CépiDc
Centre d'épidémiologie sur les causes médicales de décès

aviesan
alliance nationale pour les sciences de la vie et de la santé

Inserm
Institut national de la santé et de la recherche médicale

Appariement aux bases de donnés médico-administratives : intérêt pour la recherche et la santé publique

Séminaire : Ouverture des données de santé : comment se prémunir du risque de ré-identification ?

Grégoire Rey, Karim Bounebache

Inserm – CépiDc, Le Kremlin-Bicêtre

Plan

- Contexte
- Les bases de données médico-administratives et sociales
- Les différentes utilisations de ces bases
- Les méthodes et les freins aux appariement
- Loi de santé et perspectives

Contexte

- Paradoxe entre risque de réidentification et appariement
  - Ne fournir aucun jeu de données pour lequel une combinaison de variables est identifiante

Vs.

  - Avoir un identifiant ou une combinaison de variables permettant d'apparier les observations avec une bonne fiabilité

⇒Cohérent si cet appariement est encadré et présente un intérêt pour la santé publique

Contexte

- **La France est un des rares pays européens qui dispose de bases de données administratives et médico-administratives nationales**

- Gérées par des organismes publics
  - CNAMTS, ATIH : SNIIRAM-PMSI
  - Inserm : mortalité (CépiDc)
  - Cnav : RNIAM, SNGC
- Concernant des domaines stratégiques
  - Recours aux soins (actes médicaux, prestations effectuées)
  - Médecine de ville et Hospitalisation
  - Prestations et situation professionnelle et sociale
- Couvrant la population de façon exhaustive et permanente
  - Puissance statistique et représentativité
  - Potentiel d'utilisation très diversifiée (chaque base séparément, appariement)
- Centralisées

SNIIRAM-PMSI

- Prestations remboursées (DCIR)
  - Nature de prestation (y.c. médicaments, dispositifs, biologie)
  - Discipline de prestation
  - Codes pathologie ALD, MP
  - Professionnel de santé exécutant
  - Professionnel de santé prescripteur
  - Théoriquement exhaustif
- Hospitalisation (PMSI) – depuis 2006
  - Diagnostic principal, associés, reliés
  - Actes
  - Indicateur de gravité
  - Montant ou volume de la prestation
- Date de décès – depuis 2008
  - Montée en charge progressive
- NIR foinisé
- EGB, échantillon au 1/97ème

Informations sociales :
-CMU, AME,
-Commune de domicile

CépiDc

- Décès depuis 1968 (24 millions d'observations)
- Exhaustif pour les décès survenus sur le territoire
- Pour tous les décès
  - Causes de décès (initiale, associées)
  - Code CIM et texte renseigné
  - Quelques informations sociales (activité, profession, statut matrimonial, commune de domicile)
- Pas d'identifiant direct

D’autres données de santé

- Registres de maladie
  - cancers, maladies rares, malformations congénitales, cardiopathies ischémiques,…
  - informations nominatives
- Certificats de santé de l’enfant
- Donneurs et receveurs de sang, produits reçus (EFS)
  - pas de recueil du NIR,
- Déclaration évènements indésirables (ANSM)
  - pas d’identifiant direct
- Maladie à déclaration obligatoire
  - données non nominatives

CNAV

- SNGC (Système National de Gestion des Carrières)
  - Employeur, NAF et PCS à 4 ou 2 chiffres
  - Salaire
  - Périodes d'interruption
    - *Santé (maladie, invalidité, accidents du travail)*
    - *Famille (maternité, famille : parent au foyer, enfant ou adulte handicapé)*
    - *Activité professionnelle (chômage, préretraite, insertion)*
  - Obligations militaires
- SNSP (Système National Statistiques des Prestataires)
  - Eléments de calcul de la pension
  - Droits
  - Avantages complémentaires
- NIR (SNGI)

Autres données sociales

- INSEE
  - Recensement
  - EDP (échantillon démographique permanent)
  - Panel DADS
- CNAF (allocations familiales)
- Pôle emploi
- DGI (Direction générale des impôts)

Différentes utilisations de ces bases

- Exploitation des données d'une BDMA seule

  - Economie de la santé
    - *Estimation des coûts de différentes stratégies thérapeutiques*
    - *Recherche sur l'organisation des soins*
  - Epidémiologie descriptive
    - *Estimations de prévalence et/ou d'incidence de certains cancers à partir des données du PMSI*
    - *Variations géographiques et temporelles,*
  - Pharmacoépidémiologie
    - *Estimation de la surmortalité associée au Mediator*
    - *Suivi post-AMM*

Difficultés :

1. Pas de repérage d'évènements de santé en dehors du soin
2. Ajustement sur ou analyse du niveau socio-économique limité (CMU, AME, commune de domicile)

Différentes utilisations de ces bases

Inserm CépiDc

- Exploitation de plusieurs bases nationales

  - Economie de la santé
    - *Mesure de qualité des soins : différentiels de risque de mortalité post-hospitalière selon l'établissement (Projet AMPHI, appariement indirect causes de décès – SNIIRAM-PMSI)*
    - *Analyse des arrêts maladie (Projet HYGIE, appariement SNIIRAM-CNAV)*

  - Epidémiologie descriptive
    - *Différentiels socioéconomiques de morbidité (SNIIRAM-CNAV)*
    - *Différentiels socioéconomiques de mortalité (EDP-Cause de décès, CNAV-Cause de décès)*

  - Pharmacoépidémiologie (SNIIRAM-CNAV)
    - *Mesure de l'efficacité-efficience d'un médicament*
    - *Prise en compte des variables sociales : ajustement sur facteurs de confusion*
    - *Interaction entre variables sociales et prise de médicaments (différentiel d'observance, combinaison avec d'autres facteurs)*

Différentes utilisations de ces bases

- Appariement à des données d'enquêtes

  - Base de tirage de l'enquête
    - *Meilleure caractérisation des non répondants*
    - *Redressement par pondération ou imputation*
    - *Cohorte Constances (SNIIRAM-CNAV-CépiDc)*

  - Evite de reposer des questions sur la prise de médicaments
    - *Taille des questionnaires réduite*
    - *Limite les biais de déclaration (perception, mémoire,…)*
    - *Enquête HSM (Handicap) et ELFE (Enfants) (SNIIRAM)*

  - Suivi minimal dans le temps
    - *Présent sur le territoire, vivant, malade ou en bonne santé*

  - Analyse sur la qualité des données
    - *Comparaison à un gold standard, élaboration d'algorithmes*

→ La quasi-totalité des cohortes en population devrait être appariées aux BDMA

Méthodes d'appariement

- 3 principaux cas de figure

  - Présence du NIR (foinisé ou non) dans les deux bases
    - *Méthode simple si NIR correctement saisie (certifié)*

  - Présence de données nominatives et/ou d'adresse
    - *Méthode de cryptage des noms/adresses*
    - *Fonction de proximité phonétique et/ou orthographique*
    - *Appariements indirects*

  - Données indirectement identifiantes
    - *Exemple : sexe, date de naissance, commune de naissance, commune de domicile*
    - *Appariements indirects*

Méthodes d'appariement

- Deux types d'appariement indirect

  - Appariement déterministe
    - *utilisation d'une « métrique » spécifique pour mesurer la dissemblance, choix d'un seuil de décision apparié/non apparié.*
    - *application de règles de décision à dires d'experts (souvent les producteurs de données)*
    - *démarche non généralisable, peu reproductible et difficile à documenter*
    - *analyse des erreurs limitée*

  - Appariement probabiliste
    - *Fellegi et Sunter : formalisation rigoureuse de l'appariement probabiliste.*
    - *calcul explicite des probabilités de concordance des champs*
    - *règles de décisions et seuils théoriquement et implicitement produites dans le modèle, en optimisant des critères de classification :*
      - probabilité d'appariement à tort
      - probabilité de non appariement à tort
    - *apprentissage machine limité en pratique (aspect computationnel, orientation à priori des formes de probabilités d'erreurs)*

    → *mélange déterministe/probabiliste*

Freins actuels aux appariements

- Procédures d'accès aux données complexes inhérentes à la LIL et à des textes particuliers

  - Données sensibles : plusieurs régimes d'autorisation de la LIL dont :
    - ***Chapitre IX :*** *traitements ayant pour fin la recherche en santé autorisés par la CNIL, après avis d'un comité scientifique placé auprès du ministre de la recherche (le CCTIRS)*
    - ***Chapitre X :*** *traitements ayant pour fin l'évaluation des pratiques de soins, à partir de données indirectement nominatives (à l'exclusion des noms, prénom et du NIR) issues des dossiers médicaux des professionnels de santé libéraux, des systèmes d'information hospitaliers et des systèmes d'information des différentes caisses d'assurance maladie, autorisés par la CNIL selon une procédure d'autorisation sans avis préalable d'un comité scientifique*

Freins actuels aux appariements

- Procédures d'accès aux données complexes inhérentes à la LIL et à des textes particuliers

  - Régime particulièrement protecteur conféré au NIR (art. 27 et 29 LIL)
    - *obstacle juridique aux appariements, le NIR (ou un dérivé) étant la clé d'accès à de nombreux fichiers de l'assurance maladie ou vieillesse*
    - *Décret en CE nécessaire pour les organismes publics (même si les chercheurs ne collectent pas le NIR)*
  - Foisonnement d'organismes/juxtaposition de procédures
    - *recherche biomédicale et en soin courant relèvent des CPP + CNIL*
    - *accès au SNIRAM strictement encadré par l'arrêté actuellement en vigueur le 19 juillet 2013 modifié, pris après avis de la CNIL, et un protocole qui définit ses modalités de fonctionnement*
    - *accès aux données couvertes par le secret statistique (INSEE), nécessite un accord du comité du secret statistique*
    - *aucune structure dédiée à la recherche ne dispose d'un accès pérenne et complet aux données individuelles du SNIIRAM (DCIR)*

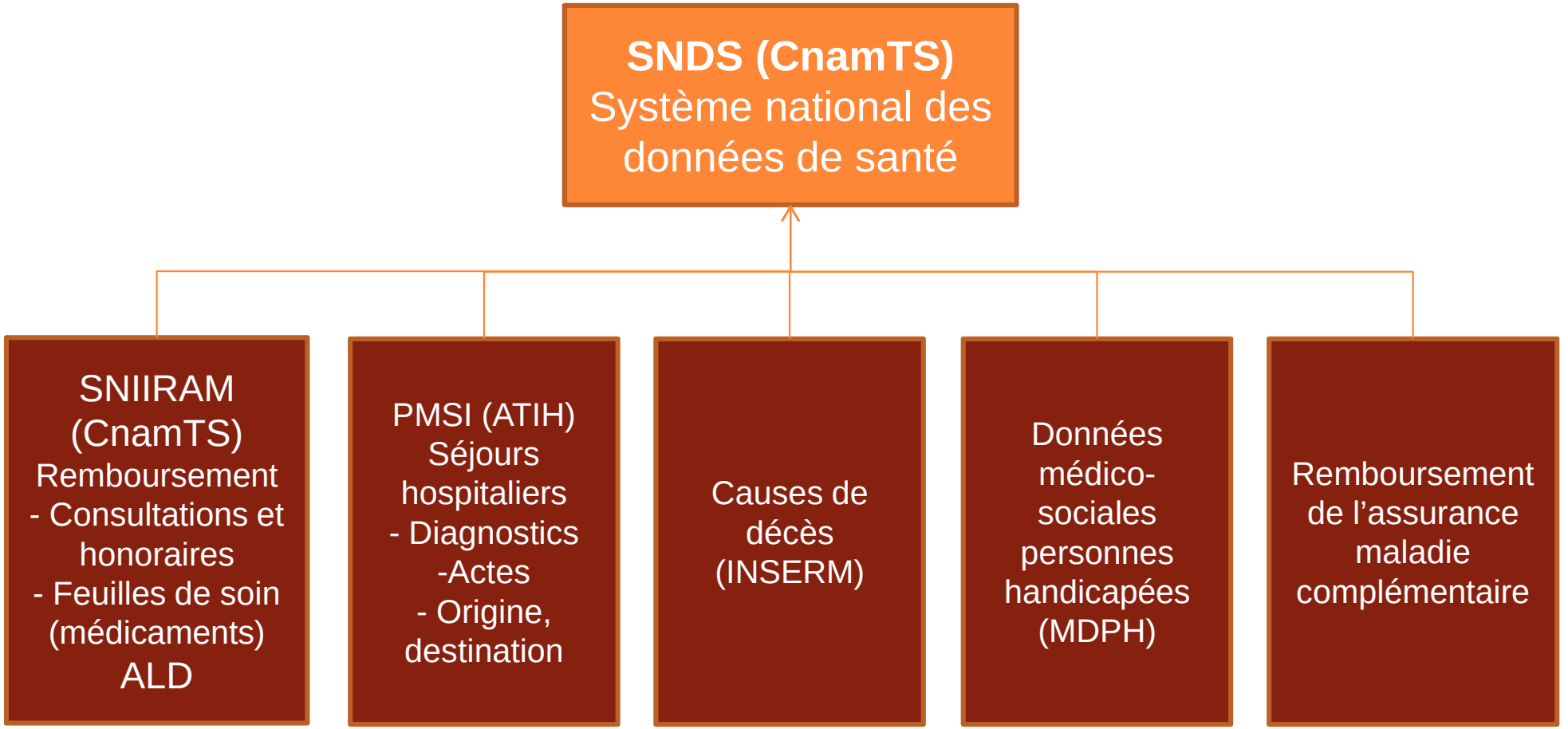
- « Parcours du combattant administratif » impossible à surmonter pour la plupart des structures

Freins actuels aux appariements

- Pas de structure identifiée dédiée aux appariements complexes pour la recherche
  - Conseil sur la sécurité SI et les aspects juridiques pour la mise en œuvre d'appariements impliquant le NIR ou des données identifiantes
  - Maîtrise des méthodes d'appariement, du contenu et de la comparabilité des champs sur lesquels apparier
  - Conseil méthodologique sur les biais potentiels et les méthodes spécifiques à mettre en œuvre sur des données lorsque l'appariement est imparfait

Perspectives : loi de santé

- Création du SNDS



Perspectives : loi de santé

- Aspects positifs
  - Centralisation accrue des données au sein d'un même système d'information : le SNDS
  - Assouplissement des conditions d'utilisation du NIR pour permettre les appariements entre les données d'enquêtes et les données issues des BDMA (Remplacement du décret en CE par une autorisation de la CNIL)
  - Choix du NIR comme Identifiant National de Santé (INS)
  - Application de procédures simplifiées pour des études demandées pour la puissance publique ou dans les cas standards
  - Fusion chapitres IX et X, LIL

Perspectives : loi de santé

- Points d'attention : 1. Utilisation du NIR

  - Elaboration d'une méthodologie de référence ou d'un guide de bonnes pratiques nécessaire.

  - Permettrait de simplifier le travail du comité d'expertise scientifique et de la CNIL et de réduire les délais d'instruction (un nombre de demandes très conséquent à prévoir)

  - L'évolution de l'article 54, alinéa 5 de la loi « Informatique et liberté » permettrait l'élaboration de méthodologies de référence (MR) y compris impliquant le NIR.

Perspectives : loi de santé

- Points d'attention : 2. Accès au SNDS

  - Des instituts de recherche doivent figurer dans la liste :
    - *« des services de l'Etat, des établissements publics ou des organismes chargés d'une mission de service public, autorisés à traiter des données à caractère personnel du SNDS pour les besoins de leurs missions »*
    - *« des organismes chargés de la mise à disposition effective des données du SNDS »*

  - A court terme, modification de l'arrêté SNIIRAM

Perspectives : loi de santé

- Points d'attention : 3. Causes de décès

  - Modification de l'article L.2223-42 du CGCT pour autoriser le transfert des données de causes médicales de décès à l'INSEE

  - Risques si appariement direct possible entre données nominatives-causes de décès :
    - *changement du statut des données,*
    - *modification de la déclaration des causes de décès par les médecins certificateurs,*
    - ***moindre pertinence des données pour une exploitation en santé publique**.*

Perspectives : infrastructure de service

- Participants à sa définition
  - ITMO Santé publique
  - Chercheurs en santé publique (épidémiologie, informatique médicale, biostatistiques,…)
  - Fournisseurs de données (CNAMTS, CépiDc)
  - Agences sanitaires (ANSM, InVS)

Déclinaison en 6 WP

WP1 Organisation de l'interface

WP2 Aspects techniques d'une interface

WP3 Interopérabilité, développement d'algorithmes et qualités des données

WP4 Appariements

WP5 Aspects légaux, réglementaires et éthiques

WP6 Typologie des besoins ; partage des besoins et des expériences

Perspectives : infrastructure de service

- IEPI (projet Interface pour l'épidémiologie) – ITMO Santé publique
  - Sécurité SI et aspects juridiques pour la mise en œuvre d'appariements impliquant le NIR ou des données identifiantes
  - Maîtrise des méthodes d'appariement, du contenu et de la comparabilité des champs sur lesquels appariés
  - Conseil méthodologique sur les biais potentiels et les méthodes spécifiques à mettre en œuvre sur des données lorsque l'appariement est imparfait

Conclusion

- Les appariements : élément essentiel pour une recherche de qualité en Santé Publique

- Des perspectives positives mais avec des points d'attention dans la loi de santé

- Propositions
  - Evolution de la PF mutualisée de services vers une infrastructure de services pour les chercheurs
  - Proposer des « méthodologies de référence » et des guides de bonnes pratiques

Remerciements

- Geneviève Chêne
- Frédérique Lesaulnier
- Charles Persoz
- Anita Burgun
- Marcel Goldberg
- Catherine Quantin
- Dina Oksen

CépiDc
Centre d'épidémiologie sur
les causes médicales de décès

Inserm
Institut national
de la santé et de la recherche médicale

Merci pour votre attention