Ministère des finances et des comptes publics
Ministère des affaires sociales, de la santé et des droits des femmes
Ministère du travail, de l'emploi, de la formation professionnelle et du dialogue social

santé protection sociale
drees
recherche, études, évaluation et statistiques

Risques de ré-identification dans les bases de données de santé et moyens de s'en prémunir

Introduction

André Loth (DREES)

Protection de la vie privée et/ ou ouverture des données de santé ?

- Positions tranchées et accusations
- Concilier les deux mais comment ?
- Présenter l'état des réflexions académiques et prolonger les travaux du GT RiRe
- **Deux sujets différents** :
  - Des données « vraiment anonymes » ouvertes à tous :
    → assurer leur anonymat (tout en les multipliant)
  - Des données personnelles déidentifiées en accès régulé
    → les protéger tout en les ouvrant davantage (à qui et sous quelles conditions ?)

Données individuelles, données personnelles, données de santé

- Toutes les données personnelles sont individuelles mais il y a des données individuelles anonymisées (*i.e.* sans caractère personnel)

- « *Constitue une donnée à caractère personnel toute information relative à une personne physique identifiée ou qui peut être identifiée, directement ou indirectement, par référence à un numéro d'identification ou à un ou plusieurs éléments qui lui sont propres. Pour déterminer si une personne est identifiable, il convient de considérer l'ensemble des moyens en vue de permettre son identification dont dispose ou auxquels peut avoir accès le responsable du traitement ou toute autre personne.* » (loi du 6 janvier 1978 article 2)

- Données personnelles de santé (= maladies) : données « sensibles »

- Traitement des données sensibles (art. 8 LIL) : interdiction sauf…

Quatre éclairages

1. Pour assurer l'anonymat il ne suffit pas de dé-identifier (pseudonymiser) les données
2. Quels critères d'anonymat pour les jeux de données en accès libre ?
3. Quelles conditions pour sécuriser l'usage des données personnelles de santé déidentifiées ?
4. Aperçu sur la question du NIR dans les bases de données de santé

Pour assurer l'anonymat il ne suffit pas de dé-identifier les données

- **Condition nécessaire** : le masquage (pseudonymisation, dé-identification) des données directement identifiantes (nom, NIR, n° de téléphone…)

- Contraintes techniques et organisationnelles (besoin de chaînage, séparation entre données et identités) et l'outil du chiffrement

- **Mais non suffisante** : l'exemple du PMSI et des « quasi identifiants » notoires dans une base précise et exhaustive

- Des personnes à caractéristiques uniques, ou à caractéristiques communes mais ayant la même maladie

Quels critères d'anonymat pour les jeux de données en accès libre ? (1)

- Objectif : éviter l'identification certaine… ou la très forte probabilité
- Assez facile pour les données agrégées mais non pour les données individuelles (« granulaires »)
- Trois exemples :
  A. 5 % des individus dans une base A présentent des caractéristiques uniques
  B. Dans une base B il n'y a jamais moins de 5 personnes qui présentent les mêmes caractéristiques
  C. Dans un échantillon C de la population, au 1/5, des individus présentent des caractéristiques uniques

La base A n'est pas anonyme mais B et C le sont-elles ?

Quels critères d'anonymat pour les jeux de données en accès libre ? (2)

- K-anonymité et L-diversité : valeurs cibles pour K et L ?
- Les moyens : appauvrir (flouter, supprimer des données), bruiter (ajouter ou recomposer des données), échantillonner…
- Arbitrer entre anonymat et utilité ? Calculer un ratio bénéfice/ risque ? Mais qui bénéficie, qui risque et qui arbitre ?
- Risques de reconstituer des jeux riches à partir de plusieurs jeux appauvris
- Echantillons protégés ? Oui sous certaines conditions
- Quel consensus ? Qui décide ? Liberté de mettre en ligne les données qu'on estime anonymes… à condition de ne pas se tromper
- Jeux anonymes en open data et solutions intermédiaires en cas de doute ?

Sécuriser l'usage des données personnelles de santé ?

- Données déidentifiées à risques de réidentification
- Gestion séparée des identifiants et des données (service du secret, tiers de confiance…)
- Accès restreint :
  - Bonnes raisons (données nécessaires à l'étude, rigueur, transparence, « intérêt public » (= intérêt général))
  - Bonnes personnes (expertise, habilitation), peu nombreuses
  - Tenues au secret (sous peine de sanctions)
  - Traçabilité (condition des sanctions et de la dissuasion)
  - Mêmes règles pour toutes les données similaires

Aperçu sur la question du NIR dans les bases de données de santé

- Le NIR un équivalent strict de « nom-prénom-date et lieu de naissance », sans les homonymies et avec moins d'erreurs
- Très utile pour les usages métier (sécurité sociale et désormais dossiers médicaux) mais critiqué parce qu'il est signifiant
- Dans les bases destinées à la connaissance, le NIR doit être masqué (comme le nom !)
- Vu il y a 35 ans comme indispensable pour croiser des fichiers (nominatifs) : ça n'est plus vrai
- Les chercheurs et assimilés sont les seuls que l'encadrement du NIR gêne aujourd'hui pour le « croisement de fichiers » parce que les fichiers en question sont déidentifiés et utilisent (généralement) un NIR chiffré comme unique identifiant
- Des solutions à l'endroit et à l'envers…

Risque de réidentification : comment s'en prémunir ?

- Les données véritablement anonymes peuvent et doivent être mises en open data, avec des jeux de données riches susceptibles de répondre à beaucoup de besoins pour lesquels on utilise aujourd'hui les bases riches. Mais il y a une zone grise

- Pour les données *personnelles* de santé, la loi protège la vie privée, limite les droits d'accès, punit, dissuade…

- Mais pas de sanction ni de dissuasion possible sans traçabilité… ce qui peut poser des problèmes d'ergonomie (et de coûts)

- Aucune ambiguïté sur l'objectif : ouverture la plus large possible (maximum de données, maximum d'utilisations) compatible avec la protection de la vie privée