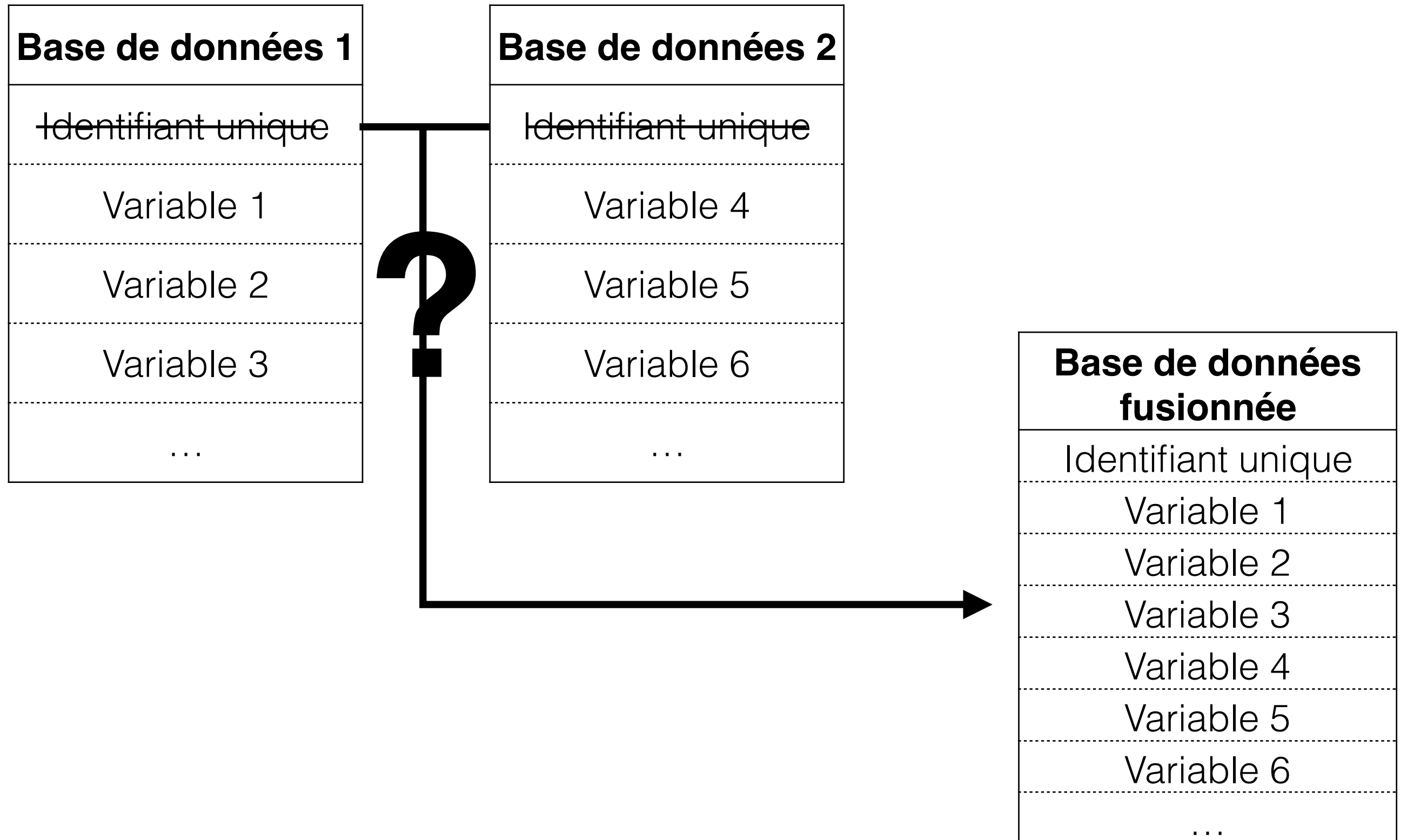


Conciliation des données du PMSI et de la base REA-RAISIN par correspondance approximative

Joris Muller, F. Séverac, S. Deboscker, P. Tran Ba Loc, F. Schneider, T. Lavigne

Médecin spécialisé en Santé Publique
Assistant Hospitalo-Universitaire
Service d'hygiène Hospitalière - CHU de Strasbourg

Problème initial



Problème initial

Base Rea-RAISIN

- Données de qualité sur les infections nosocomiales
- Mais informations manquantes
 - IGSII
 - Parcours à l'hôpital
 - Diagnostic
- Saisie manuelle
- Format variable selon l'année



| Variable | 2004-2006 | 2007-2012 | 2013-2014 |
|-------------------|-----------|---------------------|-----------|
| Nom | Complet | Tronqué à 3 lettres | Absent |
| Prénom | Complet | Tronqué à 3 lettres | Absent |
| Date de naissance | Oui | Oui | Oui |
| Date d'entrée | Oui | Oui | Oui |
| Date de sortie | Oui | Oui | Oui |
| NIP | Non | Partiellement | Oui |

Solution 1 : Retour aux dossiers



Rechercher l'information dans les dossier avec grille de recueil



Avantages

- Simple
- Utilise l'information *a priori* la plus fiable

Inconvénients

- Long et coûteux : > 600h de travail
- Non reproductible
- Saisie peut introduire de nouvelles erreurs

Solution 2 : Lier au PMSI avec correspondance approximative



| | |
|---------------|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Principe | Lier Rea-RAISIN au PMSI |
| Avantage | <ul style="list-style-type: none">• PMSI : information fiable• Exhaustive• Reproductible• Extensible |
| Inconvénients | <ul style="list-style-type: none">• Pas de clef de fusion entre les bases pour identifier le patient et le séjour• Algorithme à développer• Pas de vrai “diagnostic”, (CIM10 et GHM) |


Objectif : comment lier 2 bases de données en l'absence de clef de fusion fiable ?

Inconsistance des données

| | Base 1 | Base 2 |
|-------------------|---------------------|---------------------|
| Prénom | Jean-Michel | JMICHEL |
| Nom | Muller | Meyer |
| Date de naissance | 01/01/1980 | 31/12/1980 |
| Date d'entrée | 0 2 /01/2011 | 0 1 /02/2011 |
| ... | ... | ... |

Correspondance approximative (Fuzzy finding)

Search results





Search

[Content pages](#) [Multimedia](#) [Everything](#) [Advanced](#)

Showing results for **epidemiology**. Search instead for **epiemology**.

*The page "**Epimology**" does not exist.*

Epidemiology

For other uses, see **Epidemiology** (disambiguation). **Epidemiology** is the study and analysis of the patterns, causes, and effects of health and disease conditions

53 KB (6,572 words) - 16:44, 16 July 2016

Distance de Levenshtein

CYBERNETICS AND CONTROL THEORY

BINARY CODES CAPABLE OF CORRECTING DELETIONS, INSERTIONS, AND REVERSALS

V. I. Levenshtein

(Presented by Academician P. S. Novikov, January 4, 1965)

Translated from Doklady Akademii Nauk SSSR, Vol. 163, No. 4, pp. 845-848, August, 1965

Original article submitted January 8, 1965

Investigations of transmission of binary information usually consider a channel model in which failures of the type $0 \rightarrow 1$ and $1 \rightarrow 0$ (which we will call reversals) are admitted. In the present paper (as in [1]) we investigate a channel model in which it is also possible to have failures of the form $0 \rightarrow \Lambda$, $1 \rightarrow \Lambda$, which are called deletions, and failures of the form $\Lambda \rightarrow 0$, $\Lambda \rightarrow 1$, which are called insertions (here Λ is the empty word). For such channels, by analogy to the combinatorial problem of constructing optimal codes capable of correcting s reversals, we will consider the problem of constructing opti-

were inserted (deleted) from at least one of the words x or y to obtain z are deleted from (inserted into) the word z , then, as we can easily see, we obtain a word that can be obtained from both x and y by no more than $\max(i_2 + j_1, j_2 + i_1)$ deletions (insertions). Because x and y have the same length, $j_1 - i_1 = j_2 - i_2$ and, consequently, $i_2 + j_1 = j_2 + i_1 = \frac{1}{2}(i_1 + i_2 + j_1 + j_2) \leq s$, which proves Lemma 1.

Codes that can correct s deletions and insertions admit another, metric, description. Consider a function $\rho(x, y)$ defined on pairs of binary words and equal to the smallest number of deletions and

Distance de Levenshtein

- 1 point si délétion
- 1 point si ajout
- 1 point remplacement

| Chaine de caractères 1 | Chaine de caractères 2 | Distance | Explication |
|------------------------|------------------------|----------|-----------------|
| Marie | Mairie | 1 | 1 Insertion |
| 987761 | 98776 | 1 | 1 Suppression |
| 17/04/1984 | 17/04/1948 | 2 | 2 remplacements |

Score composite de distance =

Minimum

Distance de Levenshtein
Nom / Nom usuel

Distance de Levenshtein
Nom / Nom de naissance

+

Distance de Levenshtein **Prénoms** ÷2

+

Distance de Levenshtein **Dates de naissances**

+

Minimum

Distance de Levenshtein
Dates de sorties ×3

Nombre de jours entre
Dates de sorties ×2

Matrice de comparaison

- Algorithme “brut de force” : calcul des scores composite de distance pour tous les enregistrement d’une base face à tous de l’autre
- Matrice de $7465 * 10\ 605 = 79\ 166\ 325$ de distances calculées (500 Mo)

| | | Observations REA-RAISIN | | | | | | | | |
|-------------------|---|-------------------------|----|----|----|----|----|----|----|----|
| | | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
| Observations PMSI | 1 | 10 | 2 | 2 | 19 | 3 | 14 | 9 | 11 | 7 |
| | 2 | 13 | 8 | 3 | 11 | 9 | 5 | 6 | 20 | 3 |
| | 3 | 17 | 0 | 14 | 5 | 15 | 7 | 14 | 22 | 12 |
| | 4 | 6 | 19 | 16 | 4 | 7 | 14 | 9 | 19 | 16 |
| | 5 | 14 | 13 | 18 | 13 | 2 | 18 | 3 | 13 | 18 |
| | 6 | 0 | 2 | 18 | 3 | 6 | 18 | 5 | 2 | 18 |
| | 7 | 6 | 5 | 19 | 4 | 2 | 7 | 1 | 5 | 19 |
| | 8 | 15 | 3 | 4 | 7 | 4 | 15 | 1 | 0 | 4 |
| | 9 | 16 | 2 | 16 | 0 | 4 | 14 | 14 | 2 | 16 |

Matrice de comparaison

Correspondances parfaites

| | | Observations REA-RAISIN | | | | | | | | |
|-------------------|---|-------------------------|----|----|----|----|----|----|----|----|
| | | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
| Observations PMSI | 1 | 10 | 2 | 2 | 19 | 3 | 14 | 9 | 11 | 7 |
| | 2 | 13 | 8 | 3 | 11 | 9 | 5 | 6 | 20 | 3 |
| | 3 | 17 | 0 | 14 | 5 | 15 | 7 | 14 | 22 | 12 |
| | 4 | 6 | 19 | 16 | 4 | 7 | 14 | 9 | 19 | 16 |
| | 5 | 14 | 13 | 18 | 13 | 2 | 18 | 3 | 13 | 18 |
| | 6 | 0 | 2 | 18 | 3 | 6 | 18 | 5 | 2 | 18 |
| | 7 | 6 | 5 | 19 | 4 | 2 | 7 | 1 | 5 | 19 |
| | 8 | 15 | 3 | 4 | 7 | 4 | 15 | 1 | 0 | 4 |
| | 9 | 16 | 2 | 16 | 0 | 4 | 14 | 14 | 2 | 16 |

Matrice de comparaison

Correspondances parfaites

Aucune vérification nécessaire

| | | Observations REA-RAISIN | | | | | | | | |
|-------------------|---|-------------------------|----|----|----|----|----|----|----|----|
| | | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
| Observations PMSI | 1 | 10 | 2 | 2 | 19 | 3 | 14 | 9 | 11 | 7 |
| | 2 | 13 | 8 | 3 | 11 | 9 | 5 | 6 | 20 | 3 |
| | 3 | 17 | 0 | 14 | 5 | 15 | 7 | 14 | 22 | 12 |
| | 4 | 6 | 19 | 16 | 4 | 7 | 14 | 9 | 19 | 16 |
| | 5 | 14 | 13 | 18 | 13 | 2 | 18 | 3 | 13 | 18 |
| | 6 | 0 | 2 | 18 | 3 | 6 | 18 | 5 | 2 | 18 |
| | 7 | 6 | 5 | 19 | 4 | 2 | 7 | 1 | 5 | 19 |
| | 8 | 15 | 3 | 4 | 7 | 4 | 15 | 1 | 0 | 4 |
| | 9 | 16 | 2 | 16 | 0 | 4 | 14 | 14 | 2 | 16 |

Matrice de comparaison

**Correspondance
approximative
sous seuil de tolérance**

| | | Observations REA-RAISIN | | | | | | | | |
|-------------------|---|-------------------------|----|----|----|----|----|----|----|----|
| | | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
| Observations PMSI | 1 | 10 | 2 | 2 | 19 | 3 | 14 | 9 | 11 | 7 |
| | 2 | 13 | 8 | 3 | 11 | 9 | 5 | 6 | 20 | 3 |
| | 3 | 17 | 0 | 14 | 5 | 15 | 7 | 14 | 22 | 12 |
| | 4 | 6 | 19 | 16 | 4 | 7 | 14 | 9 | 19 | 16 |
| | 5 | 14 | 13 | 18 | 13 | 2 | 18 | 3 | 13 | 18 |
| | 6 | 0 | 2 | 18 | 3 | 6 | 18 | 5 | 2 | 18 |
| | 7 | 6 | 5 | 19 | 4 | 2 | 7 | 1 | 5 | 19 |
| | 8 | 15 | 3 | 4 | 7 | 4 | 15 | 1 | 0 | 4 |
| | 9 | 16 | 2 | 16 | 0 | 4 | 14 | 14 | 2 | 16 |

Matrice de comparaison

**Correspondance
approximative
sous seuil de tolérance**

Vérification manuelle
de 10% des 1897
correspondances

Aucune erreur

| | | Observations REA-RAISIN | | | | | | | | |
|-------------------|---|-------------------------|----|----|----|----|----|----|----|----|
| | | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
| Observations PMSI | 1 | 10 | 2 | 2 | 19 | 3 | 14 | 9 | 11 | 7 |
| | 2 | 13 | 8 | 3 | 11 | 9 | 5 | 6 | 20 | 3 |
| | 3 | 17 | 0 | 14 | 5 | 15 | 7 | 14 | 22 | 12 |
| | 4 | 6 | 19 | 16 | 4 | 7 | 14 | 1 | 19 | 16 |
| | 5 | 14 | 13 | 18 | 13 | 2 | 18 | 3 | 13 | 18 |
| | 6 | 0 | 2 | 18 | 3 | 6 | 18 | 5 | 2 | 18 |
| | 7 | 6 | 5 | 19 | 4 | 2 | 7 | 2 | 5 | 19 |
| | 8 | 15 | 3 | 4 | 7 | 4 | 15 | 2 | 0 | 4 |
| | 9 | 16 | 2 | 16 | 0 | 4 | 14 | 14 | 2 | 16 |

Matrice de comparaison

**Correspondance
approximative
au-dessus
seuil de tolérance**

Vérification manuelle
systématique
(n = 134)

14 exclus

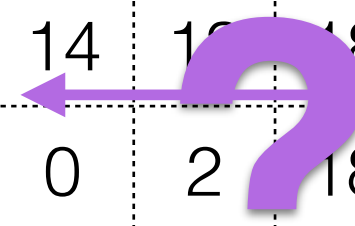
| | | Observations REA-RAISIN | | | | | | | | |
|-------------------|---|-------------------------|----|----|----|----|----|----|----|----|
| | | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
| Observations PMSI | 1 | 10 | 2 | 2 | 19 | 3 | 14 | 9 | 11 | 7 |
| | 2 | 13 | 8 | 3 | 11 | 9 | 5 | 6 | 20 | 3 |
| | 3 | 17 | 0 | 14 | 5 | 15 | 7 | 14 | 22 | 12 |
| | 4 | 6 | 19 | 16 | 4 | 7 | 14 | 9 | 19 | 16 |
| | 5 | 14 | 13 | 18 | 13 | 2 | 18 | 3 | 13 | 18 |
| | 6 | 0 | 2 | 18 | 3 | 6 | 18 | 5 | 2 | 18 |
| | 7 | 6 | 5 | 19 | 4 | 2 | 7 | 1 | 5 | 19 |
| | 8 | 15 | 3 | 4 | 7 | 4 | 15 | 1 | 0 | 4 |
| | 9 | 16 | 2 | 16 | 0 | 4 | 14 | 14 | 2 | 16 |

Matrice de comparaison

**Plus d'une
correspondance
possible**

Vérification manuelle
(< 20 cas)

| | | Observations REA-RAISIN | | | | | | | | |
|-------------------|---|-------------------------|----|----|----|----|----|----|----|----|
| | | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
| Observations PMSI | 1 | 10 | 2 | 2 | 19 | 3 | 14 | 9 | 11 | 7 |
| | 2 | 13 | 8 | 3 | 11 | 9 | 5 | 6 | 20 | 3 |
| | 3 | 17 | 0 | 14 | 5 | 15 | 7 | 14 | 22 | 12 |
| | 4 | 6 | 19 | 16 | 4 | 7 | 14 | 9 | 19 | 16 |
| | 5 | 14 | 15 | 18 | 13 | 2 | 18 | 3 | 13 | 18 |
| | 6 | 0 | 2 | 18 | 3 | 6 | 18 | 5 | 2 | 18 |
| | 7 | 6 | 5 | 19 | 4 | 2 | 7 | 1 | 5 | 19 |
| | 8 | 15 | 3 | 4 | 7 | 4 | 15 | 1 | 0 | 4 |
| | 9 | 16 | 2 | 16 | 0 | 4 | 14 | 14 | 2 | 16 |



Résultats

| Score | n | Proportion | Proportion cumulée |
|----------|------|---------------|--------------------|
| 0 | 5448 | 72,8 % | 72,8 % |
| ≤ 4 | 1897 | 25,2 % | 98,2 % |
| > 4 | 134 | 1,6 % | 99,8 % |
| Exclus | 14 | 0,2 % | 100 % |

Discussion

- Limite :
 - Incertitude
 - Algorithme non optimisé
- Avantages :
 - Logiciel libre : R (fonction *adist*)
 - Reproductible
 - Rapide



Conciliation des données du PMSI et de la base REA-RAISIN par correspondance approximative

Merci !

Joris Muller
joris.muller@jom.link

Médecin spécialisé en Santé Publique
Assistant Hospitalo-Universitaire
Service d'hygiène Hospitalière - CHU de Strasbourg