4ème journée scientifique des cohortes Constances et Gazel

11 mai 2017

Qu'est ce que le SNDS ?

L'article 193 de la Loi de modernisation de notre système de santé crée le Système national des données de santé (SNDS).

Le SNDS, géré par la Caisse nationale d'assurance maladie des travailleurs salariés, a vocation à réunir :

- les données de l'assurance maladie,
- les données des établissements de santé,
- les causes médicales de décès,
- les données des Maisons départementales des personnes handicapées,
- un échantillon représentatif des données de remboursement des organismes d'assurance maladie complémentaire.

! Les données sont pseudonymisées.

A quoi sert le SNDS ?

Le SNDS met à disposition les données pour contribuer à :

- l'information sur la santé ainsi que sur l'offre de soins, la prise en charge médico-sociale et leur qualité ;
- la définition, à la mise en œuvre et à l'évaluation des politiques de santé et de protection sociale ;
- la connaissance des dépenses de santé, des dépenses d'assurance maladie et des dépenses médico-sociales ;
- l'information des professionnels, des structures et des établissements de santé ou médico-sociaux sur leur activité ;
- la surveillance, à la veille et à la sécurité sanitaires ;
- la recherche, aux études, à l'évaluation et à l'innovation dans les domaines de la santé et de la prise en charge médico-sociale.

Qui a accès au SNDS ?

Le SNDS constitue une avancée considérable pour améliorer la connaissance sur la santé des citoyens.

Qui peut accéder aux données du SNDS ?

-> Les organismes exerçant une mission de service public cités dans le décret n° 2016-1871 du 26 décembre 2016 relatif au traitement de données à caractère personnel dénommé « système national des données de santé » disposent d'un accès permanent aux catégories de données = plus de 2000 utilisateurs.

-> Les autres acteurs publics et privés doivent déposer une demande auprès de l'Institut national des données de santé (INDS) qui se charge de la transmettre au Comité d'expertise pour les recherches, les études et les évaluations dans le domaine de la santé (CEREES) puis à la CNIL.

Quelles sont les conditions d'accès ?

- L'accès aux données est autorisé sous conditions :
  - l'intérêt public,
  - deux finalités interdites :
    - la promotion en direction des professionnels et établissements de santé,
    - l'exclusion de garanties des contrats d'assurance et la modification de primes d'assurance.
- L'accès aux données s'effectue dans des conditions assurant la confidentialité et l'intégrité des données et la traçabilité des accès et des autres traitements => un référentiel de sécurité est applicable aux données du SNDS.

Quel est l'impact du référentiel de sécurité (arrêté du 22 mars 2017) ?

- Il s'applique à l'ensemble des systèmes mettant à disposition des données du SNDS, seules ou appariées à d'autres données (ex : cohortes appariées au SNDS).
- Il a été conçu en anticipant l'entrée en vigueur du règlement européen sur la protection des données personnelles.
- Il garantit la sécurité des traitements en imposant :
    - la pseudonymisation,
    - l'authentification,
    - la traçabilité,
    - le contrôle,
    - la sensibilisation et la formation.

Ce qu'il faut retenir :

- Engagement de confidentialité et de respect du secret professionnel par chaque utilisateur.
- Tous les utilisateurs doivent être sensibilisés.
- La réidentification est proscrite dans tous les cas.
- Les fuites de données (qui peuvent avoir un impact grave sur la vie privée des personnes) sont souvent dues à une action humaine (erreur ou acte intentionnel). Il faut ainsi mettre en place une sensibilisation pour limiter les incidents de sécurité.
- La traçabilité de toutes les actions sera effective et l'information des sanctions mises encourues en cas de non-respect des consignes doit être diffusée.

Ce qu'il faut retenir :

Les données du SNDS (ou appariées à d'autres données) doivent être dans un espace garantissant le référentiel de sécurité adapté à la sensibilité des données, ce qui implique le recours à l'analyse de risque.

=> Toutes les données composant le SNDS, où qu'elles soient stockées, sont concernées par l'application du référentiel.

L'utilisateur s'engage à respecter le référentiel, notamment :

- l'absence d'actions visant la réidentification,
- l'obligation de ne diffuser que des données anonymes.

Ce qu'il faut retenir :

- **Contrôle et sanctions**

- Les habilitations doivent être régulièrement revues.

- Le SNDS élargi sera périodiquement contrôlé (CNIL, comité d'audit, Cour des comptes, ANSSI, IGAS…).

- En cas d'infraction, des sanctions adéquates peuvent être prononcées, notamment la fermeture de facto de l'accès aux données pour tout l'organisme.

- La rupture du secret professionnel est pénalement répréhensible.

Ce qu'il faut retenir :

- **L'application du référentiel de sécurité**

- Pour tout nouveau système, il s'applique en totalité,

- Pour tout système existant, un délai de deux ans est prévu pour que les gestionnaires concernés se mettent en conformité.

Quel accompagnement ?

- Un groupe de travail a mené des travaux sur la notion de risque faible et un rapport présentant des pistes d'action et des cas concrets est en cours de finalisation.

- **Objectif** : accompagner les gestionnaires de bases dans leur évaluation du niveau de risque de leurs données.

- **Position de la CNIL :**
  - A confirmé l'intérêt d'utiliser la méthodologie « PIA : Privacy Impact Assessment »
  - Risque de la base = risque de la personne la plus exposée

Étude d'impact sur la vie privée

- Evaluation du risque d'atteinte à la vie privée sur deux axes :

Gravité + **Vraisemblance** = **Niveau de Risque**

- La **gravité** estime l'ampleur des impacts :
  - **Niveau 1 :** Pas d'impact, désagréments mineurs
  - **Niveau 2 :** Désagréments significatifs mais surmontables
  - **Niveau 3 :** Désagréments significatifs difficiles à surmonter
  - **Niveau 4 :** Conséquences significatives insurmontables

- La **vraisemblance** traduit la possibilité qu'un risque se réalise.

Étude d'impact sur la vie privée

- **Problème : comment estimer la vraisemblance et la gravité ?**

- **Gravité :** c'est la caractère **discriminatoire** de l'état de santé d'une personne qui est porteur de risque.

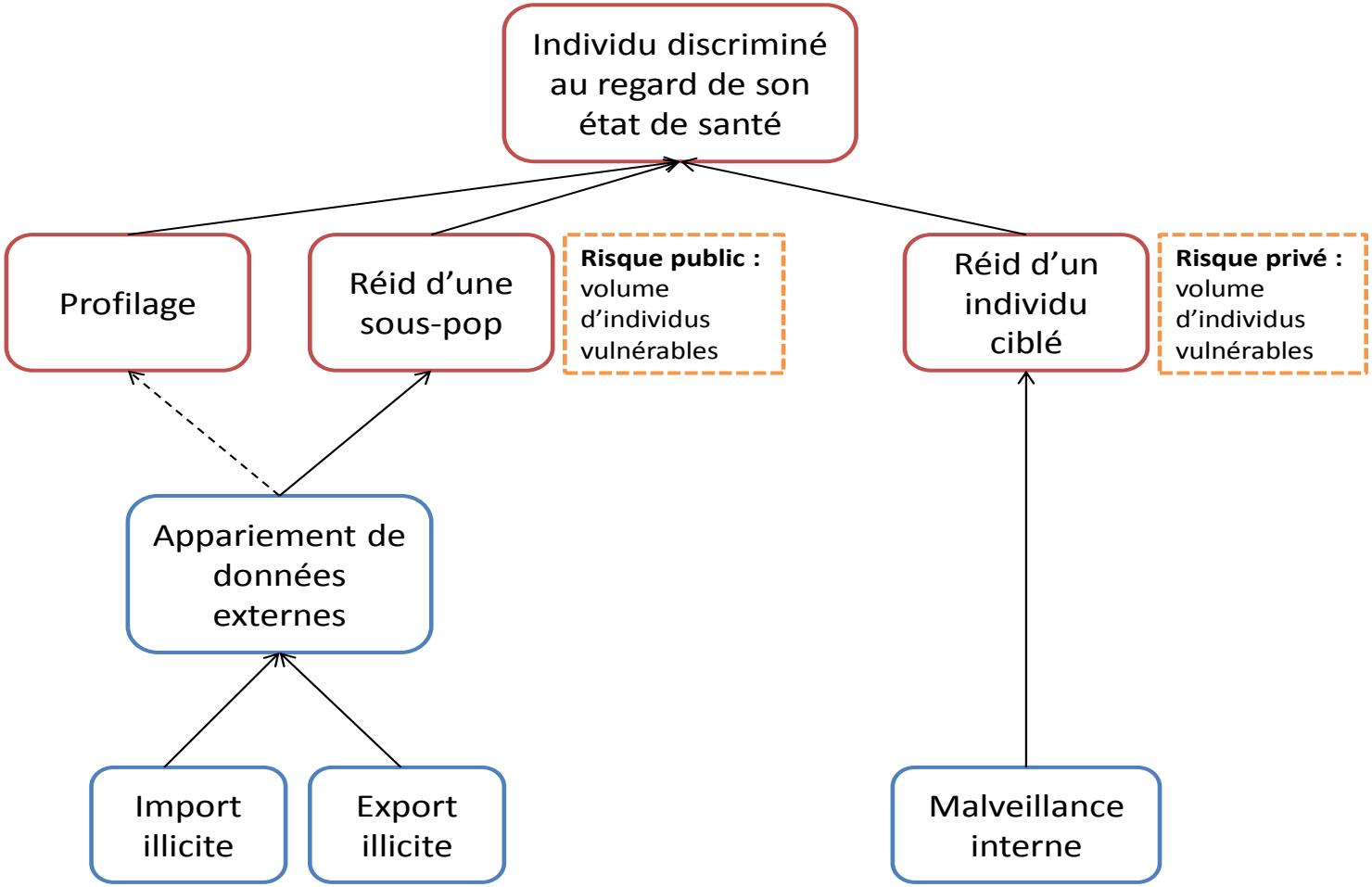
- La **vraisemblance** d'un usage malveillant se décompose en :

  Vraisemblance = Probabilité d'une tentative X Probabilité de succès.

- Quantifier la probabilité de succès nécessite de dénombrer les individus vulnérables à une attaque pouvant impacter leur vie privée.

- Probabilité d'une tentative comporte deux dimensions principales :
  - La facilité pour un attaquant à rassembler les données nécessaires
  - Le volume d'attaquants potentiels

Estimation de la vraisemblance



Risque privé vs. risque public

• **Risque privé :** risque qu'un attaquant apprenne des informations discriminantes sur un individu de son entourage.

➢ *ex : un employeur connaissant des dates de soin.*

• **Risque public :** risque qu'un attaquant apprenne des informations discriminantes sur individu inconnu, sur la base d'informations publiquement disponibles.

➢ *ex : un appariement avec des listes électorales.*

Quasi-identifiants publics	Âge, sexe, code de résidence
Quasi-identifiants privés	Lieu et établissement d'hospitalisation Certaines pathologies (grossesse, diabète…)? Durée du séjour, dates de soins , mode de sortie

Détection des individus à risque

- Nécessité de choisir une mesure imputée à un niveau individuel pour détecter des individus à risque : le risque ne se mutualise pas (cf. rencontre avec la CNIL)

- On mesure le risque de divulgation du fait : *« état de santé pouvant mener à une discrimination »*

- Par exemple, estimation en deux étapes :
  1. Mesure du risque de divulgation public sachant l'appartenance à l'échantillon
  2. Évaluation du caractère protecteur de l'échantillonnage : taux et profils spécifiques

- **Mesure « sachant que » :** part des individus affectés d'un état de santé discriminant au sein de leur sous-groupe « public »

Age	CP	Sexe	Diag	Risque de divulgation	Discriminant
43	75013	M	Brûlures	100 %	0
55	84000	F	Brûlures	25 %	0
55	84000	F	**Cancer du sein**	75%	1
55	84000	F	**Cancer du poumon**	75%	1
55	84000	F	**VIH**	75%	1

- **Échantillonnage :** part des individus uniques qui sont également uniques dans la population de référence (par exemple)

En résumé

- Quelles informations le jeu de données permet-il de **connaître sur les pathologies** des individus ?
- Le jeu de données est-il **échantillonné** et si oui à quel taux et avec quelle méthode ?
- Le risque de **réidentification public** est-il nul ?

Applications concrètes

Jeux de données étudiés :

- PMSI sans chainage patient
- Extractions de la cohorte CONSTANCES
  - **Risque public « sachant que » :** environ **1%** des individus à risque
  - **Échantillonnage :** environ **0,5%** des uniques sont également uniques dans la population de référence
- Grand EGB
- Base agrégée des CMDC
- EGB simplifié
- Enquête HSM de la DREES

Conclusions transverses :

- Le risque privé est toujours élevé.
- les jeux exhaustifs à granularité fine ne permettent pas d'atteindre un risque public nul, des dégradations sont nécessaires.
- L'échantillonnage aléatoire protège généralement bien d'une ré-identification déterministe.

Contact

- mylene.girard@sante.gouv.fr
- matthias.pigneur@sante.gouv.fr

Pour contacter l'équipe de la mission d'accès aux données de santé :

- DREES-MADS@sante.gouv.fr