Premiers sondages français dans les dossiers de Sécurité sociale et appariement avec les enquêtes auprès des ménages[1]

Andrée Mizrahi[2], Arié Mizrahi[3]

Introduction

Dans les années 60, les statistiques en micro-économie de la santé étaient issues d'enquêtes auprès des ménages ou des producteurs de soins. Très vite, le statisticiens se sont rendus compte que l'activité administrative de la Sécurité sociale produisait une masse d'information qui pouvait être utile à la statistique et à l'analyse de la consommation médicale.

> Pour analyser cette croissance (*de la consommation médicale)* et en prévoir l'évolution, il est nécessaire de mesurer les biens et services produits et consommés en quantité et en valeur ; l'observation doit être faite aux différents stades du (ou des) processus de production (enquête auprès des producteurs), de la consommation (enquête auprès des ménages) , du financement (statistiques de Sécurité sociale) ; on s'intéresse ici aux possibilités de réalisation, en France, d'une observation du troisième type, c'est-à-dire d'une enquête par sondage sur les documents de Sécurité sociale.
>
> A. et A. Mizrahi, A. Zouaoui, Projet de sondage dans les fichiers de Sécurité sociale, CREDOC, février 1977, 41 p.

> On peut faire des sondages dans des masses de documents existants, par exemple : les dossiers de Sécurité sociale ou les dossiers de malades hospitalisés. L'avantage est le moindre coût : les « questionnaires » existant déjà. L'inconvénient est que ces « questionnaires » ne sont pas conçus à priori pour répondre rigoureusement au problème posé et ne sont pas toujours « renseignés » sur la totalité des questions posées, d'où, souvent, un gros déchet de dossiers inexploitables.
>
> Economique médicale,, G. Rösch et la DEM du CREDOC, Flammarion médecine - sciences, Paris 1973

Les premiers sondages

L'enquête dans les dossiers médicaux de la SNCF (1956)

Il s'agit d'un sondage effectué sur les dossiers de la Caisse de prévoyance de la SNCF (qui joue pour les cheminots le rôle de Sécurité sociale). Sondage au 1/320, soit 1792 dossiers, examinés un par un par un médecin, qui remplissait des fiches individuelles, « exploitées mécanographiquement ». Relevé des données individuelles connues de la Caisse de prévoyance, de la morbidité, et des consommations médicales du 3ème trimestre de 1956[4].

---

1 *Colloque*

2 *Statisticienne, ARgSES, Directeur de recherche honoraire,IRDES, mizrahi@cnam.fr*

3 *Economiste, ARgSES, Directeur de recherche honoraire, CNRS, mizrahi@cnam.fr*

4 *Etude sur les vatiations de la consommation médicale en fonction de l'âge et du sexe, H. Péquignot, J.P. Etienne, N. Parmentier-Lemoigne, La semaine des hôpitaux, 36ème année, N° 9-10, 14 mars 1960, pp. 228-233*

L'enquête pilote de 1965

Le premier appariement en France entre données d'enquête et données de Sécurité sociale a été effectué en 1965 sur une petite enquête expérimentale. C'était une observation sur 549 ménages de la Région Parisienne (377 participation complète, 67 abandons, 67 refus, 38 introuvables). Nous cherchions à tester la possibilité et à mettre au point les méthodes et les documents (questionnaires, cartes réponses…) d'une enquête de plusieurs mois, avec plusieurs visites d'enquêteur. Les N° de Sécurité sociale ont été relevés au cours de l'enquête ; un attaché de l'INSEE a relevé, sur la même période tous les remboursements et les a appariés manuellement avec les données de terrain. Un travail aussi fin et détaillé n'a jamais pu être refait.

La confrontation des deux sources de données a fourni des estimations sur les taux d'erreurs dans les deux sources. Ces biais ont deux origines : (*ou plusieurs l'âge, la plus grande consommation des non-répondants, les oublis et les erreurs quid des fausses déclarations ?)*

- le lien entre niveau de soins médicaux et participation à l'enquête, les personnes qui refusent de participer à l'enquête ou abandonnent ont une consommation de soins de ville (d'après les dossiers de Sécurité sociale) globalement supérieure aux participants à l'enquête de 30 % ; de plus cette « surconsommation » dépend des soins considérés, de 3 % pour les soins de généraliste, 30 % pour la pharmacie prescrite, 46 % pour les soins de spécialistes, elle atteint 74 % pour les analyses de laboratoire (rappelons que les données datent de 1965 et que la structure de la consommation médicale n'était pas celle d'aujourd'hui *ni les taux de remboursement*). Les non participants sont plus vieux que les participants, et leur morbidité est supérieure ; si on corrige de cet effet de l'âge, la « surconsommation » des refus et abandons est réduite mais encore supérieure de 20 % aux participants (tableau 1). Au total, corriger les données d'enquête de la structure de l'échantillon en la rapportant à celle de la population générale ne corrige qu'une partie du biais
- par oubli, négligence ou mauvaise volonté, les enquêtés ne déclarent pas toutes leurs consommations. L'estimation des taux de sous déclaration est complexe et dépend de la qualité de l'appariement : un acte mal relevé peut ne pas être apparié alors qu'il a bien été déclaré, et réciproquement, un acte peut avoir été apparié à tort ; du fait de la sévérité de nos critères d'appariement, nous avons considéré négligeable les faux appariements, ce qui nous a permis d'estimer une fourchette des taux d'oublis (globalement, entre 14 % et 38 % des actes intégralement payés déclarés) ; réciproquement, on pouvait estimer le taux d'actes ne figurant pas dans les dossiers de Sécurité sociale entre 1 % et 22 %. Là encore, les taux de sous estimation dépendaient du lieu des actes et de la spécialité du personnel soignant (tableau 2). *à développer*

Cette enquête pilote a été très efficace puisqu'elle a permis de mettre au point les futures enquêtes ménages sur la santé et les soins médicaux. Elle a aussi fourni des ordres de grandeur des biais sur la consommation médicale estimée à partir des enquêtes auprès des ménages. Il faut noter que le succès de cette opération repose en partie sur l'adhésion des enquêteurs de l'INSEE, triés sur le volet et très enthousiastes à l'idée de participer à une enquête expérimentale de cette importance ; ce succès repose aussi sur le fait que l'appariement a été fait manuellement par un attaché de l'INSEE détaché dans les locaux de la Sécurité sociale dont l'aide lui a été précieuse.

Nous n'avons pas su utiliser ces informations dans nos estimations ultérieures de la consommation médicale, et surtout nous n'avons pas su renouveler cette opération ; Cet

appariement, au niveau le plus élémentaire, n'a jamais été refait à notre connaissance depuis l'enquête pilote de 1965.

Cette lacune est tout à fait regrettable, car si l'enquête auprès des ménages est sous estimée, les données de Sécurité sociale ne sont pas non plus exhaustives, et certains actes, peu remboursés ne sont pas présentés au remboursement ; de plus, la structure et les prix des consommations non connues de la Sécurité sociale sont tout à fait différents de ceux dans les dossiers, dont il découle une sous estimation de la consommation médicale et, selon les cas, sous estimation des prix (dépassements non déclarés) ou une surestimation (soins peu chers non présentés au remboursement). Ces biais d'estimation sont rendus plus gênants du fait qu'ils ne sont uniformes ni selon les types de soins, ni selon les catégories de patients (morbidité, âge, sexe, niveau social, revenu,…) ; ces biais ne sont pas non plus stables dans le temps, ce qui rend les études d'évolution fragiles, surtout lorsqu'on s'intéresse à des périodes de plusieurs années.

C'est pourquoi, nous pensons qu'il serait utile de refaire ces appariements au niveau élémentaire, en dépit du coût et de la difficulté des procédures.

L'enquête de 1966 au 60ème

Les données de Sécurité sociale donnaient lieu à des statistiques nationales exhaustives faites dans le département comptable ; elles étaient établies selon un plan statistique établi dans un esprit comptable : une série de compteurs étaient définis dans le plan statistique, correspondant aux variables dont on voulait estimer la dépense ; chaque dépense venait incrémenter les compteurs correspondants (actes médicaux de la nomenclature des actes professionnels). Ces sommations, dans les centres de paiement, étaient regroupées pour faire des tableaux dans chacune des 122 caisses primaires, puis elles remontaient au plan national et étaient regroupées avec les données des autres régimes pour établir les statistiques donnant lieu à des données mensuelles et annuelles. Ce travail, a été initié par Antoine Sanson-Carette, administrateur INSEE, détaché au Ministère des Affaires sociales et René Nathan statisticien à la FNOSS (fédération nationale des organismes de sécurité sociale[5]), et était réalisé sous l'égide du « Groupe d'harmonisation des statistiques de Sécurité sociale (Régime général, Régime agricole, Régime des travailleurs non salariés, Régime militaire, Régime de la SNCF, Régime de la RATP, Régime de la marine marchande et Régime des mines). Ces statistiques sont très lourdes à mettre en place, et les responsables du Régime Général ont donc accueilli favorablement, en 1975, le projet d'un sondage conduisant à un échantillon de taille réduite permettant d'obtenir des estimations au fur et à mesure des besoins.

L'enquête nationale de 1970

L'enquête pilote de 1965 a permis de décider les modalités de l'enquête nationale de sur la santé et les soins médicaux 1970 (qui ont été reprises en 1980 et 1991) : durée d'enquête prolongée à trois mois, nombre et durée des interviews, nature et formalisation des questionnaires (initial – orienté vers les variables socioéconomiques, ramassage – vers les soins médicaux, navette – vers les évènements dans la durée, final – plus personnel) et des autres documents.

De même, les résultats obtenus de l'enquête pilote ont incité à rééditer la procédure d'appariement, mais cette fois au plan national. L'échantillon était cette fois de plus de 20 000

5 *Les caisses primaires d'assurance maladie (CPAM) sont des organismes qui avaient une large autonomie lors de la création de la Sécurité sociale ; cette autonomie s'est réduite au cours du temps.*

personnes et ce travail ne pouvait pas être fait manuellement, il ne pouvait pas non plus être fait de manière décentralisée dans tous les centres de paiement de la Sécurité sociale. Nous avons donc tenté un appariement informatisé. Les caisses de Sécurité sociale (Régime général[6]) ont envoyé à l'INSEE des copies de tous les versements effectués pour leurs ressortissants enquêtés et retrouvés dans leurs dossiers ; ces étaient très hétérogènes (entre régimes et entre Caisses primaires de Sécurité sociale) et ont été codés à l'INSEE. Cet appariement a été un échec, non seulement au niveau élémentaire des consommations et des remboursements, mais même au niveau des personnes, puisque sur 23 197 personnes enquêtées sur le terrain, et 14 733 d'entre elles couvertes par le Régime général, seule la moitié a été retrouvée (7 393). Au niveau élémentaire, aucun appariement n'a pu être effectué et ce travail n'a jamais abouti ; les difficultés étaient insurmontables, ou n'avons-nous pas su les surmonter ; le pourrait-on aujourd'hui ? Depuis cette époque, un important effort d'homogénéisation des codes de Sécurité sociale, et les moyens informatiques sont incommensurables par rapport aux premiers rudiments à notre disposition.

Il faut souligner un élément important : à ce moment, pour nous la base de l'observation était l'enquête auprès des ménages, l'apport des données de Sécurité sociale ayant essentiellement pour but d'en mesurer l'erreur, éventuellement de les compléter.

L'échec de l'appariement de 1970 provenait en partie du fait que les personnes enquêtées étaient celles appartenant à des ménages tirés au sort à partir du recensement et dont on ne retrouvait pas trace dans les fichiers de Sécurité sociale : au biais découlant du refus venait s'ajouter celui de l'absence de dossier.

L'enquête dans les dossiers de Sécurité sociale de la CPAM de Lyon (1975)

L'échantillon permanent d'assurés sociaux

La situation statistique en 1970

Les données sur lesquelles nous travaillions provenaient des enquêtes décennales (1960, 1970, 1080,…) et il nous semblait que l'évolution était trop rapide pour pouvoir la convenablement 'en 10 ans trop de

Exploitant les données de l'enquête de 1970, il nous restait un sentiment d'insatisfaction car il nous semblait que des informations importantes nous échappaient. Deux éléments nous ont permis d'avancer : l'idée que les données de Sécurité sociales, même non reliées aux données d'enquête auprès des ménages pourraient être utiles à nos recherches et l'intention de nous éloigner d'un sondage à partir du recensement pour le faire directement sur la masse de données de Sécurité sociale. Et ce sondage présentait des difficultés.

6 *Et quelques régimes rattachés au Régime général (fonctionnaires, EDF-GDF, étudiants, invalides de guerre)*

La protection maladie en France. La protection maladie est très morcelée en France, horizontalement et verticalement.
Par ses origines professionnelles, la protection de base est ventilée en plusieurs régimes selon le statut : régimes des salariés du commerce et de l'industrie, des indépendants (artisans, commerçants, libéraux), des agricoles (exploitants et salariés), de la SNCF, de l'EDF-GDF, des étudiants, des mineurs, des marins pêcheurs, des militaires,… ; chacun de ces régimes est avait, et a encore ses règles et ses processus de remboursement (et donc l'organisation de ses fichiers), ses nomenclatures,… Au niveau du Régime général lui-même, ventilé en 122 caisses primaires autonomes, ayant chacune un conseil d'administration, les processus de remboursement n'étaient unifiées non plus que les nomenclatures. La protection de base assure un remboursement des dépenses de soins médicaux engagées variable selon différents critères liés à la personne, à la maladie ou à la nature des soins ; globalement, on estime actuellement à 70 % le taux global de remboursement par la Sécurité sociale (protection de base). Pour les mêmes motifs historiques, seuls les actifs sont assurés (ils paient des cotisations), les personnes à leur charge (conjoints inactifs et enfants) sont leurs ayants droit. Depuis la loi sur la couverture maladie universelle (CMU), les personnes adultes inactives sont également assurées.
A cette protection de base s'ajoute une protection maladie complémentaire (mutuelle ou assurance privée) qui rembourse tout ou partie des dépenses médicales non prises en charge par la Sécurité sociale.
L'EPAS ne couvre pas l'ensemble du Régime des salariés, puisque certaines mutuelles de fonctionnaires ayant une délégation de gestion, gèrent, pour le compte de la Sécurité sociale, les remboursements de leurs membres[7] ; ces mutuelles, dites « mutuelles décompteuses » ont un mode de gestion distinct de celui des CNAMTS, avec des systèmes informatiques différents, et ne désirent pas s'associer à ce sondage.

Les objectifs, les obstacles

Les objectifs de ce sondage étaient multiples :
- créer les statistiques nouvelles nécessaires à la gestion courante de la Sécurité sociale,
- établir des simulations pour évaluer l'effet des projets nouveaux dans la politique de Sécurité sociale, en particulier lors des négociations avec les syndicats des professionnels de santé,
- fournir des fichiers pour la recherche économique ; dans ce but, il avait été prévu de faire une enquête complémentaire auprès des ménages pour relever les informations nécessaires non connues de la Sécurité sociale,
- étudier l'évolution de la consommation médicale d'une même personne au cours du temps (effet panel).

L'origine principale des difficultés de ce projet résidait dans la diversité des institutions de la Sécurité sociale française[8], dans leur autonomie de gestion et de fonctionnement et corrélativement dans la multiplicité et la dispersion des fichiers des assurés sociaux (Cf. encadré).

En deuxième lieu, les documents concernant la même personne peuvent se retrouver dans différents centres de remboursement, soit qu'il travaille (simultanément ou à intervalles rapprochés) sous deux statuts différents, soit qu'il ait déménagé. Surtout le problème des ayants droit n'est pas vraiment résolu : un conjoint ou un grand enfant peut être ayant droit et en même temps assuré à titre personnel, il apparaît alors dans deux dossiers de personnes

7 *Ces mutuelles existaient avant la Sécurité sociale et ont obtenu, lors de sa création, de continuer de gérer entièrement les prestations de leurs adhérents*

8 *Du fait qu'elle a été créée et fonctionne encore sur une base professionnelle*

protégées, sans qu'on puisse les identifier. Or on a besoin d'un dénominateur correct, et donc d'identifier sans erreur chaque personne protégée et de lui attribuer toutes les prestations dont elle a bénéficié.

La troisième difficulté résidait dans le fait que la « liquidation » des dossiers (contrôle, calcul et versement des prestations) était effectué manuellement, comme d'ailleurs l'ensemble des fichiers (immatriculation).

Dans ces conditions, il était hors de question de désigner, à partir du recensement un échantillon dont les prestations seraient suivies au cours du temps, il fallait partir des fichiers de Sécurité sociale.

La méthode de sondage devait permettre de relever manuellement tous les remboursements vers les assurés appartenant à l'échantillon et regrouper les remboursements d'un même assuré (appelés « consommants ») quelqu'en soit l'organisme payeur ; nous espérions aussi que notre échantillon contiendrait des non prestataires au même titre que des prestataires.

La méthode de sondage

Le principe de la méthode consiste à définir une condition restrictive **C**, dont la mise en œuvre est possible dans les différents fichiers ; l'échantillon est constitué de toutes les personnes vérifiant la condition **C**. La proportion de personnes vérifiant la condition **C** égale le taux de sondage.

Pour être utilisable, la condition **C** doit avoir les quatre propriétés suivantes :

a. la condition **C** devait s'appliquer à un caractère se trouvant dans tous les fichiers et être facilement accessible manuellement,
b. la condition **C** doit s'appliquer à un caractère stable dans le temps de manière à contrôler l'échantillon sur lequel portera le sondage,
c. les personnes vérifiant **C** doivent être uniformément réparties dans la population, de manière que l'échantillon soit un sous ensemble représentatif,
d. **C** doit être stable dans l'espace (des fichiers), autrement dit apparaître de la même manière dans les différents fichiers ; si une personne vérifiant C dans un fichier figure dans un deuxième fichier, il faut qu'elle appartienne aux deux sous échantillons issus de ces deux fichiers : par identification et apurement, on est ainsi en mesure d'estimer la consommation totale de chaque personne quelques soient les fichiers dans lesquels elle apparaît.

L'intérêt de ce mode de sondage est qu'il conduit à un échantillon en permanence représentatif de la population de base ; en effet, la proportion est la même parmi les personnes apparaissant deux années successives, parmi les décès et parmi les naissances. L'échantillon évolue comme la population de référence.

Les seules informations existant toujours dans tous les fichiers sont le (ou les) nom et le (ou les) prénom. Nous pensions pouvoir utiliser cette particularité pour faire notre sondage à partir de l'un ou de l'autre ; par exemple, choisir l'initiale du nom. Toutes les initiales ne sont pas aussi fréquentes et nous pensions pouvoir choisir une initiale dont la fréquence correspondrait au taux de sondage visé.

L'INSEE nous a fourni deux fichiers tirés au hasard dans la base SAFARI (système automatisé pour les fichiers administratifs et le répertoire des individus), qui recense toutes les personnes résidant en France, relève leurs lieu et date de naissance et de décès, leur affecte le NNI,…, l'un de 103 300 enregistrements représentatif des personnes nées en France, l'autre

de 48 119 personnes nées hors de France. La validation du sondage retenu a été effectuée sur ces deux fichiers.

Une étude a tout de suite montré que ni les noms et ni les prénoms ne pouvaient être utilisés ; en effet les noms sont différents selon les régions de France et ils ne sont donc pas uniformément répandus sur le territoire, quant aux prénoms, ils sont soumis à des effets de mode et on ne peut retenir l'hypothèse d'une stabilité dans le temps (tableau 3)

> Remarque anecdotique : ce résultat sur les noms a été repris et étendu pour des publications sur la fréquence des noms, et sur leur évolution (noms qui se développent ou qui disparaissent, et a peut être influencé la nouvelle législation sur la transmission des noms ; quant au résultat sur les prénom, il est régulièrement repris dans des publications sur les prénoms à la mode.

Cela a été une grande déception, car il n'existait pas d'autre information commune sur les personnes protégées dans les fichiers de tous les régimes de Sécurité sociale. Cependant, un deuxième identifiant était utilisé par les trois grands régimes (salariés, agricoles, indépendants), le Numéro national d'identité (NNI). Les ambitions du sondage ont donc été réduites à ces trois régimes qui représentent environ 95 % de la population résidant en France.

Le NNI est constitué de 6 zones :

sexe	deux derniers chiffres de l'année de naissance	mois de naissance	département (ou pays) de naissance	commune de naissance	numéro d'ordre	clé de contrôle
□	□□	□□	□□	□□□	□□□	□□

La clé de contrôle a été introduite pour vérifier de manière informatique l'ensemble du NNI, elle est le resté de la division des 13 premiers chiffres par 97. La clé n'apparaissait pas sur tous les documents, comme elle le fait depuis l'informatisation complète du système.

En principe, on désire avoir les deux sexes et tous les âges ; ces deux variables ne peuvent donc pas servir au sondage.

Le mois de naissance peut être utilisé, la structure des mois de naissance est relativement stable sur le territoire français ; au cours du temps, la saisonnalité des naissances a légèrement varié et le mois de naissance n'est pas une très bonne variable de sondage. En tout état de cause, sonder sur le mois de naissance reste insuffisant, car on abouti à un taux de sondage d'environ 1/12ème, conduisant à un échantillon beaucoup trop important.

On a en France 92 départements et dans chaque département, les communes sont classées de 1 à n, en général par ordre alphabétique.

On ne peut pas sonder sur le département, car on veut avoir toute la France ; On ne peut pas sonder sur la commune, car on risquerait de perdre une ville importante.

Le numéro d'ordre est donné à la naissance, classées par jour de naissance et par ordre alphabétique ; on repart à 1 chaque mois ; dans les communes à forte densité de population, les naissances sont nombreuses, et les numéros d'ordre peuvent atteindre plusieurs centaines ; au contraire, dans les communes rurales, on a un petit nombre de naissances, voire pas du tout

certains mois. Sonder sur le numéro d'ordre revient à en choisir un (ou plusieurs) ; les numéros d'ordre élevés feraient disparaître les communes rurales, et par contre un numéro d'ordre petit leur donnerait un poids trop grand.

Nous nous donc sommes rabattus sur un sondage un peu plus compliqué : nous avons déterminé une condition **C1** satisfaisant aux conditions a, b, c, d et d'un emploi suffisamment simple pour être opérationnelle dans la sélection manuelle (et ultérieurement informatique) des assurés de l'échantillon et de tous les documents les concernant :

C1 : les deux derniers chiffres du numéro de commune sont égaux aux deux derniers chiffres du numéro d'ordre. L'idée était que, en moyenne et pour chaque mois, parmi les petites communes, avec peu de naissances, seules seraient représentées celles se terminant par 01, éventuellement par 02 si elle a 2 naissances ; parmi les communes ayant n naissances, les communes se terminant par 01 seraient par la première naissance, celles se terminant par 02 par leur deuxième naissance,… celles se terminant par n par leur $n^{ème}$ naissance ; les communes ayant plusieurs centaines de naissances seraient représentées plusieurs fois (autant que de centaines). On obtenait ainsi un taux de sondage théorique de 1/25, ce qui était insuffisant car conduisant à un échantillon d'environ 2 400 000 personnes, beaucoup trop important pour être géré avec les moyens informatiques des années 70.

C2 : nous avons donc dédoublé la condition C1 en retenant également les cas où les deux derniers chiffres du numéro de commune sont égaux aux deux derniers chiffres du numéro d'ordre dans l'ordre inverse. On dédouble ainsi l'échantillon, sauf dans les cas où les deux derniers chiffres sont égaux.

La réunion de C1 et C2, et tenant compte du cas de l'égalité des deux derniers chiffres, conduit à un taux de sondage théorique de 1/52,554, conduisant à un échantillon de 1 140 000 encore trop gros.

C3 : nous avons ajouté une condition supplémentaire : nous avons cherché parmi les mois de naissance, un mois dont la variation avait été négligeable au cours des décennies précédentes, le mois d'octobre. Les naissances au mois d'octobre sont un peu moins fréquentes que la moyenne des autres mois, se situant entre 7,83 % et 8 % des naissances annuelles au cours des dernières décennies (alors que la proportion moyenne, pour 31 jours, serait de 8,49 %).

Validation

Au total, la condition C, réunion des trois conditions **C1**, **C2**, **C3**, pouvait s'exprimer ainsi :

C : tous les numéros d'assurés nés au mois d'octobre et dont les deux derniers chiffres du numéro de naissance sont identiques aux deux derniers chiffres du numéro d'ordre dans l'ordre et dans le désordre.

On obtient ainsi, avec un taux de sondage proche de 1 / 1200, un échantillon d'environ

60 000 000 / 1200 = 50 000 personnes, qui correspondait approximativement à la taille visée.

Nous avons vérifié que ce sondage conduisait à des sous ensembles satisfaisants aux conditions présentées plus haut :

a. l'indication du NNI est obligatoire pour obtenir les remboursements, il est donc dans tous les fichiers ; par ailleurs, nous avons vérifié dans un centre de paiement que la reconnaissance des pièces à retenir est facile à faire lors du règlement des dossiers,
b. le taux de sondage est indépendant de la classe d'âge (par classes de 10 ans),

c. le taux de sondage est indépendant de la région, il est le même pour les personnes nées en France ou à l'étranger,

Les propriétés b et c ont été vérifiées sur les fichiers SAFARI fournis par l'INSEE

d. en cas d'erreur sur le NNI, les prestations ne peuvent être versées : tous les documents relatifs à une même personne figureront dans le même dossier sous le même NNI.

La réalité est ici un peu plus complexe :

- pour pouvoir être soignées, certaines personnes, non encore immatriculées à la Sécurité sociale, reçoivent un numéro provisoire : pour l'année en cours, leurs documents se partagent entre numéros provisoire et définitif, ce cas est relativement marginal,
- lorsque certaines erreurs sur le numéro d'immatriculation (NNI) ne permettent pas de retrouver la personne, les liquidateurs sont autorisés à « forcer » le versement des prestations en faisant appel à des numéros fictifs ; nous ne pensons pas que ce cas soit très fréquent,
- enfin et surtout, les personnes ne sont pas toutes directement protégées : du fait des origines historiques et professionnelles de la Sécurité sociale, seuls les assurés (actifs, retraités, chômeurs,…) sont directement protégés, les autre (enfants, époux ou épouses inactives) sont leurs ayants droit ; si ses deux parents sont actifs, un enfant peut être ayant droit de chacun d'eux, il peut donc avoir un sous dossiers dans le dossier de son père, et un deuxième sous dossier dans le dossier de sa mère ; sa consommation peut aussi bien être concentrée dans l'un des deux, que partagée entre eux. Lorsque cet enfant devient lui-même actif, la Sécurité sociale lui ouvre un dossier autonome et il disparaît des dossiers de ses parents. Seul le NNI (numéro national d'identité) de l'assuré est relevé dans le dossier, à l'exclusion des NNI des ayants droit. De même, une personne active peut apparaître à tort comme assurée et en même temps comme ayant droit de son conjoint. Ce problème sera résolu avec la prochaine réforme, puisque chacun sera alors personnellement assuré et sera donc remboursé dans son propre dossier.

La mise en place et l'extension de l'EPAS (échantillon permanent d'assurés sociaux)

Le sondage (au 1/12 000, soit 50 000 personnes) a commencé de fonctionner le 1er janvier 1978, sur sept CPAM (caisses primaires d'assurance maladie[9], organismes de base du Régime général des travailleurs salariés) de Normandie et un département de la même région pour les Régimes des agricoles et des indépendants. Dans tous les centres, la saisie était manuelle.

Lorsque, trois ans plus tard, le recueil a été informatisé, les systèmes informatiques des Régimes des agricoles et des indépendants étant différents (motif ou prétexte), ces deux régimes se sont retirés de l'opération. Au niveau du Régime général cependant, le champ géographique s'est progressivement étendu à 46 CPAM au 1er janvier 1981, 106 CPAM au 1er janvier 1985 et à l'ensemble des CPAM (122) au 1er janvier 1989 : 11 ans de montée en charge pour le seul Régime général ! La difficulté était technique (certaines CPAM avaient des systèmes informatiques différents du système national) et politique (convaincre les responsables des CPAM qu'ils ne perdaient pas une trop grande part d'autonomie, cette tâche difficile a été effectuée par René Nathan chef du service statistique de la CNAMTS et Alain Gaillard, statisticien.

En sens inverse, le Régime des indépendants (CANAM) s'est associé au sondage en 1994 et le Régime des agricoles (MSA) en 1996. Aujourd'hui, le sondage EPAS couvre environ 95 %

9 *Les CPAM sont des organismes autonomes, dotées d'un Conseil d'Administration*

de la population, en sont exclus les adhérents des « mutuelles décompeuses[10] » et les assurés des régimes spéciaux (Régime militaire, Régime de la SNCF, Régime de la RATP, Régime de la marine marchande et Régime des mines).

L'EPAS a été dédoublé en 2000 passant à 1/600, soit environ100 000 personnes.

L'enquête auprès des ménages, (ESPS, Enquête santé et protection sociale)

Les fichiers de Sécurité sociale contiennent les renseignements nécessaires au versement des prestations :

- en ce qui concerne les personnes : identité, domicile et/ou lieu de versement des prestations, âge et sexe, éléments de morbidité modifiant les droits,
- en ce qui concerne les prestations : nature et lieu de la consommation, nature du producteur.

Si on veut utiliser ces informations à d'autres fins[11] que le versement des prestations, par exemple pour la recherche, il faut les transformer, éventuellement leur adjoindre d'autres variables utiles à ces fins (variables relatives au patient ou au producteur).

Dès la création de l'EPAS, nous avions envisagé d'adjoindre aux données du sondage des informations issues de l'enquête emploi de l'INSEE, mais nous n'avions pas pu obtenir l'accord des responsables de cette enquête ; L'INSEE était en effet hostile à ce moment à la fusion de fichiers pour des raisons tant déontologiques que pratiques.

Nous avons donc prévu une enquête ad hoc auprès des assurés de l'EPAS. A partir de 1985, ce projet est devenu consistant, à partir du moment où la grande majorité (106) ont effectivement participé au sondage. Une enquête expérimentale a été effectuée en 1986 ; plusieurs points devaient être élucidés : nature du réseau d'enquêteurs (agents de la CNAMTS ou professionnels), durée de l'observation (15 jours ou un mois), mode de contact (téléphone ou face à face), champs des informations à recueillir (socio-démographie, morbidité, protection maladie complémentaire, consommation médicale déclarée) et formulation des questionnaires.

Les unités de sondage. Soulignons qu'il n'y a pas identité entre les unités de sondage dans les deux sources d'observation. Dans l'EPAS, l'unité de sondage est la « grappe assuré », groupant l'assuré et ses ayants droit. Dans ESPS, l'unité de sondage est le ménage de l'assuré : on part de l'assuré et de son adresse, et on relève les informations sur tous les membres du ménage.

Nous n'avions pas les moyens d'enquêter 50 000 personnes tous les ans, aussi a-t-il été décidé que l'enquête porterait sur un quart de l'EPAS chaque année, soit en principe environ 20 à 25 000 personnes. La clé de contrôle du NNI était à ce moment disponible dans tous les fichiers, elle varie de 0 à 97 et permet de ventiler l'échantillon en 4 sous échantillons de même taille (0 à 23, 24 à 47, 48 à 70, 71 à 93). La cinquième année, on relance un nouveau cycle de quatre ans ; les personnes enquêtées la première année, sont ré - enquêtées, les entrants (naissances, immigrés) sont introduits dans l'échantillon. A partir de 1998, l'enquête auprès des ménages ESPS est effectuée par moitié de l'EPAS tous les deux ans. L'échantillon de l'EPAS est donc enquêté tous les 4 ans, un quart tous les ans entre 1988 et 1997, une

10 *elles bénéficient d'une délégation de gestion de la Sécurité sociale et remboursent leurs adhérents aussi bien pour les prestations de Sécurité sociale que pour leurs prestations propres*

11 *Certaines données de Sécurité sociales sont transmises au fisc pour établir (ou contrôler) les revenus des personnels soignants libéraux*

moitié tous les 2 ans depuis 1998 : entre1988 et 2002, cet échantillon a donc été enquêté à 4 reprises (1988-91, 1992-95, 1996-98, 2000-2002). Notons que l'enquête est faite auprès des assurés résidant dans des « ménages ordinaires » et des autres personnes résidant dans ces ménages, qu'elles soit ayant droit des premiers ou non.

Nous désirions avoir également des informations sur les personnes résidant en institution ; elles présentent en effet un grand intérêt du fait de leur état de santé dégradé nécessitant des soins importants. L'enquête a été faite pendant les 4 premières années, 1988 à 1992. Lorsque l'institution a accepté de répondre, le taux de réponse personne n'était que de 57 %, et de plus très différent selon la nature de l'institution, 38 % pour les institutions très médicalisées (long séjour ou section de cure médicale), 64 % pour les peu médicalisées (maison de retraite ou logement foyer). En conséquence, à partir de 1993, l'ESPS s'est limitée aux ménages ordinaires. Nous enregistrons donc dans ce cas un échec conduisant à une lacune dans les observations.

Dans un premier temps, il avait été envisagé de recueillir auprès des ménages les seules données sociodémographiques (profession, niveau d'instruction, revenu, protection maladie complémentaire,…) et de morbidité (maladies, handicaps,…) ; en fin de compte, la décision a été prise de recueillir également des données de consommation médicale. Deux éléments en effet paraissaient manquer dans les données de Sécurité sociale, les consommations non remboursées et certains dépassements, et surtout on n'avait pas la certitude que ces données seraient stables : à ne mesurer la consommation médicale qu'à partir des remboursements, on se prive de la possibilité d'en connaître les évolutions en cas de modification de la part remboursée ou non parallélisme des prix et des tarifs.

Nous ne présenterons pas ici les modalités concrètes de ESPS, ce n'est pas le thème de notre exposé, nous nous contenterons d'en présenter deux difficultés.

Les taux de participation, le chaînage

Un premier problème affecte la qualité de l'enquête auprès des ménages sur le terrain : pour différents motifs (adresse incomplète ou erronée, logement vide, déménagement, déménagement, personne âgée résidant en institution, décès non encore inscrit dans les fichiers de Sécurité sociale[12],…), le contact avec l'assuré ne peut avoir lieu. Le taux de ménages non joints se situe entre 30 % et 35 % et semble stable ;

Une fois le ménage joint, les taux d'acceptation donne une indication de la bonne acceptation de l'enquête : le taux de participants (parmi les assurés joints) a régulièrement diminué depuis le premier passage, passant de 75,0 % pour le premier passage (1988-91) à 70,9 % (1992-95), 66,8 % (1996-98), 63,5 % (2000-02). Cet effet d'usure de l'échantillon est bien connu des organisateurs de panels et devrait conduire à renouveler l'échantillon. Un tel renouvellement sera possible sans grande difficulté, puisque l'EPAS a été doublé, il suffira de prendre l'autre moitié. On perdra alors malheureusement l'intérêt de panel de cette enquête, puisque l'histoire de chaque personne repartira à 0.

Le chaînage permet de relier deux ou plusieurs observations de chaque enquêté et représente l'avantage du panel sur une succession d'enquêtes transversales :

12 *le retard que met la Sécurité sociale à mettre à jour ses fichiers dans les cas de modification de décès ou de déménagement (car les remboursements peuvent être demandés de nombreux mois plus tard)*

- parmi les enquêtés du premier passage, 49 % ont été retrouvés lors du deuxième passage, 27 %lors du troisième passage, 15 % lors du quatrième passage. Rappelons que deux passages consécutifs chez un même enquêté sont espacés en général[13] de quatre ans,
- parmi les nouveaux enquêtés du deuxième passage, 43 % ont été retrouvés lors du troisième passage, 22 % lors du quatrième passage,
- parmi les nouveaux enquêtés du troisième passage, 33 % ont été retrouvés lors du quatrième passage.

Mais on rencontre d'autres configurations, certains enquêtés refusant lors d'un passage et acceptant au suivant. Ainsi, parmi les enquêtés au premier passage, 11 % refusent lors du deuxième passage et acceptent lors du troisième, parmi eux, seuls 5 % participent au quatrième passage ; de même, parmi les nouveaux enquêtés du deuxième passage, 11 % refusent lors du troisième passage et acceptent lors du quatrième ; 3 % enfin des enquêtés du premier passage refusent de participer aux deuxième et troisième passage et acceptent au quatrième.

Soulignons que ces taux ne sont pas très élevés du fait de l'organisation des fichiers de la Sécurité sociale et de la méthode d'enquête : comme nous l'avons vu plus haut, les enfants disparaissent du dossier de leurs parents lorsqu'ils deviennent autonomes ; cette lacune devrait disparaître avec la future réorganisation des fichiers de Sécurité sociale ; de même sont enquêtées les personnes assurées à titre personnel résidant avec les assurés du sondage EPAS, si lors d'un passage ultérieur, cette co-résidence disparaît, elles sortent du champ de l'enquête. Ces deux lacunes viennent s'ajouter aux pertes « naturelles », décès et entrées en institution, l'enquête n'étant faite qu'auprès des ménages ordinaires.

L'appariement, ses limites

Chaque ménage d'ESPS comprend l'assuré d'EPAS, dit assuré principal ; il peut comprendre d'autres assurés, dits assurés non principaux. Seuls l'assuré principal et ses ayants droit peuvent être appariés. En 2000, sur 20 045 personnes enquêtées, 11 257 étaient en principe appariables.

On retient 6 critères d'appariement : NNI de l'assuré, 4 premières lettres du prénom, sexe, jour, mois et année de naissance. Si on tolère une erreur, on apparie 88 % des personnes appariables ; on en ajoute 3 % si on remplace les 4 premières lettres du prénom par la première lettre seulement, et encore 2 % si on abandonne le prénom (appariement 2000). Dans ces conditions, on apparie 93 % des personnes appariables (12 025). A ce niveau, il faut regrouper les doublons, retirer les décès en cours d'année (on n'a leur consommation médicale que sur une partie de l'année), et les adhérents aux mutuelles décompteuses (Cf. encadré). On retire aussi les personnes se déclarant assurées (non principales) et les ayants droit des assurés non principaux n'ayant eu aucune prestation. En 2000, on abouti ainsi à 9 116 personnes satisfaisant aux critères de sondage de l'EPAS, résidant dans un ménage ordinaire, en France, protégée par le régime général (à l'exception des ressortissants des mutuelles décompteuses), des agricoles ou des indépendants, ayant accepté de participer à l'ESPS, dont les informations des deux sources sont cohérentes.

L'appariement ne porte que sur les personnes, à l'exclusion des consommations médicales

13 *lors de la modification du rythme de l'enquête, d'annuelle à bisannuelle, certains espacements ont pu être de 3 ans.*

L'appariement entre les données de l'EPAS et celles d'ESPS a été faite pour 1988, 1992, 1995 1997 1998, 2000, il est fait en routine depuis 2000.

Conclusion, les publications

La masse d'information recueillie est considérable et a donné lieu à un nombre de publications considérable et qui va en croissant. Les travaux ont d'abord été seulement ceux des statisticiens de la Sécurité sociale, puis du CREDES (aujourd'hui IRDES) ; depuis quelques années, les statisticiens du Ministère de la santé ont commencé à les utiliser, et plus récemment, les chercheurs universitaires et des autres centres de recherche.

la richesse de l'information recueillie et la difficulté de la traiter en mettant en valeur toute cette complexité ; en particulier, peu de travaux longitudinaux au niveau individuel sur l'aspect panel.

Lorsque nous avons commencé à travailler sur ce projet, nous n'avions conscience ni des difficultés qui nous attendaient, délais d'autorisation et de mise en œuvre des recueils, et appariements entre données de natures différentes. En effet, les fichiers de Sécurité sociale sont adaptés à la gestion des remboursements, les transformer en fichiers utilisables pour la recherche et les apparier avec une enquête ménage est complexe.