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

# Création de fichiers anonymisés à partir d'une base médico-administrative (le PMSI) : un exemple pratique de mise en œuvre des méthodes de protection des fichiers de données individuelles

**Noémie JESS – Drees, Sous direction Observation de la santé et de l'Assurance maladie**

**Maxime BERGEAT – Insee**

**Françoise DUPONT – Insee**

**XII^ième Journées de Méthodologie Statistique**

Insee
Mesurer pour comprendre

# Plan d'intervention

1. Le patient ausculté et ses antécédents
2. Le protocole opératoire
3. Les actes pratiqués
4. Bilan post-op'

Insee
Mesurer pour comprendre

# Plan d'intervention

1. Le patient ausculté et ses antécédents
   - 1.1 Ouverture des données de santé
   - 1.2 Les données du test d'anonymisation
2. Le protocole opératoire
3. Les actes pratiqués
4. Bilan post-op'

Insee
Mesurer pour comprendre

# 1.1 Ouverture des données de santé

- Réflexions sur la réforme de l'accès aux données de santé, Commission « *open data* en santé » lancée à l'automne 2013

- En aval : groupe de travail sur les risques de ré-identification pour mener une expertise technique

- Lancement d'un test d'anonymisation sur données réelles

Insee
Mesurer pour comprendre

# 1.2 Les données du test d'anonymisation

- Programme de Médicalisation des Systèmes d'Information : base médico-administrative annuelle contenant l'intégralité des séjours hospitaliers en France

- Champ :
  - Édition 2012 du PMSI-MCO , *i.e.* les courts séjours
  - Exclusion des séances

20,6 millions de séjours

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

Insee
Mesurer pour comprendre

# 1.2 Les données du test d'anonymisation

- Accès à la base PMSI-MCO sur le serveur du CASD (Centre d'accès sécurisé distant aux données) après accord Cnil

- Traitements avec le logiciel µ-argus :
  - Gratuit, développé initialement par l'INS des Pays-Bas
  - Import de fichiers plats et définition des métadonnées
  - Application des méthodes d'anonymisation
  - Export des jeux de données créés

Insee
Mesurer pour comprendre

# Plan d'intervention

1. Le patient ausculté et ses antécédents

2. Le protocole opératoire
   - 2.1 Nature des variables et clés d'identification
   - 2.2 Risque de ré-identification
   - 2.3 Choix de la méthode d'anonymisation
   - 2.4 Critères de réduction du risque (*k*-anonymat, *l*-diversité)

3. Les actes pratiqués

4. Bilan post-op'

Insee
Mesurer pour comprendre

# 2.1 Nature des variables et clés d'identification

- Point d'entrée de la démarche = distinction entre :
  - Variable(s) sensible(s) à protéger
  - Variables quasi-identifiantes dont la combinaison peut permettre la ré-identification

- Clé d'identification $c_i$ = une combinaison de modalités des quasi-identifiants

Clé $c_1$ : {Sexe=Homme ; Âge=50 ;
Lieu de résidence= Bourg-en-Bresse ;
n° Finess établissement= Centre hospitalier de Bourg-en-Bresse;
Mode d'entrée=Transfert ; Mode de sortie=Domicile ;
Durée=5 nuits}

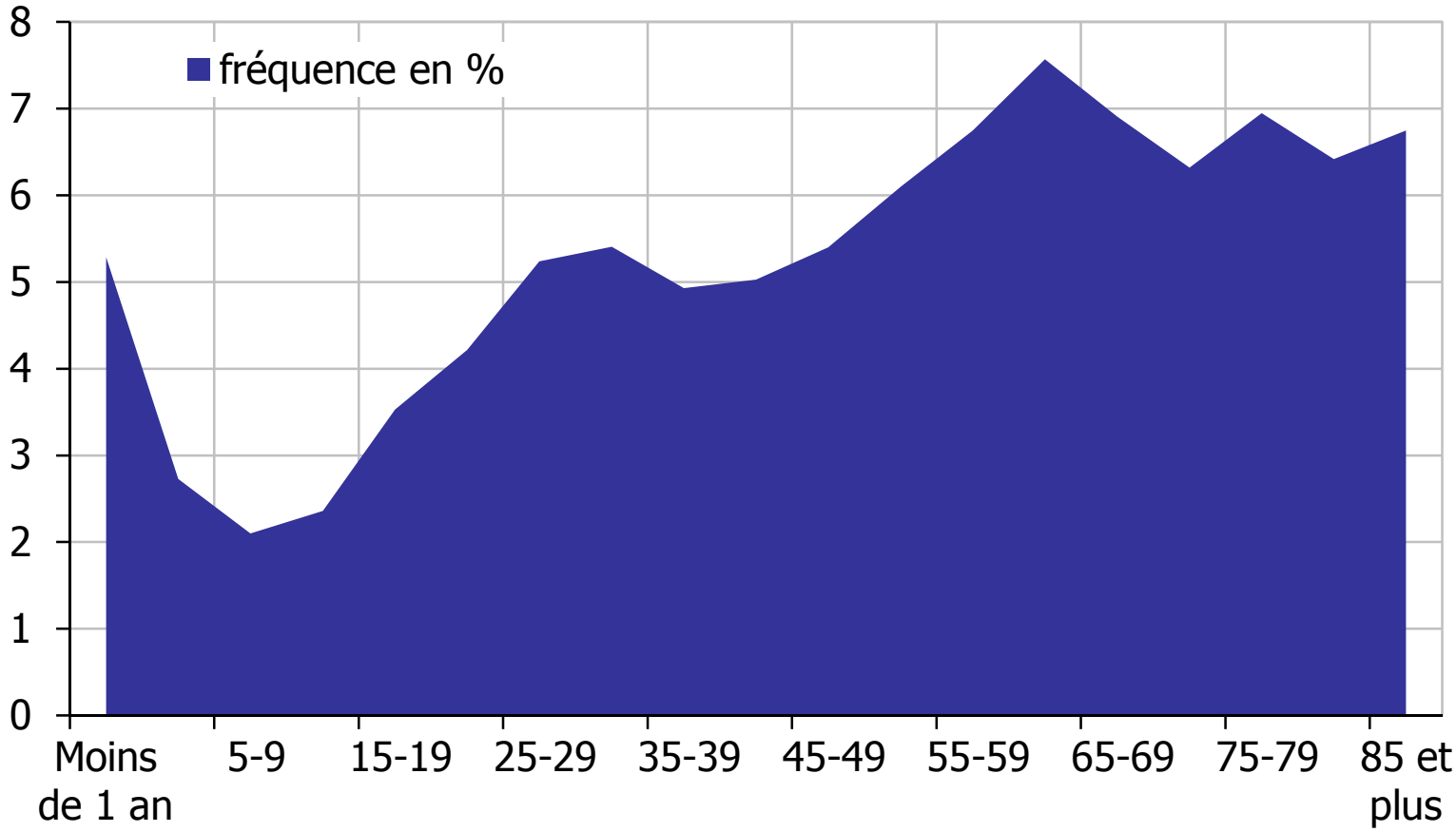
Insee
Mesurer pour comprendre

# 2.2 Risque de ré-identification

- Plusieurs risques dans la littérature:
  - révélation d'identité
  - révélation d'attribut (sensible)
  - révélation inférentielle
- PMSI-MCO chaîné : 89% de patients uniques si on combine leurs caractéristiques socio-démographiques et celles de l'hospitalisation (D. Blum, 2008)
- Approche descriptive pour détecter les modalités rares

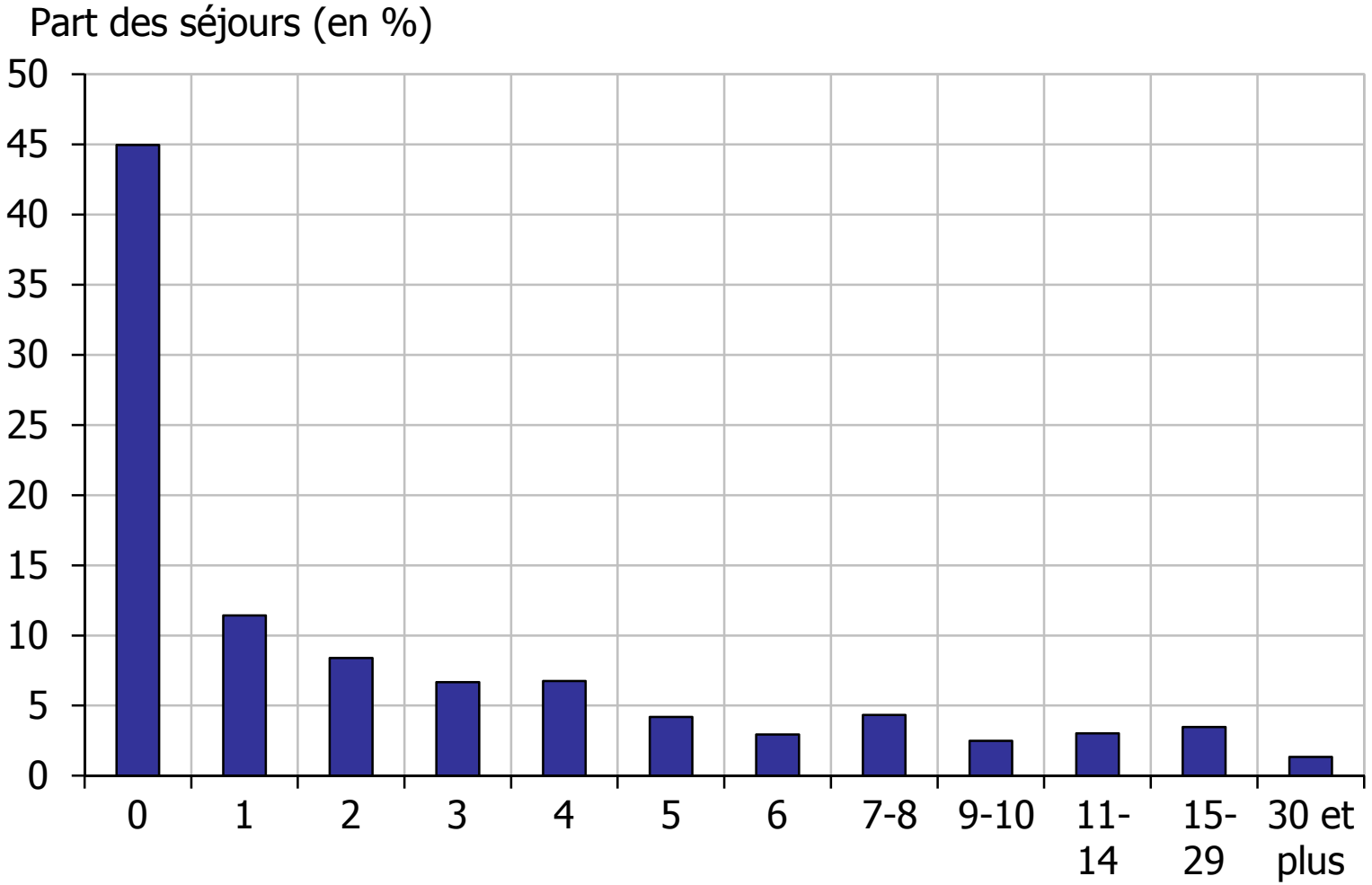
Insee
Mesurer pour comprendre

# Répartition des séjours par tranche d'âge



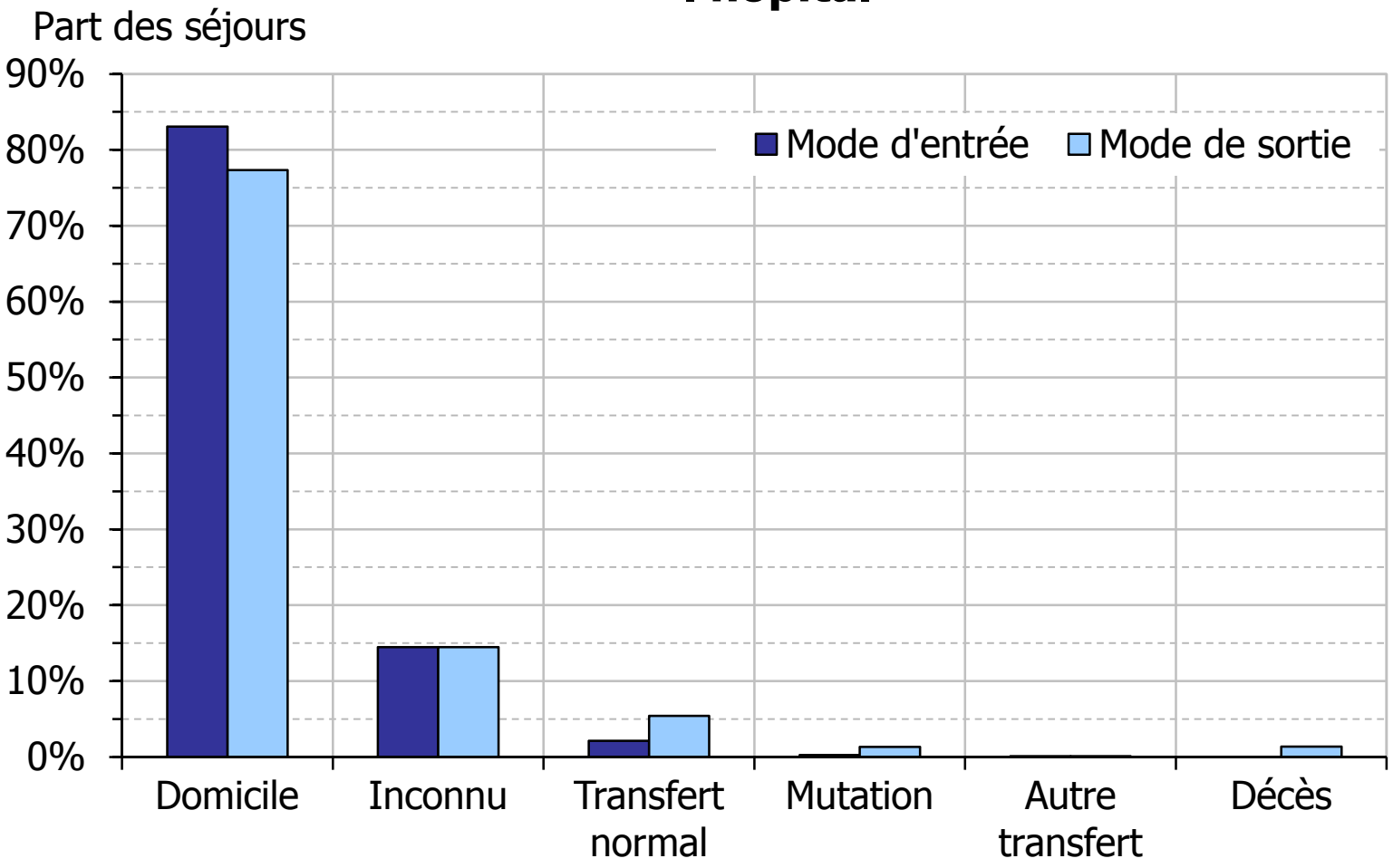

Insee
Mesurer pour comprendre

# Répartition des séjours selon le nombre de nuits




Insee
Mesurer pour comprendre

# Distribution des modes d'entrée et de sortie à l'hôpital




Insee
Mesurer pour comprendre

# 2.3 Choix de la méthode d'anonymisation

- 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

# 2.4 Critères de réduction du risque

- Le *k*-anonymat pour réduire le risque de révélation d'identité :
  - Un fichier est *k*-anonyme si pour toute clé d'identification $c_i$ , il existe au moins *k* individus possédant la clé $c_i$ :

$$n_i \geq k \; \forall \; i \; \in \{1 \ldots . J\}$$
$$où \; J = nombre \; total \; de \; clés \; d'identification$$
$$et \; n_i = effectif \; (nombre \; d'individus) \; possédant \; la \; clé \; c_i$$

- La *l*-diversité pour réduire le risque de révélation d'attribut :
  - Un fichier est *l*-divers si pour chaque clé d'identification $c_i$ , il y a au moins *l* modalités bien représentée pour chaque variable sensible

Insee
Mesurer pour comprendre

# Plan d'intervention

1. Le patient ausculté et ses antécédents
2. Le protocole opératoire
3. Les actes pratiqués
4. Bilan post-op'

Insee
Mesurer pour comprendre

# 3. Les actes pratiqués

| Nom de la variable | Nature de la variable | Nombre de modalités dans le fichier initial | Nombre de modalités dans le fichier 10-anonymisé |
|---|---|---|---|
| Sexe | Quasi-identifiante | 2 | 2 |
| Âge | Quasi-identifiante | 18 | 6 : Moins de 1 an, 1-29 ans, 30-49 ans, 50-59ans, 60-69 ans et 70 ans ou plus |
| Durée du séjour | Quasi-identifiante | 12 | 2 : + ou - d'une semaine |
| Mode d'entrée | Quasi-identifiante | 2 | 2 |
| Mode de sortie | Quasi-identifiante | 2 | 2 |
| Lieu de résidence | Quasi-identifiante | 23 | 22 : Regroupement Corse et PACA |
| **Nombre de clés d'identification** | | 39 744 | 2 112 |
| CMD (Catégorie majeur de diagnostic) | Sensible | 26 | 26 |

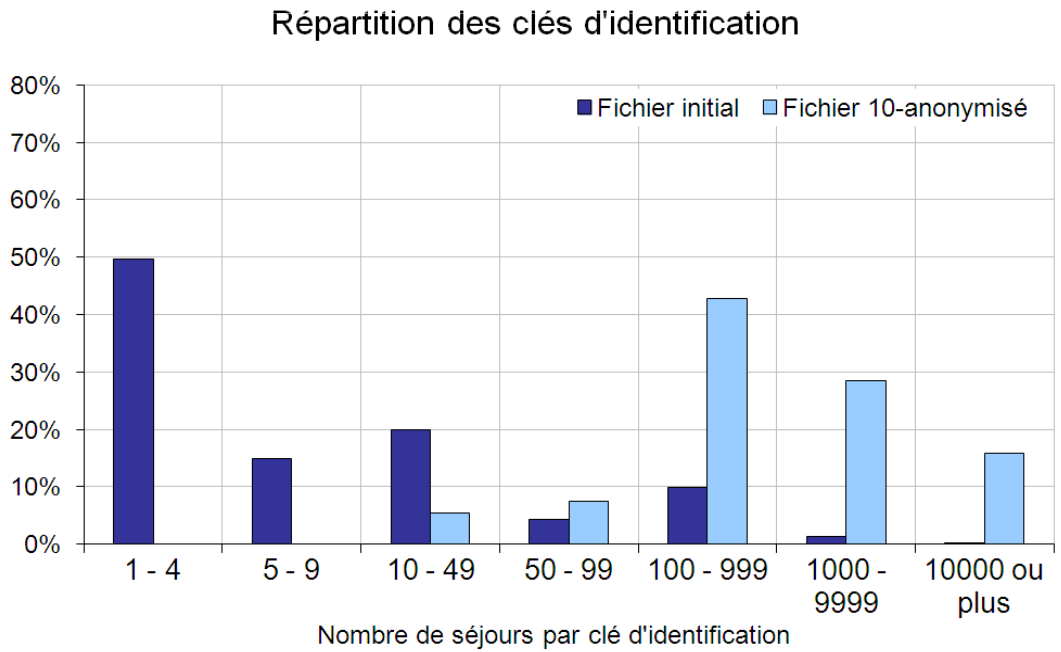
# Plan d'intervention

1. Le patient ausculté et ses antécédents
2. Le protocole opératoire
3. Les actes pratiqués
4. Bilan post-op'

Insee
Mesurer pour comprendre

# 4. Bilan post-op'

- Difficile optimisation des regroupements :
  - nomenclatures existantes, conseils des experts du domaine
  - distribution non-uniforme des séjours → risque portant sur peu de séjours mais niveau élevé d'agrégation pour tous les séjours




Insee
Mesurer pour comprendre

# 4. Bilan post-op'

- Arbitrage nécessaire entre :
  - Détail de l'information
  - Réduction du risque de ré-identification
  - Maniabilité du fichier
- D'autres pistes peuvent être exploitées : utilisation de méthodes perturbatrices, génération de données synthétiques
- Test technique : absence de définition d'un niveau de risque acceptable

Insee
Mesurer pour comprendre

# Merci de votre attention !