Recueil des articles du Symposium 2018 de Statistique Canada
Combiner pour vaincre : innovations dans l'utilisation de multiples sources de données

Mise en oeuvre de registres nationaux de santé qui préservent la vie privée

Rainer Schnell and Christian Borgs[1]

Résumé

La plupart des pays développés exploitent des registres de santé tels que les registres néonataux. Ces types de registres sont importants pour des applications en recherche médicale, telles que pour des études de suivi des traitements contre le cancer. Le couplage de ces registres à des données administratives ou à des données d'enquêtes offre des possibilités de recherche, mais peut soulever des préoccupations liées à la protection de la vie privée. En raison de la récente harmonisation des règles en matière de protection des données en Europe avec le Règlement général sur la protection des données (RGPD), il est possible d'établir des critères sur la façon d'exploiter de tels registres tout en préservant la vie privée. Un registre de données sur la santé utilisé à des fins de couplage doit être protégé contre les risques de réidentification, sans compromettre la qualité du couplage.

Nous ferons la démonstration de solutions qui offrent une forte résilience contre les attaques de réidentification tout en préservant la qualité du couplage à des fins de recherche. Plusieurs techniques de pointe pour coupler des enregistrements tout en préservant la vie privée ont été comparées pendant le développement. Pour les essais en situation réelle, nous avons apparié les données sur la mortalité provenant d'un registre administratif local (n = 14 003) avec les dossiers de santé d'un hôpital universitaire (n = 2 466). Un essai à plus grande échelle des solutions proposées a été effectué en appariant 1 million d'enregistrements simulés à partir d'une base de données nationale de noms avec un sous-ensemble corrompu (n = 205 000).

Mots clés : données administratives; données de santé; couplage d'enregistrements; protection des données; risque de réidentification

1. Introduction

Les registres nationaux de santé, comme les registres de mortalité, sont essentiels à la recherche médicale, par ex. pour les études de suivi de traitements conter le cancer. C'est pourquoi la plupart des pays exploitent de tels registres. Coupler de tels registres avec des données administratives ou des enquêtes offre des opportunités de recherche, mais pourrait soulever des préoccupations liées à la protection de la vie privée. En raison de l'harmonisation récente des règles de protection des données en Europe avec le Règlement général de protection des données (RGPD) (Council of the European Union, 2016), des critères peuvent être établis pour exploiter de tels registres tout en préservant la vie privée. Par exemple, le RGPD considère le remplacement d'identificateurs par des pseudonymes comme une façon appropriée pour réaliser la proctection des données par une solution technologique (Voigt & von dem Bussche, 2017).

Nous faison la démonstration qu'un registre national de mortalité est techniquement faisable sous les contraintes associées aux techniques de couplage préservant la vie privée (PPRL, privacy preserving record linkage). Les méthodes PPRL ont été implémentées avec succès dans plusieurs contextes, par ex. coupler les données de santé à travers les états d'Australie (Randall, Ferrante, Boyd, Bauer, & Semmens, 2014), utiliser les méthodes de couplage préservant la vie privée (PPRL, Privacy Preserving Record Linkage) pour un couplage à l'échelle nationale des naissances en Allemagne (Gemeinsamer Bundesausschuss, 2017), et pour lier 114 millions d'enregistrements d'une étude de cohorte à des enregistrements de santé pour une recherche épidémiologique au Brésil (Dantas Pita et coll., 2018).

2. Scénarios d'implémentation

Pour toutes les applications de couplage des données, il est désirable d'utiliser le plus grand nombre possible d'identificateurs stables et sans erreurs. D'habitude, les protocoles PPRL approuvés par les régulateurs de données

[1] Both authors are members of the Research Methodology Group, University of Duisburg-Essen, Lotharstr. 65, 47057 Duisburg, Germany

permettront de crypter les noms et les identificateurs numériques, tels que la date de naissance. Cependant, des demandes de protection plus stricte de la vie privée ou des informations manquantes dans les données peuvent nécessiter un scénario de rechange.

Nous considérons les deux scénarios suivants comme étant les plus probables: (1) les noms, les dates de naissance, et un second identificateur numérique (tel que la date de décès) sont disponibles et (2) les noms et *seulement* la date de naissance sont disponibles, ou il y a des erreurs possibles dans la date de naissance.

Un exemple du scénario (1) est l'étude clinique où un hopital demande la cause du décès d'un seul individu du registre central de mortalité. En utilisant la date de naissance, la date de décès et le sexe on obtient une combinaison essentiellement unique. En nous basant sur notre expérience avec les données de mortalité allemandes, nous nous attendons à avoir moins de 0.5% de doublons en terme de date de naissance et de date de décès. Avec des données complètes et sans erreur, les noms pourraient être superflus pour le couplage. Dans ce cas, le hachage de la date de naissance, de la date de décès et du sexe (avec un mot de passe) donnera un identificateur unique adéquat pour coupler les données tout en préservant la vie privée et sans nécessiter l'usage de noms.

Un autre exemple du scénario (1) est le couplage de données néonatales avec la date de naissance, le sexe, le poids de naissance et le numéro d'hôpital comme identificateurs. Pour l'Allemagne en 2017, cette combinaison a identifié 98.7% des naissances de façon unique. Dans les deux exemples, l'administrateur du couplage pourra se passer des noms comme identificateurs. Cela réduira les obstacles pour obtenir les permissions légales de coupler les enregistrements de façon substantielle.

Si la perte de 0.5% ou de 1.3% des doublons est acceptable, aucune technique PPRL spéciale n'est requise. Si la perte n'est pas acceptable, soit des identificateurs numériques supplémentaires soit des noms cryptés sont requis. Dans ce dernier cas, le scénario (1) sera un cas spécial du second scénario.

Le scénario (2) est basé sur des noms cryptés et des identificateurs numériques supplémentaires. Un exemple serait une étude de cohorte, qui est couplée à des informations administratives. Puisque les noms sont susceptibles aux erreurs, crypter les noms pour le couplage d'enregistrements requiert les techniques PPRL.

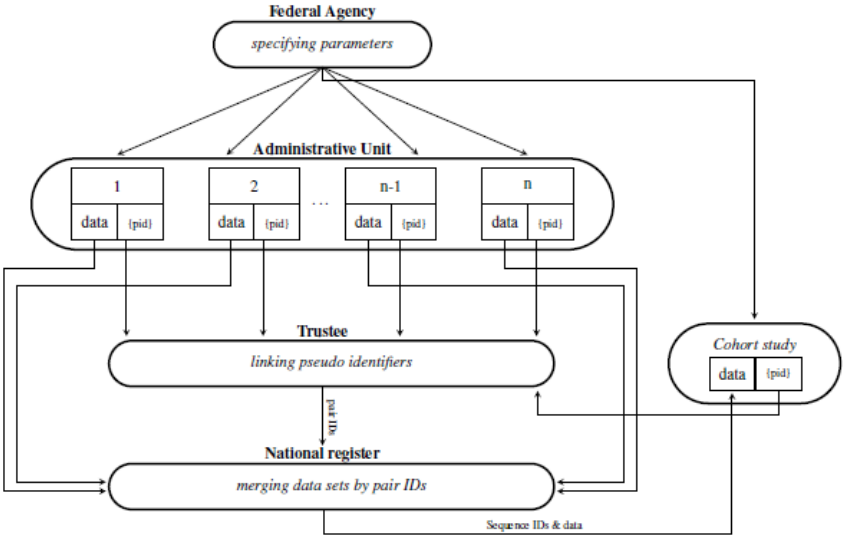
Dans cet exemple, l'étude de cohorte crypte les identificateurs selon un protocole centralisé, qui est inconnu de l'administrateur et du registre.

Dans cet article, nous étudierons si les identificateurs cryptés incluant les noms sont convenables pour identifier les patients dans les registres nationaux avec les méthodes préservant la vie privée selon le second scénario.

Nous supposons les contraintes suivantes :

1. Aucun numéro d'identification unique n'est disponible.
2. Les bases de données ne sont pas connectées à internet.
3. Nous exploitons une structure administrative et opérationnelle décentralisée.
4. Le protocole central est connu des deux parties qui cryptent.

Figure 2-1
Le protocole de couplage considéré ici. Les contraintes: pas d'identificateur unique, une structure décentralisée, pas de calculs en ligne.



Note: Simplified protocol assuming same parameter settings for all parties.

Étant donné ces contraintes, la figure 2-1 montre le processus de couplage proposé et les parties impliquées. Une agence fédérale agit comme autorité de surveillance, en spécifiant les paramètres de chiffrement utilisés pour crypter les données des unités administratives. Les mêmes paramètres doivent être connus de l'équipe de l'étude de cohorte. Les pseudo-identificateurs cryptés (PID, pseudo-identifiers) sont ensuite transmis à l'administrateur, qui couple seulement les PIDs. La paire de PIDs résultante est ensuite transmise au registre national, qui joue le rôle de conservateur des données. L'étude de cohorte recevra maintenant les PIDs du lien et les données requises, qui peuvent être fusionnées avec le fichier de données de la cohorte. La décision centralisée consiste à choisir les paramètres pour chiffrer les données sensibles, pour réaliser un couplage de haute qualité tout en prenant en compte les soucis relatifs à la vie privée.

2.1 Préoccupations relatives à la vie privée

Dans la mesure où coupler les registres requiert la divulgation d'informations personnelles à des tiers de confiance (Boyd et coll., 2012), les règlements sur la vie privée, tels que les règlements actuels de l'Union Européenne (Council of the European Union, 2016), exigent souvent l'usage des informations personnelles sous forme cryptée. Les méthodes habituelles de couplage probabiliste (Herzog, Scheuren, & Winkler, 2007) s'appuient sur les similitudes entre les chaînes de caractères et ne conviennent donc pas aux méthodes basées sur des identificateurs cryptés, dans la mesure où les similitudes sont modifiées par le hachage. Ces 15 dernières années, un nombre de méthodes ont été développées pour surmonter ce problème dans le contexte du couplage d'enregistrements. Ces techniques forment maintenant le domaine de recherche appelé couplage préservant la vie privée (PPRL). Les techniques PPRL permettent le couplage d'enregistrements à l'aide d'identificateurs cryptés. Par conséquent, aucune information de personnes physiques n'est divulguée par les dépositaires des données, dans la mesure où les identificateurs sont d'abord remplacés par des pseudonymes. Toutefois, en utilisant les techniques PPRL, il reste possible de coupler les enregistrements en tolérant des erreurs.

3. Méthodes

Un registre de données de santé utilisé pour le couplage doit être protégé des attaques de réidentification tout en permettant des couplages de haute qualité. Nous ferons la démonstration de solutions offrant une grande résistance aux attaques de réidentification tout en préservant la qualité de couplage à des fins de recherche. Plusieurs techniques

PPRL de pointe ont été comparées pendant le développement. Pour les tests sous des conditions réelles, nous avons couplé des données de mortalité d'un registre administratif régional (n = 14 003) avec des enregistrements de santé d'un hôpital universitaire (n = 909). Un test à plus grande échelle a été effectué en couplant 1 million d'enregistrements simulés à partir d'une base de données nationale de noms à un sous-ensemble corrompu (n = 205 000). Cela correspond à peu près à la taille du registre et aux décès annuels au niveau national.

3.1 Méthodes de cryptage PPRL

Pour le scénario considéré ici, seulement deux formes de cryptage sont largement employées (Randall, Ferrante, Boyd, Brown, & Semmens, 2016): clés cryptées pour le couplage statistique (ESLs, Encrypted Statistical Linkage Keys) et filtres de Bloom (BF, Bloom Filters). Les deux solutions seront décrites brièvement.

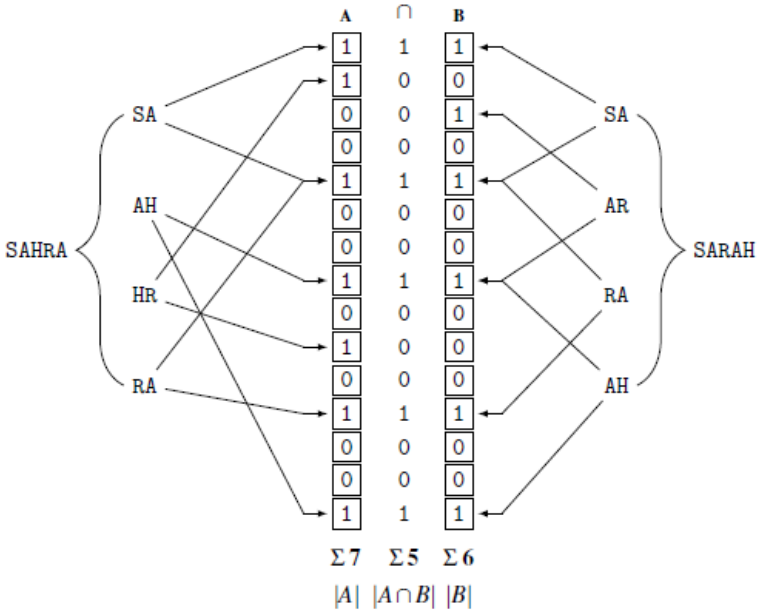
3.1.1 Clés-581

La clé-581, ou clé cryptée de couplage statistique (ESL, Encrypted statistical linkage key) (Karmel, 2005) est construite en concaténant les 2ème et 3ème lettres du prénom, les 2ème, 3ème et 5ème lettres du nom de famille, la date de naissance complète et le sexe. La chaîne de caractères résultante est cryptée avec une fonction de hachage (telle que MD5, ou mieux SHA-3) donnant la clé de couplage. Étant donné l'enregistrement John O'Shea, 1.9.1967, mâle, la chaîne de caractères OHSHA01091967M est créée. Appliquer la fonction de hachage (MD5 ici) donne le hachage ab76990b084b82d3e06701c52d02485e8e2ba9fe, qui est utilisé pour le couplage exact.

3.1.2 Filtres de Bloom

D'abord suggéré par Bloom (1970), les filtres de Bloom ont été suggérés comme technique PPRL par Schnell, Bachteler, et Reiher (2009). La Figure 3-1-2-1 donne un exemple de construction de filtres de Bloom pour deux prénoms. Dans cet exemple, les bigrammes des deux noms sont convertis en filtres de Bloom de longueur l = 15 bits avec k = 2 fonctions de hachage.

Figure 3-1-2-1:
Exemple de construction de deux filtres de Bloom de longueur l = 15 bits avec k = 2 fonctions de hachage.



SAHRA et SARAH ont en commun 3 bigrammes sur 4 (SA, AH et RA) et diffèrent par un seul bigramme (HR et AR, respectivement). Le coefficient de Dice non-crypté pour les deux noms est:

$$D(A,B) \quad = \quad \frac{2|A \cap B|}{|A| + |B|} \quad = \quad \frac{2 \times 3}{4 + 4} \quad = \quad 0.75$$

Ici, 7 bits et 6 bits sont mis à un dans les filtres de Bloom respectifs, qui s'accordent sur 5 positions.
Ainsi, le coefficicent de Dice peut être estimé par:

$$D(A,B) \quad = \quad \frac{2|A \cap B|}{|A| + |B|} \quad = \quad \frac{2 \times 5}{7 + 6} \quad = \quad 0.77$$

Cela permet d'estimer la similitude de n-grammes en texte clair à partir d'identificateurs cryptés.

3.2 Renforcer les filtres de Bloom contre les attaques

Les filtres de Bloom ont été soumis à des attaques cryptographiques (Niedermeyer, Steinmetzer, Kroll, & Schnell, 2014), ce pourquoi plusieurs techniques (méthodes de « renforcement ») ont été suggérées pour empêcher la réidentification (Niedermeyer, Steinmetzer, Kroll, & Schnell, 2014). Toutes les attaques connues sur la technique PPRL basée sur les filtres de Bloom (BF, Bloom Filters) sont soit des attaques basées sur les fréquences ou des attaques basées sur les analyses séquentielles. Ils nécessitent des filtres de Bloom fréquents ou des concomitances fréquentes de bigrammes. Selon le type d'attaque, la limite inférieure du nombre requis de séquences fréquentes diffère. Par conséquent, toutes les techniques de renforcement essaieront de réduire le nombre de séquences ou de concomitances fréquentes de bigrammes.

3.3 Exemples de techniques de renforcement de filtres de Bloom

Niedermeyer et coll. (2014) ont recommandé le salage, où une valeur statique en texte clair (telle que l'année de naissance) est utilisée comme chaîne de caractères supplémentaire pour le mot de passe utilisé lors de l'encodage des BFs. Coder différents identificateurs (prénom, nom de famille, date de naissance, sexe) dans le même filtre de Bloom permet aussi de coupler en préservant la vie privée. Cette construction est appelée clé cryptographique à long terme (CLK, Cryptographic Long-term Key, Schnell, 2014) et offre une grande protection contre les attaques. Bien sûr, le salage peut être appliqué aux CLKs pour obtenir des CLKs avec sel. Dans la mesure où le nombre de bits mis à un dans les BFs est essentiel pour plusieurs attaques de fréquence, Schnell et Borgs (2016) ont proposé l'usage d'un BF équilibré, où chaque BF avec sel de longueur $2*l$ a un poids de Hamming de l. À ce jour, aucune attaque réussie sur les CLKs avec sel n'a été signalée. Appliquer des renforcements supplémentaires, tel que l'équilibrage, rend les attaques encore plus difficiles.

3.4 Évaluation des méthodes de couplage

Pour comparer les méthodes PPRL à une méthode de référence, nous avons utilisé le couplage probabiliste avec des identificateurs non chiffrés. Cette méthode de référence a été comparée à la méthode clé-581 largement utilisée et à deux versions différentes de la méthode PPRL à base de filtre de Bloom : un CLK de base et un CLK avec sel. Les CLKs de base ont utilisé k = 20 fonctions de hachage avec une longueur l = 1000 bits, les CLKs avec sel avec la date de naissance comme sel ont utilisé k = 30 fonctions de hachage. Pour la première évaluation, nous utilisons des données réelles (n = 14 003 et n = 909; le chevauchement réel était n = 889). Pour évaluer l'impact des erreurs sur la qualité de couplage et le fonctionnement à plus grande échelle, des données simulées ont été utilisées (n = 1 million et n = 205 000) avec un pourcentage allant de 0% à 20% pour les lignes dont les identificateurs (prénoms et noms de famille, sexe et date de naissance) contenaient des erreurs.

3.5 Seuils de similitude pour les méthodes basées sur le filtre de Bloom

Comme technique de couplage PPRL basée sur le filtre de Bloom, nous utilisons les arbres multibit. Les arbres multibit ont été proposés en chimiométrie par Kristensen, Nielsen, et Pedersen (2010) et utilisés comme technique PPRL par Schnell (2014). Les paires possibles en dessous d'un seuil de similitude pré-défini ne sont pas évaluées. La similitude de Tanimoto T est utilisée comme mesure de similitude. T est défini comme le nombre de bits mis à 1 dans l'intersection des vecteurs A et B divisé par le nombre total de bits mis à 1 à la fois dans A et dans B:

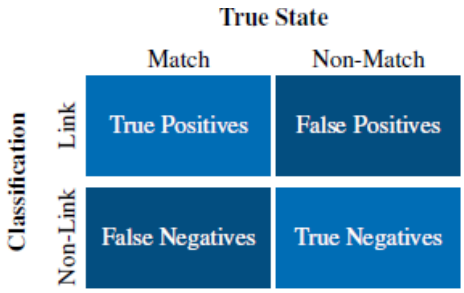
$$T(A,B) \quad = \quad \frac{\sum_i (A_i \wedge B_i)}{\sum_i (A_i \vee B_i)}$$

Des seuils de similitude plus bas conduiront à un plus gand nombre de comparaisons de paires et de décisions faussement positives. Réciproquement, le nombre de vrais positifs augmentera également. Pour les simulations, on a fait varier les seuils de Tanimoto entre 0.75 et 1.0 par incréments de 0.05.

3.6 Mesures pour l'évaluation

Toutes les méthodes de couplage classerons les paires d'enregistrements comme des liens ou des non-liens. Ce classement est soit correct soit incorrect. La Figure 3-6-1 montre la matrice de classification résultante, qui donne les vrais positifs (TP, true positive), les faux positifs (FP, false positive) ou les faux négatifs (FN, False negative).

Figure 3-6-1:
Matrice de classification pour évaluer les décisions des méthodes de couplage.



Avec ces décisions, les mesures les plus répandues de qualité de couplage, le rappel

$$Rappel = \frac{TP}{TP + FN} \qquad (1)$$

et la précision

$$Précision = \frac{TP}{TP + FP} \qquad (2)$$

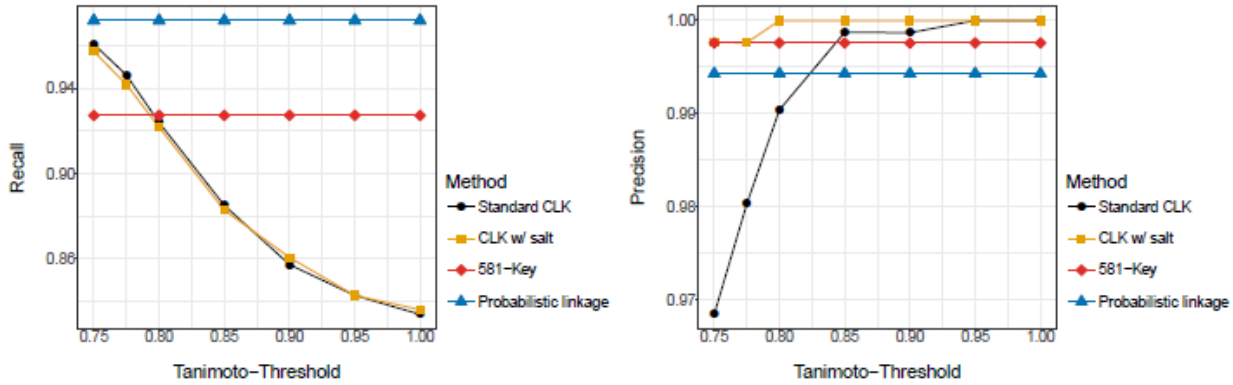
sont calculés. Nous mentionnons aussi la moyenne arithmétique de la précision et du rappel au lieu du score F traditionnel, qui a récemment été critiqué (Hand & Christen, 2017).

4. Résultats

4.1 Évaluation de stratégies de couplage avec des données réelles

Pour l'évaluation avec des données administratives réelles, la figure 4-1-1 montre la précision et le rappel pour toutes les stratégies de couplage utilisées.

Figure 4-1-1:
Rappel et précision par méthode de cryptage avec plusieurs seuils de Tanimoto.



Toutes les méthodes donnent une grande précision, ce qui indique peu de faux positifs. Pour les méthodes basées sur le filtre de Bloom, le rappel augmente quand les seuils de similitude baissent. À un seuil en deça de 0.8, la qualité de couplage des méthodes basées sur le filtre de Bloom dépasse celle avec la technique clé-581. Comme on pouvait s'y attendre, le couplage probabiliste non chiffré est supérieur à toutes les méthodes PPRL. Le cryptage d'identificateurs susceptibles aux erreurs conduira à réduire la qualité de couplage. Pour éclairer les conséquences pratiques des résultats, la Table 4-1-1 montre les nombres absolus de vrais positifs et de faux positifs pour chaque méthode.

Le vrai chevauchement des jeux de données administratifs était de 898 enregistrements. 97% de toutes les vraies paires ont été trouvées par le couplage probabiliste, avec moins de 0.006% de faux positifs. 96% de toutes les vraies paires ont été trouvées par les CLKs avec sel, avec moins de 0.003% de faux positifs. Par conséquent, nous avons perdu à peine 13 cas (environ 1.4%) du fait de la technique PPRL, tout en réglant les problèmes relatifs à la vie privée. La perte de qualité due à cette méthode PPRL semble acceptable pour la plupart des applications où des informations sensibles doivent être couplées.

4.2 Évaluer des stratégies de couplage avec des fichiers simulés volumineux comprenant des erreurs

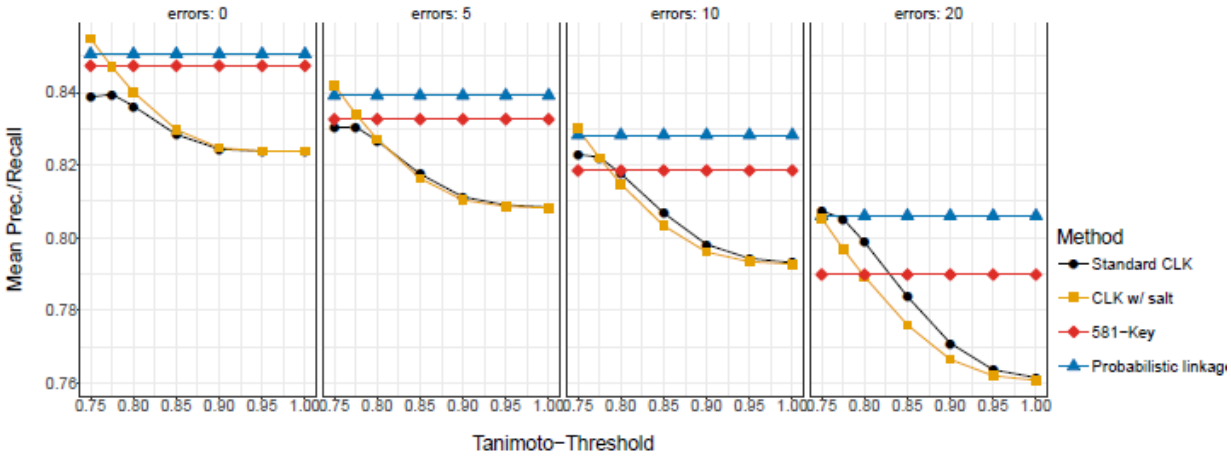
Pour évaluer l'impact des erreurs sur la qualité de couplage dans de grands fichiers, comme c'est le cas avec les registres nationaux, des données simulées ont été utilisées. Nous avons simulé les erreurs pour un pourcentage de lignes allant de 0% à 20% dans un sous-ensemble de 250 000 enregistrements d'un fichier principal de 1 million d'enregistrements.

Table 4-1-1
Vrais positifs, faux positifs, précision et rappel résultants par méthode avec un seuil de Tanimoto t = 0.75.

Method	Rec.	Prec.	TP	FP
Exact match	0.83	1.00	747	0
Probabilistic linkage	0.97	0.99	871	5
581-Key	0.93	1.00	831	2
Standard CLK	0.96	0.97	861	28
CLK w/ salt	0.96	1.00	858	2

Afin de résumer la qualité de couplage, la moyenne de la précision et du rappel est calculée. La Figure 4-2-1 montre les résultats.

Figure 4-2-1
Moyenne du rappel et de la précision par méthode de cryptage et pourcentage d'observations avec des erreurs avec plusieurs seuils de Tanimoto.



Ici aussi, le couplage probabiliste donne les meilleurs résultats. Toutes les méthodes montrent une baisse soutenue de la qualité du couplage avec un accroissement des taux d'erreur dans les identificateurs. Aux deux plus bas niveaux de la similitude de Tanimoto, la performance relative des méthodes basées sur le filtre de Bloom par rapport à la méthode clé-581 s'améliore lorsque la qualité des données diminue. Toutefois, même avec 20% d'erreurs dans 250 000 identificateurs, les CLKs avec sel produisent seulement 145 faux positifs (0.06%), tout en donnant un rappel proche de celui du couplage probabiliste. Ce niveau de qualité semble acceptable pour la plupart des problèmes nécessitant les techniques PPRL. Bien sûr, selon les coûts des faux positifs ou des faux négatifs pour une application donnée, il pourrait y avoir des exceptions.

5. Conclusion

Le couplage probabiliste basé sur les identificateurs en texte clair surpasse toutes les autres méthodes. Par comparaison, crypter des identificateurs susceptibles d'erreurs réduit toujours la qualité de couplage. Comme dans d'autres comparaisons antérieures (Randall et coll., 2016), avec des seuils de similitude bas, les CLKs surpassent les clés-581. Pour plusieurs applications, les CLKs avec sel semblent produire des résultats légèrement inférieurs au couplage probabiliste avec du texte en clair. Si un registre national en texte clair n'est pas une bonne option, les CLKs avec sel pourraient être la deuxième meilleure solution. Toutefois, nous considérons que la seule option faisable est un registre national de santé utilisant seulement des identificateurs des patients cryptés. Néanmoins, en pratique, implémenter le protocole correspondant sera difficile (Kho et coll., 2015).

Bibliographie

Bloom, B. H. (1970). Space/time trade-offs in hash coding with allowable errors. *Communications of the ACM*, *13*(7), 422–426.

Boyd, J. H., Ferrante, A. M., O'Keefe, C. M., Bass, A. J., Randall, S. M., & Semmens, J. B. (2012). Data linkage infrastructure for cross-jurisdictional health-related research in australia. *BMC Health Services Research*, *12*(1), 480.

Council of the European Union. (2016). Council regulation (EU) no 679/2016.

Dantas Pita, R., Pinto, C., Sena, S., Fiaccone, R., Amorim, L., Reis, S., Barreto, M. (2018). On the accuracy and scalability of probabilistic data linkage over the Brazilian 114 million cohort. *IEEE Journal of Biomedical and Health Informatics*, *22*(2), 346–353.

Gemeinsamer Bundesausschuss. (2017). Beschluss des Gemeinsamen Bundesausschusses über eine Beauftragung des Instituts nach § 137a SGB V: Vergleich der Methoden des Bloom-Filters und des Krebsregisterverfahrens zur Verknüpfung der Leistungsbereiche Geburtshilfe und Neonatologie und Entwicklung von entsprechenden (Follow-up-) Qualitätsindikatoren. https://www.g-ba.de.

Hand, D., & Christen, P. (2017). A note on using the F-measure for evaluating record linkage algorithms. *Statistics and Computing*. doi:10.1007/s11222-017-9746-6

Herzog, T. N., Scheuren, F. J., & Winkler, W. E. (2007). *Data quality and record linkage techniques*. New York: Springer.

Karmel, R. (2005). *Data linkage protocols using a statistical linkage key*. Canberra: AIHW.

Kho, A. N., Cashy, J. P., Jackson, K. L., Pah, A. R., Goel, S., Boehnke, J., ... Galanter, W. L. (2015). Design and implementation of a privacy preserving electronic health record linkage tool in chicago. *Journal of the American Medical Informatics Association*, *22*(5), 1072–1080.

Kristensen, T. G., Nielsen, J., & Pedersen, C. N. (2010). A tree-based method for the rapid screening of chemical fingerprints. *Algorithms for Molecular Biology*, *5*(9).

Niedermeyer, F., Steinmetzer, S., Kroll, M., & Schnell, R. (2014). Cryptanalysis of basic bloom filters used for privacy preserving record linkage. *Journal of Privacy and Confidentiality*, *6*(2), 59–69.

Randall, S. M., Ferrante, A. M., Boyd, J. H., Bauer, J. K., & Semmens, J. B. (2014). Privacy-preserving record linkage on large real world datasets. *Journal of Biomedical Informatics*, *50*, 205–212.

Randall, S. M., Ferrante, A. M., Boyd, J. H., Brown, A. P., & Semmens, J. B. (2016). Limited privacy protection and poor sensitivity: Is it time to move on from the statistical linkage key-581? *Health Information Management Journal*, *45*(2), 71–79.

Schnell, R. (2014). An efficient privacy-preserving record linkage technique for administrative data and censuses. *Journal of the International Association for Official Statistics*, *30*(3), 263–270.

Schnell, R., Bachteler, T., & Reiher, J. (2009). Privacy-preserving record linkage using Bloom filters. *BMC Medical Informatics and Decision Making*, *9*(41).

Schnell, R., & Borgs, C. (2016). Randomized response and balanced bloom filters for privacy preserving record linkage. In 2016 IEEE 16th International Conference on Data Mining Workshops (ICDM 2016), Barcelona, December 12th 2016 – December 15th 2016: IEEE Publishing.

Voigt, P., & von dem Bussche, A. (2017). *The EU general data protection regulation (GDPR): A practical guide*. Cham: Springer.