

finalité sans référence au secret statistique institué par la loi de 1951. Cette disposition fut levée en 1986 pour l'INSEE et les services statistiques ministériels et pour la recherche et les autres services statistiques en août 2004 au bénéfice de la transcription tant attendue de la directive européenne du 25 octobre 1995.

En fait, si l'EDF, n'a pas aidé l'INSEE pour la mise à jour des logements de son échantillon-maître, depuis une dizaine d'années, elle s'est montrée un excellent partenaire de l'OLAP pour mesurer les hausses de loyers à l'occasion des changements de bail dans les grandes agglomérations françaises. La qualité de cette source tient à ce qu'en France on consomme plu régulièrement de l'électricité que de l'élection.

Il convient également de rendre hommage aux remarquables estimations locales de populations réalisées à partir de 1999 par Jean-Claude Labat à partir d'une collection de statistiques des administrations et services publiques, avec le souci d'un contrôle mutuel des hoquets des statistiques administratives induits par les changements de réglementations ou de fonctionnements administratifs.

La source scolaire s'est trouvée également inopérante pour la mesure des migrations : les changements d'académies des élèves constituaient une voie intéressante de suivi des migrations des familles, mais la CNIL² s'est opposée au transfert inter-académique d'informations nominatives. Au principe de finalité s'ajoutait le droit à l'oubli, notion dont l'application à l'épidémiologie et à la démographie s'avère très invalidante.

En complément à ces éléments de contexte, précisons maintenant les difficultés opposées au chaînage de données administratives.

En 1978, quand le parlement danois a imposé aux administrations l'usage exclusif d'un l'identifiant national, la loi informatique et libertés française imposait un contrôle rigoureux de l'usage du NIR³ par la CNIL et du Conseil d'Etat. Cette garantie sérieuse n'a pas empêché que les appariements de fichiers fondés sur le NIR ont été vécus en France comme le crime statistique suprême. La CNIL a donc même limité l'usage du NIR à la sphère du travail, de la protection sociale et de la santé. L'Education nationale et le ministère des finances ont été conduits à mettre en place leur identifiant sectoriels sans validation par confrontation au NIR. Les doubles comptes d'étudiants ont alors atteint le taux extrême de 48%, ramené à 8 % quand le service statistique a pu accéder à la liste annuelle des nouveaux bacheliers. Pour sa part, le fisc ne reconnaissait plus ses contribuables entre impôts sur le revenu, sur la fortune ou locaux. Par ailleurs, des parlementaires ont dénoncé le fait que le cloisonnement administratif favorise la fraude : ils citaient la situation de personnes se déclarant mère célibataire aux services sociaux pour toucher l'allocation parent isolé, puis couple cohabitant au fisc pour bénéficier du quotient familial fiscal. Ce cloisonnement constitue également une limite pour la statistique administrative, limite peut-être en voie de dépassement grâce aux appariements sécurisés.

Après une bataille parlementaire et médiatique homérique, le ministre des finances Laurent Fabius réintroduisit à la direction des impôts le NIR que supprima naguère le secrétaire d'Etat au budget Laurent Fabius. Le Conseil constitutionnel a fait preuve d'une très grande sagesse dans son arbitrage : Au Ministère des finances, le NIR ne devra remplir que deux fonctions : valider (dédoublonner) l'identifiant fiscal, permettre les échanges d'informations

² Commission Nationale de l'Informatique et des Libertés

³ NIR : numéro d'inscription au Répertoire d'identification des personnes physiques, le RNIPP

avec les institutions de protection sociale, notamment pour la définition des droits aux allocations sous conditions de ressources. Le Conseil avait retrouvé la logique de l'institution australienne (Bourquard, à paraître) : celle-ci est chargée de valider les identifiants sectoriels, de gérer leur correspondances à partir par référence à l'identifiant national et de permettre, sous son contrôle, les appariements inter-sectoriels de données sectoriellement identifiées. A cette exception ministérielle, cette fonction manque en France, ce dont les milieux médicaux se préoccupent très vivement. Au contraire, les statisticiens officiels, se sont sentis contraints à l'auto-censure, faute d'avoir fait reconnaître la garantie apportée par le secret statistique. Ainsi, l'INSEE qui, dès le départ, avait pris soin de coordonner le sondage des panels EDP⁴ et DAS en vue de leur appariement sur la base du NIR, a attendu plusieurs décennies jusqu'en 1999 pour en présenter la demande à la CNIL de peur de mettre en danger administratif ou médiatique l'un ou l'autre panel. Autre exemple, à la fin des années 90, le Conseil national de l'information statistique n'a pas soutenu la demande de l'INED en faveur d'un dénombrement anonyme des pacs différenciant les pacs hétérosexuels des pacs homosexuels masculins et féminins. Il a fallu attendre la loi du 6 août 2004 pour que le parlement impose cette demande qui n'est d'ailleurs toujours pas appliquée. La solution réside là encore dans les appariements sécurisés qui auraient de plus l'avantage de rapprocher la rupture du pacs comme on pourrait également apparier automatiquement les jugements de divorce des bulletins de mariage de l'état civil.

Retenons enfin que le ministère de l'Intérieur est le seul grand ministère à élaborer ses statistiques sans détenir en soi sein de service statistique ministériel général au sens de la loi de 1951. Le lecteur sera porté à y voir le facteur d'inertie des statistiques de migrations internationales, y compris au moment où en ont été fixées les normes internationales.

C'est dire que la statistique officielle s'est sentie un manque de légitimité politique dans l'utilisation des sources administratives. Le renouvellement du contexte en est d'autant plus significatif. Les appariements statistiques de fichiers à partir de l'identifiant national, réalisés communément en Europe du Nord avec des procédures de sécurité bien standardisées mais reconnues, deviennent accessibles en France par le détour des appariements sécurisés. A chaque pays ses problèmes et ses méthodes !

2- Innovation des épidémiologistes

Se permettant un second détour, démographes et statisticiens constateront que les épidémiologistes sont encore plus fréquemment qu'eux confrontés à des données très sensibles. Le secret médical sous le serment d'Hippocrate interdisait tout transfert d'information médicale aux épidémiologistes avant même qu'en 1978 la loi Informatique et Libertés vienne ajouter l'obstacle du principe de finalité. Les registres du cancer étaient à la fois indispensables et deux fois illégaux jusqu'au vote de la loi bio-éthique de 1994. Curieusement, la statistique officielle au sens strict avait su discrètement rétablir la situation 8 ans avant les épidémiologistes.

Néanmoins, forte de l'obligation et de la légitimité médicales, l'épidémiologie ne semble pas avoir connu cette auto-censure : les registres du cancer ont continué leur développement sans bénédiction ni interdiction de la CNIL. Lors d'un échange relatif aux fichiers épidémiologiques présentés comme nécessaires à la lutte contre le glaucome, un ancien

⁴ Echantillon démographique permanent : appariement aux recensements successifs des bulletins individuels d'un échantillon au centième déterminé par la date anniversaire. Echantillon des Déclarations annuelles de salaire déterminé par des dates d'anniversaires coordonnées.

président de la CNIL s'est ainsi exprimé : « *Il n'est pas possible que quelqu'un devienne aveugle à cause de la CNIL* ». En fait, les juristes de la santé précisent que cette légitimité relève du préambule de la constitution traitant du respect de la personne humaine. La loi statistique de 1951 ne peut prétendre à un tel niveau de noblesse. Cette lacune invite statisticiens et démographes à se placer dans le sillage des épidémiologistes. Telle sera la recommandation de la CNIL dans son rapport de l'année 1999 en évoquant les appariements sécurisés.

Si le recensement général de population est sans contestation nominatif, ce ne peut être le cas des sous-populations de patients, en particulier pour des maladies graves cancers, VIH-sida. La mesure de l'expansion d'une épidémie et la prévision de son devenir exigent un dénombrement sans double compte des nouveaux cas. Le dépistage anonyme de la séropositivité ne satisfaisait pas à cette condition. Les épidémiologistes ont été obligés d'innover sans se réfugier derrière le « *On nous l'interdit* » auquel les statisticiens étaient confrontés, par exemple pour la mesure des discriminations.

La libéralisation du cryptage informatique, à la fin de la décennie 90, en a fourni les moyens : le cryptage de l'identifiant rend invisible l'identité. Encore faut-il qu'elle soit définitivement invisible. C'est la fonction de hachage : cette mini compression de l'identifiant est opérée sur une chaîne numériquement excessive. On évite ainsi le décryptage tout en limitant le risque de mauvais appariement (la collision) à une fraction infinitésimale de l'intervalle de confiance, de l'ordre du nombre d'Avogadro, !

Ces méthodes ont vu leur développement tant à l'hôpital et dans les registres médicaux qu'à l'assurance maladie. Ainsi, le système FOIN⁵ se fonde-t-il sur un identifiant composé du numéro NIR de l'assuré complété du sexe et de la date naissance complète du patient (soit 5 ou 7 caractères de redondance pour l'assuré lui-même) puis ce long identifiant est haché. C'est ainsi que, malgré sa pléthore de régimes et d'institutions, mais sous l'injonction d'une ordonnance du gouvernement en 1996, l'assurance maladie a pu mettre en place un système unifié et exhaustif des consommations médicales tri-annuelles des 60 millions de bénéficiaires (Lenormand, 2005). Bien sûr, cet entrepôt de données herculéen s'est construit avec une rapidité toute relative. Mais le résultat est là et les applications sont progressivement mises en place. Ainsi, par exemple, le web-médecins permet au médecin avec l'autorisation du patient, de consulter ses consommations médicales ayant donné lieu à remboursement. Les créateurs du système méritent un éloge public.

De même, cet identifiant FOIN permet de rassembler dans un même fichier statistique tous les épisodes d'hospitalisation qu'auraient connu le même patient. C'est le PMSI⁶ et ses données vont être appariées au précédent SNIIR-AM⁷.

La puissance exhaustive de ces outils n'en fait pas un outil manipulable pour les aller-retours incessants de la recherche. Le sondage s'impose donc, un sondage au 100^{ème} alimenté non plus sur trois mais sur vingt ans. C'est l'EPIBAM⁸.

L'unité statistique retenue l'individu, est familière aux épidémiologistes et démographes, alors que les économistes travaillent souvent au niveau du ménage, ou de l'assuré pour

⁵ Fonction d'occultation des informations nominatives.

⁶ Programme médicalisé du système d'information (hospitalisation publique ou privée).

⁷ Système national inter-régime d'information de l'assurance maladie

⁸ Echantillon permanent inter-régime des bénéficiaire de l'assurance maladie.

l'assurance maladie de modèle bismarkien. C'est une différence essentielle avec l'EPAS, l'échantillon permanent des assurés sociaux, dont A. et A. Mizrahi (2006) ont si magnifiquement décrit l'histoire et l'appariement à l'enquête santé et protection sociale (ESPS) de l'IRDES⁹ (ex-CREDES). Contrairement à l'Epas, l'Epibam satisfait à l'exhaustivité inter-régimes -exigée par l'ordonnance- alors que les petits régimes corporatistes avaient toujours traîné les pieds pour ne s'intégrer au dispositif antérieur.

Ce colosse statistique impressionne, mais il comprend des limites : l'EPIBAM constitue un vaste panel mais une base de sondage limitée aux enquêtes en population générale ou vastes sous-populations. Dans l'Epibam, les maladies orphelines demeureront orphelines de sondage. Seul le Dossier médical personnel pourrait dépasser cette limite grâce à son exhaustivité..

Le colosse est encore immature : un bénéficiaire de l'assurance maladie commence par être ayant droit. Son identifiant commence par le NIR de son assuré ouvrant droit, puis il s'émancipe et prend un autre identifiant commençant par son NIR personnel. Le statut de conjoint au foyer puis actif conduit au même cheminement. Quant aux enfants de divorcés, il naviguent entre les deux NIR parentaux avant de s'émanciper. Certes sur une petite population, avant cryptage de l'identifiant et grâce à sa redondance, on pourrait envisager un appariement manuel. Après cryptage, les deux identifiants n'ont plus rien de commun. Imaginons qu'une erreur soit simplement intervenue sur le sexe de la personne. (Ne voit-on pas venir accoucher des assurées dont le NIR commence par le caractère 1 de la masculinité !). La correction, aisée avant cryptage, devient impossible après le cryptage monolithique de l'identifiant. De ces remarques, on tire trois conclusions :

- l'identifiant doit être stable. Fin 2006, le « SNIIR-AM 2 » s'appuiera sur le NIR du bénéficiaire après diffusion générale d'une carte « Vitale 2 » de bénéficiaire et non plus d'assuré, toujours fondée sur le NIR personnel. C'est la maturité.
- on doit prévoir une procédure de retour des enregistrements en échec d'appariement auprès des institutions productrices des données nominatives sources ;
- un identifiant très redondant multi-modulaire facilite la correction des erreurs.

Cette dernière conclusion a été mise en oeuvre par le CHU de Dijon dans la procédure Amonymat : si le hachage par l'algorithme SHA est commun à Foin et Anonymat, le premier assure un appariement déterministe fondé sur la concordance exacte d'identifiants tandis que le second ajoute à son caractère multi-modulaire la technique d'un appariement stochastique. Cet appariement prend en compte plusieurs variables, pondérées selon la valeur discriminante de chacune d'entre elles. Le poids est d'autant plus élevé que le pouvoir discriminant de la variable est important. Il est donc recommandé lorsque les données du fichiers sont moins complètes ou précises, à cause d'erreurs.

Les modules composant l'identifiant sont assez souvent les suivants : Nom, Prénom, Date de naissance. Le sexe, compact seulement avant cryptage, est écarté pour son très faible pouvoir discriminant : on est moins souvent homonyme que de même sexe ! Pour ce type d'identifiant s'ajoute le risque des erreurs d'orthographe : une lettre fautive et les identifiants cryptés deviennent méconnaissables. On procède donc d'abord à un traitement phonétique qui simplifie les orthographes. Le relevé du sexe de la personne est assez souvent erroné.

⁹ Institut de recherche et de documentation en économie de la santé.

L'Institut national de veille sanitaire a recours à ces techniques pour l'enregistrement des maladies à déclaration obligatoire. L'identifiant est composé de l'initiale du nom, du prénom, de la date de naissance et du sexe et haché par Anonymat.

Plus compliqué, l'identifiant à composante familiale (Quantin, 2006) mérite développement pour les études internationales, notamment européennes, pour l'épidémiologie génétique et certainement la démographie : nom de naissance d'ego, premier prénom d'état civil, date de naissance, puis nom, prénom, dates de naissance du père, puis de la mère. Cette extension repose d'abord sur l'expérience pédiatrique qui conduit à associer l'identifiant de l'enfant à celui de sa mère, l'un et l'autre pouvant faire l'objet de traitements spécifiques. Pour les études génétiques, ajouter l'identifiant du père s'impose. La carte de santé électronique européenne suppose un identifiant stable indépendamment du pays actuel d'assurance maladie de la personne. Les numéros nationaux sont donc inappropriés ; cet identifiant constitue un candidat numérique adéquat. Validé par la CNIL, cet identifiant anonymisé donne accès depuis l'étranger, avec l'accord du patient, à ses informations médicales sur la plate-forme Internet sécurisée HC Forum. Cette plate-forme est destinée aux études européennes des maladies génétiques rares.

Un hachage de l'identifiant ne suffit pas à assurer la confidentialité totale, notamment parce que l'institution hacheuse détient la table de correspondance entre identifiants brut et hachés ; Pour cette raison, on pratique un double hachage amont avant transfert, aval après réception qui crée l'anonymat à l'égard de tous. Enfin le croisement des informations anonymes du fichier peut permettre une identification indirecte, mais ceci n'est pas spécifique aux données appariées qui doivent à cet égard faire l'objet des mêmes précautions que les autres fichiers.

A l'issue de cette réflexion synthétisée en 2005 par la double publication du *Courrier des statistiques* et du *Journal de la SFdS*, le débat rebondit sur l'accès des épidémiologistes aux données du Dossier médical personnel, le DMP. En juillet 2007, devrait être ouvert chez un hébergeur un dossier médical sécurisé permanent pour toute personne résidant en France. Ce défi technique majeur portera vraiment ses fruits à moyen ou long terme : d'un point de vue médical et démographique, ce sera la première source exhaustive sur la santé diagnostiquée et non plus déclarée.

Constitué à des fins exclusives de coordination des soins, il a fait l'objet d'un débat, suite à une interprétation juridique erronée, déniait le droit à son traitement statistique. C'était oublier la directive européenne, enfin transcrite en droit français : le traitement statistique de données personnelles constituées à d'autres fins est déclaré compatible avec la finalité initiale, dès lors que les garanties légales sont respectées sous le contrôle de la CNIL. Le débat a mobilisé six associations scientifiques et les académies des Sciences et de Médecine (Quantin, Guinot, 2006). Le principe d'un usage épidémiologique de cette source est maintenant acquis, ce qui, pour validation, implique des appariements de ce fichier avec d'autres sources. Se pose donc le choix de l'identifiant compatible avec ces appariements. La logique épidémiologique le place dans la famille FOIN, ce que devra confirmer un décret annoncé pour novembre 2006.

La question de l'identifiant santé est donc posée dans tout sa généralité. Elle fait l'objet d'un vaste débat international (Bourquard.). Les états fédéraux, Etats-Unis et Canada par exemple, partent en mauvaise situation par rapport aux états centralisateurs. L'Australie et la Nouvelle-Zélande figurent parmi les plus libres innovateurs, ayant échappé aux sinistres drames européens. Plusieurs logiques s'opposent : la refonte complète des fichiers médicaux de toutes

les institutions selon une norme commune. C'est le modèle anglais assez difficile à mettre en place. Les Pays-Bas ont opté pour la navigation entre les fichiers de toutes les institutions avec registre de repérage et norme d'extraction. : les flux sont intensifiés, mais les fichiers réduits et probablement plus sécurisés. La solution des identifiants sectoriels coordonnés fait son chemin, notamment en Belgique et en Autriche selon l'efficace modèle australien. Le débat déborde celui déjà difficile du fichier médical hospitalier (*Info en santé, 2006*). Si la France s'affiche comme pionnière du DMP, à tort ou à raison, elle est notée queue de classe pour le dossier hospitalier. Sans doute, l'écart de la coupe aux lèvres !

Avec le PMSI des hôpitaux, l'Institut de veille sanitaire, l'assurance maladie, bientôt peut-être le DMP, ces méthodes développées au DIM de Dijon ont pris une grande envergure dans le domaine de la santé. Qu'en est-il dans les autres secteurs de la statistique ?

3- Finalités statistiques.

Le CREDOC s'est vu confier la mission d'apparier des données sociales relatives aux allocataires parisiens du minimum social, le RMI. Ces informations sont issues d'institutions multiples, ce qui a posé un problème difficile en droit français. Dans son rapport sur l'année 1999, la CNIL a suggéré de recourir aux méthodes mises au point par le CHU de Dijon, ce qui fut fait sous le logiciel FOIN de l'assurance maladie (Aldeghi, 2004) : ces méthodes d'anonymisation fécondes en épidémiologie ne l'étaient pas moins pour la statistique publique. Cette idée simple a été développée à l'INED et par la Société française de statistique, en collaboration avec le Département d'Informatique Médicale du CHU de Dijon, par des séminaires, tables rondes, cours, présentation aux Journées de méthodologie statistique de l'INSEE.

Le service statistique de l'Education nationale a été le premier service statistique ministériel à mettre en œuvre ces techniques (Goy, 2005). Les freins imposés à l'utilisation nationale de l'identifiant INE du ministère interdisait le suivi complet d'un parcours universitaire. Chaque université ignorait le parcours de ses étudiants à la sortie de leur université ; le taux d'échec apparent des études supérieures en était gravement surestimé. Ces freins ont été abolis par le hachage des INE par la procédure Anonymat. Les 99 universités françaises sont maintenant en mesure de suivre statistiquement le devenir des étudiants de leurs diverses filières et le bilan de l'enseignement supérieur français en est réévalué. Par contre, les doublons d'étudiants ne sont pas repérés, faute d'un contrôle externe auprès du RNIPP, comme au ministère des finances, selon le modèle australien.

4- Mesure des discriminations

La publication d'un important dossier en commun par le *Courrier des statistiques* et le *Journal de la SFdS* énonçait l'ambition de la SFdS de voir ces techniques débloquent le traitement statistique des données ou institutions sensibles. La mesure des discriminations en est un exemple caractéristique. Les statisticiens français sont confrontés depuis des lustres à une demande de mesure des discriminations ethniques et à l'interdiction de collecter les informations nécessaires dans des conditions sérieuses, c'est à dire non sélectives. L'obstacle est d'abord apparu dans les institutions de protection sociale. Les entreprises sont maintenant concernées (Simon, 2006). Une directive européenne crée l'obligation de lutter contre ces discriminations. Selon la HALDE¹⁰, chaque année, les salariés ne pourraient-ils déclarer anonymement leur appartenance à une « minorité visible » de manière à mesurer la diversité

¹⁰ Haute autorité de lutte contre les discriminations et pour l'égalité.

des recrutements et l'équité dans les promotions (Bébéar, 2000). Azouz Begag a énoncé une proposition équivalente concernant les recrutements dans la police. La réticence de la CNIL serait probablement moindre, voire effacée, si cette information sociale auto-saisie était immédiatement cryptée puis traitée à l'extérieur.

Dans le passé, les services statistiques étaient inévitablement tributaires de données de gestion. Aujourd'hui, les entreprises et administrations pourraient collecter dans les bases de gestion ces informations sensibles immédiatement cryptées qui ne pourraient être déchiffrées que par des statisticiens extérieurs. Ce serait une belle avancée de l'utilité sociale de la statistique et de la technologie du secret statistique. On peut imaginer qu'il s'agit d'un nouveau métier de la statistique. Les entreprises signataires de la Charte de la diversité pourraient prendre l'initiative de créer ces bureaux.

Les institutions sociales suspectaient le fait que les familles immigrées ne bénéficiaient pas pleinement à leurs droits à aux allocations, mais le repérage de ces populations était interdit par la loi car la collecte des informations « ethniques » nécessaires était « excessive par rapport à la finalité » administrative, le versement des allocations. L'enquête par sondage était alors citée comme la seule source possible de mesure des discriminations malgré sa puissance réduite pour mesurer des dysfonctionnements supposés quand même bien minoritaires. La reconnaissance par la directive européenne de la finalité statistique comme compatible avec la finalité de collecte a résolu cette contradiction. Par contre, si l'information est plus sensible que le sexe, la nationalité ou le pays de naissance, l'obligation de l'accord exprès, évident en médecine, demeure un facteur de biais dans le domaine social. Pour les administrations s'ajoute l'interdiction constitutionnelle des traitements administratifs différenciés entre citoyens français selon leur origine. D'ailleurs, les salariés sont très réservés sur l'idée de fournir ces données personnelles sensibles à leur employeur, mais positifs pour les instituts de statistique ou de recherche (Simon, 2006).

Ne sont dispensés de l'accord exprès que l'INSEE et les services statistiques ministériels après avis du CNIS quant à l'intérêt public de la statistique concernée. La loi de statistique publique de 1951 apporte à la statistique officielle la garantie juridique du secret statistique tandis que les instituts de recherche demeurent extérieurs au périmètre d'application de cette loi (d'où, rappelons le, leur retard d'accès aux données personnelles de 1986 à 2004). Le paradoxe veut que les institutions agréées¹¹ ne sont pas nécessairement disposées à assumer cette responsabilité, même en coopération avec ces instituts.

5- Estimations démographiques

Par ces nouvelles techniques, les épidémiologistes ont appris à apparier, dédoublonner, chaîner dans le temps, pour corrélérer, dénombrer, mesurer des transitions. Ces opérations essentielles de la statistique démographique mais problématiques pour la statistique administrative française voient donc là de nouvelles solutions.

Dans le secteur de la démographie professionnelle, ils ont été confrontés à estimer et projeter des effectifs ou des pyramides professionnelles souvent à partir de plusieurs fichiers aux champs distincts sans droit de les connecter : dans le domaine de la santé, un fichier administratif de professionnels de santé et celui de l'Ordre de la profession. L'appariement

¹¹ L'INSEE est légalement en situation de tester les biais induits par l'accord exprès à partir de deux sous-échantillons comparables ; nous n'avons pas connaissance que de telles expériences aient été menées.

anonyme par cryptage des identifiants s'impose pour un dénombrer sans double compte, puis décombrer et caractériser les sous populations non appariées.

Dans un système normalement informatisé, l'enregistrement sécurisé des pacs selon l'orientation sexuelle éviterait tout débat. L'enregistrement d'une rupture de pacs avec la même clé de hachage amènerait au chaînage des deux enregistrements anonymes et donc à une analyse facile des ruptures par cohorte, durée et orientation sexuelle. Sans doute est-ce un peu anticiper par rapport à l'équipement des greffes, mais c'est évidemment une voie qui s'impose.

Le dispositif de l'assurance maladie issu des ordonnances a donné jour à quasi registre de population avec le RNIAM, certes dénué d'adresses directes. Il n'empêche qu'il produit en permanence une pyramide des âges de la population bénéficiaire. Les démographes ne peuvent se désintéresser de cette nouvelle source.

Toutefois, si l'entrée au répertoire est bien régulée, les radiations souffrent d'incomplétude. L'émigration de France, avec sa forme la plus définitive, la mortalité à l'étranger échappent à la mise à jour. De ce défaut, on rêve de faire un brillant atout : partant du principe que les morts et les émigrés ne consomment plus de soins médicaux en France, on est tenté de repérer les nouveaux non consommateurs du SNIIR-AM et, déduction faite de la pyramide des décès, de déduire celle des émigrés (Riandey, 2004).

L'entreprise relève du défi pour plusieurs raisons : la suspension de la consommation peut s'avérer fictive, due à un changement d'identifiant du bénéficiaire. Ce sera fréquemment encore le cas jusqu'à l'introduction prochaine de la carte Vitale 2 fondée sur le NIR du bénéficiaire et non plus de l'assuré ouvrant droit. Resterait toujours à vérifier l'importance des erreurs d'identifiants. En second lieu, le Sniir-am enregistre la non consommation pendant au plus trois ans ; or celle-ci est maximale pour les hommes jeunes adulte aux âges de forte émigration ; peut-on construire un modèle stable entre durée de non consommation observée et probabilité d'être émigré ? Enfin l'assurance maladie connaît des situations juridiques particulières de maintien des droits à l'étranger.

Le même souci d'utiliser les fichiers d'assurance maladie à des fins démographiques a été expérimenté en Grande-Bretagne, mais pour la mesure de la migration intérieure, avec l'espoir d'une amélioration des données à l'usage (Chappell, 2000). Les effectifs concernés sont beaucoup plus importants et donc les erreurs relatives plus faibles.

De façon plus générale, l'évaluation de la survie d'une sous-population à partir des fichiers de décès ou du statut vital au RNIPP dépasse largement le cadre de l'épidémiologie (Fournel). Les techniques de hachage y sont tout à fait utiles.

L'analyse démographique des affaires devant la Justice accuse un retard considérable en France. Ne serait-ce qu'à cause de l'extrême sensibilité des données, la statistique est bien en peine pour opérer le chaînage entre l'introduction en justice et le jugement final, chaque jugement constituant une nouvelle affaire. Le chaînage devrait même commencer à la plainte de police, enregistrée sous l'égide d'un autre ministère. On imagine l'apport que pourraient attendre leurs services statistiques de ces méthodes de traitement longitudinal dans le respect de la confidentialité.

Les champs d'application démographiques de ces techniques sont immensément variés : démographie familiale ou génétique, panels de retraités ou de cotisants, cohorte d'enfants, suivi de l'enfance en danger, démographie des pacs, statistiques d'immigration, démographie des étudiants... Une journée de formation sur les appariements sécurisés, organisée par la Société française de statistique, le 16 novembre 2006, réunissant démographes et épidémiologistes, s'est muée en un séminaire annuel, le 16 novembre ouvré de chaque année, ouvert aux volontaires. Dans quelques années, on fera le point sur l'usage et l'apport démographiques de ces méthodes.

Bibliographie

Dossier spécial « panel et appariements sécurisés », *Courrier des statistiques* n°113-114, juin 2005 et *Journal de la Société française de statistique*, vol 146 n°3, 2005 coordonnés par B. Riandey :

- Quantin C., Gouyon B, Allaert F.A., Cohen O, « Méthodologie pour le chaînage de données sensibles tout en respectant l'anonymat : application au suivi des informations médicales ».
- Goy A. « L'appariement sécurisé des fichiers d'étudiants grâce au hachage des identifiants ».
- Lenormand F. « Le système d'information de l'assurance maladie ».

Aldeghi I. et Olm Ch. (2004) « L'observatoire des entrées et sorties du RMI à Paris » in Ardilly P. (dir) « Echantillonnage et méthodes d'enquêtes », Actes du 3^{ème} colloque francophone sur les sondages de la Société Française de Statistique, Dunod, Paris.

Bébéar C. (2004) « Des entreprises aux couleurs de la France ». Rapport au Premier ministre, La Documentation française. Paris.

Bourquard K, (à paraître) « DMP : «Situation Internationale » *GMSIH*.

Chappell R., Vickers L. And Evans H.(2000). "The uses of Patient Registers to estimate migration". *Population Trends* 101, pp19-24

Fournel I. «(à paraître dans la Revue d'épidémiologie et de santé publique), « Détermination du statut vital par chaînage entre des données hospitalières et les données de mortalité nationales anonymisées »

Info en santé (2006) n° 16 « Le dossier médical personnel ». Fédération hospitalière de France.

Jaro M. A, 1995, « Probabilistic-linkage of large public health data files". *Statistics in Medicine* 14 pp.491-8.

Labat J.C (1999) « Une nouvelle méthode d'estimation de population », INSEE.

Mizrahi A. et A. (2006) « Premiers sondages français dans les dossiers de sécurité sociale et appariement avec les enquêtes auprès des ménages », in Lavallée P. et Rivest L.P. (dir) « Méthodes d'enquêtes et sondages. Pratiques européenne et nord-américaine » Acte du 4^{ème} colloque francophone sur les sondages de la Société Française de Statistique, Dunod, Paris.

Prieur C. (2006) « Le numéro de sécurité sociale pourrait devenir la clé d'accès au dossier médical ». *Le Monde* du 15 novembre, page 13.

Quantin C., Cohen O, Riandey B. Allaert F.A "Unique Patient Identifier : a world wide review" à paraître *International Journal of Medical Informatics*

Quantin C., Guinot Ch., Tursz A., Salomez JL, Rogier C., Salamon R., (2006) « Le traitement épidémiologique du Dossier Médical Personnel au service des malades », *Revue d'Epidémiologie et de santé publique, Volume 54, avril 2006, N° 2, 54 : 177-96.*

Quantin C., Gouyon B., Allaert F.A, Cohen O, (2006) « Proposition d'un identifiant à composante familiale rendu anonyme » in Lavallée P. et Rivest L.P. (dir) « Méthodes d'enquêtes et sondages. Pratiques européenne et nord-américaine » Acte du 4^{ème} colloque francophone sur les sondages de la Société Française de Statistique, Dunod, Paris.

Riandey B., 2004, « Le nouveau système d'information de l'assurance maladie en France permettrait-il une estimation des flux d'émigration ? » Colloque de l'AIDELF à Budapest, septembre 2004 (actes à paraître).

Simon P., Clément M., (2006). « Comment décrire la diversité des origines en France ? », *Populations et sociétés*, n°425.

Scott A. and Kilbey T. (1999). "Can patient Registers give an improved measure of internal migration in England and Wales" , *Population Trends* 96, pp 44-55.