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

Un test d'anonymisation des données PMSI

Maxime BERGEAT – Insee, Département des méthodes statistiques

&

Noémie JESS – Drees, Bureau des dépenses de santé et des relations avec l'assurance maladie

Insee
Mesurer pour comprendre

Plan d'intervention

1. Le patient ausculté (le PMSI)
2. Le protocole opératoire
3. Les actes pratiqués et le diagnostic final

Insee
Mesurer pour comprendre

Plan d'intervention

1. Le patient ausculté (le PMSI)
2. Le protocole opératoire
3. Les actes pratiqués et le diagnostic final

Insee
Mesurer pour comprendre

1. Le PMSI

- ***P**rogramme de **M**édicalisation des **S**ystèmes d'**I**nformation* : base médico-administrative contenant l'intégralité des séjours hospitaliers en France

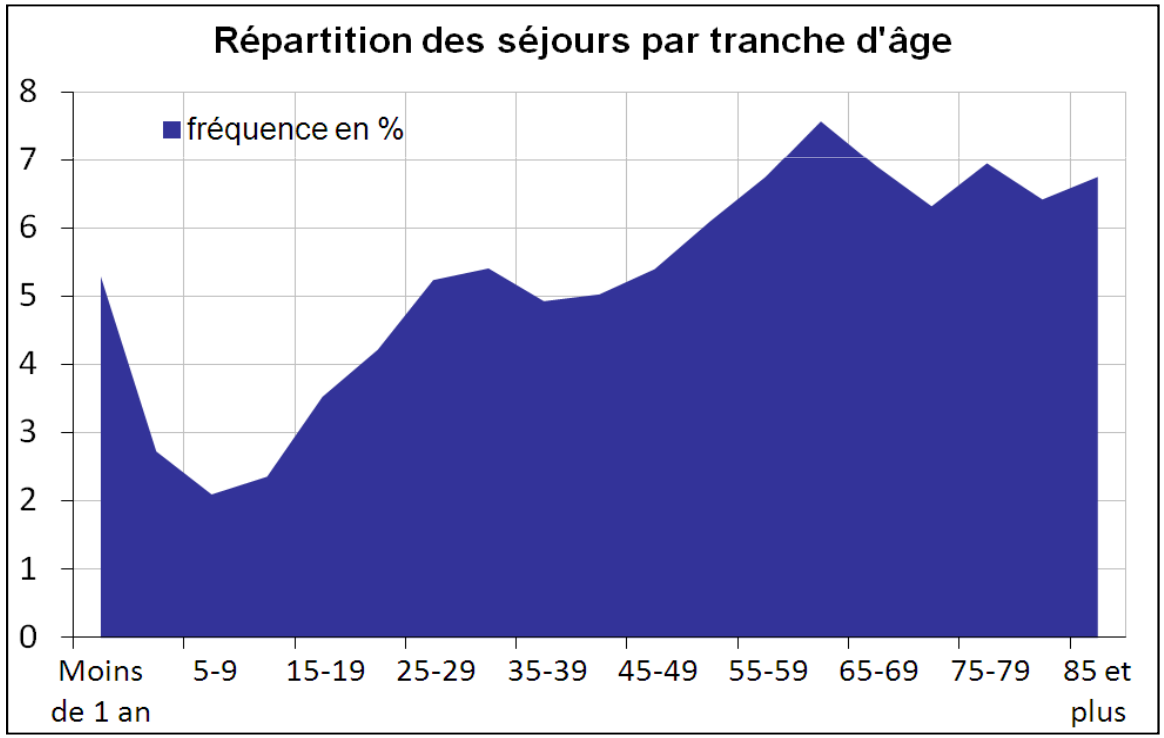
- Test porte sur le PMSI-MCO, *i.e.* les courts séjours
  - Exclusion des séances
  - Pas de chaînage des séjours entre eux

- Information médicale agrégée : la CMD (Catégorie majeure de diagnostic)

Insee
Mesurer pour comprendre

1. Le PMSI

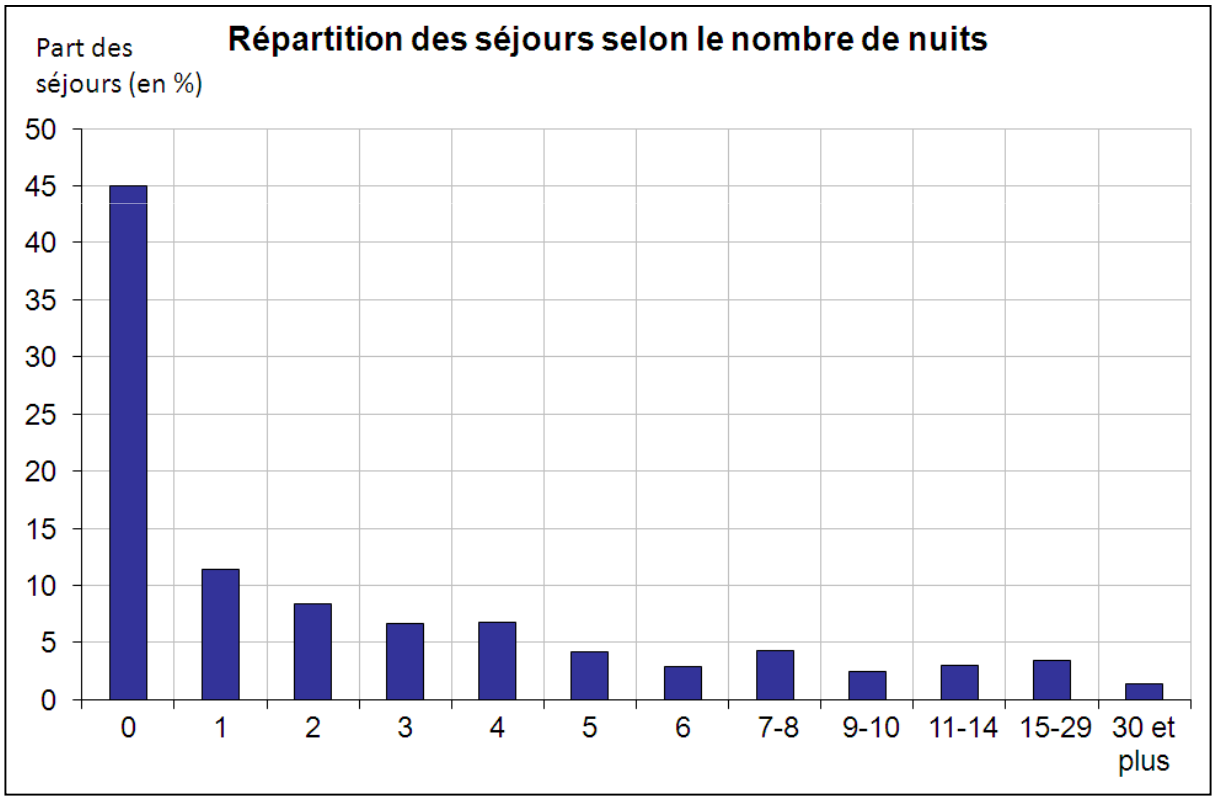
- Distribution non uniforme des séjours au sein de la population selon l'âge :



Insee
Mesurer pour comprendre

1. Le PMSI

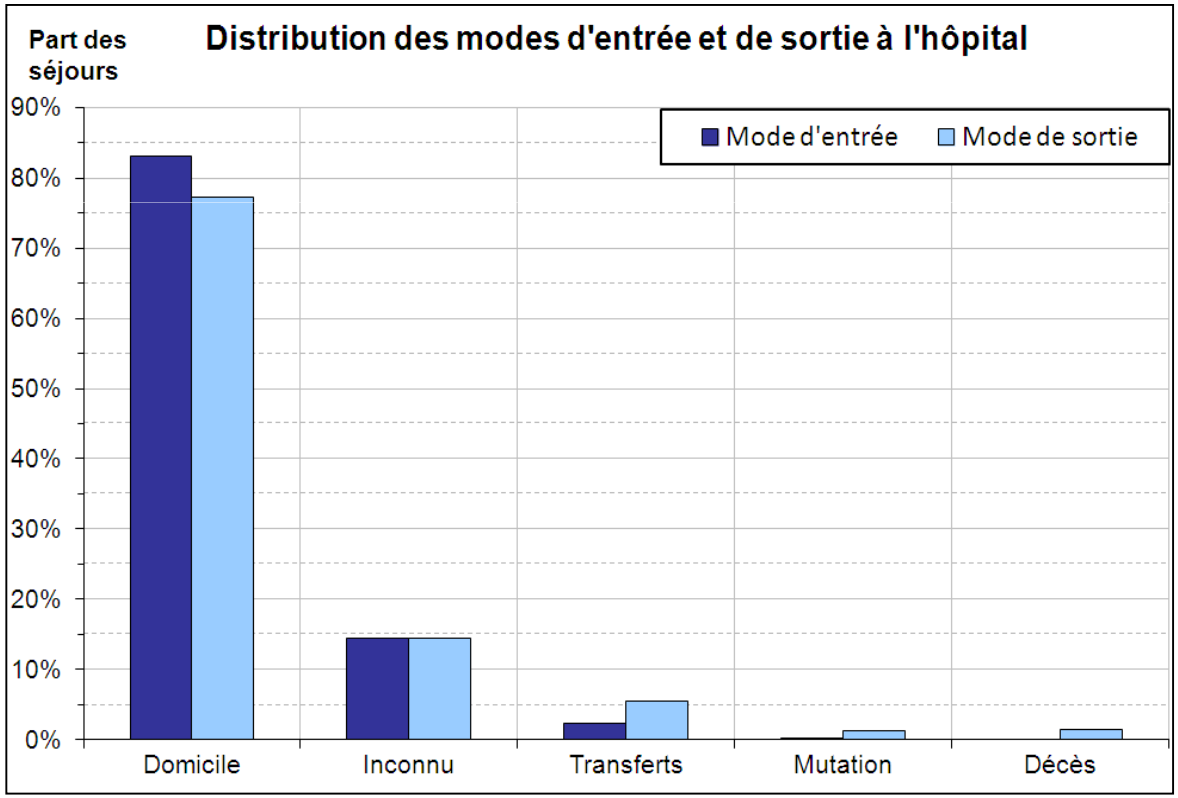
- Selon la durée du séjour :



Insee
Mesurer pour comprendre

1. Le PMSI

- Selon le mode d'entrée et de sortie :



Insee
Mesurer pour comprendre

Plan d'intervention

1. Le patient ausculté (le PMSI)

2. Le protocole opératoire
   - 2.1 Risque de ré-identification
   - 2.2 Objectifs de réduction du risque
   - 2.3 Méthodes de réduction du risque

3. Les actes pratiqués et le diagnostic final

Insee
Mesurer pour comprendre

2.1 Risque de ré-identification

Identifiant direct	Quasi-identifiants			Variable sensible non identifiante
Nom complet	**Âge**	**Sexe**	**Code postal**	**Maladie**
Ève Estampes	46 ans	Femme	42300	Cirrhose
Caroline Gérard	46 ans	Femme	73270	Bronchite
Ghislaine Bernard	68 ans	Femme	73270	Cancer du sein
Célimène Hervé	111 ans	Femme	73270	Hépatite C
Tom Chevalier	17 ans	Homme	75014	Insuffisance cardiaque
Marc Champion	31 ans	Homme	75014	Bronchite
Éric Carpentier	42 ans	Homme	93120	Grippe

Insee
Mesurer pour comprendre

2.2 Objectifs de réduction du risque

- Clé d'identification c = combinaison des modalités des variables quasi-identifiantes
- Soit f_c la fréquence d'apparition de la clé c dans la population U : $f_c,\ c \in [\![1, C]\!]$
- Un fichier est k-anonymisé si et seulement si : $f_c \geq k \ \forall c \in [\![1, C]\!]$
- Un fichier est l-diversifié si et seulement si pour chaque clé d'identification c, il y au moins l modalités « bien représentées » pour les variables sensibles

Insee
Mesurer pour comprendre

2.2 Objectifs de réduction du risque

Dans ce test, l'objectif est de construire un fichier :

- 10-anonymisé
- 3-diversifié

Âge	Sexe	Région	Maladie
+ 45 ans	Femme	Rhône-Alpes	Cirrhose
+ 45 ans	Femme	Rhône-Alpes	Bronchite
+ 45 ans	Femme	Rhône-Alpes	Cancer du sein
+ 45 ans	Femme	Rhône-Alpes	Hépatite C
- 45 ans	Homme	Île-de-France	Insuffisance cardiaque
- 45 ans	Homme	Île-de-France	Bronchite
- 45 ans	Homme	Île-de-France	Grippe

Insee
Mesurer pour comprendre

2.3 Méthodes de réduction du risque

- Les méthodes non perturbatrices ont été retenues pour ce test
- En 1ère approche volonté de conserver l'exhaustivité de la base → regroupement de modalités (agrégation)

Méthodes non perturbatrices (modifient la quantité et le détail de l'information)	**Méthodes perturbatrices (modifient la valeur des données initiales)**
Agrégation = regroupement de modalités (recodage global ou local)	Microagrégation
Suppressions locales	Bruitage
Échantillonnage	Permutations aléatoires (*swapping*)

Insee
Mesurer pour comprendre

Plan d'intervention

1. Le patient ausculté (le PMSI)

2. Le protocole opératoire

3. Les actes pratiqués et le diagnostic final
   - 3.1 Démarche
   - 3.2 Illustrations

Insee
Mesurer pour comprendre

3.1 Démarche

Deux approches complémentaires

- Logiciel µ-Argus
  - Agrégation de l'information pour les variables quasi-identifiantes
  - Possibilité de voir en temps réel :
    - le nombre de clés avec moins de 10 enregistrements
    - les modalités des variables indirectement identifiantes concernées
  - On procède par itérations jusqu'à obtenir un fichier 10-anonymisé
  - La 3-diversité est vérifiée *a posteriori*

Insee
Mesurer pour comprendre

3.1 Démarche

- Logiciel Arx
  - Définition *a priori* des différents niveaux de regroupement pour tous les quasi-identifiants
  - Le logiciel retourne l'ensemble des combinaisons respectant les critères de protection définis
  - Des approches avec différents niveaux de regroupement de l'information ont également été testées :
    - Détection des enregistrements « atypiques », dont la clé d'identification est possédée par moins de 10 séjours
    - Pour ces enregistrements, on réitère le processus en agrégeant davantage l'information

Insee
Mesurer pour comprendre

3.2 Illustrations

➢ Fichier en entrée de µ-Argus

Nom de la variable	Nature de la variable Sensible/Quasi-identifiante	Nombre de modalités
Sexe	Quasi-identifiante	2
Âge	Quasi-identifiante	19
Durée du séjour	Quasi-identifiante	12
Mode d'entrée	Quasi-identifiante	4
Mode de sortie	Quasi-identifiante	5
Lieu de résidence du patient	Quasi-identifiante	98
Nombre de clés d'identification		**893 760**
CMD (Catégorie Majeure de Diagnostic)	Sensible	26

Insee
Mesurer pour comprendre

3.2 Illustrations

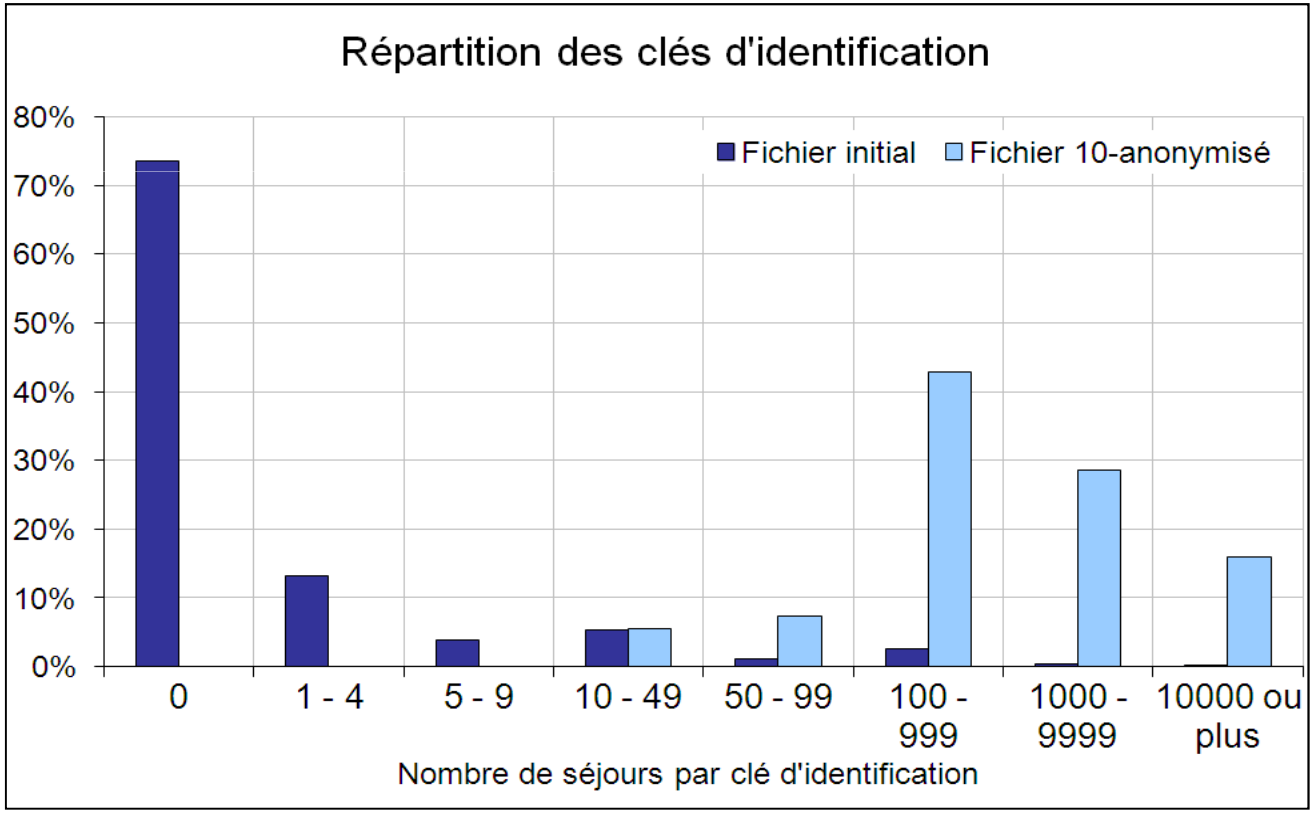
➢ Un exemple de fichier en sortie de µ-Argus

Nom de la variable	Nature de la variable Sensible/Quasi-identifiante	Nombre de modalités
Sexe	Quasi-identifiante	2
Âge	**Quasi-identifiante**	**6 : *moins de 1 an, 1-29 ans, 30-49 ans, 50-59 ans, 60-69 ans et 70 ans et +***
Durée du séjour	**Quasi-identifiante**	**2 : *+ ou - d'une semaine***
Mode d'entrée	**Quasi-identifiante**	**2 : *domicile ou autre***
Mode de sortie	**Quasi-identifiante**	**2 : *domicile ou autre***
Lieu de résidence du patient	**Quasi-identifiante**	**22 (*Régions + DOM sauf : regroupement Corse et PACA)***
Nombre de clés d'identification		**2 112**
CMD (Catégorie Majeure de Diagnostic)	Sensible	26

Insee
Mesurer pour comprendre

3.2 Illustrations

- Le risque de ré-identification avant et après 10-anonymisation avec µ-Argus



Insee
Mesurer pour comprendre

3.2 Illustrations

➢ Fichier en entrée de Arx

Nom de la variable	Nature de la variable Sensible/Quasi-identifiante	Nombre de modalités
Sexe	Quasi-identifiante	2
Âge	Quasi-identifiante	19
Durée du séjour	Quasi-identifiante	12
Lieu de résidence du patient	Quasi-identifiante	98
Lieu d'hospitalisation	Quasi-identifiante	98
Nombre de clés d'identification		**4 379 424**
CMD (Catégorie Majeure de Diagnostic)	Sensible	26

Insee
Mesurer pour comprendre

3.2 Illustrations

➢ Un exemple de fichier en sortie de Arx

Nom de la variable	Nature de la variable Sensible/Quasi-identifiante	Nombre de modalités	
		Séjours atypiques (3,8 %)	Autres séjours
Sexe	Quasi-identifiante	2	
Âge	Quasi-identifiante	19 : moins de 1 an, tranches quinquennales jusqu'à 84 ans, 85 ans et +	
Lieu de résidence du patient	Quasi-identifiante	98	
Durée du séjour	**Quasi-identifiante**	**non renseignée**	**12**
Lieu d'hospitalisation	**Quasi-identifiante**	**non renseigné**	**23 (Régions + DOM)**
Nombre de clés d'identification		**3 724**	**1 027 824**
CMD (Catégorie Majeure de Diagnostic)	Sensible	26	

Insee
Mesurer pour comprendre

Bilan

- Une démarche fondée sur l'agrégation de l'information contenue dans les variables quasi-identifiantes (recodage global et local)

- 2 logiciels d'anonymisation testés

- Conciliation difficile entre :
  - Détail de l'information
  - Réduction du risque de ré-identification
  - Maniabilité du fichier

Insee
Mesurer pour comprendre

Merci pour votre attention !

Éléments bibliographiques :

- *A. Hundepool et al. Statistical Disclosure Control, Wiley Series in Survey Methodology, 2012.*

- *Dossier Solidarité Santé « Risque de ré-identification dans les bases de données médico-administratives », dossier n°4, Drees*

Insee
Mesurer pour comprendre