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

- **Jérémy Clech**, NEHS DIGITAL
- Arnaud Gotlieb, Laboratoire SIMULA
- Florence Sève, Frédérique Didout, Groupe MNH
- Patrick Malléa, NEHS DIGITAL

APIA 2022 – Saint-Etienne

AfIA
Association française pour l'Intelligence Artificielle

MINES Saint-Étienne

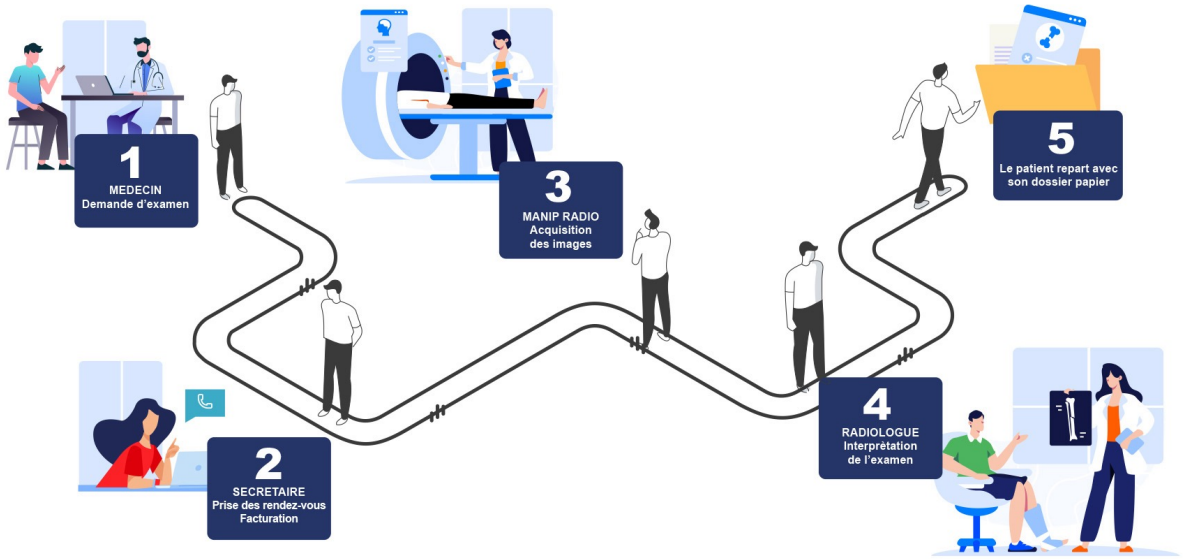
Introduction & Contexte

NEHS Digital

Constat

Difficultés de diffusion de données médicales

Activités en radiologie

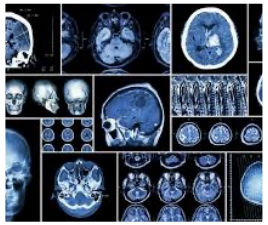


4+ Peta octets

Volume d'images produites annuellement

Rendez-vous

ERP Radio.



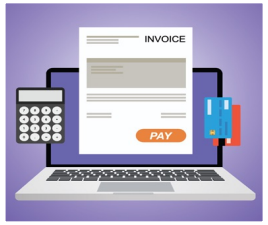
Archivage

Routage

Orchestration

Visualisation

Reporting



Facturation

Diffusion

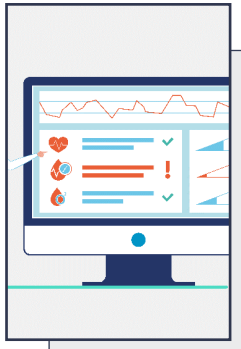
Activités en télémédecine

Téléconsultation

Fonctionne de la même manière qu'une consultation physique. Le patient, accompagné éventuellement d'un professionnel de santé, et le médecin échangent à distance et ce dernier livre son diagnostic en fonction des informations qu'on lui fournit.

Téléexpertise

Pratique qui concerne uniquement les professionnels médicaux. Le professionnel médical chargé de surveiller l'évolution de la thérapie d'un patient peut ainsi demander un ou plusieurs conseils à d'autres confrères dans le but de prendre une décision la plus juste possible.

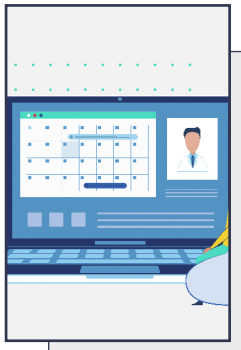


Télésurveillance médicale

Pratique au long terme qui permet au professionnel médical de suivre l'état du patient (qui est à domicile) à partir de données de suivi.

Téléradiologie

Désigne le fait qu'un professionnel médical assiste un professionnel de santé pendant qu'il agit sur le patient. Il peut s'agir notamment d'un acte de chirurgie.



Régulation médicale

Pratique de la télémédecine qui concerne les réponses données par les professionnels médicaux dans le cadre d'un appel émis en urgence (15).

Constat – 1er trimestre 2020

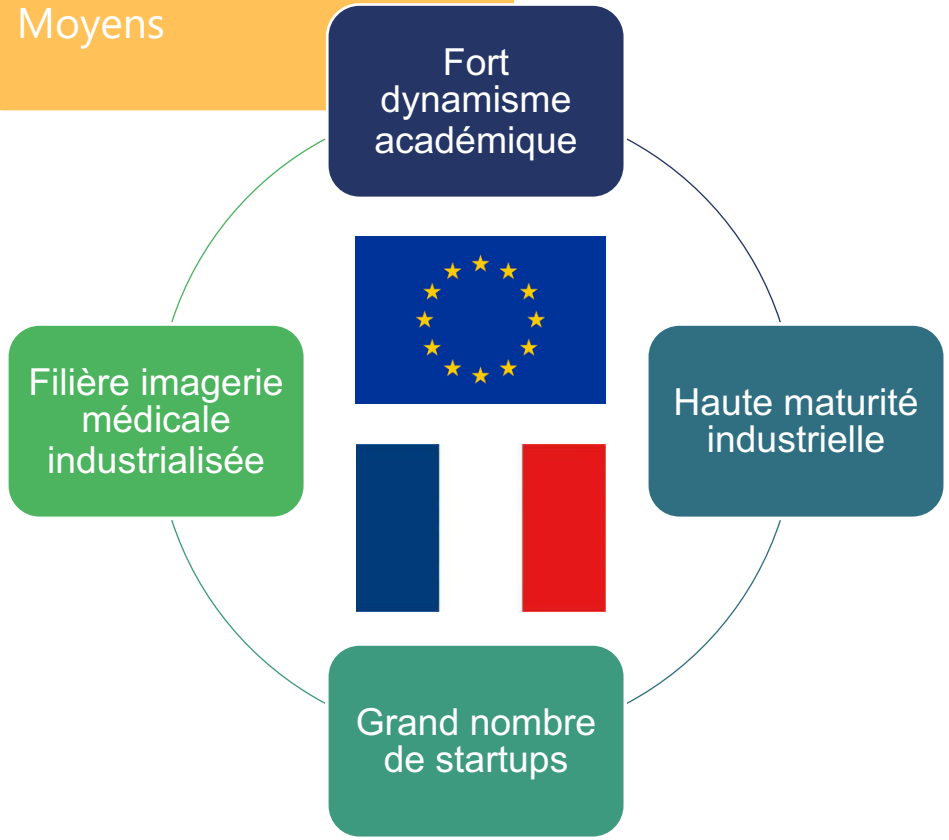
Besoin

Triage à l'hôpital des patients COVID / non COVID

Solutions

- Tests PCR
  - imprécis
  - disponibilité limité
- **Scanner thoracique**
  - fiable
  - besoin en radiologues

Moyens



Formidable opportunité pour la communauté IA de contribuer à répondre au problème

Résultats

- Base COVID
  - 3 191 Patients
- Algorithme IA
  - Infervision
- Déploiement
  - Infervision
  - On Premise

Mars 2020

- Base COVID – FIDAC
  - 5 843 Patients
  - 22 centres
- Algorithme IA
  - Thales
- Déploiement
  - NEHS DIGITAL
  - Cloud (Télémédecine)

Décembre 2020

Un écart de délai majeur

Difficulté d'accès & diffusion des données

Acquisition de données pour diagnostic initial

Réutilisation de données de santé – RGPD

Données à Caractère Personnel Sensibles – PIA

Volonté de partage ouvert du jeu de données

Anonymisation des données

Lever ce verrou en facilitant la mise à disposition de données de santé de manière licite et plus rapide

Processus d’anonymisation par construction

Données Personnelles & Imagerie médicale

Travaux antérieurs

Problématique

Description du processus d’anonymisation par construction

Retours d’expérience sur FIDAC

Données Personnelles & Imagerie Médicale

Donnée Personnelle

Toute information identifiant directement ou indirectement une personne physique

Exemples

nom, date de naissance, commune de résidence, carte d'identité, n° passeport, n° sécurité social, téléphone, photographie, empreinte digitale, coordonnées GPS…

Imagerie médicale

Ensemble de moyens d'acquisition et de restitution d'images du corps humain en exploitant les phénomènes physiques
Permettant de visualiser l'anatomie, la physiologie ou le métabolisme du corps humain

Exemples

Échographie
Radiographie générale
IRM
Scanner

DICOM

Protocole informatique permettant l'interopérabilité entre les composants de la chaîne d'imagerie : SI, modalité, console, archivage…

Format image composé de métadonnées

Pixels de l'image,
Contexte patient,
Environnement technique,
Environnement médical,
Nature et conditions de l'examen

1 scanner est une série de plusieurs centaines d'images DICOM

Plusieurs centaines de données à caractère personnel et sensibles

Critères d'effectivité de l'anonymisation

Non-individualisation

il ne doit pas être possible d'**isoler un individu** dans le jeu de données

Non-corrélation

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

Non-inférence

il ne doit pas être possible de **déduire** de façon quasi-certaine de **nouvelles informations** sur un individu

Ne peut s'évaluer que sur les données collectées

Anonymiser implique d'altérer les données pour perdre en précision

Travaux antérieurs sur l'anonymisation / confidentialité

Approches classiques

- **Principe**
  - Réduire l'espace des valeurs possibles
    - Suppression,
    - Généralisation,
    - Permutation
- **Exemple**
  - Arkhn Arx
- **Inconvénients**
  - Lenteur de réalisation
  - Intervention manuelle
  - Choix *a priori*
  - Représentativité / biais

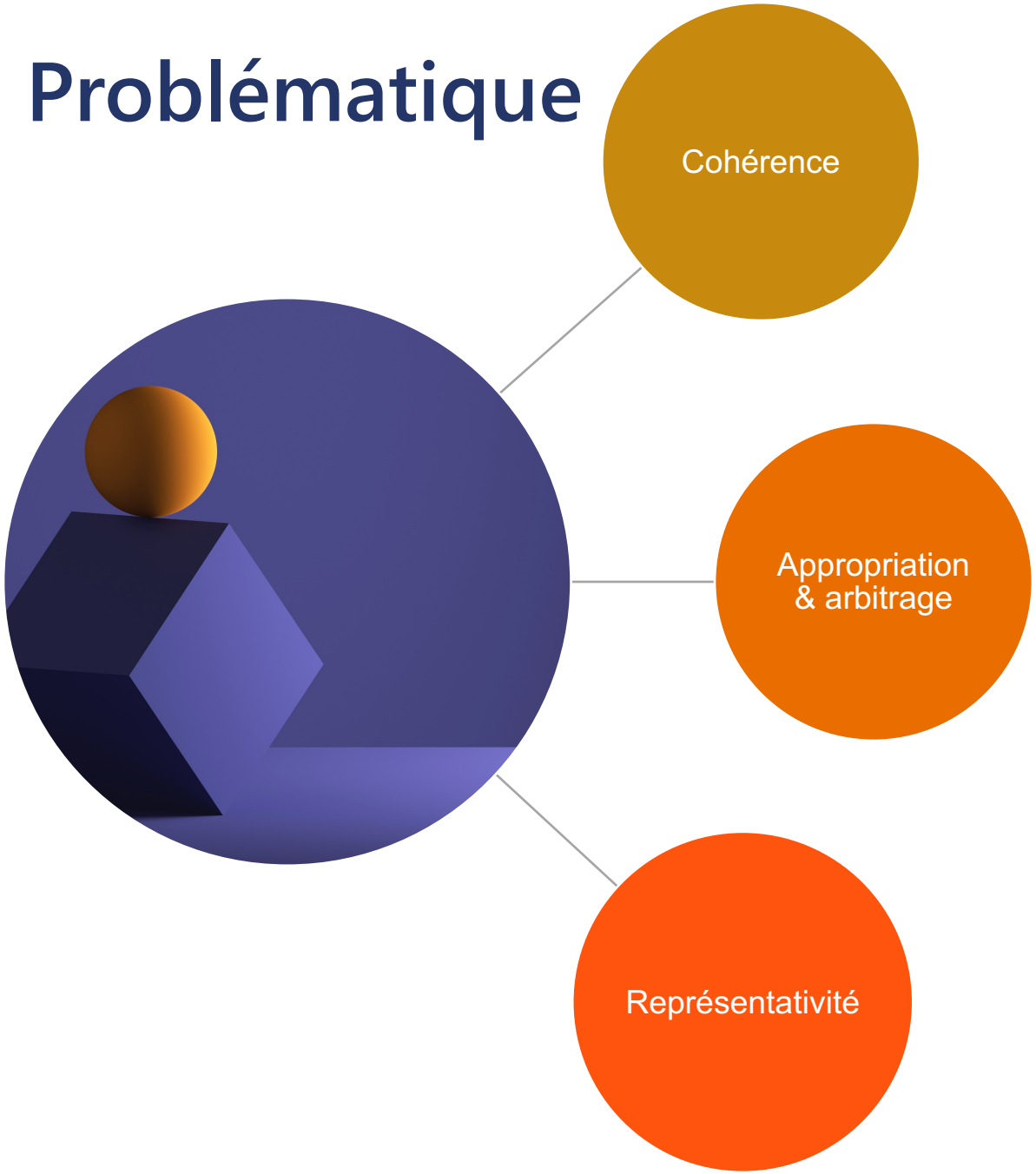
Confidentialité Différentielle

- **Principe**
  - Empêcher l'accès directe aux données individuelles
  - Ne proposer que des données agrégées
- **Exemple**
  - SmartNoise
- **Inconvénients**
  - Appropriation des données / images
  - Impossibilité de diffusion du jeu de données

Apprentissage Fédéré

- **Principe**
  - Seules les données locales sont accessibles
  - Apprentissage sur tous les sites
- **Exemple**
  - Labelia (ex Substra)
- **Inconvénients**
  - Difficulté organisationnelle
  - Prérequis de ressources sur les sites

Problématique



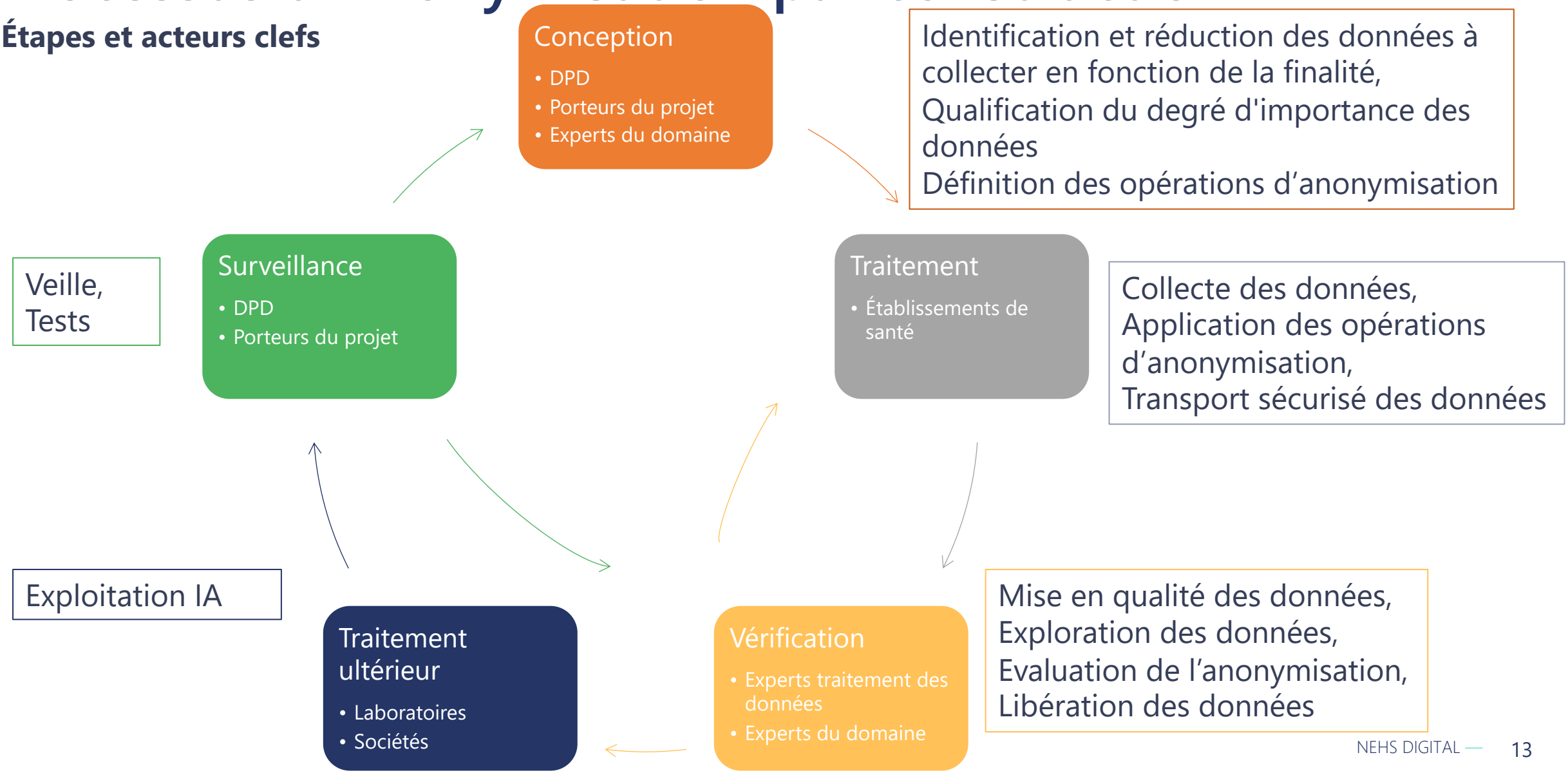
Jeu d'équilibriste car grand nombre de variables concernées et relativement peu d'individus.

Les méthodes sont appliquées *a posteriori* de la collecte des données engendrant en risque ultime : **la non pertinence du jeu de données**

Proposition
S'inspirer du *Privacy by Design* pour définir dès la conception du traitement de collecte l'objectif d'anonymisation
Et minimiser l'utilisation de ces techniques et d'en avoir une meilleure maîtrise.

Processus d'Anonymisation par construction

Étapes et acteurs clefs



DPD : Délégué à la Protection des Données

Phase de Conception

Définir la finalité de la collecte pour ensuite réaliser les arbitrages suivants :

Supprimer les éléments d'**identification directe** et les **valeurs rares**

identifiants, nom, prénom, adresse…

Distinguer les informations **importantes**, **secondaires** ou **inutiles**

quelles informations peuvent être supprimées ?

Définir le **degré de finesse** acceptable pour chacune des informations

conserver l'année de naissance ou bien la décennie correspondante ?

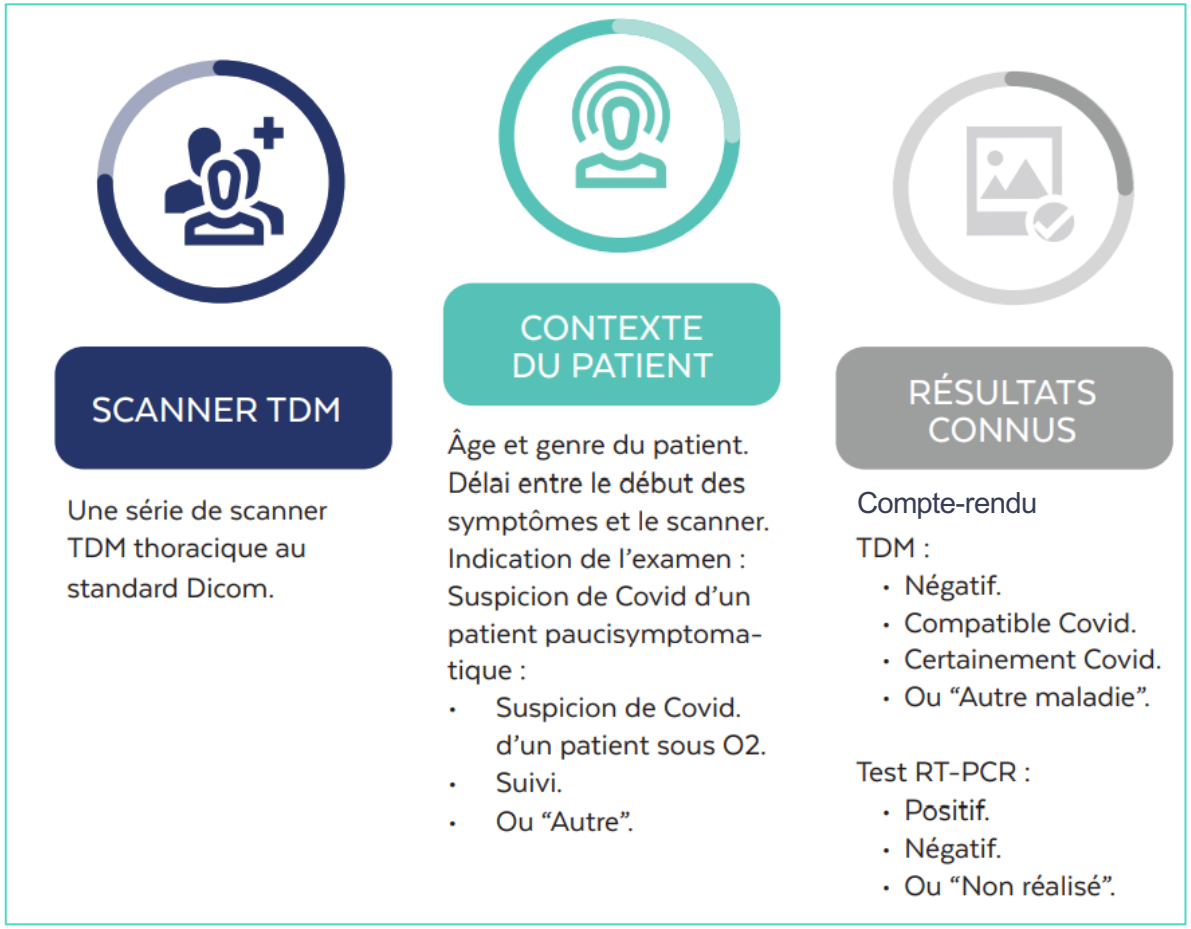
Définir les **priorités**

est-il plus important de conserver une grande finesse sur telle information ou de conserver telle autre information ?

Conception FIDAC 1/2

Finalité

Projet pour rassembler le plus grand nombre de scanners de patients atteints ou non de la COVID-19 afin de faciliter le diagnostic et favoriser les projets de recherches et d'enseignement



Données identifiantes

- Application profil anonymisation DICOM
- Régénération des identifiants ou suppression

Données principales

- Âge
- Genre
- Délai symptôme
- Diagnostic radiologique
- Image

Données secondaires

- Compte-rendu
- Caractéristique modalité

FIDAC	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

Conception FIDAC 2/2

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

Tag	Donnée	Action de retraitement
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

Étude des données techniquement requises

Phase de Vérification

Environnement sécurisé	Hébergement Données de Santé Donnée toujours réputée à caractère personnel
Contrôle qualité de la donnée	Statistiques descriptives
Contrôle anonymisation	Échantillonnage Statistiques descriptives, exploratoires Clustering
Retraitement	Dégrader la précision des données en fonction des critères définies en phase de Conception

Contrôler dans un environnement sécurisé durant la collecte et retraiter si besoin

Vérification FIDAC 1/4

Problème

- Faible effectifs dans les classes d'âges extrêmes

Nature de donnée

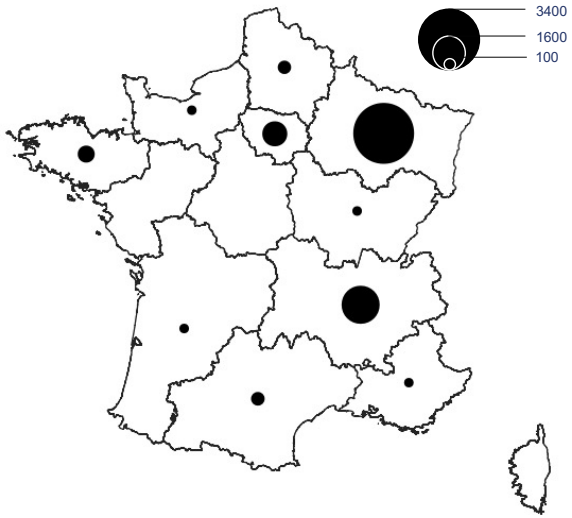
- Principale

Degré de finesse

- Les grands âges de la vie

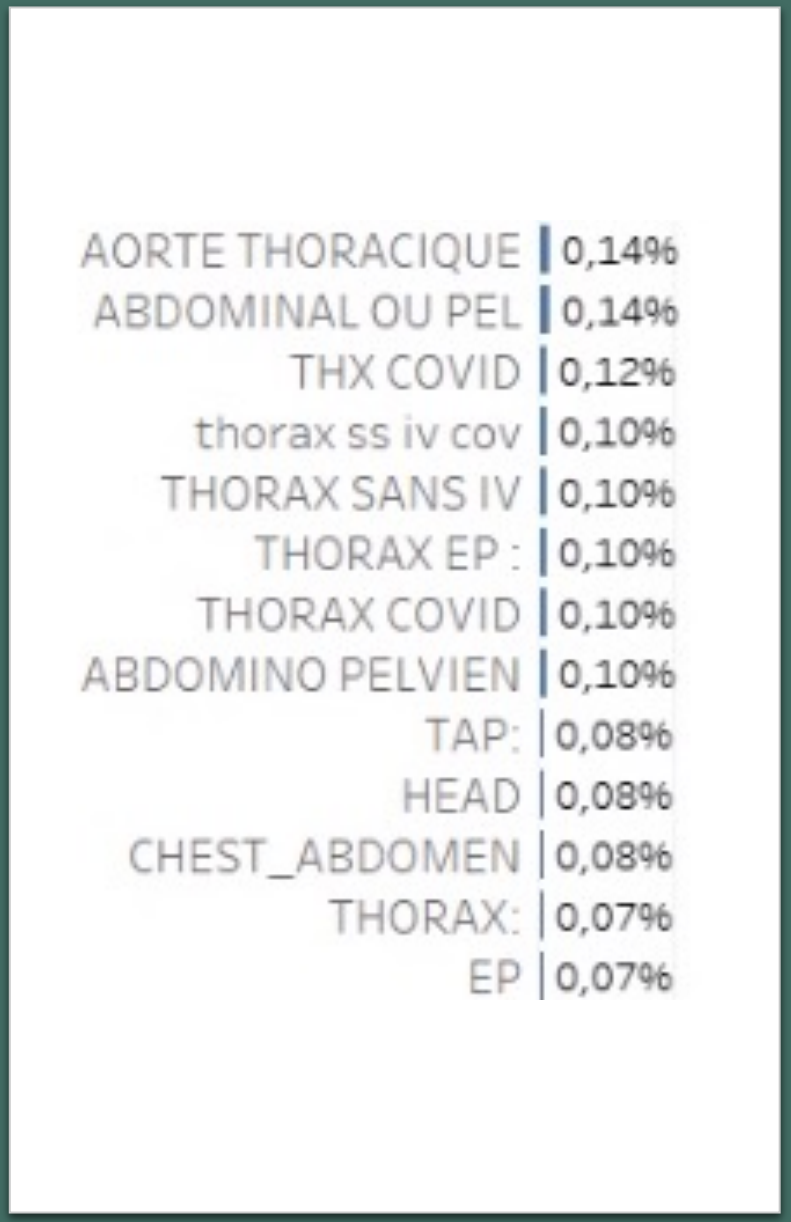
Retraitement

- Regroupement des extrêmes



Vérification FIDAC 2/4

Problème	• Faible effectifs dans certaines régions
Nature de donnée	• Secondaire
Degré de finesse	• N/A
Retraitement	• Regroupement des faibles effectifs



Vérification FIDAC 3/4

Problème

- Partie du corps non normée
- Certains libellés rares

Nature de donnée

- Requise techniquement

Degré de finesse

- N/A

Retraitement

- Regroupement en 4 catégories

Vérification FIDAC 4/4

Problème

- 140 compte-rendus analysés sur 404
- 23,6% contiennent des données sensibles : modalité, date du précédent examen, âge du patient

Nature de donnée

- Information secondaire

Degré de finesse

- N/A

Retraitement

- Suppression

Synthèse des actions correctives

- **Réencoder les modifications sur les données collectées**
- **Modifier le traitement d'anonymisation pour les données en cours de collecte**

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

Discussion

Discussion

Bilan

Perspectives

Discussion

Impact de l'anonymisation by design sur l'apprentissage ?

- Appauvrissement de la donnée *versus* toutes les données accessibles
- *data-driven* ?
  - *Deep learning* généralement appliqué sur l'image elle-même
- Qualité des données et exploration :
  - Étape chronophage et essentielle dans l'apprentissage
  - Opération réalisée lors du traitement et de la vérification de l'anonymisation

Surveillance des données libérées

- Traçabilité pour pouvoir intervenir en cas d'apparition de nouveaux évènements conduisant à évacuer les data
- En pratique ne repose que sur du contrat et donc de la confiance que le tiers réalise l'action

Bilan

Rôle clef de l'expression de la finalité et de la minimisation des données

- Principes venant de la *Data Privacy*
- Permet de s'approprier les données avant l'apprentissage
- Simplifie drastiquement les opérations d'anonymisation ultérieure car cela diminue fortement les corrélations potentielles entre les différentes variables
- Simplification de la prise de décision
- Permet de mieux exploiter les méthodes d'anonymisation

Garanties apportées à la sécurité des données

- Respect de l'approche *Privacy By Design*
- Export uniquement quand anonymisation qualifiée

Facilite les échanges avec la CNIL

- Réalisation du *Privacy Impact Assement* en parallèle
- Fourni des éléments de preuves

- Base FIDAC
- Plateforme AI4EU

Mise à disposition du jeu pour la communauté IA

- Programme CQXD de l'AID
- Partenariat Thales & NEHS DIGITAL

Un PoC de triage du COVID via téléradiologie

- *medical Imaging Report Anonymiser* (mIRA)
- Plateforme AI4EU

Docker d'anonymisation de compte-rendu

Perspectives

Évolutions réglementaires

- CNIL va faciliter l'utilisation secondaire des données avec des cadres déclaratifs et non plus soumis à validation
- AI Act

Évolution de la méthode

- Qualité de la data pour exclure automatique les infos lisibles de DM implantables (n° de série de pacemaker, implants, broche…)

#Merci
#Questions ?

nehs DIGITAL