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

Manal Ahikki, Marc Berard, Camille Bouin, Antoine Boutet, Stéphane Breant, Alice Calliger, Ariel Cohen, Jean-François Couchot, Denis Delamarre, Caroline Dunoyer, et al.

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

M. Ahikki[1], M. Berard[2], C. Bouin[2], A. Boutet[3], S. Breant[4], A. Calliger[5], A. Cohen[5], J.F. Couchot[6], D. Delamarre[7], C. Dunoyer[1], T. Fabacher[8], LW. Gauthier[9], D. Gimbert[2], C. Girard-Chanudet[10], R. Griffier[11], M. Hilka[5], Y. Jacob[5], V. Jouhet[11], D. Laiymani[6], L. Moros[1], J. Muller[8], D. Pellecuer[1], T. Petit-Jean[5], A. Richard[2], M. Salaun[8], F. Talbot[2], P. Wajsburt[5], K. Yauy[1]

[1] Health Data Science Unit, Public Health Service, CHU de Montpellier, Montpellier, France
[2] DSN Bron, Hospices Civil de Lyon, 61 boulevard Pinel, 69672 Bron, France
[3] Université de Lyon, INSA Lyon, Inria, CITI Lyon, France
[4] DSI WIND, AP-HP, Paris, France
[5] Innovation and Data Unit, IT Department, Assistance Publique-Hôpitaux de Paris, Paris, France
[6] FEMTO-ST Institute, CNRS, Université de Franche-Comté, Besançon, France
[7] INSERM U936, Université Rennes 1, IFR 140, Rennes, France
[8] Department of Public Health, Hôpitaux Universitaires de Strasbourg, 67091 Strasbourg, France
[9] Hospices Civils de Lyon, Service de Génétique -
Centre de Référence Anomalies du Développement, Bron, France.
[10] Centre d'étude des mouvements sociaux (CEMS), 4 boulevard Raspail 75006 Paris, France
[11] Pôle de Santé Publique, Service d'Information Médicale, CHU de Bordeaux, Bordeaux, France

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

Résumé

Ces dernières années, la mise en chantier de différents EDS a fait émerger des discussions, entre divers acteurs travaillant sur ces EDS, concernant la protection de la vie privée des patients. Cet article présente une synthèse des points abordés durant ces discussions. Nous y argumentons que les définitions législatives offrent une base de travail solide. Nous concluons que la multiplicité et la mixité des méthodes offrent la meilleure protection de la vie privée pour les patients, même si celles-ci doivent s'adapter en fonction des besoins des investigateurs.

Mots-clés

EDS, Protection de la Vie Privée, TAL.

Abstract

During the last years, the construction of Health Data Warehouses in France raised discussions about Patient Privacy. This article proposes a synthesis of the points covered during these discussions. We argue that legal definitions offer a strong base of work. We conclude that the diversity and the multiplicity of the methods used allow the best Patient Privacy, even if the use of these methods must be adapted according to the needs of investigators.

Keywords

Health Data Warehouse, Patient Privacy, NLP.

1 Introduction

Ces dernières années, le domaine du Machine Learning (ML), et plus spécifiquement du Deep Learning (DL), est en pleine effervescence. Les techniques de ML sont employé dans de nombreux domaines, dont celui de la santé. En parallèle, le besoin de partager des données de santé à large échelle s'est fait ressentir afin de créer des modèles prédictifs plus performants, qu'ils soient ou non basés sur du ML.

En ce sens, l'État français a mis en chantier un Entrepôt de Données de Santé (EDS) d'échelle nationale et appelé Health Data Hub [6]. Suivi ensuite par divers appels à projet pour la mise en chantier de plusieurs EDS à une échelle régionale [7].

Cependant, les EDS et les données de santé qu'ils hébergent sont soumis aux législations européennes, via la RGPD [9] et l'AIAct [1], et françaises, via la loi Informatique et Liberté [2] et le référentiel EDS de la CNIL [10].

1. https://www.aiact-info.eu/articles/
2. https://www.legifrance.gouv.fr/loda/article_lc/LEGIARTI000038888757

Des délibérations ont alors été menées afin d'établir comment stocker et partager des données de santé par l'intermédiaire des EDS sans porter atteinte à la vie privée des patients [3]. Ces délibérations s'appuient en particulier sur les travaux du groupe de travail de l'article 29 sur l'anonymisation des données de santé [8, 11].

Néanmoins, ces délibérations restent éloignées de la pratique et n'établissent pas de bonnes pratiques concrètent sur comment intégrer, dans les EDS, des méthodes permettant de protéger la vie privée des patients. Ainsi, des discussions ont été entamées entre plusieurs acteurs travaillant sur les EDS de leur Centre Hospitalier Universitaire (CHU) respectif. Ces discussions ont eu, et ont toujours, pour but de comparer les approches employées par chacun afin de traiter la problématique de la protection de la vie privée dans les EDS. Cela afin de dessiner les contours de cette problématique, d'identifier les points à éclaircir, et de dégager des bonnes pratiques.

Ce papier propose une synthèse de l'état actuel de nos discussions et est construit comme suit. En section 2, nous définissons les concepts et termes que nous utilisons lors de nos discussions. Nous y argumentons que les définitions proposées par les textes législatifs offrent une base de travail plus solide et plus clair que les définitions proposées dans la littérature scientifique. En section 3, nous détaillons les étapes et les différents cas de figures à prendre en compte pour la protection des données de santé et de la vie privée des patients. Enfin, en section 4, nous abordons les différentes approches possibles pour intégrer, dans les EDS, des méthodes pour protéger la vie privée des patients. Cela avant de conclure, en section 5, que la multiplicité et la mixité des méthodes permet de maximiser la protection des patients tout en s'adaptant aux besoins des investigateurs.

2 Concepts

Dans la littérature, scientifique comme législative, de nombreux concepts tels que "anonymisation", "pseudonymisation", ou encore "dé-identification" gravitent autour de la protection des données et de la vie privée. Il nous est apparu alors important, dans un premier temps, de définir au mieux ces concepts et, ainsi, pouvoir échanger autour de ces sujets. Pour commencer, deux positionnements ont émergé de nos discussions sur le concept de "Protection de la Vie Privée" (PVP).

Le premier positionnement, que l'on pourrait qualifier de PVP "stricte", se base sur un scénario impliquant un attaquant ayant à sa disposition toutes les informations et tous les moyens possibles et imaginables pour lever la protection sur une ou plusieurs données. Si cet attaquant ne peut, malgré ses informations et moyens, lever la protection d'une donnée, cette donnée est considérée comme protégé au sens "strict" du terme.

Le second positionnement, que l'on pourrait qualifier de plus "relaxée" ou "réaliste", se base un scénario impliquant un "investigateur curieux", un attaquant ayant à sa disposition *a minima* des connaissances sur la manière dont sont protégées les données et des moyens d'accéder aux données réelles des patients. Si cet attaquant ne peut, malgré ses informations, moyens et connaissances, lever la protection d'une donnée, cette donnée est considérée comme protégé au sens "relaxé" du terme.

À noter que chacun de ces positionnements n'implique pas une PVP plus simple ou plus complexe que pour l'autre. Ils ne sont pas non plus en opposition.

Enfin, concernant les concepts gravitant autour de la PVP, leurs définitions vont différer en fonction du domaine dans lequel ils sont présentés.

2.1 Littérature scientifique

Dans la littérature scientifique, plus spécifiquement en informatique et science de l'ingénierie, la PVP repose principalement sur les concepts d'identifiants "directs" et "indirects" [12]. Ces concepts s'ajoutent à celui de "données sensibles" qui, si associées à un individu, impliquent une atteinte à sa vie privée (ex. médicaments, antécédents médicaux, résultats d'examens, etc.).

Un identifiant "direct" est une donnée qui réfère directement à une personne physique ou morale, tels que les noms et prénoms d'une personne, le nom d'une organisation, un numéro de sécurité sociale, ou encore le numéro SIRET d'une entreprise.

Un identifiant "indirect" (ou quasi-identifiant) est, quant à lui, une donnée qui, en l'associant à d'autre donnée ou divers moyens, peut permettre de remonter à une personne physique ou morale. Tels qu'un numéro de téléphone, une adresse, ou encore une date de consultation.

À noter que le caractère "direct" ou "indirect" d'un identifiant va varier en fonction de la personne physique ou morale que l'on cherche à protéger. Par exemple, le nom d'un hôpital est considéré comme un identifiant "direct" dudit hôpital, mais un identifiant "indirect" d'un patient hospitalisé dans cet hôpital. De plus, une donnée peut à la fois être considérée comme "sensible" et "identifiante" (ex. une date de consultation).

Il existe cependant, dans cette littérature, un manque de clarté concernant les termes "anonymisation", "pseudonymisation" et "dé- identification". Tantôt bien distincts, tantôt synonymes [2]. La "dé-identification", par exemple, peut autant se référer à la suppression des identifiants direct qu'à la modification de ceux-ci (synonyme alors de "pseudonymisation"). Il apparaît alors intéressant de se baser sur les définition juridique et législatives de ces concepts.

À noter tout de même que, concernant la protection de jeux de données, il existe des notions bien définies telles que la k- anonymité et la l-diversité [12]. Un jeu de données est considéré "k-anonymisé" si il est impossible d'isoler des groupes d'individu de taille inférieur à k. Et, un jeu de données est considéré "l-diversifié" si pour chaque groupe équivalent d'individu il existe une diversité de l valeurs différentes pour chaque donnée sensible. L'objectif général de ces concepts étant de "cacher l'individu dans la foule".

2.2 Textes législatifs

Pour les définitions législatives, nous nous baserons principalement sur les textes de lois fournis par l'UE, le gouver-

nement français et la CNIL [3, 8, 9, 11], puisque les EDS mis en place en France sont soumis à ce cadre juridique.
Ces textes de loi se basent principalement sur la notion de "Données à Caractère Personnel" (DCP) définies comme : "toute information se rapportant à une personne physique identifiée ou identifiable" [9]. Cette notion englobe alors les concepts d'identifiants "direct" et "indirect", ainsi que celui de "données sensibles", vu en section 2.1.
Selon la CNIL et le rapport du G29 [8, 11], un jeu de données va être considéré comme "Anonyme" (ou "Anonymisé") si il respecte les trois conditions suivantes : (1) l'impossibilité d'isoler un individu, (2) l'impossibilité de relier le jeu de données à d'autres jeux de données pour un même individu et (3) l'impossibilité d'inférer de nouvelles informations sur un individu à partir des données présentes dans le jeu de données.
Ensuite, la pseudonymisation est définie comme : "le 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" [8, 11]. À noter que la RGPD impose la mise en place d'une table de correspondance pour lever la pseudonymisation en cas de besoin (ex. rétraction d'un patient sur son consentement à l'utilisation de ses données) [9].
La principale distinction faite par la RGPD entre "anonymisation" et "pseudonymisation" est la réversibilité des processus. Le premier étant considéré comme non-réversible, tandis que le second l'est. Ainsi, nous pouvons noter que la RGPD ne s'applique que par rapport à des jeux de données "pseudonymisés", puisque réversible et toujours à risque, et non sur les jeux de données "anonymisées", puisque non réversible est théoriquement impossible de remonter à un individu.
Enfin, la CNIL emploie généralement le terme de "masquage" des données à caractère personnel pour parler d'anonymisation et de pseudonymisation de manière indifférenciée. Nous ferons de même dans cet article.

3 Processus et méthodes de masquage

Au cours de nos discussions, le processus de masquage d'éléments identifiant nous est apparu comme relativement clair en comparaison du travail de définition présenté en section 2. Celui-ci commence par la formalisation des éléments que l'on souhaite masquer dans un jeu de données, puis par la détection de ces éléments dans les jeux de données, et enfin par le masquage de ces dits éléments. Des variations apparaissent cependant dans le choix des méthodes employées à chacune de ces étapes.

3.1 Définition et étiquetage des éléments identifiants

Avant toutes modifications des données pour en masquer les éléments identifiant, l'étape initiale consiste à déterminer quels sont les éléments que l'on peut considérer comme identifiant dans un jeu de données. Dans le domaine de la santé, la RGPD reste floue sur les éléments à masquer et la loi Health Insurance Portability and Accountability Act (HIPAA) promulguée aux États-Unis d'Amérique se révèle être une base de travail pour différents travaux [14, 15]. l'HIPAA définit 18 éléments identifiant pour les données de santé [3], allant des données nominatives (nom, prénom, numéro de sécurité sociales, etc.), aux données temporelles (dates) et données géographiques (adresses, codes postaux, villes, pays, etc.), en passant par les numéros de téléphone ou les courriels.
Cependant, ce cadre formel, s'il est adapté à une majorité de cas, peut se révéler limité pour traiter des concepts "atypiques" (ex. noms de terres en kanaky/nouvelle-calédonie, un lieu portant le nom du prioritaire de ce lieu) ou mixte (ex. personnes physico-morale, désignant à la fois ce lieu et une personne morale) [4].
Des boucles de rétroactions peuvent alors s'opérer lors de l'analyse et l'étiquetage des éléments identifiants dans les jeux de données à protéger. Ce processus a pu, par exemple, être étudié dans le domaine juridique pour la numérisation des décisions de justice [4].

3.2 Détection des éléments identifiants

Une fois déterminés les éléments identifiants que l'on souhaite masquer dans un jeu de données, il nous faut détecter ces éléments dans les jeux de données concrets. La manière dont sont structurées ces données va impacter la méthode de détection à utiliser.

3.2.1 Données structurées

Dans le cas de jeux de données dits structurés, on peut se ramener à une représentation dite "tabulaire", avec pour chaque colonne un type de données bien spécifique.
L'objectif est alors d'identifier quelles colonnes de ces jeux de données contiennent des éléments identifiants. Des logiciels, tel que ARX [4] par exemple, permettent d'identifier les colonnes contenant des données à caractère personnel. Ces logiciels permettent aussi d'appliquer des modifications spécifiques à chacune de ces colonnes.
Cependant, il peut arriver que certaines colonnes contiennent des données non structurées, tels que du texte libre ou des images.

3.2.2 Données non-structurées

Dans le domaine de la santé, un des principaux défis est d'être capable de gérer les données non-structurées, tel que les documents médicaux en texte libre [1]. Ainsi, un certain nombre de travaux se sont appuyés sur des méthodes issues du traitement automatique du langage (TAL) pour repérer la présence, ou non, de DCP dans ces documents.
Plus spécifiquement, la détection de ces données peut se formaliser sous la forme d'un problème de reconnaissance d'entités nommées (NER).
Pour cela plusieurs méthodes existent, telles que l'utilisation d'expression régulière, de méthodes statistiques, d'apprentissage machine [13], ou encore de logiciels dédiés tel que MEDINA [5].

3. https://cphs.berkeley.edu/hipaa/hipaa18.html
4. https://arx.deidentifier.org/

Bien entendu, ces méthodes ne sont pas mutuellement exclusives. Au contraire, la combinaison de plusieurs de ces méthodes est ce qui semble fournir les meilleurs résultats [14, 15].

3.3 Masquage des éléments identifiants

Une fois les DCP détectées dans un jeu de données, il nous faut les masquer. Pour cela, plusieurs techniques existent, telles que la Differential Privacy (DP), la généralisation, la "bucketisation", ou encore l'algorithme de Mondrian [12]. À noter que la CNIL recense deux approches possibles pour le masquage des données : la généralisation et la "randomisation" [11]. Cependant, les méthodes à employer vont varier selon le type de données que l'on cherche à masquer, ou plus précisément à pseudonymiser.

3.3.1 Données nominatives

Pour les données que l'on pourrait qualifier de "nominatives" (ex. noms et prénoms, numéros de sécurité sociale, courriels, etc.), la méthode généralement employée est le remplacement par une donnée factice. La principale difficulté étant de garder une cohérence dans les données modifiées lorsque cela est nécessaire. Par exemple, garder le même nom modifié pour un patient dans un document concernant son suivi.
De plus, dans le domaine de la santé, un certain nombre de maladies et de procédures médicales portent le nom de leur créateur. Ainsi, dans le masquage des données de santé, une autre difficulté est de ne pas masquer ces noms de maladie ou de procédure. Il nous faut donc des outils, à la fois extrêmement sensibles et extrêmement spécifiques, afin qu'ils soient capables de faire cette distinction.

3.3.2 Données géographiques

Concernant les données "géographiques" (ex., adresses, villes, régions, pays, etc.), plusieurs approches sont envisageables. En premier lieu, il est possible d'employer des méthodes dites de "généralisation" [11], qui consiste à regrouper des lieux selon une arborescence pré-définie afin d'augmenter la k- anonymité d'un jeu de données. Par exemple, regrouper les villes par départements, les départements par régions, etc. Cette méthode à l'avantage d'être simple à mettre en place, mais présente des effets de seuil pour les lieux à la frontières des zones pré- définis.
Une autre méthode possible est celle de la "randomisation" [11], en modifiant un lieu par un autre lieu proche géographiquement via l'usage, par exemple, d'algorithmes de Differential Privacy [15].
Cependant, ces deux méthodes peuvent masquer ou regrouper des lieux avec des caractéristiques épidémiologiques bien différentes. Il est néanmoins envisageable, dans certains cas de figure, de modifier un lieu par un autre lieu selon des métriques autres que la distance géographique [15].

3.3.3 Données temporelles

Enfin, concernant les données "temporelles" (ex. dates, années, âges, durées, etc.), il est possible d'appliquer les mêmes méthodes de "généralisation", en regroupant les dates par mois ou années, et de "randomisation", en appliquant un décalage de temps sur chaque date.
La principale difficulté pour la gestion des données temporelles étant de garder, si besoin, une cohérence dans la succession des événements et des intervalles de temps. De plus, un décalage indépendant de durées peut causer une permutation des évènement du patient, tandis qu'un décalage des intervalles conserve cet ordre.
Cependant, un point qui est ressorti lors de nos discussions est que les parcours des soins sont des données particulièrement identifiantes, même après masquage de ce type de données [15].

4 Intégration dans les EDS

En PVP, lorsqu'il est question d'intégrer les processus et méthodes de masquage vu en section 3, deux approches sont souvent opposées.
La première approche, que l'on peut nommer masquage "local" ou "décentralisé", consiste à masquer les données à la sortie des bases de données, avant le transfert vers l'entrepôt de données. Cette approche à l'avantage de limiter les risques lors du transfert. Cependant, elle nécessite l'utilisation de mécanismes de rétroaction pour lever le masquage dans certains cas de figure (ex. études de parcours de soins).
La seconde approche, que l'on peut nommer masquage "général" ou "centralisé", consiste à masquer les données dans l'entrepôt après transfert. Cette approche à l'avantage de fournir de traiter les données dans leur ensemble et donc de garder une bonne utilité après traitement. Néanmoins, elle nécessite une copie non masquée des données dans l'entrepôt et donc une forte confiance dans le tiers gérant l'entrepôt.
Dans le cas des EDS, en pratique, ce sont des approches "mixtes" ou "intermédiaires" qui sont employées. Celles-ci peuvent consister à masquer, a minima, les données les plus sensibles (ex. les identifiants "direct") localement, avant le transfert à l'EDS, et de masquer les données les moins sensibles au cas par cas dans l'EDS. Il est aussi possible de transférer les données dans une base à part dans l'EDS et d'en masquer le contenu au besoin. L'objectif ici trouver le juste équilibre entre données à masquer localement et de manière centralisé dans l'EDS.
Enfin, comme évoqué en section 2, il est important de noter que la RGPD impose l'utilisation d'une table de correspondance permettant de lever le masquage des données [9]. Cette table doit être stockée dans une base de données sécurisée, aux accès limités et à part des autres bases de données de l'EDS. Ainsi, si un patient souhaite faire valoir son droit à l'oubli ou simplement se rétracte concernant l'utilisation de ses données pour la recherche clinique, cette table permet de retrouver les données concernant ce patient et de les retirer de l'EDS.

5 Conclusion

Dans cet article, nous avons proposé une synthèse des échanges entre différents acteurs travaillant sur la problématique de la protection de la vie privée dans la mise en place de divers EDS. Chacun abordant ce sujet sous un angle différent, la multiplicité des points vue nous permet

de dessiner le contour de cette problématique, d'identifier des points à clarifier et de dégager des bonnes pratiques.
Dans un premier temps, nous avons cherché à définir notre objet d'étude. Il est rapidement apparu qu'une PVP au sens "strict" du terme était quasi-impossible à réaliser en réalité. L'objectif est donc de "faire au mieux" afin de protéger la vie privée des patients.
Il est aussi rapidement apparu que les définitions législatives liées à la PVP étaient plus standardisées que celles trouvées dans la littérature scientifique. Nous recommandons donc de se baser de préférence sur ces définitions.
Ensuite, bien que l'HIPAA fournit une liste d'éléments à masquer pour protéger la vie privée des patients, et donc une base de travail, cette liste a été faite pour un usage aux États-Unis d'Amérique et peut ne pas correspondre aux cadre de la RPGD dans l'UE. Des groupes travaillent cependant sur la standardisation d'une liste similaire adaptée au cadre français, voire européen.
Ensuite, il est apparu que la problématique de la PVP dans les EDS était avant tout une question d'équilibrage entre qualité du masquage et utilité des données masquées. Et que cet équilibrage va grandement dépendre des besoins de chaque étude menée sur des données. De plus, tant pour l'emploi de méthodes de masquage que pour l'intégration de ces méthodes dans les EDS, la multiplicité et la mixité des approches semble être ce qui conduit à une PVP de meilleure qualité.
Enfin, certaines questions restent encore en suspens telle que l'évaluation de la qualité du masquage d'un jeu de données ou encore le masquage dans des données non-structurées autres que du texte (ex. images, sons, etc.). Se pose aussi la question de si les données non-structurées, même masquées, ne constituent pas un risque pour la vie privée des patients (ex. tournure de phrases, présences de typographie spécifique, etc.).

Références

[1] Kiran Adnan, Rehan Akbar, Siak Wang Khor, and Adnan Bin Amanat Ali. Role and Challenges of Unstructured Big Data in Healthcare. In Neha Sharma, Amlan Chakrabarti, and Valentina Emilia Balas, editors, *Data Management, Analytics and Innovation*, Advances in Intelligent Systems and Computing, pages 301–323, Singapore, 2020. Springer.

[2] Raphaël Chevrier, Vasiliki Foufi, Christophe Gaudet-Blavignac, Arnaud Robert, and Christian Lovis. Use and Understanding of Anonymization and De-Identification in the Biomedical Literature : Scoping Review. *Journal of Medical Internet Research*, 21(5) :e13484, May 2019. Company : Journal of Medical Internet Research Distributor : Journal of Medical Internet Research Institution : Journal of Medical Internet Research Label : Journal of Medical Internet Research Publisher : JMIR Publications Inc., Toronto, Canada.

[3] Journal Officiel de la République Française (JORF). Délibération n° 2021-118 du 7 octobre 2021 portant adoption d'un référentiel relatif aux traitements de données à caractère personnel mis en œuvre à des fins de création d'entrepôts de données dans le domaine de la santé - Légifrance, October 2021.

[4] Camille Girard-Chanudet. *La justice algorithmique en chantier : sociologie du travail et des infrastructures de l'intelligence artificielle.* These de doctorat, Paris, EHESS, December 2023.

[5] Cyril Grouin. *Anonymisation de documents cliniques : performances et limites des méthodes symboliques et par apprentissage statistique*. phdthesis, Université Pierre et Marie Curie - Paris VI, June 2013.

[6] Jérôme Marchand-Arvier, Stéphanie Allassonniere, Aymeril Hoang, and Anne-Sophie Jannot. Fédérer les acteurs de l'écosystème pour libérer l'utilisation secondaire des données de santé, December 2023.

[7] de la Santé et des Solidarités Ministère du Travail. Dossier de presse France 2030 : 2 ans de la Stratégie "Santé numérique", January 2024.

[8] Commission nationale de l'informatique et des libertés (CNIL). Le G29 publie un avis sur les techniques d'anonymisation, January 2016.

[9] Commission nationale de l'informatique et des libertés (CNIL). Le règlement général sur la protection des données - RGPD, May 2016.

[10] Commission nationale de l'informatique et des libertés (CNIL). Le Comité européen de la protection des données (CEPD), July 2018.

[11] Commission nationale de l'informatique et des libertés (CNIL). L'anonymisation de données personnelles, May 2020.

[12] Iyiola E. Olatunji, Jens Rauch, Matthias Katzensteiner, and Megha Khosla. A Review of Anonymization for Healthcare Data. *Big Data*, March 2022. Publisher : Mary Ann Liebert, Inc., publishers.

[13] Antoine Richard, François Talbot, and David Gimbert. 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, July 2023. Association française pour l'Intelligence Artificielle (AfIA) and Université de Strasbourg and Association française d'Informatique Médicale (AIM).

[14] Xavier Tannier, Perceval Wajsbürt, Alice Calliger, Basile Dura, Alexandre Mouchet, Martin Hilka, and Romain Bey. Development and Validation of a Natural Language Processing Algorithm to Pseudonymize Documents in the Context of a Clinical Data Warehouse. *Methods of Information in Medicine*, March 2024. Publisher : Georg Thieme Verlag KG.

[15] Yakini Tchouka. *Dé-identification des comptes rendus médicaux pour les tâches d'apprentissage automatique : application à l'association des codes CIM-10*. thesis, Bourgogne Franche-Comté, December 2023.