DOSSIERS solidarité et santé

Données de santé : anonymat et risque de ré-identification

N° 64

Juillet 2015

Ce *Dossier Solidarité et Santé* présente les résultats du travail préparatoire mené par la DREES pour élaborer l'article 47 du projet de loi de modernisation de notre système de santé. Voté en avril 2015 en première lecture à l'Assemblée nationale, ce texte propose un équilibre raisonné entre ouverture et protection des données de santé.

En quoi consistent les risques de ré-identification pour des bases de données en apparence anonymes ? Et comment les anonymiser ou encadrer leur accès ? Dans son article, André Loth, co-auteur du rapport Bras de 2013, explique comment permettre l'utilisation des données pour le bénéfice de tous, sans mettre en danger le droit de chacun à la protection de sa vie privée. Sont aussi évoquées les avancées notables du projet de loi concernant la simplification des procédures, notamment le numéro national d'identification, les appariements et le rôle d'un tiers de confiance.

Afin de mieux comprendre les enjeux autour des données nominatives, Jean-Pierre Le Gléau, longtemps chargé de ces questions à l'INSEE, évoque le débat juridique sur la définition de l'anonymat : existe-t-il des critères absolus ou doit-on s'en tenir aux moyens susceptibles d'être raisonnablement mis en œuvre pour identifier une personne ? La formulation de la loi française est plus exigeante que celle de la directive européenne… mais est-ce bien raisonnable ?

Un article collectif fait le point sur les principales bases de données de santé utilisées en France pour la recherche. Des exemples d'appariement montrent l'intérêt de ces données afin de répondre à des questions cruciales pour améliorer la santé de la population.

Enfin, l'article historique du Dr. Dominique Blum sur le pouvoir de ré-identification des bases de données du PMSI est publié en annexe dans son intégralité. Cette étude alerte sur un défaut de protection des données hospitalières : il fut en partie à l'origine des rapports et du projet de loi qui ont suivi.

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

Liberté · Égalité · Fraternité
RÉPUBLIQUE FRANÇAISE

MINISTÈRE DES FINANCES ET DES COMPTES PUBLICS

MINISTÈRE DES AFFAIRES SOCIALES, DE LA SANTÉ ET DES DROITS DES FEMMES

MINISTÈRE DU TRAVAIL, DE L'EMPLOI, DE LA FORMATION PROFESSIONNELLE ET DU DIALOGUE SOCIAL

Direction de la recherche, des études, de l'évaluation et des statistiques (Drees)
Ministère des Finances et des Comptes publics
Ministère des Affaires sociales, de la Santé et des Droits des femmes
Ministère du Travail, de l'Emploi, de la Formation professionnelle et du Dialogue social

Sommaire

Avant-propos

Franck VON LENNEP, directeur de la DREES

L'ouverture des données de santé est une préoccupation ancienne du ministère chargé de la santé qui remonte au début des années 2000.

Grâce à la mise en place de l'Institut des Données de Santé en 2007, d'une part, et à l'action de l'assurance maladie qui depuis plusieurs années a construit un grand nombre d'outils nouveaux et les a progressivement ouverts à l'extérieur, de nombreux progrès ont été effectués depuis dix ans.

Mais beaucoup restait à faire. Marisol Touraine, la ministre des Affaires sociales, de la Santé et des Droits des femmes, a commandé en 2013 un rapport à Pierre-Louis Bras, membre de l'Inspection générale des affaires sociales. Suite à la remise de ce rapport en octobre 2013, elle a installé une commission, dite « commission open data », co-animée par le directeur de la recherche, des études, de l'évaluation et des statistiques (DREES) et le délégué à la stratégie des systèmes d'information de santé. Cette commission, composée de représentants des parties prenantes (producteurs et utilisateurs de données, parmi lesquels chercheurs, représentants des patients et usagers, des professionnels et des établissements de santé, des organismes complémentaires, des industriels, etc.) a remis un rapport en juillet 2014.

L'article 47 du projet de loi de modernisation de notre système de santé, voté en première lecture à l'Assemblée nationale en avril 2015, reprend et complète les orientations et préconisations de ces rapports. Il précise les principes, les modalités et la gouvernance de l'accès aux données du « système national des données de santé », le principe général étant la recherche de l'équilibre entre ouverture des données et protection contre le risque de ré-identification des personnes dans ces bases de données.

La ministre a chargé la DREES d'élaborer le texte de l'article 47, ainsi que ses futurs textes d'application, et plus largement l'a chargée de piloter la gouvernance de l'accès aux données de santé du ministère. A ce titre, la DREES a engagé un travail technique et méthodologique, peut-être unique en Europe, autour des risques de ré-identification dans les bases de données médico-administratives et des moyens de s'en prémunir.

Ce Dossier présente une partie des résultats du travail préparatoire à la nouvelle loi et un ensemble d'éléments utiles à sa compréhension. Il présente ainsi plusieurs articles portant sur l'analyse du risque de ré-identification et les moyens juridiques et techniques à mettre en œuvre pour limiter ce risque, tout en facilitant les usages utiles à la recherche et aux études. Il livre ensuite un état des savoirs sur les méthodes permettant d'élaborer des bases de données anonymes à partir de données présentant un risque de ré-identification. Enfin, des chercheurs plaident pour une multiplication des usages des bases de données médico-administratives et pour la facilitation des appariements de bases de données.

RISQUES DE RÉ-IDENTIFICATION DANS LES BASES DE DONNÉES DE SANTÉ, MOYENS DE S'EN PRÉMUNIR : UN PROJET DE LOI CONCILIANT OUVERTURE ET PROTECTION

André LOTH (DREES)

Un groupe de travail sur les risques de ré-identification dans les bases de données de santé, piloté par la Direction de la recherche, des études, de l'évaluation et des statistiques (DREES), a rendu son rapport en juin 2014 dans le contexte d'un vif débat sur l'ouverture de ces bases et dans la perspective d'un projet de loi qui est désormais l'article 47 du projet de loi de modernisation de notre système de santé, voté en première lecture le 14 avril 2015 par l'Assemblée nationale.

Comme le séminaire organisé par la DREES le 10 décembre 2014 sur le même thème, ce numéro des *Dossiers solidarité et santé* s'inscrit dans la continuité des travaux du groupe de travail et de son rapport. Le rapport lui-même, déjà publié en annexe du rapport de la Commission « Open data en santé » (juillet 2014), figure dans ce *Dossier*.

Dans le débat où le Parlement est appelé à se prononcer par son vote, les opinions sont tranchées :

- Pour les uns, le gouvernement et l'assurance maladie prennent prétexte des risques de ré-identification des personnes et les amplifient pour interdire l'accès de la société civile et des entreprises innovantes à un trésor (pour la santé, la démocratie et la croissance économique…). Ils attribuent ce comportement des pouvoirs publics à l'ignorance ou au conservatisme bureaucratique. Quand ces tenants de l'ouverture ne nient pas le risque, ils le relativisent, évoquant un nécessaire compromis entre risques et bénéfices qui justifierait selon eux une ouverture maximale ;
- D'autres à l'inverse mettent en garde le gouvernement contre les risques de ré-identification et de mésusage au regard du principe constitutionnel de protection de la vie privée. Ils affirment qu'il est impossible de rendre totalement anonymes des jeux de données individuelles pour les mettre à la disposition de tous, que l'accès aux données à caractère personnel doit être limité aux administrations concernées et aux chercheurs et qu'on ne peut mettre en balance les risques des uns et les bénéfices des autres.

Dans la ligne du rapport que Pierre-Louis Bras avait remis à la Ministre des Affaires sociales, de la Santé et des Droits des femmes en octobre 2013 et du rapport de la Commission open data en santé, le projet de loi du gouvernement propose une ouverture sécurisée des données médico-administratives, qui tient compte des arguments des uns et des autres et qui peut se résumer en deux principes :

1. Que les données vraiment anonymes soient mises librement à la disposition de tous selon un principe « d'open data »[1]. Cela implique toutefois d'avoir défini ces données vraiment anonymes, qu'il s'agisse de tableaux de données agrégées ou d'enregistrements « granulaires » individuels (données se rattachant à une seule personne même si cette personne n'est désignée ni par son nom ni son numéro de sécurité sociale) ;
2. Que les données personnelles de santé, préalablement dé-identifiées, ne soient rendues accessibles que pour de bonnes raisons (dites « d'intérêt public [2]»), à des personnes nommément identifiées, habilitées par leur

[1] Et que soient aussi rendues disponibles en open data les données sur l'activité des professionnels de santé quand ces données, nominatives, ont déjà été rendues publiques par l'assurance maladie (par exemple les tarifs moyens par acte).

[2] Intérêt public, expression consacrée dans la loi informatique et libertés, signifie ici la même chose qu'intérêt général ou bénéfice collectif. Sont exclus bien sûr des usages comme la ré-identification des malades et le ciblage des comportements de prescription individuels des médecins à des fins commerciales. Sont exclus aussi les usages dont la finalité serait *essentiellement* privée. Mais une recherche sur les effets d'un médicament peut bien sûr présenter à la fois un intérêt pour son promoteur et un intérêt pour la société. En cas de doute, il est prévu que le futur Institut national des données de santé (INDS) donne son avis à la CNIL sur le bien fondé des finalités. La notion de « bonne raison » implique en outre que chacun n'accède qu'aux données strictement nécessaires aux fins qu'il a fait valoir (la question sera soumise à un comité d'expertise qui donnera son avis à la CNIL). Le projet de loi a prévu aussi que des organismes dont les missions de service public l'exigent, désignés par un décret en conseil d'État pris sur avis de la CNIL, auront un accès permanent à certaines catégories de données personnelles du Système national des données de santé.

responsable hiérarchique, présentant des garanties suffisantes et dans des conditions techniques assurant disponibilité, intégrité, confidentialité et auditabilité[3].

ENCADRÉ 1 - DONNÉES INDIVIDUELLES ET DONNÉES PERSONNELLES

« *Constitue une donnée à caractère personnel toute information relative à une personne physique identifiée ou qui peut être identifiée, directement ou indirectement, par référence à un numéro d'identification ou à un ou plusieurs éléments qui lui sont propres. Pour déterminer si une personne est identifiable, il convient de considérer l'ensemble des moyens en vue de permettre son identification dont dispose ou auxquels peut avoir accès le responsable du traitement ou toute autre personne.* » (loi n° 78-17 du 6 janvier 1978 article 2)

Il ne faut pas confondre données individuelles et données personnelles. Ces dernières sont rattachées à une personne « identifiée ou identifiable » alors que les données individuelles sont seulement classées par individu sans qu'on connaisse nécessairement l'identité de cet individu. Par exemple une base où, pour chaque habitant d'une ville, on indique seulement son âge en années (en regroupant les plus de 100 ans) et la tranche de revenus de son ménage, est une base constituée de données individuelles mais non personnelles. Dans une base de données destinée à des fins statistiques ou de recherche, les données d'identification directe (nom-prénoms, numéro de sécurité sociale, numéro de téléphone, adresse électronique ou postale etc.) sont supprimées ou conservées séparément des autres données : l'identité des personnes est remplacée par un pseudonyme (par exemple un numéro d'ordre ou un numéro dit d'anonymat). Dans le cas du SNIIRAM[4] ou du PMSI[5] et plus généralement du futur Système national des données de santé (SNDS), ce numéro d'anonymat est obtenu par un chiffrement irréversible du NIR, effectué en amont, de sorte que même les gestionnaires de la base ne puissent pas remonter à l'identité des personnes.

On voit qu'il y a au moins deux sujets différents, appelant des réponses différentes :

- Assurer l'anonymat des données en open data c'est-à-dire en accès libre à tous ;
- Protéger adéquatement les données dont l'anonymat ne peut être garanti même après qu'elles ont été dé-identifiées.

Dans l'un et l'autre cas toutefois il faut avoir compris préalablement comment et dans quelle mesure des données dé-identifiées ne comportant ni le nom ni le NIR peuvent présenter un risque de ré-identification. A contrario il faut admettre que la dé-identification, ou pseudonymisation est une condition nécessaire mais non suffisante de l'anonymat (1).

Pour mettre des jeux de données en accès libre, il faut que les autorités compétentes définissent, en toute transparence, des critères de ce qui peut être raisonnablement considéré comme anonyme ou mis à disposition sans risque (2).

Pour les jeux de données comportant des risques de ré-identification, il faut définir des procédures d'accès qui tiennent compte des finalités et imposent des dispositifs de protection (3).

Enfin, un aspect particulier et commun à ces sujets est celui du NIR[6] et de son emploi pour apparier des données (4).

*

[3] Ces quatre qualités sont souvent désignées par leurs initiales : DICA. Auditabilité est ici synonyme de traçabilité : on conserve la trace de qui a accédé à quelles données et, autant que possible, pour quels traitements.

[4] Le Système national d'information interrégimes de l'assurance maladie est alimenté par les feuilles de soins (1,2 milliard par an) et les fichiers de l'ensemble des caisses d'assurance maladie.

[5] Le Programme de médicalisation des systèmes d'information (hospitaliers) a permis le recueil de résumés de séjours standardisés sur lesquels est fondée depuis 2004 la tarification à l'activité (T2A) des hôpitaux et cliniques.

[6] Le Numéro d'Inscription au Répertoire national d'identification des personnes physiques (NIR) est plus communément appelé numéro de sécurité sociale bien que les ayants droit, enfants ou conjoints sans profession, utilisent généralement comme numéro de sécurité sociale celui de leur père, mère ou conjoint. La CNIL craignant qu'il facilite les croisements de fichiers par les administrations, a longtemps exigé que le NIR fût cantonné à un usage unique : pour la sécurité sociale (et donc pour les échanges avec la sécurité sociale). Elle admet aujourd'hui qu'il puisse servir aussi d'Identifiant National de Santé (INS) pour indexer les dossiers médicaux notamment : cette évolution devrait être bientôt inscrite dans le code de la santé publique à l'article L. 1111-8 lorsque le projet de loi de modernisation de notre système de santé sera définitivement adopté.

ENCADRÉ 2 – DÉ-IDENTIFICATION, PSEUDONYMISATION, ANONYMISATION

Dé-identification et pseudonymisation, ces expressions un peu lourdes sont des quasi synonymes en ce sens que dans l'un et l'autre cas, la vraie identité de la personne (nom-prénoms, NIR…) est absente ou masquée. L'emploi d'un pseudonyme signifie en outre qu'on a remplacé la vraie identité par un identifiant conventionnel (souvent un « numéro d'anonymat ») qui dans un contexte donné désigne toujours la même personne afin de permettre un suivi longitudinal (suivi du parcours).

L'attribution d'un pseudonyme par un procédé qui interdit au gestionnaire des données de remonter lui-même au nom de la personne concernée (un chiffrement irréversible par exemple) était souvent appelée anonymisation et l'est parfois encore mais on sait mieux aujourd'hui que les jeux de données ainsi modifiés ne sont pas nécessairement anonymes ; c'est pourquoi il vaut mieux parler en ce cas de pseudonymisation.

Il est recommandé par ailleurs de diversifier les pseudonymes c'est-à-dire de ne pas toujours désigner la même personne par le même pseudonyme afin d'éviter qu'une ré-identification sur un jeu de données se propage à d'autres : le gestionnaire des données devra créer des pseudonymes ad hoc pour chaque utilisation des données.

Pour assurer l'anonymat des données il ne suffit pas de masquer les identités des personnes

Rappelons d'abord que les données de santé dont il s'agit ici sont des données destinées à des traitements statistiques. Certaines de ces données ont pu être recueillies directement pour cela (les données d'enquête épidémiologique ou les données de la recherche biomédicale par exemple) mais pour la plupart elles ont été recueillies initialement à d'autres fins, par exemple pour soigner des malades (ce sont les données issues de dossiers médicaux), ou pour rembourser des prestations ou pour financer des prestataires (ce sont les données dites médico-administratives comme celles recueillies par l'assurance maladie ou celles transmises par les hôpitaux dans le cadre de leur tarification à l'activité). A l'origine, ces données sont presque toujours nominatives (les dossiers médicaux et les feuilles de soins bien sûr pour soigner ou pour rembourser la bonne personne, et même les résumés de sortie hospitaliers) mais pour les réutiliser à des fins de connaissance, on s'efforce de les anonymiser.

Par hypothèse en effet, pour tous les traitements de données de santé autorisés que nous évoquons ici, l'identité des personnes concernées n'a pas d'importance puisqu'il s'agit d'établir des résultats généraux valant pour une population ou pour des segments de population. C'est pourquoi on peut et on doit remplacer les informations directement identifiantes (nom-prénoms, NIR, numéro de téléphone, adresses postales ou électroniques…) par un numéro d'ordre ou un « numéro d'anonymat » : un *pseudonyme* dont le lien avec l'identité d'origine doit rester secret, inconnu notamment des personnes, qu'il s'agisse des gestionnaires de la base ou qu'il s'agisse de tiers, qui accèdent aux données pour les traiter.

L'ignorance de l'identité ou l'ignorance du lien entre identité véritable et pseudonyme (l'impossibilité de remonter du pseudonyme à l'identité), est non seulement la condition de l'anonymat, c'en est la définition.

Bien entendu, il est le plus souvent nécessaire de relier les enregistrements différents relatifs à la même personne, pour suivre et analyser les parcours de soins et mettre en évidence des liens de causalité. Ce « chaînage » implique que la même identité soit traduite par le même pseudonyme.

Cela étant, même si les données sont dé-identifiées ou pseudonymisées et même s'il n'est pas possible aux utilisateurs de remonter d'un pseudonyme à l'identité, il n'en résulte pas nécessairement que la base ou le jeu de données soit anonyme. Le PMSI illustre bien cette difficulté. Responsable en 1991 de la *mission PMSI* au ministère de la santé, l'auteur de ces lignes avait négocié avec la CNIL la définition d'un *résumé de sortie anonyme* (RSA), sans le nom ni le NIR, ni les dates précises d'hospitalisation (seulement le mois de sortie et la durée du séjour), avec le mois et l'année de naissance mais sans le jour, et avec un code géographique de résidence correspondant à au moins mille habitants…

On était sincèrement persuadé à l'époque, du côté du ministère de la santé comme du côté de la CNIL, d'avoir ainsi défini un jeu de données parfaitement anonymes qu'on pourrait traiter et diffuser sans avoir d'autorisation à demander à quiconque. Cependant, avec la généralisation rapide du PMSI, on s'aperçut dès 1998 :

> qu'il était devenu possible de reconnaître des personnes dans la base, par recoupement, dès lors qu'on disposait sur elles d'informations assez banales (âge, sexe, code postal du domicile, dates d'hospitalisation -même approximatives- dans un

établissement hospitalier…). Si dans la base, devenue quasi exhaustive, une seule personne présentait les caractéristiques connues, on pouvait la reconnaître à coup sûr ;
> que c'était d'autant plus facile que l'établissement était petit ou que le patient venait de loin,
> et que c'était encore plus facile si on disposait, dans le jeu de données, du chaînage des hospitalisations successives des mêmes personnes (les parcours de soins présentent presque toujours des caractéristiques uniques).

Reconnaître une personne dans la base permet alors d'avoir accès à des informations particulièrement sensibles, en particulier les diagnostics motivant le ou les séjours.

A l'époque, le ministère de la santé et la CNIL en ont tiré la conclusion que ces données, bien que dé-identifiées et tenues jusque là pour anonymes, présentaient en réalité un caractère personnel (on disait alors qu'elles étaient « indirectement nominatives ») et ne devraient donc plus être diffusées et traitées sans l'autorisation de la CNIL. C'est l'origine de l'actuel chapitre X de la loi informatique et libertés.

Toutefois on a continué de considérer ces données comme « très indirectement nominatives[7] » et donc moins dangereuses pour la vie privée que celles obtenues par les chercheurs auprès des médecins (bien que ces dernières fussent elles aussi dé-identifiées dans la plupart des cas), de sorte que les données du PMSI ont été jusqu'en 2014 diffusées sur des supports aisément copiables (des CD-ROM), même si les utilisateurs devaient s'engager à ne pas le faire, et utilisées pour des fins dont l'intérêt public n'était pas toujours évident.

L'accès aux données du SNIIRAM présente des risques similaires même s'il a été jusqu'à présent plus limité : le rapport du Groupe de travail sur les risques de ré-identification, reproduit ci-après, signale notamment les risques de ré-identification de personnes dans l'Échantillon des bénéficiaires (si ces personnes ont été hospitalisées) et le défaut de traçabilité des accès pour les jeux de données extraits du SNIIRAM auxquels la CNIL a autorisé un accès.

Des jeux de données anonymes en accès libre : quels critères ?

Si les données individuelles contenues dans les bases de données médico-administratives comme le PMSI et le SNIIRAM (ou les causes de décès) présentent des risques de ré-identification, alors l'ouverture de ces données en vue de leur réutilisation ne consiste pas à prendre les données pour simplement les dévoiler en les mettant en ligne pour tous. Les données de santé anonymes que l'on mettra en ligne seront des résultats de traitements, résultats qu'il aura fallu au préalable produire à partir des données brutes individuelles.

Si en matière de données anonymes et gratuites il est évident que les organismes publics ayant pour mission de publier des données et des statistiques peuvent faire plus et mieux qu'aujourd'hui, cela pose quand même deux questions :

> La première est évoquée ici pour mémoire : quels genres de contenus et quels usages doit-on privilégier ? Traiter les données exige des ressources et oblige donc à des choix : il faudra susciter une expression des besoins[8] et gérer les priorités.
> L'autre question est comment s'assurer que les résultats ainsi produits sont bien anonymes. Différentes solutions sont évoquées dans le présent dossier et dans la littérature. L'une consiste à traiter les données pour en tirer des résultats statistiques agrégés sous la forme de tableaux (ou de « cubes de données » d'où l'utilisateur peut extraire et manipuler lui-même les dimensions et séries qui l'intéressent). Une autre consiste à mettre en ligne des échantillons. Une troisième est de produire des fichiers appauvris[9].

Des méthodes d'évaluation du risque de ré-identification sont présentées dans le rapport du groupe de travail de 2014 et, plus longuement, dans les articles suivants de ce *Dossier*. Mais il s'agit encore d'un domaine en devenir. Le cas le plus simple est celui des statistiques agrégées (où il faut éviter qu'il y ait moins de N personnes dans une case, la valeur convenable de N variant toutefois selon les interlocuteurs). Pour les cas plus complexes de jeux de données individuelles,

[7] Commission nationale de l'informatique et des libertés, *Rapport d'activité 1999*, page 144.

[8] Ce sera une des missions de l'Institut national des données de santé, dont la création est prévue par l'article 47 du projet de loi de modernisation de notre système de santé.

[9] Nous ne décrirons pas ici la quatrième solution qui consiste à brouiller les données en les modifiant ou en y ajoutant de fausses données qui sont supposées ne pas modifier les résultats des traitements.

dites aussi granulaires, il y a diverses façons de mesurer le risque mais dans le rapport du groupe de travail déjà cité on distingue un aspect « dénombrement », un aspect « classement » et un aspect « évaluation à dire d'experts » :

- Le dénombrement d'abord :
  - on compte le nombre N de personnes présentant un même ensemble de caractéristiques[10] (par exemple sexe, âge, domicile, lieux et dates d'hospitalisation avec ou sans chaînage du parcours). Toutes les personnes de la base présentant des caractéristiques uniques (N=1) sont ré-identifiables (par qui connaît ou peut connaître ces caractéristiques) ;
  - on compte aussi (parmi N personnes présentant les mêmes caractéristiques) le nombre de maladies différentes dont ces N personnes sont atteintes. Si elles ont toutes une maladie en commun, on sait de quoi la personne recherchée est atteinte sans qu'elle soit identifiable. On peut aussi raisonner en termes de probabilités (si « presque tous » ont la même maladie) ou considérer que si toutes ces personnes sont atteintes de maladies graves, c'est déjà une information confidentielle[11] ;

- Le classement ensuite : on classera les caractéristiques, appelées aussi « quasi-identifiants » (sexe, âge, hospitalisations, arrêts de travail, consultations médicales, examens biologiques…), selon leur degré de notoriété (certaines informations sont bien connues et peuvent se retrouver dans des bases de données ou avoir été mentionnées par la personne elle-même sur les réseaux sociaux, d'autres sont ignorées ou vite oubliées et présentent pour cette raison un risque moindre voire quasi nul) ;

- l'évaluation à dire d'experts enfin : évaluation de la probabilité non seulement que cela soit faisable mais que des données confidentielles soient effectivement identifiées par recoupement puis divulguées ou utilisées pour nuire. C'est la probabilité que ces données tombent en de mauvaises mains, puis que des organisations ou des individus[12] cherchent à en faire mauvais usage malgré les engagements pris, malgré le risque d'être découvert et malgré l'opprobre ou les sanctions encourues.

Avec les valeurs ci-dessus on peut évaluer la *probabilité* de divulgation.

On peut évaluer parallèlement, là aussi à dire d'experts, *l'impact* d'une divulgation, c'est-à-dire sa gravité pour les personnes considérées, si cela se produit, ou si elles en sont menacées. On en déduit le *risque* :

Risque = Probabilité x Impact

Il existe différentes méthodes pour évaluer le risque de ré-identification qui sont évoquées ailleurs dans ce *Dossier* mais il convient en tout cas de bien comprendre ce qu'on mesure. Pour prendre trois exemples simplistes, ce n'est pas du tout la même chose de dire :

- Sur dix individus dans la base de données X, il y en a toujours au moins un qui présente des caractéristiques uniques et faciles à connaître et qui est donc ré-identifiable à coup sûr ;
- Dans la base Y il y a toujours au moins dix personnes qui présentent les mêmes caractéristiques (âge, sexe, dates et lieux d'hospitalisation) ;
- Si dans un échantillon aléatoire Z, au 1/10, quelques personnes présentent des caractéristiques uniques, je peux penser qu'il y a en moyenne, pour chacune d'elles, neuf autres personnes qui présentent les mêmes caractéristiques dans la base complète.

[10] Ces caractéristiques - ou quasi identifiants - peuvent être plus ou moins précises (date de naissance en jours mois années > en mois et années > en années > en tranche d'âges). On peut ainsi les appauvrir ou en supprimer certaines (par exemple supprimer les codes géographiques de résidence ou les numéros d'établissement) en fonction des types de traitements que l'on veut rendre possibles. Les données médicales, telles que diagnostics ou actes ou médicaments révélateurs d'une maladie, ne sont généralement pas considérées comme identifiantes puisque, suppose-t-on, c'est justement l'information que recherche l'attaquant ; mais, si les épisodes de soins sont chaînés dans le jeu de données mis à disposition, un séjour hospitalier pour accouchement, qui est généralement facile à reconnaître, peut révéler les motifs d'autres hospitalisations de la même personne.

[11] A vrai dire, si l'attaquant sait déjà où et quand la personne a été hospitalisée, il sait qu'elle a un problème de santé...

[12] Le groupe de travail a considéré que les grandes institutions étaient dissuadées assez efficacement (de cibler illégalement des personnes) par la menace de sanctions ou le risque pour leur image, parce qu'il est difficile de garder durablement secrètes des pratiques de ce genre. Le danger viendrait plutôt d'autres individus (proches, collègues, ennemis politiques ou personnels…) et ne serait pas limité aux personnalités connues.

Dans le premier cas, rendre la base X accessible à tous dévoile potentiellement la vie privée d'une personne sur dix et le risque, s'il s'agit de données de santé, n'est pas acceptable : on ne peut rendre cette base accessible qu'à des chercheurs ou assimilés, si c'est très utile : sous des conditions restrictives donc et en s'entourant de protections adéquates.

Dans le second et le troisième cas, où on ne peut pas affirmer avec certitude qu'on a reconnu une personne, le décideur peut choisir de classer ces bases comme anonymes et de les mettre en accès libre, au même titre que des bases comportant seulement des données agrégées (où le risque de ré-identification peut être plus facilement écarté).

Toutefois la décision sera souvent difficile à prendre si on considère que :

- dans un jeu de données individuelles, il est facile de mesurer le « K-anonymat » (faire en sorte qu'il n'y ait jamais moins de K personnes présentant des caractéristiques semblables dans la base) mais il est plus difficile de mesurer la « L-diversité » (au moins L personnes ayant des maladies différentes dans un groupe de personnes ayant par ailleurs des caractéristiques semblables) ;
- surtout, il est difficile de se mettre d'accord sur des valeurs convenables pour K et L ; à première vue, il faut que K = 1 pour être sûr de ré-identifier une personne mais si K est petit, un attaquant obstiné pourra identifier la personne par élimination[13] ; et il y aura aussi plus de chances que ces K personnes soient atteintes de la même maladie…
- le risque peut être aussi indirect, lorsqu'on produit des jeux de données appauvris qui, pris individuellement, ne présentent aucun risque de ré-identification mais qui, pris ensemble, permettent de reconstituer les jeux d'origine[14] ;
- la protection que confère un échantillonnage n'est efficace que si l'attaquant ignore la règle de tirage (s'il la connaît, l'échantillon redevient une base exhaustive[15]) ; cette protection conférée par l'échantillonnage ne suffit pas non plus s'il est notoire que certaines sous-populations ne comportent que des personnes à caractéristiques uniques (par exemple tous les patients hospitalisés plusieurs fois dans l'année). Plus généralement, si les données individuelles sont chaînées en parcours de soins, et si les parcours sont détaillés (avec dates, lieux et natures de soins), il y a de fortes chances que chaque parcours soit unique dans l'échantillon mais aussi dans la base.

Chacun peut mettre en ligne des jeux de données qu'il juge anonymes, mais à ses risques et périls : si ces jeux de données résultent de l'appauvrissement de données personnelles en vue de leur publication, ce traitement (ou le programme de traitements dont il fait partie) doit être soumis à la CNIL. Cette dernière ne délivre pas toutefois de certificat d'anonymat... sauf dans les cas, peu fréquents où les trois conditions suffisantes mentionnées par le groupe des 29 CNIL européennes, dit G29[16], sont réunies. Cela étant, méfions nous d'une pensée magique qui considèrerait que rien n'est anonyme et qu'on trouvera toujours sur le web les informations et des méthodes permettant de tout ré-identifier : c'est faux dans le cas de bases administratives structurées où les quasi identifiants (dates et lieux notamment) sont en nombre fini. Toutefois pour prendre en ce domaine mouvant des décisions toujours assurées, il faudra davantage de travaux et davantage de consensus, impliquant tous les acteurs : la CNIL, la mission Etalab[17], les gestionnaires du futur Système national des données de santé, l'INSEE, le Conseil national de l'information statistique, les services ministériels, le futur Institut national des données de santé…

Dans le projet de loi de modernisation de notre système de santé voté par l'Assemblée nationale le 14 avril 2015, il a été ajouté à l'article 8 de la loi informatique et libertés un V ainsi rédigé : « *V. – Les jeux de données issu[e]s des traitements comportant des données à caractère personnel mentionnées au I du présent article ne peuvent être mis à la disposition du public qu'après avoir fait l'objet d'une anonymisation complète des données personnelles qu'ils contiennent. Le responsable du traitement tient à la disposition de la Commission nationale de l'informatique et des libertés les procédés*

[13] Certains experts font valoir que si K=2 et que l'une de ces deux personnes a accès à la base, elle pourra identifier l'autre et sa maladie. Plus généralement, nombreux sont ceux qui estiment qu'une chance sur 2 ou 3… voire sur 5 ou 10, de reconnaître la personne, c'est encore trop.

[14] C'est notamment le cas pour le PMSI en raison du caractère unique de la série des données médicales détaillées (diagnostics, actes etc.) dans chaque résumé de séjour : ces informations ne sont pas identifiantes en elles-mêmes mais la série unique constitue une empreinte propre à chaque séjour : si on la conserve telle quelle dans plusieurs jeux de données appauvris par ailleurs, on pourra reconstituer tout ou partie des jeux d'origine… Un problème du même genre peut se poser si la même personne est toujours désignée par le même pseudonyme (dit aussi numéro d'anonymat) dans les jeux de données mis à disposition : la vulnérabilité d'un jeu retentit sur les autres.

[15] Si « l'attaquant » sait que l'échantillon est constitué de toutes les personnes nées le 15 de chaque mois pair, que sa « cible » est née un 15 février et qu'elle est la seule dans la base à présenter les caractéristiques connues (code géographique de résidence, sexe, année de naissance, hospitalisation dans tel établissement tel mois), alors la « cible » est ré-identifiée à coup sûr dans la base exhaustive des personnes nées le 15 d'un mois pair !

[16] Voir http://www.cnil.fr/linstitution/actualite/article/article/le-g-29-publie-un-avis-sur-les-techniques-danonymisation/

[17] La mission *Etalab* est le service du Premier ministre (au sein du Secrétariat général à la modernisation de l'action publique) chargé de promouvoir l'*open data*.

mis en œuvre pour garantir cette anonymisation. La commission peut également reconnaître la conformité à la présente loi de toute méthodologie générale ou de tout procédé d'anonymisation. »

Dans le doute, ou si le risque de ré-identification semble réduit (comme c'est cas pour certains échantillons notamment), une solution intermédiaire consiste à mettre à disposition ces jeux de données mais en prenant des précautions complémentaires, comme le fait le Réseau *Quételet*[18] en réservant certains jeux à des acteurs de confiance (des chercheurs en sciences humaines et sociales dans le cas du réseau Quételet).

C'est cette solution intermédiaire ou « zone grise », que vise une autre disposition nouvelle introduite par le projet de loi. Cette disposition ajoutée à la fin de l'article 54 de la loi informatique et libertés est rédigée ainsi : « *IV bis (nouveau). – Des jeux de données agrégées ou des échantillons, issus des traitements des données de santé à caractère personnel pour des finalités et dans des conditions reconnues conformes à la présente loi par la Commission nationale de l'informatique et des libertés, peuvent faire l'objet d'une mise à disposition, dans des conditions préalablement homologuées par la commission, sans que l'autorisation prévue au I du présent article soit nécessairement requise.* »

On peut penser que la CNIL se montrera exigeante à la fois sur le sérieux du ou des organismes[19] ainsi autorisés à mettre des jeux de données à disposition, sur la nature des jeux de données concernés (échantillons tirés du SNDS, ou de certaines cohortes, présentant un risque de ré-identification réduit) et sur les précautions à prendre (contrôle des demandeurs, contrôle des demandes, contrôle des modalités techniques d'accès).

Des jeux de données comportant des risques de ré-identification, rendus accessibles, si c'est pour de bonnes raisons et avec de bonnes protections

A l'INSEE, avant 2008, l'accès aux données individuelles sur les personnes physiques, pourtant dé-identifiées, était réservé aux agents de la statistique publique. Les données fiscales ont été elles aussi réservées à l'administration et interdites aux chercheurs jusqu'à 2014. De même et sauf exception, l'accès des chercheurs à des données médicales collectées auprès des médecins suppose que le patient, dûment informé, ne s'y soit pas opposé et que la CNIL ait donné son autorisation[20]. Bref, en France comme dans d'autres pays où la protection de la vie privée n'est pas considérée comme une « notion dépassée » (pour reprendre le mot de Mark Zuckerberg, le fondateur de *Facebook*), la règle est l'interdiction, sauf consentement de la personne concernée ; et l'accès sans consentement était l'exception.

En l'absence d'accès des tiers, seuls les experts de l'administration pouvaient réaliser des recherches ou des études sur les données recueillies par les administrations. Le raisonnement sous-jacent était que plus le secret est partagé, moins il y a de secret (ce qui est vrai) et que les experts de l'administration respecteraient mieux le secret professionnel (ce qui est discutable). Toutefois :

- Cela interdisait l'évaluation contradictoire des politiques publiques ;
- et comme les ressources des administrations sont limitées, cela bridait aussi l'utilisation des données.

Ce sont principalement les chercheurs qui ont bénéficié de l'ouverture progressive des données à caractère personnel, notamment dans le domaine de la santé. Une part du débat sur le projet de règlement européen relatif à la protection des données personnelles a d'ailleurs porté sur les exceptions à l'exigence du consentement quand il s'agit de statistique et de recherche : on se dirige vers un principe d'accès aux données à caractère personnel dès lors qu'elles sont « pseudonymisées » et font l'objet de mesures de protection appropriées, quand les études et recherches (y compris la recherche appliquée) ne peuvent pas être réalisées avec des données anonymes.

[18] Voir http://www.reseau-quetelet.cnrs.fr/spip/

[19] Outre la CNAMTS ou l'ATIH, le futur INDS pourrait jouer ce rôle pour les échantillons du SNDS. L'INSERM envisage aussi de mettre à disposition une plateforme à cet effet, notamment pour des données issues de certaines cohortes.

[20] Voir l'article 57 de la loi informatique et libertés, dans sa version actuelle et tel que propose de le modifier l'article 47 du projet de loi de modernisation de notre système de santé.

Dans des bases où les données d'identification directe (noms, NIR…) ont été remplacées par des pseudonymes, il est pratiquement impossible de demander leur accord aux personnes concernées à chaque fois que l'on veut utiliser leurs données pour une étude ou une recherche. Et à supposer que ce soit possible, ce serait très coûteux et surtout contraire au but recherché : pour demander à quelqu'un s'il accepte un risque, même infime, d'être ré-identifié, on commencerait par le ré-identifier ! C'est pour la même raison (la nécessité de ré-identifier les personnes) que les droits d'accès et de rectification devraient être exclus dans de telles bases.

Pour ce qui est du droit à l'information des personnes, en revanche, le projet de loi de modernisation de notre système de santé[21] a prévu d'ajouter à l'article 57 de la loi informatique et libertés un III ainsi rédigé : « *III. – Quand la recherche, l'étude ou l'évaluation faisant l'objet de la demande utilise des données de santé à caractère personnel non directement identifiantes recueillies à titre obligatoire et destinées aux services ou aux établissements de l'État ou aux organismes de sécurité sociale, l'information des personnes concernées quant à la réutilisation possible de ces données, à des fins de recherche, d'étude ou d'évaluation, est assurée selon des modalités définies par décret en Conseil d'État, pris après avis de la Commission nationale de l'informatique et des libertés.* ». Donc le patient doit être informé que des données, si elles sont recueillies à titre obligatoire par une administration pour les besoins de sa mission (rembourser les assurés, financer les hôpitaux etc.), peuvent ensuite être réutilisées à des fins de recherche, d'étude ou d'évaluation, dans des conditions qui devront protéger le secret de la vie privée. Un droit d'opposition, s'il existe, ne peut alors être conçu que sur le mode du tout ou rien : on ne peut pas demander son avis au patient à chaque utilisation mais on peut supprimer ses données de la base (la base destinée à être mise à la disposition de tiers) s'il le demande. Techniquement, cet « *opt out* » est faisable. Cependant les textes créant le traitement de données peuvent aussi conditionner ce retrait à un motif légitime voire l'interdire. Le raisonnement sous-jacent est que ces données seront recueillies et utilisées à des fins d'intérêt public en contrepartie d'une prise en charge des soins par la collectivité et que toutes les dispositions sont prises par les pouvoirs publics (ou doivent l'être) pour les protéger des indiscrétions[22].

Quand l'accès à ces données à caractère personnel ne repose pas ou pas uniquement sur le consentement, il faut de bonnes raisons et de bonnes protections. Les bonnes raisons (raisons d'intérêt public), ce sont les finalités de recherche, d'étude ou d'évaluation, que la CNIL apprécie en s'appuyant pour ce faire sur les avis qu'elle reçoit, quand c'est nécessaire[23]. Il est exclu en tout cas de faire courir un risque, même infime, d'atteinte à la vie privée, pour des fins qui seraient de nature essentiellement privées ou commerciales.

Ensuite et si, comme on l'a vu, l'emploi de pseudonymes ne suffit pas à rendre une base de données anonyme, la question de savoir comment les données sont effectivement protégées des indiscrétions est cruciale. Certains font valoir qu'il suffirait d'un arsenal de sanctions dissuasives en cas de violations du secret professionnel ; mais encore faut-il que les indiscrétions soient repérables pour que des sanctions soient crédibles. Or la diffusion des données du PMSI sur des supports autonomes (CD) ou la possibilité d'obtenir des extraits du SNIIRAM sur son poste de travail, ont rendu possibles des copies sauvages de telle sorte que nul ne sait avec certitude où sont les données et qui en dispose.

Un moyen technique de limiter les risques de fuite est de faire en sorte que les données ne puissent être traitées ailleurs que dans un espace sécurisé : une « bulle ». Comme il est malcommode d'avoir pour cela à se rendre auprès d'un guichet physique, la meilleure solution est l'accès à distance (*remote access*) sur un serveur, où sont conservées des traces de toutes les données « sorties » : on a le droit en effet de « sortir » les résultats anonymes des traitements réalisés dans la bulle mais pas de télécharger des jeux de données permettant la ré-identification des personnes.

Il existe une autre solution encore plus sécurisée, mais aussi plus contraignante, utilisée en Australie et dans une moindre mesure en Allemagne, qui consiste à permettre le traitement des données à distance par mon programme sans que moi, chercheur, je puisse les voir (on parle en ce cas de *remote execution*). La France a préféré, pour les données de l'INSEE,

[21] Il s'agit toujours ici de l'article 47 du projet de loi de modernisation de notre système de santé, dans la version votée par l'Assemblée nationale.

[22] Certains craignent en outre que les données puissent être prises en otage par des mouvements sociaux qui appelleraient les uns ou les autres à refuser la réutilisation de leurs données, au risque que ces dernières perdent leur représentativité et donc leur utilité.

[23] Pour la plupart des demandes d'accès aux données, correspondant à des usages standard, la question des finalités d'intérêt public ne se pose pas (car elle est évidente ou a déjà été tranchée).

pour les données fiscales et bientôt pour les données de santé[24], la solution de l'accès sécurisé à distance (*remote access*) qui fait confiance aux personnes autorisées à accéder aux données mais à elles seulement. Ces privilégiés peuvent en effet "voir" les données auxquelles la CNIL ou la réglementation leur ont donné accès, et effectuer dans la « bulle » distante tous les traitements qu'ils souhaitent. Ils s'engagent seulement à ne pas photographier les données vues à l'écran ni à chercher à ré-identifier des personnes, ni à télécharger des données à caractère personnel (les données qu'ils téléchargent sont enregistrées). Ce que l'on cherche à empêcher ainsi c'est l'accès de personnes non habilitées, via des copies réalisées pour des raisons en apparence innocentes (travailler chez soi, aider un collègue…) mais qui empêche d'identifier l'origine d'une divulgation de données personnelles. Le système de bulle, évidemment plus contraignant que la diffusion sur des supports autonomes, vise à éviter la circulation de données intraçables. Le fait qu'il n'ait pas été mis en vigueur plus tôt pour le PMSI peut faire apparaître sa mise en place aujourd'hui comme un recul[25] ; c'est sans doute regrettable mais la divulgation de données de santé à caractère personnel et la perte de confiance dans le système des statistiques publiques qui en résulterait le seraient plus encore.

Ce mode d'accès sécurisé à distance pourrait être aussi une des précautions exigées pour l'accès aux données dites « grises » mentionnées plus haut (les données à risque de ré-identification réduit).

Signalons enfin qu'il existe un autre mode d'accès intermédiaire entre open data et accès restreint : cela consiste à concevoir puis mettre à disposition des batteries de requêtes préalablement définies et validées qui réalisent des traitements sur une base de données à caractère personnel mais ne peuvent produire que des résultats anonymes, sous la forme de tableaux ou de cartes, à l'usage d'un public large. Cela pourrait être le cas, par exemple, d'une visualisation interactive des données : un développeur créé une interface graphique ; cette dernière génère automatiquement des requêtes (à la demande de l'utilisateur) qui sont envoyées à la base, les résultats sont ensuite présentés graphiquement sous forme agrégée et anonyme. Une telle application nécessiterait en outre la possibilité de se connecter à la base via une interface de programmation dont les temps de réponse seraient garantis. C'est ce que font déjà en quelque sorte les producteurs de statistiques publiques (traiter les données détaillées pour en tirer des statistiques anonymes à l'usage du public) mais c'est aussi ce que font des cabinets d'études privés réalisant des traitements réservés à leurs clients. La question de savoir si ce dernier usage privatif des données de santé à caractère personnel est conforme à l'intérêt public peut se poser. La CNIL aura à en décider mais elle pourrait prendre en compte deux éléments :

> La finalité des statistiques produites (par exemple, la bonne gestion des établissements de santé est a priori d'intérêt public) ;
> l'engagement éventuel de rendre publique la méthodologie après un délai raisonnable.

Éclairage sur les risques réels ou imaginaires liés au NIR et sur les moyens de s'en prémunir

La réforme en cours vise à faciliter l'utilisation des données publiques de santé et à multiplier les études et les recherches à partir des bases de données particulièrement riches dont dispose notre pays. A cet égard, et bien que cela puisse sembler une mesure technique, la simplification des procédures d'appariement entre bases de données est sans doute l'avancée la plus significative.

[24] Pour les données de l'INSEE, les données du ministère de l'agriculture, les données de la CNAV et les données fiscales, les administrations ont fait appel au Centre d'accès sécurisé aux données (CASD), une direction du Groupement des écoles nationales d'économie et de statistique (GENES). L'ATIH à l'heure où nous rédigeons ces lignes a mis en place une procédure de marché adaptée pour choisir son prestataire.

[25] Ce recul sera plus que compensé par la mise en ligne de jeux de données anonymes plus nombreux, par l'ouverture des données du SNIIRAM à de nouveaux utilisateurs et par la simplification et l'unification des procédures d'accès. A cet égard il faut noter que la fusion des chapitres IX et X de la loi informatique et libertés (le chapitre X était celui qui organisait l'accès aux bases de données dé-identifiées comme le PMSI et le SNIIRAM) n'empêchera pas la CNIL d'autoriser des programmes d'études (et non seulement des projets d'études individuels), comme elle le faisait jusqu'à présent. Le projet de loi a même prévu de donner un fondement juridique plus solide à cette pratique en ajoutant à l'article 54 de la loi informatique et libertés un V ainsi rédigé : « *V. – La Commission peut, par décision unique, délivrer à un même demandeur une autorisation pour des traitements répondant à une même finalité, portant sur des catégories de données identiques et ayant des catégories de destinataires identiques.* »

Il faut d'abord bien comprendre que la façon la plus simple d'enrichir les données dont on dispose, pour étendre le champ des études possibles, n'est pas nécessairement de recueillir de nouvelles données mais d'utiliser celles qui existent déjà ailleurs, dans d'autres bases de données. C'est ce que montre l'article de ce *Dossier* consacré aux appariements.

Il y a plusieurs façons d'apparier ses données avec d'autres déjà recueillies par ailleurs, par exemple comparer les deux populations pour corriger d'éventuels biais de sélection dans celle qu'on étudie ; mais l'appariement dont il est surtout question ici est celui qui permet de trouver d'autres données relatives aux mêmes personnes dans l'autre base. Voici deux exemples de ce type d'appariement :

- apparier les données d'une enquête ou d'une cohorte de patients avec les données de l'assurance maladie, pour disposer d'une information fiable sur les médicaments que les personnes (interrogées au cours de l'enquête ou membres de la cohorte) ont précédemment achetés en pharmacie, si elles en ont demandé le remboursement ;
- apparier les données du SNIIRAM-PMSI avec les données du fichier national des carrières de la caisse nationale d'assurance vieillesse, pour évaluer l'impact des professions exercées sur la santé.

Donc l'appariement est très utile pour la recherche et les études en santé… mais il est aussi très difficile à réaliser en raison de l'obstacle qu'oppose la rédaction actuelle (avant la loi de modernisation de notre système de santé) de la loi informatique et libertés et notamment de son article 27 qui exige un décret en Conseil d'État pour autoriser tout usage du NIR par une administration ou un organisme chargé d'une mission de service public. Or, comme on va le voir, l'utilisation du NIR est bien souvent indispensable pour réaliser un appariement de qualité entre des données non nominatives (données dites dé-identifiées ou pseudonymisées) qui sont précisément celles qui servent aux travaux statistiques et à la recherche.

Un décret, c'est-à-dire un texte signé du Premier ministre et des ministres concernés, pris sur avis du Conseil d'État, implique une procédure longue mobilisant les services d'un ou plusieurs ministères. C'est faisable pour une administration, mais pour l'immense majorité des chercheurs, c'est simplement impossible à obtenir.

Pourtant le NIR est l'équivalent strict de l'ensemble « nom+prénoms+date et lieu de naissance » (sans les homonymies il est vrai, et avec moins d'erreurs de saisie). Utile pour des « usages métier » où il importe d'éviter toute confusion entre les personnes[26], le NIR a été critiqué parce qu'il est signifiant (il indique le sexe, la date et le lieu[27] de naissance) ; mais il était surtout vu, il y a 40 ans, comme l'instrument indispensable pour croiser des fichiers nominatifs, rendant ainsi possible, croyait-on, un fichage généralisé par l'État. Aujourd'hui, à l'heure des moteurs de recherche, il y a bien longtemps qu'un index numérique commun n'est plus nécessaire pour croiser des fichiers nominatifs (nominatifs comme le sont les fichiers « métier » des administrations, par exemple les fichiers du fisc ou ceux de la police ou des caisses d'assurance maladie). Pourtant le symbole « *SAFARI, la chasse aux Français*[28] » est resté, si bien que le NIR :

- continue d'être jusqu'ici cantonné par la CNIL à un usage « pour la sécurité sociale » (et donc pour les échanges avec la sécurité sociale)[29] ;
- et continue de « bénéficier » d'un encadrement législatif qui le distingue des autres données directement identifiantes, telles que les noms-prénoms, adresses ou numéros de téléphone, curieusement moins encadrées alors qu'elles permettent aussi facilement de croiser des fichiers nominatifs avec un taux de réussite souvent proche de 100 %.

Encadrer à la fois l'usage du NIR et le croisement des fichiers à caractère personnel, l'un étant supposé être le moyen de l'autre, c'est redondant, inefficace pour les libertés et bloquant pour la recherche. L'encadrement du NIR, on l'a vu, est inefficace pour empêcher le croisement illicite de fichiers nominatifs. En revanche, il est bloquant dans un cas : celui des bases de données dé-identifiées utilisant des pseudonymes obtenus par chiffrement du NIR… ce qui est précisément le cas des bases de données du secteur sanitaire et social.

[26] Le premier usage du NIR par la sécurité sociale a été le suivi puis la reconstitution des carrières, pour le calcul de droits à la retraite acquis tout au long de la vie.

[27] La mention « 99 » signifie « naissance hors du territoire national ».

[28] La tribune libre dans le journal *Le Monde* intitulée *Safari ou la chasse aux Français* critiquait en 1974 un projet, au nom malheureux de *SAFARI*, qui prévoyait le croisement des fichiers des administrations et qui fut analysé comme une volonté de fichage généralisé par l'État, dont le NIR aurait été le moyen. Le mouvement d'opinion qui s'en suivit est à l'origine de la loi informatique et libertés du 6 janvier 1978, modèle des lois semblables que d'autres pays adoptèrent ensuite et qui inspira la directive européenne de novembre 1995.

[29] Comme on l'a vu, la CNIL a assoupli sur ce point sa doctrine et le projet de loi de modernisation de notre système de santé prévoit d'autoriser l'usage du NIR comme Identifiant national de santé.

En effet les bases destinées à la connaissance peuvent et doivent être dé-identifiées ou pseudonymisées : NIR et nom étant supprimés, ils sont remplacés par un pseudonyme ou numéro d'anonymat qui dans une base de données doit être toujours le même pour une même personne afin de permettre un suivi longitudinal (c'est le principe du chaînage). Si le NIR est présent dans la base initiale, une manière simple de produire ce numéro d'anonymat est un chiffrement irréversible du NIR. C'est la méthode recommandée par la CNIL et elle a été utilisée pour le SNIIRAM et le PMSI (puis pour les données sur le handicap). Les principes généraux de cette méthode, dans une version dont la CNAMTS est propriétaire, sont d'ailleurs présentés dans ce *Dossier* : on peut lire à ce sujet en annexe l'article de Gilles Trouessin qui contribua naguère à sa mise au point.

La conclusion du raisonnement qui précède est d'une grande simplicité :

> Les chercheurs et assimilés, dans la sphère sociale, sont les seuls que l'encadrement du NIR gêne aujourd'hui pour « croiser des fichiers », parce que, de leurs fichiers, on a retiré le nom et les autres informations directement identifiantes et qu'il ne reste que le NIR chiffré comme unique identifiant (numéro d'anonymat, nécessaire au chaînage) ;
> La disposition conditionnant l'emploi du NIR à un décret en Conseil d'État a pour principal, voire pour seul effet de stériliser la recherche française dans les domaines sanitaire et social.

L'article 47 du projet de loi de modernisation de notre système de santé modifie à la marge l'article 27 de la loi informatique et libertés, en confiant à la CNIL elle-même le soin d'autoriser les usages du NIR lorsqu'ils s'inscrivent dans le cadre d'une recherche, d'une étude ou d'une évaluation en santé : si la loi est adoptée, un décret en Conseil d'État ne sera donc plus exigé dans ce cas. Cette modification modeste contribuera à une augmentation très significative des travaux publiés en France dans des domaines essentiels comme l'épidémiologie, la pharmaco-épidémiologie et le fonctionnement des services de santé. Plus généralement, nous faisons le pari que les données de santé françaises, plus faciles d'accès, plus faciles à enrichir par appariement, mieux documentées, serviront à une foule de travaux et d'études éclairantes dans des domaines variés, commanditées par des acteurs nombreux, publics ou privés, français ou étrangers et contribueront au développement d'une expertise française de grande valeur.

ENCADRÉ 3 - NIR À L'ENDROIT, NIR À L'ENVERS ET TIERS DE CONFIANCE

L'usage du NIR restera sous le contrôle de la CNIL. Il n'est pas question d'autoriser tout un chacun à employer le NIR (pas plus que le nom !) pour fouiller dans des bases de données. Cependant, pour comprendre la manière dont cela se passera, il faut distinguer deux cas : selon que l'on va « à l'endroit » du NIR au pseudonyme ou « à l'envers » du pseudonyme au NIR. C'est dans le second cas seulement qu'il sera obligatoire de passer par le tiers de confiance national.

Le prototype d'un usage « du NIR vers le pseudonyme » est celui, déjà cité, d'une enquête ou d'une cohorte mise en place par des chercheurs qui voudraient aussi récupérer dans la base de l'assurance maladie des données sur les consommations de médicaments de leur population. Dans ce cas, les chercheurs (ou certains membres de l'équipe de recherche, ou un prestataire de l'équipe de recherche) connaissent déjà par hypothèse les noms et les coordonnées de ces personnes, et ces dernières, en acceptant de participer à l'enquête ou de devenir membres d'une cohorte, ont aussi accepté de répondre à des questions sur leur santé ou de se soumettre à des examens de santé… Bref, ayant confié à l'équipe de recherche leur nom, leurs coordonnées et leur état de santé, ces personnes accepteront aussi de confier leur NIR pour que les chercheurs aillent récupérer auprès de l'assurance maladie des informations sur leur consommation de soins passée. Pour ce faire, les chercheurs ou un agent spécialisé au sein de leur équipe ou un prestataire « tiers de confiance » *de leur choix* se présenteront à l'assurance maladie avec un fichier composé des NIR et de numéros d'ordre[30] des personnes. Au sein de l'assurance maladie, le même dispositif qui permet le chiffrement des NIR pour l'alimentation du SNIIRAM permettra de transmettre aux gestionnaires du SNIIRAM un fichier constitué des NIR chiffrés et des numéros d'ordre correspondant. Les gestionnaires de la base SNIIRAM n'auront plus alors qu'à retourner vers les chercheurs les consommations passées de ces personnes identifiées par leur numéro d'ordre… Dans un tel circuit, qu'il faudra formaliser en bonnes pratiques admises par la CNIL, les données d'identification ne doivent jamais se trouver dans la même base et dans les mêmes mains que les données de santé. Ce circuit exige une séparation entre les personnes qui gèrent les identifiants et celles qui gèrent les données. Selon les cas, cette séparation peut être organisée au sein d'une institution (le laboratoire de recherche, la CNAMTS…) ou bien on peut faire appel à des tiers de confiance ad hoc (qui ne sont pas nécessairement le tiers de confiance national).

On peut donner deux exemples d'un usage allant du pseudonyme vers le NIR : le cas d'un appariement entre deux bases utilisant des pseudonymes différents tous deux dérivés du NIR (par exemple les données de santé et les données sur le handicap), et le cas où l'on veut prévenir une personne dont une étude sur les données de santé a montré qu'elle courait un risque. Dans ces cas, la seule possibilité est de « remonter au NIR ». Or, comme on l'a vu, il est techniquement impossible aux gestionnaires d'une base de données de santé de remonter d'un pseudonyme (un NIR chiffré) vers le NIR d'origine. En revanche, le tiers de confiance national sera lui autorisé à chiffrer tous les NIR à l'aide des clés secrètes ad hoc qu'il sera autorisé à

[30] Il est de bonne pratique que chaque personne ayant répondu à l'enquête ou chaque membre de la cohorte soit désigné par un numéro d'ordre dans la base constituée par les chercheurs. Les vrais noms et coordonnées sont conservés séparément : il serait inutile et imprudent de les conserver parmi les données utilisées pour la recherche.

détenir et à générer ainsi une table de correspondance entre NIR et pseudonymes. Les services de ce tiers de confiance national (qui sera nécessairement unique[31]) seront requis dans de tels cas. C'est ce que prévoit le nouvel article L. 1461-5 du code de la santé publique tel qu'il résulte de l'article 47 du projet de loi de santé, article L. 1461-5 dont il faut admettre que la rédaction est complexe :

« Art. L. 1461-5. – I. – Le système national des données de santé ne contient ni les noms et prénoms des personnes, ni leur numéro d'inscription au répertoire national d'identification des personnes physiques, ni leur adresse. Les numéros d'identification des professionnels de santé sont conservés et gérés séparément des autres données.

« II. – Un décret pris en Conseil d'État après avis de la Commission nationale de l'informatique et des libertés détermine les données à caractère personnel qui, en raison du risque d'identification directe des personnes concernées, sont confiées à un organisme distinct du responsable du système national des données de santé et des responsables des traitements.

« Cet organisme est seul habilité à détenir le dispositif de correspondance permettant de ré-identifier les personnes à partir des données du système national des données de santé. Il assure la sécurité de ce dispositif.

« III. – La Commission nationale de l'informatique et des libertés peut autoriser l'accès aux données détenues par l'organisme mentionné au II du présent article, dans les conditions prévues par la loi n° 78-17 du 6 janvier 1978 précitée, quand il est nécessaire :

« 1° Pour avertir une personne d'un risque sanitaire grave auquel elle est exposée ou pour lui proposer de participer à une recherche ;

« 2° Pour la réalisation d'un traitement à des fins de recherche, d'étude ou d'évaluation si le recours à ces données est nécessaire, sans solution alternative, à la finalité du traitement et proportionné aux résultats attendus. »

[31] Ont vocation à se porter candidat à être le tiers de confiance national des organismes de la sphère publique (par exemple l'INSEE, l'Imprimerie nationale…).

Résumé de l'article 47 « données de santé » du projet de loi de modernisation de notre système de santé après la première lecture à l'Assemblée nationale

Le texte du projet de loi est disponible sur le site de l'Assemblée nationale (http://www.assemblee-nationale.fr/14/ta/ta0505.asp) mais la lecture de l'article, qui procède par modifications de textes existants, n'est pas commode. Le résumé qui suit en expose les principales dispositions.

Le cadre des mesures proposées est un équilibre d'ensemble :

- ouvrir au public et multiplier les jeux de données complètement anonymes (et permettre aussi la réutilisation des données publiées par l'assurance maladie sur l'activité des professionnels de santé),
- autoriser les traitements des données comportant un risque de ré-identification mais seulement pour des projets d'intérêt public et dans des conditions garantissant le respect de la vie privée des personnes.

Les mesures nouvelles modifient notamment le Code de la santé publique puis certains articles de la loi informatique et libertés dont les chapitres IX et X sont fusionnés.

Dans le Code de la santé publique on expose d'abord les grands principes, puis le nouveau système national des données de santé (SNDS) est défini dans son périmètre et ses finalités.

Ce SNDS est constitué essentiellement de la grande base déjà assemblée par la Caisse nationale de l'assurance maladie des travailleurs salariés (les feuilles de soins, chaînées avec les séjours hospitaliers), qui sera complétée par des bases de taille plus réduite : causes de décès, données sur le handicap et échantillon représentatif des données de l'assurance maladie complémentaire. Elles sont réunies à des fins de connaissance, pour être mises à la disposition de personnes autorisées à les traiter, dans les conditions définies par la loi. Elles contribuent ainsi à l'information du public et aux autres finalités mentionnées.

Ces données ne comportent ni noms et prénoms ni numéros de sécurité sociale ni aucune information permettant une identification directe. Néanmoins, en raison des recoupements possibles, elles demeurent des données à caractère personnel.

La gouvernance des données de santé se caractérisera par :

- une implication plus forte de l'État dans le pilotage stratégique,
- une responsabilité de l'assemblage et de la gestion des données confiée à la Caisse nationale de l'assurance maladie des travailleurs salariés, en tant qu'opérateur principal, en concertation avec les autres gestionnaires de systèmes d'information contribuant au système national,
- une continuité entre l'actuel Institut des données de santé et le futur Institut national des données de santé (INDS) qui lui succèdera avec une composition élargie et des missions plus étendues, missions consultatives mais aussi délibératives (il pourra donner un avis sur l'intérêt public de projets pour lesquels une autorisation de traitement est demandée à la CNIL) et missions opérationnelles (il sera notamment un guichet commun pour l'enregistrement et l'orientation diligente des demandes).

Les modalités d'accès aux données seront rationalisées.

Un régime particulier d'accès permanent aux données du SNDS est prévu pour certains services de l'État, ou établissements publics ou organismes chargés d'une mission de service public, désignés par décret en Conseil d'État pris sur avis de la CNIL, pour autant que la mission de service public de ces services, établissements ou organismes l'exige.

Le régime commun demeure toutefois celui de l'autorisation par la CNIL de projets de traitement. Il est décrit au nouveau chapitre IX de la loi informatique et libertés (résultant de la fusion des actuels chapitres IX et X).

L'autorisation de traitement doit être justifiée par une finalité d'intérêt public et par le besoin, pour ce faire, d'utiliser des données à caractère personnel. L'INDS pourra donner son avis à la CNIL sur l'intérêt public et un comité d'expertise indépendant donnera lui son avis sur la méthode, et donc sur le besoin de recourir aux données à caractère personnel. Au total, l'avis de la CNIL rendu dans ces conditions vaudra avis d'un comité d'éthique pour les projets qui le nécessitent.

Des dispositions ambitieuses, conçues en concertation avec les services de la CNIL, permettront d'étendre l'usage de méthodes d'autorisation simplifiées, afin que cette étape du processus soit rendue fluide voire quasi instantanée pour les demandes d'accès respectant un cadre prédéfini par la CNIL ou pour des demandes répétitives ou pour la mise à disposition d'échantillons.

Le Code de la santé publique ajoute toutefois quelques particularités à la procédure pour ce qui concerne le SNDS lui-même. Deux finalités interdites d'abord : les assureurs et les industriels en produits de santé notamment devront apporter des garanties supplémentaires pour éviter que l'utilisation des données conduise à une sélection du risque pour les premiers, et au ciblage commercial des professionnels et établissements de santé pour les autres. A défaut d'avoir démontré que les données demandées ne permettent pas les traitements interdits, ils devront passer par des prestataires : laboratoires de recherche publics ou cabinets d'études privés accrédités par la CNIL.

Des précautions sont prises également pour s'assurer que les données ne puissent servir qu'aux finalités autorisées. Un référentiel technique garantira notamment que chaque accès aux données à caractère personnel est enregistré et que les données sont tracées. En outre les finalités d'accès et les méthodes utilisées devront être transparentes.

L'usage à des fins d'études et de recherche en santé du Numéro d'inscription au répertoire (dit NIR ou « numéro de sécurité sociale ») pourra être directement autorisé par la CNIL, qui vérifiera que les procédures mises en œuvre protègent bien la confidentialité, sans qu'un décret en Conseil d'État soit nécessaire : cela libérera la recherche en santé qui pourra s'appuyer sur des bases rendues plus riches du seul fait qu'elles seront appariées. Dans les cas exceptionnels où il faudra remonter du pseudonyme au NIR de la personne, il pourra être fait appel à un dispositif géré par un tiers de confiance national, désigné par décret, qui sera le seul organisme à détenir les secrets permettant de réaliser cette opération.

Enfin et par ailleurs, la mise en place du NIR en tant qu'identifiant national de santé (à l'article L. 1111-8 du Code de la santé publique) va permettre de faciliter l'échange et le partage de données de santé sous forme numérique en s'appuyant sur un identifiant opérationnel et certifié.

« Est-ce bien raisonnable ? »

Jean-Pierre LE GLÉAU

Dès qu'il s'est agi de protéger les données personnelles, il a fallu définir ce que l'on entendait par donnée nominative. Très vite, on a constaté qu'il ne suffisait pas de retirer le nom et le prénom pour obtenir une information anonyme. Des périphrases (le meilleur buteur de la coupe du monde de football en 1958) ou des accumulations d'informations (la personne qui a habité successivement au 26 rue du Labrador à Bruxelles puis au château de Moulinsart et qui est reporter) suffisent souvent à identifier des individus de façon unique.

La loi Informatique et libertés et la directive européenne

C'est pourquoi, la loi dite « Informatique et libertés » de janvier 1978[1] a posé la définition suivante, qui reconnaît les possibilités d'identification directe et indirecte : « Sont réputées nominatives les informations qui permettent, sous quelque forme que ce soit, directement ou non, l'identification des personnes physiques auxquelles elles s'appliquent ».

Cependant, le Parlement et le Conseil européen ont voulu affiner encore cette définition en précisant, dans le considérant n°26 d'une directive européenne de 1995[2] , les éléments à prendre en considération pour déterminer si une personne est, ou non, identifiable. Pour cela, dit la directive, il faut « *considérer l'ensemble des moyens susceptibles d'être raisonnablement mis en œuvre, soit par le responsable du traitement, soit par une autre personne, pour identifier ladite personne* ».

Pour éviter de considérer comme identifiantes certaines informations difficiles à rattacher à un individu, le législateur européen a introduit l'adverbe « raisonnablement ». Cette disposition permet d'éviter de considérer comme indirectement identifiante une information telle que : « La personne qui a obtenu son permis de conduire en 1970, qui a été abonnée pendant huit années consécutives à la Comédie française et qui a habité pendant trois ans à l'étranger » Même si cette personne est unique, il faudrait pour l'identifier recourir à des fichiers d'accès peu commode, pas toujours informatisés et éparpillés dans divers organismes ; autrement dit, mettre en œuvre des moyens déraisonnables. En conséquence, au sens de la directive de 1995, ces données ne suffisent pas pour considérer que l'on a affaire à des informations identifiantes.

De plus, la terminologie « donnée à caractère personnel » s'est substituée à celle de « donnée directement ou indirectement nominative ».

La transposition de la directive par la France

Comme toute directive européenne, celle de 1995 doit être transposée dans le droit de chaque État membre pour devenir effective. Des délais sont imposés aux États pour effectuer cette transcription, qui doit respecter le texte de la directive.

Première lecture à l'Assemblée nationale

Lorsque la France a engagé en 2002 cette transposition, le texte du projet de loi déposé par le gouvernement indiquait simplement que :

[1] Loi n° 78-17 du 6 janvier 1978 relative à l'informatique, aux fichiers et aux libertés.

[2] Directive 95/46/CE du Parlement européen et du Conseil, du 24 octobre 1995, relative à la protection des personnes physiques à l'égard du traitement des données à caractère personnel et à la libre circulation de ces données.

« *Constitue une donnée à caractère personnel toute information relative à une personne physique identifiée ou qui peut être identifiée, directement ou indirectement, par référence à un numéro d'identification ou à un ou plusieurs éléments qui lui sont propres* »

Il reprenait en cela, façon un peu simplifiée, la définition suivante figurant à l'article 2a de la directive européenne de 1995 :

« *donnée à caractère personnel : toute information concernant une personne physique identifiée ou identifiable (personne concernée); est réputée identifiable une personne qui peut être identifiée, directement ou indirectement, notamment par référence à un numéro d'identification ou à un ou plusieurs éléments spécifiques, propres à son identité physique, physiologique, psychique, économique, culturelle ou sociale* ».

Première lecture au Sénat

Lorsque ce texte a été examiné au Sénat en première lecture, en avril 2003, le rapporteur de la commission des lois, Alex Türk, qui était à l'époque vice-président de la Cnil, a proposé d'inclure dans cette définition la précision figurant dans le considérant n°26 de la directive de 1995. Il a été suivi par le Sénat qui a donc adopté la rédaction suivante :

« *Constitue une donnée à caractère personnel toute information relative à une personne physique identifiée ou qui peut être identifiée, directement ou indirectement, par référence à un numéro d'identification ou à un ou plusieurs éléments qui lui sont propres. Pour déterminer si une personne est identifiable, il convient de considérer l'ensemble des moyens susceptibles d'être raisonnablement mis en œuvre, soit par le responsable du traitement, soit par une autre personne* ».

Seconde lecture à l'Assemblée nationale

Quand ce texte est revenu devant l'Assemblée nationale en avril 2004, le rapporteur de la commission des lois, Francis Delattre, a considéré que l'emploi de l'adverbe « raisonnablement » n'était pas sans ambiguïté et risquait de provoquer de réelles difficultés d'interprétation, source de contentieux. Par ailleurs, il notait que la nouvelle rédaction proposée par le Sénat reproduisait partiellement, non pas une disposition de la directive de 1995 elle-même, mais son considérant n°26 qui, comme tout considérant, n'a pas de valeur normative mais explicative de l'intention du législateur européen. En conséquence, il proposait de supprimer, dans la transposition française, l'adverbe « raisonnablement »

Cette suppression, si lourde de conséquences, de l'adverbe incriminé, a fait l'objet, en séance publique, d'un débat qui a duré moins de trente secondes. Était-ce bien raisonnable ?

Seconde lecture au Sénat

Lorsque le texte est arrivé devant la commission des lois du Sénat, présidée par Alex Türk, devenu entretemps président de la Cnil, celle-ci a jugé que la suppression de l'adverbe « raisonnablement » devait effectivement permettre de prévenir des difficultés d'interprétation et a donc souscrit à cette volonté de sécurité juridique.

Aucun amendement n'a été déposé visant à rétablir l'adverbe « raisonnablement » et le texte définitif a donc été adopté sans celui-ci.

Situation actuelle et évolution prévue

Le résultat de ces turbulences législatives est que, aujourd'hui, la définition d'une donnée à caractère personnel n'est pas strictement identique selon que l'on se place du point de vue du droit français ou du droit européen. Les données pour lesquelles l'identification des individus requiert la mise en œuvre de moyens déraisonnables sont considérées comme nominatives par la loi française, mais anonymes selon la directive européenne.

Cela a des conséquences pratiques. Du point de vue la loi française, l'anonymat est considéré comme quelque chose d'absolu. Il n'est pas question de mettre en balance le coût nécessaire pour parvenir à une éventuelle levée d'anonymat avec le tort qui peut en résulter ou le bénéfice qui peut en être tiré.

Au contraire, la directive européenne invite implicitement à prendre en compte le caractère déraisonnable de certaines possibilités de levée de l'anonymat qui, pour aboutir, devraient mettre en jeu des moyens sans rapport avec les potentiels bénéfices escomptés ou torts portés.

Aujourd'hui, la directive européenne de 1995 fait l'objet d'un nouveau débat au niveau européen.

Le nouveau texte visé sera un règlement, donc directement applicable dans le droit de chacun des États membres, sans qu'il soit besoin d'attendre une loi de transposition.

Pour la définition d'une donnée à caractère personnel, le texte élaboré par la commission a repris pour l'essentiel celui de la directive de 1995. Les définitions proposées à l'article 4 du projet étaient les suivantes :

> « - *données à caractère personnel : toute information se rapportant à (…) une personne physique identifiée ou une personne physique qui peut être identifiée, directement ou indirectement, par des moyens raisonnablement susceptibles d'être utilisés par le responsable du traitement ou par toute autre personne physique ou morale, notamment par référence à un numéro d'identification, à des données de localisation, à un identifiant en ligne ou à un ou plusieurs éléments spécifiques, propres à son identité physique, physiologique, génétique, psychique, économique, culturelle ou sociale* ».

Mais le Parlement européen a proposé, le 12 mars 2014 un texte un peu différent, qui donne la définition suivante d'une donnée à caractère personnel :

> « *toute information se rapportant à une personne physique identifiée ou identifiable ; est réputée identifiable une personne qui peut être identifiée, directement ou indirectement, notamment par référence à un identifiant, par exemple à un nom, à un numéro d'identification, à des données de localisation, à un identifiant unique ou à un ou plusieurs éléments spécifiques, propres à l'identité physique, physiologique, génétique, psychique, économique, culturelle, sociale ou de genre de cette personne. »*

On notera que, dans cette nouvelle définition, l'adverbe « raisonnablement » a disparu…

Cependant, il reste présent dans le considérant n°23 de ce projet de règlement, qui précise : « *Pour déterminer si une personne est identifiable, il convient de considérer l'ensemble des moyens raisonnablement susceptibles d'être mis en œuvre, soit par le responsable du traitement, soit par une autre personne, pour identifier ou isoler directement ou indirectement ladite personne. Pour établir si des moyens sont raisonnablement susceptibles d'être mis en œuvre afin d'identifier une personne physique, il convient de considérer l'ensemble des facteurs objectifs, tels que le coût de l'identification et le temps nécessaire à celle-ci, en tenant compte à la fois des technologies disponibles au moment du traitement et de l'évolution de celles-ci.* »

Ce considérant ne semble pas tout à fait cohérent avec le texte du règlement lui-même.

On attend la réaction du Conseil européen sur ce nouveau texte….

Conclusions du groupe de travail sur les risques de ré-identification dans les bases de données médico-administratives

Annexe 9 du rapport de la Commission Open Data en santé - Juillet 2014

A la suite du rapport Bras[1] qui lui a été remis le 3/10/2013, la ministre, Madame Marisol Touraine, a demandé à Franck von Lennep, directeur de la recherche, des études, de l'évaluation et des statistiques (DREES), de diligenter une expertise technique sur la sécurité des données concernant le risque de ré-identification des personnes à partir de données [supposées] anonymes.

L'étude a été réalisée de novembre 2013 à avril 2014 par un groupe de travail animé par André Loth (DREES) et composé de : Alireza Banaei et Max Bensadon (ATIH), Hélène Caillol (CNAMTS), Nora et Frédéric Cuppens (Institut Telecom Bretagne), Emmanuelle Denis (DSS), Françoise Dupont (INSEE et CASD), Noémie Jess (DREES), J-Pierre Le Gléau (expert), Grégoire Rey (INSERM-CépiDc).

Y ont également contribué (pour le test de jeux de données anonymes) Maxime Bergeat (INSEE), Dominique Blum (expert) et les équipes techniques du CASD.

Le comité de pilotage, présidé par Franck von Lennep, s'est réuni 3 fois. Ont participé à tout ou partie de ces réunions Max Bensadon (ATIH), Philippe Burnel (DSSIS), Anne Coat (ANSSI), Claude Gissot (CNAMTS), Nora et Frédéric Cuppens (Institut Télécom Bretagne), François Godineau (DSS), Philippe Cuneo et Michel Isnard (INSEE), Jean-Pierre Le Gléau (expert), André Loth (DREES), Christian Saout (CISS) ainsi que Marc-André Beaudet (expert de la CNIL, en observateur).

Ce rapport constitue l'annexe n°9 du rapport de la Commission open data en santé du 9 juillet 2014 (disponible sur le site de la DREES : http://www.drees.sante.gouv.fr/IMG/pdf/annexes_rapport_open_data.pdf)

L'objet de l'étude

En substance la mission confiée au groupe de travail était :

- d'identifier et évaluer les risques de ré-identification dans les bases médico-administratives considérées (SNIIRAM, PMSI…) notamment pour les jeux de données issus de ces bases et ayant donné lieu à une diffusion relativement large (données hospitalières du PMSI, échantillon généraliste des bénéficiaires) ou dont l'anonymat faisait débat (datamart des consommations inter-régimes de l'assurance maladie) ;
- d'établir une ligne de démarcation entre
  - > jeux de données anonymes (pouvant donc être mis en open data ou publication sans restriction)
  - > jeux de données présentant un risque de ré-identification et devant pour cette raison être en accès restreint…
- De recommander des moyens pour élargir, dans cette offre de données, la part et le volume des données en accès libre ;
- De préciser pour les données en accès restreint (parce que comportant des risques de ré-identification) sous quelles conditions techniques les personnes habilitées à accéder à ces données devraient pouvoir le faire afin de limiter le risque.

Le groupe a entendu des chercheurs et parcouru la littérature sur ce sujet. Il a été surpris et déçu de constater qu'il n'existe pas vraiment encore de doctrine établie et que la plupart des responsables de bases de données, publiques et

[1] P-L Bras Rapport sur la gouvernance et l'utilisation des données de santé (sur les risques de ré-identification dans les bases de données médico-administratives, voir notamment pp. 26 à 30 et 56 à 58 du rapport).

privées mais aussi les responsables des autorités de contrôle (les équivalents de la CNIL) en sont encore à tâtonner. Le groupe des 29 CNIL européennes (Encadré 1) vient ainsi de définir une liste de trois conditions suffisantes pour qu'un jeu de données puisse être qualifié d'anonyme mais ces trois conditions ne sont pratiquement jamais réunies et on est renvoyé dès lors à une évaluation au cas par cas[2] : la veille technologique et scientifique sur ces questions devra être poursuivie demain, à l'initiative des gestionnaires des bases mais aussi dans les instances de gouvernance et dans les instances de contrôle afin que la doctrine s'affirme (et évolue quand c'est nécessaire).

ENCADRÉ 1 - L'AVIS DES AUTORITES DE PROTECTION DES DONNEES EUROPEENES SUR LES PRINIPALES TECHNIQUES D'ANONYMISATION

M-A Beaudet, A. Rousseaux (CNIL)

Le service de l'expertise technologique de la Commission nationale de l'informatique et des libertés (CNIL) a été associé aux travaux menés par le groupe de travail sur le risque de ré-identification en tant qu'observateur. Lors des auditions, il a notamment présenté l'avis[3] du G29, qui regroupe les autorités de protection des données européennes, du 16 d'avril 2014 relatif à la notion d'anonymisation. Cet avis précise la définition de l'anonymisation et rappelle que les principes de protection des données ne s'appliquent pas aux données rendues anonymes d'une manière telle que la personne concernée n'est plus identifiable.

Trois critères permettent de s'assurer du caractère anonyme d'un jeu de données :

- L'individualisation : il ne doit pas être possible d'individualiser une personne,
- La corrélation : il ne doit pas être possible de relier plusieurs données au sein d'un même jeu ou entre plusieurs jeux de données, et
- L'inférence : il ne doit pas être possible de déduire des informations.

Dès lors que ces 3 critères ne sont pas réunis, il est nécessaire de mener une analyse des risques de ré-identification propre à chaque jeu de données, afin de considérer l'exhaustivité des risques (possibilité de recoupements différents, historiques des jeux de données ouverts, capacité et motivation des personnes souhaitant ré-identifier un jeu…).

Ces analyses doivent permettre de déterminer les techniques à utiliser pour obtenir un jeu de données anonymes. En revanche, si les personnes restent identifiées ou identifiables, le jeu de données est dit pseudonymisé. Des mesures de protection doivent alors être définies en fonction des risques de ré-identification identifiés.

Ainsi, trois niveaux d'ouverture ont pu être identifiés : le premier, relatif à des données anonymes, permet d'offrir des jeux de données en Open Data ; le second, relatif à des données pseudonymisées présentant un faible risque de ré-identification, permet d'offrir des jeux de données en accès limité et encadrés par quelques mesures de sécurité ; enfin, le troisième niveau, propre aux données pseudonymisées, soit présentant un fort risque de ré-identification, soit directement identifiantes, pour lequel un accès restreint et hautement sécurisé est nécessaire.

L'anonymisation d'un jeu de données peut donc s'avérer complexe. La CNIL reste attentive à ce sujet et sera bien évidemment à la disposition des instances impliquées sur ce sujet pour les accompagner dans la mise en œuvre des outils facilitant la mise en œuvre des techniques d'anonymisation.

L'évaluation du risque de ré-identification dans le cas des données médico-administratives de santé

Il est entendu qu'il n'est question ici que des bases de données dé-identifiées ou « pseudonymisées », dites parfois, à tort, anonymes, ou de « jeux » de données issus de ces bases.

Le risque de ré-identification des personnes dans ces bases ne vient pas tant du pseudonyme utilisé[4] que de la possibilité pour des tiers de reconnaître, dans des jeux de données, des personnes sur lesquelles ils disposent d'informations par ailleurs.

[2] http://www.cnil.fr/linstitution/actualite/article/article/le-g-29-publie-un-avis-sur-les-techniques-danonymisation/

[3] http://ec.europa.eu/justice/data-protection/article-29/documentation/opinion-recommendation/files/2014/wp216_en.pdf

[4] Dans les bases SNIIRAM et PMSI, le pseudonyme est le NIR chiffré deux fois de manière irréversible. Une bonne organisation (par exemple des clés de chiffrement secrètes ou une table de correspondance, confiées à plusieurs personnes différentes extérieures à la base de données) évite le risque que le pseudonyme permette de remonter à la vraie identité de la personne, identité qui est parfaitement inutile pour les traitements ordinaires. Dans les cas où remonter à l'identité des personnes est nécessaire (pour rendre possible des appariements avec d'autres jeux de données, d'enquête par exemple), le principe de cloisonnement évite que les gestionnaires de la base connaissent les vraies identités. Ce principe de gestion séparée des données directement identifiantes (nom, NIR, numéro de téléphone…) est celui qui est posé par le projet de règlement européen sur la protection des données personnelles (art. 81 et 83).

Classiquement (cf. le Modèle d'évaluation harmonisée du risque –MEHARI- du Club de la sécurité de l'information français), le risque, dépend à la fois :

- de la probabilité ou potentialité que l'événement se réalise, malgré des mesures dissuasives ou préventives (potentialité généralement évaluée sur une échelle de 1, faible, à 4, très fort) ;
- de l'impact (noté de 1 à 4 également) qu'a sa réalisation, compte tenu d'éventuelles mesures compensatoires ou réparatrices…

Risque = Potentialité x Impact

S'agissant de l'impact d'une divulgation non souhaitée de données personnelles de santé, le groupe a considéré qu'il était *très fort*, s'agissant de données classées comme sensibles dans le droit français et européen, et protégées par le droit constitutionnel à la vie privée. Des exemples de conséquences potentiellement très dommageables pour la personne concernée peuvent être cités (portant sur la carrière professionnelle ou l'image ou la vie de famille notamment).

Le fait que certaines personnes puissent rendre publiques elles-mêmes des informations sur leur propre santé ne change rien à l'évaluation de l'impact pour les victimes d'une divulgation non souhaitée. En outre, une fois l'information divulguée, le mal est fait et on ne peut ni remettre l'information dans la boîte ni considérer qu'une compensation financière réparera le préjudice.

Si on s'accorde à évaluer l'impact comme très fort, cela conduit à refuser en l'espèce une conception utilitariste de l'analyse bénéfice/risque où on admettrait que l'utilité économique pour le plus grand nombre compense le préjudice subi par quelques uns (supposés peu nombreux) dont la vie privée serait divulguée. Les membres du groupe ont préféré une autre conception de l'analyse bénéfice-risques, consistant à maximiser le niveau de bénéfice pour un niveau de risque maîtrisé. Cela permet de conclure à l'intérêt :

- de rendre les données personnelles de santé accessibles – notamment aux chercheurs ou à des bureaux d'études – dans des conditions convenables (rapidité, assistance…) - mais avec l'accord de la CNIL et en contrôlant le risque de divulgation ;
- d'ajuster le niveau des précautions au niveau du risque ;
- et de maximiser la production et la diffusion des données anonymes issues - par divers procédés - des données individuelles brutes.

S'agissant de la **probabilité de ré-identification** et de l'efficacité de la dissuasion ou des mesures préventives, le groupe a considéré sans s'y appesantir que la diffusion de données agrégées issues des bases médico-administratives ne présentait pas en général de risque de ré-identification et qu'il n'y avait donc pas lieu d'en limiter l'accès à telle ou telle catégorie de personnes ou d'organismes comme c'est encore actuellement le cas, et qu'il y avait en outre certainement des moyens d'accroître la richesse et la pertinence des statistiques mises en ligne.

Comme on le verra, le groupe a également travaillé sur la possibilité d'étendre la production et l'usage de jeux de données individuelles -jeux exhaustifs ou échantillons- ayant pour caractéristiques que le risque de ré-identification y soit a priori inexistant et que ces données puissent donc elles aussi être publiées et réutilisées par tous.

Cependant, il lui était d'abord demandé de se prononcer sur la politique de diffusion des données médico-administratives individuelles par le ministère de la santé et les agences ou établissements publics placés sous sa tutelle. S'agissant des données individuelles à vocation exhaustive du SNIIRAM, du PMSI et de l'ensemble SNIIRAM-PMSI, il confirme l'évaluation concluant à un niveau élevé du risque :

- le risque que présente une diffusion mal contrôlée des données hospitalières du PMSI exhaustif (dans les conditions et avec les précautions insuffisantes[5] qui caractérisent son mode de diffusion actuel) ;

[5] Les calculs de D. Blum, confirmés par l'ATIH, donnent les ratios de 89 % et 100 % de séjours où une personne est seule à présenter les caractéristiques cherchées (respectivement sur l'ensemble des séjours de 2008 et sur les séjours des patients hospitalisés au moins 2 fois dans l'année). Ces calculs ont été effectués en 2011sur une base présentant le même type de « floutage » que les bases diffusées au cours des années récentes (imprécision volontaire sur la date de sortie et code géographique correspondant à au moins 1000 habitants). La diffusion à plusieurs centaines d'exemplaires de copies de cette base complète, sous la forme de disques, la diffusion systématique du fichier de chaînage des séjours, l'existence avérée de copies sauvages (et l'intérêt d'en obtenir en raison de leur valeur marchande et pour éviter des procédures et un paiement), l'impossibilité pratique d'empêcher ces copies ou de contrôler le respect des engagements pris par les titulaires des autorisations de la CNIL ont été les principaux arguments cités.

- le risque que cela entraîne pour l'Échantillon généraliste des bénéficiaires (EGB) au 1/97, accessible à un nombre important d'organismes, où on trouve les parcours de soins de 600 000 personnes sur (potentiellement) 20 ans. Cet échantillon a été malheureusement compromis par la diffusion trop large du PMSI lequel permet comme on l'a vu d'identifier des personnes hospitalisées et de les retrouver, avec les mêmes caractéristiques, dans l'EGB ;
- le risque que présenterait un accès trop large au datamart de consommations inter-régimes (DCIR) exhaustif de l'assurance maladie, si par exemple il devenait ouvert à toutes les personnes ayant le statut de chercheurs, même sans possibilité de croisement des variables sensibles et même sans croisement avec le PMSI (ce qui lui retirerait d'ailleurs une bonne part de son intérêt).

Dans tous ces cas, le nombre de personnes présentant les mêmes caractéristiques au regard des quasi-identifiants (date en mois des soins, durée d'hospitalisation, arrêt de travail, âge en années…) est très petit, tendant rapidement vers la valeur 1 si les données sont chaînées dans le temps : tous les parcours de soins sont différents.

Si on prend l'exemple de l'INSEE, celui-ci ne prend pas le risque de diffuser un fichier exhaustif présentant ce type de risques autrement que via un dispositif sécurisé à des personnes dont le besoin de connaître les données a été préalablement vérifié et évalué (c'est également ce qui est prévu pour les données fiscales).

Dans cette évaluation de la probabilité de ré-identification, il convient de distinguer trois temps :

> Le comptage du nombre de « personnes dans une case » (personnes présentant les mêmes caractéristiques au regard des quasi identifiants que sont l'âge, le sexe, le lieu de résidence, le lieu, la nature et les dates de soins, la date de décès éventuellement). Il présente un caractère objectif. Ce nombre est fortement réduit jusqu'à devenir vite égal à un pour toute la population dès que les données relatives à une même personne sont chaînées[6] ;

> Le classement des quasi-identifiants par degrés de notoriété n'est guère contestable non plus. Les dates de naissance et de décès par exemple sont aisément connues de tous ; les dates et lieu d'hospitalisation ou les dates d'arrêt de travail restent assez faciles à connaître, même si pour une personne donnée, le nombre de personnes susceptibles de les connaître est généralement petit (proches, employeur, assureur). En revanche, les dates de consultations médicales ou de soins par des auxiliaires médicaux sont plutôt mal connues et vite oubliées ;

> L'estimation de la probabilité que ces informations donnent lieu à une divulgation par les personnes qui en ont eu connaissance est plus subjective (« à dire d'experts ») :

  - cette probabilité augmente évidemment avec le nombre de personnes mises dans le secret, directement ou indirectement, et dépend donc de la procédure de sélection des demandes d'accès ;
  - elle est réduite avec l'emploi de dispositifs techniques permettant le confinement des données, la traçabilité des accès et des requêtes et l'interdiction des copies, alors qu'au contraire la diffusion des données au moyen de copies et d'extractions augmente la probabilité de divulgation par négligence et/ou malveillance ;
  - elle est réduite aussi avec le caractère dissuasif des sanctions pénales pour divulgation illicite, que le groupe a jugé
    - assez efficace dans le cas d'institutions telles que les assurances (secret de pratiques illicites difficile à garder, souci d'image…)
    - mais peu efficace devant d'autres « attaquants » (par exemple adversaires politiques, malveillance dans un cadre familial ou professionnel, presse à scandale, démonstration qu'on peut le faire ou simple curiosité…) surtout si la probabilité d'être pris est très faible (pas de traçabilité pour les données extraites). On notera que les cas de divulgation des données peuvent alors rester confinés à un petit cercle.

En tout état de cause, l'argument « jusqu'ici tout va bien » ne dispense pas, selon le groupe de travail, de se conformer aux standards professionnels en vigueur et ne peut justifier la continuation du mode actuel de diffusion des données du PMSI et de l'EGB, ni l'ouverture incontrôlée des accès au DCIR, qui sont porteuses de risques pour les personnes concernées et qui seraient de nature, en cas d'incidents publics, à saper la confiance dans le système d'information.

Il appartient à la CNIL de faire savoir quelle sera désormais son attitude concernant la diffusion des données du PMSI, mais le GT recommande au ministère de la santé de revoir sa politique de diffusion des données qui est jugée trop

[6] Même sans chaînage des données, certains cas facilitent la ré-identification des personnes (patients soignés loin de chez eux, décès…).

restrictive dans le cas des données agrégées que chacun s'accorde à classer comme anonymes et trop laxiste dans le cas des données du PMSI.

Pour mémoire, le groupe de travail ne s'est pas penché sur le cas des usages opérationnels des données (par exemple dans les caisses d'assurance maladie pour les feuilles de soins ou dans les ARS pour les données du PMSI) : ces usages sont ceux qui ont justifié initialement le recueil des données et sont donc par hypothèse nécessaires au service public. Le groupe rappelle seulement que dans ces cas aussi l'accès aux données doit être encadré (personnes habilitées individuellement et tenues au secret professionnel[7], tenue de registres, accès sécurisé…).

Une réponse graduée selon le risque de ré-identification

L'offre que le groupe de travail propose de faire à ce stade (travaux à poursuivre, en liaison avec les services de la CNIL et la future instance de gouvernance s'articule autour de trois axes :

a) une offre pour tous les publics en accès direct sur internet

La manière la plus simple de mettre à disposition des données anonymes issues de bases indirectement nominatives est évidemment de produire des tableaux statistiques répondant par des données agrégées aux questions que se posent les différentes catégories d'utilisateurs. Le groupe n'a pas exploré cette piste faute de temps mais elle ne pose pas de problème de principe : seulement de faisabilité technique et de coûts. La limite de l'exercice est que la demande est potentiellement infinie, alors que la production de statistiques agrégées a un coût et répond souvent mal ou trop lentement aux besoins particuliers des uns et des autres, notamment aux besoins des chercheurs ou à des questions d'une grande technicité ou exigeant le recours à plusieurs sources de données.

Une solution est de rendre plus « agile » la production de ces tableaux statistiques à l'aide d'outils techniques par exemple en mettant à la disposition des utilisateurs des bibliothèques de requêtes en libre service ; à la limite cela revient à laisser interroger la base à l'aide d'outils logiciels paramétrables fournissant des réponses anonymes sans que l'auteur de la requête ait lui-même accès aux données. Le même résultat peut être obtenu, plus classiquement en organisant des collaborations (un « guichet ») entre une équipe d'experts attachés à la base de données et les experts « métier » des demandeurs, pour fabriquer les sorties agrégées adaptées aux besoins, sachant que la mobilisation d'agents pour des travaux à façon a un coût et que là aussi on ne répond pas toujours aux besoins les plus pointus.

L'autre solution explorée par le Groupe de travail « risques de ré-identification » consiste en la production d'un grand nombre de « jeux » de données rendus anonymes, tels que pour chacun de ces jeux, le nombre de personnes présentant des caractéristiques identiques soit suffisant (K=10 au moins dans le test) et que parmi ces personnes il y en ait toujours un nombre suffisant (L=3 au moins dans le test) qui aient des maladies différentes. On a démontré que c'est possible et que les critères d'anonymat peuvent être vérifiés… Des précautions doivent être prises cependant pour qu'il ne soit pas possible de reconstituer ainsi des jeux de données identifiants… Une sécurité supplémentaire peut être obtenue en utilisant des techniques d'échantillonnage (si la règle de tirage est secrète).

Les fichiers qui seront mis à disposition de cette façon doivent être absolument anonymisés, pour tout type de consultant qui n'a d'ailleurs pas besoin de se faire connaître (assureur, employeur, voisins, famille…).

Le groupe de travail est parti du fichier PMSI et a construit, par généralisation ou suppression de variables, des jeux répondant aux critères d'anonymisation : pour toute combinaison de variables quasi identifiantes, il y a au moins dix individus. Parmi ces individus, il y a au moins trois types de « maladies » différentes.

À titre d'exemple, on a construit un fichier comprenant pour chaque séjour d'hospitalisation : le sexe et l'âge (en 18 groupes d'âge) de la personne hospitalisée, son mode d'entrée et de sortie (en 2 groupes : « domicile » ou « autre ») et la durée d'hospitalisation (12 modalités) et le groupe homogène de malades (indicateur de la maladie).

[7] Le groupe rappelle qu'il n'est pas besoin d'être médecin pour respecter et faire respecter le secret professionnel sur des données personnelles de santé.

Un autre fichier donnant un peu moins de précision sur la durée d'hospitalisation (2 modalités), mais comprenant la région de résidence du malade semble aussi répondre aux critères. D'autres exemples sont décrits en annexe mais un grand nombre de jeux répondant aux critères d'anonymisation peut a priori être mis à disposition de la même manière,

Des fichiers analogues pourront, selon la même méthode, être construits à partir du SNIIRAM.

Le groupe ne donne pas de recommandation définie sur le nombre minimal de personnes présentant les mêmes caractéristiques que doit comporter un jeu de données pour être considéré comme anonyme (k-anonymat)… Certains membres du groupes préconisaient d'examiner les seuils K=3 et L=3 qui préservent en théorie l'anonymat en admettant qu'il serait peut-être nécessaire de se fixer sur un seuil un peu plus élevé (K=5, L=3 par exemple). Les travaux devront être poursuivis à ce sujet, car certains membres du groupe considéraient à l'inverse que ces valeurs sont trop basses pour garantir l'anonymat.

Il est mis à l'étude aussi la proposition que des fichiers plus détaillés sur les caractéristiques médicales du séjour (médicaments en sus, DMI, diagnostics et actes…) soient fournis mais sans indication du nom de l'établissement et sous la forme de sondage (au 1/10 voire 1/3). Certes, les données médicales ne sont pas considérées comme un quasi identifiant (c'est l'information que « l'assaillant » est supposé rechercher dans la base : s'il en disposait déjà , il n'aurait pas besoin de la base) mais elles doivent être appauvries elles aussi dans les jeux de données anonymes parce que la partie médicale du fichier de résumé des sorties anonymisés (RSA) est très discriminante (chaque RSA est à cet égard différent des autres et on pourrait, si on conservait cette information commune dans des jeux de données par ailleurs appauvris, reconstituer le jeu d'origine).

La technique de sondage confère en tout état de cause une sécurité supplémentaire et peut être appliquée partout où un fichier exhaustif n'est pas indispensable, à la condition bien sûr qu'on ne parte pas d'un fichier où chaque personne a déjà notoirement des caractéristiques uniques et que la règle de tirage soit gardée secrète.

Le caractère anonyme de ces fichiers, notamment l'impossibilité de remonter au fichier d'origine, devra être validé par des instances à désigner (on pense notamment à un avis de la CNIL sur le mode de construction de ces fichiers et non sur chaque fichier pris isolément).

La piste du « bruitage » (ajout ou substitution de données rendant « fausses » les données individuelles mais conservant des valeurs telles que moyenne, médiane, écart type etc.) n'a pas été suivie, notamment parce qu'elle nous a paru supposer une utilisation prédéterminée du jeu de données alors qu'en open data, les usages sont à la main des utilisateurs. Elle suppose également une compétence forte des utilisateurs pour en maitriser les limites d'utilisation.

b) une offre plus limitée, destinée à un public ciblé : un nouvel EGB ?

S'agissant des données à faible risque d'identification, la proposition faite est de constituer au moins un échantillon de données de grande taille tel qu'on ne puisse y reconnaître personne de manière certaine et dont l'accès soit ouvert à des organismes publics ou privés dont les missions et les compétences le justifient. Le principe de l'accès leur étant acquis, une limitation s'impose cependant en raison d'un risque résiduel tenant au caractère très particulier de certains parcours de soins[8] : il conviendra qu'ils souscrivent un certain nombre d'engagements sur le respect des règles d'accès, d'usage des données et de publication des résultats.

Il appartiendrait à la CNIL après avis de l'instance technique décrite au point précédent de valider le niveau de risque. Dès lors, l'accès à tel ou tel organisme (chercheurs, organisations professionnelles, administrations) pourrait relever comme aujourd'hui d'un arrêté ministériel.

Un fichier spécifique pourrait être construit à partir du SNIIRAM, chaîné avec le PMSI. Ce fichier ne serait pas exhaustif, mais issu d'un sondage (par exemple au 1/10[9]). Il est important que le mode de sondage et les paramètres retenus pour celui-ci demeurent strictement confidentiels. Il sera alors impossible à quiconque de savoir si telle ou telle personne est présente dans ce fichier.

[8] Dans l'échantillon, si la règle de tirage est secrète et si certaines données sont rendues suffisamment imprécises, on ne pourra jamais affirmer qu'une personne est reconnaissable même si elle présente (âge, sexe, parcours de soins) des caractéristiques uniques.

[9] Un échantillon au 1/10 a été suggéré dès 2006 dans un rapport des Professeurs Bégaud et Costagliola.

Si l'on découvre alors une personne ayant exactement les caractéristiques d'un individu (et d'un seul) présent dans ce fichier, on ne pourra pas en conclure à son identification, car :

- on ne sait pas si la personne visée est ou non présente dans le tirage ;
- il n'y a qu'une certaine probabilité (dans l'exemple, une chance sur dix) pour que la personne repérée dans le fichier soit la personne connue.

Cette remarque doit cependant être tempérée par le fait qu'une accumulation de ressemblances avec une personne connue pourrait, au fil de l'addition des millésimes diffusés, au fur et à mesure de l'enrichissement annuel de la base, conduire à une probabilité d'identification proche de l'unité.

C'est pourquoi, ce type de fichier devrait être

- rendu imprécis dans certaines de ses dimensions (telles que dates et lieu de soins, adresse, âge, date de décès ?) ce qui le rendrait inutilisable pour le suivi précis des parcours de soins mais très utile pour des études de pharmaco-épidémiologie par exemple…
- mis en accès limité, avec identification de la personne qui y a accès et engagement de sa part à n'utiliser le fichier qu'à des fins déclarées à l'avance, lesquelles n'incluent pas la tentative d'identifier une personne (modèle du réseau Quetelet ou de l'accès actuel des chercheurs INSERM aux données de l'EGB).

c) une offre restreinte pour les fichiers permettant une identification indirecte des malades

Comme cela a été vu, certains fichiers, utiles pour la recherche, sont extrêmement détaillés et permettent une identification indirecte de nombreuses personnes.

Les fichiers PMSI et EGB ayant été assez largement diffusés, ils ne devraient dorénavant être disponibles que moyennant des procédures d'accès très strictes.

Les personnes demandant un accès doivent justifier de la raison qui leur fait effectuer cette demande. Elles doivent démontrer les garanties de sérieux pour elles-mêmes et pour leur environnement professionnel. Au moment de leur accès, une trace des opérations effectuées doit être conservée. Enfin, un organisme doit vérifier que les sorties effectuées à partir de ces fichiers ne mettent pas en danger la protection de la vie privée des personnes.

Des systèmes et des procédures existent qui permettent ce type d'accès à des données confidentielles sans que cela implique un site unique.

Mais si ces conditions sont réunies et qu'on admet que le fichier est indirectement10 nominatif, il n'y a sans doute pas lieu d'appauvrir les données fournies : la distinction actuelle entre les personnes autorisées ou non à croiser les quatre données sensibles (dates de soins, mois de naissance, commune de résidence et date de décès) n'a a priori plus de raison d'être.

Ce système à trois niveaux devrait permettre de présenter une offre aussi vaste que possible, tout en préservant la confidentialité des données.

[10] Bien sûr, puisque les traitements autorisés visent exclusivement à tirer des conclusions générales à partir de données individuelles, il n'y a pas lieu de mentionner dans les données traitées les données directement identifiantes (nom, NIR, adresse etc.)

Le centre d'accès sécurisé aux données, du groupe des écoles nationales d'économie et statistique

Françoise DUPONT (CASD), Kamel GADOUCHE (INSEE – CASD), Antoine FRACHOT (Genes)

Un équipement d'accès sécurisé aux données confidentielles

L'accès des chercheurs à des données individuelles très détaillées est un enjeu scientifique majeur. Ces données sont relatives à toutes les catégories d'agents du monde économique et social, que ce soient des individus, des ménages, ou des entreprises et couvrent des informations aussi diverses que les revenus, le patrimoine, la santé, les données socio-démographiques, la localisation géographique, les parcours scolaires, les trajectoires professionnelles, etc. Elles peuvent donc être extrêmement sensibles et sont naturellement protégées par un dispositif de lois (Encadré 1) visant à garantir leur confidentialité et leur conformité au respect des libertés individuelles.

En 2008, l'Insee a souhaité permettre l'accès des chercheurs à ses données individuelles très détaillées couvertes par le secret statistique. Il s'agissait alors à la fois de lever un verrou juridique et de trouver une solution technologique adaptée aux besoins des chercheurs qui garantisse la non-dissémination des données.

Le verrou juridique a été levé via un changement législatif en 2008, mettant en place une procédure d'habilitation adossée au Comité du secret statistique. Dans le même ordre d'idée, un verrou juridique semblable concernant les données fiscales a été levé récemment, en 2013, et pourrait peut-être servir prochainement de modèle pour les données de santé.

S'agissant du verrou technologique, il a été levé ces dernières années par les équipes du Groupe des Écoles Nationales d'Economie et Statistique (Genes) grâce à la mise au point de la technologie d'accès sécurisé à distance du Centre d'accès sécurisé distant aux données (CASD), développée en collaboration étroite avec ses utilisateurs (principalement des chercheurs). Le CASD est donc un équipement permettant aux chercheurs de travailler à distance, de manière hautement sécurisée, sur des bases de données individuelles très détaillées.

Le principe de protection de la confidentialité des données indirectement nominatives entre potentiellement en conflit avec leurs utilisations par le plus grand nombre. Pour résoudre ce problème, certains pays (USA, Canada, Allemagne…) ont d'abord mis en place des centres d'accès sécurisé sous forme de locaux isolés où les chercheurs se rendent physiquement. Pour protéger la confidentialité des données, les chercheurs ne peuvent récupérer, après vérification par des opérateurs, que des tableaux suffisamment agrégés préservant ainsi le secret statistique.

Depuis les années 2000, d'autres pays (Danemark, Pays Bas...) ont mis en place des solutions d'accès sécurisé à distance pour les chercheurs. Ces solutions reposent sur l'installation sur l'ordinateur du chercheur de logiciels particuliers d'accès distant. Comme les installations sur les postes des chercheurs sont gérées au niveau du chercheur (décentralisées), les solutions s'appuyant dessus sont apparues comme peu sécurisées, complexes à mettre en œuvre et très coûteuses en termes de gestion et de maintenance.

En France, le développement du CASD repose sur un cahier des charges très strict :

- une sécurité maximale empêchant toute évasion incontrôlée de fichiers de données, ce qui exclut les solutions classiques via des logiciels de sécurité directement installés sur les postes de travail non maîtrisés des utilisateurs ;
- un coût minimal d'installation, de gestion et de maintenance, afin de ne pas faire supporter aux utilisateurs des coûts prohibitifs comme ceux constatés dans d'autres pays (États-Unis, Pays-Bas...) ;
- une adaptation maximale aux besoins des utilisateurs, ce qui exclut des systèmes de centre physique (dans lesquels les utilisateurs sont obligés de se déplacer) ou les dispositifs de *remote execution*[1].

[1] *Remote execution* : le chercheur conçoit ses programmes à partir d'un jeu de données fictives (ayant une structure similaire aux données réelles) et les soumet au centre de données pour exécution. Les résultats sont ensuite vérifiés et renvoyés au chercheur.

Le principe général de l'architecture du CASD repose sur deux éléments indissociables :

- une infrastructure sécurisée (étanche, parfois appelée « bulle ») de serveurs de calculs et de stockage ; les données restant protégées au sein de cet environnement ;
- un ensemble de boitiers (SD-Box) permettant un accès distant sécurisé à la « bulle » dans laquelle sont hébergées les bases de données confidentielles. L'accès des utilisateurs aux boitiers ne peut se faire qu'à l'aide d'une authentification forte par carte à puce.

Les utilisateurs peuvent, s'ils le souhaitent, faire une demande d'insertion de leurs propres données ou d'autres logiciels par l'intermédiaire d'une procédure sécurisée dédiée. De la même manière, toute demande de sortie de résultats intermédiaires ou finaux (tableaux, graphiques, fichiers…) du système est réalisée via une procédure spécifique de vérification de non-rupture de la confidentialité. Ainsi, tout ce qui « entre » ou « sort » de l'environnement protégé est vérifié.

Cet ensemble cohérent et étanche permet une isolation forte du dispositif, prévenant ainsi les risques d'intrusion interne ou externe, les attaques virales ainsi que les risques d'évasion de fichiers de données.

La levée simultanée du verrou juridique et du verrou technologique a permis à la plateforme CASD de prendre son essor et d'être aujourd'hui au service d'un grand nombre d'utilisateurs.

En résumé, le CASD :

- 800 utilisateurs en France et dans la zone Euro
- Près de 300 projets de recherche
- 200 SD-Box installées en France et en Europe
- 80 sources de données mises à disposition

Quels sont les retours des utilisateurs ?

Une enquête de satisfaction a été menée en mars 2014 dans le but de collecter les opinions de l'ensemble des utilisateurs du CASD sur différents critères, allant de la procédure d'accès au support informatique, en passant bien sûr par les dimensions statistiques. Les résultats de l'enquête sont globalement très positifs avec une satisfaction moyenne de 7/10 et un niveau de satisfaction record pour le support informatique (9/10 de moyenne). En revanche, les notations concernant la documentation des données en cours de réalisation, les possibilités d'entrées de programmes et de données, de sorties des résultats et la facturation (passage du gratuit au payant) mise en place récemment sont inférieures à celles des services informatiques.

Une utilisation plus large qu'initialement prévue

Un hébergement plus large que la sphère Insee

La technologie du CASD a été initialement conçue dans le but de fournir un accès sécurisé aux données individuelles très détaillées de l'Insee. Cet objectif n'est pas spécifique aux données de l'Insee. Au fil du temps, d'autres organismes détenteurs de données ont demandé, parfois sous l'impulsion des chercheurs, au CASD d'héberger et de mettre à disposition leurs données. C'est par exemple le cas des ministères de la justice et de l'agriculture, de la banque publique d'investissement, et sera bientôt le cas des données fiscales du ministère des finances

Dans le secteur privé, une grande banque ainsi qu'une grande compagnie d'assurance ont sollicité le CASD. Elles désiraient mettre des données confidentielles à disposition d'utilisateurs extérieurs (chercheurs, consultants) pour qu'ils puissent réaliser à distance des études sur leurs données. Pour cela, le CASD a créé, selon la même technologie, deux

environnements séparés de l'environnement des données mises à disposition des chercheurs, permettant d'offrir un accès sécurisé aux données confidentielles de ces compagnies.

Un hébergement pour des données volumineuses : Big Data

L'émergence des technologies du Big data qui permettent le traitement à grande échelle d'importantes quantités de données ne doit pas occulter la nature même de ces données, surtout lorsqu'il s'agit, pour la plupart, de données à caractère personnel. La collecte massive, et parfois en temps réel, de ce type de données ne peut se faire sans élargir le champ des possibilités en matière de protection technique ou juridique. Le CASD a commencé en 2012 à investir dans la recherche et le développement autour de ces nouvelles technologies de traitement de données volumineuses.

L'Institut Mines Telecom et le Genes se sont associés dans le cadre du projet Teralab[2] qui consiste à mettre en œuvre une plateforme Big Data. Dans ce projet, le CASD intègrera dans son environnement sécurisé un cluster Hadoop (ensemble de serveur pour le traitement et le stockage distribué).

Le CASD peut constituer ainsi une structure d'avenir pour l'hébergement sécurisé de données confidentielles de grand volume comme cela est souvent le cas pour les données de santé.

Un rôle de tiers de confiance

Le CASD est une direction du GENES, établissement public à caractère scientifique, culturel et professionnel (EPSCP). Le GENES a donc une personnalité juridique lui conférant par là même une indépendance juridique et lui permettant d'engager sa responsabilité en cas de défaillance.

En plus de mettre à disposition des données confidentielles de manière sécurisée, le CASD documente les données qu'il met à disposition, aide les chercheurs dans l'utilisation de la box, forme les utilisateurs et réalise les appariements de données.

En 2011, le centre d'accès sécurisé aux données (CASD) a été lauréat de l'appel à projet national « équipement d'excellence » (Equipex). Le projet soumis présentait six grands axes de développement dont un sur les méthodologies d'appariement et la possibilité pour le CASD d'être tiers de confiance pour les appariements de données[3].

Pour remplir son rôle de tiers de confiance, le CASD apporte des garanties en matière de sécurité et de confidentialité des données. De plus, les agents du CASD n'ont pas d'intérêt direct dans l'utilisation des données sources ou résultantes. Il n'y a pas de chercheur ou chargé d'étude au sein du CASD qui pourrait utiliser indument les données ainsi enrichies ou les résultats en découlant.

Ainsi, le CASD agit comme un tiers de confiance entre le producteur des données et l'utilisateur. Néanmoins, le producteur de données reste propriétaire des données et définit lui-même les modalités d'accréditation des utilisateurs lorsqu'elles ne sont pas encadrées par des textes législatifs.

Quelques exemples à l'étranger

Les données médico-administratives constituent un gisement important d'informations pour la compréhension de notre système de santé, son pilotage mais aussi la sécurité sanitaire, l'épidémiologie et la recherche scientifique en général.

L'accès aux données de santé est une question qui se pose dans plusieurs pays.

[2] Le projet est financé à hauteur de 5,6 millions d'euros dans le cadre des investissements d'avenir.

[3] Le CASD est déjà impliqué dans plusieurs projets d'appariement de données que ce soit avec des données relevant du champ de la santé : enquête sur la santé et la protection sociale de l'Institut de recherche et documentation en économie de la santé (Irdes) ou l'enquête Capacités, aides et ressources des séniors de la direction de la recherche, des études, de l'évaluation et des statistiques (Drees) ou avec des données socio-économiques. Des expérimentations sont aussi menées avec l'Institut thématique multi-organismes de santé publique de l'alliance nationale pour les sciences de la vie et de la santé (Inserm) pour l'enrichissement de cohortes.

En France, certaines données sont déjà mises à disposition auprès d'un nombre restreint d'utilisateurs sous la forme de fichiers chiffrés. La procédure d'obtention des données est souvent jugée lourde, longue et complexe. Aux États-Unis, le CMS (Center for Medicare and Medicaid Services) propose depuis de nombreuses années l'accès aux données médico-administratives également sous la forme de fichiers chiffrés. Outre la procédure elle-même, le coût d'acquisition des données est un facteur le plus souvent dissuasif pour les utilisateurs (100 000 $ par année de données).

Ces dernières années, la mise à disposition des données a été repensée de manière à accroître la sécurité des données, simplifier les procédures d'accès et réduire les coûts de mise à disposition. Ainsi de nombreux pays ont investi ou investissent dans la mise en place de centres d'accès sécurisés.

Par exemple, le CMS a ouvert en 2013, un nouveau système d'accès sécurisé distant aux données médico-administratives avec comme objectifs, entre autres, de réduire les coûts d'accès ainsi que les coûts d'hébergement pour les utilisateurs. Le SSI (Statens Serum Institut) au Danemark travaille également sur la mise en place de procédures d'accès sécurisé distant pour les données de santé. En Grande Bretagne, plusieurs initiatives sont en cours pour ouvrir des accès distants aux données de santé. C'est ainsi qu'en Écosse, l'eDRIS (electronic Data Research and Innovation Service) a mis en place un système d'accès sécurisé aux données de santé permettant aux chercheurs de travailler à distance à l'aide logiciels statistiques tels que SPSS, SAS, STATA et R. En Irlande, le Health Research Board démarre un projet de mise en place d'une infrastructure sécurisée pour faciliter la mise à disposition, le partage et les appariements de données de santé. En Allemagne, les données médico-administratives (équivalent du Programme de médicalisation des systèmes d'information PMSI et d'un grand échantillon généraliste des bénéficiaires en France EGB) sont mises à disposition au sein des centres d'accès sécurisé de l'institut national de statistiques allemand (Destatis). Des réflexions ont aussi lieu au niveau européen pour faciliter l'accès aux données de santé. Le programme de la commission européenne « Horizon 2020 » prévoit plusieurs appels à projets dans cet objectif. L'émergence des technologies du Big Data, qui facilitent le stockage et l'analyse de grands volumes de données, rend ce sujet d'autant plus prometteur qu'il ouvre de nouvelles possibilités d'exploitation des données de santé.

COMMENT ANONYMISER LES DONNÉES : UN PANORAMA NON EXHAUSTIF DES MÉTHODES D'ANONYMISATION

Maxime BERGEAT (INSEE), Dominique BLUM (Expert PMSI), Nora CUPPENS (CNRS, IMT), Frédéric CUPPENS (CNRS, IMT), Françoise DUPONT (INSEE, CASD), Noémie JESS (DREES)[1]

À l'heure actuelle, dans un contexte politique et social qui prône l'ouverture des données ou *open data*, la volonté de diffuser des fichiers de données individuelles va croissant. Pour mettre à disposition de telles informations, tout en protégeant la vie privée des individus et le secret des affaires des entreprises, il est nécessaire de masquer certaines informations. Cette anonymisation est essentielle pour conserver la confiance des participants au recueil des données quelle que soit leur origine (enquête statistique, recueil administratif, collecte d'information dans le cadre de l'exercice d'une activité commerciale).

Les risques de ré-identification sont inséparables du contexte d'utilisation du fichier. En effet la tentation de ré-identifier n'est pas la même selon les profils des personnes qui accèdent au fichier, les informations disponibles ne sont pas les même, la capacité de rapprocher le fichier qu'on cherche à protéger d'autres fichiers dépend du niveau de sécurité autour de l'accès au fichier. De ce point de vue l'open data est le contexte d'utilisation le plus risqué puisqu'il permet à tous les profils d'accéder aux informations et de les croiser avec toutes les autres informations disponibles sur les sites internet ou que l'individu possède soit dans un fichier soit par connaissance d'éléments de la vie de la personne qu'il souhaite ré-identifier. Par ailleurs, une fois diffusé, le fichier qui ne peut être croisé au moment de sa diffusion, peut l'être plus tard, une fois d'autres informations diffusées sur internet par ailleurs. On est donc dans un contexte où le risque d'attaque est très fort.

Ce dossier propose un tour d'horizon rapide des méthodes les plus connues à ce jour. Ce domaine fait l'objet de nombreuses recherches et de nombreux écrits. Avec le développement de l'Open data et des mégadonnées (Big Data) la protection de la confidentialité est un domaine en pleine effervescence pour lequel il est difficile à ce jour de disposer d'un consensus clair sur les méthodes à utiliser en pratique.

Les risques de ré-identification

Avant toute chose, il faut expliciter le risque de ré-identification.

Définitions

Les personnes cherchant à ré-identifier une personne dans un fichier de données (appelées « attaquants » dans la suite) peuvent avoir des objectifs assez différents :

- rechercher une personne parce qu'elle est connue et divulguer cette information dans la presse,
- avoir des raisons personnelles ou professionnelles d'en savoir plus sur une personne ou une entreprise particulière,
- ou encore chercher à démontrer une faille dans la sécurité du système (dans ce cas, ce qui est visé, c'est la divulgation d'information d'une personne quelconque de la base).

Ces attaquants ne disposent pas nécessairement tous au départ de la même information. Une démarche de protection des données repose sur des hypothèses sur ce que les attaquants connaissent au départ et sur ce qu'ils recherchent et que l'on veut protéger. Ces hypothèses amènent ensuite à préciser la nature des variables contenues dans le fichier de données.

[1] Les auteurs ont bénéficié d'échanges avec D. Domingo-Ferrer (Universitat Rovira i Virgili, Spain, UNESCO) et Benjamin Nguyen (INRIA- Université de Versailles).

Un fichier peut contenir plusieurs types d'informations (Tableau 1) :

- des variables directement identifiantes : numéro SIREN pour une entreprise, NIR (numéro de sécurité sociale) ou adresse complète pour un individu,
- des informations indirectement identifiantes, dites « quasi-identifiants », qui, lorsqu'on les combine, permettent à un utilisateur du fichier de retrouver une personne donnée (sexe, âge, commune d'habitation…), même si la donnée d'une seule variable indirectement identifiante ne permet généralement pas la ré-identification.
- des informations sensibles et non identifiantes qui sont des informations relatives à un individu (maladie, mode de vie, informations fiscales, etc.) ou à une entreprise (prix d'achat, marges, clients, etc.) ne devant en aucun cas, pour un individu (respectivement une entreprise), être révélées. Ces informations peuvent potentiellement faire l'objet d'une « attaque ». Elles peuvent être rendues publiques mais à condition qu'il soit impossible de les relier[2] à l'individu (respectivement l'entreprise) auquel elles se rapportent.
- des informations non sensibles et non identifiantes.

Les méthodes de réduction du risque de divulgation présentées dans la suite de ce document reposent sur une répartition des variables du fichier initial entre les variables sensibles, les variables indirectement identifiantes et les variables directement identifiantes (supprimées ou pseudonymisées). Il peut être difficile en pratique d'effectuer cette distinction entre les variables. Un employeur pourrait par exemple chercher à connaître les dates de séjours de l'un de ses employés de manière à contrôler ses justifications d'absences. Dans les tests d'anonymisation menés à partir des données du PMSI, cette donnée a été considérée comme étant une variable quasi-identifiante (Cf. Dossier 4)

TABLEAU 1

Différents types d'informations contenues dans le fichier de données initial avant toute démarche de protection

Pseudonyme	**Identifiants directs**	**Identifiants indirects**			**Variable sensible non identifiante**
	Nom complet	*Age*	*Sexe*	*Code postal*	*Maladie*
290388	Frida Kahlo	46 ans	Femme	42300	Cirrhose
276209	Niki de Saint Phalle	46 ans	Femme	73270	Bronchite
251057	Louise Moillon	68 ans	Femme	73270	Cancer du sein
186704	Berthe Morisot	111 ans	Femme	73270	Hépatite C
219687	Pablo Picasso	17 ans	Homme	75014	Insuffisance cardiaque
223818	Joan Miro	31 ans	Homme	75014	Bronchite
182604	Georges Braque	42 ans	Homme	93120	Grippe

Exemple de rubriques d'un fichier nominatif à partir duquel on fabrique des fichiers pseudonymisé (par suppression des identifiants directs) puis des fichiers diffusables. (par exemple en rendant moins précis les identifiants indirects

La littérature sur les méthodes d'anonymisation distingue différents risques [1] :

- Le risque de révélation de l'identité (« record linkage ») : on retrouve l'identité d'une personne qui est présente dans le fichier, sans forcément en déduire de l'information supplémentaire par rapport à ce qu'on sait déjà. Par exemple, un journaliste reconnaît dans le fichier que l'enregistrement « xxxxx » appartient au président de la république, mais sans apprendre de nouvelles informations sur ce dernier.
- Le risque de révélation d'attribut (« attribute linkage ») : on obtient des informations sensibles (non perturbées), comme par exemple la maladie, sur un individu reconnu (à partir de variables quasi-identifiantes, plus ou moins notoires, telles que l'âge, le sexe, le lieu de résidence par exemple). Ces informations sont relatives à la personne et leur divulgation peut dans certains cas lui être préjudiciable. Par exemple, si dans le fichier diffusé, tous les hommes de plus de 70 ans vivant dans une commune donnée ont eu une prescription pour un somnifère, alors si l'attaquant sait que Monsieur D. a 75 ans et vit dans cette commune, il apprendra que Monsieur D. a pris des somnifères, même si l'enregistrement correspondant lui est inconnu.

[2] Soit de façon directe, soit en dévoilant le groupe (patients atteints d'une maladie par exemple) auquel un individu ou une entreprise appartient. Pour plus de détails voir plus bas la présentation des critères de k-anonymisation et de l-diversité.

- Le risque de révélation inférentiel (« probabilistic attack ») : l'attaquant infère, avec une probabilité importante, de l'information significative par rapport à ce qu'il savait au départ. Par exemple, s'il ne dispose pas initialement d'information sur la maladie d'une personne et qu'il apprend qu'il s'agit d'une maladie grave, il a obtenu une information très précise par rapport à ce qu'il savait déjà.

Idéalement, avant d'entreprendre une démarche d'anonymisation pour un fichier, il faut expliciter le risque de divulgation d'informations confidentielles qu'il contient en analysant les variables qu'il contient et l'information extérieure disponible (autres fichiers de données, informations personnelles sur les réseaux sociaux, agendas en ligne etc.) ; cette analyse permet d'établir un classement des variables selon leur nature (variables indirectement identifiantes selon leur degré de notoriété et informations sensibles que l'on cherche à protéger). Cette analyse qui peut comporter une part de subjectivité, est réalisée à un moment donné avec un état de l'information auxiliaire mobilisable pour ré-identifier. Cette information est susceptible de s'enrichir au fil du temps. Dans un contexte de diffusion régulière dans un environnement d'open data, un fichier déjà diffusé a pu être téléchargé : il n'est pas possible de revenir en arrière concernant cette diffusion. Il faut donc tenir compte de ce fait pour mettre au point le dispositif de diffusion.

Différents critères de protection des données

Au fil du temps, différents critères de protection ont été mis au point. Ce dossier ne détaille que les plus utilisés[3] (usuels). Pour protéger un fichier, on est amené à réduire le risque associé aux individus les plus facilement ré-identifiables.

Le k-anonymat et la l-diversité

Un premier critère de protection des données est le k-anonymat**.** Un fichier est dit k-anonymisé si, pour toute clé d'identification[4], il existe au moins *k* individus indistinguables du point de vue des variables quasi-identifiantesdans le fichier diffusé. La k-anonymisation protège du risque de révélation de l'identité (« *record linkage* »), mais pas des autres risques. Sans information, autre que la clé d'identification, un individu a moins d'une chance sur *k* d'être retrouvé pour un fichier k-anonymisé.

Toutefois, dans un fichier k-anonymisé, si tous les individus ayant la même clé partagent la même valeur d'une variable sensible (par exemple s'ils sont tous atteints de la même maladie), même s'il existe moins d'une chance sur *k* d'identifier un individu, il y a divulgation d'information sur la variable sensible (dans notre exemple, la maladie). On a ici un cas de divulgation d'attribut pour un groupe d'individus (« *group disclosure* »). Pour protéger davantage le fichier, le concept de l-diversité a été défini. Un fichier est dit l-diversifié si, pour chaque clé d'identification *c*, il existe au moins *l* modalités représentées pour chaque variable sensible [2]. Néanmoins, le risque de divulguer une information sur la variable sensible dépend des effectifs pour chacune des *l* modalités. Si la distribution de la variable sensible n'est pas uniforme au sein d'une clé d'identification, alors elle peut être divulguée avec une forte probabilité. Par exemple, pour une clé d'identification donnée, composée de 10 individus (k=10) avec 2 maladies A et B (l=2)) représentées, si un seul individu à la maladie A et que tous les autres ont la maladie B, alors on peut inférer avec une probabilité de 9/10 qu'un individu ayant la même clé d'identification et dont on chercherait à identifier la maladie est atteint de la maladie B. Si cette répartition est très différente de celle dans la population générale, on a augmenté significativement son information.

D'autres critères ont été élaborés pour augmenter le niveau de protection du fichier.

[3] Pour une liste plus complète voir [1].

[4] On appelle clé d'identification, notée *c*, une combinaison des différentes modalités possible des variables quasi-identifiantes. Par exemple, si l'on considère les variables quasi-identifiantes suivantes : l'âge, le sexe, le code postal du lieu de résidence ou d'hospitalisation, soit pour un homme de 50 ans résidant dans l'Ain à Bourg-en-Bresse *c* := {Homme ; 50 ans ; 01000}.

La t-proximité

La t-proximité permet d'augmenter le niveau de protection d'un fichier qui serait k-anonymisé et l-diversifié [3]. Ce critère requiert que la distribution des variables sensibles pour les individus ayant une même clé d'identification soit assez proche de la distribution sur la totalité de la population (distance entre les deux distributions inférieure à *t)*. Concrètement, dans l'exemple précédent, pour chaque clé d'identification, la répartition des effectifs du fichier selon les variables sensibles (dans notre cas, la maladie) doit être suffisamment proche de celle de la population totale. La diffusion de ce fichier n'apporterait donc pas d'information supplémentaire sur un individu. On peut se poser la question de l'utilité effective pour les utilisateurs du fichier résultant.

La confidentialité différentielle

La confidentialité différentielle est un objectif de réduction du risque généralement utilisé dans un contexte où la perturbation est introduite lorsque l'utilisateur du fichier effectue une requête (en demandant le calcul d'une statistique, par exemple) sur le fichier original (qui n'est alors pas diffusé, seul le résultat perturbé est transmis à l'utilisateur) [4] [5]. Afin de satisfaire à la propriété de confidentialité différentielle, la réponse à la requête demandée par l'utilisateur est bruitée[5]. Le bruit ajouté est calculé de façon à ce qu'un individu rentrant en compte dans le calcul de la requête ne puisse être identifié. L'objectif est que la réponse à la requête ne soit que très peu modifiée (le "très peu" étant une quantité à définir par la personne en charge de la protection des données) lorsqu'on ajoute ou enlève une observation pour la calculer. Par exemple, si l'utilisateur demande de calculer le revenu moyen des habitants de Lille, la quantité de bruit introduite dans la réponse est calculée de façon à ce que le revenu moyen par habitant soit "très peu" différent de celui établi en excluant l'individu qui possède le revenu le plus élevé. Dans ce cas-là, plus la distribution des revenus lillois est étendue, plus la quantité de bruit à ajouter lors du calcul du revenu moyen doit être importante.

Une démarche d'anonymisation des données consiste à analyser le risque de ré-identification d'un ou plusieurs individus (ou entreprises) du fichier que l'on souhaite mettre à disposition, puis à adopter une méthode de protection, en choisissant astucieusement la perte d'information à consentir, donc les critères de protection à respecter, pour préserver au maximum son utilité auprès des différents utilisateurs.

Méthodes de protection

Il existe différentes méthodes de protection des données : certaines méthodes perturbatrices altèrent les données initiales, d'autres non perturbatrices (ou « restrictives ») réduisent la quantité d'information mise à disposition de l'utilisateur [7]. Ce chapitre dresse un panorama non exhaustif de ces méthodes.

Les limites de la pseudonymisation dans un contexte d'open data

Un premier réflexe qui vient à l'esprit lorsqu'on cherche à anonymiser des données consiste à remplacer l'identifiant initial d'une personne (par exemple le numéro de sécurité sociale ou NIR pour un individu, le numéro SIREN pour une entreprise) par un autre identifiant arbitraire - appelé **pseudonyme** - qui ne contient pas d'information connue par les futurs utilisateurs du fichier.

Il existe plusieurs méthodes pour « **pseudonymiser** » un jeu de données.

L'identifiant initial peut être remplacé par une simple numérotation incrémentale, pourvu que l'ordre du fichier associé ne donne aucune information. Pour cela, il peut par exemple être trié au préalable sur la base d'un nombre généré aléatoirement. La correspondance entre chaque identifiant et son pseudonyme peut être stockée dans une table de passage. La protection des données repose sur le fait que cette table reste confidentielle, conservée uniquement de façon sécurisée, pour assurer la traçabilité des informations.

On peut également appliquer une fonction de hachage sur l'identifiant initial (un exemple de procédé sécurisé utilisant le hachage (dénommée Fonction d'Occultation des Informations Nominatives (FOIN) est donné en Annexe). Ces fonctions

[5] Il existe également des extensions du concept permettant de construire des jeux de données qui respectent l'objectif de confidentialité différentielle [6].

transforment de façon irréversible une chaîne de caractère numérique en une autre chaîne de caractère de longueur fixe, appelée empreinte. Les fonctions de hachage sont irréversibles, c'est-à-dire qu'il est impossible de retrouver l'identifiant initial à partir du seul pseudonyme, même si l'on connaît la fonction de hachage utilisée.

Cette façon de faire permet facilement de mettre en place une organisation qui permet de maitriser qui accède à des informations sur l'identifiant.

Néanmoins, la pseudonymisation a ses limites : si le remplacement ou la suppression de l'identifiant individuel initial est nécessaire pour protéger des données confidentielles, il n'est pas suffisant. En effet, il est souvent possible d'identifier la personne recherchée, même si le fichier ne contient pas d'identifiants directs, grâce à d'autres informations très identifiantes également présentes dans la base.

Deux exemples de fichiers insuffisamment protégés sont célèbres dans la littérature sur les méthodes de protection des données :

- En 2006, l'entreprise américaine AOL, fournisseur d'accès à internet, a publié en ligne une base de données qui rassemblait 20 millions de recherches effectuées sur son site par 650 000 utilisateurs sur trois mois. La base ne contenait pas d'information directement identifiante (comme l'adresse IP par exemple ou le nom de l'internaute) mais conservait les liens entre toutes les recherches d'un même utilisateur. En observant la liste des requêtes quotidiennes sur plusieurs mois, des journalistes du New York Times sont parvenus à retrouver F. Thelma Arnold, 62 ans, habitant Lilburn, Georgie, créant ainsi de sérieux dégâts d'image et entrainant la démission de deux hauts responsables et la mise à disposition d'un antivirus gratuit pour les abonnés à partir de cette date.
- La société américaine Netflix, qui offre un service de location en ligne de DVD et permet ensuite à ses utilisateurs de noter les films qu'ils ont visionnés (ces appréciations permettant ensuite à Netflix de recommander des films en fonction des goûts de ses clients), a également connu des problèmes de sécurité se soldant par un procès de la part de ses utilisateurs. Dans le cadre d'un concours visant à améliorer ses algorithmes d'analyse des préférences des utilisateurs, la société Netflix a publié en 2010 un fichier ne contenant pas d'information directement identifiante mais seulement la note donnée par un utilisateur à un film et sa date, afin que des développeurs indépendants proposent des algorithmes plus performants que ceux qu'elle utilisait. Des chercheurs en sécurité ont réussi à montrer qu'à partir de ce fichier, il était possible de retrouver certains individus en recoupant leurs appréciations sur trois films et leurs dates avec des appréciations identiques faites aux mêmes dates, disponibles dans la base.

Ces deux exemples montrent qu'il faut réfléchir différemment en termes d'anonymisation d'un fichier de données, la simple pseudonymisation n'étant pas suffisante. Ainsi, l'anonymisation d'un fichier doit intégrer la distinction entre les informations que l'on cherche à protéger et les variables pouvant être connues par un tiers et servir pour ré-identifier un individu. Des objectifs de réduction du risque de ré-identification peuvent alors être introduits. Ensuite, il faut savoir quantifier la perte d'information que l'on est prêt à consentir pour limiter le risque de ré-identification. Enfin, en faisant un arbitrage entre ces critères de réduction du risque et l'utilité du fichier résultant, une procédure de réduction du risque peut être définie [8].

Méthodes non perturbatrices

Regroupement de modalités (« généralisation »)

Un procédé couramment utilisé pour protéger un fichier de données individuelles consiste à publier une information moins précise en généralisant ou regroupant des données en plus grandes catégories (Tableau 2), l'objectif étant de réduire le nombre d'enregistrements comportant une combinaison rare de variables permettant potentiellement la ré-identification (quasi-identifiants). Par exemple, on peut remplacer l'âge exact d'un individu par une tranche d'âge choisie pour limiter le risque de ré-identification et en fonction de sa pertinence par rapport aux utilisations des données. On recode ainsi les variables en utilisant une nomenclature plus agrégée. On peut également recoder les variables temporelles : par exemple en annualisant les dates précises d'un séjour à l'hôpital. Enfin, pour éviter le risque de divulgation lié aux valeurs extrêmes très identifiantes, des recodages peuvent être effectués pour les valeurs extrêmes de la distribution (création d'une tranche « 80 ans ou plus » par exemple pour l'âge).

Plus le niveau d'agrégation est important, moins il y a de risques de ré-identification, mais moins l'information est précise. Le producteur de données doit donc arbitrer entre le risque encouru et la perte d'information liée aux regroupements de modalités sous la contrainte de respecter le niveau de risque fixé[6]. Pour les variables dont les modalités possibles correspondent à une nomenclature à plusieurs niveaux, des possibilités naturelles d'agrégation existent (par exemple le code postal peut être agrégé aux niveaux départemental et régional). Pour les variables spécifiques au domaine d'étude, le recours à des experts métiers permet de conserver l'information la plus pertinente pour les utilisateurs. Cette méthode est adaptée pour obtenir un fichier k-anonymisé et même l-diversifié, voire t-proche.

TABLEAU 2

Exemple de fichier, 3- anonymisé et 3-diversifié à partir du regroupement de modalités

Le fichier présenté dans le tableau 1 devient :

Pseudonyme	*Tranche d'âge*	*Sexe*	*Région*	*Maladie*
290388	Plus de 45 ans	Femme	Rhône-Alpes	Cirrhose
276209	Plus de 45 ans	Femme	Rhône-Alpes	Bronchite
²251057	Plus de 45 ans	Femme	Rhône-Alpes	Cancer du sein
186704	Plus de 45 ans	Femme	Rhône-Alpes	Hépatite C
219687	Moins de 45 ans	Homme	Ile-de-France	Insuffisance cardiaque
223818	Moins de 45 ans	Homme	Île-de-France	Bronchite
182604	Moins de 45 ans	Homme	Île-de-France	Grippe

Il y a, en pratique, plusieurs façons d'appliquer la généralisation. Si on met sous forme d'arbre les différents niveaux de regroupement possibles pour les quasi-identifiants, la généralisation consiste à « remonter dans l'arbre » (du plus détaillé au plus général). On peut remonter pour tous les individus à un niveau donné, on parle alors de recodage global (« *global recoding* »), ou seulement pour les individus trop peu nombreux dans leur catégorie, on parle alors de recodage local («*local recoding* »). Dans ce dernier cas, la perte d'information est moindre, mais le niveau de détail diffusé n'est pas homogène dans tout le fichier et l'utilisation en est plus délicate pour l'utilisateur. Celui-ci pourra être amené à faire un recodage global pour pouvoir utiliser l'ensemble du fichier pour certains de ses traitements.

Suppressions locales

Pour réduire le risque de ré-identification, on peut également éliminer les valeurs des variables quasi-identifiantes pour les individus présentant un risque de ré-identification trop fort (Tableau 3). Typiquement, pour obtenir le k-anonymat, on remplace par des valeurs manquantes les modalités de certaines variables quasi-identifiantes pour les individus possédant des clés d'identification pour lesquelles il y a moins de *k* individus ayant la même clé. Il est alors possible de définir des programmes de minimisation de la perte d'information (en attribuant un coût de suppression par variable, par exemple) sous contrainte de respecter un objectif de réduction du risque, par exemple le k-anonymat. Des logiciels comme µ-Argus et Arx permettent de réaliser ces suppressions « optimisées » (Cf. Dossier 4).

TABLEAU 3

Exemple de fichier 2-anonymisé et 2-diversifié à partir de suppressions locales

Le fichier présenté dans le tableau 1 devient :

Pseudonyme	*Âge*	*Sexe*	*Région*	*Maladie*
290388	46 ans	Femme	-	Cirrhose
276209	46 ans	Femme	-	Bronchite
251057	-	Femme	73 270	Cancer du sein
186704	-	Femme	73 270	Hépatite C
219687	-	Homme	75 014	Insuffisance cardiaque
223818	-	Homme	75 014	Bronchite

[6] Certains pays disposent en effet d'un règlement en matière de diffusion de données, définissant un cadre d'où on ne peut pas sortir.

Pseudonyme	*Âge*	*Sexe*	*Région*	*Maladie*
182604	-	Homme	-	Grippe

L'avantage de cette méthode est qu'elle ne réduit l'information apportée par les données que pour les individus qui posent problème. En revanche, elle rend le maniement du fichier résultant plus difficile pour les utilisateurs, en particulier pour les non experts.

Diffusion d'un échantillon de données

La création d'un échantillon à partir du fichier initial (qu'il soit exhaustif ou non) que l'on cherche à protéger est une méthode introduisant une incertitude supplémentaire.

En effet, si un attaquant trouve dans l'échantillon une personne présentant les mêmes caractéristiques (dates de séjours hospitaliers, date de décès par exemple) que celles qu'il recherche, il ne peut avoir la certitude qu'il n'y a pas dans la base de données (à partir de laquelle l'échantillon a été tiré) un autre individu présentant les mêmes caractéristiques.

Néanmoins, pour garantir une protection suffisante, la règle de tirage de l'échantillon doit rester secrète, il ne doit pas être possible d'accéder à la base originale à partir de laquelle a été tiré l'échantillon et on ne doit pas disposer d'information statistique fine qui puisse être croisée avec des statistiques issues de l'échantillon. Si ces règles sont respectées, on se protège contre le risque de divulgation de l'identité et de l'attribut des individus.

Néanmoins, cette méthode ne protège pas si l'échantillon contient des observations éminemment atypiques :

- si l'on trouve un individu présentant des caractéristiques uniques notoires[7], tout le monde peut l'identifier,
- si un attaquant constate que l'individu qu'il recherche a des caractéristiques rares uniques dans l'échantillon et si atypiques[8] qu'il est très peu probable (proche de zéro) qu'une autre personne ait les mêmes.

La méthode d'échantillonnage peut être combinée à d'autres méthodes de protection des données [9]. L'échantillonnage des données peut apporter une protection supplémentaire qui permet potentiellement de couper le lien entre des jeux de fichiers successifs dans le cadre d'une diffusion régulière sur la même population de référence.

Méthodes perturbatrices

De façon générale, ces méthodes consistent à altérer les données originales (quasi-identifiants ou données sensibles) en les remplaçant par des données perturbées de manière à ce que les statistiques calculées à partir de ce nouveau jeu de données ne diffèrent pas significativement de celles obtenues à partir des données originales. Les enregistrements perturbés ne correspondent plus à ceux du fichier initial et l'attaquant ne peut recouvrer des données sensibles à partir des données rendues publiques. En d'autres termes, l'utilisateur du fichier diffusé ne peut plus faire un lien certain entre les données réelles (initiales) et les données perturbées.

Ce chapitre présente un bref descriptif des principales méthodes de perturbation de données : la microagrégation, l'ajout de bruit, la permutation des données et la génération de données synthétiques.

Bruitage des données

Cette technique est souvent utilisée pour masquer des données numériques sensibles (la date de naissance par exemple) et peut être également appliquée à des données catégorielles. Le principe de cette méthode est d'ajouter un aléa aux variables quasi-identifiantes ou sensibles, tout en conservant certaines propriétés de leur distribution (par exemple médiane, écart type, etc.). Pour des variables quantitatives, on peut ajouter une quantité aléatoire issue d'une loi de probabilité maîtrisée par le propriétaire du fichier et d'espérance nulle, afin de ne pas introduire de biais dans les estimations. Par exemple, un nombre aléatoire de jours peut être ajouté à la date de naissance pour en créer une version bruitée.

[7] Si on trouve par exemple dans un échantillon d'enquête, réalisée auprès de la population française, une personne ayant l'âge du doyen de la population française, on peut aisément en déduire qu'il s'agit précisément du doyen de la population française.

[8] Si un attaquant sait que son voisin, âgé de 15 ans, a consulté plus de 10 fois au cours de l'année N un pédicure-podologue, alors s'il trouve dans l'échantillon un bénéficiaire présentant les mêmes caractéristiques, il y aura une probabilité importante (proche de 1) que l'individu identifié soit son voisin.

Pour les variables qualitatives comme le sexe, la profession, le lieu d'habitation, les maladies, on peut modifier une faible proportion des enregistrements en les remplaçant par une valeur différente choisie grâce à un mécanisme défini à l'avance. En pratique, pour chaque variable, des probabilités de transition sont définies (par exemple pour la variable « sexe », les hommes ont 90 % de chance conserver la valeur « homme » et 10 % de se voir attribuer la valeur» femme »), et ce mécanisme de perturbation est ensuite appliqué afin d'introduire de l'incertitude dans le fichier résultant, pour réduire le risque de ré-identification.

En jouant sur la valeur de ces probabilités de transition, on modifie le niveau de protection sur les variables et la perte d'utilité du fichier pour l'utilisateur.

Toutefois, l'ajout d'un aléa mal déterminé peut conduire à des incohérences, qui pourraient permettre à un attaquant de déconstruire le processus de protection du fichier et de ré-identifier des individus. Un exemple d'incohérence est donné dans le Tableau 4 où des perturbations aléatoires ont été réalisées en prenant en fichier de départ le Tableau 1. On obtient ici un homme qui a le cancer du sein, ce qui est très peu probable[9]. Pour résoudre ce problème, il est nécessaire au départ de bien définir le mécanisme de perturbation (via les probabilités de transition) afin de ne pas créer d'individus « impossibles ».

TABLEAU 4

Exemple de fichier anonymisé par ajout de bruit avec une incohérence

Pseudonyme	Age	*Sexe*	*Code postal*	*Maladie*
290388	46 ans	Femme	42300	Cirrhose
276209	46 ans	Femme	73270	Bronchite
251057	68 ans	Homme	73270	Cancer du sein
186704	111 ans	Femme	73270	Hépatite C
219687	17 ans	Homme	75014	Insuffisance cardiaque
223818	31 ans	Homme	75014	Bronchite
182604	42 ans	Homme	93120	Grippe

Cependant, en pratique, il n'est pas aisé de donner des éléments simples et pédagogiques pour que l'utilisateur s'approprie les conséquences du bruitage des données sur les analyses qu'il va réaliser. En pratique, l'imprécision apportée sur les variables peut perturber les analyses et mener à des conclusions erronées. Par exemple, si nous considérons la donnée « femme de 45 ans atteinte d'un cancer », il n'est pas possible de savoir si elle est âgée de 50 ans ou de 40 ans. La microagrégation pose le même type de problème (âge moyen au lieu de l'âge réel).

La microagrégation

Cette technique est fondée sur une classification des individus en plusieurs groupes dont l'effectif est supérieur à un seuil *k*. Après avoir regroupé les individus semblables (par exemple avec des techniques de classification), on remplace chacun des individus du groupe par un individu dont les valeurs des variables quasi-identifiantes sont remplacées par exemple par la valeur de la « moyenne », la « médiane » (pour des variables quantitatives ou qualitatives ordinales), ou par le mode (pour des variables qualitatives non ordinales) du groupe. Le fichier obtenu en sortie est k-anonymisé. Un exemple de microagrégation reprenant les données du Tableau 1 est présenté dans le Tableau 5 ci-dessous.

TABLEAU 5

Exemple de fichier, 3-anonymisé et 3-diversifié par microagrégation

Pseudonyme	Age	*Sexe*	*Code postal*	*Maladie*
290388	*68 ans*	Femme	73270	Cirrhose
276209	*68 ans*	Femme	73270	Bronchite
251057	*68 ans*	Femme	73270	Cancer du sein
186704	*68 ans*	Femme	73270	Hépatite C
219687	*30 ans*	Homme	75014	Insuffisance cardiaque

[9] Moins de 1 % des cancers du sein touchent des hommes, cf. http://www.e-cancer.fr/cancerinfo/les-cancers/cancer-du-sein/quelques-chiffres.

Pseudonyme	Age	*Sexe*	*Code postal*	*Maladie*
223818	*30 ans*	Homme	75014	Bronchite
182604	*30 ans*	Homme	75014	Grippe

Échange, permutation (« Swap »)

Cette méthode consiste, pour une ou plusieurs variables quasi-identifiantes, à échanger entre deux individus leurs modalités (un exemple reprenant les données du Tableau 1 est présenté dans le Tableau 6). Ainsi, on se protège contre le risque de divulgation de l'identité. L'échange peut également se faire entre les modalités des variables sensibles à protéger. On se protège dans ce cas contre le risque de divulgation d'attribut.

La permutation des données peut être ainsi considérée comme une forme d'ajout de bruit, à la différence que les domaines des valeurs et les distributions univariées restent inchangés.

TABLEAU 6

Exemple de fichier, 3-anonymisé et 3-diversifié à partir de permutations

Le fichier présenté dans le Tableau 1 devient :

Pseudonyme	*Âge*	*Sexe*	*Code postal*	*Maladie*
290388	46 ans	Femme	42300	Cirrhose
276209	*46 ans*	*Homme*	73270	Bronchite
251057	68 ans	Femme	73270	Cancer du sein
186704	111 ans	Femme	73270	Hépatite C
219687	17 ans	Homme	75014	Insuffisance cardiaque
223818	*31 ans*	*Femme*	75014	Bronchite
182604	42 ans	Homme	93120	Grippe

Note : Les données du tableau anonymisés ont restées ordonnées comme dans le fichier initial. Or, le fait de n'avoir permuté que la variable sexe (sans la variable d'âge par exemple) et de ne pas avoir trié le fichier avant sa diffusion renseigne sur le mécanisme de perturbation utilisé. Cet exemple vise uniquement à illustrer le mécanisme de permutation.

En revanche, les corrélations et les distributions croisées des variables ayant subi des permutations sont modifiées.

Comme pour le « bruitage », cette méthode peut créer des incohérences. Par exemple, dans le tableau 7, il est évident que le salaire peut être deviné et que la permutation dans ce cas est inefficace. Il est très probable que le directeur général ait le salaire maximal et le chômeur le plus bas.

Ces incohérences jettent ainsi un doute sur la qualité des données pour des utilisateurs peu avertis et peuvent aider à retrouver le procédé de permutation utilisé et ainsi rompre la protection. Pour limiter la perte d'information engendrée par ces modifications, on peut ordonner le fichier selon un critère et réaliser les échanges par lignes (avec l'ensemble des variables quasi-identifiantes par exemple) entre individus proches. Les permutations peuvent également être réalisées sur la base de variables croisées (permutation conjointe du salaire et de la profession, par exemple).

TABLEAU 7

Exemple de permutation n'assurant pas la protection de la vie privée

Profession	*Sexe*	*Salaire (Attribut permuté)*
Ingénieur	Homme	450 000
Directeur général	Homme	6 000
Chômeur	Homme	100 000
Responsable des ressources humaines	Homme	50 000
Ingénieur	Homme	80 000

SOURCE : ILLUSTRATIONS DES AUTEURS.

Génération de données synthétiques

Une modélisation des données est effectuée en se basant sur le fichier de données initial. On simule ensuite des données selon le modèle estimé. Les données simulées sont ensuite mises à disposition des utilisateurs. Il est important de noter que la modélisation doit prendre en compte le plus possible les liens entre variables pour préserver l'information contenue dans le fichier initial. En effet, il ne faut pas raisonner indépendamment pour chaque variable, au risque de briser les corrélations. L'utilisation de ce type de fichier est limitée. Ces techniques sont par exemple utilisées pour mettre à disposition des jeux d'essais afin de tester des développements informatiques ou mettre au point des programmes d'analyses par des chercheurs avant de faire tourner les programmes sur les vraies données dans des environnements aux accès plus restreints.

Il est possible de mélanger dans le fichier diffusé des données simulées et des données réelles : on parle alors de données hybrides.

Conclusion

Aucune méthode de sécurisation des données n'arrive seule à satisfaire les objectifs conflictuels d'assurer à la fois la richesse des données mises à disposition, un coût faible de prétraitement et qui soit applicable à n'importe quel environnement de base de données statistiques, tout en s'assurant une protection infaillible de la vie privée. Il n'y a pas à ce jour de consensus sur la meilleure méthode. Pour assurer un taux de divulgation très faible de données personnelles (permettant la ré-identification), les techniques de perturbation de données développées jusqu'ici devraient être combinées avec celles ne modifiant pas les données comme la généralisation ou l'agrégation. L'échantillonnage des données peut apporter une protection supplémentaire qui permet potentiellement de couper le lien entre des jeux de fichiers successifs dans le cadre d'une diffusion régulière sur la même population de référence. La protection de la confidentialité qui s'avère particulièrement ardue dans un contexte d'open data fait l'objet de nombreuses recherches et d'une abondante littérature. L'état de l'art a évolué rapidement ces dernières années et les années à venir permettront probablement de dégager plus facilement un consensus sur des solutions opérationnelles outillées.

Bibliographie

Par ordre d'apparition dans le texte.

[1] B.C.M, WANG K.FUNG, R. CHEN and P.S Yu, 2010, «Privacy preserving data publishing: a survey of recent developements». ACM computing surveys, 42, 4, article 14, june, 53 pages

[3] N. Li, T. Li, et S. Venkatasubramanian. t-closeness: Privacy beyond k-anonymity and l-diversity, dans International Conference on Data Engineering, 2007.

[4] Dwork C., 2006, « Differential privacy », Proc. ICALP.

[8] Ninghui Li, Wahbeh Qardaji, and Dong Su. 2012. On sampling, anonymization, and differential privacy or, k-anonymization meets differential privacy. In *Proceedings of the 7th ACM Symposium on Information, Computer and Communications Security* (ASIACCS '12). ACM, New York, NY, USA, 32-33. DOI=10.1145/2414456.2414474 http://doi.acm.org/10.1145/2414456.241447

[6] Sánchez D., Domingo-Ferrer J., Martínez. S., 2014, « Improving the Utility of Differential Privacy via Univariate Microaggregation », lecture notes of the international Privacy in Statistical Databases's conference, Springer, LNCS 8744, 130-142.

[8] Hundepool A., Domingo-Ferrer J., Franconi L., Giessing S., Schulte Nordholt E., Spicer K., de Wolf P-P., 2012, Statistical Disclosure Control, Wiley Series in Survey Methodology.

Benedetti, R. et Franconi, L. (1998), Statistical and technological solutions for controled data dissemination, Pre-proceedings of New Techniques and Technologies for Statistics,1, 225-232.

Blum A., Ligett K., Roth A., 2013, « A Learning Theory Approach to Non-Interactive Database Privacy », STOC 2008, JACM, Volume 60, Issue 2, April.

En complément

[2] Allard T., Nguyen B., Puchétral P., 2013, « Comment préserver l'anonymat. » Pour la science, n°433, novembre.

[7] Bergeat M., 2014, « Données individuelles : bien les protéger pour mieux les diffuser », actes des Journées de Statistique de la SFdS.

[9] Smith, S. (2014), Data and privacy: Now you see me; New model for data sharing; Modern governance and statisticians. Significance, 11: 10–17. doi: 10.1111/j.1740-9713.2014.00762.x

Article 29 Data protection working party Opinion 05/2014 on Anonymisation Techniques : avis du groupe de l'Article 29 qui regroupe les autorités de protection des données européennes sur les principales techniques d'anonymisation.

Conférence régulière sur le sujet UNECE : Work session on statistical data confidentiality http://www.unece.org/index.php?id=31958

Dupriez O., Boyko E., 2010,» Diffusion des fichiers de microdonnées : Principes, procédures et pratiques ».

Introduction to Statistical Disclosure Control (SDC), IHSN (International Household Survey Network) working paper n°7, august 2014

Prada S., Gonzalez C., Borton J., Fernandes-Huessy J., Holden C., Hair E., Mulcahy T., 2011, « Avoidingdisclosure of individually identiable health information: a literaturereview », SAGE Open published online, 14 December, DOI: 10.1177/2158244011431279.

Rapport d'information du sénat, « La protection des données personnelles dans l'open data : une exigence et une opportunité ».

Urbatas C., 2009, « Les fichiers de données synthétiques et les FMGD, deux applications tirées d'enquêtes post-censitaires », Assemblée nationale de la société de statistiques canadienne, recueil de la section des méthodes d'enquêtes, juin.

Résultats d'un test mené sur l'anonymisation des données du PMSI

Maxime BERGEAT (INSEE), Nora CUPPENS (CNRS, IMT), Frédéric CUPPENS (CNRS, IMT), Noémie JESS (DREES), Françoise DUPONT (INSEE, CASD)

Ce dossier présente les résultats d'un test mené sur l'anonymisation des données du Programme de médicalisation des systèmes d'information (PMSI) avec une méthode non perturbatrice.

Ce travail avait pour vocation d'éclairer la mise au point de jeux de données anonymes pouvant être diffusés en Open Data. Il a été réalisé dans le cadre des travaux du groupe de travail sur les risques de ré-identification (RIRE) créé suite à la demande de la ministre Madame Marisol Touraine, faite à la Direction de la recherche, des études, de l'évaluation et des statistiques (DREES), de diligenter une expertise technique sur la sécurité des données concernant le risque de ré-identification des personnes à partir de données « supposées » anonymes [1].

Cette étude a été réalisée par Nora Cuppens et Frédéric Cuppens (CNRS LabSTICC et Institut Mines-Télécom Télécom Bretagne), Maxime Bergeat (INSEE), Françoise Dupont (INSEE et CASD) et Noémie Jess (DREES) [2]. Les auteurs tiennent à remercier les membres du groupe de travail ainsi que C.Quantin et D. Blum pour leur apport.

Les tests menés par le groupe de travail ont été réalisés dans un calendrier particulièrement restreint : l'accord de la Cnil (pour 3 mois renouvelables une fois) a été obtenu le 30 janvier 2014 et les données et les nomenclatures des variables ont été rendues disponibles un mois plus tard. Les travaux n'ont donc pu commencer que début mars et les conclusions ont été rendues fin avril. Pour ces raisons, le test ne représente qu'un éclairage partiel de la démarche d'anonymisation. Il ne peut constituer à lui seul une réflexion complète sur la démarche d'anonymisation dans le cadre d'une diffusion régulière de données.

Présentation des données et des logiciels

Les données du PMSI

Les tests d'anonymisation ont été menés sur les données du PMSI (Programme de médicalisation des systèmes d'information), une base médico-administrative annuelle qui rassemble la totalité des séjours hospitaliers publics et privés en France[1]. Le périmètre des tests a été restreint au secteur Médecine, chirurgie et obstétrique (MCO, également dit « secteur de court séjour »), et donc à la base PMSI-MCO. Chaque enregistrement (près de 26 millions en 2012) dans la base correspond à un seul séjour à l'hôpital dans le secteur de court séjour. Nous n'avons pas pris en considération le chaînage des séjours entre eux, c'est-à-dire l'ensemble des séjours pour un patient donné. Des travaux antérieurs[2] ont montré que la quasi-totalité des personnes ayant eu plus d'un séjour dans l'année étaient uniques si l'on combinait leurs caractéristiques sociodémographiques (âge, sexe, code géographique de résidence[3]) et celles de leur hospitalisation (établissement d'hospitalisation, mois de sortie, mode de sortie, durée des hospitalisations et délais entre les hospitalisations). Enfin, pour ne pas rendre plus complexe l'analyse[4], nous avons exclu de la base les séjours comportant des séances (chimiothérapie, dialyse ou radiothérapie). En effet, ceux-ci peuvent indifféremment faire l'objet d'un seul enregistrement cumulant l'ensemble des séances, ou d'un enregistrement par séance, et cette liberté organisationnelle est susceptible de perturber l'évaluation du risque de ré-identification.

[1] Pour les besoins du test, ces données ont été mises à disposition au Centre d'accès sécurisé distant aux données (CASD).

[2] Cf. travaux de Dominique Blum sur les données du PMSI-2008. Des travaux similaires ont été menés en 2000 mais sur des données démographiques

[3] Il s'agit d'un code spécifique au PMSI permettant de repérer le lieu de résidence du patient hospitalisé avec un niveau plus agrégé que celui des codes postaux. Le même code PMSI est associé à des codes postaux différents lorsque leur population est inférieure à 1000 habitants (Source : ATIH).

[4] L'exclusion des séjours avec séances est une pratique courante pour les études menées à partir des données du PMSI.

La base PMSI-MCO pour l'année 2012 à partir de laquelle les jeux de données anonymisés (selon les critères de réduction du risque retenus dans le test) ont été construits, contient ainsi 20,6 millions de séjours hospitaliers. Les informations disponibles pour chaque séjour sont très détaillées et l'on peut distinguer d'une part les informations « sensibles » couvertes par le secret médical (essentiellement le diagnostic principal, les diagnostics associés et les actes chirurgicaux) qu'un « attaquant » potentiel cherche à connaître, et d'autre part deux catégories d'informations indirectement identifiantes (ou quasi-identifiantes)[5], pouvant être également « sensibles » : les informations administratives (principalement le numéro Finess d'identification de l'établissement, la durée du séjour, les dates de séjour, les modes d'entrée et de sortie) et les informations sociodémographiques (notamment âge, sexe et lieu de résidence du patient). On dispose également pour chaque séjour de son classement dans un GHM (groupe homogène de malades) qui est également une variable sensible médicale. La classification des GHM est une classification médico-économique hiérarchique dont le niveau le plus agrégé comporte 28 CMD (catégories majeures de diagnostic). Sur ce millésime 2012, outre la CMD des séances, nous avons exclu la CMD regroupant les séjours inclassables (informations manquantes, incomplètes ou erronées), ne conservant donc que 26 CMD. Considérant que la CMD constituait une bonne synthèse médicale du séjour, nous n'avons pas analysé le détail des diagnostics et des actes.

Les variables analysées comprennent donc finalement la CMD, le numéro Finess d'identification de l'établissement (permettant d'identifier le lieu d'hospitalisation), le code géographique de résidence du patient, son âge, son sexe, la durée de son hospitalisation, et ses modes d'entrée et de sortie de l'hôpital ou de la clinique.

Les logiciels utilisés pour le test

Les tests menés par le groupe de travail ont souffert d'un calendrier restreint. Pour cette raison, la création de jeux de données anonymes à partir du PMSI a été testée avec 2 logiciels : Mu-argus [5] et ARX [6], gratuits et immédiatement disponibles (Encadré 1).

Mu-argus est un logiciel développé par les statisticiens publics des Pays-Bas (institut CBS), initialement dans le cadre du projet européen CASC (Computational Aspects of Statistical Confidentiality) entre 2000 et 2003. Depuis, de nouvelles versions ont vu le jour grâce à plusieurs projets européens menés sous l'égide d'Eurostat[6] et une version open-source du logiciel devrait bientôt être disponible. Plusieurs techniques d'anonymisation sont implémentées dans Mu-argus. La marche à suivre peut être résumée en 3 étapes. En premier lieu, il faut importer les données et définir les métadonnées : nom, position et nature des variables (variables quasi-identifiantes ou variables sensibles dont on cherche à éviter la divulgation). Différentes méthodes d'anonymisation, perturbatrices ou non, peuvent ensuite être mises en œuvre (cf. dossier 3 pour une présentation détaillée de ces méthodes). On peut maitriser la k-anonymisation du fichier, mais pas les autres critères de protection comme la l-diversité ou la t-proximité (cf. article précédent de ce *Dossier* pour une définition de ces critères). Enfin, une fois que le niveau de protection ajouté à la base initiale est jugé suffisant, le logiciel permet d'exporter le jeu de données créé. L'interface de Mu-argus est relativement intuitive et le logiciel est utilisé par plusieurs Instituts de Statistique publique européens, souvent en combinaison avec d'autres outils.

ARX est un logiciel libre implémenté en Java. Il permet de protéger contre le risque de ré-identification dans les jeux de données en fonction de différents critères (par exemple k-anonymat, l-diversité, t-proximité) que l'utilisateur peut combiner. Les techniques de protection s'appliquent ensuite selon les critères de confidentialité retenus et l'ensemble des jeux de données vérifiant ces critères sont générés. Des mesures de la perte d'information engendrée sont également disponibles. Comme dans Mu-argus, la démarche d'anonymisation peut être résumée en plusieurs étapes. D'abord il faut importer les données au format .csv et la définition des variables (les variables sensibles, les variables quasi-identifiantes) puis des critères de confidentialité (par exemple la valeur de k pour le k-anonymat). Avant de mettre en œuvre les techniques d'anonymisation, il faut préciser ou créer, pour chaque quasi-identifiant, les différents niveaux de nomenclature, i.e. les différents niveaux de regroupement et d'agrégation des données, ainsi que le nombre maximal d'enregistrements qu'on

[5] Terme utilisé dans la littérature traitant des risques de ré-identification, une variable quasi-identifiante est un attribut qui, pris individuellement, ne permet pas d'identifier un individu de manière certaine, mais la ré-identification peut devenir possible lorsqu'on s'intéresse à la combinaison de plusieurs quasi-identifiants. L'âge, le sexe, la localisation géographique sont des variables quasi-identifiantes.

[6] http://neon.vb.cbs.nl/casc/..%5Ccasc%5CESSNetTask1.htm.

s'autorise à supprimer de la base initiale. Après la mise en œuvre des techniques d'anonymisation, les différents jeux de données en sortie peuvent être visualisés et les jeux retenus exportés.

ENCADRÉ 1 - DES LOGICIELS D'ANONYMISATION DE BASES DE DONNÉES

Il existe de nombreux logiciels permettant d'anonymiser et de garantir la confidentialité des données mais peu peuvent être utilisés pour l'objet qui nous intéresse ici, à savoir le retraitement des variables quasi-identifiantes sans modification des autres variables du fichier initial. D'autres outils que ceux utilisés pour le test sont moins employés car développés par des chercheurs ou des instituts publics pour un jeu de données ou un projet qui leur est propre. Actuellement, il existe peu de travaux menés ou publiés sur l'évaluation des outils d'anonymisation (généralement, les chercheurs comparent plutôt les algorithmes qu'ils ont développés à d'autres qu'ils ont sélectionnés dans l'état de l'art du domaine).

µ-Argus

- Développé par l'Institut National de Statistique des Pays-Bas dans le cadre de projets européens associant d'autres instituts de statistique et des universitaires.
- Téléchargeable à l'adresse : http://neon.vb.cbs.nl/casc/mu.htm.
- Une version open source sera bientôt disponible [7].
- Il met en œuvre les méthodes de recodage des modalités, de suppression locale, de perturbation aléatoire des données, de bruitage des données, de micro-agrégation, de « swap » (permutation de valeurs) et d'arrondi.
- Il est utilisé par une partie des instituts de statistique publique en Europe [8], les autres ayant recours aux outils qu'ils ont développés pour leurs propres besoins.
- Les poids de sondage dans le cas d'un échantillon peuvent être pris en compte pour l'évaluation du risque de divulgation.

sdcMicro

- Développé par l'Université de Vienne
- Package du logiciel R, téléchargeable à l'adresse : http://cran.r-project.org/web/packages/sdcMicro
- Il met en œuvre les méthodes de micro-agrégation, de bruitage, de « swap » (permutation de valeurs), de suppression locale, et de génération de données synthétiques.

ARX

- Développé par l'université de Munich.
- Téléchargeable à l'adresse : http://arx.deidentifier.org.
- Permet d'appliquer différents critères de protection comme le k-anonymat, la l-diversité ou la t-proximité.
- Crée des jeux de données, fournit une liste exhaustive des solutions possibles pour une agrégation des modalités des variables quasi-identifiantes donnée. Fournit une mesure de la perte d'information associée à chaque solution.

CAT

- Développé par l'Université de Cornell, USA.
- Téléchargeable à l'adresse http://sourceforge.net/projects/anony-toolkit/.
- Permet d'appliquer des critères de protection tels que le k-anonymat ou la l-diversité.

UTD

- Développé par l'Université du Texas
- Permet d'appliquer les critères de protection tels que le k-anonymat, la l-diversité, la t-proximité
- La dernière version, implémentée en Java, date de 2012.

L'ensemble de ces logiciels ont par ailleurs l'avantage d'être gratuit. Un autre logiciel (PARAT), cette fois-ci payant, développé par la société canadienne (Ottawa) *PrivacyAnalyticsInc pour le domaine de la santé* traite également l'anonymisation des données.

Première approche du risque de ré-identification dans la base initiale

Évidemment, les séjours hospitaliers en MCO ne se répartissent pas, pour une pathologie donnée, de manière uniforme parmi les tranches d'âges quinquennales. Les variables sociodémographiques (âge et sexe) mais également celles relatives au séjour (le mode d'entrée et de sortie et la durée du séjour) sont des variables très discriminantes.

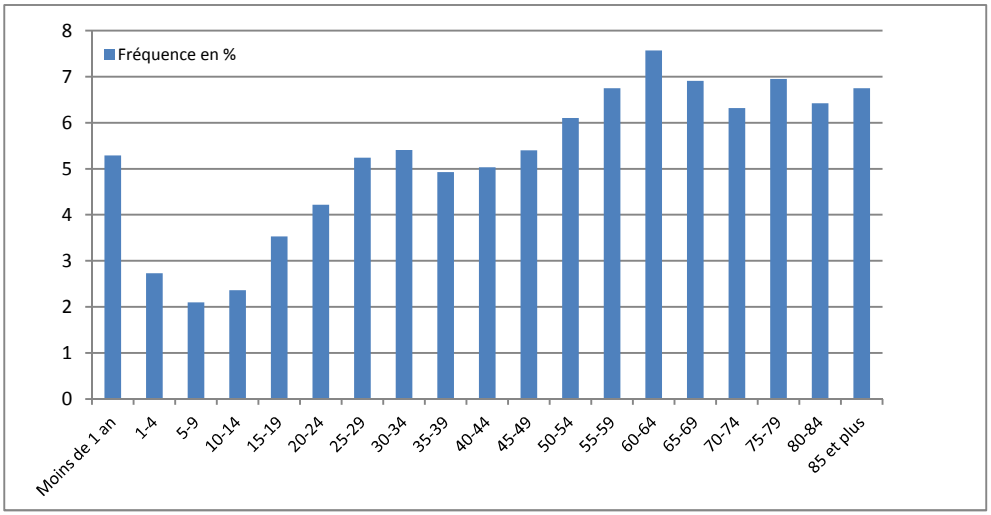
En effet, concernant l'âge, les séjours hospitaliers sont plus fréquents aux âges élevés ainsi que chez les nouveau-nés alors que les patients âgés de 1 à 14 ans ne représentent que 7 % des séjours (Graphique 1). La plupart des séjours en MCO sont évidemment de courte durée : 45 % sont des hospitalisations de jour (0 nuit), et ceux supérieurs à une semaine

sont peu fréquents (Graphique 2). Par ailleurs, dans la plupart des cas, les patients arrivent à l'hôpital depuis leur domicile (83 %), et quand ils sont hospitalisés (qu'importe leur provenance) les patients quittent l'hôpital pour rejoindre leur domicile (77 % de l'ensemble des séjours, Graphique 3). Les autres modes d'entrée et de sortie sont rares, et le décès est ainsi particulièrement identifiant. Enfin, la distance entre le lieu de résidence du patient et celui de son hospitalisation est également une dimension très discriminante : les patients qui sont hospitalisés dans un département éloigné ou dans une autre région que celle où ils habitent sont assez atypiques et donc facilement ré-identifiables. Par exemple dans l'Ain, 43 % des séjours sont effectués par des patients résidant dans ce même département, et 86 combinaisons « *Ain comme département d'hospitalisation ×département de résidence autre que l'Ain* » comportent moins de 10 séjours.

Ces premières statistiques descriptives permettent d'identifier *a priori* les modalités rares, celles qui pourraient être à l'origine des risques de ré-identification, et donc d'établir une première stratégie de regroupement de modalités pour diminuer ce risque. Néanmoins, le recours à ces statistiques pour généraliser les variables quasi-identifiantes présente de fortes limites, au sens où les méthodes se basant sur le critère du k-anonymat (Cf. article précédent du *Dossier*), pour se prémunir du risque de ré-identification, analysent les croisements de toutes les variables quasi-identifiantes. Ainsi, si le mode de sortie « décès » apparaît comme étant une modalité rare, elle sera a priori plus rare chez les moins de 25 ans que chez les personnes âgées de plus de 75 ans.

GRAPHIQUE 1

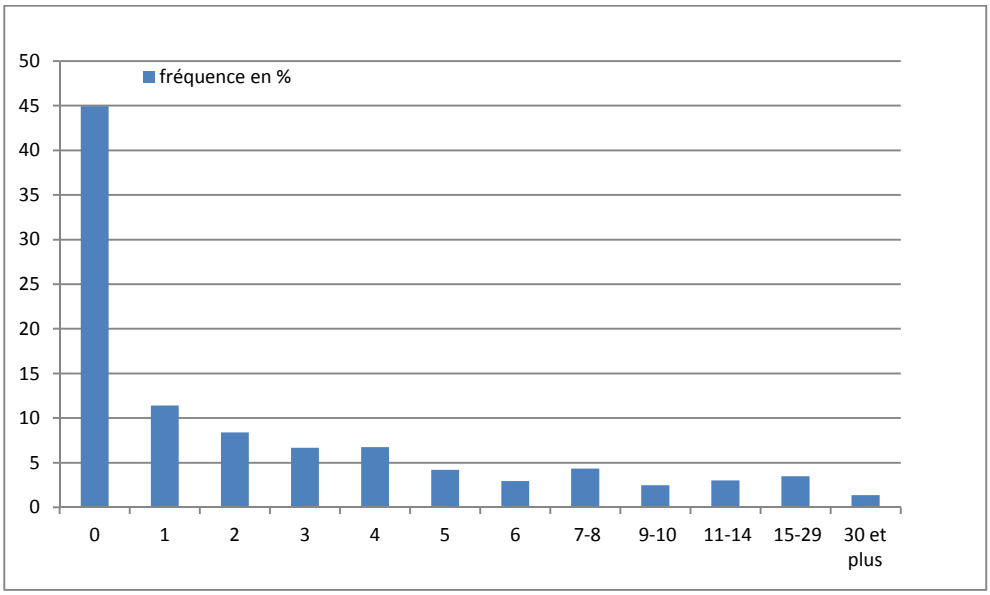
Répartition des séjours par tranche d'âge (en années)



SOURCE : PMSI-MCO 2012.

GRAPHIQUE 2

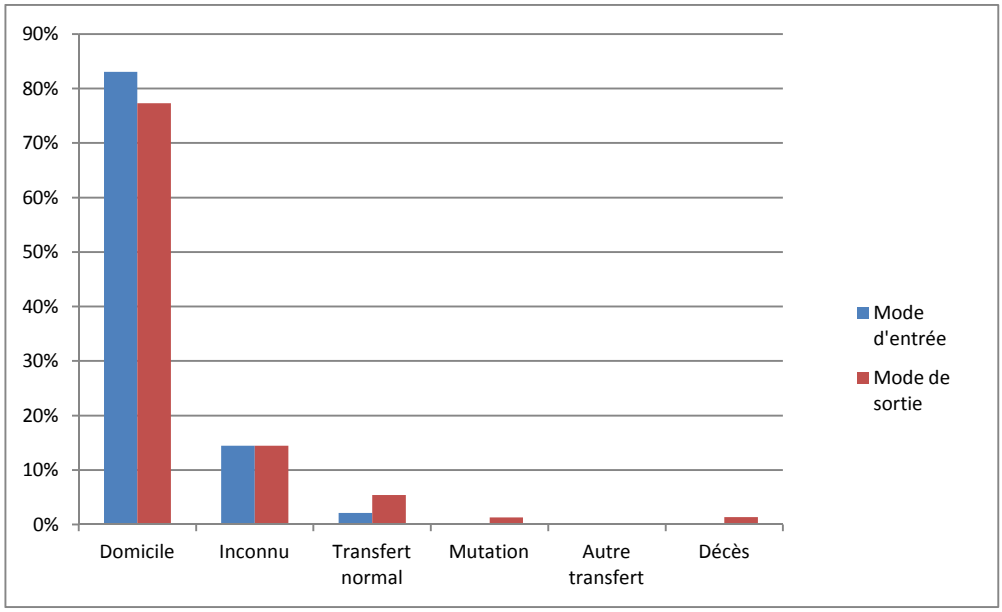
Répartition des séjours selon le nombre de nuits passées à l'hôpital



SOURCE : PMSI-MCO 2012.

GRAPHIQUE 3

Répartition des modes d'entrée et de sortie à l'hôpital



SOURCE : PMSI-MCO 2012.

NOTE :

Mode d'entrée « domicile » : le malade vient de son domicile ; **Mode de sortie « domicile »** : le malade retourne à son domicile ; **Mode d'entrée « transfert normal » :** pour une hospitalisation à part entière (c'est-à-dire hors entrée pour une prestation demandée par un autre établissement) le malade vient d'une autre entité juridique ou d'un établissement différent appartenant à la même entité juridique selon le caractère privé ou public de l'établissement ; **Mode de sortie « transfert normal »** : le malade sort pour une hospitalisation à part entière (hors sortie après prestation réalisée pour le compte d'un autre établissement) dans une autre entité juridique ou d'un établissement différent appartenant à la même entité juridique selon le caractère privé ou public de l'établissement ; **Mode de sortie par « mutation »** : le malade sort vers une unité médicale (autre que le court séjour) du même établissement ; **Mode de sortie « décès »** : le malade est décédé dans l'unité médicale ; **Mode de sortie « autre transfert »** : un établissement « prestataire » reçoit un patient pour une durée de moins de 48 heures dans le seul but de réaliser un acte que l'établissement d'origine ne peut réaliser lui-même.

Méthodologie et techniques d'anonymisation retenues pour le test

L'objectif du test sur la base exhaustive PMSI-MCO 2012 (à l'exclusion des séances) est d'évaluer puis de réduire le risque de ré-identification pour construire des fichiers qui respectent les critères d'anonymisation définis. Les discussions sur l'état de l'art des méthodes de protection ont conduit à retenir comme critères de protection le *k*-anonymat et la *l*-diversité ([9] et article précédent de ce *Dossier* pour une présentation formelle de ces méthodes et paragraphes suivants pour la mise en œuvre).

Les méthodes d'anonymisation dites perturbatrices (qui altèrent les données, comme le bruitage ou la permutation aléatoire de valeurs) n'ont pas été retenues. Elles complexifient l'usage du fichier pour des analyses et accroissent les risques d'utilisation erronée ou de mauvaise interprétation auprès des utilisateurs non experts. Dans une démarche d'*open data*, les utilisateurs peuvent être très divers, et les échanges entre producteurs et utilisateurs (de profils différents) des données difficiles. Les suppressions locales (remplacement par des valeurs manquantes des modalités de certaines variables quasi-identifiantes pour des individus ne remplissant pas les critères d'anonymisation - voir aussi article précédent de ce *Dossier*) ont également été écartées par le groupe de travail pour le test. En effet, les suppressions de valeurs risquent de conduire l'utilisateur à éliminer des observations de ses analyses, ce qui peut fausser les conclusions (par exemple pour l'analyse de la prévalence d'une maladie selon l'âge ou le département).

Dans le cadre de ce test, nous avons donc regroupé les modalités des quasi-identifiants.

Le groupe a discuté le choix des variables quasi-identifiantes et des variables sensibles (pour lesquelles on ne veut pas qu'il y ait divulgation, Tableau 1) ; point crucial dans la démarche de protection du fichier.

Les valeurs de *k* et de *l* ont ensuite été longuement discutées dans le groupe de travail : k=10 et l=3 ont finalement été retenues[7] pour le test. Puis, grâce à des regroupements pertinents de modalités, des fichiers 10-anonymisés et 3-diversifiés ont pu être construits.

Le travail préliminaire sur les nomenclatures ainsi que les éléments descriptifs sur la base de données ont permis de faciliter les traitements des logiciels visant à définir les regroupements des modalités des variables quasi-identifiantes à privilégier. Des regroupements conformes aux pratiques usuelles dans le domaine de l'épidémiologie (par exemple des tranches d'âges quinquennales pour les individus âgés de plus d'un an de manière à isoler les nourrissons) ont été pratiqués. Cette approche permet par la suite de maximiser l'utilité des jeux de données construits ou d'arbitrer entre les différents jeux de données qui satisfont les critères d'anonymisation.

La k-anonymisation

L'âge, le sexe, le code postal du lieu de résidence ou d'hospitalisation sont des variables, qui peuvent être facilement connues (on parle de variables avec un niveau élevé de notoriété) ou retrouvées par des tiers non autorisés (employeurs, assureurs, journalistes, conjoints, voisins par exemple). Pour le test, ont donc été retenues comme quasi-identifiantes les variables indiquées comme telles figurant dans le Tableau 1.

Le croisement des modalités de chacune de ces variables forme la clé d'identification de l'individu. Soit, pour un homme de 50 ans résidant dans l'Ain à Bourg-en-Bresse arrivé au centre hospitalier de Bourg-en-Bresse par un « transfert normal » et rentré à son domicile après 11 jours d'hospitalisation :

Clé : {Sexe=1 ; Âge=50 ; Lieu de résidence= 01053 ; Finess= 01 000 962 9, Mode d'entrée=7 ; Mode de sortie=8 ; Durée=11}

Un fichier est dit k-anonymisé si et seulement si chaque clé d'identification est partagée par au moins k séjours. En d'autres termes, une personne qui connaît l'ensemble des modalités prises par un individu pour tous les quasi-identifiants,

[7] En pratique, le seuil de k=5 est souvent utilisé [10]. Néanmoins, par précaution, la valeur de k=10 a été retenue pour le test [11].

et donc sa clé d'identification, trouvera dans la base au moins k individus partageant cette clé et ne pourra pas ré-identifier l'individu concerné.

Dans ce test, nous cherchons à construire un fichier 10-anonymisé, où chaque individu est donc indistinguable (au vu de sa clé d'identification) d'au moins 9 autres individus possédant les mêmes caractéristiques quasi-identifiantes.

TABLEAU 1

Description des variables utilisées pour le test

Nom de la variable	Nature de la variable
Sexe	**Quasi-identifiante**
Âge	**Quasi-identifiante**
Le code géographique du lieu de résidence	**Quasi-identifiante**
Le lieu d'hospitalisation (numéro Finess d'identification de l'établissement)	**Quasi-identifiante**
Durée d'hospitalisation	**Quasi-identifiante**
Mode d'entrée (suivant les tests)	**Quasi-identifiante**
Mode de sortie (suivants les tests)	**Quasi-identifiante**
CMD, catégorie majeure de diagnostic	**Donnée sensible**

SOURCE : PMSI-MCO 2012.

La l-diversité

L'objectif du test est la protection des données dites « sensibles » ; dans le test effectué, il s'agit de la pathologie pour laquelle le patient a été hospitalisé. Celle-ci est présente dans le PMSI à travers plusieurs variables de diagnostics (principaux, associés, reliés) et synthétisée pour chaque séjour par la variable GHM ou Groupe Homogène de Malade (cf. point 1.1). Les GHM forment une nomenclature hiérarchique dont le niveau le plus agrégé est la Catégorie Majeure de Diagnostic (CMD) à 26 modalités[8], qui a été retenue provisoirement et pour les besoins du test comme l'information sensible à protéger[9].

Si les patients (au moins 10) partageant la même clé d'identification ont tous la même CMD, alors, même sans retrouver de façon certaine l'individu dans la base, un attaquant peut en déduire sa CMD (spécialité ou discipline) dans laquelle il a été pris en charge). L'attribut sensible est divulgué du fait de l'homogénéité du groupe d'individus partageant la même clé d'identification[10].

Un fichier est dit l-diversifié si et seulement si, pour chaque clé d'identification, chaque variable sensible est suffisamment diversifiée, avec au moins l modalités différentes.

Dans ce test, l'unique variable sensible sur laquelle la l-diversité est mesurée est la CMD. Nous cherchons à construire un fichier 3-diversifié.

Si le fichier de départ ne vérifie pas ces deux critères (10-anonymat et 3-diversité), il faut alors limiter le nombre de croisements (i.e. le nombre de clés d'identification) possibles en appauvrissant le niveau de détail des quasi-identifiants.

[8] En excluant les séances et les séjours au GHM inconnu, la CMD à 26 modalités correspond le plus souvent à un système fonctionnel, une région anatomique (affections du système nerveux, de l'œil, de l'appareil respiratoire…).

[9] L'appartenance des séjours à des CMD différentes n'est pas en réalité un bon indicateur de la diversité des maladies. Il est évident qu'un mélanome et un abcès cutané (diagnostics d'entrée dans la même CMD pourtant) sont des maladies très différentes. Inversement, deux séjours hospitaliers où la même maladie aurait été signalée dans les codes de diagnostic, peuvent être classés dans deux CMD différentes si pour l'un au moins de ces séjours, la maladie en question n'était pas le diagnostic principal (celui qui a motivé l'admission du patient dans l'unité médicale).

[10] Par exemple si on apprend (à partir des quasi-identifiants que l'on connaît) que l'individu que l'on recherche appartient à une classe d'au moins 10 individus dont les séjours relèvent tous de la CMD25 (SIDA et infections au VIH) alors on apprendra de fait qu'il est porteur du virus.

Cela revient à regrouper certaines modalités de manière pertinente, en minimisant la perte d'information, c'est à dire en travaillant sur les modalités qui sont les plus identifiantes.

Résultats ARX

La mise en œuvre dans ARX suppose de définir *a priori* les différents niveaux de regroupements de modalités (les différents niveaux de nomenclature) possibles pour chaque variable quasi-identifiante. Le logiciel détermine ensuite l'ensemble des solutions (avec différents niveaux de détails pour chacune des nomenclatures) qui permettent de construire un fichier k-anonymisé et l-diversifié. Les résultats présentés ici n'intègrent pas les modes d'entrée et de sortie en l'absence de nomenclature *a priori*. Une fois les données au format .csv chargées, les variables quasi-identifiantes (sexe, âge, lieu de résidence, numéro Finess, durée d'hospitalisation) et la variable sensible (la CMD, catégorie majeure de diagnostic) ont été définies. Le fichier ainsi défini est ensuite 10-anonymisé et 3-diversifié. Pour chaque quasi-identifiant, les nomenclatures issues des différents regroupements sont définies *a priori* (tableau 2). Ces chemins d'agrégation doivent être hiérarchisés et monotones, i.e. chaque groupe au niveau agrégé doit s'obtenir en regroupant des groupes du niveau inférieur plus détaillé. ARX applique ensuite l'ensemble des combinaisons de regroupements possibles et vérifie les critères, ici 10 séjours et 3 CMD différentes par clé d'identification. L'ensemble des solutions possibles est ensuite renvoyé à l'utilisateur.

Certains niveaux d'anonymisation ne peuvent pas être atteints sans suppression de données. Dans ce cas, ARX permet de définir le pourcentage de séjours atypiques (pour une clé d'identification donnée, on comptabilise moins de 10 séjours et/ou le critère de diversité n'est pas vérifié) qui pourraient être supprimés pour atteindre le niveau d'anonymisation souhaité. Toutefois, la suppression de données atypiques peut créer un biais d'analyse. C'est la raison pour laquelle la solution qui a été retenue consiste à produire un fichier contenant les séjours atypiques, mais avec un niveau de détail des variables quasi-identifiantes différent (plus faible) que celui des autres séjours.

En interdisant toute suppression de séjours dans la base de données obtenue, ARX produit 127 solutions (sur les 1 000 différents chemins d'agrégation[11]) permettant de construire des fichiers 10-anonymisés et 3-diversifiés. Mais aucun d'entre eux ne contient l'ensemble des quasi-identifiants, en d'autres termes dans chacun des fichiers un quasi-identifiant a dû être supprimé (cf. variable positionnée à « non renseigné » dans la nomenclature du tableau 2).

[11] 5 niveaux pour le numéro Finess x 5 niveaux pour l'âge x 2 niveaux pour le sexe x 4 niveaux pour la durée de séjour x 5 niveaux pour le lieu de résidence du patients, cf. Tableau 1.

TABLEAU 2

Les différents niveaux de nomenclatures des quasi-identifiants retenus pour le test

Quasi-identifiant	Niveau 0	Niveau 1	Niveau 2	Niveau 3	Niveau 4
Numéro Finess de l'établissement	En clair	Département	Région	Région en regroupant les DOM, et la Corse avec la région Provence-Alpes- Côte d'Azur	Non renseigné
Âge	En clair	Moins de 1 an 1-4 ans 5-9ans 10-14 ans … 75-79 ans 80-84 ans 85 ans ou plus	Identique au Niveau 1 avec le regroupement de 5-9 ans et 10-14 ans en 5-14 ans	0-39 ans 40-69 ans 70 ans ou plus	Non renseigné
Sexe	En clair	Non renseigné	Non renseigné	Non renseigné	Non renseigné
Durée d'hospitalisation	En clair	En jours jusqu'à 6 jours 7-8 jours 9-10jours 11-14 jours 15-29 jours 30 jours ou plus	0 jours 1-2 jours 3-4 jours 5-6 jours 7-10 jours 11 jours ou plus	Non renseigné	Non renseigné
Code géographique de résidence	En clair	Département	Région	Région en regroupant les DOM, et la Corse avec PACA	Non renseigné

SOURCE : PMSI-MCO 2012, EN EXCLUANT LES CMD SÉANCES ET ERREUR, SOIT 17,6 MILLIONS DE SÉJOURS.

Pour isoler (limiter) l'effet des observations avec un risque de ré-identification, il est intéressant de voir les fichiers construits en raisonnant à part pour ces séjours atypiques. Pour cela, il faut dans un premier temps définir un seuil maximal d'enregistrements à supprimer de la base initiale, choisir une solution satisfaisant les critères d'anonymisation sur les séjours non supprimés. Ensuite, on travaille uniquement sur les séjours atypiques, et on réitère le processus (en agrégeant davantage l'information portée par les variables quasi-identifiantes pour ces derniers) afin que le fichier global soit 10-anonymisé et 3-diversifié. On obtient ainsi un fichier avec deux niveaux de détail selon que le séjour fait ou non partie des cas atypiques.

Le Tableau 3 présente un exemple de solution obtenue avec cette méthode (le seuil maximal d'enregistrements à supprimer est de 4 %) : pour les séjours atypiques, le fichier ne contient ni la durée de séjour ni le lieu d'hospitalisation. En revanche, pour les autres séjours, l'ensemble des quasi-identifiants est conservé et deux dimensions géographiques sont incluses. On note que le nombre de clés d'identification différentes dans le cas des séjours atypiques est de 3 762, celui

pour les autres séjours est de 1 038 312. On voit clairement sur cet exemple qu'un nombre restreint de séjours atypiques (correspondant à 3,8 % de l'ensemble des séjours) a un impact considérable sur le niveau d'anonymisation des séjours.

TABLEAU 3

Description d'un fichier 10-anonymisé, avec deux niveaux de détail selon la rareté du séjour

Nom de la variable	Nature de la variable	Niveau de nomenclature			
		Séjours atypiques (3,8 %)		*Autres séjours*	
		Niveau	*Nombre de Modalités*	*Niveau*	*Nombre de Modalités*
Sexe	Quasi-identifiant	Niveau 0	2	Niveau 0	2
Âge	Quasi-identifiant	Niveau 1	19	Niveau 1	19
Lieu de résidence	Quasi-identifiant	Niveau 1	99	Niveau 1	99
Numéro Finess	Quasi-identifiant	Niveau 4	valeur unique	Niveau 2	23 (22 régions+ les DOM regroupés)
Durée d'hospitalisation	Quasi-identifiant	Niveau 4	Valeur unique	Niveau 1	12
Nombre de clés d'indentification		3 762		1 038 312	
CMD, catégorie majeure de diagnostic	Donnée sensible	En clair	26	En clair	26

SOURCE : PMSI-MCO 2012 EN EXCLUANT LES CMD SÉANCES ET ERREUR, SOIT 17,6 MILLIONS DE SÉJOURS.

650 388 séjours atypiques ont été agrégés à un niveau plus fin, soit 3,8 % de l'ensemble des séjours.

Résultats Mu-Argus

La mise en œuvre concrète

La mise en œuvre de l'évaluation du risque de ré-identification puis de l'anonymisation des données avec le logiciel Mu-Argus est itérative. Une fois les données chargées dans le logiciel et les métadonnées définies, Mu-argus calcule la distribution du nombre d'occurrences de chaque clé d'identification. Pour permettre ce calcul et afin de limiter le nombre de croisements, certaines variables quasi-identifiantes ont été discrétisées *a priori*. En particulier, l'âge a été recodé en 19 tranches et la durée d'hospitalisation en 12 modalités (nombre de jours détaillé jusqu'à 6 jours, puis 7 à 8 jours, 9 à 10 jours, 11 à 14 jours, 15 à 29 jours et 30 jours ou plus) et le lieu de résidence du patient mis au niveau régional (Tableau 4). En première approche, nous avons considéré une seule dimension géographique afin de limiter l'appauvrissement du fichier dans les autres variables[12]. À noter que les valeurs et modalités manquantes ne sont pas ici considérées comme ré-identifiantes[13] et n'entrent donc pas dans le dénombrement des occurrences des clés d'identification.

µ-Argus permet d'identifier les modalités des variables quasi-identifiantes pour lesquelles il y a un fort pouvoir de ré-identification, en indiquant celles impliquées dans un nombre important de clés d'identification « à problème » (pour lesquelles il y a entre 1 et 9 séjours représentés). Ainsi, il apparaît que la diffusion des données pour la Corse peut faciliter grandement la ré-identification : il a donc été décidé de diffuser au niveau « Corse ou Provence-Alpes-Côte d'Azur ».

Nous appliquons une méthode itérative. Les modalités qui sont à l'origine d'un important risque de ré-identification sont détectées. Ces modalités sont regroupées avec une autre en respectant les usages habituels des experts du domaine de la santé, réduisant ainsi le niveau de détail contenu de la variable. Si aucun regroupement n'est possible ou pertinent (regrouper 2 départements très éloignés par exemple n'aurait aucun sens), il faut revenir à l'étape de détection des modalités à l'origine de risque de ré-identification. Lorsque les regroupements successifs ont permis d'obtenir un fichier 10-anonymisé (où chaque clé d'identification est partagée par au moins 10 séjours), la l-diversité est vérifiée à l'aide d'un

[12] En effet, les statistiques descriptives ont montré que le croisement de ces 2 variables quasi-identifiantes que sont le département de résidence et d'hospitalisation générait des modalités rares dès qu'un patient résidait dans un département très éloigné de son lieu d'hospitalisation. Ainsi, avoir un effectif suffisant supposerait d'agréger fortement l'information sur ces dimensions et celle des autres quasi-identifiants.

[13] C'est un paramétrage que permet Mu-argus. En effet, quand une variable quasi-identifiante est manquante, c'est une modalité dont ne peut pas se servir un attaquant potentiel pour essayer de ré-identifier un individu.

programme SAS. Il s'agit de s'assurer que pour chaque clé d'identification, le groupe de séjours associé présente au moins 3 CMD différentes.

TABLEAU 4

Description d'un fichier 10-anonymisé par le producteur des données

Nom de la variable	Nature de la variable Sensible/Quasi-identifiante	Nombre de modalités
Sexe	Quasi-identifiante	2
Âge	Quasi-identifiante	19
Durée du séjour	Quasi-identifiante	12
Mode d'entrée	Quasi-identifiante	4
Mode de sortie	Quasi-identifiante	5
Lieu de résidence du patient	Quasi-identifiante	23 (22 régions + les DOM regroupés)
CMD (Catégorie Majeure de Diagnostic)	Sensible	26

Source : PMSI-MCO 2012, en excluant les CMD séances (soit 20,6 millions de séjours)

Deux exemples de fichiers 10-anonymisés à partir de la base PMSI-MCO

Les Tableaux 5 et 6 présentent les 2 fichiers obtenus avec Mu-argus qui vérifient les critères présentés plus haut (10-anonymisation et 3-diversité).

En première approche, sans dimension géographique, Mu-argus vérifie le respect du k-anonymat pour les 9 120 clés d'identifications[14]. Sur ces clés d'identification, 1 132 (12 %) comportent moins de 10 séjours. La démarche itérative décrite plus haut conduit à regrouper d'une part les tranches d'âges jeunes (les 5-9 ans et les 10-14 ans) et d'autre part les modes d'entrée et de sortie autres que « domicile » (cf. Tableau 5). Ces regroupements concernent logiquement les modalités rares déjà entrevues par les statistiques descriptives sur la base initiale (cf. paragraphe 1.3). Par ailleurs, la tranche d'âge des 5-14 ans correspond à la pédiatrie, ce qui reste pertinent d'un point de vue de l'analyse en santé. Chacune des 1 728 clés d'identification comporte bien 3 CMD différentes, le fichier est donc 3-diversifié.

[14] 2 modalités pour le sexe x 19 tranches d'âge x 12 durées de séjour x 4 modes d'entrée x 4 modes de sortie.

TABLEAU 5

Premier exemple de fichier 10-anonymisé avec Mu-Argus, (sans dimension géographique)

Nom de la variable	Nature de la variable	Nombre de modalités dans le fichier en entrée de Mu-Argus	Nombre de modalités dans le fichier 10-anonymisé
Sexe	Quasi-identifiant	2	2
Âge	Quasi-identifiant	19 : les moins de 1an puis tranches quinquennales jusqu'à 85 ans ou plus	18 : regroupement des 5-9ans et des 10-14ans
Durée du séjour	Quasi-identifiant	12	12
Mode d'entrée	Quasi-identifiant	4	2 : Domicile ou Autre
Mode de sortie	Quasi-identifiant	5	2 : Domicile ou Autre, et les DOM sont regroupés ensemble
Nombre de clés d'identification		9 120	1 728
CMD (Catégorie majeur de diagnostic)	Donnée sensible	26	26

Source : PMSI-MCO 2012 en excluant les CMD séances soit 20,6 millions de séjours.

Note : les modalités ayant fait l'objet de regroupements sont en rose.

Un deuxième fichier incluant une dimension géographique, la région de résidence du patient, a également été construit. Sur les 39 744 clés d'identification possible, 9 553 (24 %) ne vérifient pas le critère de k-anonymisation. La même démarche itérative, avec comme contrainte de conserver un haut degré de granularité sur la région de résidence sous peine de rendre cette variable inutilisable dans les études, a conduit à diminuer nettement le détail d'information sur l'âge, notamment en agrégeant les tranches jeunes ainsi que sur la durée d'hospitalisation (cf. tableau 6). La région de résidence du patient est ainsi laissée en clair, sauf pour la Corse regroupée avec la région Provence-Alpes-Côte d'Azur (PACA). Les 2 112 combinaisons finales vérifient la l-diversité, à savoir 3 CMD différentes parmi les patients présentant la même clé d'identification[15].

TABLEAU 6

Second exemple de fichier 10-anonymisé avec Mu-argus (avec une dimension géographique)

Nom de la variable	Nature de la variable	Nombre de modalités dans le fichier initial	Nombre de modalités dans le fichier 10-anonymisé
Sexe	Quasi-identifiant	2	2
Âge	Quasi-identifiant	18	6 : Moins de 1 an, 1-29 ans, 30-49 ans 50-59ans, 60-69 ans et 70 ans ou plus
Durée du séjour	Quasi-identifiant	12	2 : +ou- d'une semaine
Mode d'entrée	Quasi-identifiant	2	2
Mode de sortie	Quasi-identifiant	2	2

[15] Trois clés d'identification sur les 2112 (hors modalités manquantes) possibles aboutissent à seulement 2 CMD différentes, dont 2 avec 1 séjour ayant 1 CMD, tous les autres ayant la seconde. Les croisements "seulement" 2-diversifiés concernent la CMD 15 à savoir les nouveau-nés et prématurés (les autres clés concernant les séjours pour une naissance). A la limite, on peut penser que pour ces cas, la CMD est une variable plutôt identifiante que sensible : si un individu est la seule personne avec l'unique CMD (et qu'il a accès au fichier), alors il peut en déduire la CMD pour les autres individus du même âge, région, sexe et durée d'hospitalisation.

Lieu de résidence	Quasi-identifiant	23	22 : Regroupement Corse et PACA
Nombre de clés d'identification		39 744	2 112
CMD	Donnée sensible	26	26

Source : PMSI-MCO 2012.Note : les modalités ayant fait l'objet de regroupements sont en rose.

Il n'est pas possible d'obtenir un fichier 10-anonymisé semblable qui contienne en plus le lieu d'hospitalisation (même au niveau région). Certains établissements étant spécialisés (les centres de lutte contre le cancer par exemple), le critère de l-diversité ne sera pas respecté. De plus, comme déjà mentionné plus haut, les cas très atypiques où le patient est hospitalisé très loin de son lieu de résidence posent inévitablement des problèmes de ré-identification (en termes de k-anonymisation). Abaisser le critère de k-anonymat, par exemple en prenant k = 5 ne réduit pas d'autant le nombre de clés d'identification à problèmes. En outre, avec un nombre minimal d'individus par clé plus petit, la l-diversité est plus difficile à obtenir, la probabilité d'avoir 3 CMD différentes parmi 5 séjours est plus faible que parmi 10 séjours.

Conclusions des tests

Dans le cadre de ce test, deux approches différentes ont été suivies pour produire des fichiers 10-anonymisés et 3-diversifiés. Avec Mu-argus, on regroupe les modalités en conservant le même niveau de détail d'information pour tous les séjours, en agrégeant à chaque étape les modalités les plus problématiques tout en conservant la pertinence des données diffusées (bon sens et selon la priorité donnée aux variables). Cette démarche longue et exploratoire permet de ne pas agréger trop rapidement mais demande l'analyse d'un expert du domaine. Avec ARX, on définit des niveaux de regroupement *a priori* pour chaque variable et on applique différents niveaux d'agrégation selon le risque de ré-identification des séjours.

Du fait de la distribution non uniforme des modalités des variables quasi-identifiantes (cf. paragraphe « première approche du risque de ré-identification dans la base initiale »), construire des jeux de données 10-anonymisés sans recourir à des méthodes perturbatrices n'est pas aisé. La perte d'utilité engendrée par les regroupements de modalités des variables quasi-identifiantes peut être significative. Notamment, nous ne sommes pas parvenus à construire des jeux de données (selon les critères fixés pour ce test) en *open data* avec 2 dimensions géographiques dans le temps imparti. Pour conserver plus d'information « géographique », il faudrait agréger autrement les informations sur la localisation du patient pour mieux rassembler les modalités rares, par exemple en utilisant comme critère la distance au lieu d'hospitalisation, quitte à orienter les utilisations possibles du fichier. En effet, dans tous les fichiers proposés, on fournit au plus une dimension géographique à l'utilisateur. Seul le fichier détaillé dans le Tableau 3 permet de donner, pour une partie des enregistrements, des informations à la fois sur le lieu d'hospitalisation et le lieu de résidence du patient.

L'optimisation des regroupements à opérer pour maximiser l'utilité du fichier 10-anonymisé est complexe. Des algorithmes[16] existent mais ils ne sont pas implémentés dans les logiciels utilisés dans ce test. Ils ne permettent pas d'intégrer des contraintes de regroupement au cas par cas, par exemple pour éviter de regrouper des unités géographiques trop distantes. La méthode itérative et les conseils et avis des experts du domaine sont plus opérationnels ici.

Les outils testés réduisent le risque de ré-identification. Le test mené est essentiellement technique. Les logiciels d'anonymisation utilisent des techniques pour protéger les données personnelles, mais les données générées après anonymisation doivent être compatibles avec le type d'analyse qui va être entrepris sur ces données pour en préserver l'utilité. Cela nécessite des échanges et discussions avec les futurs utilisateurs pour s'assurer que les outils et les techniques d'anonymisation sélectionnés s'adaptent bien à l'usage qui en sera fait.

[16]FUNG B.C.M, WANG K., CHEN R and Yu P.S Privacy preserving data publishing : a survey of recent developments. ACM computing surveys, 42, 4, article 14, june 2010, 53 pages

Dans le cadre des tests, nous avons considéré comme critères de protection d'avoir un fichier 10-anonymisé et 3-diversifié [11]. Avant de rendre des données accessibles, il est nécessaire d'évaluer le niveau de risque du jeu de données et de déterminer s'il est acceptable. Cependant, il est difficile de savoir comment mesurer le risque et « qu'est-ce qui constitue un risque acceptable ? », car il n'y a pas à l'heure actuelle de mesure du risque de ré-identification et de niveau de risque qui fasse largement consensus dans la communauté des chercheurs ou des producteurs de données.

Le sujet du risque de ré-identification fait l'objet de nombreuses recherches. La littérature existante incite à la plus grande prudence en matière de protection et d'ouverture des données.

Interrompus un temps après la remise du rapport du Groupe de travail (rapport reproduit dans le présent *Dossier*), les tests sur la base PMSI en vue de produire des jeux de données individuelles suffisamment appauvris pour que le risque de ré-identification puisse y être raisonnablement considéré comme nul (mais suffisamment riches pour qu'on puisse en tirer des analyses utiles) ont été repris au printemps 2015 par la DREES.

Bibliographie

[1] Bras, P.L., Loth, A., 2013, Rapport sur la gouvernance et l'utilisation des données de santé.

[2] Bergeat M., Cuppens-Boulahia N., Cuppens F., Jess N., Dupont F., Oulmakhzoune S., de Peretti F., 2014, « A French Anonymization Experiment with Health Data », Privacy in Statistical Databases conference, Eivissa, septembre.

[3] D. Blum, Congrès Emois de 2011 à Nancy, http://www.canal-u.tv/video/canal_u_medecine/emois_nancy_2011_anonymat_du_patient_dans_le_pmsi_quel_leurre_est_il.6824

[4] Sweeney, L.,2000, «Simple Demographics Often Identify People Uniquely », Data Privacy working paper.

[5] Hundepool, A.et al., 2008, « μ-Argus User's Manual », disponible en ligne.

[6] http://arx.deidentifier.org/anonymization-tool/configuration/

[7] Wolf, P.P., 2013, « Open source software Argus », UNECE, Work session on statistical data confidentiality.

[8] Eurostat (2013).Results of the questionnaire on SDC tools, 5th meeting of the Expert Group on Statistical Disclosure Control, Luxembourg, octobre 2013.

[9] Bergeat M., 2014, « Données individuelles : bien les protéger pour mieux les diffuser », actes des Journées de Statistique de la SFdS, Rennes, juin.

[10] El Emam, K. & al., 2009, « A Globally Optimal k-Anonymity Method for the De-Identification of Health Data », Journal of the American Medical Informatics Association. http://jamia.bmjjournals.com/content/16/5/670.full

[11] Rapport de la commission Open Data, 2014, juillet.

L'APPARIEMENT AUX BASES DE DONNÉES MÉDICO-ADMINISTRATIVES : UN ATOUT POUR LA RECHERCHE ET LA SANTÉ PUBLIQUE

Marcel GOLDBERG[1], Marie Aline CHARLES[2], Catherine QUANTIN[345], Grégoire REY[6], Marie ZINS[1]

Les bases de données publiques administratives et médico-administratives nationales : une richesse insuffisamment exploitée

La France est un des rares pays qui dispose de bases de données médico-sociales et socioéconomiques nationales centralisées, constituées et gérées par des organismes publics, couvrant de façon quasi exhaustive et permanente l'ensemble de la population dans divers domaines liés à la santé : recours aux soins, hospitalisation, handicaps, prestations et situation professionnelle et sociale. Pratiquement toutes les bases de données nationales, dont la constitution repose sur des activités liées aux missions de l'administration et d'organismes publics, utilisent actuellement un identifiant individuel unique (le numéro d'identification au répertoire ou NIR) directement ou sous forme anonymisée. Pour les bases médico-administratives, la collecte de l'ensemble des données, incluant le cas échéant le NIR, est très encadrée, notamment par le code de la santé publique

Ces bases de données présentent évidemment des limites diverses en termes de couverture, de qualité et de validité des données, variables selon les types d'utilisation qu'on peut envisager. Ces bases de données, concernant jusqu'à 65 millions de personnes, constituent néanmoins un patrimoine immatériel considérable, vraisemblablement sans équivalent au monde à cette échelle. D'autres pays ont su depuis longtemps mettre au service de la santé publique et de la recherche leurs systèmes d'information médico-sociaux, notamment les pays scandinaves ou le Canada, en créant de véritables « *Population Data Centers* », largement ouverts à la communauté scientifique qui permettent de très nombreuses études de grande qualité dans des domaines divers (voir par exemple le centre mis en place à la *British Columbia University* [1]).

Étudier la santé ne se résume pas à analyser des données de santé proprement dites : il faut bien souvent mettre celles-ci en relation avec d'autres données concernant les déterminants de la santé et les facteurs de risque, afin de mieux comprendre ce qui explique l'état de santé des personnes et de la population. Les caractéristiques socioprofessionnelles (diplômes, revenus, situation sociale et familiale), les comportements ou le style de vie (tabac, alcool, exercice physique, alimentation…), l'environnement physique et social, les conditions de travail, les caractéristiques physiologiques, biologiques et génétiques des personnes ou de leur entourage familial ou social, sont autant d'éléments qu'il faut souvent prendre en compte dans l'étude de la santé. C'est pourquoi il faut considérer non seulement les bases de données de santé, mais aussi celles qui concernent l'ensemble des autres catégories de données, et mettre en place des dispositifs complémentaires directement auprès des personnes pour le recueil de données qu'on ne peut trouver dans les bases médico-administratives, mais qui pourront être enrichies par les données de celles-ci afin de disposer de l'ensemble des informations pertinentes pour étudier la santé.

[1] UMS 011 Inserm-UVSQ, Unité Cohortes épidémiologiques en population, Villejuif, France
[2] Unité mixte Ined-Inserm-EFS Elfe, Paris, France
[3] CHRU Dijon, Service de Biostatistique et d'Informatique Médicale (DIM), Dijon, F-21000, France
[4] Inserm, U866, Université de Bourgogne, Dijon, France
[5] Inserm, CIC 1432, Dijon, France ; Dijon University Hospital, Clinical Investigation Center, clinical epidemiology/ clinical trials unit, Dijon, France
[6] CépiDc, Inserm, Le Kremlin-Bicêtre, France

Cet article décrit les principales bases de données nationales pour la recherche et la santé publique et donne des exemples d'appariement pour montrer l'intérêt de ces données pour répondre à des questions cruciales pour améliorer la santé de la population.

Les principales bases de données nationales pour la recherche et la santé publique

Données de santé

Plusieurs bases de données nationales enregistrent des données concernant la santé (incluant le recours au système de soins).

La base de données du Centre d'épidémiologie des causes de décès de l'Inserm (CépiDc) enregistre les causes de décès, codées selon la Classification internationale des maladies.

L'ATIH (Agence technique de l'information sur l'hospitalisation) gère la base de données nationale issue du fonctionnement du PMSI (Programme de médicalisation du système d'information) qui recueille des informations administratives et médicales pour chaque séjour hospitalier.

Les différents régimes d'assurance maladie se sont associés pour centraliser leurs données de remboursement de soins au sein du Système national d'information inter-régimes de l'assurance maladie (SNIIRAM), géré par la CNAMTS. Créé en 2003, il inclut depuis maintenant plusieurs années les données du PMSI transmises par l'ATIH (Encadré 1). L'accès à ces données a été facilité avec la mise en place, en 2005, d'un échantillon au 1/97ème dédié aux institutions publiques, aux agences et aux chercheurs, l'Échantillon généraliste des bénéficiaires (EGB). Les données du SNIIRAM incluent aujourd'hui tous les régimes de l'assurance maladie et concernent aussi bien la médecine de ville que les hospitalisations. Il s'agit de données individuelles extrêmement riches : données de remboursement de soins avec codage des actes et des médicaments ; identifiants des professionnels et des établissements de santé qui ont participé aux soins du patient ; informations sur la pathologie traitée pour les patients en Affection de longue durée (ALD) et en accidents du travail et maladies professionnelles (AT-MP) ; données d'hospitalisation du PMSI.

Bien qu'elles n'informent évidemment pas sur de nombreuses données personnelles et environnementales pouvant être indispensables pour la recherche et la surveillance (comportements, expositions à des facteurs de risque de nature diverse, etc.), les bases de données du PMSI et de l'assurance maladie ont à l'évidence un intérêt majeur pour documenter les évènements de santé et les parcours de soins. Elles présentent cependant certaines limites : absence de données concernant la situation socioprofessionnelle des personnes (en dehors de la notion de CMU-C), de résultat d'examens cliniques ou paracliniques, diagnostics médicaux insuffisamment fiables [2-8], impliquant un important travail méthodologique, de contrôle et de validation pour leur utilisation, notamment dans une optique de recherche épidémiologique.

Outre les dispositifs cités, d'autres bases de données plus spécifiques, enregistrent de façon exhaustive les personnes présentant une caractéristique de santé particulière. Certaines couvrent la totalité de la population française, d'autres un territoire plus restreint (département ou région). Sans chercher à être exhaustif, on peut citer parmi les bases de données ayant un intérêt pour la santé publique : les registres de maladie (cancers, maladies rares, malformations congénitales, cardiopathies ischémiques, accidents vasculaires cérébraux, tentatives de fécondation *in vitro*, interruptions médicales de grossesse), les certificats de santé de l'enfant, les bases concernant les donneurs de sang et produits reçus par des patients, ou les patients en insuffisance rénale chronique traitée par un traitement de suppléance.

ENCADRÉ 1 - « L'ANONYMISATION DANS LE SNIIRAM »

Le SNIIRAM est habituellement présenté comme une base de données individuelles anonymes. En effet, les données enregistrées pour une personne sont identifiées dans le SNIIRAM sous un numéro chiffré de façon irréversible qui ne permet pas de retrouver la personne concernée.

Pour cela, une fonction d'anonymisation appelée FOIN (*Fonction d'Occultation des Informations Nominatives*) a été conçue et fourni dès 1996 par le Centre d'études des sécurités du système d'information (CESSI) de la CNAMTS [1], pour la mise en place du PMSI établissements privés, sur recommandation de la CNIL, qui a suggéré l'utilisation de l'algorithme développé par le DIM du CHU de Dijon, après expertise du Service central de la sécurité des systèmes d'informations (SCSSI). Cette fonction d'anonymisation, dite de « hachage » permet de remplacer l'identité des personnes par des numéros d'anonymat, ou clés de chaînage, pérennes dans le temps (tant que le secret de la fonction à sens unique est inchangé) et dans l'espace [2,3]. Le numéro anonyme est calculé à partir du NIR de l'assuré (ou ouvrant-droit), de la date de naissance et du sexe de la personne. Ce système de hachage a été étendu à l'ensemble des établissements publics soumis au PMSI en 2001 [4].

L'anonymisation du SNIIRAM a ensuite été assurée par le passage successif de deux fonctions d'anonymisation utilisant la procédure FOIN [5] : (1) une fonction d'anonymisation de premier niveau, appliquée avant transmission des données d'alimentation du SNIIRAM, et donc implantée dans l'ensemble des sites alimentant le SNIIRAM en informations (c'est-à-dire actuellement dans l'ensemble des Centres de traitement informatique, CTI du régime général de sécurité sociale, et dans l'ensemble des autres régimes) ; (2) une fonction d'anonymisation de second niveau, appliquée après réception des données d'alimentation de l'entrepôt et implantée uniquement sur le site du gestionnaire du SNIIRAM (le CENTI, Centre national de traitement informatique).

Selon D. Blum [6], il faut cependant nuancer la notion d'anonymat dans le SNIIRAM. Si l'utilisation de la procédure FOIN garantit en effet qu'il n'est pas possible d'identifier une personne sélectionnée dans le SNIIRAM sur la base des données enregistrées (diagnostic, établissement, etc.) si on ne sait rien d'autre sur elle, ce n'est pas le cas si on dispose déjà de certaines informations sur cette personne (cas par exemple d'un journaliste qui chercherait à connaître la maladie dont souffrirait une personnalité qui a été hospitalisée, ou d'un employeur vis-à-vis de ses salariés). Pour cela, on procède à un appariement « indirect » (par opposition à l'appariement direct avec un identifiant unique) en cherchant dans le SNIIRAM les personnes correspondant à certaines caractéristiques présentes dans la base de données : si elles sont suffisamment nombreuses et discriminantes, la probabilité que deux personnes partagent l'ensemble de ces caractéristiques peut être très faible, voire nulle. Ainsi, si on prend l'exemple des séjours hospitaliers du PMSI (MCO), avec pour seules informations l'hôpital, le code géographique du domicile, l'âge, le sexe, le mois de sortie et la durée du séjour, 89 % des personnes hospitalisées dans l'année 2008 (et 100 % des personnes hospitalisées deux fois cette année) sont identifiables si on connait ces informations pour chacune de ces personnes [6]. L'exemple du projet AMPHI, détaillé plus loin, montre bien qu'en croisant deux bases de données (SNIIRAM et causes de décès), qui partagent un petit nombre de données communes (sexe, mois et année de naissance, date de décès, et commune de domicile), on obtient un excellent taux d'appariement.

Le fait de ne pas disposer du NIR (en clair ou « haché » par la procédure FOIN) ne procure donc qu'une anonymisation incomplète, particulièrement pour des utilisations malveillantes.

1 - Trouessin G and Allaërt FA. FOIN : a nominative information occultation function. MIE, 1997, 3, pp. 196-200.

2- Trouessin G. Rapport « qualité diagnostique et thérapeutique en cancérologie : communication d'informations multimédia dans un réseau sécurisé multidisciplinaire. Sécurité de l'information médicale en télémédecine », étude du ministère de la recherche.

3- Quantin C et al. Méthodologie pour le chaînage de données sensibles tout en respectant l'anonymat : application au suivi des informations médicales. Courrier des statistiques, 113-114, mars-juin 2005.

4 - Circulaire DHOS-PMSI-2001 n° 106 du 22 février 2001

5 - Lenormand F. Le Système d'information de l'assurance maladie, le SNIIRAM et les échantillons de bénéficiaires. Courrier des statistiques n° 113-114, mars-juin 2005.

6 - Blum D, Trouessin G. Association française des correspondants à la protection des données à caractère personnel, 27 janvier 2012.

Situation socio-professionnelle

Dans un contexte où les inégalités sociales de santé sont devenues un enjeu majeur des politiques de santé publique et de la recherche en prévention, les bases de données de la Caisse nationale d'assurance vieillesse (CNAV) sont un élément essentiel pour disposer de données sur les principales caractéristiques socioprofessionnelles des personnes. Le rôle de cet organisme est notamment d'assurer le droit au paiement de la retraite pour toute personne ayant appartenu durant sa vie au moins une fois au régime général de sécurité sociale. Pour cela, la CNAV a mis en place un système permettant de collecter et traiter les données sociales issues de différents organismes : employeurs ayant un numéro Siret, employeurs de personnel de maison, Unedic (chômage), CNAMTS (arrêts maladie), CNAF (maternité, …), des régimes particuliers ou spéciaux (SNCF, EDF, RATP…).

La principale base pertinente pour la recherche et la santé publique est le Système national de gestion des carrières (SNGC) qui permet de retracer pour chaque individu dès la première validation d'un droit (premier salaire, etc.) et jusqu'à la liquidation de ses droits à la retraite, ses différentes périodes d'activité : périodes d'activité professionnelle ou assimilées

(chômage, maladie, maternité ou congés parentaux, …). Le SNGC contient donc l'ensemble des données inhérentes à la carrière des personnes tout au long de leur vie active.

D'autres bases de données peuvent être utilement mobilisées, comme celles de l'administration fiscale concernant les revenus, ou l'Échantillon démographique permanent de l'Insee (EDP), créé en 1967, qui correspond actuellement à un sondage au 1/25ème de la population (personnes nées un des 16 jours de l'année choisis par l'Insee). Pour chaque personne incluse, il contient des informations issues des bulletins d'état-civil depuis 1968, des recensements de 1968, 1975, 1982, 1990 et 1999 et du panel tous salariés. Un appariement avec les données socio-fiscales est en cours. Par sa taille, l'EDP permet des analyses fines qui peuvent notamment prendre en compte les effets de génération et des différenciations selon les qualifications, l'origine...

Quelques exemples d'utilisation des bases de données administratives et médico-administratives nationales pour la recherche et la surveillance

Il est clair que les bases de données administratives et médico-administratives nationales ne peuvent être une panacée qui apporterait l'information permettant de répondre à toutes les questions qui se posent, mais elles peuvent apporter une aide importante. Ceci est particulièrement vrai pour les études et recherches de grande dimension et de longue durée, comme les études de cohorte qui vont continuer de se développer dans beaucoup de domaines, notamment en épidémiologie, où l'effectif envisagé de certaines cohortes se compte désormais en dizaines, voire en centaines de milliers de sujets, une envergure comparable à ce qui existe dans plusieurs autres pays développés. Les très grandes études cas-témoins en population générale, les systèmes de surveillance épidémiologique, les études concernant le recours aux soins, peuvent bénéficier de ces bases de données. De plus, dans le domaine de la surveillance, la rapidité de la remontée des informations dans le SNIIRAM, en constante amélioration, en fait un outil spécifique pouvant contribuer à la surveillance générale de l'état de santé de la population et à des investigations spécifiques, comme le suivi d'épidémies, l'investigation d'agrégats spatio-temporels de cas de maladies diverses, la surveillance autour d'une catastrophe environnementale. Sa disponibilité devrait être un atout pour répondre aux fortes contraintes de temps inhérentes aux missions de surveillance des agences sanitaires.

Il est évidemment impossible d'imaginer toutes les utilisations possibles de bases de données aussi riches en informations et couvrant des domaines aussi différents que les grandes bases nationales médicales et socioéconomiques. Cependant, afin d'illustrer l'apport potentiel d'une large ouverture de ces bases à des utilisateurs diversifiés, on peut évoquer quelques utilisations typiques, qui ont déjà ou vont faire l'objet d'expérimentations. Il peut s'agir d'une utilisation de chaque base de données indépendamment des autres, de l'appariement de bases de données entre elles ou de l'enrichissement d'enquêtes avec recueil de données auprès des personnes.

Utilisation de chaque base de données indépendamment des autres

Depuis longtemps, les données de l'assurance maladie sont utilisées pour estimer la fréquence de divers paramètres d'intérêt concernant les consommations de soins et la santé, malgré diverses limites concernant la qualité des données. Le SNIIRAM, qui permet de combiner, pour les mêmes sujets, données de l'assurance maladie et données d'hospitalisation, et l'introduction du chaînage des données individuelles dans le PMSI ont permis d'améliorer très nettement la qualité des estimations, notamment grâce à des travaux développant des algorithmes destinés à identifier avec une bonne validité des pathologies spécifiques. On voit ainsi depuis peu des travaux présentant des estimations de la prévalence et/ou de l'incidence de certains cancers à partir des données du PMSI [9] au moyen d'algorithmes combinant diagnostics et actes techniques, ou des résultats concernant la maladie de Parkinson ou l'asthme à partir des données d'ALD et de consommations de certains médicaments [10].

Dans le domaine de la pharmacoépidémiologie, il est possible de réunir des échantillons d'effectif important de sujets correspondant à un ou plusieurs critères d'intérêt et de suivre les sujets sélectionnés de façon longitudinale. Un exemple

récent et largement médiatisé est celui de l'étude du risque de valvulopathies cardiaques chez les patients diabétiques utilisateurs de benfluorex (Médiator®). Une cohorte exhaustive des diabétiques affiliés au régime général, âgés de 40 à 69 ans et ayant présenté au moins trois remboursements d'antidiabétiques oraux et/ou d'insuline à des dates différentes, a été constituée à partir du SNIIRAM ; plus d'un million de sujets a ainsi été inclus, et des comparaisons entre exposés (consommation de benfluorex en 2006) et non-exposés (aucune consommation de benfluorex en 2006, 2007 ou 2008) ont porté sur les hospitalisations pour insuffisance mitrale ou aortique ou chirurgie de remplacement valvulaire pour insuffisance valvulaire survenus en 2006 ou 2007 recherchées dans le PMSI [11].

Au-delà de ce travail particulièrement démonstratif de l'intérêt du SNIIRAM, il est clair que de très nombreuses études de pharmacoépidémiologie et de suivi post-AMM (étude sur un médicament ayant obtenu une Autorisation de mise sur le marché) peuvent être réalisées uniquement à partir de cette base. Ceci est particulièrement vrai dans le cas de l'étude de situations peu fréquentes, comme des maladies rares, ou des traitements très spécifiques, qui peuvent nécessiter l'étude de la totalité des sujets concernés : dans de tels cas, le recours au SNIIRAM est la seule méthode possible. D'une façon plus générale, face aux difficultés opérationnelles d'un suivi détaillé des patients bénéficiaires de traitements spécifiques en termes de consommations de soins et d'événements de santé, il faut souligner qu'il existe une forte demande de l'Agence nationale de sécurité du médicament (ANSM), de la Haute autorité de santé (HAS) et du ministère chargé de la santé pour ce type d'utilisation du SNIIRAM, comme le montre le récent financement par l'ANSM de deux plateformes en pharmacoépidémiologie et évaluation des usages du médicament.

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

À titre d'exemple dans le domaine environnemental, une étude de l'Institut de veille sanitaire (InVS) sur les gastro-entérites d'origine hydrique montre que les données du SNIIRAM peuvent contribuer au repérage de secteurs vulnérables quant à la qualité de leurs ressources en eau [12]. Dans le domaine de la surveillance autour des catastrophes, les données de l'assurance maladie ont été mobilisées pour suivre l'impact de la catastrophe d'AZF [13] ; une étude est actuellement en cours sur les données du SNIIRAM afin d'évaluer l'impact de la tempête Xinthia sur la consommation de psychotropes.

Appariement de bases de données entre elles

Le SNIIRAM, à l'instar de pratiquement toutes les sources médicales, ne contient pas de données sur la situation socioprofessionnelle des personnes ; de leur côté, les bases de la CNAV ne contiennent pas de données sur la santé (en dehors de données concernant certaines prestations sociales occasionnées pour raisons de santé). Dans un contexte de fort développement des études concernant les inégalités sociales et territoriales de santé, les risques professionnels ou la pénibilité du travail, des appariements à l'échelle des individus permettant de combiner des données en provenance du SNIIRAM et de la CNAV sont indispensables et particulièrement nécessaires pour venir en appui aux politiques publiques en matière de santé et d'emploi. Ceci a notamment un intérêt particulier pour constituer un « système permanent de surveillance des inégalités de santé » comme l'ont recommandé les rapports des groupes de travail sur les inégalités de santé et sur les systèmes d'information pour la santé publique du Haut conseil de la santé publique [14,15].

Quelques projets d'appariement de bases de données ont déjà été réalisés ou sont en cours (Encadré 2). On peut citer le projet HYGIE mené par l'Institut de recherche et de documentation en économie de la santé (IRDES) dont l'objectif initial est d'explorer les conditions dans lesquelles les indemnités journalières sont versées, dans le but de caractériser les populations les plus concernées. Comme il n'existe pas en France de base de données permettant l'étude des indemnités journalières, l'IRDES a construit une base de données individuelles contenant des informations à la fois sur les arrêts de travail, les consommations de soins associés, les contextes individuel et professionnel des salariés. La création de cette

base a été possible grâce à l'appariement à partir des données issues du SNIIRAM de deux fichiers de la CNAV : le SNGC et le Système national statistiques prestataires (SNSP). L'échantillon comprend actuellement 538 870 bénéficiaires (actifs et retraités), pour lesquels on dispose ainsi de données provenant des différentes bases : caractéristiques socioprofessionnelles (sexe, date de naissance, statut d'emploi, employeur, salaire), consommations médicales (recours ambulatoires, consommations de médicaments, montant des dépenses de santé totales…), affection de longue durée, prestations AT-MP, arrêts de travail. Parmi les travaux en cours, on peut citer : (i) l'analyse des mécanismes d'arrêts de travail des salariés du privé, en lien avec la nature et les spécificités des établissements qui les emploient ; (ii) la connaissance de l'impact des maladies chroniques psychiatriques sur les parcours professionnels [16].

Par ailleurs, un premier appariement entre les causes médicales de décès et l'EDP d'une part, le panel DADS (déclarations annuelles de données sociales) d'autre part a été réalisé dans le cadre de la surveillance systématique de la mortalité par profession et par secteur d'activité en population générale (InVS, projet COSMOP [17]). Enfin, le projet EDISC (Inserm Unité 1018) a également réalisé diverses analyses sur les inégalités sociales de mortalité à partir de l'appariement de l'EDP avec la base des causes de décès du CépiDc [18].

ENCADRÉ 2 - AMPHI (2010-2013)

Le projet AMPHI (*Analyse de la Mortalité Post-Hospitalisation et recherche d'Indicateurs par établissement*) a été mené au CépiDc-Inserm dans l'objectif initial d'investiguer l'apport des causes de décès dans la construction d'indicateurs de mortalité post-hospitalière visant à refléter la qualité des soins. Il était en effet établi que les indicateurs de mortalité intra-hospitalière étaient insuffisamment informatifs, car influencés par les politiques de transfert et de durée de séjour des établissements. Il était par ailleurs vraisemblable que l'utilisation d'indicateurs prenant en compte la mortalité jusqu'à un délai fixe incluant des décès post-hospitalisation pouvait entrainer la prise en compte de décès dont la cause était sans lien avec l'hospitalisation.

Pour réaliser cette étude, les données du SNIIRAM-PMSI, pour les bénéficiaires du Régime Général (RG) de l'Assurance Maladie décédés dans l'année suivant une hospitalisation en 2008 ou 2009 ont été appariées aux causes de décès (base du CépiDc), à l'aide des variables communes aux deux bases : sexe, mois et année de naissance, date de décès, et commune de domicile. En utilisant un algorithme prenant en compte la faible qualité de la variable commune de domicile dans le SNIIRAM à cette date, le taux d'appariement obtenu était de 96,4 % [1].

Une seconde phase visait à repérer les séjours pour lesquels la cause initiale de décès pouvait être qualifiée d'indépendante de la pathologie principale traitée, afin de limiter le biais de mesure des indicateurs de mortalité post-hospitalière. Pour chaque patient décédé, le diagnostic principal (DP) de chaque séjour a été comparé à la cause initiale de décès (CI) à l'aide d'un algorithme et d'un logiciel s'appuyant sur des standards internationaux. La relation DP/CI a été analysée pour le dernier séjour de chaque patient. Même pour les décès intra-hospitaliers, les deux codes n'étaient similaires que dans 40 % des cas, soulignant la complémentarité des deux informations. Pour les décès extrahospitaliers, les diagnostics ont été classés indépendants dans 14 % des décès survenus dans le mois suivant la sortie, et dans 28 % des décès survenus entre 6 et 12 mois suivant la sortie. Cependant, la prise en compte des causes de décès en supprimant les décès indépendants changeait de façon négligeable la distribution des indicateurs de mortalité post-hospitalière par établissement [2].

Ce projet a ainsi mis en évidence la faisabilité d'un appariement entre le SNIIRAM et la base des causes de décès. Les pistes d'exploitations d'une telle base sont multiples, permettant en particulier d'observer des associations entre différentes consommations de soin et la survenue de décès par cause spécifiques, renforçant ainsi le niveau de preuve des associations mises en évidence.

1 - Appariement du PMSI-MCO aux causes médicales de décès via le Sniiram (2008-2009), France. Lamarche-Vadel A, Weill A, Blotiere Po, Moty-Monnereau C., Jougla E, Rey G. ADELF-EMOIS, Dijon, mars 2012.

2 - Automated comparison of last hospital main diagnosis and underlying cause of death ICD10 codes, France, 2008-2009. Lamarche-Vadel A, Pavillon G, Aouba A, Johansson LA, Meyer L, Jougla E, Rey G. BMC Med Inform Decis Mak. 2014 Jun 5;14(1):44.

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

Enrichissement d'enquêtes avec recueil de données auprès des personnes

Dès 2006, le Conseil national de l'information statistique (CNIS) a souligné la complémentarité entre sources administratives et données d'enquêtes [19]. L'appariement de données administratives avec des données d'enquête s'est développé ces dix dernières années [20], mais reste encore relativement rare.

Depuis plusieurs années, l'Insee et l'ensemble du système statistique public se sont engagés dans un mouvement croissant de recours aux sources administratives et médico-administratives. Dans cette perspective, il a été décidé de compléter le dispositif des enquêtes Handicap-Santé (HSM) 2008 et 2009 en les appariant pour la première fois avec les données de remboursements de l'assurance maladie (SNIIRAM). Parmi les avantages, on peut citer l'intérêt pratique, l'appariement permettant notamment de réduire le temps d'enquête et d'alléger la charge de réponse pour les enquêtés, ainsi que l'amélioration de la qualité des études par le recueil de données sans « biais de mémoire » ou de « biais de désirabilité sociale[7] » comme cela peut être le cas pour les enquêtes en face-à-face, par téléphone ou par auto-questionnaire. Cette base de données issue de l'appariement HSM-SNIIRAM a permis de développer des travaux originaux, en particulier un chiffrage des dépenses de santé des personnes âgées dépendantes dans le cadre de la préparation de la réforme de la dépendance.

Le suivi de cohortes épidémiologiques longitudinales peut également bénéficier de l'enrichissement des données recueillies directement auprès des sujets par des données de consommation de soins et d'hospitalisation dans les bases de données de santé. Un exemple est celui de l'étude Entred (*Échantillon national témoin représentatif des personnes diabétiques*) coordonnée par l'InVS, qui s'intéresse à la qualité de prise en charge médicale des diabétiques, à l'évolution du contrôle des facteurs de risque vasculaire et de la fréquence des complications chez les personnes diabétiques [21]. D'autres projets complétant les recueils auprès de personnes par les bases de données nationales sont menés en collaboration avec la CNAMTS (Encadré 3). A titre d'exemple, on peut citer l'enquête sur la santé et la protection sociale (ESPS) de l'Irdes [22], le projet Cesir (étude de l'influence de la consommation de médicaments et de l'état de santé sur l'insécurité routière, réalisée par l'Unité 897 de l'Inserm [23]) ou les programmes de suivi post-professionnel de retraités ayant été exposés à l'amiante Spirale (UMS 011 Inserm-UVSQ et CNAMTS [24]) et ESPrI (DST-InVS et RSI [25]).

ENCADRÉ 3 - CONSTANCES

La cohorte Constances est une très grande cohorte épidémiologique destinée à contribuer au développement de la recherche épidémiologique et à fournir des informations à visée de santé publique. Réalisée dans le cadre d'un partenariat avec la Caisse nationale d'assurance maladie des travailleurs salariés (CNAMTS) et la Caisse nationale d'assurance vieillesse (CNAV), labellisée *Infrastructure nationale en biologie et santé* par les Investissements d'avenir, cette cohorte a vocation à constituer une infrastructure accessible à la communauté de recherche conçue pour aider à analyser une large gamme de problèmes scientifiques. Constances a été également conçue comme un outil d'information de santé publique et de surveillance, par le caractère particulièrement complet du dispositif de suivi et de recueil d'informations très diversifiées auprès d'un échantillon de 200 000 sujets représentatif de la population adulte couverte par le Régime général de Sécurité sociale.

De très nombreuses données sont recueillies concernant la santé : antécédents médicaux, échelles de santé et de qualité de vie, pathologies, affections de longue durée et hospitalisations, absence au travail, handicaps, limitations, incapacités et traumatismes, cause médicale de décès, comportements de santé (tabac, alcool, alimentation, activité physique, cannabis, orientation sexuelle), problèmes de santé spécifiques des femmes, recours aux soins et prise en charge. Un examen de santé réalisé dans les Centres d'examens de santé (CES) de la CNAMTS permet de recueillir également des données complémentaires : poids, taille, rapport taille-hanches, tension artérielle, fréquence cardiaque, vision, audition, spirométrie, investigations biologiques ; une batterie de tests physiques et cognitifs est réalisée pour les personnes âgées de 45 ans ou plus, et des échantillons biologiques seront conservés dans une biobanque. Dans le but de mieux connaître les déterminants sociaux de la santé et les inégalités sociales de santé, des données sociodémographiques et professionnelles sont collectées : situation d'emploi, niveau d'études, revenus, situation matrimoniale, composition du ménage, conditions de vie matérielles, histoire professionnelle, expositions professionnelles, stress au travail.

Ces données sont recueillies de façon prospective à des sources multiples : autoquestionnaires, examens de santé et appariement avec les bases de données du SNIIRAM, de la CNAV et du CépiDc.

[7] Le biais de désirabilité sociale est une tendance, consciente ou inconsciente, qui consiste à vouloir se présenter sous un jour favorable à ses interlocuteurs.

Le dispositif mis en place pour la constitution et le suivi de la cohorte est complexe et fait appel à divers acteurs : (1) pour assurer la représentativité de l'échantillon, la CNAV tire au sort dans ses bases de données des assurés sociaux les personnes éligibles ; (2) la liste des NIR de ces personnes est alors transmise à la CNAMTS pour constituer un fichier d'adresses postales qui est transmis à un tiers de confiance afin de permettre l'envoi par courrier des invitations à participer ; la CNAMTS applique également la procédure FOIN (Encadré 1) pour extraire du SNIIRAM les données de santé concernant les personnes volontaires ; (3) les personnes volontaires complètent ensuite des questionnaires concernant leur santé, leurs modes vie, leur histoire professionnelle, etc., et se rendent dans un CES pour l'examen de santé initial dont les résultats sont transmis à l'équipe Constances ; (4) par la suite, chaque année un questionnaire (postal ou Internet) est complété par les participants pour suivre dans le temps l'évolution de l'état de santé, de la situation socio-économique et professionnelle, de l'environnement familial, social et de lieu de vie, des facteurs de risque personnels et environnementaux ; une invitation à revenir au CES tous les 5 ans est proposée ; chaque année, les principaux événements socioprofessionnels sont extraits des bases de données de la CNAV, les données de santé sont extraites du SNIIRAM et les causes de décès proviennent du CépiDc.

Toutes ces procédures sont hautement sécurisées ; les données identifiantes (nom, adresse) sont conservées par un tiers de confiance, toutes les données transmises sont cryptées, et le NIR des participants reste confiné à la CNAV et à la CNAMTS.

Constances est une infrastructure de recherche ouverte à la communauté scientifique. Actuellement, une quarantaine de demandes ont été reçues, provenant de nombreuses équipes de recherche et portant sur des thèmes diversifiés : pathologies spécifiques (pathologie respiratoire chronique, troubles musculo-squelettiques, cirrhose, dépression…), états de santé (vieillissement, fonctionnement physique et cognitif, hypertension, ….), comportements, facteurs psychologiques, facteurs de risque professionnels et environnementaux, inégalités sociales de santé.

Grâce à la participation de la CNAMTS et de la CNAV, ainsi que du CépiDc, qui réalisent les appariements avec leurs bases de données, il est possible de recueillir de très nombreuses informations sur la santé et les trajectoires socioprofessionnelles des participants de la cohorte dans des conditions exceptionnelles : données structurées et constamment actualisées, suivi dans le temps assuré sans « perdus de vue », car les participants sont toujours présents dans les bases du SNIIRAM et de la CNAV. De plus ces appariements évitent de surcharger les volontaires de lourds questionnaires, et permettent d'obtenir des données souvent plus fiables que celles qu'on peut recueillir auprès des personnes (listes de médicaments, dates précises de consultation ou d'hospitalisation, diagnostics portés par des médecins, etc.).

Le suivi de cohortes d'enfants pose des problèmes particuliers, du fait de problèmes techniques pour l'identification des enfants dans le SNIIRAM, et d'aspects juridiques et éthiques. L'exemple de la cohorte ELFE (Encadré 4) illustre à la fois le très grand intérêt de l'appariement de la cohorte avec le SNIIRAM et les difficultés rencontrées.

ENCADRÉ 4 - ELFE

La cohorte ELFE (*Étude Longitudinale Française depuis l'Enfance*) est la première cohorte nationale d'enfants suivis depuis leur naissance. Constituée par plus de 18 000 familles avec un enfant né en 2011, son objectif général est de comprendre comment les conditions périnatales, l'environnement dans ses différentes dimensions (familiale, socio-économique, géographique, expositions physico-chimiques) affectent, de la période intra utérine à l'adolescence, le développement, la santé et la socialisation des enfants.

Les données de santé recueillies dans le cadre du suivi des enfants reposent essentiellement sur les questionnaires posés périodiquement aux parents et, lors de certaines phases de l'enquête, au médecin qui suit l'enfant. Les limites de ces sources de données sont la fiabilité du report des parents, les biais de mémorisation, la disponibilité des professionnels de santé pour répondre à une enquête, mais également de celle des parents qui peuvent temporairement ou définitivement ne plus participer au suivi. De plus comme ses objectifs très larges le montrent, l'étude ELFE est une étude multidisciplinaire et le nombre de questions que l'on peut poser aux parents sur une thématique donnée est obligatoirement limité. Il est donc exclu d'obtenir des informations détaillées sur, par exemple, les consommations de médicaments.

Le croisement des données recueillies sur les enfants avec les données du SNIIRAM permettra de compléter et/ou de valider les informations qui sont collectées auprès des parents sur les pathologies, les séjours hospitaliers et les achats de médicaments.

En ce qui concerne la pharmacovigilance, l'effectif de la base de données de l'étude ELFE ne permettra pas de détecter des effets secondaires dont la fréquence est inférieure à environ 1/6000, ce qui est souvent le cas pour des effets secondaires graves et non connus de médicaments. En revanche, en cas de détection d'anomalies par la pharmacovigilance, la base de données ELFE pourra servir à l'exploration des conditions de prescription chez les femmes enceintes et les enfants en France. Une dernière utilisation potentielle des données du SNIIRAM associées à Elfe serait de valider certains algorithmes utilisés en pharmacovigilance. Par exemple dans l'étude des effets secondaires des vaccins dans les études de pharmacovigilance, la date de réalisation d'un vaccin est imputée au moyen d'un algorithme en fonction de sa date d'achat. La base de données Elfe appariée comprendrait la date d'achat du vaccin à partir des données SNIIRAM et la date de sa réalisation à partir du relevé du carnet de vaccination.

Lors de l'inclusion dans la cohorte Elfe, un consentement a été obtenu auprès de 95 % des familles pour effectuer l'appariement entre les données directement recueillies auprès d'elles et celles du SNIIRAM. L'appariement concernera les hospitalisations et consommations de soins des enfants tout au long de l'étude et des mères uniquement pendant la grossesse. Comme dans toute étude longitudinale, un certain nombre de familles désirent au cours du suivi interrompre leur participation. Lorsqu'elles le notifient à l'équipe ELFE, un courrier leur est adressé confirmant la prise en compte de leur demande et leur indiquant que, sauf si elles s'y opposent, l'appariement avec les données du SNIIRAM se poursuivra. Il permettra de poursuivre le recueil d'information sur la santé de leur enfant sans les importuner et de valoriser au mieux les données qu'elles nous ont communiquées lors des phases précédentes.

La réalisation de l'appariement pose cependant des problèmes particuliers quand il s'agit d'enfants. En effet leur consommation de soins dans le SNIIRAM est rattachée à celle de leurs ouvrant-droits (mais il est prévu une évolution sur ce point). Celui-ci peut être indifféremment la mère ou le père

et changer au cours du temps. Le père peut également être l'ouvrant-droit de la mère ou inversement. En ce qui concerne l'étude ELFE, en l'absence d'autorisation du recueil du NIR de l'enfant (qui du reste n'est pas connu à la naissance), ni de celui de ses parents, une procédure alternative est possible, consistant à : (1) recueillir précisément des données d'état-civil pour chacun des parents ; (2) confier ces informations à la CNAV, complétées de la date de naissance et du sexe de l'enfant ainsi que d'un numéro d'identification Elfe spécifique. A partir de ces éléments, la CNAV est en mesure : (3) d'interroger le Répertoire national inter-régimes de l'assurance maladie (RNIAM), afin de dresser la liste des identifiants SNIIRAM possible des enfants, et (4) de transmettre ces informations à la CNAMTS de manière cryptée. Le rôle de la CNAMTS consistera alors ensuite à : (5) appliquer le double algorithme d'anonymisation des NIR (Foin 1 et 2) et (6) extraire les données individuelles de l'assurance maladie de l'enfant et de la mère. Au final, seules les données de consommation de soins, associées à l'identifiant ELFE, seront transmises à l'équipe ELFE.

Les enfants étant des personnes vulnérables, la sécurité des données les concernant recueillies dans une étude demande une vigilance toute particulière. Les données du SNIIRAM seront conservées indépendamment des autres données de l'étude et seuls des indicateurs synthétiques construits spécifiquement pour la recherche seront intégrés dans le système de stockage des données ELFE.

Une utilisation encore trop restreinte des bases médico-administratives

Il est clair que les bases de données administratives et médico-administratives nationales ont un potentiel considérable. Elles répondent à des besoins d'information très diversifiés de surveillance, d'études et de recherches, dépassant largement les préoccupations des organismes qui les constituent et les gèrent, et peuvent rendre de grands services à la communauté de santé publique et de recherche. Les quelques exemples rapportés ici l'illustrent, mais n'offrent qu'une vision très partielle des multiples usages qu'on pourrait envisager si les restrictions actuelles à l'utilisation des bases médico-administratives étaient levées et leur utilisation facilitée. Ainsi, il existe des situations où il est nécessaire de recontacter les personnes ayant une maladie rare ou bénéficiant d'un dispositif médical ou d'un médicament dont il apparaît qu'ils présentent des risques pour la santé et qu'il faut alerter ou pour lesquelles un suivi médical spécifique doit être mis en place ; alors que les informations concernant ces personnes sont disponibles dans le SNIIRAM, il est actuellement difficile de les faire bénéficier d'une prise en charge adéquate, du fait des textes encadrant l'usage du NIR. On peut aussi souligner qu'à l'heure du « Big Data », une exploitation plus facile du SNIIRAM permettrait sans doute des analyses originales de cet ensemble gigantesque de données, grâce à des méthodes statistiques et épidémiologiques en plein développement.

Dans notre pays, les bases de données administratives et médico-administratives nationales restent encore, malgré leur considérable apport, insuffisamment exploitées en dehors des organismes qui les constituent et les gèrent, même si plus d'une centaine de publications référencées, se rapportant à des travaux réalisés sur les données de remboursement de l'assurance maladie, ont été recensées en juin 2009 [26]. Si l'utilisation de l'EGB a beaucoup augmenté ces dernières années grâce aux efforts de la CNAMTS pour en faciliter l'accès (plus de 22 000 requêtes effectuées en 2012 par divers acteurs de l'assurance maladie, des Agences régionales de santé (ARS), du ministère de la santé, des agences sanitaires, des organismes de recherche), le SNIIRAM complet, qui inclue les données de l'ensemble de la population, reste notoirement sous-utilisé : à la fin 2012, seules 26 demandes d'extraction avaient été enregistrées [27].

Cette situation peut s'expliquer du fait d'obstacles divers, dont les plus importants sont de nature juridique et opérationnelle : (1) l'identifiant utilisé de façon directe ou cryptée par les bases de données nationales étant le NIR, la quasi-interdiction de fait de le recueillir auprès des personnes ou des organismes qui en disposent limite très fortement les possibilités d'accès aux bases de données, qui n'est possible que dans certaines circonstances et qui, de plus, nécessite une participation active et un travail conséquent des gestionnaires de ces bases ; (2) la mise à disposition des données à la communauté de santé publique et de recherche dans des conditions en permettant l'exploitation, nécessite des ressources scientifiques, techniques et organisationnelles complexes et de haut niveau de compétence.

Pour une meilleure utilisation des bases médico-administratives

L'existence dans notre pays de bases de données médico-administratives de très grande qualité est une chance unique pour que la France joue un rôle majeur dans la recherche en épidémiologie et en santé publique, car (exception faite des pays scandinaves, mais avec des tailles de population plus réduites) aucun pays ne dispose d'un tel dispositif à l'échelle

de 65 millions d'habitants [28]. Comme on l'a souligné, ce potentiel reste encore trop peu utilisé, et il est urgent de lever les divers obstacles qui entravent le développement de leur utilisation tout en respectant strictement la protection des données des personnes et en améliorant la fiabilité des données.

Ainsi, la possibilité d'utiliser le NIR serait une amélioration cruciale, car celui-ci constitue la clé d'entrée quasiment indispensable dans les bases de données nationales. Si, comme on l'a montré, il est parfois possible de procéder par appariement indirect et s'exonérer ainsi de la disponibilité du NIR, ceci nécessite de disposer déjà de certaines informations enregistrées dans le SNIIRAM concernant les personnes pour lesquelles on veut obtenir des données, ce qui n'est pas le cas dans de nombreuses situations. De plus, la réalisation d'un appariement indirect demande un important travail aux chercheurs et les résultats ne sont jamais complètement satisfaisants. Les textes actuels encadrant l'utilisation du NIR, qui constituent de fait une protection peu efficace des données à caractère personnel, doivent être modifiés pour en permettre un usage facilité dans des conditions strictement encadrées et respectueuses de la confidentialité[8].

Restent les importantes difficultés méthodologiques et techniques d'utilisation des bases de données médico-administratives pour la recherche, du fait de leur très grande complexité et de diverses limites. La CNAMTS a fait ces dernières années de très importants efforts pour faciliter l'utilisation du SNIIRAM, et diverses initiatives ont par ailleurs vu récemment le jour pour développer des outils qui permettront aux équipes de recherche d'en bénéficier dans de meilleures conditions : mise en place du réseau REDSIAM (réseau de données du SNIIRAM) qui associe de nombreuses équipes sous l'égide des principaux organismes producteurs et utilisateurs des bases, réalisation de logiciels d'interface et de plateforme d'accès comme la plateforme d'interface pour les chercheurs en cours de réflexion sous l'égide de l'ITMO (Institut thématique multi-organismes) de Santé publique d'Aviesan (Alliance nationale pour les sciences de la vie et de la santé), établissement de catalogues de données, appels à projets de l'Institut de recherche en santé publique (IReSP) et de l'Agence nationale de la recherche (ANR) pour l'usage secondaire des bases de données et pour la recherche méthodologique, etc. Ces efforts doivent être poursuivis et développés afin que les bases de données médico-administratives fassent partie de la panoplie des outils « ordinaire » des chercheurs.

[8] Cette disposition est prévue par l'article 47 du projet de loi de modernisation de notre système de santé.

Bibliographie

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

2. Couris CM, Forêt Dodelin C, Rabilloud M, Colin C, Bobin JY, Dargent D, Raudran D, Schott AM. Sensibilité et spécificité de deux méthodes d'identification des cancers du sein incidents dans les services spécialisés à partir des données médico-administratives. Rev Epidemiol Sante Publique 2004, 52, 151-60.

3. Couris CM et al. Method of correction to assess the number of hospitalized incident breast cancer cases based on claims databases. Journal of Clinical Epidemiology, 2002, 55 : 386-391.

4. Incidence médico-sociale des ALD30 en 1999. CNAMTS-DSM-Mission des Soins de ville-Mission Statistique. Avril 2004. Disponible sur le site www.ameli.fr/245/doc/1391/article_pdf.html.

5. Grosclaude et al et Lauzeille et al, BEH 2012 (à paraître).

6. Deprez Ph-H, Chinaud F, Clech S, Germanaud J, Weill A, Cornille JL, Fender P. La population traitée par médicaments de la classe des antihistaminiques en France métropolitaine : données du régime général de l'assurance maladie, 2000. Revue médicale de l'assurance maladie Avril-juin 2004, 35 (1), 3-11.

7. Lecadet J, Vialaret K, Vidal P, Baris B, Fender P. Mesure à l'échelle d'une région des effets d'un programme national d'information sur le bon usage des antibiotiques. Revue médicale de l'assurance maladie Avril-Juin 2004, 35 (2) ,81-91.

8. Fender P, Weill A. Epidémiologie, santé publique et bases de données médico-tarifaire. (Éditorial) Rev Epidemiol Santé Publique, 2004, 52,113-117.

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

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

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

12. Utilisation des données de l'assurance maladie pour évaluer l'impact sanitaire d'une épidémie de gastro-entérites d'origine hydrique, Bourg-Saint-Maurice (Arc 1800), 2006. BEH 31, 6 septembre 2011.

13. Diène E, Geoffroy-Perez B, Cohidon C, Gauvin S, Carton M, Fouquet A, Fatras JY, Imbernon E. Psychotropic drug use in a cohort of workers 4 years after an industrial disaster in France. J Trauma Stress. 2014;27:430-7.

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

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

16. http://www.irdes.fr/EspaceRecherche/Partenariats/Hygie/Presentation.html

17. DST-InVS. Analyse de la mortalité et des causes de décès par secteur d'activité de 1968 à 1999 à partir de l'Échantillon démographique permanent. InVS, Septembre 2006.

18. Saurel-Cubizolles MJ, Chastang JF, Menvielle G, Leclerc A, Luce D; EDISC group. Social inequalities in mortality by cause among men and women in France. J Epidemiol Community Health. 2009;63:197-202.

19. Cnis. Chroniques n° 5, Enquêtes statistiques et sources administratives : une complémentarité à exploiter, 2006.

20. Gensbittel M.-H., Riandey B., Appariements sécurisés et statistiques (2000-2011) : Une décennie d'expériences. Courrier des statistiques, 131, Septembre 2011.

21. http://www.invs.sante.fr/entred/

22. http://www.irdes.fr/EspaceRecherche/Enquetes/ESPS/index.html.

23. Orriols et al. (2010) Prescription Medicines and the Risk of Road Traffic Crashes: A French Registry-Based Study. PLoS Med 7(11): e1000366. doi:10.1371/journal.pmed.1000366.

24. Carton M, Bonnaud S, Nachtigal M, Serrano A, Carole C, Bonenfant S, Coste D, Lepinay P, Varsat B, Wadoux B, Zins M, Goldberg M. Post-retirement surveillance of workers exposed to asbestos or wood dust: first results of the French national SPIRALE Program. Epidemiol Prev. 2011;35:315-23.

25. http://www.invs.sante.fr/surveillance/espri/default.htm.

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

27. Source : CNAMTS

28. Goldberg M, Jougla E, Fassa M, Padieu R, Quantin C. The French public health information system. Stat J Int Assoc Official Statistics 2011. 27: 1–11 1.doi: 10.3233/sji-2011-0747.

ANNEXE 1 : LE POUVOIR DE RÉ-IDENTIFICATION DES BASES NATIONALES DE DONNÉES DU PMSI

(Article présenté le 18 mars 2011 à Nancy lors des Journées ÉMOIS par le Dr. Dominique Blum)

AVERTISSEMENT DE LA REDACTION

Cet article rédigé et présenté en 2011 n'avait jamais été publié dans son intégralité (seul un résumé avait paru en juin 2011 dans la « Revue d'épidémiologie et de santé publique », volume 59, page S54). Il s'agissait à la fois d'une analyse de fond sur un risque sous-estimé et d'un « lancement d'alerte » avec ce que cela impliquait de critiques contre ce qui apparaissait comme une forme de négligence de la part des autorités compétentes. Comme on le sait, cependant, cette alerte a été entendue, prise en compte dans le rapport Bras de 2013 puis dans le projet de loi de modernisation de notre système de santé, en cours d'examen par le Parlement, si bien que les recommandations à la fin de l'article n'ont plus aujourd'hui qu'un intérêt historique.

Il a semblé utile de reproduire ici cet article, même s'il est déjà ancien, en raison de son intérêt historique, et surtout parce qu'il est en partie à l'origine des travaux présentés dans ce *Dossier solidarité santé*. Il aurait certes été envisageable de rédiger une version actualisée et « lissée », mais il a finalement paru préférable d'en rester à la version initiale et de laisser à son auteur la responsabilité de certaines de ses appréciations. Par rapport à la version originale, quelques corrections de forme ont été apportées, et des notes de bas de page de la rédaction complètent ou actualisent l'analyse.

Objectif et contexte technique de l'étude

La réflexion qui a conduit à cette étude est la suivante : en raison de la richesse informationnelle des bases de données nationales anonymes du PMSI, il risque de devenir possible d'y ré-identifier un patient à condition de connaître de lui quelques traits caractéristiques, et d'accéder ainsi à son insu aux informations confidentielles afférentes à sa santé, contenues dans ces bases.

Les objectifs de cette étude sont donc de quantifier le pouvoir de ré-identification des patients dans les bases nationales du PMSI, et d'attirer sur le risque encouru l'attention des acteurs du PMSI, notamment institutionnels, de s'interroger sur ses causes et de proposer des pistes pour y remédier à l'avenir.

Depuis le début des années 1990, dans le cadre du programme de médicalisation des systèmes d'information (PMSI) les hôpitaux et les cliniques recueillent et codent au fil de l'eau un résumé de sortie standardisé (RSS) pour chaque séjour réalisé en secteur de médecine, chirurgie, obstétrique et odontologie (MCO). Ces établissements mettent en œuvre ensuite un logiciel national développé pour l'État par l'Agence technique de l'information sur l'hospitalisation (ATIH), qui classe chaque séjour dans l'un des groupes que compte la classification française des groupes homogènes de malades (GHM). Puis un dispositif d'anonymisation développé par l'ATIH transforme chaque RSS en un résumé de sortie anonymisé (RSA). Enfin un processus de centralisation des RSA comportant une seconde anonymisation transfère l'intégralité de ceux-ci dans les serveurs de l'ATIH, via une plateforme internet spécifique.

La fonction première de cette centralisation est le financement des établissements puisque le PMSI est devenu en 2005 le support de la tarification à l'activité (T2A). Le processus de ce calcul sort du cadre de notre étude. Notons simplement que la réalisation de cet objectif nécessite de disposer de l'intégralité des séjours et qu'en pratique la base nationale de données du PMSI est effectivement exhaustive : elle comporte un RSA pour chaque séjour en secteur MCO de quelque structure d'hospitalisation publique ou privée que ce soit.

Avec les deux anonymisations consécutives précédemment décrites, combinées à l'absence de table de correspondance entre les trois identifiants successifs (identifiant d'origine, identifiant anonyme de premier niveau et identifiant anonyme de second niveau), la base de données des RSA constituée à l'ATIH est réputée anonyme : quel que soit l'étage de la chaîne de production auquel on se place, il est impossible d'établir la correspondance entre l'identifiant administratif du séjour-patient contenu dans le RSS et l'identifiant anonyme du RSA enregistré dans la base nationale.

Partant donc du principe que cette base nationale de données est anonyme, la CNIL et les pouvoirs publics ont autorisé depuis la fin des années 1990 la diffusion de copies sur cédéroms : outre la fonction économique et budgétaire pour laquelle elle a été conçue, elle se révèle en effet être une mine d'informations exhaustive et d'excellente qualité pour tous les chercheurs – au sens large – qui consacrent leurs travaux à l'offre de soins, à la santé publique, et à l'épidémiologie hospitalières.

Car si le RSS de 1985 ne comptait en tout et pour tout que 22 informations, il s'est tellement enrichi que le RSA de 2008 en compte 72 dans sa partie fixe, et jusqu'à un maximum de 100 677 dans sa partie variable. Chacune des trois catégories d'informations qu'il véhicule s'est en effet accrue pour comporter désormais :

- au titre des informations administratives : l'identifiant de l'établissement, le numéro d'ordre du RSA, le sexe du patient, son âge, la durée de son séjour, le mois de sa sortie, son mode d'entrée et sa provenance, son mode de sortie et sa destination, le nombre de services fréquentés pendant son séjour, la durée de son séjour dans chacun d'entre eux, son code géographique de résidence ;
- au titre des informations médicales : d'une part la liste des diagnostics pris en charge lors du séjour (un au minimum et 101 au maximum), d'autre part celle des actes médicaux et chirurgicaux réalisés au cours de celui-ci (9 999 au maximum) ainsi que le délai écoulé entre la date d'entrée et la réalisation de chacun d'eux, et enfin quelques informations diverses : nombre de séances, poids de naissance et âge gestationnel pour les nouveau-nés, score de gravité simplifié utilisé en réanimation, qualification médicale de chaque service fréquenté au cours du séjour ;
- au titre des informations dites médico-économiques :
  - le résultat de l'algorithme de classement national des séjours MCO (dit « fonction groupage MCO ») qui se compose de deux éléments : la catégorie majeure de diagnostic (CMD) et le groupe homogène de malades (GHM). Pour mémoire, la classification complète comporte près de 3 000 GHM distincts répartis en 28 CMD ;
  - le groupe homogène de séjour (GHS), groupe tarifaire correspondant au GHM ;
  - le nombre de journées ouvrant droit à un supplément ;
  - un renseignement déterminant un éventuel abattement du tarif ;
  - une série de 24 renseignements ouvrant éventuellement droit au versement de suppléments ;
  - la valorisation partielle de chaque service fréquenté au cours du séjour ;

L'anonymisation du RSA consiste à remplacer l'identifiant individuel du RSS par une clef cryptée, dite parfois numéro d'anonymat. C'est un logiciel fourni par l'ATIH qui s'en charge. Cette clef est le résultat d'un algorithme non réversible, qui a les caractéristiques d'une application injective au sens mathématique du terme : d'une part pour un patient déterminé cette clef est la même quels que soient les établissements d'hospitalisation public ou privé qui l'ont pris en charge et quelles que soient les dates de ses séjours (autrement dit, tous les séjours hospitaliers d'un patient sont identifiés par une clef unique), d'autre part cette clef est distincte pour deux patients distincts.

En pratique la clef n'est pas enregistrée dans le RSA lui-même, mais dans un fichier dit « de chaînage » qui établit la correspondance entre le numéro d'ordre de chaque RSA dans la base nationale et la clef de chaînage (voir figure n°1).

Outre la clef de chaînage, le fichier de chaînage comporte une information essentielle mais non documentée que nous nommons « index chronologique », destinée à ordonner les séjours multiples d'un même patient, soit pour étudier les parcours de soins, soit pour détecter les recouvrements de séjours, involontaires ou non (erreurs de saisie de dates, fraude). Cette information représente le nombre de jours écoulés entre une date de référence propre à chaque patient (date fictive, dérivée de sa clef de chaînage) et la date d'entrée du séjour mentionnée dans le RSS. Par soustraction elle permet donc de calculer le délai écoulé entre deux hospitalisations d'un patient, sans toutefois permettre de dater précisément chaque séjour, ni de calculer un délai entre les séjours de patients distincts puisque leurs dates de référence sont distinctes (voir figure n°2).

FIGURE N°1

Principe de transformation du RSS en RSA (anonymisation)

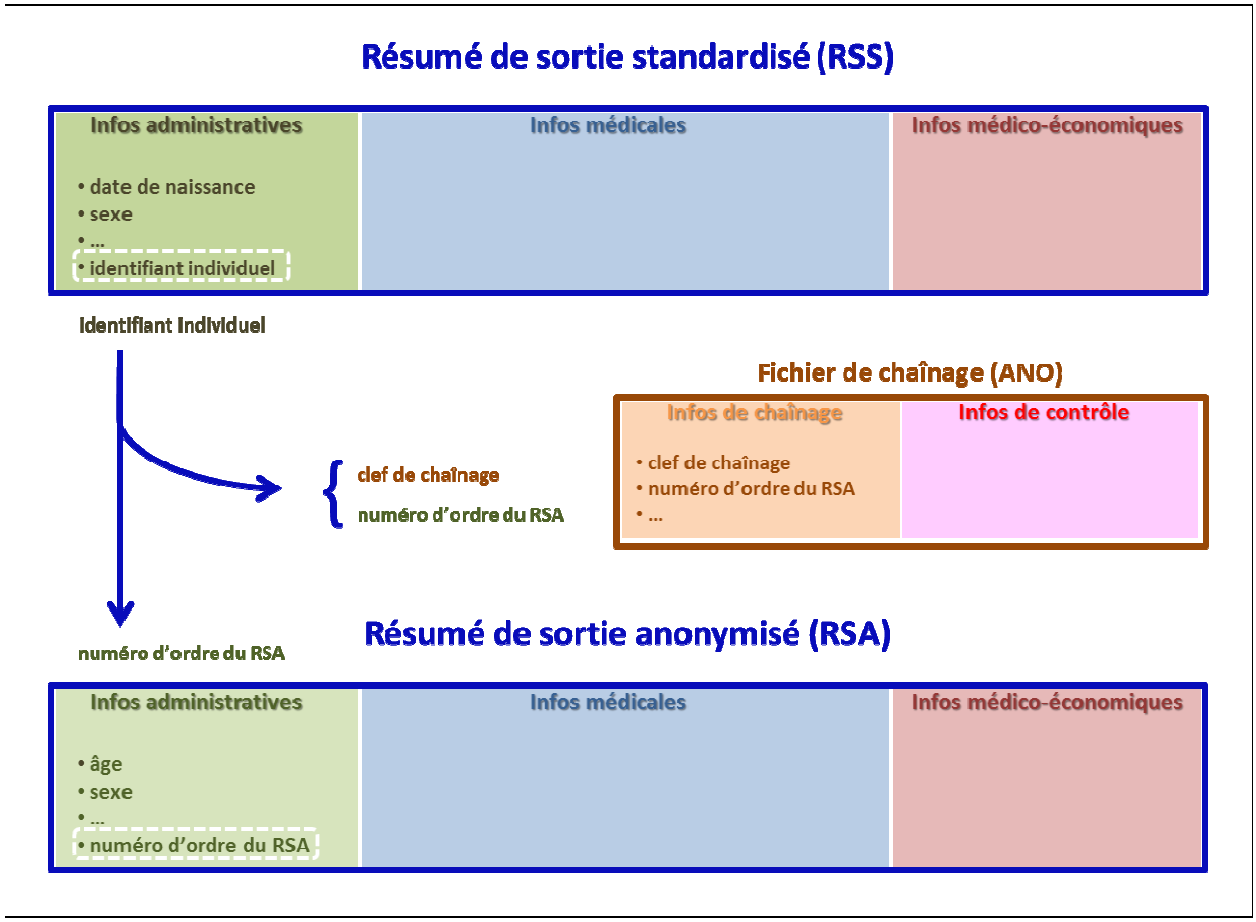
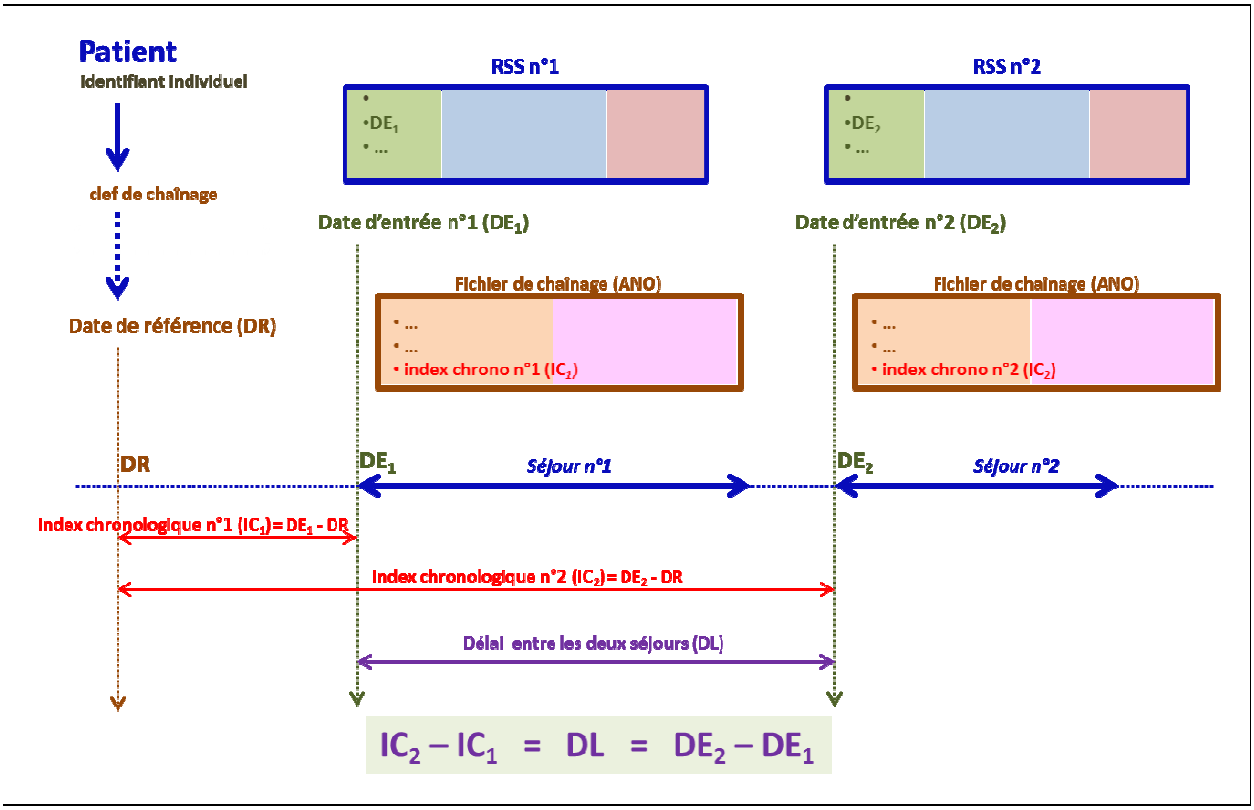


FIGURE N°2

L'index chronologique et son utilisation



Matériel et méthode

Matériel

L'analyse porte sur la base nationale du PMSI-MCO de 2008 fournie par l'ATIH. Cette base de données comporte 23 828 892 RSA et un fichier de chaînage comportant le même nombre d'enregistrements, totalisant 12 351 219 clefs de chaînage distinctes.

Les RSA dérivés de RSS erronés ou mal conformés, et signalés comme tels par l'ATIH, ont une clef de chaînage inexploitable. Cela représente 2 698 864 RSA (soit 11,3 % du fichier complet) qui sont supprimés de l'étude.

D'autre part nous considérons que les séjours comportant des séances ne doivent pas être pris en compte. En effet, plus le nombre de venues pour séances s'accroît, plus la suite complète des délais entre deux venues devient singulière, donc discriminante pour la ré-identification d'un patient. Or par essence les séances sont répétitives et de surcroît le nombre total de RSA mentionnant des séances (radiothérapie, chimiothérapie, dialyse) est très important, de sorte que leur prise en compte constituerait indiscutablement un biais pour l'étude. Par ailleurs la tolérance qui permet aux établissements d'opter pour un recueil « global » des séances rend inexploitable dans ce cas le délai entre deux venues. Pour toutes ces raisons, nous supprimons de l'étude 4 603 831 RSA comportant une ou plusieurs séances.

Les 16 526 197 RSA restant correspondent à des séjours sans séance réalisés par 10 837 919 patients distincts, identifiés chacun par sa clef de chaînage individuelle. Parmi eux, 248 259 ont également eu des séjours avec séances, supprimés précédemment. Ils totalisent 613 573 séjours sans séance, soit 2,5 séjours en moyenne par patient. Les 10 589 460 autres patients totalisent pour leur part 15 912 624 séjours, soit 1,5 séjour en moyenne par patient. Il semble donc que d'une manière générale les patients « susceptibles de bénéficier de séances » soient plus souvent hospitalisés que les

autres. Pour cette raison, étant donné que nos estimations du pouvoir de ré-identification prendront en compte la succession des délais entre séjours et des durées de séjour de chaque patient, nous choisissons de supprimer tous les séjours de ces patients afin de ne pas majorer ces estimations.

C'est donc finalement un ensemble de 15 912 624 séjours totalisés par 10 589 460 patients qui sont intégrés à l'étude. Parmi eux, 7 887 767 patients (74,5%) n'ont séjourné qu'une seule fois à l'hôpital en 2008, ce qui représente 49,6% des séjours. Autrement dit, trois quarts des patients représentent la moitié des séjours sans séance.

Méthode

Nous mesurons le pouvoir de ré-identification d'une manière directe sur l'intégralité des dossiers retenus. Pour ce faire, nous reconstituons d'abord le parcours hospitalier de chaque patient, en agrégeant tous les RSA portant la même clef de chaînage et en les ordonnant grâce à l'index chronologique décrit plus haut. Comme indiqué précédemment, seules les informations administratives sont susceptibles d'être connues par un tiers cherchant à identifier un patient déterminé. Aussi nous ne retenons que les informations administratives des RSA, auxquelles nous ajoutons, pour chaque RSA, le délai écoulé entre sa date d'entrée et celle du suivant par ordre chronologique. Ce délai est obtenu par soustraction entre les index chronologiques des deux RSA concernés. Pour le dernier RSA d'un parcours individuel, par convention nous fixons ce délai à zéro.

Ayant décidé de nous concentrer non pas sur chaque séjour mais sur le parcours complet des patients, nous avons retenu pour chaque parcours : le sexe du patient, son âge à l'entrée dans le parcours, le mois de sortie du dernier séjour du parcours, le code géographique de son domicile, l'identifiant de l'établissement d'entrée dans le parcours, le mode de sortie du dernier séjour du parcours et « l'empreinte chronologique » du parcours.

Nous avons appelé empreinte chronologique du parcours la combinaison des durées de séjour du premier, deuxième et troisième séjours du parcours et du délai inter-séjour du premier et deuxième séjours[1]. Quand le nombre de séjours du parcours est inférieur à trois, les informations « manquantes » (durée et délai) sont fixées à zéro par convention.

En imaginant la démarche d'un intrus cherchant à ré-identifier un patient[2] à partir de traits caractéristiques qu'il connaît à son sujet, nous pouvons considérer que toutes les informations retenues sont autant de critères potentiellement utiles à cette fin. Nous allons donc envisager diverses associations d'un ou plusieurs de ces critères, selon leur intérêt présumé. Chacune des modalités d'un des critères pourrait, ou non, être observée en association avec chacune des modalités de chacun des autres critères, déterminant ainsi des combinaisons réellement observées dans la base nationale du PMSI. Nous dénombrerons alors la proportion des combinaisons observées qui ne comportent qu'un seul patient, ou deux patients, ou trois patients : plus cette proportion est élevée, plus la ré-identification est facilitée (et plus le nombre de combinaisons atteint, voire dépasse le nombre total de patients) ; à l'inverse plus les combinaisons regroupent un nombre important de patients, moins il est possible de ré-identifier ceux-ci.

Pour illustrer la méthode, prenons deux exemples :

- pour tester le pouvoir de ré-identification simultané du sexe et du mois de sortie du premier séjour, à l'exclusion de toute autre information, nous établirions une association ne comportant que ces deux critères. En passant en revue la totalité des parcours, nous observerions qu'ils se ventilent dans vingt-quatre combinaisons distinctes (une par mois de l'année et par sexe) et que ces deux critères ayant des distributions à peu de choses près uniformes et indépendantes l'une de l'autre, chaque groupe compte finalement plus de 400 000 individus. Il est clair qu'en ne connaissant que la date de sortie du premier séjour et le sexe d'une personne déterminée, un intrus ne pourrait pas l'identifier dans la base nationale ;
- supposons à présent que l'intrus connaisse également l'âge et le département de la personne recherchée et posons l'hypothèse que les âges sont compris entre 0 et 99 ans, le département entre 01 et 99, et que les distributions sont uniformes et indépendantes les unes des autres (ce qui à l'évidence est faux) : nous aurions alors une équi-répartition

[1] 2015 : mais l'attaquant potentiel ne connaît pas toujours cette durée et ces délais exacts. On trouvera plus loin, en note, les résultats obtenus par la DREES avec les données de 2012, sans tenir compte du délai inter-séjour.

[2] Cette ré-identification permettant alors d'accéder à des informations médicales inconnues de l'attaquant, notamment le diagnostic, les comobrbidités et les actes pratiqués.

entre les 237 600 combinaisons distinctes (12 x 2 x 100 x 99), soit 45 patients par combinaison. Dans la réalité nos hypothèses sur les distributions n'étant pas vérifiées, on observe une distribution des effectifs par combinaison avec un minimum inférieur à la dizaine pour une poignée d'entre elles, et approchant quelques centaines pour la combinaison la plus fréquente.

En fait, pour peu qu'un intrus dispose d'informations administratives ou chronologiques supplémentaires, et sachant qu'en réalité les distributions ne sont ni uniformes, ni indépendantes, on conçoit aisément que le nombre de combinaisons observées puisse être très important, avec une forte proportion de combinaisons d'effectif faible.

En pratique, nous avons testé les quinze associations suivantes :

1. empreinte chronologique seule ;
2. empreinte, âge, sexe ;
3. empreinte, âge, sexe, mois de sortie ;
4. empreinte, âge, sexe, département de l'hospitalisation initiale ;
5. empreinte, âge, sexe, département de résidence du patient ;
6. empreinte, âge, sexe, département du patient, département de l'hospitalisation ;
7. empreinte, âge, sexe, identification de l'établissement initial ;
8. empreinte, âge, sexe, département de l'hospitalisation initiale, mois de sortie ;
9. empreinte, âge, sexe, département du patient, mois de sortie ;
10. empreinte, âge, sexe, département du patient, département de l'hospitalisation, mois de sortie ;
11. empreinte, âge, sexe, code géographique de résidence du patient ;
12. empreinte, âge, sexe, identification de l'établissement initial, mois de sortie ;
13. empreinte, âge, sexe, code géographique de résidence du patient, mois de sortie ;
14. empreinte, âge, sexe, code géographique de résidence du patient, identification de l'établissement initial, mois de sortie ;
15. empreinte, âge, sexe, code géographique de résidence du patient, identification de l'établissement initial, mois de sortie, mode de sortie

Nous avons effectué les dénombrements en premier lieu sur l'ensemble des parcours hospitaliers retenus. Au vu des résultats, il nous a paru indispensable de renouveler l'opération en ne prenant en compte que les patients ayant effectué au moins deux séjours hospitaliers dans l'année et qui, comme nous l'avons dit plus haut, représentent un quart des patients et la moitié des séjours sans séances.

Résultats3

Le tableau qui suit récapitule les dénombrements effectués pour chacune des quinze associations de critères énumérées ci-dessus, observées parmi les 10 589 460 parcours hospitaliers.

[3] 2015 : considérant que l'attaquant ne connaît pas nécessairement avec précision la durée des séjours et surtout le délai inter-séjour, la DREES a renouvelé récemment les calculs, sur la base MCO 2012 sans séances (18 millions de séjours), avec des critères plus courants et sans chaînage. La proportion de séjours présentant des caractéristiques uniques selon les associations de critères retenus s'établit comme suit :

- 81% pour sexe, âge, code de résidence, FINESS, mode d'entrée, mode de sortie, mois de sortie, durée de séjour ;
- 79% pour sexe, âge, code de résidence, FINESS, mois de sortie, durée de séjour ;
- 53% pour sexe, âge, code de résidence, FINESS, mois de sortie ;
- 10% pour sexe, âge, département de résidence, FINESS, mois de sortie.

Appliqués aux seuls patients décédés à l'hôpital (285 000 séjours en 2012), les mêmes critères produisent respectivement les pourcentages suivants : 99%, 99%, 94% et 65%.

TABLEAU N°1

Dénombrement et effectifs des combinaisons pour chaque association de critères, pour l'ensemble des parcours hospitaliers

association de critères n°	nombre de combinaisons distinctes	effectif le plus élevé	proportion de combinaisons avec			
			1 seul patient	2 patients	3 patients	4 et plus
1	1 076 771	2 877 667	9,1%	0,8%	0,4%	89,6%
2	2 056 251	136 133	17,1%	2,5%	1,3%	79,1%
3	2 554 607	11 931	22,2%	2,0%	0,8%	75,0%
4	2 980 174	7 555	25,2%	1,7%	1,0%	72,2%
5	3 014 361	6 259	25,5%	1,6%	1,0%	72,0%
6	3 517 028	6 147	28,9%	2,7%	1,8%	66,6%
7	4 152 165	4 726	31,0%	5,0%	4,0%	60,0%
8	4 450 339	726	32,8%	5,9%	4,7%	56,6%
9	4 552 498	558	33,2%	6,3%	5,1%	55,4%
10	5 324 843	549	40,3%	7,6%	5,3%	46,8%
11	6 099 758	1 092	45,6%	11,0%	7,0%	36,3%
12	7 128 608	463	54,3%	13,9%	8,2%	23,7%
13	8 973 865	112	76,3%	11,3%	4,3%	8,0%
14	9 866 933	46	88,8%	6,6%	1,9%	2,7%
15	9 880 536	46	89,0%	6,5%	1,9%	2,6%

On constate que l'empreinte chronologique à elle seule (association n°1) détermine un nombre très élevé de combinaisons distinctes, dont plus de neuf pour cent (9,1%) sont uniques dans la base de données. Dans plus d'un parcours sur dix (10,4%) les modalités particulières de ce critère réunissent moins de quatre patients.

Si l'on tient compte de l'âge et du sexe en plus de l'empreinte chronologique (association n°2), la proportion de combinaisons uniques est presque doublée (17,1%) et plus de deux parcours sur dix (20,9%) présentent moins de quatre patients par combinaison.

Avec le département d'hospitalisation en complément (association n°3), c'est un quart (25,0%) des parcours qui se ventilent dans des combinaisons d'effectif inférieur à quatre patients, dont la majorité (22,2%) sont même des combinaisons uniques.

Sans le département d'hospitalisation mais avec le lieu de résidence du patient (association n°11), près des deux tiers des parcours (63,7%) présentent des combinaisons de moins de quatre patients, et presque la moitié (45,6%) ont un profil unique.

Avec les mêmes critères auxquels on ajoute le mois de sortie (association n°13) on atteint près de neuf millions de combinaisons distinctes, et moins d'un parcours sur dix (8,0%) relève d'une combinaison d'effectif supérieur à trois patients, plus de trois quarts des parcours (76,3%) ayant même un profil unique.

En tenant compte de l'identification précise de l'établissement d'accueil du premier séjour en plus des critères précédents (association n°14) près de neuf patients sur dix (88,8%) ont un profil de parcours unique ; le mode de sortie (association n°15) n'apporte alors qu'une information discriminante mineure.

La figure n°3 est une représentation graphique de l'ensemble de ces résultats (voir figure n°3).

A titre documentaire nous présentons en annexe le dénombrement détaillé de l'association n°15 (empreinte, âge, sexe, code géographique de résidence du patient, identification de l'établissement initial, mois de sortie, mode de sortie) qui

comporte 46 combinaisons. Nous tenons le détail de toutes les autres associations à la disposition des lecteurs intéressés, sous forme informatique en raison du volume des tableaux.

Le caractère déterminant de l'empreinte chronologique mis en évidence dans le tableau n°1 nous impose de renouveler les calculs en ne sélectionnant que les patients hospitalisés plusieurs fois. En effet cette empreinte est par construction d'autant plus discriminante que le nombre de séjours est élevé. Dans le tableau suivant nous récapitulons donc les mêmes dénombrements, effectués uniquement sur les 2 701 693 parcours comportant deux séjours ou plus, totalisant 8 024 857 séjours sans séances. Ces résultats sont également représentés sous forme graphique (voir figure n°4).

TABLEAU N°2

Dénombrement et effectifs des combinaisons pour chaque association de critères, pour les parcours hospitaliers comportant au moins deux séjours

association de critères n°	nombre de combinaisons distinctes	effectif le plus élevé	proportion de combinaisons avec			
			1 seul patient	2 patients	3 patients	4 et plus
1	1 076 381	17 930	35,9%	3,1%	1,8%	59,3%
2	2 039 871	459	66,7%	9,8%	5,0%	18,6%
3	2 459 804	63	86,0%	7,1%	2,4%	4,5%
4	2 605 522	56	94,3%	3,2%	0,8%	1,7%
5	2 620 547	28	95,1%	2,9%	0,7%	1,3%
6	2 635 929	25	96,0%	2,4%	0,6%	1,0%
7	2 662 933	26	97,6%	1,4%	0,4%	0,6%
8	2 681 176	11	98,6%	1,0%	0,2%	0,1%
9	2 686 377	8	98,9%	0,9%	0,1%	0,1%
10	2 689 290	8	99,1%	0,7%	0,1%	0,0%
11	2 693 225	12	99,4%	0,4%	0,1%	0,1%
12	2 694 482	10	99,5%	0,4%	0,1%	0,0%
13	2 700 495	6	99,9%	0,1%	0,0%	0,0%
14	2 701 077	6	100,0%	0,0%	0,0%	0,0%
15	2 701 086	6	100,0%	0,0%	0,0%	0,0%

L'empreinte chronologique à elle seule (association n°1) détermine pratiquement le même nombre de combinaisons distinctes que précédemment (1 076 391 contre 1 076 771) ce qui était prévisible, par construction. Plus d'un tiers sont uniques (35,9%) dans la base de données, et dans près de quatre parcours sur dix (40,7%) il y a moins de quatre patients par profil distinct de ce critère.

Avec l'âge et du sexe en plus de l'empreinte chronologique (association n°2), la proportion de combinaisons uniques est de plus de deux parcours sur trois (66,7%). En complétant avec le département d'hospitalisation en complément (association n°3), moins d'un parcours sur vingt (4,5 %) se ventile dans des combinaisons d'effectif supérieur à trois patients.

Avec le lieu de résidence du patient à la place du département d'hospitalisation (association n°11), la quasi-totalité des parcours (99,4%) ont un profil unique, et la totalité présentent des combinaisons de moins de quatre patients.

En substituant le seul département de résidence au lieu de résidence détaillé (association n°4) moins d'un parcours sur cinquante (1,7%) se trouve dans une combinaison de plus de trois patients et dix-neuf parcours sur vingt (94,3%) ont un profil unique. On obtient des proportions semblables, voire plus élevées, avec l'identification exacte de l'établissement d'hospitalisation (association n°6 : 1,0% et 96,0% respectivement), ou avec les départements d'hospitalisation ou de

résidence complétés par le mois de sortie (association n°7 : 0,6% et 97,6% ; association n°8 : 0,1% et 98,6%, respectivement).

A titre documentaire nous présentons en annexe le dénombrement détaillé des associations de critères n°9, 10, 13, 14 et 15, pour lesquelles l'effectif de la combinaison la plus fréquente ne dépasse par la dizaine. Nous tenons sous forme informatique le détail de toutes les autres associations à la disposition des lecteurs intéressés.

Discussion

Interprétation des résultats[4]

Les résultats présentés ci-dessus ne laissent aucun doute : même en se limitant aux parcours de patients ne bénéficiant jamais de séances, le pouvoir de ré-identification de la base nationale de données du PMSI est très élevé : l'âge, le sexe, le code postal du patient, le numéro FINESS de l'établissement et le mois de sortie, ajoutés à quelques informations relatives à son parcours hospitalier (durée de séjour, délai inter-séjour), suffisent à ré-identifier à coup sûr neuf patients sur dix, et même la totalité des patients s'ils ont subi plus d'une hospitalisation dans l'année.

Puisqu'il suffit de disposer de quelques informations relativement simples à obtenir pour cibler dans la base anonyme les deux ou trois patients dont le profil correspond au critères sélectionnés, voire pour localiser à coup sûr celui qu'on recherche, si la base de données est accessible à des personnes ou des institutions qui ont un intérêt à connaître le contenu médical du dossier d'un individu, alors ce pouvoir de ré-identification devient un risque.

Or toutes les conditions sont réunies pour que ce risque se manifeste :

- la quasi-totalité des acteurs du PMSI méconnaissent l'existence de ce risque, donc ne prennent aucune des précautions propres à s'en prémunir ;
- la base nationale de données du PMSI est largement diffusée, sous forme de cédéroms, à de nombreuses institutions publiques et entreprises privées ;
- cette diffusion, bien que réglementée, ne fait l'objet d'aucun contrôle (copies des cédéroms en chaîne) ni d'aucune tenue de registre par les services de la CNIL ;
- le fichier de chaînage, dont l'index chronologique est un élément constitutif essentiel bien que non documenté, est distribué de manière quasi-systématique avec le fichier des RSA ;
- les séjours multiples sont les plus vulnérables, or c'est pour des patients avec des séjours multiples qu'on peut redouter le plus la « curiosité » de certains intrus : des employeurs, des banquiers, des assureurs, des héritiers, etc. pour le « patient lambda », auxquels il faut ajouter des « journalistes people » pour le patient célèbre, l'élu local ou le notable du cru ;
- les informations nécessaires à la ré-identification sont aisément disponibles en général, et encore plus aisément pour les intrus évoqués précédemment : plus facilement que d'autres, l'employeur, l'assureur, le banquier, l'héritier, peuvent connaître les dates d'arrêt maladie, donc les durées de séjours et délais inter-séjours exacts ou approchants ;
- la ré-identification dans le PMSI est à la fois permanente et ubiquitaire : la clef de chaînage, diffusée sans précaution particulière avec les cédéroms, est immuable au cours du temps, et identique dans tous les champs du PMSI.

[4] 2015 : ces résultats prennent en compte certains cas dans lesquels l'attaquant potentiel ne connaît qu'une partie de l'information ré-identifiante, par exemple le département de résidence et non pas le code complet. Mais concernant les données de l'empreinte chronologique, ils se fondent sur l'hypothèse d'un attaquant qui connaîtrait de manière certaine le nombre exact d'hospitalisations du patient, leurs durées respectives et les délais entre deux séjours successifs. Cette approche « favorable » à l'attaquant pourrait faire l'objet d'analyses de variantes moins favorables : en supposant par exemple qu'il ne dispose que d'informations approximatives (à quelques jours près) sur les durées et les délais ; ou bien, pour les patients revenus plusieurs fois, qu'il ignore l'existence d'au moins un séjour intermédiaire.

La sous-estimation du risque dans les années 2000

Parmi les points signalés, certains feront sans doute polémique, mais il nous semble nécessaire d'en développer un en particulier. Il s'agit de la sous-estimation du risque réel par les autorités chargées de l'autorisation et des modalités de diffusion[5]. Or ce risque s'est considérablement accru, nous l'avons vu, en raison du dispositif de chaînage et d'indexation chronologique.

En fait, de 2002 à 2010, la CNIL a pris sept délibérations relatives au PMSI. Trois seulement concernent le recueil et la transmission systématiques des données du PMSI aux services de l'État :

- délibération 2005-127 du 14 juin 2005
- délibération 2008-051 du 21 février 2008
- délibération 2009-643 du 26 novembre 2009

Pourtant, durant la même période le contenu des RSS recueillis dans les établissements a été modifié à sept reprises (formats 007, 008, 011, 012, 013, 014 et 015) justifiant la diffusion par le ministère de cinq guides méthodologiques successifs (2004/2 bis, 2006/2bis, 2007/4bis, 2009/5bis et 2010/5bis) et d'un guide modificatif non numéroté en 2008. Quant au contenu des informations individuelles transmises aux services centraux de l'État dans les RSA il a fait l'objet de onze évolutions distinctes au cours de la même période (formats 207, 208, 209, 210, 211, 212, 213, 214, 215, 216 et 217).

Quels que soient le nombre et la teneur des informations nouvelles introduites dans les RSS ou transmises dans les RSA, il aurait été souhaitable d'en analyser la portée et les conséquences afin d'en autoriser respectivement le recueil et la transmission aux services centraux. Or le seul document de présentation générale du dispositif technique du PMSI disponible sur le site internet de l'Agence technique de l'information sur l'hospitalisation (ATIH) ne fournit qu'un exposé très succinct (7 lignes) de la procédure de chaînage des résumés de séjour, introduite en 2002, qui indique ceci :

> *« une procédure de chaînage […] permet de relier entre elles, grâce à un numéro de chaînage anonyme, les différentes hospitalisations d'un même malade. »*

On ne trouve pas trace de délibération de la CNIL relative à cette procédure de chaînage avant 2008.

De surcroît voici ce qu'on lit dans la délibération 2008-051 de la CNIL :

> *« Les fichiers transmis à l'ARH sont anonymisés. Une clé de chaînage permet de lier les séjours ou les prestations non suivis d'hospitalisation (résumés de sortie anonymisés – RSA) avec les fichiers de résumés standardisés de facturation (RSF). »*

Ainsi la CNIL a probablement pu croire que le chaînage est un dispositif permettant de lier anonymement et simplement les données médicales d'un séjour avec les informations de facturation de celui-ci. Autrement dit, ce serait un système qui pose une sorte d'étiquette unique, dont une partie est collée sur le résumé médical et l'autre sur la facture, afin de pouvoir les réunir anonymement. Pour saisir l'invraisemblance de cette interprétation, une précision s'impose : seuls les établissements privés produisent des RSF. En tout état de cause, si l'objectif de la clef de chaînage était de lier la partie médicale du dossier à un résumé de facturation transmis séparément, on devait en limiter l'application aux seuls établissements privés. Or tous les établissements, sans exception, produisent depuis 2002 la fameuse clef de chaînage des résumés anonymisés.

En réalité, ainsi que le précise la documentation de sept lignes citée plus haut et que les auteurs de la délibération de la CNIL n'ont peut-être pas eue en mains, ce ne sont pas les éléments d'une seule et même hospitalisation que le chaînage vise à rassembler, mais toutes les venues et hospitalisations d'un même patient dans n'importe quel établissement français, à quelque date que ce soit.

Un document de trois pages relatif au chaînage, plus détaillé et accessible sur le site internet de l'ATIH, est cité dans les sept lignes de la présentation générale. On peut y lire ceci :

[5] 2015 : rappelons que ces lignes ont été écrites en 2011.

> *« Le principe du chaînage anonyme consiste en la création d'un numéro anonyme commun à toutes les hospitalisations d'un même malade […]. Sur une période donnée, les différents épisodes d'hospitalisation d'un même malade peuvent ainsi être identifiés et liés entre eux. »*

Plus loin, on apprend que :

> *Un numéro anonyme premier est créé à l'échelon de l'établissement […] par un procédé sécurisé dénommé fonction d'occultation des informations nominatives (FOIN) […]*

et que

> *un numéro anonyme second est créé au niveau de l'agence régionale de l'hospitalisation par l'application une seconde fois de FOIN. C'est ce numéro anonyme second qui est utilisable au plan national.*

En réalité, pour être précis, l'application de FOIN est réalisée par la plateforme nationale de l'ATIH (e-PMSI) pour le compte des ARH. Mais le document est muet sur un point important et méconnu car non documenté ailleurs : le chaînage ne se limite pas à l'attribution d'un identifiant unique à tous les enregistrements anonymisés relatifs à un même patient, quelle que soit la date et quel que soit l'établissement d'hospitalisation, puisqu'il calcule aussi et enregistre l'index chronologique, ce délai écoulé entre une date de référence propre au patient et la date de son hospitalisation que nous avons présenté plus haut.

Cette « astuce technique » permet, pour qui en est informé, de calculer le délai entre deux hospitalisations successives et de vérifier l'absence de chevauchement des hospitalisations déclarées. Il s'agit donc d'une information très utile. Mais ce chaînage chronologiquement précis au jour près apporte un élément discriminant essentiel qui permet de lever l'anonymat d'un nombre considérable de patients.

Dès 1996 une étude interne du ministère avait déjà identifié le problème : même débarrassé d'identifiants, de dates de séjour précises et de date de naissance exacte, le résumé de sortie anonymisé pouvait dans de nombreux cas être désanonymisé, notamment en raison de la connaissance de l'établissement d'hospitalisation, du code géographique de domicile du patient, et surtout du mode de sortie, particulièrement le décès. La CNIL avait alors, à juste titre, imposé des règles strictes imposant des transformations de données (âge et code géographique principalement) afin d'obtenir des agrégats de taille suffisante. Il était donc plus que probable, sans même effectuer des études approfondies des données, qu'en connaissant la durée exacte séparant deux hospitalisations successives d'un patient et les durées de chacun des séjours on pouvait être en mesure d'aller « pêcher » à coup sûr dans la base nationale de données un patient déterminé, et donc obtenir la liste exacte des diagnostics pour lesquels il avait été pris en charge.

Le caractère permanent et ubiquitaire de la ré-identification dans le PMSI

On pourrait utiliser les termes « longitudinal » et « transversal » pour signifier la même chose. Quels que soient les termes employés, ce point mérite d'être développé.

Si le pouvoir de ré-identification constitue un risque, c'est en raison des informations médicales du patient ré-identifié auxquelles un intrus pourrait accéder. Certes, mais ce qui aggrave la situation, c'est que l'intrus pourrait aussi tirer parti des caractéristiques « universelles » de la clef de chaînage qu'il récupérerait dans le fichier : muni de cette clef, il la rechercherait dans toutes les bases de données PMSI passées et à venir, et dans les champs MCO, SSR, HAD et PSY du PMSI. S'il la retrouve, c'est qu'il s'agit du même patient.

La meilleure illustration de la gravité de ce constat apparemment anodin est fournie par une situation d'une grande banalité, puisqu'il s'agit de l'accouchement. Observons au préalable que la clef de chaînage du PMSI étant constante, s'il en est de même de la date de référence individuelle qui en est dérivée, alors on peut reconstituer le parcours hospitalier des patients sans se cantonner aux limites annuelles. Or les séjours pour accouchements sont des séjours comme les autres, donc toutes les mères de famille ayant accouché au moins deux fois seraient ré-identifiables, ce qui permettrait à un intrus de récupérer tous les séjours antérieurs et postérieurs, y compris pour d'autres motifs qu'un accouchement. Cela suppose bien entendu que l'intrus connaisse les durées de séjour et délais entre les séjours pour accouchements. Mais précisément, c'est l'un des cas où il est particulièrement simple de les connaître, puisque les dates d'hospitalisation correspondent aux dates de naissance des enfants, et qu'il n'y a rien de moins secret que la date de naissance des

enfants de n'importe quelle mère de famille. Voilà donc chaque année environ 2,2% de la population féminine française dont on peut récupérer la clef de chaînage avec laquelle consulter ensuite les données médicales de ses séjours passés et à venir.

Première recommandation : surveiller l'évolution du risque, même de manière imparfaite

Pour qui souhaite effectivement prendre à temps les mesures propres à limiter les risques de ré-identification des bases nationales de données, il faut surveiller l'évolution de leur pouvoir ré-identifiant au fur et à mesure qu'on les enrichit. Cette anticipation peut être réalisée d'une manière approximative et empirique.

L'effet conjoint du marquage chronologique précis des RSA et de leur enrichissement considérable au fil du temps accroît de manière exponentielle le nombre théorique de combinaisons distinctes que les informations qu'ils comportent peuvent constituer. Même en sous-estimant délibérément le nombre de modalités possibles de chacune de ces informations, on obtient un nombre théorique de « RSA uniques » plusieurs milliers de fois supérieur au nombre réel de RSA dans la base nationale. Intuitivement cela semble sauter aux yeux puisque chaque diagnostic, par exemple, peut prendre une valeur parmi 20 000 codes distincts, de même que chaque acte parmi 10 000 codes environ.

En pratique cependant le véritable risque ne consiste pas à utiliser les informations médicales ou médico-économiques pour ré-identifier un RSA : au contraire, la motivation d'un intrus serait d'obtenir des informations médicales qu'il ignore concernant un patient qu'il connaît. Les traits caractéristiques pouvant permettre la ré-identification du RSA relèvent donc uniquement des informations dites administratives du RSA et du dispositif de datation relative des séjours.

Le tableau ci-dessous présente dans la deuxième colonne le dénombrement réel des modalités de ces informations. Mais ces dernières n'ont pas une distribution uniforme et ne sont pas indépendantes les unes des autres. La troisième colonne fournit donc une valeur minorée du nombre de modalités, afin de calculer par leur produit général une estimation globale du nombre de combinaisons théoriques distinctes. Les valeurs retenues, très basses, reflètent notre désir de ne surtout pas surestimer le nombre de ces combinaisons. Elles sont fixées de manière empirique.

TABLEAU N°3

Surveiller l'évolution du risque

	modalités ou cardinal exact	nombre de modalités minoré
identifiant de l'établissement	2000	50 modalités
sexe du patient	2	2 modalités
Age	2 à 120 ans 0 à 24 mois	20 modalités
durée du séjour	0 à 366 jours	4 modalités
mois de sortie	1 à 12	10 modalités
mode d'entrée	0,6,7 et 8	2 modalités
provenance	1,2,3 et 4	1 modalité
mode de sortie	0,6,7,8 et 9	2 modalités
destination	1,2,3 et 4	1 modalité
nombre de services fréquentés	1 à 99 services	90% : 1 seul 10% : 2
code géographique de résidence	6 500 codes	50
nombre de séjours dans l'année	1 à 50 séjours	75% : 1 seul 25% : 2
délai entre deux séjours	2 à 364 jours	10 modalités (si plus d'un séjour)

Pour les calculs, quatre cas de figure découlent de ce tableau selon que le patient à fréquenté un seul service (90% des cas) ou plusieurs (10% des cas), et qu'il séjourné à l'hôpital une seule fois dans l'année (75% des cas) ou plus d'une fois (25% des cas).

Le résultat confirme notre intuition, l'ensemble des informations totalisant 101,2 millions de combinaisons théoriques distinctes, qui se répartissent ainsi :

- 10,8 millions de combinaisons distinctes pour 67,5% des cas (1 séjour, 1 service) ;
- 72,0 millions pour 22,5% des cas (2 séjours, 1 service) ;
- 2,4 millions pour 7,5% des cas (1 séjour, plusieurs services) ;
- 16,0 millions pour 2,5% des cas (2 séjours, plusieurs services) ;

Certes ce ne sont que des intuitions et des spéculations, et seul le dénombrement exact de la véritable base nationale de données va permettre de vérifier ces chiffres. Mais même de manière approximative et sans aucune donnée encore recueillie, cette méthode aurait permis de prédire qu'avec plus de 100 millions de combinaisons théoriques distinctes on dépassait largement le nombre de patients de la base nationale, et donc que chaque combinaison avait une très forte probabilité de se trouver unique en pratique. C'est d'ailleurs ce rapide calcul qui nous a conduit à décider de réaliser cette étude.

Deuxième recommandation : suspendre la distribution du fichier ANO[6]

Le rôle discriminant de certaines informations dans le pouvoir de ré-identification est évident : la distinction des deux sexes n'a un pouvoir résolutif intrinsèque que de 2, tandis que le jour de sortie dans l'année (s'il existait dans la base de données) a un pouvoir résolutif intrinsèque de 366. Nous avons vu qu'à elle seule, l'empreinte chronologique détermine plus d'un million de combinaisons distinctes. Le pouvoir résolutif intrinsèque de l'empreinte est donc exceptionnellement élevé : même si la distribution des parcours selon chaque modalité de l'empreinte n'est pas uniforme, ce pouvoir résolutif intrinsèque atteint au moins la centaine de milliers.

C'est la raison pour laquelle on doit s'interroger sur la légitimité de diffuser une telle donnée, et sur les solutions éventuelles pour lui en substituer une autre, plus pauvre mais moins discriminante.

De notre point de vue, pour les usages courants des détenteurs de copies de la base nationale sur cédéroms, il n'y a aucune légitimité à disposer de l'empreinte chronologique. La seule raison qui peut expliquer le nombre de demandes à disposer du système de chaînage conjointement à la fourniture des RSA – toutes demandes d'ailleurs validées par la CNIL – est que la description inexacte de son intérêt et les lacunes de la documentation relative à son contenu ont induit les demandeurs en erreur, de même que la CNIL.

Nous recommandons donc de cesser purement et simplement, dès à présent, cette distribution intitulée « fichier ANO » dans les formulaires de l'ATIH[7].

Troisième recommandation : traçabilité des copies

Le principe des copies de cédéroms ne permet pas de suivre à la trace les copies illicites qui, malgré des engagements écrits signés par les détenteurs de copies légales, se multiplient de façon incontrôlée, au vu et au su de tous ceux qui se penchent sur la question.

Il faudrait donc mettre au point un autre système de diffusion, incluant un système de traçage dissuasif. La question est techniquement complexe, car on peut difficilement « marquer » une base de données. En attendant de trouver une réponse technique adaptée, il est indispensable de constituer un registre des copies diffusées[8].

[6] 2015 : le choix opéré finalement dans le projet de loi discuté au Parlement est plus nuancé et plus opérationnel, puisqu'il permet l'usage du fichier pour l'accès contrôlé aux données du PMSI, réservé aux chercheurs habilités. En revanche, le contenu du fichier ANO ne sera pas disponible pour l'accès en mode ouvert.

[7] Rappelons que ces propos datent de 2011. Cette recommandation n'aurait plus lieu d'être en 2015 dans le cadre de l'article 47 du projet de loi de modernisation de notre système de santé qui prévoit un référentiel sur la sécurité du Système national des données de santé (dont le PMSI fait partie). L'option retenue ne sera pas de tracer les copies des bases de données mais d'interdire leur diffusion (et leur copie), ces bases n'étant plus accessibles (aux seules personnes autorisées) qu'en accès à distance (cf. dans ce *Dossier* l'article sur le CASD). Les accès aux données seront tracés et le caractère anonyme des résultats issus de ces traitements sera vérifiable.

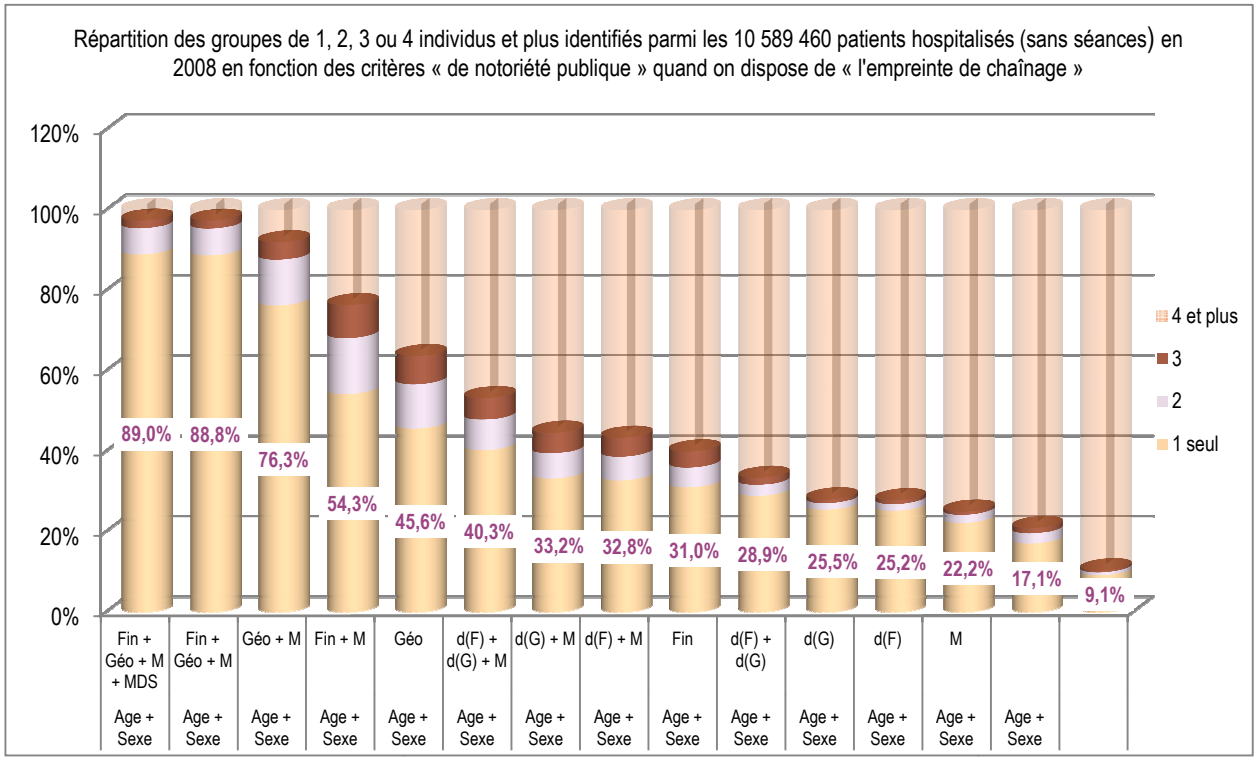
Dernière recommandation : encadrement règlementaire

Le principe qui consistait à publier un arrêté fixant le contenu détaillé du RSS à chaque fois que celui-ci évoluait était un bon principe, qui avait pour avantage de mobiliser en temps et en heure tous les acteurs concernés. S'il avait été respecté, outre le filtre des services du ministère, celui de la CNIL aurait été activé six ans plus tôt. Il faut le rétablir et le faire respecter.

Cela ne règle pas un autre point essentiel : les déclarations à la CNIL doivent être sincères et complètes, et les intitulés qui désignent les champs des fichiers de RSS, de RSA ou de chaînage devraient être explicites et sans ambiguïté, ce qui n'est pas le cas par exemple du champ intitulé « N° de séjour » dans le fichier de chaînage… alors qu'il s'agit précisément de l'index chronologique, c'est-à-dire une datation déguisée.

FIGURE N°3

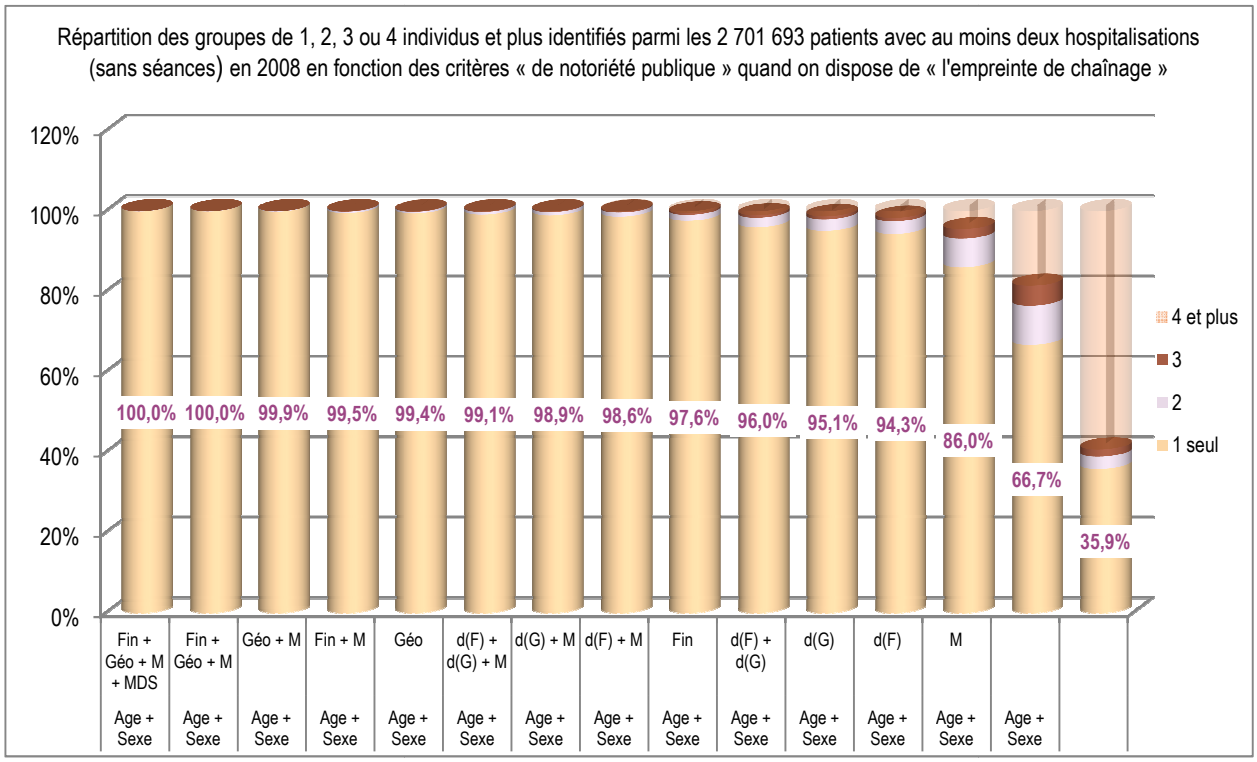
Associations de critères pour l'ensemble des parcours hospitaliers



[8] En mai 2015, l'ATIH a annoncé qu'en accord avec les observations du rapport Bras et du rapport sur les risques de ré-identification dans les bases de données de santé (publié dans ce *Dossier*) indiquant que le mode actuel de communication des données issues du PMSI, fondé sur l'extraction de données et la retranscription sur cédérom, ne présentait pas les garanties de fiabilité nécessaires pour assurer la protection de ces données, elle a décidé d'y mettre fin.

FIGURE N°4

Associations de critères pour l'ensemble des parcours hospitaliers comportant au moins deux séjours



Annexe

Les tableaux qui suivent n'ont qu'une vocation documentaire, et leur absence ne nuit pas à la compréhension de l'exposé. Pour des raisons de volume nous n'avons pas pu placer ici le détail de toutes les associations de critères, que nous tenons à la disposition des lecteurs intéressés sous forme informatique.

Ci-après nous présentons le détail du dénombrement pour les cinq associations de critères qui totalisent moins de 10 patients par combinaison (associations n°9, 10, 13, 14 et 15) pour les parcours comportant au moins deux séjours.

TABLEAU N°4

Dénombrement détaillé des combinaisons observées pour l'association de critères n°15 pour les parcours hospitaliers comportant au moins deux séjours

effectif de la combinaison	nombre de combinaisons ayant cet effectif	effectif total (colonne 1 x colonne 2)	proportion de l'effectif total cumulé
1	2 700 517	2 700 517	100,0%
2	542	1 084	0,0%
3	20	60	0,0%
4	4	16	0,0%
5	2	10	0,0%
6	1	6	0,0%
cumul	**2 701 086**	**2 701 693**	**100,0%**

TABLEAU N°5

Dénombrement détaillé des combinaisons observées pour l'association de critères n°14 pour les parcours hospitaliers comportant au moins deux séjours

effectif de la combinaison	nombre de combinaisons ayant cet effectif	effectif total (colonne 1 x colonne 2)	proportion de l'effectif total cumulé
1	2 700 501	2 700 501	100,0%
2	547	1 094	0,0%
3	22	66	0,0%
4	4	16	0,0%
5	2	10	0,0%
6	1	6	0,0%
cumul	**2 701 077**	**2 701 693**	**100,0%**

TABLEAU N°6

Dénombrement détaillé des combinaisons observées pour l'association de critères n°13 pour les parcours hospitaliers comportant au moins deux séjours

effectif de la combinaison	nombre de combinaisons ayant cet effectif	effectif total (colonne 1 x colonne 2)	proportion de l'effectif total cumulé
1	2 699 363	2 699 363	99,9%
2	1 078	2 156	0,1%
3	46	138	0,0%
4	5	20	0,0%
5	2	10	0,0%
6	1	6	0,0%
cumul	**2 700 495**	**2 701 693**	**100,0%**

TABLEAU N°7

Dénombrement détaillé des combinaisons observées pour l'association de critères n°10 pour les parcours hospitaliers comportant au moins deux séjours

effectif de la combinaison	nombre de combinaisons ayant cet effectif	effectif total (colonne 1 x colonne 2)	proportion de l'effectif total cumulé
1	2 678 514	2 678 514	99,1%
2	9 494	18 988	0,7%
3	1 018	3 054	0,1%
4	203	812	0,0%
5	46	230	0,0%
6	11	66	0,0%
7	3	21	0,0%
8	1	8	0,0%
cumul	**2 689 290**	**2 701 693**	**100,0%**

TABLEAU N°8

Dénombrement détaillé des combinaisons observées pour l'association de critères n°9 pour les parcours hospitaliers comportant au moins deux séjours

effectif de la combinaison	nombre de combinaisons ayant cet effectif	effectif total (colonne 1 x colonne 2)	proportion de l'effectif total cumulé
1	2 673 117	2 673 117	98,9%
2	11 641	23 282	0,9%
3	1 292	3 876	0,1%
4	242	968	0,0%
5	66	330	0,0%
6	14	84	0,0%
7	4	28	0,0%
8	1	8	0,0%
cumul	**2 686 377**	**2 701 693**	**100,0%**

ANNEXE 2 : FOIN : un exemple de système de pseudonymisation sécurisé

Gilles TROUESSIN

Cette contribution vise à donner une description générale des fonctionnalités essentielles pour un système de pseudonymisation (ou système d'occultation) et à décrire de façon détaillée les propriétés de base des fonctions de hachage utilisées. Un procédé sécurisé utilisant le hachage, dénommé Fonction d'Occultation des Informations Nominatives (FOIN) a été mis au point en 1996 par la Caisse Nationale de l'Assurance Maladie des Travailleurs Salariés (CNAMTS). Il permet la mise en œuvre fiable et sécurisée d'une version pseudonymisée des résumés de sortie hospitaliers issus du programme de médicalisation des systèmes d'information (PMSI), le but étant de collecter ces résumés de sortie dans une base nationale et d'y chaîner les informations relatives aux prises en charge successives d'un même patient.

Ce qui suit est une description de la FOIN originelle ; les développements postérieurs éventuels n'ont pas été pris en compte. La description générale des fonctionnalités essentielles de la FOIN originelle (ne permettant pas la moindre implémentation immédiate) est complétée par une description affinée des propriétés de base des fonctions de pseudonymisation (ne permettant pas non plus l'implémentation informatique immédiate mais permettant de lister les précautions, ou exigences de sécurité à prévoir, pour parer à certaines attaques vis-à-vis de ce type de fonctions).

Caractéristiques fonctionnelles de la FOIN

Les caractéristiques fonctionnelles de *FOIN* découlent directement de la demande initiale formulée par le Ministère de la Santé.

Très simplement, cette fonction doit répondre à un critère essentiel : permettre un suivi longitudinal des trajectoires de soins de tous les patients via un numéro d'identification anonyme servant de lien entre tous les comptes rendus de soins d'un même patient. Afin d'assurer le respect de la vie privée des patients, la génération de ce numéro doit se faire via un processus irréversible qui assure l'anonymat du numéro (anonymat en ce sens que le numéro seul ne permet pas de remonter jusqu'à l'identité du patient : il s'agit en ce sens d'une pseudonymisation anonyme). Enfin, des mesures doivent être prises pour assurer la robustesse du procédé ainsi qu'une certaine résistance aux tentatives de ré-identification.

Les caractéristiques fonctionnelles de *FOIN* peuvent être réparties en trois catégories : anonymisation, cryptographie et sécurisation.

RAPPEL DU CONTEXTE LIÉ AU PMSI

Le Programme de Médicalisation des Systèmes d'Information (PMSI) est un système d'analyse de l'activité des établissements de santé dont la finalité est l'allocation des ressources tout en diminuant les inégalités budgétaires. Le PMSI a été expérimenté depuis 1983, et généralisé dans les hôpitaux publics et privés participant au service public par la circulaire du 24 juillet 1989[1] pour l'activité de MCO (Médecine, Chirurgie, Obstétrique). Il a été étendu aux établissements privés par les ordonnances du 24 avril 1996[2]. Son utilisation à des fins budgétaires a été formalisée par la circulaire du 7 décembre 1996. La circulaire du 9 mars 1998[3] a généralisé le PMSI aux établissements publics ayant une activité de soins, de suite et de réadaptation. Une multitude de textes ont été élaborés pour réglementer le fonctionnement du PMSI. Rappelons à titre indicatif la loi du 31 juillet 1991[4], suivie du décret du 27 juillet 1994 ainsi que des arrêtés des 20 septembre 1994, 22 juillet 1996 et 29 juillet 1998 avec, quelques années plus tard, la loi n° 2003-1199 du 18 décembre 2003 de financement de la sécurité sociale pour 2004 (art. 22 à 34) modifiant profondément les modalités de financement des établissements de santé, passant d'une logique de moyens à une logique de produit.

1 Circulaire DH/PMSI n° 303 du 24 juillet 1989 relative à la généralisation du Programme de médicalisation (BOMS n° 89/46), Ministère de l'emploi et de la solidarité, France.

2 Ordonnance n° 96-346 du 24 avril 1996 portant réforme de "l'hospitalisation publique et privée des systèmes d'information et à l'organisation médicale dans les hôpitaux publics".

3 Circulaire n° 153 du 9 mars 1998 relative à la généralisation dans les établissements de santé sous dotation globale et ayant une activité de soins de suite ou de réadaptation d'un recueil de RHS, ministère de l'emploi et de la solidarité, France.

4 Loi n° 91-748 du 31 juillet 1991 portant réforme hospitalière et décret n° 92-329 du 30 mars 1992 relatif au dossier médical et à l'information des personnes accueillies dans les établissements de santé publics et privés.

Dans la pratique, chaque séjour d'un patient donne lieu à un recueil standardisé de données de nature administrative (dates d'entrée et de sortie, date de naissance, nom et prénom) et médicale (diagnostics, actes codés). Les séjours sont ensuite classés selon l'indicateur médico-économique ou « Groupe Homogène de Malades (GHM) ». Les patients d'un GHM donné sont considérés comme ayant mobilisé des ressources de même ampleur. Chaque année une échelle des coûts affecte un coût relatif à chaque GHM. Les données du PMSI des établissements publics sont [dites] « anonymisées », pour pouvoir être transmises semestriellement aux Agences Régionales de Santé (ARS), qui ont succédé aux Agences Régionales de l'Hospitalisation (ARH), et qui les utilisent pour l'allocation budgétaire. Celles des établissements privés sont transmises trimestriellement à l'Assurance Maladie. Plus précisément, tout séjour hospitalier effectué dans la partie court séjour d'un établissement fait l'objet d'un Résumé de Sortie Standardisé (RSS), constitué d'un ou plusieurs Résumés d'Unité Médicale (RUM). Le RUM contient des données (administratives et médicales) concernant le séjour d'un patient dans une unité médicale. À partir des RUM, le Département d'Information Médicale (DIM) construit le fichier des Résumés de Sortie Standardisés (RSS) puis des Résumés de Sortie Anonymes (RSA).

Bien que considérée, depuis son origine comme « fonction d'anonymisation », la fonction dite *« F.O.I.N. »* doit plutôt être vue comme une fonction de génération de pseudonymes anonymes ; des pseudonymes qui restent anonymes tant que leur utilisation ne permet pas, à partir de corrélations diverses et autres appariements nécessaires, d'inférer l'identité exacte de tout patient ainsi rebaptisé provisoirement à travers son propre pseudonyme anonyme : c'est en cela qu'il faut voir la fonction *FOIN* non pas comme une fonction d'anonymisation (irréversible, en théorie) ou de pseudonymisation (réversible, en pratique), mais plutôt comme une fonction d'occultation des éléments identifiant un individu donné (un patient, en l'occurrence).

Anonymisation

L'exigence sous-jacente à l'anonymat dans les bases est la possibilité de pouvoir chaîner tous les compte-rendus de soins d'un même patient via un pseudonyme anonyme. Or, permettre de suivre un patient tout au long de son parcours de soins sans révéler son identité exige :

- de transformer chaque identifiant initial dit « nominatif » en un identifiant final dit « anonymisé » que l'on appellera pseudonyme de manière irréversible,
- d'associer toujours-et-partout à un identifiant initial un seul et même pseudonyme, quel que soit l'environnement (transformation sans possibilité de générer des doublons),
- que chaque pseudonyme ne soit associé qu'à un seul et même identifiant initial, quel que soit l'environnement (transformation sans possibilité de produire des « collisions »).

Afin d'y parvenir techniquement, il faut :

- trouver les traits discriminatoires, disponibles pour tous les patients (NIR et sexe, par exemple) et qui peuvent être mis à disposition de la fonction de pseudonymisation,
- s'assurer que ces traits sont suffisamment pérennes et universels pour éviter les doublons dans la base. En effet, si les traits représentatifs d'une même personne changent au cours du temps ou d'une base à l'autre, alors une même personne est susceptible d'être représentée par deux pseudonymes dans la base anonyme.

On peut donc désormais définir plus précisément les caractéristiques techniques auxquelles doit répondre le processus de génération des pseudonymes. Ce processus doit être à la fois universel, pérenne, fiable et facile à manipuler . Cela implique :

- d'obtenir des pseudonymes dont l'espérance de vie est de plusieurs années. Cela nécessite de faire appel à une fonction cryptographique récente et reconnue comme robuste,
- d'obtenir des pseudonymes générés à partir du NIR complété de variables discriminantes entre les différents membres d'une même famille (dans la décennie 90, seul le NIR assuré et non pas le NIR individu était utilisé) ; d'où la nécessité de le compléter avec le champ « rang gémellaire » (qui a finalement été remplacé par le champ « sexe »), complété par le champ « date de naissance » (ce champ permet un premier tri au sein d'un même NIR) soit NIR+DateNaiss+sexe,

En résumé, pour assurer la possibilité d'un chaînage anonyme, il faut créer une fonction robuste irréversible prenant en argument la chaîne de caractères NIR+DateNaiss+Sexe (chaîne considérée comme unique pour chaque individu) avec des garanties sur l'impossibilité de phénomènes de doublonnage ou de collision.

Cryptographie

L'enjeu cryptographique de ce projet réside dans la création d'une fonction d'anonymisation irréversible, c'est-à-dire qu'il doit être impossible de remonter à l'identité du patient à partir du seul pseudonyme. Les besoins associés à cet objectif peuvent se décliner en deux points :

- disposer d'une transformation, depuis tout identifiant initial (nominatif) vers le pseudonyme correspondant;
- ne pas disposer de transformation-retour, depuis un pseudonyme vers l'identifiant initial (nominatif) correspondant (i.e. il n'existe pas de transformation réciproque qui permettrait une désanonymisation)

Il faudra donc chercher parmi les fonctions cryptographiques dites « Fonctions à Sens Unique (FSU) » (ou *One Way Functions* (OWF), en anglais). Plus techniquement et pratiquement, afin d'assurer une irréversibilité « totale » et une utilisation courante, il faudra :

- disposer d'une fonction pour laquelle les attaques les plus connues ne sont pas exploitables, afin de garantir une durée de vie de plusieurs années au système à déployer,
- disposer d'une fonction dont les résultats sont standardisés de manière à faciliter son implémentation et, surtout, son utilisation sur les différents systèmes d'information concernés.

Cela consiste à retenir la sous-famille des fonctions cryptographiques dites « fonctions de hachage » ou « fonction de condensation » (ou « One Way Hash Functions », en anglais), telles que MDC, MD4, MD5 ou le récent SHA (Secure Hash Algorithm), dans les années 1990, conforme au SHS (*Secure Hash Standard*). Après diverses investigations menées dans le milieu des années 90 par le Centre d'Études des Sécurités du Systèmes d'Information (CESSI) de la CNAM-TS en coopération étroite avec la Commission Nationale de l'Informatique et des Libertés (CNIL) et après expertises demandées auprès du Service Central de la Sécurité des Systèmes d'Information (SCSSI), devenu aujourd'hui l'Agence Nationale de la Sécurité des Systèmes d'Information (ANSSI), il a été convenu d'adopter l'algorithme SHA1. Ce dernier, tout récent successeur du SHA (ou SHA0) est arrivé en fin de période probatoire et bénéficie ainsi d'améliorations appréciables et d'une expertise au niveau mondial. Il a donc été choisi d'implémenter SHA1 (Encadré 2) qui présente les particularités suivantes :

- taux de collision extrêmement faible et considéré nul (considérant que, grâce à SHA1, deux entrées distinctes ne génèrent que des résultats distincts),
- taux de doublon nul (considérant que pour un patient donné une seule entrée est possible (NIR+DateNaiss.+Sexe), alors il est impossible de générer des doublons pour ce patient),
- très bon effet avalanche[5].

Les exigences en cryptographie prennent en considération toutes les exigences-en-sécurité qui seront détaillée plus loin. En particulier, la possibilité de pouvoir choisir la valeur d'initialisation de la fonction de hachage SHA1 offre les possibilités de régulation de l'utilisation de SHA1.

Sécurisation

Comme nous l'avons vu aux travers des différentes propriétés de la fonction ainsi sélectionnée, la FOIN garantit l'impossibilité d'obtenir l'identité d'un patient à partir de son numéro d'anonymat. Cependant, puisque le sens identité vers anonymat est possible, rien d'empêche la création d'une table de correspondance qui mettrait en relation le NIR+DateNaiss+Sexe d'une personne et son numéro d'anonymat, autorisant l'attaquant à la retrouver dans la base anonyme. Il y a donc un besoin en sécurisation en ce sens.

La première condition de sécurisation est d'empêcher l'utilisation de FOIN par des personnes non-autorisées, cela implique :

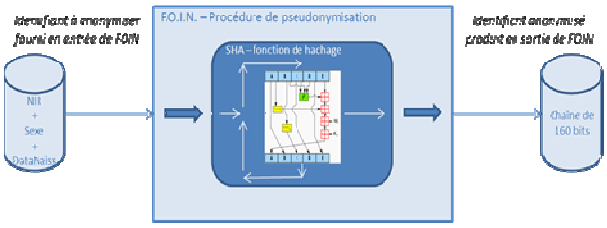
[5] Pour éviter toute erreur de re-saisie lors de manipulations humaines, même deux entrées considérées très proches génèreront des résultats (pseudonymes) très différents : il importe que la qualité des saisies initiales soit parfaitement assurée.

- d'une part de mettre en place une sécurisation de l'usage de la fonction grâce un système de « clé » secrète nécessaire au lancement de la fonction,
- et d'autre part de définir une politique cohérente d'utilisation de cette procédure via une divulgation maitrisée de cette « clé » secrète.

La « clé » secrète choisie pour empêcher le lancement de FOIN, est en réalité une valeur secrète. Il s'agit de la valeur d'initialisation de la fonction de hachage. Chaque valeur d'initialisation correspond à une fonction. Deux fonctions avec des valeurs d'initialisation différentes fournissent des résultats différents à la même entrée. Alors en conservant comme secret national une valeur d'initialisation spécifique, on rend impossible l'accès à la fonction de hachage et donc à la procédure de pseudonymisation qui en découle.

ENCADRE 2 - PRINCIPE GÉNÉRAL DE LA FONCTION *FOIN* ET DE L'ALGORITHME SHA1 POUR *FOIN*

Principe général de la fonction *FOIN* en tant que fonction de pseudonymisation

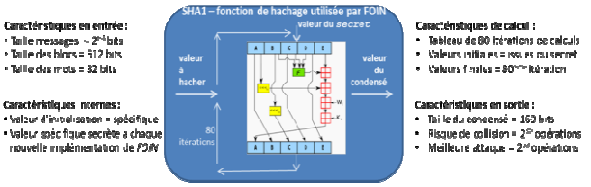


La procédure *FOIN* « d'anonymisation » (ou procédure de "pseudonymisation" si appliquée identiquement toujours-et-partout), prend en entrée des chaînes de caractères qui doivent être représentatives (sans équivoque) de l'identité de chaque patient, ici : NIR + sexe + DateDeNaissance.

Elle fait appel, en interne, à une fonction de hachage (de la famille SHA) pour transformer les champs fournis en entrée (ici : NIR + sexe + DateDeNaissance) en une chaîne 160 bits (qui peuvent être recodés en hexadécimal).

La procédure *FOIN* fournit alors, en sortie, pour chaque entrée de type NIR + sexe + DateDeNaissance la valeur de hachage (ou le condensé) unique ; ici : le pseudonyme *FOIN*.

Principe général de l'algorithme SHA1 pour FOIN en tant que fonction irréversible



La fonction de type SHA (ici : SHA1), utilisée par FOIN :

- prend en entrée chacune des valeurs à anonymiser (dans le cas du PMSI : NIR + sexe + DateDeNaissance), complétée par l'ajout de "0" (bourrage) jusqu'à obtenir, pour chaque message fourni en entrée, une taille égale au prochain multiple de 512 bits (taille des blocs) ;

- puis elle est initialisée avec une valeur de démarrage fournie de façon secrète (dans le cas de l'utilisation sécurisée pour les besoins de la procédure FOIN) ;

- et elle produit, après la 80ème itération (de décalages de bits et de sommes inter-blocs et autres fonctions binaires, etc.), le résultat du hachage, sous la forme d'une chaîne de 160 bits (que l'on peut coder de manière ad hoc pour en raccourcir la taille : moins de 20 octets, par exemple).

La suite consiste à s'assurer que la fonction FOIN ne pourra être exploitée que dans les conditions prévues dans le cadre du « PMSI anonymisé » ou du « SNIIRAM anonymisé » ; ce qui signifie :

- prévoir un contrôle d'accès nominatif pour limiter son utilisation au médecin responsable du DIM (en établissement de soin) ou de ses délégataires (si cela est prévu) ; mais compte tenu de la disparité des organisations internes et des moyens informatiques déployés localement, d'autres précautions de sécurité sont nécessaires ;
- prévoir de sécuriser le lancement de la fonction en utilisant une valeur d'initialisation secrète (ou un secret national) générée et stockée de façon sécurisée car devant être diffusée et partagée entre près de mille établissements de santé participant au PMSI.

Il faut alors définir une stratégie d'approche pour protéger la valeur d'initialisation de la fonction *FOIN*. Il a été choisi :

- d'appliquer une fragmentation lors de la génération de la valeur secrète : le choix de chaque fragment du secret initial est confié à chacune des autorités nationales représentatives et parties prenantes dans la mise en place du « PMSI anonymisé » : par exemple un représentant dûment désigné de l'Assurance Maladie, un représentant dûment désigné du Ministère de la Santé et un représentant national des établissements de santé publics et/ou privés ;
- d'appliquer une technique de fragmentation très spécifique lors de la reconstruction du secret national : puisque ce secret doit être le plus pérenne possible et ne doit pas pouvoir être trop aisément compromis, il est nécessaire de le stocker de façon très sécurisée avant et après toute utilisation par la procédure FOIN (ou plus exactement par la fonction SHA1). Ainsi, il a été choisi d'appliquer la technique dite du « schéma à seuil de A.Shamir » qui consiste à reconstituer une valeur secrète à partir d'un certain nombre, appelé seuil S, de ces différentes projections mathématiques (appelées ses images), choisies parmi un nombre total d'images préalablement générées (appelé total T) : ainsi, il existe T images de la valeur secrète et il nécessaire d'avoir au moins S de ces valeurs pour remonter jusqu'à la valeur du secret, il existe donc une redondance de T - S images.

Les exigences de sécurité consistent donc à mettre en œuvre la méthode du schéma à seuil de A. Shamir de manière à combiner des solutions à base de contrôle d'accès (une image accessible sur autorisation du responsable du DIM), et aussi de contrôle d'utilisation (une image fournie, à tous les établissements, en mode externe : une disquette (à l'époque), une clé USB (désormais)), et enfin de contrôle d'usage (une image stockée dans le code exécutable de *FOIN*, code exécutable fourni à tous les établissements).

Synthèse conceptuelle de la procédure FOIN

Récapitulation des concepts de base de FOIN

- **Transformation déterministe** : tout identifiant nominatif initial doit toujours-et-partout être transformé de façon à obtenir le même identifiant anonymisé final : que l'on appelle donc « pseudonyme » ;
- **Transformation injective** : un pseudonyme ne peut correspondre qu'à ce seul et unique identifiant nominatif initial ;
- **Faible taux de collision** : avec un condensat résultant de l'application d'une fonction de hachage d'une longueur de 160 bits (telle que SHA1), FOIN, transformation injective, présente un taux de collision infime et négligeable ;
- **Faible taux de doublon** : en s'assurant que les données fournies en entrée sont correctement saisies : puisque la transformation est déterministe, il ne peut pas y avoir de doublons (sauf en cas d'erreur de saisie) ;
- **Contrôle d'accès** : il repose sur la politique d'autorisation pour accéder à l'environnement de FOIN ;
- **Contrôle d'utilisation** : il faut posséder une des T images du secret pour exécuter la fonction FOIN ;
- **Contrôle d'usage** : il faut posséder une autre des T images du secret pour pouvoir lancer SHA1 ;
- **Contrôle d'intégrité** : le code exécutable contient une autre des images pour reconstituer le secret.

Récapitulation des propriétés de base correctement satisfaites par FOIN :

- La transformation faite par FOIN est assimilée à de l'anonymisation car ne traitant que des identifiants nominatifs en entrée pour fournir des identifiants totalement dé-identifiés ou « anonymisés » en sortie ; toutefois, puisqu'il est considéré que cette transformation sera toujours-et-partout la même, alors il peut être considéré que FOIN est, de fait, une fonction de pseudonymisation sécurisée ;
- La transformation effectuée par FOIN est bien irréversible car robuste aux réversions directes (pas de transformation inverse possible), comme aux réversions indirectes (pas de constitution de table de correspondance « identifiant nominatif/identifiant anonymisé » possible pour tout utilisateur autorisé du système).
- **En revanche, la généralisation à grande échelle de ce type de système de pseudonymisation fait courir des risques de mise en correspondance par des techniques d'inférence plus ou moins sophistiquées : autrement dit, en croisant des informations recueillies en dehors du système d'information informatisé, il est possible de « reconnaître » un individu (par recoupement de traits symptomatiques) ; cela fait que, pseudonymisation sécurisée ou pas, ce ne sont pas les nouveaux identifiants qui permettent une ré-identification, mais**

l'accumulation de traits significatifs, tous reliés à un même individu par le biais d'un même pseudonyme anonymisé ;

- Les résultats des transformations sont manipulables à l'échelle individuelle lorsqu'ils sont fournis par l'application de FOIN à base de SHA1, à partir d'identifiants nominatifs initiaux. En effet, les identifiants anonymisés finals sont codés sur 160 bits (sur 20 octets, par codage en octal), sinon sur moins de 20 signets (si codage dans un alphabet étendu), ce qui est « humainement pratique ».

Caractérisation des processus de pseudonymisation (générique)

Cette seconde partie consiste à analyser les éventuels besoins/objectifs/exigences à combler car intrinsèquement manquant au niveau des propres performances de la fonction *FOIN*.

Comme vu au cours de la partie précédente, tous les aspects d'ordre cryptographique (reposant sur l'utilisation sécurisée de SHA1) ont été traités. Toutes les mesures relevant de la sécurité d'utilisation et d'usage ont également été exploitées au maximum de leur potentiel. En revanche, certains aspects de l'ordre de « l'anonymisation » – tels que les risques de reconstruction de tables de correspondances par des utilisateurs internes habilités à accéder et à exécuter la procédure *FOIN* – n'ont pas tous été pleinement couverts.

Expression de besoins fonctionnels supplémentaires pour la pseudonymisation

- le premier besoin-en-anonymisation-supplémentaire consiste à interdire toute constitution de table de correspondance qui permettrait de mettre en relation chaque identifiant nominatif initial avec son identifiant anonymisé final correspondant ;
- le deuxième besoin-en-anonymisation-supplémentaire consiste à empêcher toute ré-identification de patient, via la découverte de son identifiant nominatif initial par exemple, par des logiques d'inférence qualifiées de « simples » (déduction, induction) ;
- le troisième besoin-en-anonymisation- supplémentaire consiste à prévenir toute ré-identification de patient, via la découverte de son identifiant nominatif (initial) ou de certains de ces traits significatifs discriminants, par des logiques d'inférence qualifiées de « complexes » (abduction, adduction).

Identification d'objectifs fonctionnels supplémentaires pour la pseudonymisation

- le premier objectif-en-anonymisation-supplémentaire consiste à définir et mettre en place une étape intermédiaire entre le système d'information manipulant les identifiants nominatifs initiaux et le système d'information manipulant les identifiants anonymisés finals. Ceci a pour but de rendre plus difficile la constitution d'une table de correspondance.
- Le deuxième objectif-en-anonymisation-supplémentaire consiste à interdire toute utilisation conjointe, par une seule et même entité, des systèmes d'information nominatif et anonymisés. Ceci a pour but d'interdire toute ré-identification de patient via la découverte de son identifiant nominatif initial ; par exemple, par des logiques d'inférence qualifiées de « simples » (cf. inférence par déduction ou induction) ;
- Le troisième objectif-en-anonymisation-supplémentaire consiste à interdire tout rapprochement entre, d'un côté, des informations accessibles librement et légalement et, d'un autre côté, tout système d'information même anonymisé véhiculant ou manipulant les mêmes informations (cela comprend certains traits discriminants comme la date et le lieu de naissance ou la date et le lieu d'hospitalisation).Ceci a pour but d'interdire toute ré-identification de patient via la découverte de son identifiant nominatif initial ou de certains de ses traits significatifs discriminants par des logiques d'inférence qualifiées de « complexes » (cf. inférence par abduction ou adduction, cette fois).

Sélection d'exigences fonctionnelles supplémentaires pour la pseudonymisation

- La première exigence-en-anonymisation-supplémentaire consiste à définir et mettre en place un deuxième niveau de pseudonymisation de type FOIN, de façon à établir une rupture intermédiaire entre les identifiants initiaux nominatifs et les identifiants finals anonymisés, à travers la ré-anonymisation (2nde anonymisation) des identifiants pré-anonymisés (1ère anonymisation).
- Cela implique de mettre en place deux procédures FOIN (une en amont, en sortie de DIM : FOIN1; et une en aval, en entrée des bases nationales et avant intégration dans le SNIIR-AM : FOIN2, Graphique 1). Ceci qui rend significativement plus difficile l'établissement d'une table de correspondance car cela exigerait une collusion de deux intrus (l'un au niveau de FOIN1 et l'autre au niveau FOIN2) pour mettre en correspondance leurs deux tables de passage, fabriquées de façon illicite, pouvant entrainer les mises en correspondances suivantes :
  - > correspondance « identifiant initial fourni en entrée de $FOIN_1$ » et « identifiant intermédiaire obtenu par $FOIN_1$ » ;
  - > correspondance « identifiant intermédiaire réinjecté dans $FOIN_2$ » et « identifiant final obtenu par $FOIN_2$ ».

$FOIN_1$ et $FOIN_2$ doivent être compatibles afin que les sorties de la première procédure ($FOIN_1$) puissent être injectées en entrée de la seconde procédure ($FOIN_2$).

En revanche ces deux fonctions $FOIN_1$ et $FOIN_2$ doivent avoir des secrets d'initialisation différents (plus exactement, des secrets d'initialisation des appels de la fonction SHA1 différents) pour être implémentées par des systèmes/processus/utilisateurs différents (responsables du DIM, pour $FOIN_1$; services nationaux de l'Assurance Maladie ou de la Direction des Hôpitaux ou de la Direction Générale de Santé ou de représentations nationales du monde hospitalier, pour $FOIN_2$) ; avec, par conséquent, chacune, leur propre secret dédié ($Secret_1$ pour $FOIN_1$, $Secret_2$ pour $FOIN_2$).

Une procédure de pseudonymisation utilisant deux fonctions *FOIN* doit prévoir de protéger les deux secrets nationaux associés par la même technique dite du « schéma à seuil de A.Shamir » avec le même algorithme SHA1 mais avec deux instances différentes :

- > $FOIN_1$ et $FOIN_2$ doivent alors être différenciées par les valeurs du $Secret_1$ (pour $FOIN_1$) et du $Secret_2$ (pour $FOIN_2$;
- > mais on peut choisir de prendre les mêmes valeurs de seuils ($S_1 = S_2$) ou pas ($S_1 \neq S_2$), avec les mêmes choix de niveaux de redondance ($T_1 - S_1 = T_2 - S_2$ ou $T_1 - S_1 \neq T_2 - S_2$).

- La deuxième exigence-en-anonymisation-supplémentaire consiste à définir précisément les usages (i.e., "Pour faire quoi au juste ?"), et donc les autorisations d'utilisation (i.e., "Qui autorise qui à faire quoi ?"), et les droits d'accès des utilisateurs (i.e., "Qui est habilité à accéder à quoi ?"), pour les « systèmes utilisant des identifiants nominatifs initiaux » ainsi que pour les « systèmes utilisant des identifiants anonymisés finals ».

Cela permet d'interdire formellement d'utiliser conjointement, par un même processus (système d'information, applicatif-métier, utilisateurs ayant double-habilitation), un système d'information nominatif manipulant les identifiants nominatifs initiaux et un système d'information anonymisé manipulant les identifiants anonymisés finals.L'objectif est d'interdire toute ré-identification de patient, *via* la découverte de son identifiant nominatif initial, par exemple par des logiques d'inférence qualifiées simples (déduction, induction (cf. précédemment)).

En résumé : C'est à travers une *exigence-en-sécurité* que la solution pour satisfaire à une *exigence-en-anonymisation* peut être trouvée ; via une politique d'autorisation appropriée (de type « Muraille de Chine » entre différentes catégories de systèmes) et donc par la mise en place d'une politique de contrôle d'accès *ad hoc*.

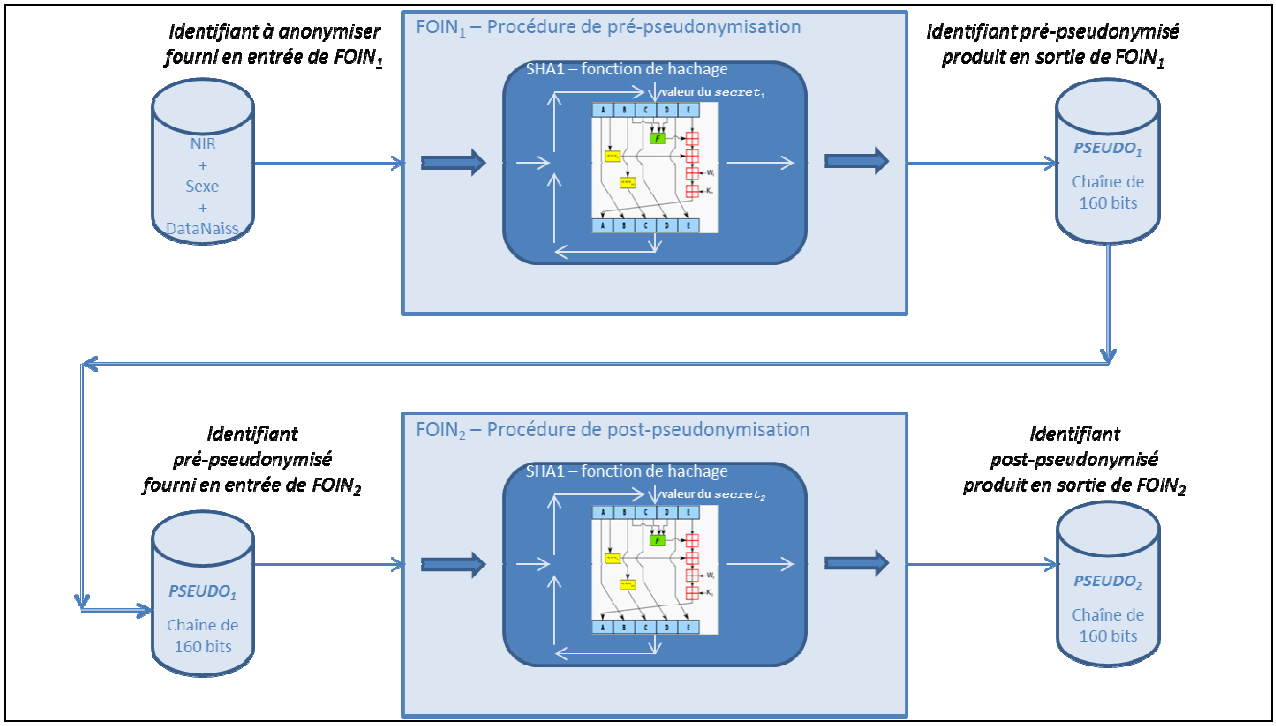
- La troisième exigence-en-anonymisation-supplémentaire consiste à généraliser la ré-anonymisation en cascade (ou, pour être exact, la « re-pseudonymisation » en cascade), en tant que procédure artificielle pour interdire tout rapprochement possible entre, d'un côté des informations provenant de sources externes aux systèmes d'information, et accessibles librement et légalement (cf. abduction et adduction), et de l'autre tout système d'information même anonymisé véhiculant ou manipulant les mêmes informations par ailleurs publiquement connues (tels que certains traits discriminant de certains patients).

Ceci a donc bien pour but d'interdire toute ré-identification de patient, via la découverte de son identifiant nominatif initial ou de certains de ses traits significatifs discriminants, par des logiques d'inférence qualifiées de « complexes » ou de « sophistiquées » (abduction, adduction).

Pour prévenir ce type d'attaque, il convient de cacher au mieux les pseudonymes anonymisés. La bonne politique semble de ne dévoiler les pseudonymes de la base qu'en cas de nécessité absolue. Une première manière de procéder est de remplacer le numéro d'anonymat par un numéro d'ordre lors de la divulgation de données (la table de correspondance pouvant être conservée en lieu sûr pour remonter, en cas de nécessité, jusqu'au numéro d'anonymat). Il est aussi envisageable d'appliquer une fois de plus la procédure FOIN, avec un nouveau secret, différent de ceux de $FOIN_1$ et $FOIN_2$.

GRAPHIQUE 1

Principe général d'une procédure globale de pseudonymisation permettant de compenser les besoins / objectifs / exigences non traités par l'utilisation d'une simple procédure *FOIN*



*à travers l'utilisation d'une double-procédure *FOIN* : notamment pour supprimer le risque de reconstitution, sans collusion, d'une table de correspondance « identifiants nominatifs / identifiants anonymisés ».
Note : Ce graphique schématise l'utilisation qui a été faite de la procédure FOIN, dans le cadre du PMSI : avec (dans sa partie haute) une première mise en œuvre de *FOIN*, avec $FOIN_1$, par les DIM, en sorties des établissements de santé (participant au PMSI anonymisé) ; puis (dans sa partie basse) une seconde mise en œuvre de *FOIN*, avec $FOIN_2$, en entrée immédiate du système national et, donc, en amont du (désormais) système SNIIR-AM.

Remerciements

La Drees remercie tous les auteurs et contributeurs de ce dossier qui, avec la collaboration de Marie Cavillon, consultante, ont permis de rassembler les connaissances et les pistes de réflexion pour mieux comprendre le tournant que représente l'article 47 du projet de loi de modernisation de notre système de santé.

DOSSIERS SOLIDARITÉ ET SANTÉ
Directeur de la publication : Franck von Lennep
ISSN : 1958-587X