Proceedings of Statistics Canada Symposium 2018
Combine to Conquer: Innovations in the Use of Multiple Sources of Data

Implementing Privacy-preserving National Health Registries

Rainer Schnell and Christian Borgs[1]

Abstract

Most developed nations operate health registers such as neonatal birth registries. These kind of registries are important for medical research applications, for example, follow-up studies of cancer treatments. Linking such registers with administrative data or surveys offers research opportunities, but may raise privacy concerns. Due to the recent harmonisation of data protection rules in Europe with the General Data Protection Regulation (GDPR), criteria for operating such registers in a privacy-preserving way can be derived. A health data register used for linking needs to be secured against re-identification attacks while retaining high linkage quality.

We will demonstrate solutions providing strong resilience against re-identification attacks while preserving linkage quality for research purposes. Several state of the art privacy-preserving record linkage (PPRL) techniques were compared during the development. For real-world testing, we matched mortality data from a local administrative registry (*n*= 14,003) with health records of a university hospital (*n*= 2,466). Scaling of the proposed solutions was tested by matching 1 million simulated records from a national database of names with a corrupted subset(*n*= 205,000).

Key Words: Administrative data; Health data; Record linkage; Data protection; Re-identification risk.

1. Introduction

National health registers like mortality registries are essential for medical research, e.g. for follow-up studies of cancer treatments. Therefore, most nations operate such registers. Linking such registers with administrative data or surveys offers research opportunities, but may raise privacy concerns. Due to the recent harmonisation of data protection rules in Europe with the General Data Protection Regulation (GDPR, Council of the European Union, 2016), criteria for operating such registers in a privacy-preserving way can be derived. For example, the GDPR considers the use of pseudonymization of identifiers as an appropriate measure for achieving data protection through technology (Voigt & von dem Bussche, 2017).

We demonstrate that a national mortality registry is technically feasible under the given constraints with privacy preserving record linkage (PPRL). PPRL methods have been successfully implemented in several settings, e.g. to link health data across states in Australia (Randall, Ferrante, Boyd, Bauer, & Semmens, 2014), using PPRL methods used for a nation-wide record linkage of births in Germany (Gemeinsamer Bundesausschuss, 2017), and to link 114 million records from a cohort-study to health records for epidemiological research in Brazil (Dantas Pita et al., 2018).

2. Implementation scenarios

For all data linkage applications, it is desirable to use as many stable and error-free identifiers for linking as possible. Usually, PPRL protocols approved by the data regulators will allow encrypting names and numeric identifiers, such as the date of birth (DOB). However, stricter privacy demands or missing information in the data can require a fall-back scenario.

We consider two scenarios as most likely: (1) Names, dates of birth, and a second numerical identifier (such as the date of death) are available and (2) Names and *only* the date of birth is available, or there are likely errors in the DOB.

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

An example for scenario (1) is a clinical study where a hospital requests the cause of death for a single individual from a central mortality register. Using date of birth and date of death (DOD) together with sex will form a nearly unique combination. In our experience with German mortality data, we expect less than 0.5% duplicates in terms of DOB and DOD within a year. Given complete and flawless data, names might not be needed for linking. In this case, hashing DOB, DOD and sex (with a password) will give a unique identifier suitable for privately linking data without requiring names to be used.

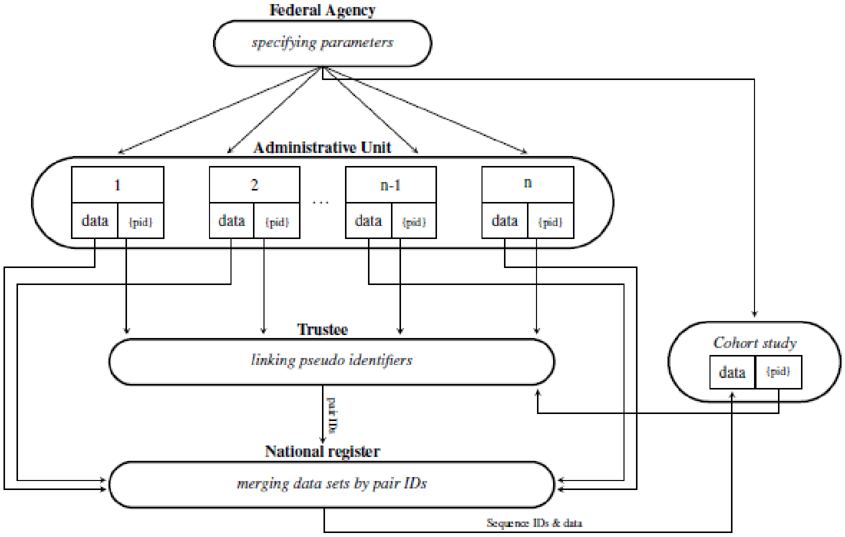
Another example for scenario (1) is linking neonatal data using DOB, sex, birth weight and hospital ID as identifiers. For Germany in 2017, this combination identified 98.7% of all births uniquely. In both examples, the linking trustee will not require names as identifiers. This will reduce the obstacles for attaining legal permissions to link records substantially. If the loss of about 0.5% resp. 1.3% duplicates is acceptable, no special PPRL technique is required. If the loss is not acceptable, either additional numerical identifiers or encrypted names are needed. In the latter case, scenario (1) will be a special case of the second scenario.

Scenario (2) is based on encrypted names and additional numerical identifiers. An example would be a cohort study, which is linked to administrative information. Since names are error-prone, encrypting names for record linkage requires PPRL. In this example, the cohort study encrypts identifiers according to a centrally specified protocol which is unknown to the trustee and the registry.

In this paper, we will study whether encrypted identifiers including names are suitable for identifying patients within national registers using privacy-preserving methods under the second scenario. We assume the following constraints:

1. No unique identifying number is available.
2. The data bases are not connected to the internet.
3. We operate in a decentralised administrative and operational structure.
4. The central protocol is known to both encrypting parties.

Figure 2-1
The linkage protocol considered here. Constraints: No unique ID, decentralized structure, no on-line computations.



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

Given these constraints, figure 2-1 shows the proposed linking process and the parties involved. A federal agency acts as an overseeing authority, specifying encryption parameters used for encrypting data in the administrative units. The same parameters must be known to the team of the cohort study. The encrypted pseudo-identifiers (PID) are then sent to a trustee, which only links the PIDs. The resulting pair IDs are then sent to the national register, which acts as the data custodian. The cohort study will now receive the link IDs and the data requested, which they can merge with their cohort data file. The central decision to make is the choice of the parameters for encrypting the sensitive data, to achieve high linkage quality while addressing privacy concerns.

2.1 Privacy concerns

Since linking registries requires the release of personally identifying information to trusted third parties (Boyd et al., 2012), privacy regulations, such as the current EU regulations (Council of the European Union, 2016), often mandate using encrypted personal information. Standard probabilistic record linkage methods (Herzog, Scheuren, & Winkler, 2007) depend on string similarities, and are therefore unsuitable for methods based on encrypted identifiers, as similarities are not preserved by hashing. In the last 15 years, a number of new methods have been developed to overcome this problem in record linkage settings. These techniques now form the research field called *privacypreserving record linkage* (PPRL). PPRL techniques enable linking records using encrypted identifiers. Therefore, no information of natural persons is released by data custodians, as identifiers are subject to prior pseudonymization. However, using PPRL techniques, error-tolerant record linkage is still possible.

3. Methods

A health data register used for linking needs to be secured against re-identification attacks while retaining high linkage quality. We will demonstrate solutions providing strong resilience against re-identification attacks while preserving linkage quality for research purposes. Several state of the art privacy-preserving record linkage (PPRL) techniques were compared during the development. For real-world testing, we matched mortality data from a local administrative registry (n = 14,003) with health records of a university hospital (n = 909). Scaling of the proposed solutions was tested by matching 1 million simulated records from a national database of names with a corrupted subset (n = 205,000). This roughly corresponds to the size of a state-wide register and all deaths of a single year.

3.1 Encryption methods for PPRL

For the scenario considered here, only two variants of encryption are widely used (Randall, Ferrante, Boyd, Brown, & Semmens, 2016): Encrypted Statistical Linkage Keys (ESLs) and Bloom filters (BF). Both will be described briefly.

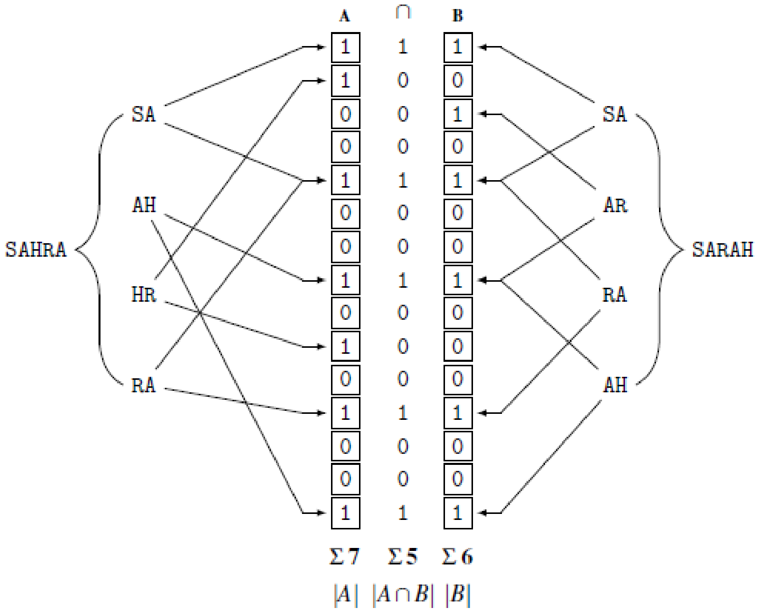
3.1.1 581-Keys

The 581-Key, or Encrypted statistical linkage key (ESL) (Karmel, 2005) is built by concatenating the second and third letter of the first name, the second, third and fifth letter of the last name, the full date of birth and sex. The resulting string is encrypted using a hash function (such as MD5, or better SHA-3), giving the linkage key. Given the record John O’Shea, 1.9.1967, male, the string OHSHA01091967M is created. Applying a hash function (here: MD5), gives the hash ab76990b084b82d3e06701c52d02485e8e2ba9fe, which is used for exact linking.

3.1.2 Bloom filters

First suggested by Bloom (1970), Bloom filters have been suggested for PPRL by Schnell, Bachteler, and Reiher (2009). Figure 3-1-2-1 gives an example of constructing Bloom filters for two first names. In this example, the bigrams of both names were mapped to Bloom filters with a length of l = 15 bits using k = 2 hash functions.

Figure 3-1-2-1:
Exemplary construction of two Bloom filters with a length of l = 15 bits using k = 2 hash functions.



A	∩	B
1	1	1
1	0	0
0	0	1
0	0	0
1	1	1
0	0	0
0	0	0
1	1	1
0	0	0
1	0	0
0	0	0
1	1	1
0	0	0
0	0	0
1	1	1
Σ7	Σ5	Σ6
$\|A\|$	$\|A \cap B\|$	$\|B\|$

SAHRA and SARAH both share 3 out of 4 bigrams (SA, AH and RA) and differ on one single bigram (HR and AR, respectively). The unencrypted Dice coefficient for the bigrams of both names is:

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

Here, 7 resp. 6 bits are set to one in the Bloom filters, while having 5 common bit positions. This way, the Dice coefficient can be estimated as:

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

This allows estimating the clear-text *n*-gram similarity using encrypted identifiers.

3.2 Hardening Bloom filters against attacks

Bloom filters have been subject to cryptographic attacks (Niedermeyer, Steinmetzer, Kroll, & Schnell, 2014), which is why several techniques to prevent re-identification ("hardening" methods) have been suggested. All known attacks on Bloom filters (BF) for PPRL are either frequency or pattern-based attacks. They require frequent Bloom filters or frequent co-occurrences of bigrams. Depending on the kind of attack, the lower limit on the number of required frequent patterns differs. Therefore, all hardening techniques will try to reduce the number of frequent patterns or co-occurrences.

3.3 Examples for hardening techniques for Bloom filters

Niedermeyer et al. (2014) recommended salting, where a stable plaintext value (such as year of birth) is used as an additional substring for the password used during encoding BFs. Mapping different identifiers (first name, last name, date of birth, sex) into the same Bloom filter will still permit PPRL. This construction is called a Cryptographic Long-term Key (CLK, Schnell, 2014) and offers stronger resilience against attacks. Of course, salting can be applied to CLKs, resulting in salted CLKs. Since the number of bits set to one in BFs is essential for many frequency attacks, Schnell and Borgs (2016) proposed the use of Balanced BF, where each salted BF of length $2*l$ has a hamming weight of l. No successful attacks on salted CLKs have been reported yet. Using additional hardenings such as balancing makes attacks even harder.

3.4 Evaluation of linkage methods

To compare PPRL methods with a baseline, we used probabilistic record linkage on unencrypted identifiers.

This baseline was compared to the widely used 581-key and two different versions of Bloom filter-based PPRL: A standard CLK and a salted CLK.

The standard CLKs used $k = 20$ hash functions with a length of $l = 1000$ bits, the salted CLKs with DOB as a salt used $k = 30$ hash functions.

For the first evaluation we use real-world data ($n = 14{,}003$ and $n = 909$; true overlap was $n = 889$). To test the influence of errors on linkage quality and the scalability of the approach, simulated data was used ($n = 1$ million and $n = 205{,}000$) with 0% – 20% of rows with errors in the identifiers (first and last names, sex and date of birth).

3.5 Similarity thresholds for Bloom filter-based methods

For linking the Bloom filter-based PPRL methods, we used Multibit trees. Multibit trees were suggested for chemometrics by Kristensen, Nielsen, and Pedersen (2010) and used for PPRL by Schnell (2014). Possible pairs below a pre-set similarity threshold are not evaluated. The Tanimoto similarity T is used as a similarity measure. T is defined as number of bits set to 1 in both vectors A and B divided by the total number of bits set to 1 in A and B:

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

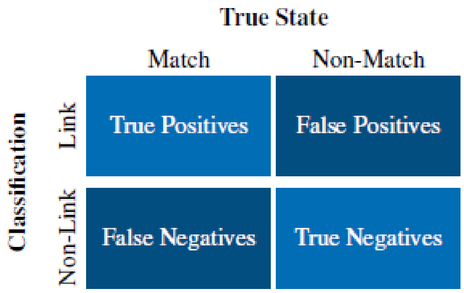
Lower similarity thresholds will result in more pair comparisons and a higher number of false positive classifications. Conversely, the number of true matches will increase as well. For the simulations, Tanimoto-thresholds were varied between 0.75 and 1.0 in steps of 0.05.

3.6 Evaluation metrics

All linking methods will classify record pairs as links or non-links. This classification is either correct or incorrect. Figure 3-6-1 shows the resulting classification matrix, which can give true positive (TP), false positive (FP) or false negative (FN) classifications.

Figure 3-6-1:

Classification matrix for evaluating classification results of linkage methods. True



		True State	
		Match	Non-Match
Classification	Link	True Positives	False Positives
	Non-Link	False Negatives	True Negatives

With these classifications, the widely used metrics for linkage quality, recall

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

and precision

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

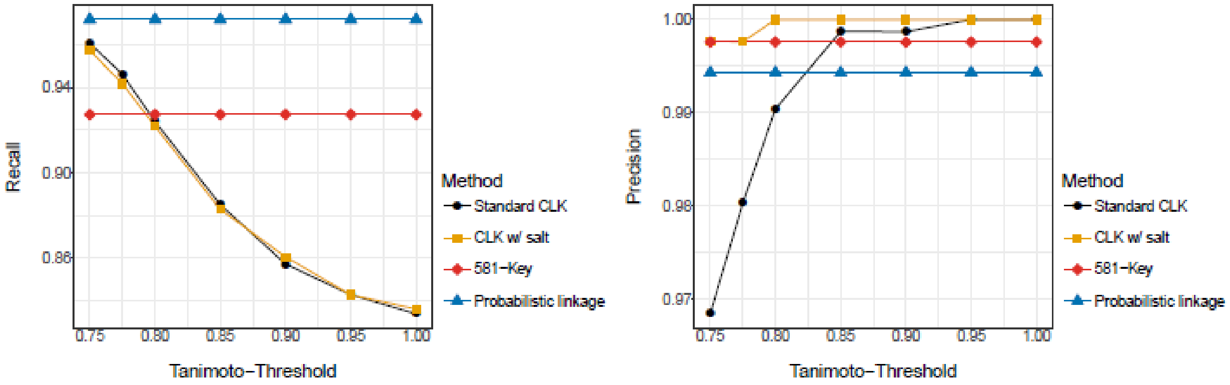
are calculated. We also report the arithmetic mean of precision and recall instead of the traditionally reported F-score, which has recently been under scrutiny (Hand & Christen, 2017).

4. Results

4.1 Evaluating linkage strategies using real-world data

For the evaluation using real-world administrative data, figure 4-1-1 shows precision and recall for all linkage strategies tested.

Figure 4-1-1:
Recall and precision by encryption method using several Tanimoto-thresholds.



All methods yielded high precision, which indicates a low amount of false positive matches. Recall increases with decreasing similarity thresholds for Bloom filter-based methods. At a threshold below 0.8, linkage quality of both Bloom filter-based methods exceeds the 581-key. As could be expected, unencrypted probabilistic record linkage is superior to all PPRL methods. Encrypting error-prone identifiers will lead to a decrease in linkage quality. To clarify the practical implications of the results, Table 4-1-1 shows the absolute numbers of true and false positives for each method.

The true overlap of the administrative data sets was 898 records. 97% of all true pairs were found by probabilistic linkage, with less than 0.006% false positives. 96% of all true pairs were found by salted CLKs, with less than 0.003% false positives. Therefore we lost a mere 13 cases (approx. 1.4%) due to PPRL, while addressing privacy concerns. The loss of quality using this PPRL method seems to be acceptable for most applications when sensitive information has to be linked.

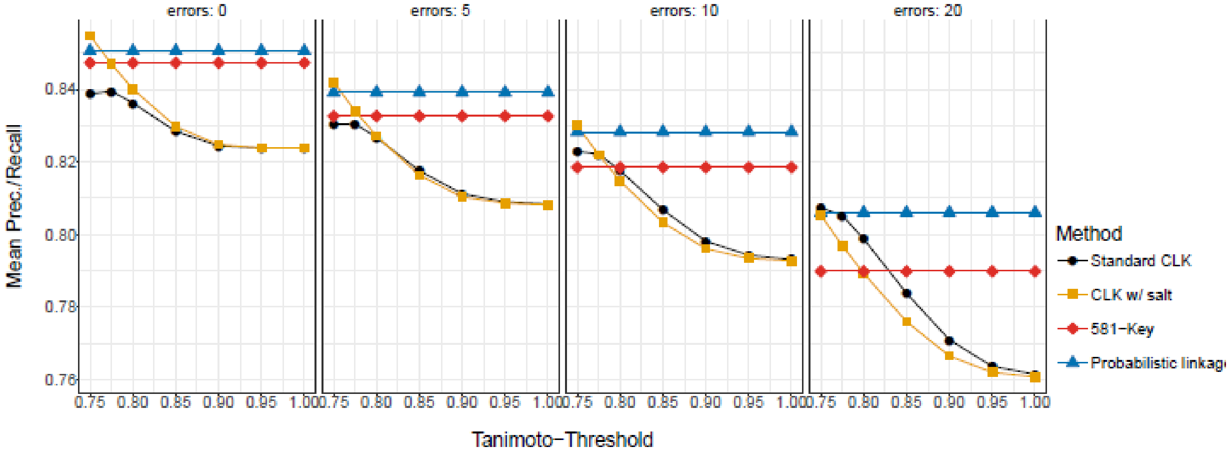
4.2 Evaluating linkage strategies using simulated large files with errors

To assess the effect of errors on linkage quality in files with a larger size, as encountered in national registries, simulated data was used. We simulated errors for 0% to up to 20% of rows of a 250,000 record subset of a 1 million record master file. To summarise linkage quality, the mean of precision and recall was computed. Figure 4-2-1 shows the results.

Table 4-1-1
True positives, false positives, resulting precision and recall by method at a Tanimoto-threshold of *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

Figure 4-2-1
Mean of recall and precision by encryption method and percent of rows with errors using several Tanimoto-thresholds.



Here too, probabilistic record linkage yields the best results. All methods show a steady decrease in linkage quality with increasing error rates of the identifiers. At the two lowest levels of the Tanimoto similarity, the relative performance of Bloom filter-based methods to the 581-Key increases with decreasing data quality. However, even with 20% errors in 250.000 identifiers, salted CLKs yield only 145 false positive matches (0.06%), while giving a recall very close to a probabilistic linkage. This quality level seems to be acceptable for most problems requiring PPRL. Of course, depending on the costs for false positives or false negatives in a given application, there might be exceptions.

5. Conclusion

Probabilistic linkage using clear-text identifiers outperforms all other methods. By comparison, encrypting error prone identifiers always reduced linkage quality. As in other comparisons before (Randall et al., 2016), CLKs with low similarity thresholds outperform 581-keys. For many applications, salted CLKs seem to yield only slightly inferior results than clear-text probabilistic linkage. If a clear-text national registry is no suitable option, salted CLKs might be the second-best solution. Therefore, we consider a national health register using only encrypted identifiers of patients as feasible. Nevertheless, implementing a corresponding protocol in practical settings will be challenging (Kho et al., 2015).

References

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.