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

Domaine : *confidentialité, protection des données individuelles, ouverture des données de santé*

Auteur : Noémie Jess[1] (Drees) avec la collaboration de Maxime Bergeat (Insee) et de Françoise Dupont (Insee-CASD)

La ministre de la santé Marisol Touraine a demandé au directeur de la Drees[2] 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[3] lancée à l'automne 2013 pour répondre à la demande croissante de l'ouverture des données de santé. L'amélioration de l'accès aux données de santé fait l'objet d'une mesure dans la proposition de loi de santé publique. Cette expertise a été menée par un groupe de travail constitué de personnalités qualifiées dans différentes disciplines (informatique, statistique, confidentialité et diffusion de données, santé, épidémiologie) et de représentants de organismes producteurs et/ou utilisateurs de données de santé (Atih, Cnamts, Inserm-CépiDc, Drees).

L'objet de la communication est de présenter le test d'anonymisation des données hospitalières du PMSI-MCO[4], réalisé dans le cadre de ce groupe de travail. Elle s'appuie sur le Dossier Solidarité et Santé « *Risque de ré-identification dans les bases de données médico-administratives* » à paraître d'ici décembre ainsi que sur le séminaire « *Frontière entre données quasi-identifiantes et données anonymes : comment se prémunir du risque de ré-identification ?* » organisé par la Drees le 10 décembre prochain.

Le PMSI-MCO contient des informations détaillées (médicales, administratives et sociodémographiques) sur la totalité des courts séjours effectués en France. Des travaux antérieurs ont montré que les données de cette base exhaustive, malgré la suppression des identifiants directs, sont indirectement nominatives du fait de la distribution non-uniforme des modalités des séjours. Les bases du PMSI sont actuellement diffusées aux structures qui en font la demande après autorisation de la Cnil.

En premier lieu, cette communication présente le risque de ré-identification contenu dans les données et décrit la distinction opérée entre la variable sensible à protéger (l'information médicale : le motif du séjour résumé dans la catégorie majeure de diagnostic) et les variables indirectement identifiantes (âge, sexe et lieu de résidence du patient, mode d'entrée et de sortie de l'hôpital, durée et lieu d'hospitalisation).

La méthode d'anonymisation sélectionnée (regroupements de modalités) pour la construction de fichiers diffusables en open data à partir du PMSI-MCO est ensuite discutée. La méthode choisie est non perturbatrice, c'est-à-dire qu'elle ne modifie pas la valeur des données initiales mais diminue le niveau de détail et la quantité d'information dans le fichier. Ce choix découle de la volonté de conserver le caractère exhaustif de la base et d'éviter de complexifier voire de biaiser les analyses – ce qui est conforme aux bon usages de la diffusion des données de la statistique publique.

---

[1] noemie.jess@sante.gouv.fr, Sous-direction observation de la santé et de l'assurance maladie, Bureau des dépenses de santé et des relations avec l'assurance maladie.

[2] Direction de la recherche, des études, de l'évaluation et des statistiques du ministère de la santé.

[3] Installée fin novembre 2013, la commission a rendu ses conclusions sous forme d'un rapport en le 9 juillet 2014 http://www.drees.sante.gouv.fr/rapport-de-la-commission-open-data-en-sante,11323.html.

[4] 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), soit (hors séances) 20,6 millions d'enregistrements en 2012.

Enfin, les critères de protection retenus par le groupe de travail (k-anonymat, l-diversité[5]) sont explicités à travers les exemples de jeux de données construits avec les logiciels Arx et Mu-argus. En effet, ces fichiers sont :

(i) 10-anonymes : chaque individu est indistinguable (du point de vue de l'ensemble de ses caractéristiques indirectement identifiantes) d'au moins 9 autres individus, ce qui permet de protéger contre le risque de révélation d'identité.

(ii) 3-diversifiés : parmi chaque groupe d'individus possédant les mêmes caractéristiques indirectement identifiantes, on trouve au moins 3 motifs d'hospitalisation différents, ce qui protège contre le risque de révélation d'attribut (l'information sensible à protéger étant le motif d'hospitalisation).

Enfin, des perspectives générales sont présentées sur la démarche d'anonymisation et le difficile arbitrage entre le degré de protection apporté et le détail de l'information dans le fichier diffusé. 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 apparu indispensable de réfléchir aux possibilités d'analyse offertes par les fichiers mis à leur disposition.

[5] Les valeurs appropriées de k et de l font encore l'objet de débat.