Entrepôts de Données de Santé et Protection de la Vie Privée : Synthèse de discussions Inter-CHU

PFIA 2024 - Journée Santé & IA

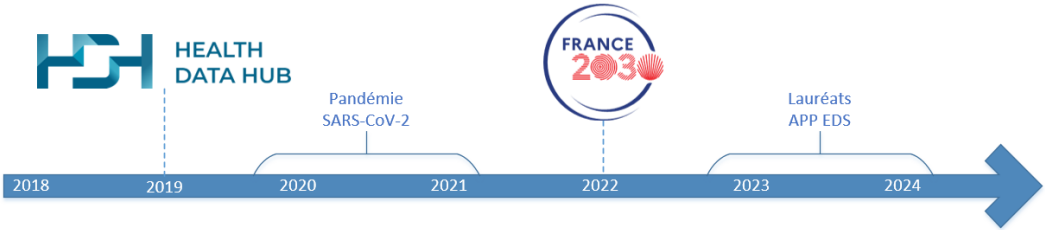
1 Juillet 2024

Manal Ahikki[1], Marc Berard[2], Camille Bouin[2], Antoine Boutet[3], Stéphane Breant[4], Alice Calliger[4], Ariel Cohen[4], Jean-François Couchot[5], Denis Delamarre[6], Caroline Dunoyer[1], Thibaut Fabacher[7], Lucas W. Gauthier[2], David Gimbert[2], Camille Girard-Chanudet[8], Romain Griffier[9], Martin Hilka[4], Yannick Jacob[4], Vianney Jouhet[9], David Laiymani[5], Leonardo Moros[1], Joris Muller[7], David Pellecuer[1], Thomas Petit-Jean[4], Antoine Richard[2], Maxime Salaun[7], François Talbot[2], Perceval Wajsburt[4] et Kevin Yauy[1]

1 2 3 4 5 6 7 8 9

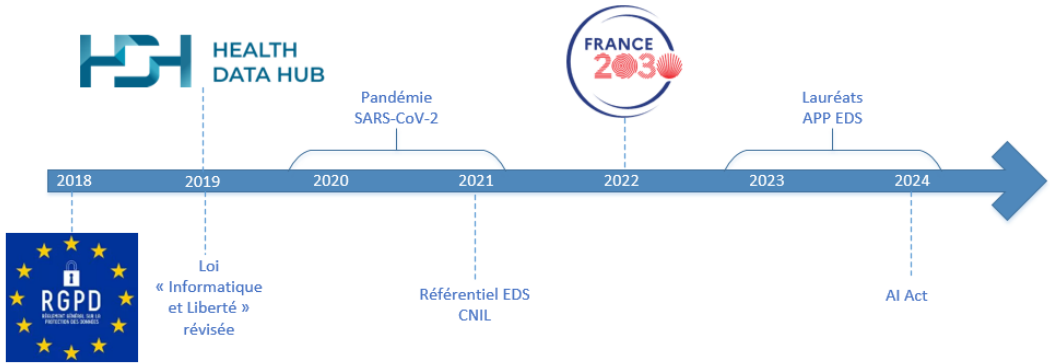
antoine.richard@chu-lyon.fr / joris.muller@chru-strasbourg.fr / perceval.wajsburt-ext@aphp.fr

Contexte [1] [2]



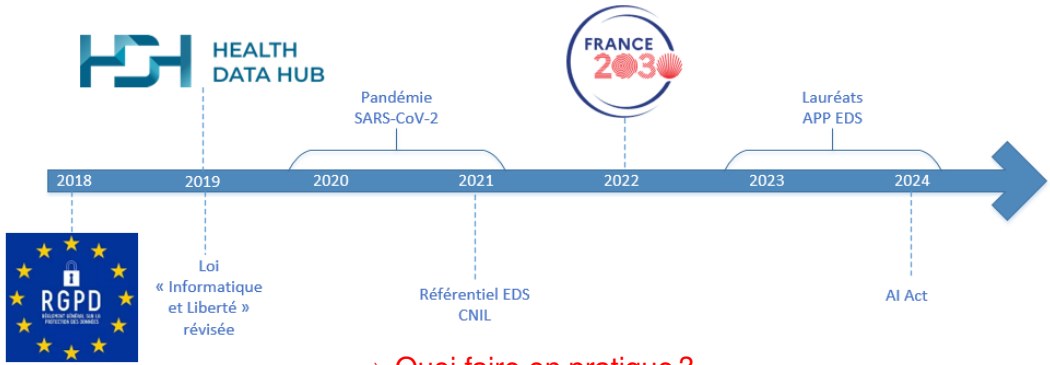
1. MINISTÈRE DU TRAVAIL, DE LA SANTÉ ET DES SOLIDARITÉS 2024 - Dossier de presse France 2030 : 2 ans de la Stratégie "Santé numérique"
2. MARCHAND-ARVIER et al. 2023 - Fédérer les acteurs de l'écosystème pour libérer l'utilisation secondaire des données de santé

Problématique [3] [4] [5] [6]



3. COMMISSION NATIONALE DE L'INFORMATIQUE ET DES LIBERTÉS (CNIL) 2016 - Le règlement général sur la protection des données - RGPD
4. Loi nř 78-17 du 6 janvier 1978 relative à l'informatique, aux fichiers et aux libertés
5. JOURNAL OFFICIEL DE LA RÉPUBLIQUE FRANÇAISE (JORF) 2021 - Délibération nř 2021-118 du 7 octobre 2021 sur le référentiel EDS
6. EUROPEAN UNION 2024 - Artificial Intelligence Act

Problématique [3] [4] [5] [6]



⇒ Quoi faire en pratique ?

3. COMMISSION NATIONALE DE L'INFORMATIQUE ET DES LIBERTÉS (CNIL) 2016 - Le règlement général sur la protection des données - RGPD
4. Loi nř 78-17 du 6 janvier 1978 relative à l'informatique, aux fichiers et aux libertés
5. JOURNAL OFFICIEL DE LA RÉPUBLIQUE FRANÇAISE (JORF) 2021 - Délibération nř 2021-118 du 7 octobre 2021 sur le référentiel EDS
6. EUROPEAN UNION 2024 - Artificial Intelligence Act

Objectifs

1. Établir quelles sont les méthodes de protection de la vie privée à notre disposition

Objectifs

1. Établir quelles sont les méthodes de protection de la vie privée à notre disposition

2. Déterminer comment intégrer ces méthodes dans un EDS

Objectifs

1. Établir quelles sont les méthodes de protection de la vie privée à notre disposition

2. Déterminer comment intégrer ces méthodes dans un EDS

3. Proposer des bonnes pratiques

Protection de la Vie Privée [7]

7. Images générée avec Stable Diffusion XL 1.0

Protection de la Vie Privée [7]

PVP "stricte"

Suppose un attaquant quasi-omniscient et quasi-omnipotent

7. Images générée avec Stable Diffusion XL 1.0

Protection de la Vie Privée [7]

PVP "stricte"

Suppose un attaquant quasi-omniscient et quasi-omnipotent

PVP "relaxée"

Suppose un investigateur honnête mais curieux

7. Images générée avec Stable Diffusion XL 1.0

Protection de la Vie Privée [7]

PVP "stricte"

Suppose un attaquant quasi-omniscient et quasi-omnipotent

⇒ Maximiser le temps d'attaque

PVP "relaxée"

Suppose un investigateur honnête mais curieux

7. Images générée avec Stable Diffusion XL 1.0

Protection de la Vie Privée [7]

PVP "stricte"

Suppose un attaquant quasi-omniscient et quasi-omnipotent

⇒ Maximiser le temps d'attaque

PVP "relaxée"

Suppose un investigateur honnête mais curieux

⇒ Minimiser les risques

7. Images générée avec Stable Diffusion XL 1.0

Identifiants (in)directs et données sensibles [8]

Numéro de Sécurité Sociale	Nom	Prénom	Âge	Adresse	Diagnostic	Allergies	Médicaments
123-45-6789	Durand	Marie	45	123 Rue de Paris, 75001 Paris	Diabète	Arachides	Metformine
234-56-7890	Martin	Jean	58	45 Avenue des Champs, 75008 Paris	Hypertension	Pollen	Lisinopril
345-67-8901	Bernard	Clara	30	78 Boulevard Saint-Germain, 75006 Paris	Asthme	Aucun	Salbutamol
456-78-9012	Petit	Louis	67	22 Rue de Rivoli, 75004 Paris	Cancer	Latex	Paclitaxel
567-89-0123	Robert	Sophie	52	15 Rue de la Paix, 75002 Paris	Insuffisance cardiaque	Antibiotiques	Digoxine

8. Données factices générées avec ChatGPT 3.5

K-anonymité et l-diversité [9]

K-anonymité :

Impossible d'isoler des groupes de taille plus petit que *k* sur les identifiants indirects

Âge	Adresse	Diagnostic	Allergies	Médicaments
40-50	Paris	Diabète	Arachides	Metformine
40-50	Paris	Maladie rénale	Aucun	Énalapril
50-60	Paris	Insuffisance cardiaque	Antibiotiques	Digoxine
50-60	Paris	Dépression	Acarien	Sertraline
50-60	Paris	Cholestérol élevé	Aucun	Atorvastatine
50-60	Paris	Hypertension	Pollen	Lisinopril
30-40	Paris	Asthme	Aucun	Salbutamol
30-40	Paris	Arthrite	Poussière	Ibuprofène
60-70	Paris	Cancer	Latex	Paclitaxel
60-70	Paris	Alzheimer	Aucun	Donepezil

9. Exemple avec données factices générées avec ChatGPT 3.5

K-anonymité et l-diversité [9]

K-anonymité :

Impossible d'isoler des groupes de taille plus petit que *k* sur les identifiants indirects

L-diversité :

Avoir une diversité d'au moins *l* éléments différents pour chaque groupe sur les données sensibles

Âge	Adresse	Diagnostic	Allergies	Médicaments
40-50	Paris	Diabète	Arachides	Metformine
40-50	Paris	Maladie rénale	Aucun	Énalapril
50-60	Paris	Insuffisance cardiaque	Antibiotiques	Digoxine
50-60	Paris	Dépression	Acarien	Sertraline
50-60	Paris	Cholestérol élevé	Aucun	Atorvastatine
50-60	Paris	Hypertension	Pollen	Lisinopril
30-40	Paris	Asthme	Aucun	Salbutamol
30-40	Paris	Arthrite	Poussière	Ibuprofène
60-70	Paris	Cancer	Latex	Paclitaxel
60-70	Paris	Alzheimer	Aucun	Donepezil

9. Exemple avec données factices générées avec ChatGPT 3.5

Données à Caractère Personnel [10] [11]

Données à Caractère Personnel

Numéro de Sécurité Sociale	Nom	Prénom	Âge	Adresse	Diagnostic	Allergies	Médicaments
123-45-6789	Durand	Marie	45	123 Rue de Paris, 75001 Paris	Diabète	Arachides	Metformine
234-56-7890	Martin	Jean	58	45 Avenue des Champs, 75008 Paris	Hypertension	Pollen	Lisinopril
345-67-8901	Bernard	Clara	30	78 Boulevard Saint-Germain, 75006 Paris	Asthme	Aucun	Salbutamol
456-78-9012	Petit	Louis	67	22 Rue de Rivoli, 75004 Paris	Cancer	Latex	Paclitaxel
567-89-0123	Robert	Sophie	52	15 Rue de la Paix, 75002 Paris	Insuffisance cardiaque	Antibiotiques	Digoxine

10. Données factices générées avec ChatGPT 3.5
11. COMMISSION NATIONALE DE L'INFORMATIQUE ET DES LIBERTÉS (CNIL) 2016 - Le règlement général sur la protection des données - RGPD

Anonymisation [12]

12. (CNIL) 2020 - L'anonymisation de données personnelles

Anonymisation [12]

Risque d'individualisation :

Il ne doit pas être possible d'isoler un individu dans un jeu de données

12. (CNIL) 2020 - L'anonymisation de données personnelles

Anonymisation [12]

Risque d'individualisation :

Il ne doit pas être possible d'isoler un individu dans un jeu de données

Risque de corrélation :

Il ne doit pas être possible de relier entre eux des ensembles de données distincts concernant un même individu

12. (CNIL) 2020 - L'anonymisation de données personnelles

Anonymisation [12]

Risque d'individualisation :

Il ne doit pas être possible d'isoler un individu dans un jeu de données

Risque de corrélation :

Il ne doit pas être possible de relier entre eux des ensembles de données distincts concernant un même individu

Risque d'inférence :

Il ne doit pas être possible de déduire, de façon quasi certaine, de nouvelles informations sur un individu

12. (CNIL) 2020 - L'anonymisation de données personnelles

Anonymisation [12]

Risque d'individualisation :

Il ne doit pas être possible d'isoler un individu dans un jeu de données

Risque de corrélation :

Il ne doit pas être possible de relier entre eux des ensembles de données distincts concernant un même individu

Risque d'inférence :

Il ne doit pas être possible de déduire, de façon quasi certaine, de nouvelles informations sur un individu

$\Rightarrow$ La RGPD ne s'applique pas sur un jeu de données "anonymisé"

12. (CNIL) 2020 - L'anonymisation de données personnelles

Pseudonymisation [13] [14] [15]

Définition :

"Traitement de données à caractère personnel de telle façon que celles-ci ne puissent plus être attribuées à une personne concernée précise sans avoir recours à des informations supplémentaires"

13. (CNIL) 2016 - Le G29 publie un avis sur les techniques d'anonymisation
14. (CNIL) 2020 - L'anonymisation de données personnelles
15. Where's Wally ? 4,626 people dressed as Waldo break a record in Japan

Pseudonymisation [13] [14] [15]

Définition :

"Traitement de données à caractère personnel de telle façon que celles-ci ne puissent plus être attribuées à une personne concernée précise sans avoir recours à des informations supplémentaires"

$\Rightarrow$ "Cacher dans la foule"

13. (CNIL) 2016 - Le G29 publie un avis sur les techniques d'anonymisation
14. (CNIL) 2020 - L'anonymisation de données personnelles
15. Where's Wally ? 4,626 people dressed as Waldo break a record in Japan

Pseudonymisation [13] [14] [15]

Définition :

"Traitement de données à caractère personnel de telle façon que celles-ci ne puissent plus être attribuées à une personne concernée précise sans avoir recours à des informations supplémentaires"

$\Rightarrow$ "Cacher dans la foule"

$\Rightarrow$ Modification des données

13. (CNIL) 2016 - Le G29 publie un avis sur les techniques d'anonymisation
14. (CNIL) 2020 - L'anonymisation de données personnelles
15. Where's Wally ? 4,626 people dressed as Waldo break a record in Japan

Pseudonymisation [13] [14] [15]

Définition :

"Traitement de données à caractère personnel de telle façon que celles-ci ne puissent plus être attribuées à une personne concernée précise sans avoir recours à des informations supplémentaires"

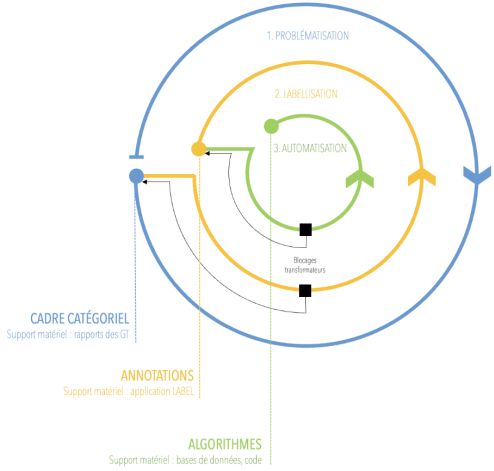
$\Rightarrow$ "Cacher dans la foule"

$\Rightarrow$ Modification des données

$\Rightarrow$ Processus réversible

13. (CNIL) 2016 - Le G29 publie un avis sur les techniques d'anonymisation
14. (CNIL) 2020 - L'anonymisation de données personnelles
15. Where's Wally ? 4,626 people dressed as Waldo break a record in Japan

Processus annotation [16]



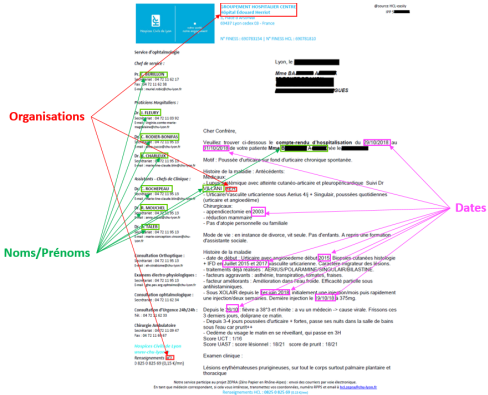
16. GIRARD-CHANUDET 2023 - La justice algorithmique en chantier : sociologie du travail et des infrastructures de l'intelligence artificielle

Données Structurées [17] [18]

Numéro de Sécurité Sociale	Nom	Prénom	Âge	Adresse	Diagnostic	Allergies	Médicaments
123-45-6789	Durand	Marie	45	123 Rue de Paris, 75001 Paris	Diabète	Arachides	Metformine
234-56-7890	Martin	Jean	58	45 Avenue des Champs, 75008 Paris	Hypertension	Pollen	Lisinopril
345-67-8901	Bernard	Clara	30	78 Boulevard Saint-Germain, 75006 Paris	Asthme	Aucun	Salbutamol
456-78-9012	Petit	Louis	67	22 Rue de Rivoli, 75004 Paris	Cancer	Latex	Paclitaxel
567-89-0123	Robert	Sophie	52	15 Rue de la Paix, 75002 Paris	Insuffisance cardiaque	Antibiotiques	Digoxine

17. Données factices générées avec ChatGPT
18. ARX - Data Anonymization Tool

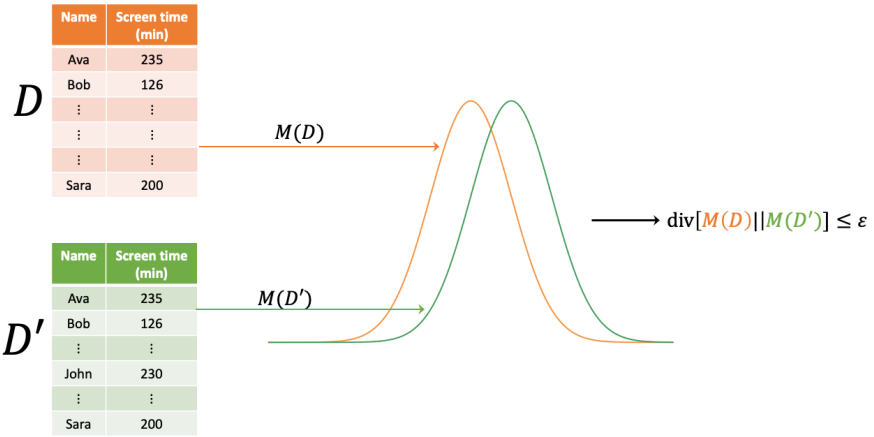
Données Non-structurées



Reconnaissance d'entités nommées via :

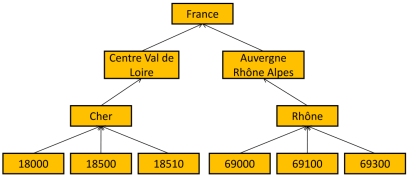
- Systèmes de règles GROUIN 2013
- Réseaux de neurones RICHARD, TALBOT et GIMBERT 2023
- Systèmes mixtes TCHOUKA 2023 ; TANNIER et al. 2024

Randomisation [19] [20] [21]

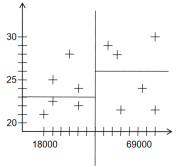


19. (CNIL) 2016 - Le G29 publie un avis sur les techniques d'anonymisation
20. (CNIL) 2020 - L'anonymisation de données personnelles
21. The ABCs of Differential Privacy

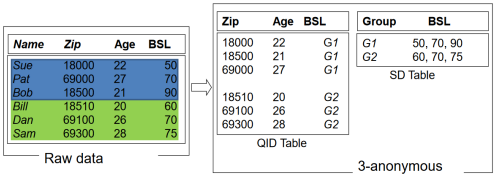
Généralisation [22] [23]



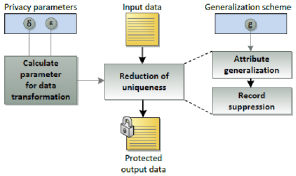
Generalization Algorithm (SWEENEY 2002)



Mondrian Algorithm (LEFEVRE, DEWITT et RAMAKRISHNAN 2006)



Bucketization (XIAO et TAO 2008)

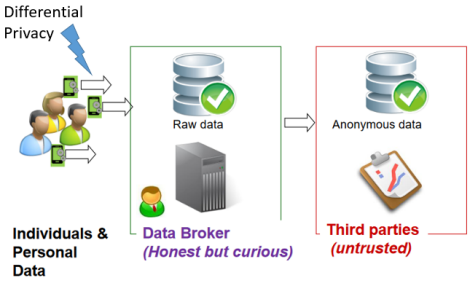


ARX : SafePub Algorithm (BILD, KUHN et PRASSER 2018)

22. (CNIL) 2016 - Le G29 publie un avis sur les techniques d'anonymisation
23. (CNIL) 2020 - L'anonymisation de données personnelles

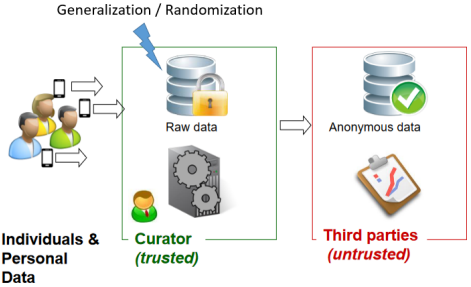
En théorie [24]

PVP "décentralisée"



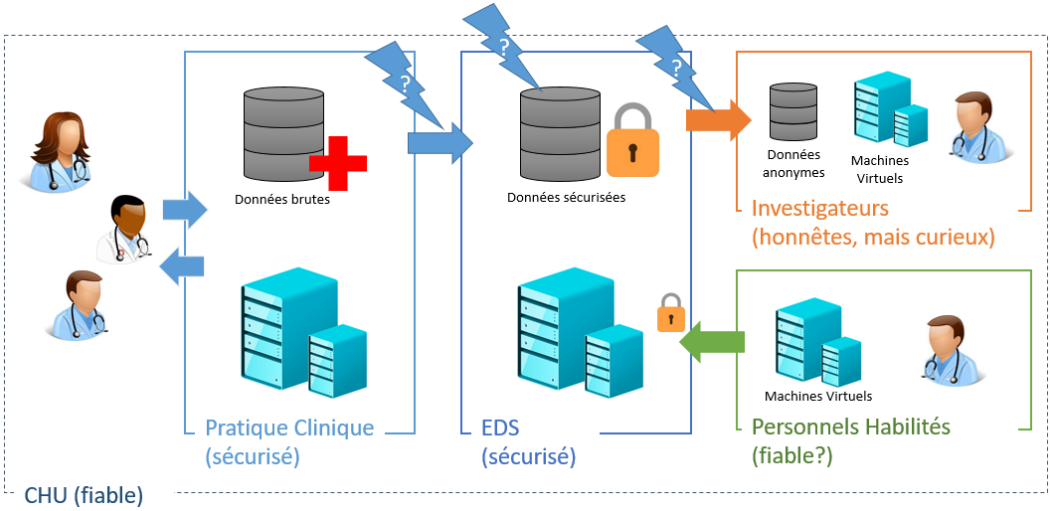
VS

PVP "centralisée"



24. Benjamin Nguyen - Privacy, Data Protection and Security

En pratique


Données brutes
Données sécurisées
Données anonymes
Machines Virtuels
Investigateurs
(honnêtes, mais curieux)
Machines Virtuels
Pratique Clinique
(sécurisé)
EDS
(sécurisé)
Personnels Habilités
(fiable?)
CHU (fiable)

Pour les travaux en IA :

Pour les travaux en IA :

- Contactez votre DPO

Pour les travaux en IA :

- Contactez votre DPO
- Déterminez les données, la granularité et le bruit

Pour les travaux en IA :

- Contactez votre DPO
- Déterminez les données, la granularité et le bruit
- Évaluez les risques

Pour les travaux en IA :

- Contactez votre DPO
- Déterminez les données, la granularité et le bruit
- Évaluez les risques

$\Rightarrow$ Faire au mieux pour trouver un équilibre entre protection des données pour les patients et utilité des données pour les investigateurs

Pour les travaux en IA :

- Contactez votre DPO
- Déterminez les données, la granularité et le bruit
- Évaluez les risques

$\Rightarrow$ Faire au mieux pour trouver un équilibre entre protection des données pour les patients et utilité des données pour les investigateurs

Pour les travaux sur les EDS :

Pour les travaux en IA :

- Contactez votre DPO
- Déterminez les données, la granularité et le bruit
- Évaluez les risques

$\Rightarrow$ Faire au mieux pour trouver un équilibre entre protection des données pour les patients et utilité des données pour les investigateurs

Pour les travaux sur les EDS :

- Les approches mixtes sont les plus sûre

Pour les travaux en IA :

- Contactez votre DPO
- Déterminez les données, la granularité et le bruit
- Évaluez les risques

Pour les travaux sur les EDS :

- Les approches mixtes sont les plus sûre
- Supprimez les identifiants directs par défauts

$\Rightarrow$ Faire au mieux pour trouver un équilibre entre protection des données pour les patients et utilité des données pour les investigateurs

Pour les travaux en IA :

- Contactez votre DPO
- Déterminez les données, la granularité et le bruit
- Évaluez les risques

Pour les travaux sur les EDS :

- Les approches mixtes sont les plus sûre
- Supprimez les identifiants directs par défauts
- Permettre d'ajuster la protection des données aux besoins

$\Rightarrow$ Faire au mieux pour trouver un équilibre entre protection des données pour les patients et utilité des données pour les investigateurs

Pour les travaux en IA :

- Contactez votre DPO
- Déterminez les données, la granularité et le bruit
- Évaluez les risques

Pour les travaux sur les EDS :

- Les approches mixtes sont les plus sûre
- Supprimez les identifiants directs par défauts
- Permettre d'ajuster la protection des données aux besoins

$\Rightarrow$ Faire au mieux pour trouver un équilibre entre protection des données pour les patients et utilité des données pour les investigateurs

Perspectives :

- Masquage des images, sons, etc.
- Référentiel commun
- Sûreté des modèles

Pour les travaux en IA :

- Contactez votre DPO
- Déterminez les données, la granularité et le bruit
- Évaluez les risques

$\Rightarrow$ Faire au mieux pour trouver un équilibre entre protection des données pour les patients et utilité des données pour les investigateurs

Pour les travaux sur les EDS :

- Les approches mixtes sont les plus sûre
- Supprimez les identifiants directs par défauts
- Permettre d'ajuster la protection des données aux besoins

Perspectives :

- Masquage des images, sons, etc.
- Référentiel commun
- Sûreté des modèles

Merci pour votre attention :)

Références I

- BILD, Raffael, Klaus A KUHN et Fabian PRASSER (2018). “SafePub : A Truthful Data Anonymization Algorithm With Strong Privacy Guarantees.”. In : **Proc. Priv. Enhancing Technol.** 2018.1, p. 67-87. DOI : 10.1515/popets-2018-0004.
- (CNIL), Commission nationale de l’informatique et des libertés (jan. 2016). **Le G29 publie un avis sur les techniques danonymisation**. fr. URL : https://www.cnil.fr/fr/le-g29-publie-un-avis-sur-les-techniques-danonymisation.
- — (mai 2020). **Lanonymisation de données personnelles**. fr. URL : https://www.cnil.fr/fr/lanonymisation-de-donnees-personnelles.
- COMMISSION NATIONALE DE L’INFORMATIQUE ET DES LIBERTÉS (CNIL) (mai 2016). **Le règlement général sur la protection des données - RGPD**. fr. URL : https://www.cnil.fr/fr/reglement-europeen-protection-donnees.
- EUROPEAN UNION (mars 2024). **EU Artificial Intelligence Act**. URL : https://www.aiact-info.eu/articles/.
- GIRARD-CHANUDET, Camille (déc. 2023). “La justice algorithmique en chantier : sociologie du travail et des infrastructures de l’intelligence artificielle.”. These de doctorat. Paris, EHESS. URL : https://theses.fr/2023EHES0141.

Références II

- GROUIN, Cyril (juin 2013). “Anonymisation de documents cliniques : performances et limites des méthodes symboliques et par apprentissage statistique”. fr. Thèse de doct. Université Pierre et Marie Curie - Paris VI. URL : https://theses.hal.science/tel-00848672 (visité le 29/03/2024).
- JOURNAL OFFICIEL DE LA RÉPUBLIQUE FRANÇAISE (JORF) (oct. 2021). **Délibération nř 2021-118 du 7 octobre 2021 portant adoption d’un référentiel relatif aux traitements de données à caractère personnel mis en uvre à des fins de création d’entrepôts de données dans le domaine de la santé - Légifrance**. URL : https://www.legifrance.gouv.fr/jorf/id/JORFTEXT000044239566.
- LEFEVRE, Kristen, David J DEWITT et Raghu RAMAKRISHNAN (2006). “Mondrian multidimensional k-anonymity”. In : **22nd International conference on data engineering (ICDE’06)**. IEEE, p. 25-25. DOI : 10.1109/ICDE.2006.101.
- MARCHAND-ARVIER, Jérôme et al. (déc. 2023). **Fédérer les acteurs de lécosystème pour libérer lutilisation secondaire des données de santé**. URL : https://sante.gouv.fr/IMG/pdf/rapport_donnees_de_sante.pdf.
- MINISTÈRE DU TRAVAIL, DE LA SANTÉ ET DES SOLIDARITÉS (jan. 2024). **Dossier de presse France 2030 : 2 ans de la Stratégie "Santé numérique"**. URL : https://sante.gouv.fr/IMG/pdf/dp_2ans_sasn_18janvier2024.pdf.

Références III

RICHARD, Antoine, François TALBOT et David GIMBERT (juill. 2023). “Anonymisation de documents médicaux en texte libre et en français via réseaux de neurones”. In : **Plate-forme Intelligence Artificielle 2023 (PFIA2023) - Journée Santé & IA**. Starsbourg, France : Association française pour l’Intelligence Artificielle (AfIA), Université de Strasbourg et Association française d’Informatique Médicale (AIM). URL : https://hal.science/hal-04139391.

SWEENEY, Latanya (2002). “k-anonymity : A model for protecting privacy”. In : **International journal of uncertainty, fuzziness and knowledge-based systems** 10.05, p. 557-570. DOI : 10.1142/S0218488502001648.

TANNIER, Xavier et al. (mars 2024). “Development and Validation of a Natural Language Processing Algorithm to Pseudonymize Documents in the Context of a Clinical Data Warehouse”. en. In : **Methods of Information in Medicine**. Publisher : Georg Thieme Verlag KG. ISSN : 0026-1270, 2511-705X. DOI : 10.1055/s-0044-1778693. URL : http://www.thieme-connect.de/DOI/DOI?10.1055/s-0044-1778693.

TCHOUKA, Yakini (déc. 2023). “Dé-identification des comptes rendus médicaux pour les tâches d’apprentissage automatique : application à l’association des codes CIM-10”. fr. thesis. Bourgogne Franche-Comté. URL : https://theses.fr/s257929.

Références IV

XIAO, Xiaokui et Yufei TAO (2008). “Dynamic anonymization : Accurate statistical analysis with privacy preservation”. In : **Proceedings of the 2008 ACM SIGMOD international conference on Management of data**, p. 107-120. DOI : 10.1145/1376616.1376630.