

# Appariement entre données hospitalières et certificats de décès en combinant moteur de recherche et apprentissage automatique

**Sébastien Cossin<sup>1,2</sup>, Sérigne Diouf<sup>1,2</sup>, Romain Griffier<sup>1,2</sup>, Philippine Le Barrois d'Orgeval<sup>2</sup>, Gayo Diallo<sup>1</sup>, Vianney Jouhet<sup>1,2</sup>**

<sup>1</sup> Univ. Bordeaux, Inserm, Bordeaux Population Health Research Center, Equipe ERIAS, UMR 1219.

<sup>2</sup> Unité IAM, Service d'information médicale. CHU de Bordeaux

**4ème Journée Dataquitaire - 25 février 2021**



# Auteurs

Sébastien Cossin<sup>1,2</sup>, Sérigne Diouf<sup>1,2</sup>, Romain Griffier<sup>1,2</sup>,  
Philippine Le Barrois d'Orgeval<sup>2</sup>, Gayo Diallo<sup>1</sup>, Vianney  
Jouhet<sup>1,2</sup>

- ① Erias: Equipe de recherche en informatique appliquée à la santé. Centre INSERM U1219



- ② IAM: unité hospitalière mettant en oeuvre l'entrepôt de données biomédicales du CHU de Bordeaux en étroite collaboration avec la DSI



L'entrepôt  
de données



- 1 Introduction
- 2 Etat de l'art
- 3 Méthodes
- 4 Résultats
- 5 Discussion

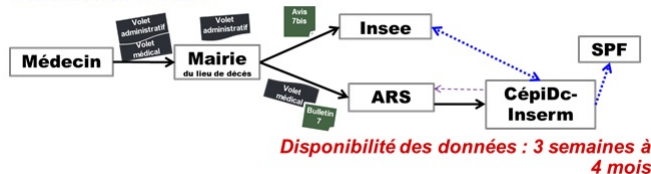
# *Importance de connaître le statut vital*

- Recherche clinique
  - Identification de patients éligibles
  - Etudes de cohorte
- Gestion des archives
  - Papier
  - Electronique

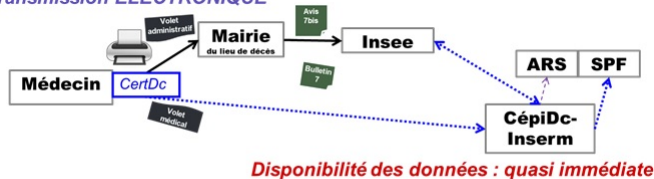


# Circuit des certificats de décès<sup>1</sup>

## Transmission PAPIER



## Transmission ELECTRONIQUE



→ Transmission papier    .....> Flux informatique    <---> Accès web

<sup>1</sup> <https://www.cephdc.inserm.fr/le-circuit-administratif-du-certificat-de-deces>

## *Statut vital au CHU de Bordeaux*

- 2,2 millions de patients venus au CHU de Bordeaux
- 58.000 décès

Les hôpitaux enregistrent les **décès intra-hospitaliers** mais ne reçoivent aucune information sur les décès extra-hospitaliers.

Aucune solution simple pour connaître les décès extra-hospitaliers jusqu'au 5 décembre 2019

## *Fichier des personnes décédées<sup>1</sup>*



etalab gouv.fr



25 millions certificats de décès (>01/01/1970):

- nom de famille
- prénoms
- sexe
- date de naissance
- code du lieu de naissance
- lieu de naissance
- pays de naissance en clair
- date du décès
- code du lieu de décès

---

<sup>1</sup><https://www.data.gouv.fr/en/datasets/fichier-des-personnes-decedees/>

# Objectif

Identifier les décès extra-hospitaliers en rapprochant les identités de la base de données patients du CHU de Bordeaux et la base de données open data de l'INSEE



# Record Linkage

*Processus visant à identifier si des enregistrements concernent la même entité (patient) en utilisant des identifiants communs entre les jeux de données.<sup>1</sup>*



Base CHU



- Nom: Chirac
- Prenom: Jacques
- DDN: 29/11/1932
- Sexe: M
- LastVisit: 20/01/2002
- ...



Base INSEE



- Nom: Chirac
- Prenom: Jacques René
- DDN: 29/11/1932
- Sexe: M
- Deces: 26/09/2019
- ...

<sup>1</sup>Padmanabhan S et al. Approach to record linkage of primary care data from Clinical Practice Research Datalink to other health-related patient data: overview and implications. Eur J Epidemiol. 2019;34(1):91-9.

# Définition

- Appariement (= alignement): le fait de relier une entité de la base A à une entité de la base B
  - Mauvais appariement (faux positif): quand l'entité de la base A n'est pas la même que celle de la base B
  - Bon appariement (vrai positif): quand l'entité de la base A est la même que celle de la base B
  - Appariement raté (faux négatif): quand aucun appariement n'a été réalisé alors que l'entité de la base A est la même que celle de la base B
- Identifiant: une caractéristique d'une entité (nom, prénom, sexe, lieu de naissance...)

# Méthodes

3 principales approches:

- Déterministe
- Probabiliste
- Machine Learning

# Déterministe

Approche consistant à fixer les règles d'appariement.

Par exemple:

- Nom (identique)
- Prénom (identique)
- Date de naissance (+/- 1 jour)
- Sexe (identique)

⬆️ règles/conditions :

- Les faux positifs ⬇️
- Les faux négatifs ⬆️

# Probabiliste

Modèle de Fellegi-Sunter (1969) reposant sur 2 probabilités:

- $u$ : probabilité de valeur identique par chance (sexe:  $1/2$ ; mois de naissance:  $1/12$ )
- $m$ : probabilité de valeur identique pour les mêmes entités (1 - taux d'erreurs)

Estimation des probabilités par un gold standard ou par algorithme espérance-maximisation(EM)

# Machine Learning

Apprendre à pondérer l'importance de chaque identifiant (nom, prenom, sexe...) à partir d'exemples (**gold standard**)

- bons appariements
- mauvais appariements

Algorithmes de classification:

- Forêts aléatoires
- Régression logistique
- SVM
- ...

# Blocking stratégie

Le nombre d'appariements possibles est  $n_A \times n_B$

Blocking stratégie: toute stratégie visant à diminuer le nombre de comparaison.

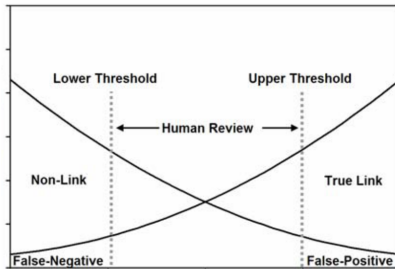
Exemple: "on compare une entité A avec une identité B si au moins 3 identifiants identiques"

# Seuils

Traditionnellement 2 seuils sont choisis<sup>1</sup>:

- Seuil haut qui maximise la valeur prédictive positive (précision)
- Seuil bas qui maximise la sensibilité (rappel)

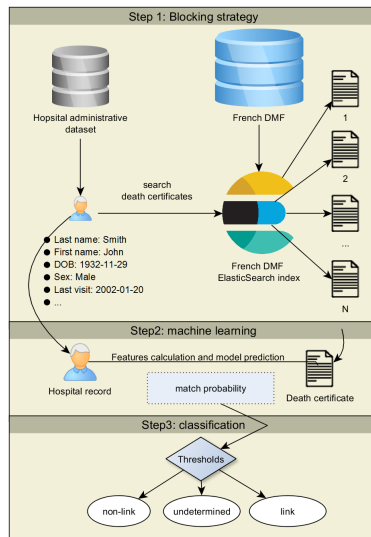
Les appariements au-dessus du seuil haut n'ont pas besoin d'être revus par un humain.



<sup>1</sup>Grannis SJ, Overhage JM, Hui S, McDonald CJ. Analysis of a Probabilistic Record Linkage Technique without Human Review. AMIA Annu Symp Proc. 2003;2003:259-63.





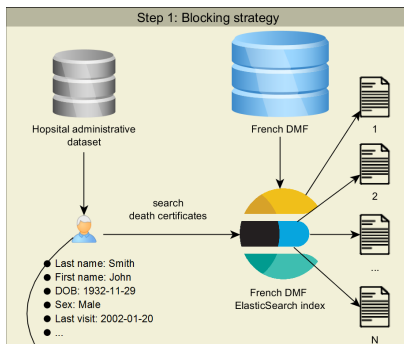
# Pipeline



## Blocking stratégie avec Elasticsearch<sup>1</sup>

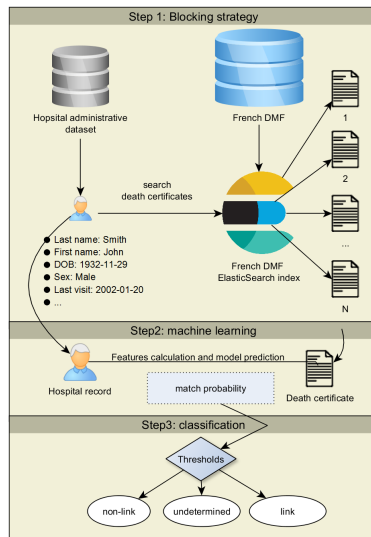
Limiter le nombre de comparaison aux N premiers certificats retournés par Elasticsearch, ordonnés par leur score basé sur le TF-IDF:

- Nombre d'identifiants en commun 
- Valeurs rares (nom, prénom ...) 



<sup>1</sup><https://matchid.io/>

# Pipeline



# Gold standard: bons appariements

Création par approche déterministe avec les **décès intra-hospitaliers 2005-2018**



- Nom: Chirac
- Prenom: Jacques
- DDN: 29/11/1932
- Sexe: M
- LastVisit: 20/01/2002
- Date de décès: 26/09/2019
- Departement de décès
- ...



- Nom: Chirac
- Prenom: Jacques René
- DDN: 29/11/1932
- Sexe: M
- Date de décès: 26/09/2019
- Departement de décès

Blocking stratégie:

- Date de décès (+/- 2 jours)
- Département de décès (Nouvelle-Aquitaine)

4 identifiants identiques parmi 5:

- Nom
- Prenom
- Sexe
- DDN
- Departement de naissance

# Gold standard: bons appariements

44,127 décès alignés. 89% identiques sur les 7 identifiants.

**11%** avaient un identifiant différent:

- 3,7% le nom de famille est différent
- 2,7% le département de naissance est différent
- 2,5% le prénom est différent
- 0.7% la date de naissance est différente
- 0.1% le sexe est différent

# Gold standard: faux appariements

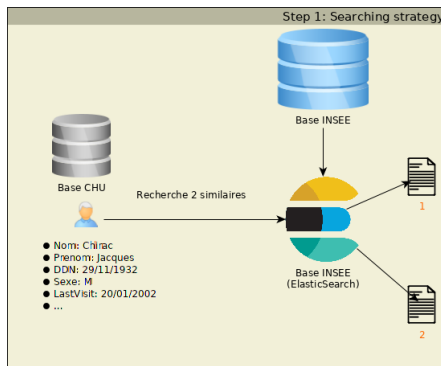
Choisir des faux appariements en "zone grise"<sup>1</sup>.

	Vrai	Faux n°1	Faux n°x
Nom	Chirac	Aupetit	Pompidou
Prénoms	Jacques Rene	Jacques	Georges
Date de naissance	1932-11-29	1932-11-29	05/07/1911
Sexe	Homme	Homme	Homme
Lieu de naissance	Paris 5	Paris 12	Montboudif
Pays de naissance	France	France	France

<sup>1</sup> Capuani L, Bierrenbach AL, Abreu F, et al. Accuracy of a probabilistic record-linkage methodology used to track blood donors in the Mortality Information System database. Cad Saude Publica 2014;30:1623–32.

## Gold standard: faux appariements

On utilise les bons alignements pour rechercher des faux alignements. On sélectionne le faux alignement qui a le score le plus élevé.



# Création des features

Une feature:

$$f : (identifiant_{chu}, identifiant_{insee}) \rightarrow \mathbb{R}$$

Ex: `soundex_nom("Chirac","Chirak") -> 1`

Au total 40 features ont été créées pour l'ensemble des variables:

- Similarité de chaînes de caractères
- Comparaisons de valeurs (0 ou 1)
- Différence de dates: dernière visite - date de décès



# Création de la matrice

- 1 ligne: un couple ( $\text{entité}_{chu}, \text{entité}_{insee}$ ) (bon ou faux)
- 1 colonne: une feature

Ce bon alignement:

	nom	prenom	sexe	DDN	Dep
$\text{entité}_{chu}$	Chirac	Jacques	M	29/11/1932	75
$\text{entité}_{insee}$	Chirak	Jacques Rene	M	30/11/1932	92

Est transformé en:

exact_nom	son_nom	sexe	DDN	Annee	Dep	Region	target
0	1	1	0	1	0	1	1

# *Algorithmes de classification*

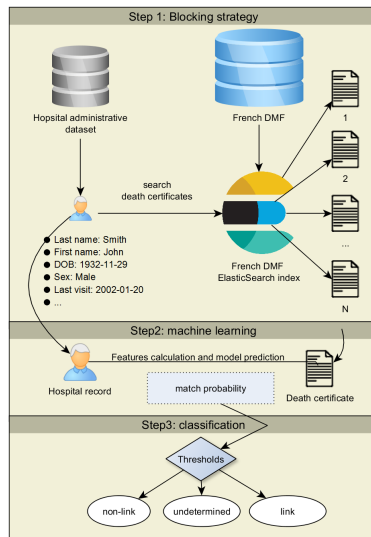
Préparation du jeu de données:

- Sous-échantillonnage (Downsampling)  
(imbalanced dataset 89% / **11%** -> 50% / 50%)
- Normalisation des variables
- Jeu de développement / validation / test (60:20:20)

"Fine-tuner" 2 modèles non linéaires:

- Forêts aléatoires
- Réseau de neurones "fully connected"

# Pipeline



# Evaluation de la pipeline complète

L'année 2019 a été utilisée pour évaluer la pipeline complète:

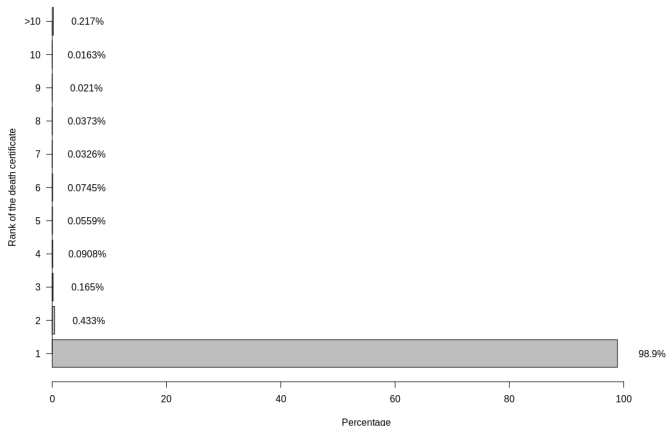
- Décès intra-hospitaliers 2019: évaluer la sensibilité
- Personnes non décédées en 2019: évaluer la précision

Détermination de 2 seuils:

- Seuil haut: minimiser le nombre de faux positifs
- Seuil bas: maximiser la sensibilité

# Nombre N = 10

Parmi les décès intra-hospitaliers liés à un certificat de décès(N=44,127), le certificat de décès apparait dans 99.8% des cas dans les 10 premiers résultats d'Elasticsearch.



# Machine Learning

Jeu de test de 3294 couples ( $record_{chu}$ ,  $record_{insee}$ )

- Forêts aléatoires<sup>1</sup> réalise 34 erreurs<sup>3</sup> (17FP-17FN)
- Réseau de neurones<sup>2</sup> réalise 35 erreurs (21FP-14FN)

---

<sup>1</sup> 1.39% OOB, ntrees=2500, mtry=6

<sup>2</sup> 3 hidden layers, 40:10:20 nodes per layer, 0.2:0.4:0.1 drop out rate

<sup>3</sup> seuil à 0,5. FP: faux positif, FN: faux négatif

# Pipeline

Parmi 3.565 décès intra-hospitaliers en 2019 et 15.000 patients non décédés en 2019.

- Seuil haut: probabilités  $> 0.95$   
Sensibilité: 97.5%, VPP: 99.97%, 1 faux positif<sup>1</sup>
- Seuil bas: probabilités  $> 0.4$   
Sensibilité: 99.4%, VPP: 98.9%

---

<sup>1</sup>jumeau décédé avec le même premier prénom

# En production

Parmi 2,2 millions de patients venus au CHU de Bordeaux:

- 207.507 appariements > seuil haut
  - dont 159.640 (75%) décès extra-hospitaliers
- 29.152 entre les 2 seuils

Performances: 4'30 pour  $10^3$  patients soit 3 jours pour  $10^6$



# Méthode déterministe (baseline)

**200.824 alignements** sur les critères suivants:

Nom, Prénom, Date de naissance, Sexe

- 195.465 appariements en commun avec le seuil haut (Sensibilité = **91.8%**)
- 3.885 entre les 2 seuils
- 1.474 faux alignements (homonymes)

**La pipeline proposée a une sensibilité supérieure à la méthode déterministe et une meilleure spécificité**

# Principaux résultats

- Bonne sensibilité de la pipeline (97.5%) pour un faible taux de faux positif
- 159.640 décès extra-hospitaliers non connus du CHU de Bordeaux, dont 40.000 survenus avant 2011

## Limites:

- Francisation des prénoms (Maria => Marie)
- Inversion nom/prénom
- Etude monocentrique
  - Qualité des données du CHU ?
  - Reproductibilité dans un autre établissement ?

# Merci de votre attention

- [https://github.com/scossin/record\\_linkage\\_insee](https://github.com/scossin/record_linkage_insee)
- Cossin S, Diouf S, Griffier R, Le Barrois d'Orgeval P, Diallo G, Jouhet V. Linkage of Hospital Records and Death Certificates by a Search Engine and Machine Learning. JAMIA Open. doi:10.1093/jamiaopen/ooab005 (accepté).
- Contact: [sebastien.cossin@chu-bordeaux.fr](mailto:sebastien.cossin@chu-bordeaux.fr)