Données synthétiques SNDS & développement d'outils et méthodes :

Cas d’usage en pharmaco-épidémiologie

Emmanuel OGER EA 7449 REPERES
André HAPPE EA 7449 REPERES
Erwan DREZEN CUBR

EA 7449 REPERES

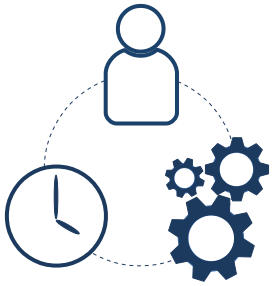
- **REcherche en Pharmaco-Epidémiologie et REcours aux Soins**
  - Université de Rennes 1 & EHESP
- **Domaines**
  - Pharmaco-épidémiologie
  - Organisation des soins

→ Recherche d'association entre exposition médicamenteuse et évènement(s) indésirable(s)

- **Spécificités**
  - Multi-disciplinarité
  - SNDS
- **Issue du consortium PEPS** (financé par ANSM)
  - Composante R&D (méthodes & outils)
  - Organisation de 2 colloques autour du SNDS (2017 & 2019)

CUBR

- **Startup qui développe des algorithmes**
  - Traitement ultra rapide de grands volumes de données
  - Focus sur l'aspect temporel des données
  - Vélocité permettant de faire sauter des verrous en terme de use cases



- **Cas d'usage naturel : l'épidémiologie**
  - Appariement de bases de données
  - Outils de visualisation pour l'exploration interactive
- **Un domaine d'application naturel : le SNDS**
  - Grande volumétrie
  - Aspect temporel fondamental en épidémiologie

Partie I

Cas d’usage de données synthétiques SNDS

Use case 1 : REPERES & la formation

Former

- **Historique : formation continue de l'EHESP**
  - « Extraire et Manipuler le SNDS pour l'épidémiologie et la santé publique »
- **Mise en place d'un TP**
  - Possibilité pour les étudiants de mener une étude épidémiologique à partir des données brutes
  - Nécessité d'un jeu de données de synthèse SNDS
- **Caractéristiques de ce jeu de synthèse**
  - Conforme au format fourni par la CNAM / DEMEX
    - Archive contenant des fichiers CSV (un fichier par table)
  - Couverture du SNDS
    - DCIR (partiel)
    - PSMI / MCO (partiel)

Use case 1 : REPERES & la formation

Former

Données synthétiques

SANDBOX.7z
(tables CSV)

A

décompression
vérification

```
Name
-----------------------
SANDBOX
SANDBOX/er_bio_f.csv
SANDBOX/er_cam_f.csv
SANDBOX/er_pha_f.csv
SANDBOX/er_prs_f.csv
SANDBOX/t_mco13a.csv
SANDBOX/t_mco13b.csv
...
```

B

chargement
vérification

SANDBOX
Base primaire
MySQL

C

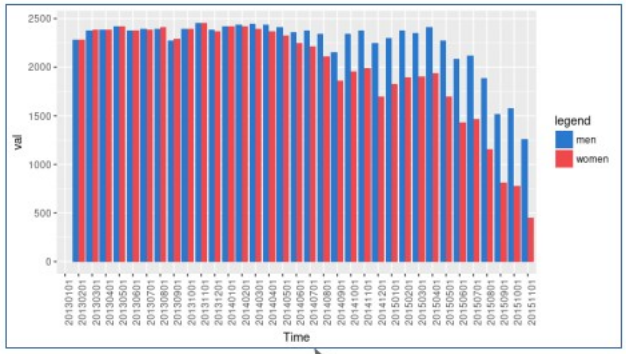
création variables
statistiques

SANDBOX
Base statistique
MySQL

DMB

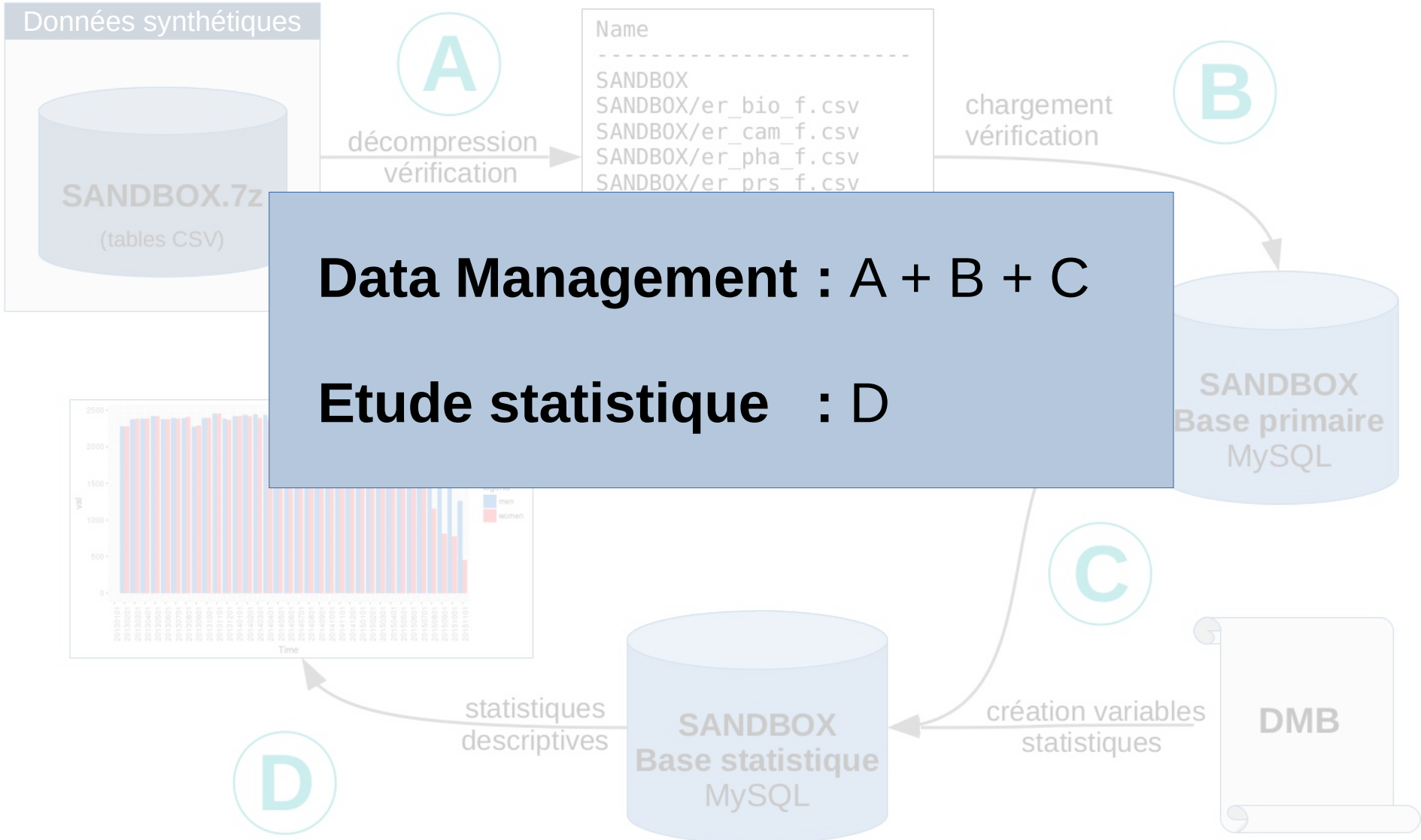
D

statistiques
descriptives



Use case 1 : REPERES & la formation

Former



Use case 2 : REPERES & HDH

Former

- **Convention REPERES / HDH**
  - Mise à disposition par REPERES à la communauté d'un jeu de données synthétiques SNDS
  - Issu de l'expérience de REPERES sur le SNDS
  - Centralisation des demandes d'utilisation par le HDH
- **Caractéristiques du jeu de synthèse**
  - Couverture similaire à celle de la formation EHESP
  - 20 000 patients, entre 2015 et 2018
- **Besoins des utilisateurs**
  - Familiarisation avec le SNDS (avant accès aux vraies données)
  - Compréhension des données du SNDS

https://documentation-snds.health-data-hub.fr/formation_snds/donnees_synthetiques.html#donnees-de-synthese-du-lab-sante-de-la-drees

Use case 3 : CUBR & outils

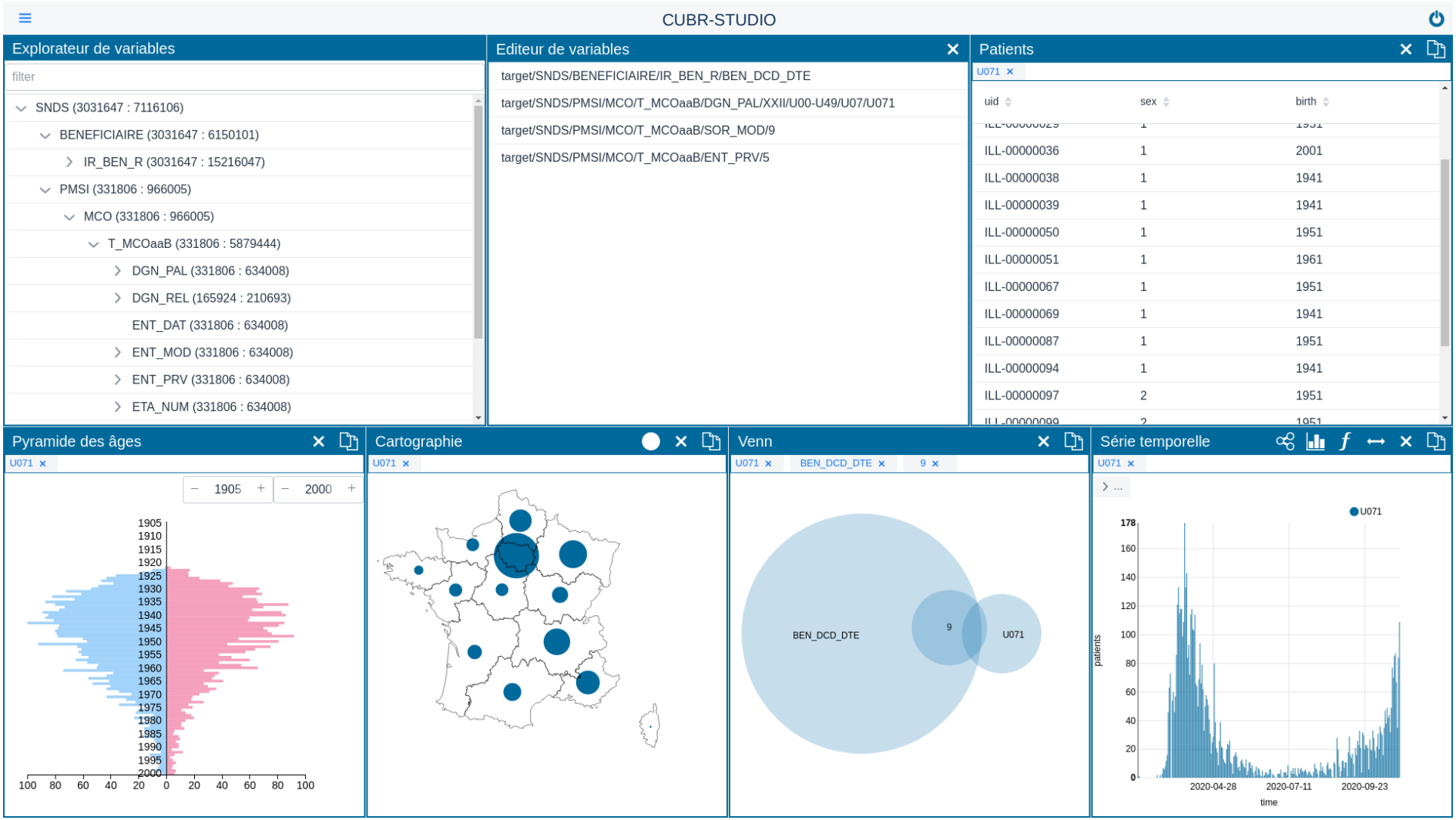
Tester

- **R&D chez CUBR**
  - Développement d'outils de visualisation
  - Mettant en valeur la dimension temporelle
- **Objectifs**
  - Supporter de très larges populations
  - Donner des résultats de requêtes quasi instantannément
- **Un bon candidat pour utilisation : le SNDS**
  - CUBR traite nativement le schéma du SNDS
  - **Mais CUBR n'a pas accès à des données réelles !**
- **Solution : des données synthétiques SNDS**
  - Développement en interne par CUBR d'un tel générateur

Use case 3 : CUBR & outils

Tester

Exploration de cohorte sur données synthétiques



Use case 4 : CUBR & appariement

Valider

- **Appariement de bases de données**
  - Enrichir une base de données par une autre
  - 1 +1 > 2
  - Besoin important en épidémiologie
- **Apparier le SNDS avec d'autres sources**
  - SNDS : base de remboursements uniquement !
  - Grande utilité pour enrichir un registre, une cohorte, ...
- **R&D chez CUBR**
  - Développement d'un algorithme d'appariement de bases de données dit « combinatoire »
  - Un cas d'utilisation de choix : apparier le SNDS avec une autre base de données

Use case 4 : CUBR & appariement

Valider

- **Comment valider un tel algorithme ?**
  - Impossible sur des bases réelles (résultat inconnu à l'avance)
  - Important d'avoir un « gold standard »
- **Solution : utiliser des donnés synthétiques**
  - Usage du générateur SNDS développé en interne par CUBR
  - Permet de savoir *a priori* que le patient « A3197 » de la base A correspond au patient « B123 » de la base B
- **Résultats grâce aux données synthétiques**
  - Validation de l'algorithme et caractérisation de ses performances (bonnes spécificité / sensibilité)
  - Utilisation par REPERES de l'algorithme de CUBR pour apparier le registre des AVC de Brest avec le SNDS
  - Soumission d'une publication commune REPERES et CUBR

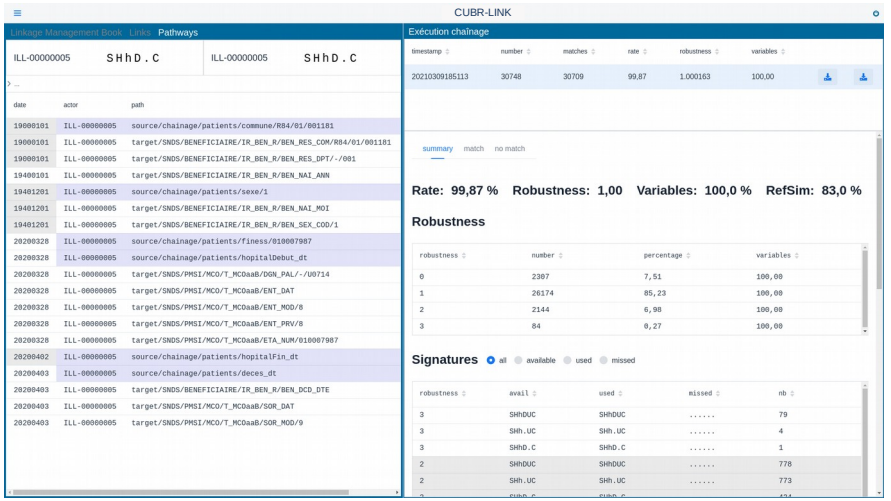
Use case 5 : Besoin de communiquer

Communiquer

- **Communication sur un outil et/ou un algo**
  - Démontrer son intérêt spécifiquement pour le SNDS
  - Pour une publication, un site web, etc...
- **Moyens classiques**
  - Captures d'écran, vidéo
  - **Mais impossibilité d'utiliser de vraies données !**
- **Recours à des données de synthèse SNDS**

Exemple d'un snapshot utilisé par CUBR pour sa communication

Données synthétiques obligatoires, d'autant plus que des données individuelles sont ici affichées !



Partie II

Retours sur la création d'un générateur de données SNDS

Principes de génération de données SNDS

Générer

- **Besoins de REPERES**
  - Utilisation de statistiques descriptives pour la génération
  - Cohérence structurelle entre les tables
  - Cohérence du parcours de soins (socio-démographique)
  - Pas de cohérence multivariables
  - Information « médicale » uniquement (pas de médico-éco)
  - Couverture du schéma SNDS partielle
- **Besoins de CUBR**
  - Approche globalement similaire à celle de REPERES
  - Couverture quasi complète du schéma SNDS
  - Possibilité de définir des parcours type (ex : covid19)

Conclusion

Retour d’expérience(s)

- **Besoins variés**
  - Former
  - Tester
  - Valider
  - Communiquer
- **Dépendant des acteurs**
  - Equipe de recherche (ayant accès aux données du SNDS)
  - Startup (sans accès aux données du SNDS)
- **Pas de jeu de données « universel »**
  - Une problématique -> un jeu de données « sur mesure »

Merci de votre attention.