Alexis Astruc[1,2,3], Adeline Jouannin[4,5,6], Erik Lootvoet[1], Timothée Bonnet[7], Frederic Chevallier[1,8]

1.. Université Sorbonne Paris Nord, Département universitaire de Médecine générale, DUMG, F-93000 Bobigny
2.. Centre de santé Ellasanté, F-75008, Paris
3. Maison de santé Orly Santé, F-94310, Orly
4. Département de médecine générale, Université de Rennes, F-35000 Rennes.
5. Centre d'investigation clinique de Rennes (CIC Inserm 1414), Université de Rennes, CHU Rennes, Inserm, F-35000 Rennes.
6. Centre de recherche des Cordeliers (UMRS1138), INSERM, Sorbonne Université, USPC, Université Paris-Descartes, Université Paris-Diderot, équipe ETREs, F-75006 Paris.
7. Université Sorbonne Paris Nord, Service des archives, F-93430 Villetaneuse
8. MUSSP, 14 rue de la république, 95120 Ermont

alexis.astruc@yahoo.fr
exercer 2021 ; 172:178-84.

Les données à caractère personnel : quelles formalités règlementaires pour les travaux de recherche en médecine générale ?

Personal data: what are the regulatory requirements for research work in general medicine?

Annexe 1 - Les différences entre pseudonymisation et anonymisation

La pseudonymisation est une méthode permettant de remplacer des données directement identifiantes par un attribut spécifique appelé "pseudonyme". La réidentification est donc possible, en comparant avec une table de concordance, voire par recoupement d'informations. Ce niveau de désidentification n'est pas suffisant pour se soustraire au RGPD.

Exemple : un patient n'est pas identifié dans la base de données par son nom et son prénom, mais par son année de naissance suivie de ses initiales.

L'anonymisation est l'ensemble de méthodes permettant d'exclure les données identifiantes d'une base de données afin d'empêcher toute réidentification des personnes concernées. Il faut que les données soient anonymisées avant leur traitement statistique.
Il faut faire attention au risque de réidentification. Par exemple, un patient avec une maladie rare sera facilement identifiable avec un nombre limité de données supplémentaires. De manière générale, il ne faut collecter que ce qui est nécessaire pour mener à bien la recherche autant dans une optique de minimisation du traitement de données que vis-à-vis de l'intégrité scientifique.
Il est possible de réidentifier les personnes avec précision et peu de faux-positifs, même à partir de jeux de données complètement anonymes et échantillonnés. Les résultats montrent qu'en utilisant certains modèles statistiques, 99,98 % des Américains seraient correctement réidentifiés dans n'importe quel ensemble de données en utilisant 15 attributs démographiques. Ainsi, même des ensembles de données anonymisés fortement et échantillonnés sont peu susceptibles de satisfaire aux critères d'anonymisation établis par le RGPD et remettent sérieusement en question l'adéquation technique et juridique de l'anonymisation des données.

Exemple :le patient en question n'est pas identifiable, les données ont été agrégées pour empêcher une désanonymisation. Par exemple l'âge est converti en classe d'âge, des données aléatoires ont été intégrées au jeu de données. Plusieurs patients ont des données similaires ou proches.

1. Rocher, Luc, Julien M. Hendrickx, et Yves-Alexandre de Montjoye. « Estimating the Success of Re-Identifications in Incomplete Datasets Using Generative Models ». *Nature Communications* 10, nº 1 (23 juillet 2019): 1▫9. https://doi.org/10/gf5d5h.