CRÉATION DE FICHIERS ANONYMISÉS À PARTIR D'UNE BASE MÉDICO-ADMINISTRATIVE (LE PMSI) : UN EXEMPLE PRATIQUE DE MISE EN ŒUVRE DES MÉTHODES DE PROTECTION DES FICHIERS DE DONNÉES INDIVIDUELLES

Noémie JESS[1] (*), Maxime BERGEAT[2] (**) et Françoise DUPONT[3](**)

(*) Drees[4], Sous-direction Observation de la santé et de l'Assurance maladie
(**) Insee, Direction de la méthodologie et de la coordination statistique et internationale

Résumé

À la suite de la remise du rapport Bras sur la gouvernance et l'ouverture des données de santé, la ministre de la santé Marisol Touraine a demandé au directeur de la Drees de mener une expertise technique sur le risque de ré-identification des individus dans les bases médico-administratives. Ces travaux ont alimenté la Commission « *open data* en santé » lancée à l'automne 2013 pour répondre à la demande croissante de l'ouverture des données de santé.

Cette expertise a été menée par un groupe de travail animé par A. Loth et constitué de personnalités qualifiées et de représentants des organismes producteurs et/ou utilisateurs de données de santé. Le groupe a lancé un test d'anonymisation sur les données hospitalières exhaustives du PMSI afin de réfléchir à la démarche de protection permettant de construire des fichiers diffusables en *open data*. Cet article retrace le test mené à l'aide du logiciel µ-argus.

Après un retour rapide sur le contexte dans lequel s'est déroulé le test d'anonymisation, les variables quasi-identifiantes et l'information sensible à protéger, ainsi que les risques de ré-identification, sont présentés. La méthode d'anonymisation utilisée dans ce test (regroupements de modalités) est ensuite discutée. La méthode choisie est non perturbatrice, c'est-à-dire qu'elle ne modifie pas la distribution des données initiales mais en diminue le niveau de détail. Le choix du regroupement des modalités découle de la volonté de conserver le caractère exhaustif de la base et, usage commun pour la diffusion des données de la statistique publique, d'éviter de complexifier voire de biaiser les analyses. Les critères de protection retenus par le groupe de travail (k-anonymat, l-diversité) sont ensuite définis puis illustrés par deux exemples de jeux de données construits. Enfin, des perspectives générales sont dressées concernant la démarche d'anonymisation et le difficile arbitrage entre le degré de protection et l'utilité finale du fichier.

Mots-clés

Risque de ré-identification, protection des données individuelles, k-anonymat, l-diversité, ouverture des données de santé

[1] noemie.jess@sante.gouv.fr
[2] maxime.bergeat@insee.fr
[3] francoise.dupont@insee.fr, conseillère scientifique au CASD lors des travaux.
[4] Direction de la recherche, des études, de l'évaluation et des statistiques du ministère de la santé.

Abstract

The French Health minister asked F. von Lennep, head of the statistical department of the Ministry of Health and Solidarity, to investigate safety and risk re-identification in individual health database. A taskforce made up of experts, health data users and producers worked on health data specifics and anonymization techniques. In order to get a practical perspective on building *open data* files, a test on real data was carried out on the comprehensive medical-administrative file containing hospital stays.

This paper presents this anonymization test. After underlining the test framework, anonymization techniques are discussed. Examples of derived microdata files that reached the anonymization criteria applied in this test (*l*-diversity and *k*-anonymity) are then detailed. Finally, we lay out general prospects on the anonymization process and the complicated balance between disclosure risk and data utility.

Keyword

Disclosure risk, personal data protection, k-anonymity, l-diversity, open health data

Introduction

L'ouverture et le partage des données publiques, en d'autres termes l'*open data*, est un enjeu majeur pour la transparence démocratique, l'innovation ou encore la croissance économique. Les institutions et administrations publiques, à la faveur de la modernisation de leur action, sont encouragées dans cette voie [1] [2].

La diffusion et la libre réutilisation des données publiques sont encadrées par le principe de protection des données personnelles et de respect de la confidentialité de la vie privée. Les cadres juridiques français (loi du 7 juin 1951 sur le secret statistique, loi informatique et liberté du 6 janvier 1978 et loi CADA sur l'accès aux documents administratifs) et européen [3] définissent les règles de traitement des données à caractère personnel[5]. Pour mettre à disposition des données individuelles, tout en protégeant la vie privée des individus et le secret des affaires des entreprises, il est nécessaire d'anonymiser les données au préalable.

En matière de données de santé, les débats de la Commission « *open data* en santé » installée par la ministre à l'automne 2013 ont alimenté la réflexion sur la réforme de l'accès aux données de santé. L'ouverture des données de santé fait partie du projet de loi de santé qui sera discuté au Parlement en avril 2015.

L'objet de l'article est de présenter le test d'anonymisation des données hospitalières exhaustives du PMSI-MCO[6]. Ce test a été réalisé dans le cadre des travaux du groupe de travail sur les risques de ré-identification dans les bases médico-administratives destinés à alimenter la réflexion de la Commission « *open data* en santé ». Il a pour but de proposer un éclairage concernant la construction de jeux de données anonymes pouvant être diffusés en *open data*.

1. Contexte

1.1. Ouverture des données de santé

À la suite de la remise du rapport Bras sur la gouvernance et l'ouverture des données de santé en octobre 2013 [4], la ministre de la santé et des affaires sociales Marisol Touraine a créé la Commission « *open data* en santé » le 21 novembre 2013. Cette Commission, animée par F. von Lennep, directeur de la Drees, et P. Burnel, délégué à la stratégie des systèmes d'information de santé au Ministère de la santé et des affaires sociales, s'est réunie de novembre 2013 à mai 2014. Elle avait pour mission de débattre, « dans un cadre pluraliste associant les parties prenantes, des enjeux et des propositions en matière d'accès aux données de santé » [5]. Elle a rendu ses conclusions sous la forme d'un rapport en juillet 2014 [5].

Conjointement, la ministre de la santé a demandé au directeur de la Drees de mener une expertise technique sur le risque de ré-identification des individus dans les bases médico-administratives. Ces travaux ont naturellement alimenté la Commission « *open data* en santé ».

La réforme de l'accès aux données de santé fait partie du projet de loi de santé, présenté en conseil des Ministres le 15 octobre 2014 et qui sera discuté au Parlement en avril 2015 [6]. L'article 47 de ce projet de loi contient les propositions pour « la création d'un accès ouvert aux données de santé », avec la création d'un système national des données de santé et des mesures (notamment concernant l'utilisation du NIR et les appariements) facilitant l'accès aux données par les chercheurs.

---

[5] Une donnée à caractère personnel est, d'après la législation française, *« une 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 tout autre personne »*, article 2 de la loi informatique et liberté, http://www.cnil.fr/documentation/textes-fondateurs/loi78-17/ .

[6] Programme de médicalisation des systèmes d'information, sur le champ des courts séjours, *i.e.* dans les domaines de la Médecine, la chirurgie et l'obstétrique (MCO).

1.2. Groupe de travail sur le risque de ré-identification dans les bases de données médico-administratives

Le groupe de travail sur les risques de ré-identification dans les données de santé était constitué de personnalités qualifiées dans différentes disciplines (informatique, statistique, confidentialité, diffusion de données, santé, épidémiologie) et de représentants d'organismes producteurs et/ou utilisateurs de données de santé (Atih, Cnamts, Inserm-CépiDc, Drees). La mission confiée au groupe de travail était d'identifier et d'évaluer les risques de ré-identification dans les bases médico-administratives, d'établir une ligne de démarcation entre les jeux de données anonymes et ceux présentant un risque de ré-identification et de faire des recommandations pour élargir l'offre et le volume des données en accès libre [7].

Au fil des échanges et des discussions sur l'état de l'art des techniques et des critères d'anonymisation, une mise en œuvre concrète sur les données est apparue nécessaire pour affiner la réflexion sur les moyens d'élargir l'offre des données en *open data*.

Le test a été réalisé 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 sur le serveur du Centre d'accès sécurisé distant aux données (CASD). Les travaux ont donc commencé 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.

1.3. Données et logiciels

1.3.1. Le PMSI

Le test d'anonymisation a été mené sur l'édition 2012 du PMSI, Programme de médicalisation des systèmes d'information. C'est une base médico-administrative annuelle qui rassemble la totalité des séjours hospitaliers publics et privés en France. Chaque enregistrement dans la base (près de 26 millions en 2012) correspond à un séjour. Le PMSI contient des informations détaillées de nature médicale (comme les actes chirurgicaux réalisés et les diagnostics posés), administrative (comme le type d'établissement hospitalier et sa localisation) et sociodémographique (comme l'âge, le sexe et le lieu de résidence du patient). Les bases du PMSI sont actuellement diffusées aux structures qui en font la demande après autorisation de la Cnil. Pour les besoins du test, ces données ont été mises à disposition sur le serveur du CASD.

Le CASD[7], Centre d'accès sécurisé aux données, a été initialement créé en 2009 pour la diffusion de fichiers de données individuelles de l'Insee notamment aux chercheurs. Cette entité du Genes[8] fournit un équipement permettant d'accéder à distance à des données confidentielles non anonymisées. Le CASD assure ainsi l'interface entre plusieurs producteurs de données (Insee, ministère de la justice, des finances, de l'agriculture, de l'éducation, Inserm, BPI) et les chercheurs. Conçu comme un serveur étanche, l'accès aux données se fait à distance *via* un boîtier, une authentification forte par empreinte digitale et dans le cadre d'un projet précis habilité par l'autorité (ou les autorités) compétente(s) (Cnil, Comité du secret statistique, producteur de données). L'environnement étant fermé, toutes les importations et toutes les sorties sont soumises au contrôle du CASD, même si la responsabilité concernant le respect du secret statistique est déléguée aux chercheurs par la signature d'un contrat valant engagement à respecter la confidentialité.

Le périmètre du test a été restreint au secteur de la médecine, la chirurgie et l'obstétrique (MCO), également dit « secteur de court séjour ». On parle de la base PMSI-MCO. Les séjours comportant des séances (de chimiothérapie, de radiothérapie ou de dialyse par exemple) ont également été exclus de la base. En effet, ces derniers peuvent indifféremment faire l'objet d'un seul enregistrement cumulant l'ensemble des séances ou d'un seul enregistrement par séance, et cette liberté organisationnelle est susceptible de perturber l'évaluation du risque de ré-identification. La

[7] https://casd.eu/
[8] Groupe des Écoles Nationales d'Économie et de Statistique

base PMSI-MCO pour l'année 2012 sur laquelle le test d'anonymisation a été effectué contient ainsi 20,6 millions de séjours hospitaliers. On dispose pour chaque séjour de son classement dans un GHM (groupe homogène de malades) qui est une variable médicale. Les GHM sont une classification médico-économique hiérarchique[9] dont le niveau le plus agrégé comporte 28 catégories majeures de diagnostic, les CMD. La CMD représente la discipline de prise en charge du séjour et constitue une bonne synthèse médicale du séjour. Dans ce test qui constitue une première approche, le détail des diagnostics et des actes n'a pas été analysé. Avec l'exclusion des séances et des CMD inconnues, l'information médicale sensible à protéger comporte 26 modalités[10].

1.3.2. µ-argus

La création de jeux de données anonymisés à partir du PMSI a été testée avec le logiciel gratuit µ-argus [8]. Une équipe de chercheurs a travaillé en parallèle avec le logiciel ARX.

µ-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[11]. L'interface de µ-argus est relativement intuitive et le logiciel est utilisé par plusieurs Instituts de statistique publique européens, souvent en combinaison avec d'autres outils [9]. La mise en œuvre des techniques d'anonymisation nécessite au préalable l'import de fichiers plats et la définition des métadonnées. Différentes méthodes d'anonymisation, perturbatrices ou non, peuvent ensuite être appliquées. Le logiciel permet d'exporter le jeu de données créé.

2. Nature des variables et risques de ré-identification

Supprimer les identifiants directs est une étape nécessaire mais non suffisante pour garantir le respect de la confidentialité et protéger contre le risque de ré-identification. La pseudonymisation est le fait de remplacer l'identifiant direct d'une personne (nom, adresse ou NIR pour un individu, numéro SIREN pour une entreprise) par un autre identifiant arbitraire - appelé pseudonyme - qui ne contient pas d'information connue des futurs utilisateurs du fichier. Cette étape ne peut à elle-seule garantir l'anonymisation car la combinaison d'informations indirectement nominatives du fichier (âge, sexe, lieu de résidence par exemple) peut, lorsqu'elle aboutit à un individu unique, permettre de ré-identifier les individus.

2.1. Les clés d'identification résultent du croisement des variables indirectement nominatives

Les personnes cherchant à ré-identifier un individu dans un fichier de données (appelées « attaquants ») 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, 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 » potentiels connaissent ou peuvent trouver et sur ce qu'ils recherchent et que l'on veut protéger.

Ces hypothèses amènent à préciser la nature des variables contenues dans le fichier de données. Le choix et la distinction des variables quasi-identifiantes et des variables sensibles (celles qu'il faut protéger du risque de divulgation) est le point d'entrée crucial dans la démarche de protection du fichier dans laquelle s'inscrit ce test. On considère qu'on peut catégoriser les variables du fichier selon 4 types :

[9] Elle classe chaque séjour à la suite d'un algorithme.
[10] La CMD à 26 modalités correspond ainsi le plus souvent à un système fonctionnel, une région anatomique (affections du système nerveux, de l'œil, de l'appareil respiratoire…).
[11] http://neon.vb.cbs.nl/casc/..%5Ccasc%5Cindex.htm

- des identifiants directs : numéro SIREN pour une entreprise, NIR (numéro de sécurité sociale) ou adresse complète pour un individu...
- des informations indirectement nominatives, dites « quasi-identifiants », qui, lorsqu'on les combine, peuvent permettre à un utilisateur du fichier de retrouver une personne donnée, même si la donnée d'une seule variable quasi- identifiante ne permet généralement pas la ré-identification. Par exemple, le sexe, l'âge ou le lieu de résidence sont généralement considérés comme des quasi-identifiants.
- 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 par un « attaquant » s'il est possible de les relier à l'individu (respectivement l'entreprise) auquel elles se rapportent.
- des informations non sensibles et non identifiantes.

Une clé d'identification c_i , $i \in \{1 \ldots . . J\}$ est une combinaison de modalités des variables indirectement nominatives, les quasi-identifiants. Le croisement des valeurs prises pour chaque variable quasi-identifiante forme la clé d'identification de l'individu.

Pour le test de réduction du risque de ré-identification sur le PMSI-MCO, les variables analysées sont :

0-la CMD, catégorie majeure de diagnostic qui est la variable sensible médicale

1-le numéro Finess d'identification de l'établissement (permettant d'identifier le lieu d'hospitalisation)

2-le code géographique de résidence du patient[12]

3-l'âge du patient

4-le sexe du patient

5-la durée de son hospitalisation

6-son mode d'entrée

7-et de sortie de l'hôpital ou de la clinique

Ces 7 dernières variables ont été retenues comme quasi-identifiants.

Par exemple dans le cadre de ce test, 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 5 jours d'hospitalisation, la clé d'identification est la suivante :

Clé c_1: {Sexe=1 ; Âge=50 ; Lieu de résidence= 01053 ; n° Finess d'établissement= 01 000 962 9 ; Mode d'entrée=7 ; Mode de sortie=8 ; Durée=5}

2.2. Risque de ré-identification

Des précédentes analyses [10] [11] ont montré, que malgré la suppression des identifiants directs, des données peuvent être indirectement nominatives, la combinaison de plusieurs modalités permettant d'identifier de façon unique des individus. Ainsi, sur les données du PMSI-MCO 2008 chaînées, c'est-à-dire en considérant l'ensemble des séjours sur l'année pour chaque patient, D. Blum [12] trouve que 89 % des patients sont uniques si l'on combine leurs caractéristiques sociodémographiques (âge, sexe, code géographique de résidence) et celles de leur(s) hospitalisation(s) (établissement d'hospitalisation, mois de sortie, mode de sortie, durée des hospitalisations et délais entre les hospitalisations). Ce taux monte à 100 % pour les patients qui ont effectué plus d'un séjour en 2008. Notons que, dans le fichier de données utilisé dans le cadre de ce test, l'identifiant permettant de chaîner les différents séjours effectués par un même patient n'est pas considéré.

[12] 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 de façon à ce que chaque code PMSI corresponde à au moins 1 000 habitants.

2.2.1. Nature des risques

La littérature sur les méthodes d'anonymisation distingue différents risques [13].

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 n°XXX appartient au président de la République.

Le risque de révélation d'attribut (*attribute linkage*) : on obtient des informations sensibles, 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 sensibles sont relatives à la personne et leur divulgation peut lui être préjudiciable. Par exemple, si dans le fichier exhaustif 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, vit dans cette commune et a reçu une prescription, il apprendra que Monsieur D. a pris des somnifères, même si l'enregistrement correspondant lui est inconnu.

Le risque de révélation inférentielle (*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.

2.2.2. Approche descriptive dans la base

La distribution non-uniforme des séjours et de leurs modalités au sein de la population ainsi que le caractère exhaustif de la base PMSI rendent la ré-identification possible. Les variables sociodémographiques mais également celles relatives au séjour sont discriminantes. En effet, les séjours hospitaliers sont plus fréquents aux âges élevés, l'état de santé se dégradant, ainsi que chez les nouveau-nés, tandis que les 1-14 ans sont sous-représentés parmi les patients hospitalisés en MCO. Les séjours longs (au-delà d'une semaine) sont logiquement peu fréquents en MCO, secteur dit de court séjour. Les modes d'entrée et de sortie de l'hôpital autres que domicile sont relativement rares et le décès est particulièrement identifiant (*cf. tableau 1*). Enfin, 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. En prenant par exemple 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.

Tableau 1 :

Variables	**Modalités**			
	Rares (en % des séjours)		**Fréquentes (en % des séjours)**	
Age	1-4 ans	3 %	Moins de 1 an	5 %
	5-9 ans	2 %	Tranches quinquennales à partir de 50 ans	6 % à 8 %
	10-14 ans	2 %		
durée	7 nuits ou plus	15 %	0 nuit	45 %
			1 à 6 nuits	40 %
mode d'entrée	transferts	2 %	domicile	83 %
mode de sortie	décès	1 %	domicile	77 %
	transferts	7 %		

Ces premières statistiques descriptives univariées permettent d'identifier *a priori* les modalités rares, celles qui pourraient être à l'origine des risques de ré-identification. Cela permet d'établir une première stratégie de regroupement de modalités pour diminuer ce risque. Néanmoins, le recours à ces statistiques pour réduire le niveau de détail des variables quasi-identifiantes présente de fortes

limites car les méthodes se basant sur le critère du k-anonymat (*cf. paragraphe 3.2*) pour diminuer le risque de ré-identification analysent les croisements de toutes les variables quasi-identifiantes, sans se limiter aux distributions univariées.

L'objectif du test est de construire à l'aide des méthodes d'anonymisation (*cf. paragraphe 3.1*) des fichiers présentant un risque de ré-identification réduit, risque évalué à l'aune de 2 critères de protection usuels retenus : le k-anonymat et la l-diversité (*cf. paragraphes 3.2 et 3.3*). Les choix des valeurs seuils k et l font l'objet de débats ; celles-ci ne sont pas définies institutionnellement.

3. Méthodes d'anonymisation et critères de réduction du risque

3.1. Choix des méthodes « non perturbatrices »

Les méthodes d'anonymisation dites perturbatrices (qui altèrent les données, comme le bruitage ou la permutation aléatoire de valeurs, *cf. tableau 2*) n'ont pas été retenues dans le cadre de ce test. En effet, 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. En effet, dans une démarche d'*open data*, les utilisateurs peuvent être très divers, et les échanges entre producteurs et utilisateurs (aux profils variés) des données sont difficiles.

Les méthodes d'anonymisation « non perturbatrices » ou « restrictives » consistent à publier une information moins précise en modifiant le détail (regroupement de modalités) et/ou la quantité d'information (suppressions locales, échantillonnage). 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) 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). Enfin, en première approche, le groupe a également souhaité conservé l'exhaustivité de la base PMSI-MCO, écartant les techniques d'échantillonnage.

La généralisation consiste à regrouper ou recoder, localement pour les individus trop peu nombreux dans leur clé d'identification ou globalement pour l'ensemble des individus, les modalités des variables quasi-identifiantes. Le regroupement peut se faire en passant à un niveau plus agrégé de nomenclature ou en agrégeant les valeurs extrêmes de la distribution (*top coding*) et permet de réduire le nombre d'enregistrements comportant une combinaison rare de modalités des variables quasi-identifiantes. Plus le niveau d'agrégation est important, moins il y a de risques de ré-identification, mais moins l'information est précise. Après regroupement, *i.e.* diminution du nombre de modalités et donc de croisements, il y aura j clés avec $j < J$.

Tableau 2 :

Méthodes non perturbatrices ou restrictives (modifient la quantité et le détail de l'information)	**Méthodes perturbatrices (modifient la valeur des données initiales)**
Agrégation/Généralisation = regroupement de modalités (recodage global ou local)	Microagrégation
Suppressions locales	Bruitage
Échantillonnage	Permutations aléatoires (*swapping*)

3.2. Le k-anonymat pour réduire le risque de révélation d'identité

Un fichier est dit k-anonyme si, pour toute clé d'identification *i*, il existe au moins k individus possédant la clé *i*. Tout individu est par conséquent indistinguable d'au moins $k-1$ autres [14]. Sans information autre que la clé d'identification, un individu a moins de une chance sur k d'être retrouvé pour un fichier k-anonyme.
Dans un fichier k-anonyme on a donc :

$$n_i \geq k \; \forall \; i \; \in \{1 \ldots . J\}$$
$$où \; J = nombre \; total \; de \; clés \; d'identification$$
$$et \; n_i = effectif \; \left(nombre \; d'individus\right) \; possédant \; la \; clé \; c_i$$

Dans le cadre de ce test, on fixe $k = 10$: un fichier est 10-anonyme si chaque individu est indistinguable (du point de vue de l'ensemble de ses caractéristiques quasi-identifiantes) d'au moins 9 autres individus. En d'autres termes, une personne qui connaît l'ensemble des modalités prises par un individu pour tous les quasi-identifiants, et donc sa clé d'identification, trouvera dans la base au moins 10 individus (et 9 autres individus) partageant cette clé et ne pourra pas ré-identifier l'individu concerné. La k-anonymisation (transformation d'un fichier pour atteindre le k-anonymat) protège du risque de révélation de l'identité (*record linkage*), mais pas forcément des autres risques mentionnés au paragraphe 2.2.1.

3.3. La l-diversité pour réduire le risque de révélation de l'attribut sensible

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. Or, dans un fichier k-anonyme, si tous les individus ayant la même clé partagent également 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). En d'autres termes, il y a divulgation d'attribut pour un groupe d'individus (*group disclosure*). Si les patients (au moins 10 dans ce test) 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 pour son séjour). L'attribut sensible est divulgué du fait de l'homogénéité du groupe d'individus partageant la même clé d'identification.

Pour protéger davantage le fichier du risque de révélation d'attribut, le concept de l-diversité a été utilisé. Un fichier est dit l-divers si, pour chaque clé d'identification c_i, il existe au moins l modalités représentées pour chaque variable sensible [15]. En d'autres termes, un fichier est dit l-divers si et seulement si, pour chaque clé d'identification, chaque variable sensible est suffisamment diversifiée, avec au moins l modalités différentes.

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 estimée 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 l'information possédée par un « attaquant ». Il existe d'autres critères de protection comme la t-proximité ou la confidentialité différentielle permettant de prendre en compte cet aspect [16].

Dans le cadre de ce test, $l = 3$ a été retenu. Ainsi un fichier est 3-divers si, parmi chaque groupe d'individus possédant les mêmes caractéristiques quasi-identifiantes, on trouve au moins 3 spécialités d'hospitalisation (3 CMD) différentes.

Les valeurs de k et de l ont été longuement discutées dans le groupe de travail. En pratique, le seuil de $k = 5$ est parfois utilisé [17]. Néanmoins, par précaution, la valeur de $k = 10$ a été retenue pour le test [7]. Les paramètres définissant les objectifs de réduction du risque ($k = 10\ et\ l = 3$) sont fixés dans la suite de ce papier.

4. Mise en œuvre sous µ-argus

4.1. Démarche itérative

La mise en œuvre des méthodes d'anonymisation sous µ-argus pour construire des fichiers respectant les critères de k-anonymat et de l-diversité est une démarche itérative. Une fois les données chargées dans le logiciel et les métadonnées définies, µ-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[13], 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 départemental (*cf. tableau 3*). Cette transformation préalable permet d'éviter un nombre de croisement trop grand au vu des capacités computationnelles du logiciel (notamment en cas de variables continues).

Tableau 3 : Fichier de données individuelles en entrée du µ-Argus

Nom de la variable	Nature de la variable	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	98
Nombre de clés d'identification		**893 760**
CMD (Catégorie Majeure de Diagnostic)	Sensible	26

Si le fichier de départ ne vérifie pas les deux critères de réduction du risque retenus dans ce test (*10-anonymat et 3-diversité*), il faut réduire le nombre de croisements (*i.e.* le nombre de clés d'identification) possibles en appauvrissant le niveau de détail des quasi-identifiants. Cela revient à regrouper certaines modalités de manière pertinente, conformément aux usages habituels en santé et en minimisant la perte d'information, c'est à dire en travaillant sur les modalités qui sont les plus identifiantes. µ-Argus permet de voir en temps réel le nombre de clés avec moins de 10 enregistrements, de détecter les modalités des variables quasi- identifiantes concernées (*i.e.* qui sont à l'origine d'un important risque de ré-identification) et leur fréquence d'apparition dans la base. On procède par itérations jusqu'à obtenir un fichier 10-anonyme : chaque individu (en l'occurrence chaque séjour et donc chaque patient[14]) est indistinguable (au vu de sa clé d'identification) d'au moins 9 autres individus possédant les mêmes modalités quasi-identifiantes.

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 des risques de ré-identification.

[13] Il y a potentiellement plus de clés d'identification que de séjours dans la base car il y a beaucoup de « vide », de clés qui ne sont prises par aucun séjour.
[14] Hors chaînage : dans la base PMSI-MCO sur laquelle est réalisé ce test, l'information permettant de relier, le cas échéant, les différents séjours pour un même patient.

La l-diversité est vérifiée *a posteriori* à l'aide d'un programme SAS. Il s'agit de s'assurer que pour chaque clé d'identification, le groupe de séjour associé présente au moins 3 CMD différentes.

4.2. Exemple de fichiers construits

En première approche, sans dimension géographique, µ-argus vérifie le respect du k-anonymat pour les 9 120 clés d'identifications (*cf. tableau 4*). Sur ces clés d'identification, 1 132 (12 %) comportent moins de 10 séjours. La démarche itérative d'agrégation 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 4*). Ces regroupements concernent logiquement les modalités rares déjà entrevues par les statistiques descriptives sur la base initiale (*cf. paragraphe 2.2.2*). Par ailleurs, la tranche d'âge des 5-14 ans correspond à la pédiatrie, ce qui reste pertinent pour les analyses et les études dans le domaine de la santé. Chacune des 1 728 clés d'identification comporte bien 3 CMD différentes, le fichier est donc 3-divers.

Tableau 4 : Niveau de détail dans le 1er fichier 10-anonyme en sortie de µ-argus

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

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-anonymat. 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 5*). 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].

[15] Trois clés d'identifications sur les 2112 (hors modalités manquantes) possibles aboutissent à seulement 2 CMD différentes. Les croisements "seulement" 2-diversifiés concernent la CMD 15 à savoir les nouveau-nés et prématurés. On peut penser que pour ces cas, la CMD est une variable plutôt identifiante que sensible.

Tableau 5 : Niveau de détail dans le fichier 10-anonyme avec une dimension géographique en sortie de µ-argus

Nom de la variable	Nature de la variable	Nombre de modalités dans le fichier initial	Nombre de modalités dans le fichier 10-anonyme
Âge	Quasi-identifiante	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-identifiante	12	2 : +ou- d'une semaine
Mode d'entrée	Quasi-identifiante	2	2
Mode de sortie	Quasi-identifiante	2	2
Lieu de résidence	Quasi-identifiante	23	22 : Regroupement Corse et PACA
Nombre de clés d'identification		39 744	2 112
CMD (Catégorie majeur de diagnostic)	Sensible	26	26

5. Enseignements et perspectives

5.1. Démarche d'anonymisation : connaissances métiers et spécificité du fichier

L'optimisation des regroupements à opérer pour maximiser l'utilité du fichier 10-anonymisé est complexe. Des algorithmes existent [13] 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. 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.

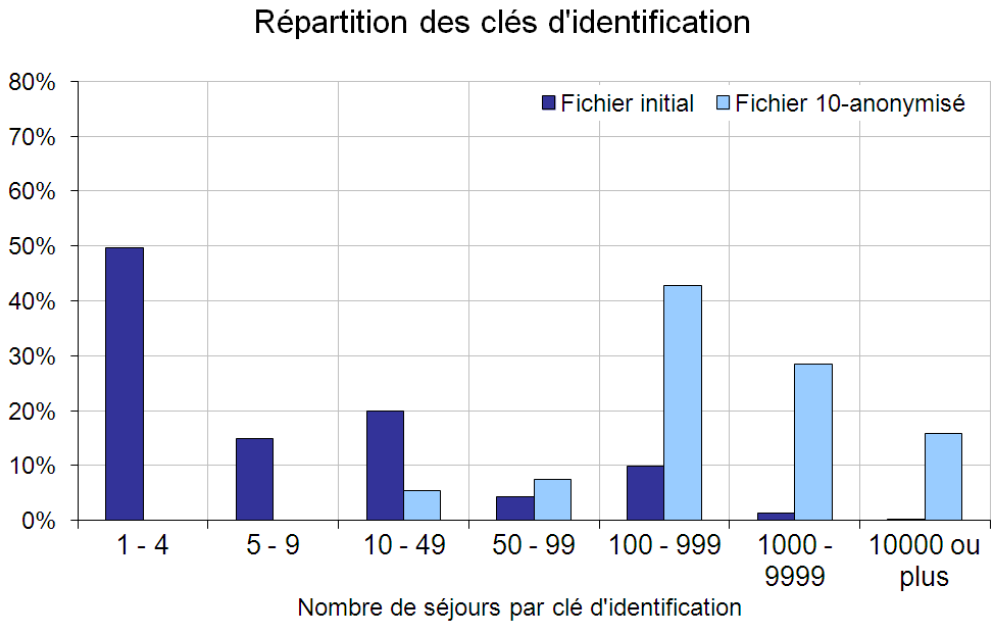
Des regroupements conformes aux pratiques usuelles dans le domaine de l'épidémiologie (par exemple des tranches d'âges quinquennales pour les individus âgés de plus d'un an de manière à isoler les nourrissons) ont été pratiqués. Cette approche permet par la suite de maximiser l'utilité des jeux de données construits ou d'arbitrer entre les différents jeux de données qui satisfont les critères d'anonymisation.

La spécificité du fichier, en l'occurrence la distribution non uniforme des séjours au sein de la population, joue également sur les possibilités et le niveau de regroupement. Dans le fichier extrait du PMSI-MCO avec le département de résidence du patient, son âge et la durée de son hospitalisation discrétisés, 65 % des clés d'identification[16] apparaissant dans le fichier sont possédées par moins de 10 séjours, et 50 % par moins de 5 (*cf. graphique 1*). Les clés qui ne vérifient pas le critère de 10-anonymat ne représentent cependant que 2 % des 20,6 millions de séjours contenus dans la base. Dans le fichier 10-anonyme et 3-divers avec la région de résidence construit à l'aide du logiciel µ-Argus (*cf. tableau 5*), 56 % des clés d'identification contiennent au moins 500 séjours, et moins de 1 % comportent entre 10 et 19 séjours. Il y a donc beaucoup de clés d'identification qui sont

[16] On ne considère pas ici les combinaisons de quasi-identifiants qui ne sont prises par aucun patient, en d'autres termes les combinaisons théoriquement possibles mais qui n'apparaissent pas dans le fichier.

possédées par un très grand nombre d'individus, ce qui signifie que le risque de ré-identification a nettement été réduit, mais au prix d'une perte d'utilité importante.

Graphique 1 :



5.2. Protection et utilité : la quadrature du cercle ?

Du fait de la distribution non uniforme des modalités des variables quasi-identifiantes, construire des jeux de données 10-anonymes sans recourir à des méthodes perturbatrices n'est pas aisé. Le producteur de données doit donc arbitrer entre le risque encouru et la perte d'information liée aux regroupements de modalités pour respecter les critères de réduction du risque retenus. En effet, même si, conceptuellement, l'*open data* s'adresse à tous les citoyens indépendamment de leurs motifs d'utilisation des données publiques, il est indispensable de réfléchir aux possibilités d'analyse offertes par les fichiers mis à leur disposition.

Il n'a pas été possible d'obtenir avec la démarche présentée ici un fichier 10-anonyme qui contienne en plus de la région de résidence du patient, le lieu d'hospitalisation (même au niveau régional). 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-anonymat). Abaisser le critère de k-anonymat, par exemple en prenant $k = 5$, ne réduit pas d'autant le nombre de clés d'identification problématiques. 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. Enfin, certains établissements étant spécialisés (les centres de lutte contre le cancer par exemple), le critère de l-diversité est d'autant plus difficile à atteindre si on ajoute une variable concernant l'établissement d'hospitalisation.

Une autre approche peut être d'avoir 2 niveaux de détail d'information différents selon la rareté du séjour en utilisant les recodages locaux. L'équipe travaillant avec le logiciel Arx a présenté un tel fichier : pour les séjours atypiques (3,8 % de l'ensemble des séjours), 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 (région d'hospitalisation et département de résidence) sont incluses. On voit clairement sur cet exemple qu'un nombre restreint de séjours atypiques (qui ont une clé d'identification possédée par moins de 10 patients) a un impact considérable sur le niveau d'anonymisation de l'ensemble des séjours.

Si la perte d'information est moindre avec le recodage local, 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. En outre, le fait de mettre à blanc certaines modalités pour une partie des séjours donne une indication sur le caractère atypique du séjour et peut être utilisée comme une information pour ré-identifier.

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. Les possibilités de croisement à l'intérieur d'un fichier et d'appariements entre différents fichiers peuvent accentuer le risque de ré-identification. Dans ce cas, les objectifs de réduction du risque présentés dans ce papier et conçus pour un unique fichier (non apparié) peuvent être insuffisants.

Au-delà des critères retenus dans le cadre de ce test, il est difficile de savoir comment mesurer le risque ou définir un risque acceptable. Il n'y a pas à l'heure actuelle de mesure du risque de ré-identification et de niveau de risque acceptable qui fasse largement consensus dans la communauté des chercheurs ou des producteurs de données.

Bibliographie et références

[1] Gorce G. et Pillet F., « La protection des données personnelles dans l'*open data* : une exigence et une opportunité », Rapport de la mission d'information du Sénat sur l'*open data*, avril 2014 .
http://www.senat.fr/notice-rapport/2013/r13-469-notice.html

[2] https://www.data.gouv.fr/fr/

[3] Directive européenne du 24 octobre 1995.
http://eur-lex.europa.eu/legal-content/FR/TXT/?uri=CELEX:31995L0046

[4] Bras P.-L., « Rapport sur la gouvernance et l'utilisation des données de santé », Drees, octobre 2013.
http://www.drees.sante.gouv.fr/rapport-sur-la-gouvernance-et-l-utilisation-des-donnees-de,11202.html

[5] Rapport de la Commission *open data* remis le 9 juillet 2014 à Marisol Touraine.
http://www.drees.sante.gouv.fr/rapport-de-la-commission-open-data-en-sante,11323.html

[6] Projet de loi santé présenté au conseil des ministres le 15 octobre 2014.
http://www.legifrance.gouv.fr/affichLoiPreparation.do?idDocument=JORFDOLE000029589477&type=general&typeLoi=proj&legislature=14

[7] Annexe 9 du rapport de la commission *open data* : Rapport du groupe de travail « risque de ré-identification », juillet 2014.
http://www.drees.sante.gouv.fr/annexes-du-rapport-de-la-commission-open-data-en-sante,11324.html

[8] Hundepool A. *et al.*, « µ-Argus User's Manual », 2008.

[9] Eurostat, *Results on the questionnaire on SDC tools*, *5th meeting of the Expert Group on Statistical Disclosure Control*, octobre 2013

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

[11] Tarran B., « Does data anonymisation work? », *Significance*, pp. 11-16, 2014.

[12] Blum D., 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

[13] Fung B.C.M, Wang K., Chen R., Yu P.S., « Privacy preserving data publishing: a survey of recent developments »., *ACM computing surveys*, 42, 4, article 14, 2010.

[14] Bergeat M. « Données individuelles : bien les protéger pour mieux les diffuser », 46ièmes journées de Statistique de la SFdS, juin 2014.
http://papersjds14.sfds.asso.fr/submission_204.pdf

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

[16] Hundepool et al., *Statistical Disclosure Control*, *Wiley Series in Survey Methodology*, 2012

[17] El Emam, K. *et al.*, « A Globally Optimal k-Anonymity Method for the De-Identification of Health Data », *Journal of the American Medical Informatics Association*, vol. 16, pp. 670-682, 2009.
http://jamia.bmjjournals.com/content/16/5/670.full