Méthodologie d'anonymisation dès la conception d'un jeu de données en imagerie médicale

J. Clech[1,5], A. Gotlieb[2], F. Sève[3,5], F. Didout[4,5], P. Malléa[1,5]

[1] NEHS Digital, 1 rue Augustine Variot 92240 Malakoff, France
[2] Simula Research Laboratory, KA 23, 0164 Oslo, Norway
[3] Kalhyge, 4-6 rue Truillot 94200 Ivry sur Seine, France
[4] MNH, 331, avenue d'Antibes, 45200 Amilly, France
[5] groupe NEHS, 185, rue de Bercy 75012 Paris, France

jeremy.clech@groupe-nehs.com

Résumé

La recherche en santé s'appuie notamment sur des bases de données d'imagerie médicale. Les données personnelles qu'elles contiennent doivent être évacuées afin d'empêcher toute réidentification ultérieure des patients. Dans cet article, nous présentons notre méthodologie d'anonymisation de données d'imagerie médicale. Les leçons apprises dans cette expérience nous ont permis 1) de créer un premier outil qui peut anonymiser ce type d'imagerie et 2) de mettre à la disposition de la communauté IA cette base de données au travers de la plateforme européenne AI4Europe.

Mots-clés

Imagerie médicale, Apprentissage automatique, Qualité des données, Anonymisation, RGPD.

Abstract

Health research relies in particular on medical imaging databases. The personal data they contain must be removed in order to prevent any subsequent re-identification of patients. In this article, we present our methodology for anonymizing medical imaging data. The lessons learned in this experience allowed us 1) to create a first tool that can anonymize this type of imagery and 2) to make this database available to the AI community through the European platform AI4Europe.

Keywords

Medical imaging, Machine learning, Data quality, Anonymization, GDPR.

1 Introduction

Le Règlement Général sur la Protection des Données (RGPD) apporte à travers ses principes et ses obligations un cadre juridique pour l'exploitation de données à caractère personnel. L'objectif de ce papier est de proposer un retour d'expérience sur la mise en application d'un traitement d'anonymisation sur des données à caractère personnel liée à une base de données d'imagerie médicale : comment conjuguer les contraintes réglementaires, de structuration et mise en qualité des données collectées afin de pouvoir proposer en un temps court des volumes de données et les exploiter à des fins de recherche en IA.

L'imagerie médicale produit annuellement plusieurs dizaines de millions d'examens en France. Ainsi en 2019, 56,7 millions d'actes d'imageries médicales ont été réalisés par les radiologues libéraux[1] [1, p. 134]. Ces données, stockées en France dans des infrastructures informatiques sécurisées contre la violation de données (divulgation, perte et altération), sont un formidable vivier et sont sources d'innovations médicales afin de lutter contre la perte de chance d'un patient en réalisant trop tardivement des examens, d'améliorer la productivité des radiologues en garantissant la qualité et la constance du diagnostic ou encore de personnaliser les soins en fonction du patient, de ses souhaits et de son contexte.

Alors même que l'apprentissage automatique est arrivé à un haut niveau de maturité méthodologique et industrielle, l'accès à ces larges volumes de données est malaisé. En effet, tant en France qu'en Europe, leur accès requiert d'une part des déclarations et autorisations auprès des autorités compétentes (e.g. CNIL) et information auprès des patients et d'autres part que le recours à des sous-traitants (e.g. prestataires, intervenants) garantisse le bon respect des exigences du RGPD. Dès lors, accéder à ces bases de données requiert un investissement important sur les plans juridique et financier et sur un temps long. Les difficultés rencontrées par le gouvernement français pour la mise en place du *Health Data Hub*, avec la remise en cause de l'hébergeur de données de santé Microsoft en raison de risques de transferts de données vers les États-Unis [2], sont un exemple flagrant.

La pandémie du COVID-19 apporte un éclairage fort sur cette problématique : la Chine a pu proposer dès mars 2020 une solution de triage des patients à partir d'un jeu de données composé de 3 191 patients (dont 1 000 non atteints du COVID-19) [3] alors même que de nombreuses sociétés

[1] Ce nombre d'actes ne prend donc pas en compte ceux réalisés par les hôpitaux publics et ni ceux réalisés par les autres spécialités comme par exemple la cardiologie, la médecine nucléaire ou encore la gynécologie.

existent en France (startup et PME établies) en imagerie médicale, et que l'écosystème français maîtrise toute la chaîne de traitements de l'IA (laboratoires de recherche, centres de calculs…).

Nous souhaitons lever ce verrou à l'innovation en proposant un cadre méthodologique permettant un accès à des données de qualité aux acteurs de cet écosystème, dans des délais courts, de manière maîtrisée et respectueuse de la réglementation européenne. Ainsi, nos travaux visent à réaliser dès la conception du projet une collecte anonymisée de données. La finalité de cette démarche a permis de mettre à disposition une base de données d'imagerie médicale de forte volumétrie mais également une preuve de concept avec un outil d'anonymisation automatique de rapports radiologiques.

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 ce de manière irréversible. De fait, l'anonymisation permet de sortir du cadre du RGPD au prix d'une perte relative d'information puisque supprimant les données à caractère personnel. Comme rapidement décrit en section 2, de nombreuses méthodes existent afin d'altérer ces données en vue d'écarter la possibilité de réidentification ultérieure. Toutefois, ces dernières sont réalisées *a posteriori* de la collecte et peuvent introduire des biais.

Après présentation en section 3 des motivations et enjeux à la collecte d'une base de scanners thoraciques, nous décrivons en section 4 les éléments clefs du traitement portant sur des données anonymisées. Nous abordons en section iv la conception et les opérations mises en œuvre pour la collecte des données anonymisées. La section 6 est consacrée à la vérification de l'efficacité de l'anonymisation. La section 7 décrit les éléments mis à la disposition de la communauté IA tandis que la section 8 discute de la pertinence de l'approche proposée consistant à définir l'anonymisation des données lors de la définition du projet de recherche. Enfin, la section 9 conclue ce travail et propose quelques perspectives.

2 Rappel des travaux antérieurs

Le RGPD met en place des contraintes, telles que le recueil du consentement préalable des personnes, rendant parfois impossible l'exploitation des données et l'anonymisation est la seule méthode (lorsque ce recueil ne peut être envisagé) permettant de réaliser ces exploitations dans un cadre licite [4]. De nombreuses techniques, rappelées dans [5] ont été développées et existent aujourd'hui. Les approches classiques de suppression [6] et de généralisation [7] ou encore de recodage global [8] visent en premier objectif à supprimer le caractère personnel des données respectivement par suppression pure ou simple ou par réduction de l'espace des valeurs possibles afin de diminuer les valeurs singulières. Ces approches ont été enrichies afin de mieux intégrer la diversité des données dans ces données généralisées comme avec « *Anotomy* » [9] ou bien par des permutations d'une valeur au sein d'un groupe d'individus évalués comme similaire [10].

Ces techniques réalisées *a posteriori* de la collecte sont utiles en fonction des cas d'usage et des spécificités des données à anonymiser mais présentent plusieurs inconvénients comme i) la lenteur de réalisation, car les traitements imposent une intervention manuelle pour les prises de décision critique ; ii) l'impact sur les modèles d'apprentissage car une généralisation trop importante peut diminuer drastiquement la pertinence d'un modèle et que des permutations mal maîtrisées peuvent entraîner un biais d'apprentissage.

Pour contrer ces problématiques et partant du principe que les données ne peuvent pas être anonymisées tout en restant utiles, l'approche de la confidentialité différentielle[2] [11] propose de contourner l'obstacle en empêchant d'accéder directement à la donnée initiale. Ces solutions sont de plus en plus utilisées, notamment par les GAFAM[3], pour garantir la confidentialité. Cette approche est intéressante mais conduit à la production de biais potentiels liés à l'utilisation d'un générateur, qui rendent les approches d'apprentissage automatique inopérantes.

L'Apprentissage Fédéré[4] est utilisé sur les données de santé [12] car proposant également de ne pas accéder directement aux données en permettant de distribuer l'apprentissage sur les différents sites gérant les données et sous la responsabilité des DPD de chacun de ces sites. Cette approche offre un solide niveau de sécurité et de traçabilité des accès mais induit pour ce faire un prérequis technique non négligeable avec la mise en place d'une infrastructure sur chacun des sites concernés. Ceci génère 2 difficultés : une organisationnelle et une de ressources. En effet, il est nécessaire d'impliquer chacune des Direction des Systèmes Informatiques (DSI) des établissements participants afin qu'elles mettent à disposition les données dans une base distincte de celle de production, accorder de la puissance de calcul et autoriser et contrôler les flux.

Nous pensons que de définir dès la conception du traitement de collecte cet objectif clair d'anonymisation permet de minimiser l'utilisation de ces techniques ou du moins d'en avoir une meilleure maîtrise. C'est cette approche de la confidentialité par construction[5] pour les données d'imagerie médicale que nous présentons et défendons dans cet article

3 Genèse de FIDAC

Lors de la première vague de la pandémie de maladies liée à la propagation du coronavirus, il était difficile de déterminer rapidement si un patient était atteint du COVID-19 ou bien d'une autre maladie pulmonaire. Or, à ce stade limité de nos connaissances sur cette nouvelle maladie, il était crucial de pouvoir séparer les flux de patients (covid et non covid) au sein des établissements de santé pour réguler au mieux la propagation de l'épidémie. Toutefois, nous ne disposions ni de tests antigéniques, ni d'autotests permettant une réponse rapide. À cette période, seuls les tests PCR étaient disponibles mais ceux-ci étaient en nombre limité et rendent leur résultat sous 24h. Dans ce contexte, le scanner thoracique a beaucoup été employé car il permet de caractériser la pathologie du patient par la présence de signes radiologiques typiques, comme par exemple de « *crazy-paving pattern* » [13].

[2] *Differential Privacy*
[3] Google Apple Facebook Amazon Microsoft
[4] *Federated Learning*
[5] *Privacy-by-design*

Afin de favoriser l'effort collectif de lutte contre la crise liée à la pandémie, NEHS Digital, la Société Française de Radiologie (SFR) et le Collège des Enseignants en Radiologie de France (CERF) se sont mobilisés pour mettre en place une base de données anonymisées de référence nationale et européenne permettant l'amélioration des connaissances et le développement de solutions innovantes pour le diagnostic, le pronostic et le suivi des conséquences de la COVID-19. Cette base de données anonymisées a été baptisée FIDAC pour ***F**rench **I**maging **D**atabase **A**gainst **C**oronavirus*.

Nous avons mis en place un système de collecte de scanners thoraciques avec des données complémentaires de type clinique, virologique et radiologique sur une cohorte de patients présentant des signes cliniques d'infection au COVID-19. Dès le départ du projet, les partenaires avaient pour exigence que ces données soient anonymisées afin de favoriser les échanges et le partage des connaissances. Au final, la base anonymisée est constituée de l'imagerie de 5 843 patients adultes.

Ces scanners sont enrichis de métadonnées médicales et les données sont anonymisées par les établissements de santé volontaires. Cette base est destinée à être mettre mise à disposition de tiers à des fins statistiques ou de recherche.

Afin de rendre cette base exploitable pour la recherche en IA, nous avons proposé un mécanisme d'anonymisation qui assure l'impossibilité de réidentification d'une personne physique et ce, en considérant les principes de régulation du RGPD tels que définis dans l'article 26 [14]. Il est important de noter que le traitement amont de ces données (de leur collecte à leur anonymisation) doit lui répondre strictement aux exigences du RGPD. Dans le reste de ce document, nous nous focalisons sur le processus d'anonymisation et proposons un retour d'expérience sur son utilisation dans cette base de données d'imagerie médicale.

4 Conception du processus d'anonymisation des données

4.1 Les grands principes

Pour concevoir un processus d'anonymisation pertinent, nous avons suivi les recommandations de l'autorité nationale, c'est-à-dire la CNIL [15] [16] :

1. Supprimer les éléments d'identification directe ainsi que les valeurs rares qui pourraient permettre une réidentification aisée des personnes ;
2. Distinguer les informations importantes des informations secondaires ou inutiles (i.e., supprimables) ;
3. Définir la finesse acceptable pour chaque information conservée (e.g., conserver l'année de naissance des patients n'est pas possible, mais conserver la décennie dans laquelle ils sont nées est acceptable) ;
4. Définir les priorités (e.g. est-il plus important de conserver une grande finesse sur telle information ou de conserver telle autre information ?).

La suppression d'éléments d'identification est triviale lorsqu'ils sont explicites (e.g. nom et prénom) et ceux-ci sont soit écartés lors de la collecte, soit ils sont immédiatement supprimés. Évaluer le degré d'importance des données ainsi que définir leur finesse et priorité relative requièrent de réaliser des arbitrages. C'est pourquoi, il est indispensable de bien définir les objectifs d'études ainsi que le domaine métier concerné.

Ce processus d'anonymisation correspond ainsi à identifier et réduire les données à collecter mais également de baisser leur granularité informationnelle. La conséquence de l'anonymisation correspond donc une nécessaire perte de précision des données.

Cette seule phase de conception est loin d'être suffisante. En ce sens, le RGPD [17, pp. 11-12] définit trois critères qui permettent de s'assurer qu'un jeu de données est véritablement anonyme :

1. La non-individualisation : il ne doit pas être possible d'isoler un individu dans le jeu de données ;
2. La non-corrélation : il ne doit pas être possible de relier entre eux des ensembles de données distincts concernant un même individu ;
3. La non-inférence : il ne doit pas être possible de déduire de façon quasi-certaine de nouvelles informations sur un individu.

Cette seconde phase de robustesse de l'anonymisation ne peut pleinement s'évaluer qu'au regard des données collectées. Elle est donc à réaliser *a minima* à la fin de la collecte, bien qu'il puisse être intéressant de la faire en cours de collecte pour disposer de premières tendances. Enfin, les techniques d'anonymisation et de réidentification étant amenées à évoluer régulièrement, il est indispensable d'effectuer une troisième phase de surveillance afin de préserver dans le temps le caractère anonyme des données produites.

4.2 Un processus d'anonymisation par construction

La FIGURE 1 ci-dessous illustre la méthodologie d'anonymisation par construction que nous avons mise en place pour le traitement des données personnelles issues des jeux de données en imagerie médicale. Cette méthodologie, comme nous allons le voir, s'appuie sur les principes énoncés ci-dessus.

En s'inspirant des bonnes pratiques utilisées en confidentialité par construction, nous distinguons cinq phases successives, c'est à dire :

i. la phase de conception du projet de diffusion des données d'imagerie médicale, essentiellement portées par les experts du domaine concerné et par les responsables du projet
ii. la phase de traitement qui porte sur la collecte des données en établissements de santé ;
iii. la phase de vérification de l'anonymisation des données, principalement réalisée par des experts du traitement des données ;

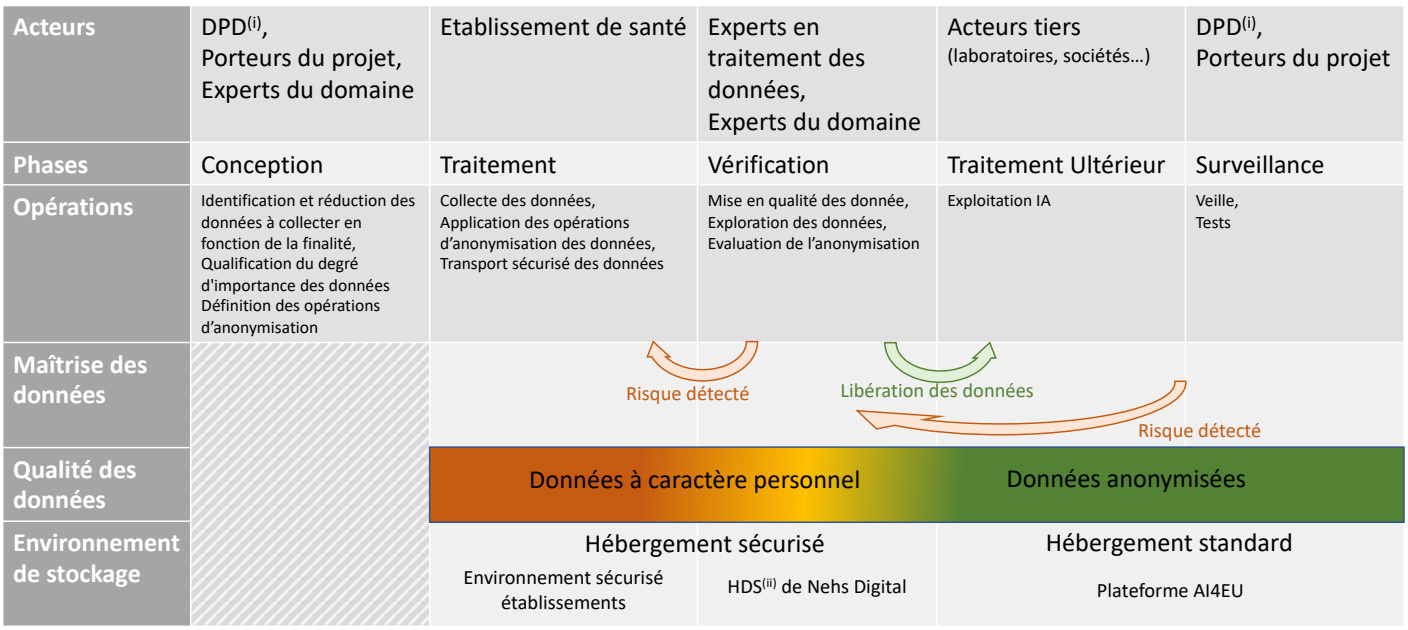

FIGURE 1 - METHODOLOGIE D'ANONYMISATION PAR CONSTRUCTION

iv. la phase de traitement ultérieur des données utilisant massivement l'apprentissage automatique et qui concerne les acteurs tiers tels que les laboratoires privés ou publiques, les sociétés privés qui développent des solutions innovantes, etc. ;

v. la phase de surveillance qui incombe principalement aux Délégués à la Protection des Données (DPD)[6] des entités concernées et qui interagissent de concert avec les responsables de projet.

Les opérations concernées par ces différentes phases sont indiquées dans la figure car elles représentent véritablement le cœur des activités du projet de diffusion des données d'imagerie médicale. Il est intéressant de noter que le basculement entre l'hébergement sécurisé et privé vers un hébergement des données accessible aux acteurs finaux et au public (par exemple, par l'entremise de plateformes d'IA) n'est réalisé qu'après la phase de vérification (libération des données). En effet, il n'est pas rare que cette phase identifie des risques majeurs qu'il faut traiter en appliquant des processus d'élimination ou bien d'offuscation des données.

Les risques identifiés dans les environnements de production sont quant à eux surveillés et conduisent à une amélioration de la phase de vérification.

5 Mise en œuvre du processus d'anonymisation pour la base FIDAC

5.1 Identification et réduction des données à collecter

Pour définir la pertinence des données à collecter, NEHS Digital s'est appuyé sur la Société Française de Radiologie (SFR) et le Collège d'Enseignants Radiologues de France (CERF) qui ont défini, outre la série de scanner et leur compte-rendu radiologique, un ensemble de 4 informations médicales supplémentaires à récolter : (i) l'indication, (ii) le délai entre le début des symptômes et la réalisation du scanner thoracique, (iii) le diagnostic radiologique et (iv) le résultat du test PCR. En outre, 2 informations complémentaires relatives à l'établissement ayant réalisé l'examen ont également été demandées : (v) le nom de l'établissement et (vi) son identifiant.

Un travail de conception a été mené afin de modéliser les opérations d'anonymisation des données, le « transport » de ces dernières ainsi que leur hébergement. La sélection des données à retenir a principalement porté sur les métadonnées présentes dans les images composant les séries de scanner thoracique.

Concernant FIDAC, les données collectées sont : (i) des données médicales définies par la SFR, (ii) le compte-rendu radiologique produit par le radiologue et (iii) l'imagerie médicale au standard DICOM [18].

La finalité est de mettre à disposition un jeu de données conséquent et adapté pour la recherche en IA sur l'identification automatique de la COVID. En s'appuyant sur les experts du domaine que sont la SFR et le CERF, les données médicales sont jugées prioritaires ainsi que l'imagerie alors que les données techniques sont jugées secondaires. Par ailleurs, la conservation du format DICOM a également été jugée prioritaire car elle permet d'exploiter toute la dynamique de l'image acquise. Pour résumer, les données suivantes ont été ciblées :

— Les 6 données définies par les radiologues (sans identification directe des patients) et mentionnées dans le premier paragraphe de cette sous-section ;

— 1 donnée textuelle sous la forme du Compte Rendu Radiologique (ne devant pas contenir d'entête ni de paragraphe d'identification du patient) ;

[6] *Data Protection Officer (DPO)*

— 1 série d'images au standard DICOM contenant 1 donnée image et plus d'une centaine de données associées (métadonnées ou tags DICOM).

La partie conséquente du travail d'analyse des données à conserver ou non, a donc principalement porté sur les métadonnées DICOM. Ce standard encapsule l'image acquise dans un fichier contenant en outre un ensemble de métadonnées d'identification (patient, médecin), de réalisation de l'examen (date et heure, dosimétrie), de dynamique et caractéristiques de l'image acquises (résolution, mot-machine alloués), de description du matériel (marque, modèle, composants) et d'un ensemble de descripteurs permettant l'interopérabilité avec les systèmes d'information hospitaliers et radiologiques. Ces métadonnées sont nombreuses, de l'ordre de centaines, et certaines sont obligatoires pour conserver la conformité à ce standard[7] [19].

Le standard DICOM prévoit ces opérations de déidentification et propose un cadre [20] qui définit les éléments obligatoires de manière la plus minimaliste possible. Suite à l'application de ce profil anonymisé, il en résulte un sous-ensemble de 93 métadonnées à analyser en dehors de l'image elle-même.

5.2 Qualification du degré d'importance des données

À partir de ce travail d'identification et de réduction des données à collecter, nous avons passé en revue chacune de ces données afin de qualifier leur niveau d'importance. Le résultat de cette revue est synthétisé dans la TABLE 1.

Les critères retenus sont les suivants :

— Information principale pour l'objectif : l'absence de la donnée dégrade fortement la pertinence du projet ou le remet en cause ;

— Information secondaire pour l'objectif : l'absence de la donnée diminue la finesse des analyses mais sans remettre en cause l'objectif principal ;

— Information inutile pour l'objectif : l'absence de la donnée n'impacte pas les objectifs principaux ou secondaires définis ;

— Information techniquement requise pour conformité au standard DICOM : la donnée est nécessaire pour exploiter l'image au standard DICOM.

	Données médicales	Compte-rendu	Image
Information principale	4	0	25
Information secondaire	2	1	23
Information inutile	0	0	20
Information techniquement requise	0	0	26

TABLE 1 – Qualification de l'importance des données

Les 20 informations évaluées comme inutiles ont été soit supprimées soit mises à la valeur par défaut en fonction des exigences du standard DICOM. Par exemple la description de l'examen a été supprimée alors que sa date et heure de réalisation ont été vidées.

5.3 Données techniquement requises

Les informations requises techniquement ont nécessité une analyse complémentaire pour déterminer si elles pouvaient contribuer à réidentifier le patient. Pour cela, nous les avons passées en revue pour évaluer s'il était nécessaire de retraiter leurs valeurs. Le résultat de cette évaluation est présenté dans la Table 2.

Par exemple, le type de modalité (ici un scanner) induit la valeur CT[8] pour la donnée *Image Type*. Cette valeur étant constante par construction du jeu de données, aucun retraitement n'est donc nécessaire. Ou encore, les données relatives au patient (son nom et son identifiant) sont automatiquement inscrites dans le fichier image. Étant contraint de disposer de nom unique de patients pour une bonne comptabilité DICOM, nous avons ré-encodé ces valeurs.

Le point d'attention a porté sur les données de type *Instance UID*. Le standard DICOM [21] a défini l'UID selon le schéma d'identification basé sur l'identification objet de l'OSI comme défini par le standard ISO 8824. Chaque identifiant est unique et enregistré selon l'ISO 9834-1 afin d'assurer son unicité. Chaque UID est ainsi composé de 2 parties : une racine provenant de l'organisation émettrice et un suffixe :

UID = <organisation émettrice>.<suffixe> (1)

Un type « *Instance UID* » est un UID mais utilisé pour chacune des instances d'un élément DICOM. Dès lors, par construction chacune des images d'une série d'un scanner possède un identifiant unique à travers le monde. Ce mécanisme permet d'assurer la traçabilité et de garantir que 2 images distinctes ne soient pas associées par erreur à un même patient. En raison de sa construction, il est possible de retrouver beaucoup d'informations comme le fabricant de la modalité mais également la période au cours de laquelle les images ont été réalisées.

Pour évacuer toute possibilité de déduire ces diverses informations, ces identifiants ont été régénérés par séquence aléatoire lors de leur export à partir des établissements contributeurs vers les serveurs NEHS. Un même préfixe d'UID a été utilisé afin de généraliser cette valeur.

6 Vérification de l'efficacité de l'anonymisation

La réception des données anonymisées s'effectue à ce stade dans un environnement certifié Hébergeur de Données de Santé (HDS) requis par la législation française pour toute entité hébergeant des données de santé. Cette certification s'appuie principalement sur les normes ISO 27001 [22] et ISO 27018 [23].

Au cours de la collecte, nous procédons à des opérations de contrôle qualité des données (e.g. détection de valeurs aberrantes, manquantes) et de l'anonymisation. D'une part,

[7] Ces métadonnées permettent par exemple de garantir la bonne interprétation des données avec les appareils compatibles (console de diagnostic ou de revue) et de réaliser des mesures (distances, volumes, angles) correctes.

[8] *Computed Tomography*

nous avons mis en place une procédure portant sur les métadonnées contenues dans les images DICOM et d'autre, nous procédons à une vérification portant sur les compte-rendus radiologiques.

6.1 Vérification des métadonnées

Comme dans tout contrôle de qualité des données, les statistiques descriptives sont utilisées afin d'appréhender et comprendre le jeu de données.

Typiquement, la distribution de l'âge par décennie a fait apparaître rapidement une rareté d'individus aux classes d'âges extrêmes (18-19 ans et plus de 100 ans) comme illustré dans la FIGURE 2 réalisé au cours de la collecte. Pour minimiser le risque de réidentification pour ces classes extrêmes, nous avons évalué la possibilité de les regrouper avec les classes suivantes ou précédentes auprès des radiologues. Ces derniers ont validé que cette baisse de finesse n'avait pas d'impact majeur sur l'interprétation des images.

Tag	Donnée	Action de retraitement
0002:0001	File Meta Information Version	aucune : valeur non spécifique ne permettant pas d'identifier la modalité
0002:0002	Media Storage SOP Class UID	aucune : valeur identique pour tout le jeu de données
0002:0003	Media Storage SOP Instance UID	valeur régénérée lors de l'anonymisation
0002:0010	Transfer Syntax UID	aucune : valeur non spécifique ne permettant pas d'identifier la modalité
0002:0012	Implementation Class UID	aucune : valeur générique pour ce type d'image
0002:0013	Implementation Version Name	aucune : valeur non spécifique ne permettant pas d'identifier la modalité
0008:0005	Specific Character Set	aucune : valeur courante sans être exclusive à une modalité
0008:0008	Image Type	aucune : valeur non spécifique ne permettant pas d'identifier la modalité
0008:0016	SOP Class UID	aucune : valeur identique pour tout le jeu de données
0008:0018	SOP Instance UID	valeur régénérée lors de l'anonymisation
0008:0050	Accession Number	valeur régénérée lors de l'anonymisation
0008:0060	Modality	aucune : valeur identique pour tout le jeu de données
0010:0010	Patient's Name	valeur régénérée lors de l'anonymisation
0010:0020	Patient ID	valeur régénérée lors de l'anonymisation
0010:0021	Issuer of Patient ID	valeur régénérée lors de l'anonymisation
0018:0015	Body Part Examined	aucune : valeur normalisée
0018:9345	CTDIvol	aucune : valeur non spécifique ne permettant pas d'identifier la modalité
0020:000D	Study Instance UID	valeur régénérée lors de l'anonymisation
0020:000E	Series Instance UID	valeur régénérée lors de l'anonymisation
0028:0002	Samples per Pixel	aucune : valeur non spécifique ne permettant pas d'identifier la modalité
0028:1050	Window Center	aucune : valeur non spécifique ne permettant pas d'identifier la modalité
0028:1051	Window Width	aucune : valeur non spécifique ne permettant pas d'identifier la modalité
0028:1052	Rescale Intercept	aucune : valeur non spécifique ne permettant pas d'identifier la modalité
0028:1053	Rescale Slope	aucune : valeur non spécifique ne permettant pas d'identifier la modalité
0028:1054	Rescale Type	aucune : valeur non spécifique ne permettant pas d'identifier la modalité
0028:1055	Window Center & Width Explanation	aucune : valeur non spécifique ne permettant pas d'identifier la modalité

TABLE 2 – RETRAITEMENT DES DONNEES REQUISES TECHNIQUEMENT

Nous avons également évalué le volume de de données transféré par établissement pouvant donner lieu à un risque de réidentification. Il existe une forte disparité dans les contributions et la base de données comporte plusieurs établissements ayant transférés moins de 50 patients. Cette information ayant été évaluée comme secondaire, nous avons décidé de la généraliser et de ne retenir que la région d'appartenance comme illustré en FIGURE 3. Toutefois, à la fin de la collecte, certaines régions comportaient encore un nombre faible de patients (moins d'une centaine), nous les avons également regroupées dans une catégorie *Autre*.

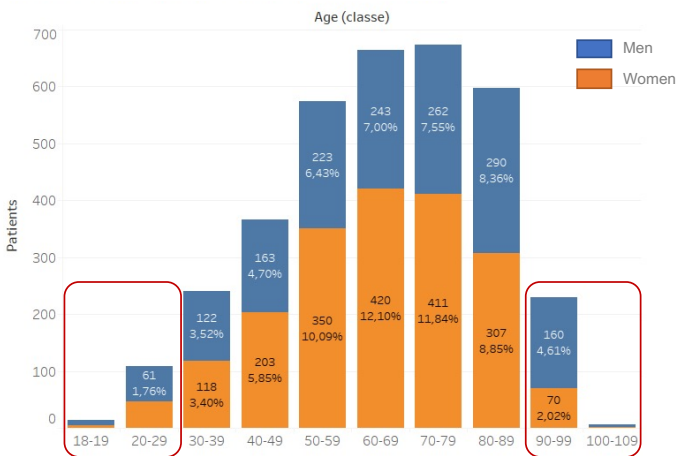

FIGURE 2 – Distribution de l'âge des patients par genre et décennie

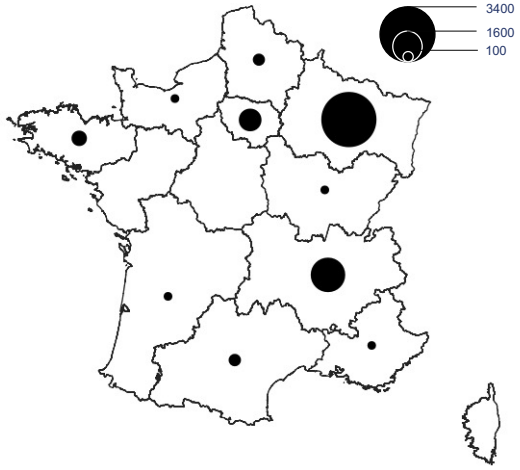

FIGURE 3 – Répartition des patients par région

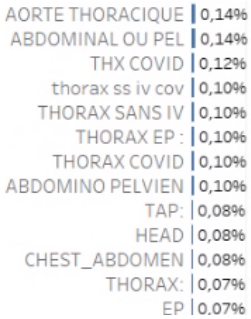

FIGURE 4 – Extrait de parties du corps recensée (tag 0018:0015) et leur proportion dans la base

D'autres analyses statistiques sont menées comme des Classifications Ascendantes Hiérarchiques, ou des mesures de corrélation afin de tenter d'identifier des croisements de données pouvant faire apparaître des moyens d'inférences pour réidentifier des patients. De ces travaux, nous avons identifié que pour les établissements ayant transmis peu de cas, il pouvait y avoir un risque (faible) de réidentification par une personne en contact avec les établissements concernés. Puisque nous avons supprimé l'information directe du nom de l'établissement, nous aurions pu croire avoir écarter ce risque. Néanmoins, nos travaux ont montré plusieurs corrélations fortes entre le nom de l'établissement et d'autres informations comme par exemple le modèle de scanners employés ou encore les libellés des parties du corps. En effet, comme l'illustre la FIGURE 4, certains noms sont très peu utilisés et présents uniquement dans un seul cas, malgré la standardisation DICOM.

6.2 Vérification des compte-rendus radiologues

Un compte-rendu radiologique (CRR) contient par nature beaucoup d'informations à caractère personnel de santé et donc sensibles. Les contributeurs avaient pour consigne d'extraire le corps du compte-rendu car cette partie ne contient que les informations médicales et sans données personnelles.

Afin de garantir l'absence de ces éléments, une vérification automatique à la soumission s'assure de l'absence de termes introduisant le nommage d'une personne (e.g. Monsieur, Mme, Dr, Confrère). En cas de détection de ces éléments, le contributeur est alors alerté et cette soumission refusée.

Afin de s'assurer qu'aucune autre donnée à caractère personnel ne se trouve présente, des contrôles aléatoires suivant une stratégie d'échantillonnage par établissement et médecin sont réalisés par des personnels soumis à un engagement de confidentialité. La procédure prévoit une vérification quotidienne de 5% des compte-rendus de la veille, avec un minimum de 5 CRR pour attester de l'anonymisation de ces derniers.

Après plusieurs semaines, 404 CRR ont été collectés dont 140 ont été vérifiés. Parmi ces 140, 33 cas (23,6%) ont été remontés comme contenant les informations suivantes :

- Modèle et marque du scanner utilisé ;
- Date du précédent examen ;
- Nom de l'établissement de l'examen ;
- Âge du patient.

Puisqu'à l'occasion du travail sur les métadonnées, nous avons déterminé un risque de réidentification sur ces données et qu'il n'est pas possible de retraiter l'ensemble des documents, nous avons décidé de supprimer définitivement ces derniers.

6.3. Synthèse des actions correctives

Tag	Donnée	Action corrective
0008,0070	Fabricant	Suppression
0008,1090	Modèle	Suppression
0010,1010	Âge du patient	Regroupement en tranche d'âges élargies sur les extrêmes
0018,0015	Partie du corps examinée	Regroupement en 4 catégories
0018,1020	Version logicielle	Suppression
NA	Établissement	Regroupement en grandes régions
NA	Rapport Radiologique	Suppression

TABLE 3 – Actions correctives réalisées pour baisser le risque de réidentification

Suite à l'évaluation des risques de réidentification, nous avons entrepris au cours de la collecte de supprimer certaines données ou bien de diminuer leur finesse en les regroupant (cf. TABLE 3). Dès lors, une opération de réencodage a été menée sur les données collectées, et les outils d'anonymisation mis à disposition des établissements contributeurs ont été mis à jour.

7 Mise à disposition du jeu de données pour la communauté IA

Le jeu de données anonymisé que nous avons créé a été mis à la disposition de la communauté de recherche en IA sous forme d'un accès indirect depuis la plateforme AI4Europe[9].

En effet, dans le cadre du projet H2020 AI4EU, un pilote industriel en santé a été conduit par Nehs avec ses partenaires académiques afin d'évaluer la capacité de la plateforme à mobiliser les acteurs européens de l'IA autour de problématiques fortes lié à la santé et démontrer le potentiel de la plateforme dans sa capacité à mettre en œuvre des solutions IA pour le monde industriel. Ainsi, ce jeu de données a fait l'objet d'une présentation et d'une mise à disposition.

La base de données FIDAC est ainsi accessible à travers la plateforme AI4Europe [24] et ouverte aux projets d'IA. Afin d'en assurer la surveillance, les acteurs souhaitant exploiter ces données doivent s'identifier et accepter une licence d'utilisation des données à titre non-exclusif qui leur donne le droit d'accéder, d'utiliser et d'exploiter la base de données et les données qu'elle contient conformément à leur destination.

Les retours de ces acteurs sur les données et leurs analysent contribuent à maintenir un niveau de veille sur ces dernières et nous pouvons alerter les licenciés de tout problèmes ultérieurs.

De cette mise à disposition, Thales et NEHS Digital ont exploité la base FIDAC dans le cadre du projet d'aide au diagnostic COVID-19 par IA appliquée sur des scanners tomographiques thoraciques (CT SCAN) financé par l'Agence de l'Innovation de Défense. L'IA permet d'apporter une aide à la décision pour le triage des cas covid / non covid en effectuant un 1er diagnostic probable.

Nous avons réalisé également un prototype de recherche pour anonymiser automatiquement les compte-rendus radiologiques qui systématise certains des principes mentionnés ci-dessus. En ce sens, nous avons développé le projet *medical Imaging Report Anonymiser* (mIRA). Ce projet au statut de PoC propose une API REST permettant de soumettre un compte-rendu et de recevoir sa version anonymisée. Ce dernier est quant à lui disponible à travers la plateforme AI4Europe [25].

8 Discussion

Le travail de conception sur les projets de données est une étape clef depuis l'avènement du RGPD. Dans le contexte d'un traitement en vue d'anonymiser les données, cette phase en devient majeure : l'effort de minimisation des données simplifie drastiquement ces opérations d'anonymisation ultérieure car cela diminue fortement les corrélations potentielles entre les différentes variables.

Pour mener à bien ce travail, il convient de définir clairement les objectifs de l'exploitation future de ce jeu de données à anonymiser. En outre, nous pensons que cette démarche permet de mieux exploiter les outils d'anonymisation existants en motivant les choix et paramétrages de ces algorithmes en fonction de la finalité d'exploitation plutôt que par une décision *a posteriori*.

Cette démarche peut sembler de prime abord rentrer en opposition avec les méthodes d'apprentissage automatique. En effet, diminuer de fait les données d'entrées pourrait laisser à penser que l'on va appauvrir les espaces des possibles et perdre en performance. Toutefois, une grande phase réalisée au cours des projets d'*apprentissage automatique* consiste à améliorer la qualité les données d'entraînement, de réaliser des opérations d'analyse d'impact des variables, de sélection de variables, d'échantillonnage ou encore de construction de variables synthétiques. Ce sont typiquement les opérations que nous avons menées lors de l'évaluation de l'efficacité de l'anonymisation.

Nous avons vu que les analyses de l'efficacité de l'anonymisation peuvent mettre en exergue des faiblesses permettant de réidentifier les individus et donc amener à réaliser des actions correctives sur le jeu de données. En conséquence, tant que les données ne sont pas libérées, il nous paraît primordial que ces données anonymisées mais non encore vérifiées disposent du même niveau de sécurité que si ces données n'étaient pas anonymisées. Dans notre cas, nous avons continué à utiliser l'environnement labellisé Hébergement de Données de Santé (HDS) jusqu'à la libération des données.

Une fois les données libérées, elles sont réputées anonymisées jusqu'à ce qu'elles ne le soient plus. Cela peut arriver par l'apparition de nouvelles techniques d'apprentissage, mais également par l'apparition ultérieure d'autres données qui, mises en lien avec le jeu initial, permet de faire émerger des schémas de réidentification. L'apparition de ces données complémentaires peut faire suite à des mises à disposition légales ou bien via des fuites suite à des violations de données. Ces pourquoi maîtriser la traçabilité des acteurs exploitant les jeux de données est nécessaire.

9 Conclusion et perspectives

Anonymiser des données sensibles n'est pas une opération triviale tant elle peut être lourde de conséquences. L'investissement en temps et en moyen peut paraître conséquent mais finalement les opérations dédiées s'intègrent bien aux étapes de conduite d'un projet basé sur la donnée. En effet, la phase de conception permet d'intégrer l'identification et la réduction des données à collecter alors que la phase de mise en qualité des données peut être utilisée pour s'assurer de l'efficacité de l'anonymisation.

L'étape de conception est importante et permet de faciliter les prises de décisions en cas de risque de réidentification. Toutefois, lorsque la finalité est générique comme dans notre cas (mettre à disposition des données pour la recherche en IA), il peut être difficile d'arbitrer quelles données conserver et avec quelle finesse.

Ce projet nous a permis de confirmer qu'il est primordial de

[9] https://www.ai4europe.eu/node/107

prévoir un espace sécurisé pendant la phase de collecte et tant que l'efficacité de l'anonymisation n'a pas été éprouvée sur le jeu de données à publier. Cela permet le cas échéant de mettre en place des actions de retraitement ou de suppression avant d'effectivement libérer les données.

La poursuite de nos travaux vise à appliquer cette méthodologie à d'autres jeux de données de santé afin d'en éprouver la capacité de généralisation.

10 Remerciements

Nous remercions nos partenaires que sont la SFR et le CERF pour leur volonté et dynamisme à contribuer à aider la recherche et la prise en soin des malades. Un grand merci également aux radiologues, DSI et personnels des établissements qui ont participé activement à la constitution de la base FIDAC.

11 Références

[1] J.-P. Laboureix, «Les comptes de la sécurité sociale 2020-2021,» 2021.

[2] CNIL, «La Plateforme des données de santé (Health Data Hub),» 09 02 2021. [En ligne]. Available: https://www.cnil.fr/fr/la-plateforme-des-donnees-de-sante-health-data-hub. [Accès le 03 03 2022].

[3] M. Wang, C. Xia, L. Huang, S. Xu, C. Qin, J. Liu, Y. Cao, P. Yu, T. Zhu, H. Zhu, C. Wu, R. Zhang, X. Chen, J. Wang, G. Du, C. Zhang, S. Wang, K. Chen, Z. Liu, L. Xia et W. Wang, «Deep learning-based triage and analysis of lesion burden for COVID-19: a retrospective study with external validation,» *Lancet Digit Health,* vol. 2(10), pp. e506-e515, Oct. 2020.

[4] L.-P. Sondeck, «Anonymisation des données, une nécessité à l'ère du RGPD,» *Sécurité des systèmes d'information,* 10 Nov. 2019.

[5] F. Ben Fredj, «Méthode et outil d'anonymisation des données sensibles,» Conservatoire national des arts et metiers - CNAM; Université de Sfax (Tunisie). Faculté des Sciences économiques et de gestion, 2017.

[6] L. H. Cox, «Suppression methodology and statistical disclosure analysis,» *Journal of the American Statistical Association,* vol. 75(370), p. 377–385, 1980.

[7] P. Samarati, «Protecting respondents identities in microdata release,» *IEEE transactions on Knowledge and Data Engineering,* vol. 13(6), pp. 1010-1027, 2001.

[8] J. Domingo-Ferrer et V. Torra, «A quantitative comparison of disclosure control methods for microdata,» *Confidentiality, disclosure and data access: theory and practical applications for statistical agencies,* pp. 111-134, 2001.

[9] X. Xiao et Y. Tao, «Anatomy: Simple and effective privacy preservation,» chez *Proceedings of the 32nd international conference on Very large data bases*, Seoul (Korea), 2006.

[10] T. Dalenius et S. P. Reiss, «Data-Swapping: A Technique for Disclosure Control,» *Journal of Statistical Planning and Inference,* vol. 6(1), pp. 73-85, 1982.

[11] C. Dwork, F. McSherry, K. Nissim et A. Smith, «Calibrating Noise to Sensitivity,» *Private Data Analysis Journal of Privacy and Confidentiality,* vol. 7(3), 2016.

[12] T. S. Brisimi, R. Chen, T. Mela, A. Olshevsky, I. C. Paschalidis et W. Shi, «Federated learning of predictive models from federated Electronic Health Records,» *International Journal of Medical Informatics,* vol. 112, pp. 59-67, 2018.

[13] S. E. Rossi, J. J. Erasmus et M. Volp, «Crazy-Paving Pattern at Thin-Section CT of the Lungs: Radiologic-Pathologic Overview,» *RadioGraphics Vol. 23, No. 6,* 2003.

[14] Parlement européen et du conseil, «Règlement 2016/679 du parlement Européen et du Conseil,» *Journal officiel de l'Union européenne,* 27 04 2016.

[15] CNIL, «L'anonymisation des données, un traitement clé pour l'open data,» 17 10 2019. [En ligne]. Available: https://www.cnil.fr/fr/lanonymisation-des-donnees-un-traitement-cle-pour-lopen-data. [Accès le 10 10 2021].

[16] CNIL, «Fiche n°1 : Identifier les données à caractère personnel,» chez *Guide RGPD du développeur, v1.0.1*, LaboCNIL, 2020, pp. 5-6.

[17] Article 29 Data Protection Working Party, «Opinion 05/2014 on Anonymisation Techniques,» European Commission, 2014.

[18] NEMA, «DICOM,» The Medical Imaging Technology Association, [En ligne]. Available: https://www.dicomstandard.org. [Accès le 11 10 2021].

[19] Innnolitics, LLC, «DICOM Standard Browser,» 2016. [En ligne]. Available: https://dicom.innolitics.com/ciods. [Accès le 11 10 2021].

[20] DICOM Standards Committee, «DICOM PS3.15 2022a - Security and System Management Profiles,» DICOM, [En ligne]. Available: http://dicom.nema.org/medical/dicom/current/output/html/part15.html. [Accès le 11 10 2021].

[21] DICOM Standards Committee, «Unique Identifiers (UIDs),» [En ligne]. Available: https://dicom.nema.org/dicom/2013/output/chtml/part05/chapter_9.html.

[22] ISO/IEC JTC 1/SC 27, ISO/IEC 27001:2013, 2013, p. 23.

[23] ISO/IEC JTC 1/SC 27, ISO/IEC 27018:2019, 2019, p. 23.

[24] NEHS Digital, «Covid-19 chest CT-scan Dataset,» 12 2021. [En ligne]. Available: https://www.ai4europe.eu/research/ai-catalog/covid-19-chest-ct-scan-dataset-0.

[25] J. Clech et G. Martial, «mIRA - medical Imaging Report Anonymiser,» 11 2021. [En ligne]. Available: https://www.ai4europe.eu/research/ai-catalog/mira-medical-imaging-report-anonymiser.