Anonymisation, chaînage

Maxence Guesdon

CHU Dijon - Département d'Information Médicale

8 octobre 2014

Plan

1. Introduction
2. Anonymisation ?
3. Logiciel ANONYMAT
4. Chaînage probabiliste
5. Conclusion

Cadre général

Analyse de données médicales pour

- Etudes statistiques,
- Recherches épidémiologiques,
- Recherches d'antécédents,
- ....

Problématique

- Multiples sources de données, nationales et locales : PMSI, Registres, ...,
- Pas d'identifiant national utilisable,
- Mais nécessité de croiser (chaîner) les données de plusieurs sources : Naissances et transfusions, parcours d'un individu, ...,
- Respect de la loi Informatique et Libertés (et transposition européenne).

Plan

1. Introduction
2. Anonymisation ?
3. Logiciel ANONYMAT
4. Chaînage probabiliste
5. Conclusion

Anonymisation ?

- Rendre impossible l'identification d'une personne,
- Pseudonymisation : aléatoire ou déterministe, doit être irréversible,
- Assez d'information non directement identifiante peut permettre d'identifier (ex : parcours de soins),
- Techniques : hachage, chiffrement, agrégation de données, dégradation des données, ...

Plan

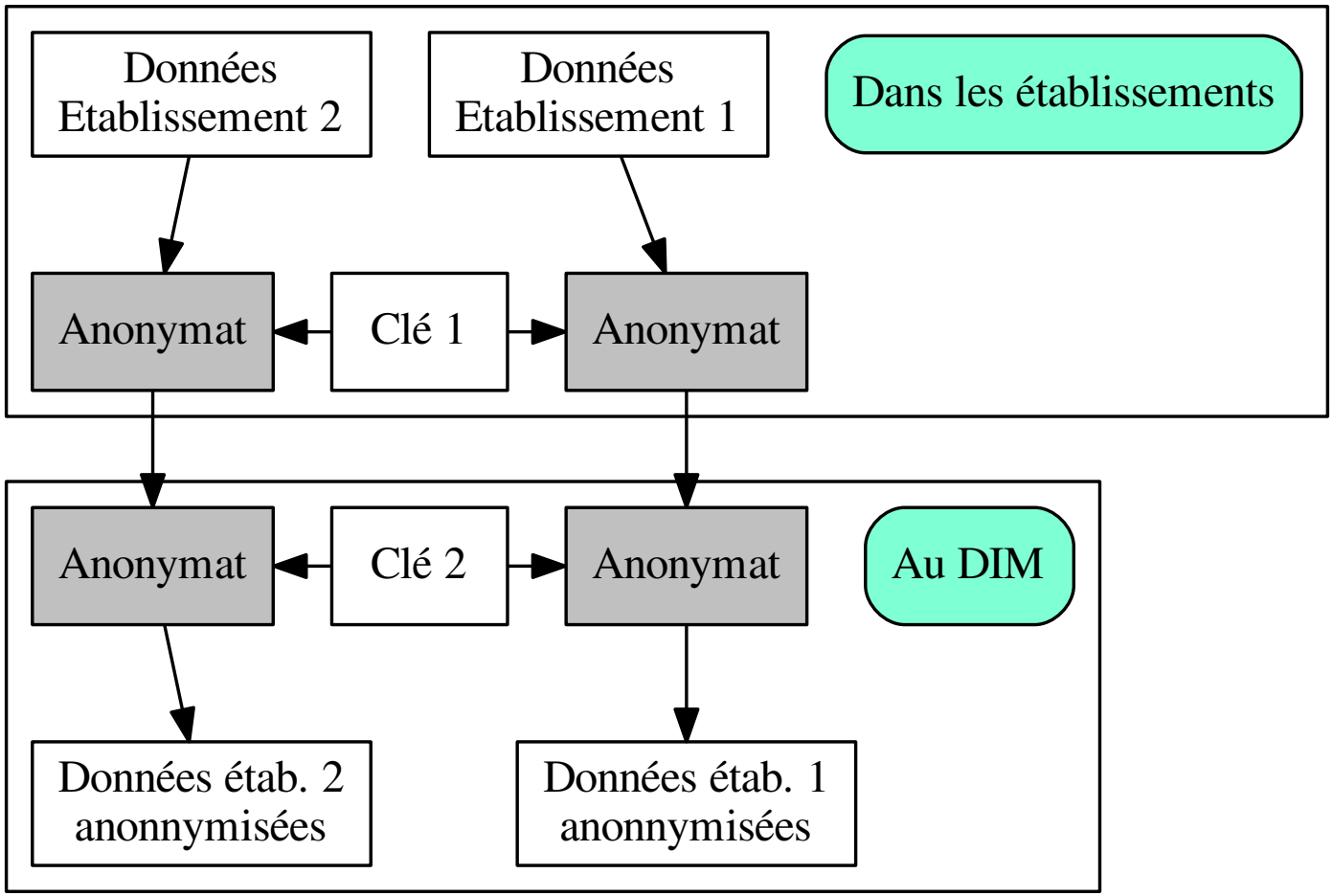
1. Introduction
2. Anonymisation ?
3. Logiciel ANONYMAT
4. Chaînage probabiliste
5. Conclusion

Logiciel ANONYMAT

- Logiciel développé au DIM,
- Accord de la CNIL pour son utilisation,
- Application sur les données d'établissements,
- Normalisations de noms,
- Hachages avec clé secrête dans l'établissement pour pouvoir sortir les données,
- En amont d'un éventuel chaînage en utilisant les informations identifiantes hachées.

Anonymisation des données

Anonymat : Hachage de champs (SHA-1 ou SHA-256) avec clé secrête.
Déterminisme : x = y ⇒ Anonymat(x, clé) = Anonymat(y, clé)



Plan

1. Introduction
2. Anonymisation ?
3. Logiciel ANONYMAT
4. Chaînage probabiliste
5. Conclusion

Objectifs

Rapprocher les données d'un même patient en limitant les erreurs :

- Doublons : Ne pas associer les informations du même individu (changement de nom, erreur de saisie, ...),
- Collisions : Associer à tort les informations de 2 personnes différentes.

Méthode de chaînage

Chaînage d'enregistrements composés de plusieurs variables (champs).
Méthode probabiliste de Jaro :

- Association d'un poids aux différents champs identifiants utilisés pour le chaînage,
- Ce poids diffère selon le pouvoir discriminant du champ ; par exemple, le sexe est moins discriminant que la date de naissance,
- Selon des seuils, traitement automatique aboutissant à chaînage, non chaînage ou zone d'indécision,
- Champs hachés $\Rightarrow$ pas d'utilisation de distances (on peut découper par groupes de lettres, mais plus long).

Calcul de la pondération pour chaque variable

Par exemple le nom :

	même individu	individus $\neq$
même nom	VP	FP = Collision
noms $\neq$	FN = Doublon	VN

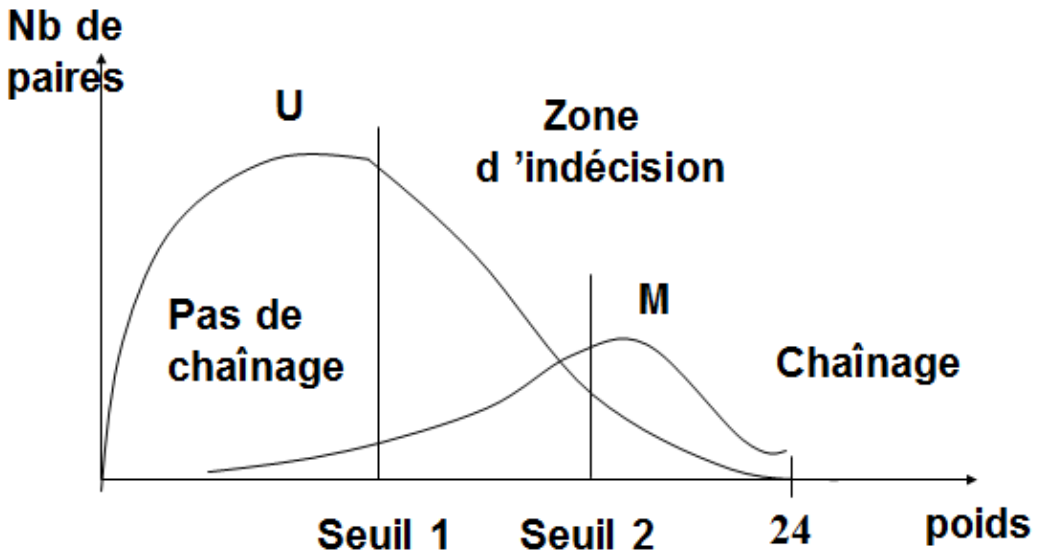
N paires d'enregistrements.
Rapport de vraisemblance $= \frac{1-\text{taux de doublons}}{\text{taux de collisions}}$

Modélisation statistique

2 sous populations théoriques :

- ensemble M des paires correspondant :
  - au même individu,
  - en proportion p,
- ensemble U des paires correspondant à des individus différents.

Modèle par mélange des 2 distributions pour estimer les taux de doublons et de collisions.



Exemple de calcul des taux

	Doublons (%)	Collison (%)
Nom	6.053	0.021
Prénom	3.081	0.335
Date de naissance	4.469	0.003

Exemple de calcul des poids

	Nom	Prénom	Date de naissance	
	Dupont	François	29/01/1940	
	Dupont	François	29/03/1940	
Poids	+8.4	+5.7	-3.1	= 11

- Poids attribué à chaque paire d'enregistrements en fonction
  - de la concordance des différents champs (nom, prénom, ...),
  - de la quantité d'information apportée par chaque champ.

  La concordance peut être calculée selon une méthode adaptée à chaque champ (par exemple comparaison après traitement).
- Décision de chaînage par rapport à un seuil de poids.

	Nom	Prénom	DdN	Total
Sans discordance (111)	+8.4	+5.7	+10.3	+24.4
Discordance sur le nom (011)	-2.8	+5.7	+10.3	+13.2
Discordance sur la DdN (110)	+8.4	+5.7	-3.1	+11
Discordance sur tous les champs (000)	-2.8	-3.5	-3.1	-9.4

Exemple

Concordance Nom	Prénom	DdN	Fréquence	Seuils	Poids	$P(m)$	$G(u)$
0	0	0	1 452 966 248		-9.4	6e-08	99.99
0	1	0	4 880 218		-0.2	5e-04	99.99
1	0	0	304 887		1.8	4e-03	99.99
0	0	1	46 081		1.4	0.04	99.96
1	1	0	1 438	Seuil non chaînage	11	28.79	71.21
0	1	1	725		13.2	78.66	21.34
1	0	1	291		15.2	96.68	3.32
1	1	1	8 852	Seuil chaînage	24.4	99.99	4e-04

$P(m)$: Proba. que les 2 enregistrements de la paire correspondent au même individu.

$G(u)$: Proba. que les 2 enregistrements correspondent à 2 individus différents.

Seuils d'appariement

On chaîne les couples dont les poids composés sont supérieurs à 10 (ad hoc).
La décision finale est **fonction du contexte de l'étude** ; on souhaite :

- un chaînage exhaustif ? $\Rightarrow$ baisser le seuil, donc accepter des faux positifs ou pratiquer une validation
  - automatique ; exemple : les dates de distribution PSL compatibles avec dates d'hospitalisations,
  - ou manuelle ; exemple : retour au dossier médical pour vérifier le diagnostic (cancer, ...).
- identification de vrais positifs ?

Choix des variables

Les variables les plus utiles à garder pour l'appariement :

- Dépend des variables disponibles et de l'étude (ex : date de naissance chez nouveau-nés VS pop. générale),
- Celles ayant la meilleure valeur discriminante (à estimer en fonction de la population étudiée et de la qualité des données),
- Sexe non pertinent (faible valeur discriminante), mais permet de conforter/valider le chainage, pas dans le modèle,
- Commune de résidence, oui si le nombre de communes est important et si les deux sources de données concernent les mêmes dates (pas de déménagement entre temps).

Utilité ou non des traitements phonétiques

- Dépend de la qualité des données.
- Sur nos données, il nous a paru préférable de ne pas appliquer de traitement phonétique car l'algorithme de chainage permet de retrouver les cas douteux sans perte d'information. Appliquer toutefois une normalisation (accents, ponctuation, majuscules/minuscules, etc..).
- Attention, risque de collisions si l'on applique des traitements phonétiques.

Utilité ou non des mesures de distances

Distances de Jaro-Winkler, de Hamming ou de Levenshtein.

- Dans nos travaux, difficile, à cause du hachage.
- Toutefois, il est possible de hacher des n-uplets (n-grams, sous-chaines de la chaine initiale) tout en gardant la possibilité d'un calcul de distance (nombre de n-uplets identiques entre deux chaines de caractères).
- Si non anonyme, combiner le modèle avec le calcul des distances pour les cas en indécision.

Plan

1. Introduction
2. Anonymisation ?
3. Logiciel ANONYMAT
4. Chaînage probabiliste
5. Conclusion

Conclusion

S'il n'y a qu'une seule chose à retenir, c'est que tout (champs identifiants, seuils) dépend du contexte de l'étude :

- besoins,
- données et leur qualité,
- possibilités de vérifications,
- ...

De plus, d'autres traitements peuvent avoir lieu :

- en amont pour limiter la taille du produit cartésien (*blocking*),
- en aval pour conforter les appariements, en utilisant d'autres champs (sexe, commune, ...) non utilisés pour le chaînage avec le modèle.

Merci