Epidémiologie et Biostatistique sur un Entrepôt de Données hospitalier

03 juin 2025

Introduction à l'appariement des bases

Luca THIEBAUD & Adam REMAKI
adam.remaki@centralesupelec.fr
luca.thiebaud@centralesupelec.fr

L'appariement (*record linkage*):

Déf: tâche consistant à trouver des données qui se réfèrent à la même entité à travers différentes sources de données.

Objectif:

- Lier deux bases de données de **source différente** en comparant des paires de dossiers patients
- Dédupliquer un identifiant censé être unique d'une base de données

L'appariement déterministe:

- Utilisation d'**identifiers**: variables communes aux deux bases de données
- **Match** si accord sur les identifiers (cas trivial: jointure sur une clé)
- Processus possiblement itératif

  →match sur IPP

  → **si absent** match sur NIR

  → **si absent** match exact sur (prénom, nom, date de naissance)

Appariement déterministe
⇒ différents niveaux de *confiance*

L'appariement probabiliste:

- Introduction d'une **probabilité de match** et d'une **frontière de décision**
- Article princeps de Fellegi et al. [1]

Direction des Systèmes d'Informations

[1] Fellegi I, Sunter A. A theory for record linkage. J Am Stat Assoc. 1969;64.

ASSISTANCE PUBLIQUE HÔPITAUX DE PARIS

International Journal of Population Data Science

Journal Website: www.ijpds.org

IJPDS International Journal of Population Data Science

Swansea University Prifysgol Abertawe

Demystifying probabilistic linkage: Common myths and misconceptions

Doidge, JC[1,2*] and Harron, K[3]

- **Méthodes probabiliste et déterministes sont différentes ?**
  - Pas forcément, on a équivalence entre trouver un ensemble de règles de décisions déterministes et trouver des ensembles (poids / threshold) dans un contexte probabiliste

- **Méthodes probabilistes basées sur la "probabilité qu'une paire soit un match" ?**
  - En réalité, les poids sont des scores censés être en corrélation avec cette probabilité qu'une paire soit un match, mais la probabilité exacte est souvent inconnue.

- **Les méthodes probabilistes induisent plus d'erreurs ?**
  - Réduisent en général la précision, mais augmentent ainsi le rappel

Précison = nombre de bons match prédits / nombre total de match prédits (bons ou mauvais)
Rappel = nombre de bons match prédits / nombre total de bons match
***La précision** peut être comprise comme une mesure de l'exactitude ou de la qualité.*
***Le rappel** est une mesure de l'exhaustivité ou de la quantité.*

2 ensembles A et B

On travaille sur A x B, qui est l'union de 2 sous-ensembles disjoints

- M : sous ensemble des matchs de A x B
- U : sous ensemble des non-matchs de A x B

Pour une paire issue de A x B:

- construction d'un **vecteur de comparaison** (one-hot encoding) en fonction de la concordance ou non sur **K** variables communes
- Le modèle estime ensuite le **poids** de chaque variable, en supposant leur indépendance (Naive Bayes)

Michel	SCOTT	13-10-1950	PARIS
Jean-Michel	SCOTT	13-10-1950	PARIS
0	1	1	1

$$\gamma = [0, 1, 1, 1]$$

But: estimer $\Pr(\text{record match} \mid \gamma)$

m-probabilities et **u-probabilities**

Intuition

- Si une variable match, soit la paire match, soit on a une **collision**

- Si une variable ne match pas :
  - soit la paire match (typo, différence de format, …)
  - soit la paire ne match pas non plus

Formalisation

m-probabilities et **u-probabilities**

	Var. match	Var. mismatch
Rec. match	$m_{i,1} = \Pr(\text{variable } i \text{ matches} \mid \text{records match})$	$m_{i,0} = \Pr(\text{variable } i \text{ mismatches} \mid \text{records match})$
Rec. mismatch	$u_{i,1} = \Pr(\text{variable } i \text{ matches} \mid \text{records do not match})$	$u_{i,0} = \Pr(\text{variable } i \text{ mismatches} \mid \text{records do not match})$

Intuition

- Si une variable match, soit la paire match, soit on a une **collision**
  → On compare $m_{i,1}$ vs $u_{i,1}$
- Si une variable ne match pas :
  - soit la paire match (typo, différence de format, …)
  - soit la paire ne match pas non plus

  → On compare $m_{i,0}$ vs $u_{i,0}$

$$b_{i,1} = \frac{m_{i,1}}{u_{i,1}}$$

$$b_{i,0} = \frac{m_{i,0}}{u_{i,0}}$$

Formalisation → Introduction du **Bayes Factor**

Michel	SCOTT	13-10-1950	PARIS
Jean-Michel	SCOTT	13-10-1950	PARIS
0	1	1	1

$$\gamma = [0, 1, 1, 1]$$

- **Pour une variable i**, calcul de la probabilité à posteriori

- **Pour le vecteur de comparaison γ**

Michel	SCOTT	13-10-1950	PARIS
Jean-Michel	SCOTT	13-10-1950	PARIS
0	1	1	1

$$\gamma = [0, 1, 1, 1]$$

- **Pour une variable i**, calcul de la probabilité à posteriori

$$\Pr(\text{records match} \mid \gamma_i) = \frac{\Pr(\gamma_i \mid \text{records match}) \Pr(\text{records match})}{\Pr(\gamma_i)}$$

$$= \frac{m_{i,l}\lambda}{m_{i,l}\lambda + u_{i,l}(1-\lambda)}$$

$$\text{with} \begin{cases} \lambda = \Pr(\text{records match}) \\ l = 0 \text{ if } \gamma_i = 0 \\ l = 1 \text{ if } \gamma_i = 1 \end{cases}$$

- **Pour le vecteur de comparaison γ**

Michel	SCOTT	13-10-1950	PARIS
Jean-Michel	SCOTT	13-10-1950	PARIS
0	1	1	1

$$\gamma = [0, 1, 1, 1]$$

- **Pour une variable i**, calcul de la probabilité à posteriori

$$\begin{aligned}\Pr(\text{records match} \,|\, \gamma_i) &= \frac{\Pr(\gamma_i \,|\, \text{records match}) \Pr(\text{records match})}{\Pr(\gamma_i)} \\ &= \frac{m_{i,l}\lambda}{m_{i,l}\lambda + u_{i,l}(1-\lambda)}\end{aligned}$$

$$\text{with} \begin{cases} \lambda = \Pr(\text{records match}) \\ l = 0 \text{ if } \gamma_i = 0 \\ l = 1 \text{ if } \gamma_i = 1 \end{cases}$$

- **Pour le vecteur de comparaison γ**

$$\begin{aligned}\Pr(\text{records match} \,|\, \gamma) &= \Pr(\text{records match} \,|\, \gamma_1, \ldots, \gamma_K) \\ &= \frac{\prod_{i=1}^{K} \Pr(\gamma_i \,|\, \text{records match}) \cdot \Pr(\text{records match})}{\Pr(\gamma_1, \ldots, \gamma_K)} \quad \text{par indépendance} \\ &= \frac{\lambda m_{1,l} \cdot m_{2,l} \ldots m_{K,l}}{\lambda m_{1,l} \cdot m_{2,l} \ldots m_{K,l} + (1-\lambda) u_{1,l} \cdot u_{2,l} \ldots u_{K,l}}\end{aligned}$$

En termes de cote…

En termes de cote…

$$\begin{aligned}\text{odds} &= \frac{p}{1-p} \\ &= \frac{\lambda}{1-\lambda} \cdot \frac{\prod m_{i,l}}{\prod u_{i,l}} \\ &= \frac{\lambda}{1-\lambda} \cdot b_{1,l} \dots b_{K,l}\end{aligned}$$

Bayes Factor

En termes de cote…

$$\begin{aligned}\text{odds} &= \frac{p}{1-p} \\ &= \frac{\lambda}{1-\lambda} \cdot \frac{\prod m_{i,l}}{\prod u_{i,l}} \\ &= \frac{\lambda}{1-\lambda} \cdot b_{1,l} \ldots b_{K,l}\end{aligned}$$

Bayes Factor

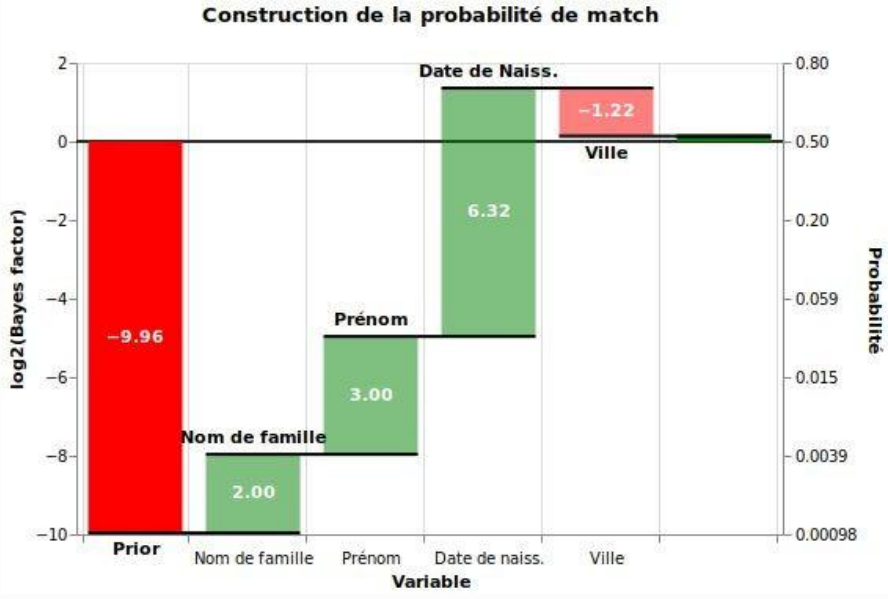
$$\log_2(\text{odds}) = \log_2\left(\frac{\lambda}{1-\lambda}\right) + \log_2(b_{1,l}) + \ldots + \log_2(b_{K,l})$$

En termes de cote…

$$\begin{aligned}\text{odds} &= \frac{p}{1-p} \\ &= \frac{\lambda}{1-\lambda} \cdot \frac{\prod m_{i,l}}{\prod u_{i,l}} \\ &= \frac{\lambda}{1-\lambda} \cdot b_{1,l} \ldots b_{K,l}\end{aligned}$$

Bayes Factor

Construction itérative de la probabilité de match



$$\log_2(\text{odds}) = \log_2\left(\frac{\lambda}{1-\lambda}\right) + \log_2(b_{1,l}) + \ldots + \log_2(b_{K,l})$$

- L'idée générale du modèle est donc d'utiliser une **estimation du poids relatif de chaque variable**

	m	u	b
Date de naissance	0.95	0.001	950
Nom	0.9	0.01	90
Sexe	0.95	0.5	1.9

- Limites:
  - Concordance binaire
  - Indépendance des variables
  - Comment estimer m et u ? ⟶ Empiriquement ou bien via un algorithme EM

A retenir:

- **Certaines variables ont un poids (bien) plus importants pour déterminer le statut d'une paire**
- **Variable clé:** Date de naissance
- Nécessité d'une étape de **blocking**: présélectionner des candidats pour limiter le nombre de comparaisons

- Cas d'usage: déduplication d'une base de N = 10M de patients

- Approche naïve: comparaison complète
  - N^2 paires de patient $\longrightarrow$ **100 billiards ici**
  - Il faut présélectionner des **candidats**

- *Standard Blocking*: utilisation d'une *clé*
  - Par exemple: initiale du nom, année de naissance
  - Avec b blocs, le nombre de comparaisons est alors en $\mathcal{O}\left(\frac{N^2}{b}\right)$
  - Attention à la taille des blocs
  - Attention au choix de la clé

- Le choix d'une technique de blocking adaptée est **essentiel**

- 2 métriques classiques pour mesurer la qualité du blocking:
  - **Pairs Completeness:** $$PC = \frac{Card(\text{matches détectés})}{Card(\text{matches existants})}$$
  - **Pairs Quality:** $$PQ = \frac{Card(\text{matches détectés})}{Card(bloc)}$$

- Tradeoff entre
  - Optimisation du temps de calcul
  - Optimisation de la procédure d'appariement

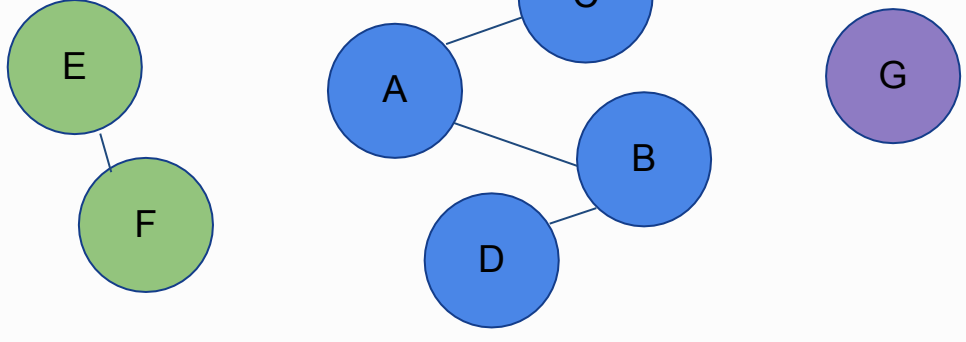
- En sortie d'algorithme, on a un ensemble de paires de matches

- De cet ensemble, il est possible de créer un graphe
  - Les noeuds sont les patients
  - Les arêtes déterminent si 2 patients sont appariés

- On en retire alors les **composantes connexes**

A	B
A	C
B	D
E	F

Paires de matchs



Composantes connexes

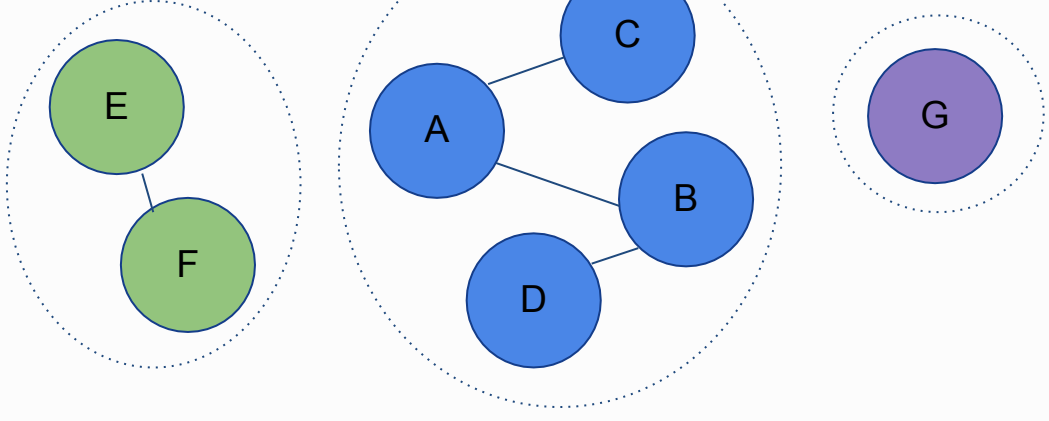
- En sortie d'algorithme, on a un ensemble de paires de matches

- De cet ensemble, il est possible de créer un graphe
  - Les noeuds sont les patients
  - Les arêtes déterminent si 2 patients sont appariés

- On en retire alors les **composantes connexes**

A	B
A	C
B	D
E	F

Paires de matchs



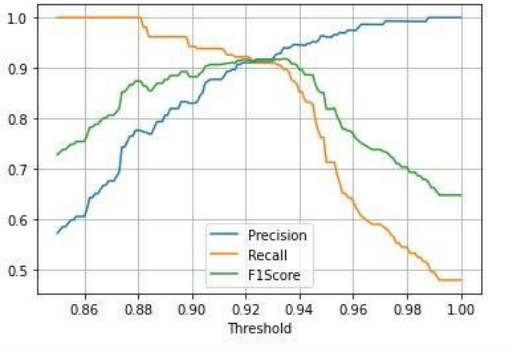
Composantes connexes

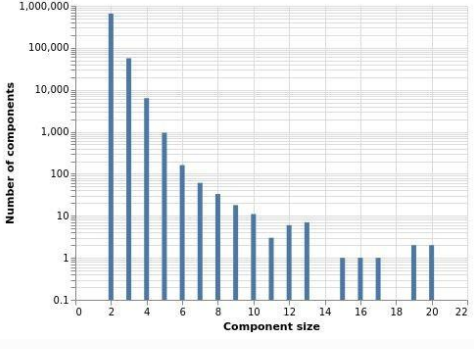
Un exemple d'algorithme: Déduplication des patients de l'EDS

- But de l'algorithme:
  - Proposer un 1ère version d'un algorithme de dédoublonnage moins conservateur que le l'algorithme existant (destiné au soin) à destination de la recherche
  - Doit tourner sur les ~13M de patients de l'EDS en un temps court / avec des ressources raisonnables

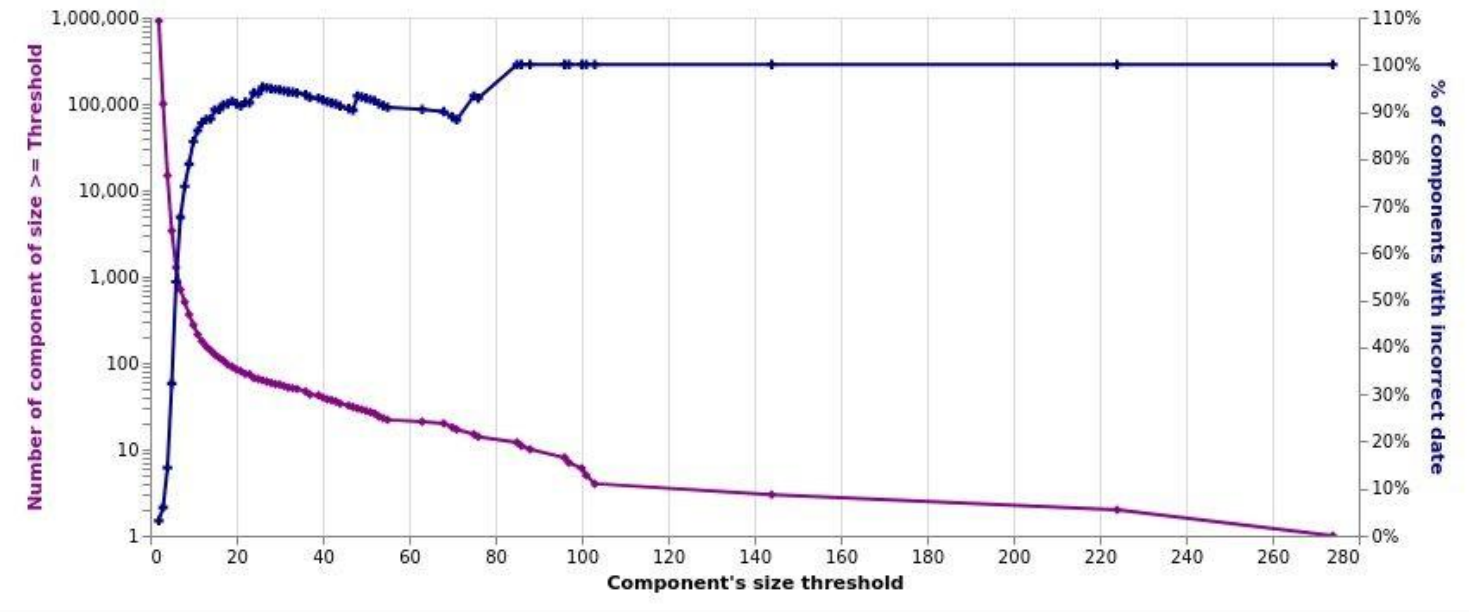
- Fonctionnement global
  - Utilisation des variables avec une complétude élevée
  - Blocking sur la date de naissance
    - On passe de 6.4×10^{13} à 2.2×10^{9} comparaisons
  - Fuzzy-matching sur les patronymes
    - Métrique de **Jaro-Winkler**
    - Détermination du threshold via une campagne d'annotation

	Colonne	Entrées non-nulles (%)
1	nom	99.999373
0	prenom	99.994017
5	dt_nais	91.531307
3	cd_genre	90.703285
2	nom_nais	65.168952
4	lieu_nais	20.267825





- Exemple de biais identifié sur les dates de naissance
  - Sur-représentation du 1er Janvier et 31 décembre
    - Origine identifiée
      - 1er Janvier: attribuée par défaut par l'état civil en l'absence de date précise
      - 31 décembre: attribuée par défaut à l'arrivée à l'AP-HP si absence de papiers
  - Sur-représentation des décennies
  - **Cause de FP importante, visible dans les "grands" composants**



- Gestion des valeurs manquantes
  - Approches classiques:
    - éliminer les patients avec au moins 1 valeur manquante du processus d'appariement
    - Considérer qu'une valeur manquante correspond à un match exact sur cette valeur
    - Considérer qu'une valeur manquante correspond à un non-match total sur cette valeur

And… what about ML and DL ?

- Peu de recherche à ce niveau !
- Pourtant, [1] montre qu'un simple MLP est meilleur que le Naive Bayes présenté précédemment.
- Dans le futur, usage d'embeddings et de réseaux siamois [2] ?

[1] Wilson et al, Beyond probabilistic record linkage: Using neural networks and complex features to improve genealogical record linkage
[2] Anna Jurek-Loughrey; (2020). Siamese Neural Network for Unstructured Data Linkage

Exercice 2 – Déduplication des identités

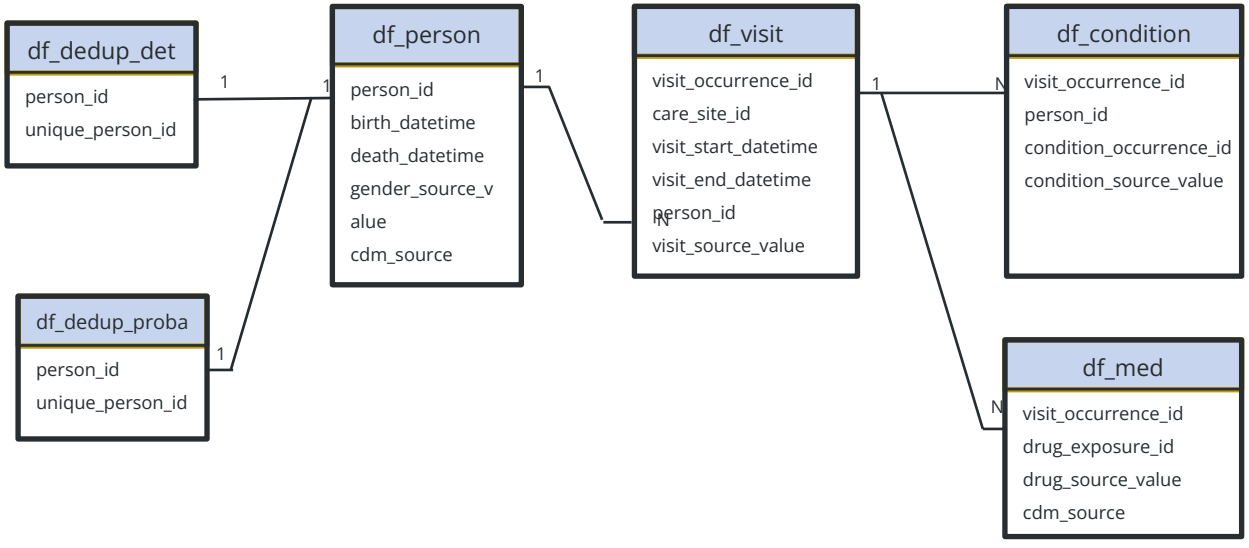
Direction des Systèmes

ASSISTANCE PUBLIQUE HÔPITAUX DE PARIS

Application – Déduplication

- **Pull le sujet de l'exercice 2 depuis le repo GitHub**
- **Ouvrir le notebook *exercises/exercise-2***
  - Contexte :
    - *Mise à disposition d'une table de déduplication déterministe pour les patients*
    - *Mise à disposition d'une table de déduplication probabiliste pour les patients*
  - Objectif : Comparer l'impact de la déduplication sur les résultats

Schéma de la base mise à disposition



Projet

Direction
des Systèmes

ASSISTANCE PUBLIQUE HÔPITAUX DE PARIS

Projet – Exploration des données

Question scientifique

- Identification des facteurs de risque associés au cancer du sein
- Analyse rétrospective à partir des données de l'EDS

Objectifs globaux

- Synthèse de la littérature sur le sujet
- Construction des objectifs techniques pour répondre à la question scientifique
- Nettoyage/Traitement de la BDD mise à disposition
- Analyses statistiques
- Restitution
  - *Court article scientifique (~1500 mots)*
  - *Notebook reproductible*

Objectifs première journée

- Lecture de la base
- Exploration & mapping des données
- Revue documentaire (introduction de l'article scientifique)
- Elaboration d'un patient set et des facteurs de risques à évaluer

Objectifs seconde journée

- Implémenter la déduplication du patient set
- Analyse comparative des premiers facteurs de risques

Bibliographie

- *Kohane, Isaac S, Bruce J Aronow, Paul Avillach, Brett K Beaulieu-Jones, Riccardo Bellazzi, Robert L Bradford, Gabriel A Brat, et al. « What Every Reader Should Know About Studies Using Electronic Health Record Data but May Be Afraid to Ask ». Journal of Medical Internet Research 23, n° 3 (2 mars 2021): e22219. https://doi.org/10.2196/22219.*
- *Agniel, Denis, Isaac S Kohane, et Griffin M Weber. « Biases in Electronic Health Record Data Due to Processes within the Healthcare System: Retrospective Observational Study ». BMJ, 30 avril 2018, k1479. https://doi.org/10.1136/bmj.k1479.*
- *Wilson, Greg, D. A. Aruliah, C. Titus Brown, Neil P. Chue Hong, Matt Davis, Richard T. Guy, Steven H. D. Haddock, et al. « Best Practices for Scientific Computing ». Édité par Jonathan A. Eisen. PLoS Biology 12, n° 1 (7 janvier 2014): e1001745. https://doi.org/10.1371/journal.pbio.1001745.*
- *Benchimol, Eric I., Liam Smeeth, Astrid Guttmann, Katie Harron, David Moher, Irene Petersen, Henrik T. Sørensen, Erik von Elm, Sinéad M. Langan, et RECORD Working Committee. « The REporting of Studies Conducted Using Observational Routinely-Collected Health Data (RECORD) Statement ». PLOS Medicine 12, n° 10 (6 octobre 2015): e1001885. https://doi.org/10.1371/journal.pmed.1001885.*