HEALTH
DATA HUB

Health Data Hub

Qu’est-ce qu’une donnée anonyme en santé ?

2022.08.31

Objectifs

L'objectif de ce document est d'aider la communauté des utilisateurs des données de santé à cerner ce que recouvre le processus d'anonymisation de ces données, son contexte d'application et les possibilités qu'il ouvre, notamment par rapport la pseudonymisation. Pour ce faire, ce guide est structuré en suivant les grandes étapes préalables à l'anonymisation des données de santé :

- **Comprendre** les notions d'anonymisation des données et d'identification des personnes
- **Choisir** entre anonymisation et pseudonymisation
- **Cadrer** le processus d'anonymisation
- **Vérifier** que les données anonymisées ne présentent aucun risque de réidentification

Ce guide est dérivé des ressources officielles faisant autorité en France sur la thématique de l'anonymisation, et en reprend synthétiquement certains extraits. En complément de cette synthèse, il est très utile de prendre connaissance de ces documents pour approfondir le sujet.

(a) Règlement Général sur la Protection des Données (RGPD)

(b) Avis 05/2014 du groupe de travail "article 29" sur les Techniques d'anonymisation

(c) Ressource CNIL "L'anonymisation de données personnelles"

Les références aux ressources listées ci-dessus, faites au travers du guide, seront identifiées via les symboles (a), (b) et (c).
Par souci de clarté, les liens ne seront pas répétés en bas de page.

Comprendre

Définition

L'anonymisation est un traitement qui consiste à utiliser un ensemble de techniques de manière à rendre impossible, en pratique, toute identification de la personne par quelque moyen que ce soit et de manière irréversible.[a] **Les données anonymisées ne sont plus considérées comme personnelles.**

Identifier une personne physique

Une personne physique peut être **directement identifiée** à partir d'informations prises isolément telles que *nom, adresse, numéro de téléphone, adresse email, NIR, etc.*

Une personne physique peut être **indirectement identifiée** en croisant plusieurs informations telles que *âge, code postal, médecin traitant, lieu de travail, etc.*

Anonymisation vs Pseudonymisation

La **pseudonymisation** est un traitement de données personnelles réalisé de manière à empêcher toute identification directe, typiquement en remplaçant les informations directement identifiantes par des numéros non signifiants.

À l'inverse de l'anonymisation, la pseudonymisation ne permet donc pas d'empêcher l'identification indirecte, et reste généralement réversible : **les données restent personnelles**.[c]

(a) (b) (c) : voir slide 2

Choisir

Le choix entre anonymisation et pseudonymisation résulte d'un compromis entre les questions suivantes :

1. Est-il nécessaire de préserver la finesse des données au niveau des individus composant le jeu de données ?
2. Est-il nécessaire de pouvoir partager simplement et largement le jeu de données ?

Quand choisir l'anonymisation ?

Quand l'utilisation ne justifie pas le recours à des données identifiantes ou qu'il est plus important de permettre un partage large et simple des données (car n'ayant plus à respecter la législation relative à la protection des données personnelles).

Exemples d'utilisation :

- Statistiques génériques (e.g. populationnelles)
- Exploitation large via des data challenges
- Partage libre de données pour assurer la reproductibilité de la recherche par exemple

Drug ID	Age group	Period	Patient count
A	20-30	Q1 2020	225
A	30-40	Q1 2020	137
B	20-30	Q2 2020	51

Exemple de jeu de données anonymisé

Quand choisir la pseudonymisation ?

Quand il est plus important de préserver la finesse des informations individuelles composant le jeu de données tout en limitant les risques liés à leur traitement.

Exemples d'utilisation :

- Études nécessitant des données fines à l'échelle individuelle (e.g. étude de parcours de soin individuels)
- Appariement de jeux de données à l'aide d'informations individuelles (e.g. NIR masqués)

Patient ID	Birth date	Drug ID	Date
EJ2I004D	1993/07	A	2020/01/04
EJ2I004D	1993/07	B	2020/05/23
FEO40JO	1974/01	A	2020/01/06

Exemple d'un jeu de données pseudonymisé

(a) (b) (c) : voir slide 2

Voir également la ressource CNIL "Recherche scientifique (hors santé) : enjeux et avantages de l'anonymisation et de la pseudonymisation"

Cadrer

Quel encadrement réglementaire ?

Les données résultant d'un traitement d'anonymisation, n'étant plus considérées comme personnelles, sortent du champ de la législation relative à la protection des données[c], en particulier du RGPD au niveau européen et de la LIL au niveau français.

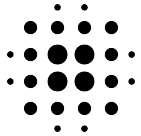
Toutefois, le processus d'anonymisation reste quant à lui un traitement de données personnelles, et doit donc respecter les dispositions prévues par ces textes, en particulier les principes de licéité[1] et de finalité[2].

Le détail des démarches réglementaires applicables au traitement d'anonymisation est exploré et exposé dans un guide dédié (à venir).

Comment anonymiser ?

Le choix du processus d'anonymisation dépend des réutilisations futures envisagées, et peut donc considérablement varier d'un projet à l'autre et d'un jeu de données à un autre.

On peut distinguer deux familles de techniques fréquemment utilisées pour anonymiser des données : la généralisation et la randomisation.[b]



La généralisation altère la finesse des données en modifiant leur échelle

k-anonymat
l-diversité
t-proximité

La randomisation altère la véracité des données pour affaiblir le lien avec l'individu

Ajout de bruit
Confidentialité différentielle
Permutation

(a) (b) (c) : voir slide 2

(1) https://www.cnil.fr/fr/les-bases-legales/liceite-essentiel-sur-les-bases-legales
(2) https://www.cnil.fr/fr/definition/finalite-dun-traitement

Vérifier

Les autorités de protection des données européennes retiennent **trois critères** qui, s'ils sont parfaitement vérifiés, permettent d'assurer qu'un jeu de données est véritablement anonyme. À défaut de remplir parfaitement ces trois critères, le responsable de traitement qui souhaite anonymiser un jeu de données doit démontrer, via une **analyse approfondie des risques d'identification**, que le risque de ré-identification avec des moyens raisonnables est négligeable.(c)

Individualisation

Il ne doit pas être possible d'isoler une partie ou la totalité des enregistrements liés à un individu dans le jeu de données.

Exemple :
Dans une base de données de parcours de soin individuels, les enregistrements (événements de soin) liés à un individu peuvent être isolés, puisqu'ils sont par construction liés dans un même parcours de soin.

Corrélation

Il ne doit pas être possible de relier deux enregistrements distincts concernant un même individu (que les bases de données soient distinctes ou non)

Exemple :
Dans un extrait d'une base de données hospitalières, constitué uniquement du numéro de séjour et du diagnostic principal, les enregistrements peuvent être reliés à la base source par le numéro de séjour.

Inférence

Il ne doit pas être possible de déduire, avec un degré de probabilité élevé, de nouvelles informations sur un individu

Exemple :
Une étude indique que dans la cohorte C, tous les hommes de plus de 50 ans sont diabétiques. Connaissant un homme de 50 ans participant à la cohorte, on peut en déduire qu'il est atteint de diabète.

(a) (b) (c) : voir slide 2

HEALTH
DATA HUB

Suivez-nous sur les réseaux sociaux !