HEALTH DATA HUB

Guide d'évaluation du caractère anonyme d'un jeu de données dans le cadre d'un projet de recherche

[NOM]

Aide au remplissage

*(*Cette partie est à effacer avant envoi du document*)*

Les projets utilisant des données parfaitement anonymisées, contrairement aux projets utilisant des données pseudonymisées, ne se voient pas appliquer les dispositions relatives à la protection des données (RGPD et loi "Informatique et Libertés"). L'accès et le traitement des données peuvent donc se faire sans avoir recours à une demande d'autorisation auprès de la CNIL ou à une procédure simplifiée telle qu'une méthodologie de référence.

L'évaluation du caractère anonyme des données reste toutefois de la responsabilité du porteur de projet, et il est donc nécessaire de ne pas la prendre pour acquise.

Il est donc proposé de vous accompagner ici dans la conduite d'une analyse de risque de réidentification afin de démontrer l'impossibilité de réidentifier les individus à partir des données considérées.

À noter que ce guide :

- **propose une méthode analytique** pour conduire une analyse de risque de réidentification d'un jeu de données mais **ne prescrit pas de méthode à employer** pour anonymiser un jeu de données ; cette activité reste sous la responsabilité du porteur de projet et correspond à un traitement de données personnelles nécessitant d'être respectueux de la réglementation ;
- pourra faire l'objet de mises à jour en fonction d'éventuelles **nouvelles recommandations de la CNIL** en matière d'anonymisation des données.

Dans la suite du document, nous vous invitons à effacer systématiquement les parties "Aide au remplissage", "Nos conseils" et "Définitions" pour ne laisser apparaître que vos réponses.

Nos conseils

*(*RAPPEL : cette partie est à effacer avant envoi du document*)*

Dans ce document, nous vous invitons à renseigner toute information pertinente sur les données considérées, que vous en soyez à **l'origine** (vous avez constitué la base de données), ou qu'elles soient **mises à votre disposition** par un tiers. Dans ce dernier cas, nous vous invitons à vous rapprocher du responsable de données afin qu'il vous

communique, si possible, la documentation de sa base - description des données, cycle de vie (sélection de l'échantillon et opérations d'anonymisation) - qui vous permettra de vous assurer du caractère anonyme des données..

Ces éléments peuvent alimenter les sections 2 et 3 mais l'évaluation et les analyses présentées restent de la responsabilité du porteur de projet même si vous faites intervenir des tiers dans l'écriture ou la relecture du document.

Enfin, afin de ne pas réaliser de traitement de données personnelles non encadré, nous vous invitons à n'inclure dans ce document **aucune des données réelles concernées**, même à titre illustratif.

HEALTH
DATA HUB

[TITRE COMPLET DU PROJET]

HEALTH
DATA HUB

Sommaire

1. Cadre du projet

Aide au remplissage

*(*Cette partie est à effacer avant envoi du document*)*

Le contexte du projet peut aider à comprendre la nature précise des données utilisées et aider le délégué à la protection des données à apprécier l'exposition à un éventuel risque de réidentification.

Présenter le projet et l'équipe en charge du projet :

- Finalité du traitement des données (objectif du projet mobilisant les données anonymes) ;
- Présentation de l'équipe ;
- Partenaires impliqués.

2. Traitements de génération de la base de données anonymes

Aide au remplissage

*(*Cette partie est à effacer avant envoi du document*)*

L'objectif de cette section est d'expliciter le processus de génération du jeu de données anonymes.

Décrivez dans un premier temps les données sources et leur processus de collecte, avec un niveau de détail suffisant pour qu'un novice du domaine puisse apprécier leur sensibilité. Dans le cas où vous récupérez les données anonymes de la part d'un responsable de données et que vous n'avez pas pu récupérer de documentation auprès de celui-ci : décrivez aussi précisément que possible ce que vous connaissez de ce processus.

Décrivez ensuite les étapes de traitement des données ayant permis la production de la base anonyme, en vous aidant des questions exposées ci-dessous et en explicitant à chaque fois les transformations appliquées (suppression d'information, généralisation, ajout de bruit, etc).

Décrivez enfin la base de données anonymes résultant du processus.

Pour chaque étape, vous devez vous poser les questions suivantes :

- Quel est le jeu de données impliqué (e.g. base source, base pseudonymisée, base anonymisée) ?
- Qui y a accès ?
- Quel est le traitement appliqué à ces données ?
- Qui applique ce traitement, avec quel outil / support informatique ?
- Quelles sont les données résultantes ?
- Quel lien (e.g. ID patient) est possible entre les jeux de données impliqués ?

La description de ce workflow, qui peut avantageusement être illustrée par un schéma, est essentielle pour identifier les possibilités pour un attaquant de "remonter" la chaîne de traitement jusqu'à une personne physique, i.e. de réidentifier la personne.

Pour conclure et résumer, listez les données identifiantes et/ou sensibles qui sont particulièrement à protéger, en indiquant leur présence / absence dans les différents jeux de données mentionnés au fil du processus (voir l'exemple joint, dont un extrait est repris ci-dessous).

Informations sensibles	***Base source***	***Base intermédiaire***	***Base anonymisée***
Nom & Prénom	*oui*	*Initiales*	***non***
Date de naissance	*oui*	*Année*	***non***
Diagnostic	*oui*	*oui*	***oui***

Nos conseils

*(*RAPPEL : cette partie est à effacer avant envoi du document*)*

Décrire précisément la mise en œuvre des techniques d'anonymisation en explicitant pour chaque étape :

- L'objectif de la transformation
- La ou les techniques d'anonymisation : méthode statistique, modèle, outils utilisés si pertinent, efficacité de la technique si connue, références ou certifications d'un fabricant, approche déjà utilisée dans un projet autorisé par la CNIL, caractère réversible, gestion des tables de passage si attribution d'identifiant, contrôle mis en oeuvre pour vérifier sa bonne application …
- Si une seule technique ne suffit pas à anonymiser suffisamment le jeu de données alors plusieurs techniques doivent être cumulées et toutes doivent être décrites une par une.

3. Evaluation du caractère anonyme des données

Aide au remplissage

*(*Cette partie est à effacer avant envoi du document*)*

Pour déterminer si les données visées sont anonymes, on procède en deux temps :

1- Vérifier le respect parfait et cumulatif des trois critères d'évaluation définis par les autorités européennes ci-dessous.

2- Si les trois critères ne sont pas réunis, le risque résiduel de réidentification avec des moyens raisonnables doit être évalué. Pour l'évaluer vous devez mener **une analyse de risque**. S'il est estimé nul, alors les données sont considérées comme anonymes.

3.1. Vérification du respect des trois critères d'évaluation définis par les autorités européennes.

Pour chaque critère, explicitez en quoi vous avez la garantie qu'il est parfaitement vérifié.

Définitions

*(*RAPPEL : cette partie est à effacer avant envoi du projet*)*

Critères d'évaluation définis par les autorités européennes (avis du G29[1]), qui, s'ils sont parfaitement respectés, garantissent l'anonymat d'un jeu de données[2] :

- **Non-individualisation** : il ne doit pas être possible d'isoler un individu dans le jeu de données.
  Par exemple, une base de données de CV où seuls les noms et prénoms d'une personne auront été remplacés par un numéro qui ne correspond qu'à elle permet d'individualiser cette personne. Cette base serait considérée comme pseudonymisée et non comme anonymisée.
- **Non-corrélation** : il ne doit pas être possible de relier ensemble deux jeux de données distincts et concernant un même individu.
  Par exemple, une base de données contenant des numéros de téléphones portables ne peut être considérée comme anonyme si d'autres bases de données, existantes par ailleurs, contiennent ces numéros avec d'autres données identifiantes telles que le prénom et le nom.
- **Non-inférence** : il ne doit pas être possible de déduire de façon quasi certaine, depuis des informations internes ou externes, de nouvelles informations sur un individu.
  Par exemple, si une base de données supposément anonyme contient des informations sur le montant des impôts de personnes ayant répondu à un

[1] https://ec.europa.eu/justice/article-29/documentation/opinion-recommendation/files/2014/wp216_fr.pdf
[2] Source CNIL : https://www.cnil.fr/fr/lanonymisation-de-donnees-personnelles

questionnaire, que tous les hommes ayant entre 20 et 25 ans ayant répondu sont non imposables, il sera possible de déduire, si on sait que M. X, homme âgé de 24 ans, a répondu au questionnaire, qu'il est non imposable.

3.2. Analyse de risque

Détaillez les scénarios de risque amenant à l'événement redouté de réidentification de votre jeu de données, les mesures de mitigation du risque éventuellement mises en œuvre, et le niveau de risque résiduel.

Définitions

*(*RAPPEL : cette partie est à effacer avant envoi du projet*)*

L'analyse de risque consiste à envisager des scénarios de risque amenant à l'événement redouté, à savoir la réidentification d'une personne physique par un attaquant qui accéderait aux jeux de données. Cette analyse doit présenter les éléments suivants :

- Description du scénario envisagé
- Évaluation du niveau de risque initial de réidentification sans mesure corrective : il s'agit ici d'évaluer la gravité de chaque scénario de risque, croisé avec la probabilité de survenue.
- Description des mesures correctives mises en oeuvre (par exemple, les nouvelles techniques d'anonymisation appliquées pour empêcher la réalisation de ce risque)
- Évaluation du niveau de risque résiduel de réidentification {impact x probabilité}, après application des mesures correctives : celui-ci doit être acceptable.

Nos conseils

*(*RAPPEL : cette partie est à effacer avant envoi du document*)*

Vous pouvez vous servir de la représentation ci-dessous, les chiffres correspondent aux échelles fournies en annexe.

Scénario et mesure corrective		I	P	C	Justification de la cotation pour I et P, C est déduite (cf annexe)
Sc #1	Exemple : La base de données anonymisées n'est plus constituée que des images. Cependant, il existe une base non-anonyme comportant ces images (base source) : ces deux bases pourraient être	2	1	2	Exemple : **Evaluation de la probabilité** (a) Probabilité qu'un attaquant dispose des données : **Nulle (1)** car : - Les bases de données sont stockées sur le serveur de l'entreprise. Il s'agit d'un serveur physique sécurisé, hébergé dans les locaux de la société, dont l'accès est strictement réservé aux collaborateurs autorisés, par

	reliées entre elles par les images (non-corrélation pas respectée)				identifiant unique et mot de passe. - Les bases sources (pseudonymisées) sont stockées sur un répertoire dédié à l'équipe clinique, seule habilitée à les consulter. - Les bases anonymes sont stockées sur un autre répertoire dédié à l'équipe IA. - La même base en version pseudonymisée et anonymisée n'est jamais conservée simultanément sur le serveur de l'entreprise. - Les bases de données anonymes qui seront exportées depuis le serveur de l'entreprise vers le serveur du HDH ne seront pas conservées sur le serveur de l'entreprise. (b) Probabilité de réussite d'une attaque tentée : **Fréquente (3)** : si une image présente une spécificité morphologique marquée, une comparaison 2 à 2 de toutes les images permettrait de retrouver les 2 images identiques, et donc l'identité du patient. Il n'est pas très compliqué pour un homme du métier qui disposerait des 2 bases de comparer une grande quantité d'images 2 à 2 de manière automatique. Par conséquent, la combinaison de (a+b) fournit une probabilité : **Possible (2)** **Evaluation de l'impact** (c) Impact de la réidentification : **Négligeable (1)** Recorréler les images entre elles nécessiterait d'avoir déjà accès à la base source et n'apporterait donc aucune information supplémentaire sur le patient.
MC #1	***NA***				
Sc #2					
MC #2					
Sc #3					

3.3. Conclure sur le caractère anonymes des données

Les risques identifiés sont-ils acceptables (impact négligeable et/ou probabilité nulle des scénarios de risque), une fois les éventuelles mesures correctives appliquées ?

4. Signature du porteur de projet

Aide au remplissage

*(*RAPPEL : cette aide au remplissage est à effacer avant envoi du projet*)*

Formaliser l'avis du DPO, et la validation et la signature du document par le porteur de projet.

Cette partie permet d'attester de la cohérence de la démarche, et des moyens mis en œuvre pour vérifier le caractère anonyme des données.

ANNEXE 1 - échelles utilisées

Afin de faciliter l'analyse des risques, une échelle de valeur est attribuée pour chaque critère d'évaluation des scénarios de réidentification (exemples ci-dessous adaptables).

Probabilité de réalisation du scénario de réidentification *Prendre en compte la probabilité qu'un attaquant dispose des données (i.e. qu'une attaque soit tentée),* *et la probabilité qu'une attaque tentée soit réussie*		
1	Nulle	Nécessiterait des moyens déraisonnables à mettre en oeuvre
2	Possible	Pourrait se produire à l'aide de moyen raisonnables
3	Fréquente	Peut se produire systématiquement

Impact en cas de réalisation du scénario de réidentification		
1	Négligeable	Aucun impact sur la personne
2	Mineur	Impact mineur, e.g. informations peu sensibles dévoilées
3	Significatif	Impact significatif, e.g. informations très sensibles dévoilées

		CRITICITÉ = IMPACT x PROBABILITÉ		
		Négligeable	**Mineur**	**Significatif**
		1	2	3
Fréquente	3	**3**	**6**	**9**
Possible	2	**2**	**4**	**6**
Nulle	1	**1**	**2**	**3**

NIVEAU D'ACCEPTABILITÉ DES RISQUES		
	Acceptable	Aucune mesure corrective nécessaire
	Indésirable	Mesure corrective ou justification
	Inacceptable	Mesure corrective obligatoire