au seuil de 5 %.

Lecture : un homme sans mutuelle a 2,2 fois plus de risques d'être non-consommant qu'un homme de même âge, CS, et déclarant le même recours aux soins dans l'enquête. Parmi les hommes réellement non-consommant dans ESPS 2006, 46% sont retrouvés par le modèle (sensibilité).

La régression logistique n'est pas utilisée ici à des fins d'analyse mais de prédiction. Les risques relatifs du modèle n'ont donc que peu d'intérêt : c'est surtout la capacité du modèle à bien prévoir le fait d'être non-consommant qui importe (statistiques de test C, sensibilité et spécificité).

Annexe 2 : Proportion d'individus appariés avec les données du SNIIR-AM

	Part d'appariés
Sexe	
Hommes	72%
Femmes	68%
Classe d'âge	
Moins de 16 ans	46%
16-29 ans	67%
30-44 ans	77%
45-59 ans	75%
60-74 ans	74%
75 ans et plus	71%
Catégorie socioprofessionnelle	
Agriculteur	74%
Artisan, commerçant	75%
Cadre	76%
Profession intermédiaire	76%
Employé	73%
Ouvrier	76%
Autres	57%
Statut d'activité	
Actif occupé	77%
Etudiant	56%
Chômeur	74%
Retraité	74%
Inactif	71%
Régime de Sécurité Sociale	
Régime général des salariés	69%
Fonctionnaires et ouvriers de l'état	71%
Agents des collectivités locales ou de la fonction publique hospitalière	71%
Etudiants	70%
CMU de base	62%
Adultes handicapés bénéficiant de l'allocation adulte handicapé (AAH)	71%
Régime local Alsace-Moselle	79%
Autre régime particulier dépendant du régime général	65%
Régime général sans autre indication	65%
Régime agricole AS (salariés agricoles)	72%
Régime agricole AMEXA (exploitants agricoles)	74%
RSI	72%
SNCF, MINES, RATP	72%
Autres régimes particuliers	66%
MSA sans autre indication	74%
Ensemble	69%

Source handicap-santé en ménage 2008. Calculs Drees.

Lecture : parmi les enfants de moins de 16 ans enquêtés dans HSM 2008, seuls 46% sont retrouvés dans le Sniiram, contre 69% en moyenne