Building a national perinatal database without the use of unique personal identifiers

Rainer Schnell
Centre for Comparative Social Surveys
City University London
London, United Kingdom
Email: Rainer.Schnell@city.ac.uk

Christian Borgs
German Record Linkage Center
University of Duisburg-Essen
Duisburg, Germany
Email: christian.borgs@uni-due.de

***Abstract*—To assess the quality of hospital care, national databases of standard medical procedures are common. A widely known example are national databases of births. If unique personal identification numbers are available (as in Scandinavian countries), the construction of such databases is trivial from a computational point of view. However, due to privacy legislation, such identifiers are not available in all countries. Given such constraints, the construction of a national perinatal database has to rely on other patient identifiers, such as names and dates of birth. These kind of identifiers are prone to errors. Furthermore, some jurisdictions require the encryption of personal identifiers. The resulting problem is therefore an example of *Privacy Preserving Record Linkage* (PPRL). This contribution describes the design considerations for a national perinatal database using data of about 600,000 births in about 1,000 hospitals. Based on simulations, recommendations for parameter settings of Bloom filter based PPRL are given for this real world application.**

I. Background

For the medical assessment of German hospitals, a federal institution (GBA)[1] is obliged by law to link administrative records of more than 600,000 births yearly. The records are scattered across about 1,000 independent perinatal and neonatal units. The linked data is used for monitoring hospital performance and epidemiological analyses like spatial prevalence of very low birth weights. Due to privacy regulations, patient databases of hospitals are not linked by an electronic network. The hospitals use different electronic medical record systems, but have to use the same data exchange format. All details of the data exchange are part of a mandatory regulation. Because the German health insurance system has no common unique personal identifier number, other patient identifiers have to be used. Since the current regulations do not allow names in any form, encrypted or not, current linkage is based on different combinations of health insurance numbers, birth weight and hospital identifiers. Given the described constraints, only about 80% of the records can be linked [1].

From a statistical point of view, non-linked records might cause a missing data problem [2]. If the fact, that a true link is missed, depends on variables of interest, this is referred to as *differential linkage error* [3], [4]. This might result in biased estimates of causal effects and population parameters [5], [6]. Concerning our field of application, evidence of bias caused by differences between linked and non-linked maternal data sets has been published [7], [8]. The easiest way to reduce differential linkage bias is improving the linkage rate. Therefore, using additional identifiers has been proposed to the regulatory authority [9]. As previous research has shown [10], using first and last names in combination with date of birth would give acceptable results. However, in many legal frameworks, names have to be encrypted before linkage across different agencies is allowed. Since identifiers as names are prone to errors [11], [12], standard cryptographic methods such as keyed HMACs (*Hash Message Authentication Code* [13], for example, SHA-256 or MD5) would result in missed links. This problem has created the field of *privacy preserving record linkage* (PPRL) [14].

For real-world medical applications, the most widely used methods today are variants of statistical linkage keys based on phonetic codes, such as Soundex or NYSIIS [15]. After the phonetic encoding, the resulting codes are usually combined with additional identifiers, such as date of birth and sex, which are finally encrypted with keyed HMACs. Examples for these kind of linkage keys are given by [16] and [17]. Statistical linkage keys usually produce acceptable results and seem to be safe against most cryptographic attacks.[2] However, previous research on the performance of statistical linkage keys [18] showed very few false positive links, but also high levels of missed links (larger number of false negative links).

Therefore, the search for better PPRL solutions in real-world settings is a very active field of research. During the last decade, an impressive list of PPRL methods have been suggested [14]. However, very few of these techniques are suitable for large scale linkage operations under the restrictions described above [19].

One such method using Bloom filters [20] for error-tolerant privacy preserving record linkage was proposed by [21]. This method has been used internationally in several different settings [22], [23], [24], [25] and is discussed extensively in recent publications concerning privacy preserving record linkage [26], [27], [28] [29], [30]. In this paper, we will describe the design of a national perinatal database using Bloom filters for PPRL.

[1] For details, see www.english.g-ba.de.

[2] We are not aware of any published attack against encrypted statistical linkage keys. However, since these codes result in unique bit patterns given a combination of names, date of birth and sex all statistical linkage keys are vulnerable for the most frequent names. An exploration of this problem is subject of ongoing research by our group.

A. Outline of the paper

In our application setting, all national record linkage operations in health research are based on the linkage protocol used for the cancer registries [31], [32]. Because local legal authorities consider this protocol as gold standard, as a first step the current protocol will be described before we suggest our modifications and extensions. The next section reports on a simulation using records with error types and error rates empirically observed in the intended field of application. The suggested procedure is tested against several well-established alternative procedures. Finally, we will discuss aspects of re-identification risks and problems in real-world implementation of the procedure.

II. Methods

A. The current German standard for linking registries

Information is classified into three groups:

1) sensitive identifiers (including first and last name, date of birth, date of diagnosis and date of death),
2) epidemiological data (sex, month and year of birth, residential code, diagnosis),
3) linkage identifiers (a set of 22 fields of standardized[3] identifiers and phonetically encoded identifiers).

Sensitive identifiers are encrypted by RSA [33], followed by IDEA encryption [34]. Epidemiological data is stored unencrypted. Linkage identifiers are encrypted by MD5 followed by IDEA encryption. The IDEA encryption is removed during the linkage process. The details are described by [31], [32]. Due to the federal organisation of registries, the number of records in a registry is noticeable below the number of records in a national registry. Cases are usually not linked across different regional registries, therefore the number of cases within a block (for example: a given day of birth) is small. To the best of our knowledge, the resulting record linkage process has never been evaluated independently using real data with known links.

B. The new proposal

The procedure described above was designed 20 years ago. The only protection against errors is due to the phonetic encoding (by a Soundex variant) of some linkage identifiers. For the reasons given above, there is a demand for a revised procedure for a number of different medical registries (National Death Index, Neonatal Registry). However, in the case of the neonatal registry considered here, neither names of children nor mothers are currently available for linkage. Due to this lack of information and the considerable amount of missing data in other identifiers (see table I), currently about 20% of records can not be linked, mostly due to the lack of discriminating information for very similar records [1]. Therefore, the inclusion of first names and surnames of mothers has been suggested [9]. Due to the local privacy legislation, names have to be encrypted. The form of the encryption is not regulated by law. In practice, either phonetic codes or Bloom filter encodings might be used for the encryption. For efficient blocking and linkage, keyed HMACs for all numerical identifiers (hospital ID, state, sex, multifetal sequence number, date of birth and hour of birth) might be considered. To summarize the proposal:

1) inclusion of first and last name of the mother,
2) encrypted phonetic codes and/or Bloom filter encoding of names,
3) keyed HMACs for numerical identifiers.

A simulation study was conducted to clarify if the inclusion of additional identifiers would resolve the problem of insufficient discriminating power between similar records and if Bloom filters or phonetic codes would yield better results.

C. Simulation study

1) Databases: The perinatal registry is the result of the linkage of a neonatal database with about 100,000 records and a perinatal database with about 600,000 records. Each database is the result of independent data deliveries due to separate mandatory requirements. Both databases may contain different hospital episodes of the same child, therefore there may be valid multiple records for identical patients. We simulated two corresponding data files of the mentioned size. The perinatal data contains about 95% of the records included in the neonatal data. To account for this reduced overlap, 5% of the neonatal records were replaced with additional simulated records. To ensure the same real-world marginal distributions for all identifiers of each data set, we used marginal distributions of the real-world data set provided on request by the institute commissioned for the national linkage of this data.[4]

2) Simulated identifiers: The number of births per hospital was simulated according to the empirical distribution, resulting in a hospital ID per simulated record. The federal state of residence of the mother was simulated using the marginal distribution of residents according to Official Statistics. Date (day, month, year) and time (hour and minute) of birth, sex and multifetal sequence number[5] of the child were generated according to the empirical distribution in the neonatal database. Last names were sampled proportional to their counts in the national social security database. We used the same administrative database for the first names. Sampling of first names was stratified by maternal year of birth to reflect changing cohort preferences in name selection. With this exception, all other identifiers were simulated independently.

3) Missing identifiers: Missing identifiers were generated according to the observed distribution (see table I). Since the documentation of the date of birth is currently not legally required in all states, it is missing in about 23% of the cases. Because the neonatal database does currently not contain any names, we used the proportions of missing names supplied by the data generation module of Febrl [35][6]. The local legal framework on neonatal data is currently under revision. One of the most likely results of this revision will be a reduction in the amount of missing identifers. Therefore, additional data sets with reduced proportions (50% of the currently observed proportion) of missing identifiers were created. To study the

[3] The standardization includes reformatting values, removal of special characters, within-word punctuation and professional titles.

[4] https://www.aqua-institut.de/en/home/

[5] The multifetal sequence number is always one for singleton births, while multifetal gestation usually requires a numbering according to sequence of births.

[6] Febrl can be used to generate and link data and is freely available from http://sourceforge.net/projects/febrl/.

TABLE I. PROPORTION OF SIMULATED MISSING IDENTIFIERS

Identifier	Missing (%)
First/Last Names (Mother)	2
Hospital ID	0
Multifetal gestation	0
Multifetal sequence number	0
Sex of the child	0
Date of birth	23.41
Time (Hour/Minute) of birth	0.43
Birth weight	$1.3{\cdot}10^{-5}$

effects of varying proportions of missing identifiers, additional data sets assuming an increase (200% of the currently observed proportions) were generated.

4) Errors in identifiers: Different types of errors (for deletions, insertions and character swapping) were applied to simulated first as well as last names. Error probabilities were taken from Febrl. Error levels were set in 5% steps between 0% to 25%.

D. PPRL methods

To test the newly proposed protocol, different PPRL methods were used.

1) Anonymous Linking Codes: Previous work [18] suggested at least two widely used and suitable encryption methods for privacy preserving record linkage in the given context: The Swiss Anonymous Linking Code (*Swiss ALC* [16]) and the Encrypted Statistical Linkage Key (*Australian ALC* [17]). The Swiss ALC concatenates the Soundex codes of first and last name, sex and date of birth. The resulting string is encrypted with MD5. The Australian ALC uses a string consisting of the second and third character of the first name and the second, third and fifth character of the last name, sex and date of birth. Again, the resulting string is encrypted with MD5.

2) Bloom filters combined with exact encodings of numerical identifiers: The newly suggested protocol (denoted as *GBA*) uses single field Bloom filter encryptions for the name fields, as described in [21]. First and last names were each mapped to separate Bloom filters with a length of $l = 1000$ bits using $k = 10$ (GBA k=10), $k = 20$ (GBA k=20) and $k = 30$ (GBA k=30) hash functions for mapping bigrams to Bloom filters. Testing for an optimal number of hash functions is recommended, since the linkage quality is directly affected by this choice. Automatic choice of optimal parameters is subject of ongoing research.

All other identifiers (hospital ID, state, sex, multifetal sequence number, date of birth, hour of birth) were initially coded numerically and then encrypted with a keyed MD5 hash function.[7]

3) Cryptographic Long-term Keys: The final type of encryption methods are combinations of separate Bloom filters resulting in a single bit array called Cryptographic Long-term Keys (*CLKs* [36]). For this simulation, CLKs were built using the first and last name, sex, multifetal sequence number, date of birth and hour of birth.The number of hash functions was $k = 10$ (CLK k=10), $k = 20$ (CLK k=20) and $k = 30$ (CLK k=30) for each identifier. For names, bigrams were used for hashing, for all other identifiers, unigrams were used. Finally, another CLK version (CLK + birthweight) additionally using the unigrams of birth weight (with $k = 10$) was tested.

E. Linkage methods

The two ALC variants were linked using exact matching, as MD5-encoded strings are ill-suited for probabilistic linkage techniques. All GBA variants were matched using probabilistic linking, blocked with a combination of the MD5 codes of a hospital ID, state, sex, multifetal sequence number, date of birth and hour of birth.

The CLK variants were linked using Multibit trees [37]. Initially developed in chemoinformatics, they have been proposed for PPRL using bit vectors by [38] for the classification of matches and non-matches using a given similarity threshold. Simulations have shown that Multibit Trees are among the fastest methods for finding nearest neighbours in data sets consisting of millions of bit vectors [39]. Multibit trees are built by finding a bit position in a bit array where approximately half of the records have a value set to one for this bit position and the other half of the records are set to zero for this exact bit position. These halved records form *leaves*, which are then split again. This process is repeated (here: 8 times) for all leaves. The resulting data structure is the Multibit tree. This tree is queried by a second set of bit arrays. For each query array, a minimum and maximum bound for a similarity coefficient can now be calculated for each leaf in the Multibit tree. This reduces the amount of pairs considered for similarity calculations dramatically. As suggested by [37], similarity is assessed by the Tanimoto coefficient T [40]:

$$T(A, B) = \frac{\Sigma_i(A_i \wedge B_i)}{\Sigma_i(A_i \vee B_i)} \qquad (1)$$

T is 1 for exact matching bit arrays and 0 for bit arrays where every bit position differs. For the use of Multibit trees in PPRL a threshold for T has to be selected. This threshold controls the number of candidate pairs of a leaf to be considered for similarity calculation. A lower threshold considers less similar pairs. Therefore, a threshold of 1.0 will only consider exact matches, while lower thresholds allow for more errors. Lower thresholds will obviously result in a higher number of false positive classifications and increase the computing time. For the simulations, Tanimoto thresholds were varied between 0.8 and 1.0 in steps of 0.05.

Simulations were done using R 3.2.0 [41] on a machine with 64 GB RAM, a six-core Intel i7-4930K CPU with 3.4 GHz, running Ubuntu 12.04.

III. ANALYSIS AND RESULTS

To assess linkage quality, the standard record linkage criteria (precision, recall and F-score) were used.

[7] Although MD5 was used in the simulations, replacing MD5 with a different HMAC will have no impact on precision and recall. Different hash functions will generate different bit patterns, but the distribution of hash values should be uniform regardless which hash function is used. Since the linkage depends only on the overall number of common bits, not on local similarities or pattern similarities, the choice of the HMAC is uncritical with regard to precision and recall.

A. Linkage quality measures

Precision is defined as the number of correctly classified pairs (true positive classifications tp) divided by the number of all classified pairs (tp and false positives fp):

$$\text{Precision} = \frac{tp}{tp + fp} \tag{2}$$

Recall is defined as the number of true positive matches divided by the number of factual pairs, including pairs falsely classified as non-matches (false negatives fn) by the linkage algorithm:

$$\text{Recall} = \frac{tp}{tp + fn} \tag{3}$$

Finally, F-score is defined as harmonic mean of recall and precision:

$$\text{F-score} = \frac{2 \cdot \text{Recall} \cdot \text{Precision}}{\text{Recall} + \text{Precision}} \tag{4}$$

B. Results

Figure 1 shows the performance of all procedures with regard to F-scores. Since the Tanimoto threshold is only relevant for the CLK variants, all other variants are shown as constants. The GBA approach proposed in this paper is superior to all alternative approaches tested here. GBA variants seem to be resilient against errors in identifiers and missing identifiers, whereas most other methods perform poorly under these circumstances.

Precision for each linkage procedure considered here is shown in figure 2. Obviously, nearly all methods have a very high precision, whereas the CLKs seem to be sensitive to errors and Tanimoto thresholds below 0.95. A high precision implies a low false positive rate.[8]

The recall shown in figure 2 is very low for the ALC methods, especially in a setting with many errors and missing identifiers. Regarding recall, CLKs perform increasingly better with lower Tanimoto thresholds, since more pairs are considered for the similarity calculation. However, even the best CLK variants are consistently outperformed by the GBA approach suggested here. The recall is close or equal to 1.0 for all scenarios except for the scenario with high amounts of missing identifiers. Under these conditions, GBA drops slightly below 1.0, but still outperforms all other procedures by far.

IV. Discussion

Using F-scores as criterion, the simulations reported above demonstrated the superior performance of Bloom filter variants compared to anonymous linkage keys in nearly all simulated scenarios. Exceptions are conditions with a doubled proportion of missing identifiers as currently observed and – at the same time – higher error rates in names. Using a low similarity threshold of 0.85 for the Multibit tree linkage as suggested by [19], even in this (unlikely) worst case scenario the best CLK variant is still slightly superior to the Australian Statistical Linkage Key (AUS ALC). In general, variants of CLKs with either $k = 20$ or $k = 10$ hash functions show the best performance.

A. Security of Bloom filter-based approaches

Two partially successful attacks on Bloom filters using a Constraint Satisfaction Solver have been reported by [42], [43]. In the more recent study, the authors used combined Bloom filters (a CLK variant). After reducing the problem to the most common 20 surnames and using considerable computing time, they reported the identification of 4 of the 20 most frequent names. Therefore, they concluded that the combination of independent Bloom filters seem to withstand cryptographic attacks.

However, their attack was purely computational and not based on the specific properties of the encryption method. In contrast to this, an analytical cryptanalysis has been reported by [44]. The initial construction of Bloom filters was based on the use of only two independent HMACs [45]. This *double hashing scheme* reduces the number of possible encodings. Using this idea, [44] demonstrated a manual attack on double hashing-based Bloom filters resulting in a near complete re-identification of all records. This kind of attack has been used by [46] for attacking double hashing based combined Bloom filter (CLKs).

Because of these unsatisfactory cryptographic properties of linear combinations of HMACS [44], the use of k randomly selected values (for example, by using seeded random number streams) for each n-gram is currently recommended practice [44]. This modification makes the described analytical attack nearly impossible. By the use of randomly selected hash values in combined Bloom filters (CLKs) in conjunction with additional hardenings like salting [19], all cryptographic attacks on Bloom filters known so far will fail. Consequentially, no successful attack on hardened Bloom filters has been reported up to now.

For the application considered here, the separate Bloom filters should be salted with blocking variables. Hereby, the number of records with common encodings can be reduced below the number necessary for a frequency attack.[9]

[8] The false positive rate is the ratio of false positives to the sum of true negatives and false positives ($FPrate = \frac{fp}{fp+tn}$). The false positive rate is zero for all scenarios and methods except for the CLKs. Here, false positive rates for a CLK with $k = 20$ hash functions under the assumption of 10% errors in the files using a Tanimoto threshold of 0.9 gives a false positive rate ranging from 0.065 (missings 50% of observed) to 0.342 (missings 200% of observed). Note that the false positive rate depends on the Tanimoto threshold used and rises sharply with an increase in missing values. If the application demands very low false positive rates, the set of personal identifiers used for building CLKs should be extended. This effect can be seen in figure 2, where the precision of the CLK plus birth weight is always higher than all other CLKs.

[9] Using this strategy, the re-identification risk by any frequency based attack will be greatly reduced since only few cases share the same encoding. Using Date of Birth (365) and Sex (2) will yield about $600.000/(365 * 2) \approx 822$ births using the same encoding. This set has to be used for estimating the frequency of the most frequent bigrams. Due to the low number of cases available, the resulting estimated frequencies will have large large standard errors and re-identification will be more difficult.

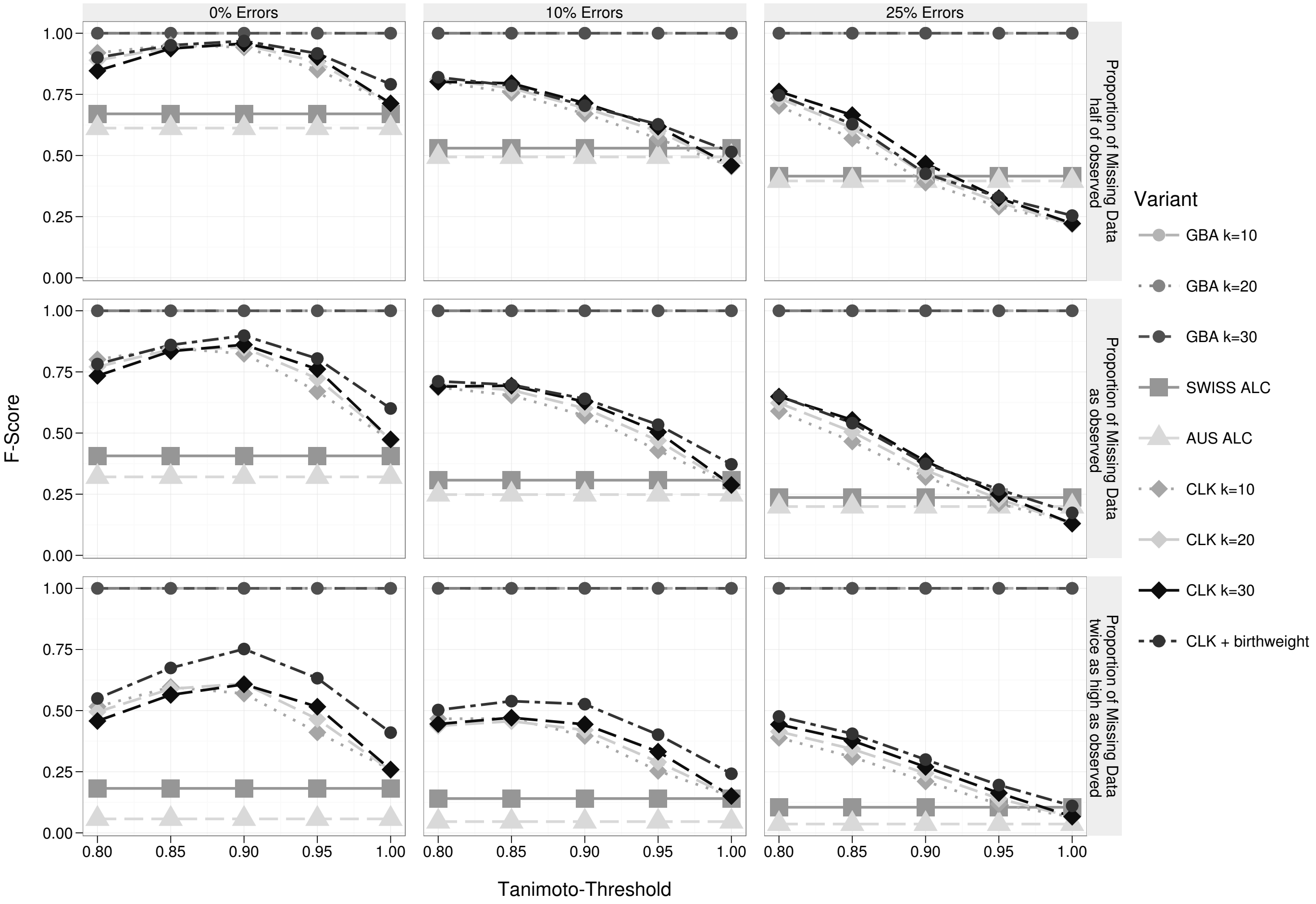

Fig. 1. F-scores of different linkage procedures depending on proportion of missing identifiers and errors in identifiers. Because the Tanimoto threshold is only relevant for CLKs using Multibit trees, all other encryption methods are shown as constants. Note that the differences between the GBA variants are too small to be visible.

B. Implementation in the real-world

Implementation of the suggested procedure in about 1,000 hospitals using about 30 different hospital information systems will need some additional efforts. For example, organisational safeguards regarding the key management are required [9]. To prevent building a database of all newborn over time, the keys for generating the encryptions have to be changed yearly. However, because the hospitals are not linked electronically, data delivery will be delayed, sometimes for months. To link all records of two consecutive years despite these delays will require the separate encoding of each record both by the current key and the key of the previous year.

V. Conclusion

The findings reported here demonstrate that it is possible to achieve good results with privacy preserving record linkage even under very strict privacy jurisdictions. The proposed inclusion of additional personal identifiers to the set of characteristics to be encrypted will yield sufficient discriminating power between very similar cases, resulting in only a small number of false positive links. A further decrease of false positive links can be achieved by including additional identifiers. In the context of neonatal registries, obvious additional identifiers are birth weight and place of birth of the mother. If organisational and legal constraints permit the use of additional identifiers, their inclusion will be useful. Using the recommended parameters for the encoding procedure and similarity thresholds will find most true links despite missing or misspelled names.

However, the simulations have shown that the performance of Bloom filter-based PPRL strongly depends on carefully chosen parameters. If different identifiers or different distributions of identifiers have to be used, fine tuning parameters (especially for the CLKs) might be necessary. For example, a population with a different mixture of ethnic names will certainly need different pre-processing routines and different parameter choices. To facilitate further applications, we are investigating the automatic choice of optimal parameters for Bloom filter-based PPRL.

Acknowledgements

We would like to thank the three anonymous reviewers for their helpful comments.

The authors are indebted to Sabrina Torregroza and Günther Heller for comments on an earlier draft. This study was supported by the research grant SCHN 586/17-1 of the German Research Foundation (DFG) to the first author.

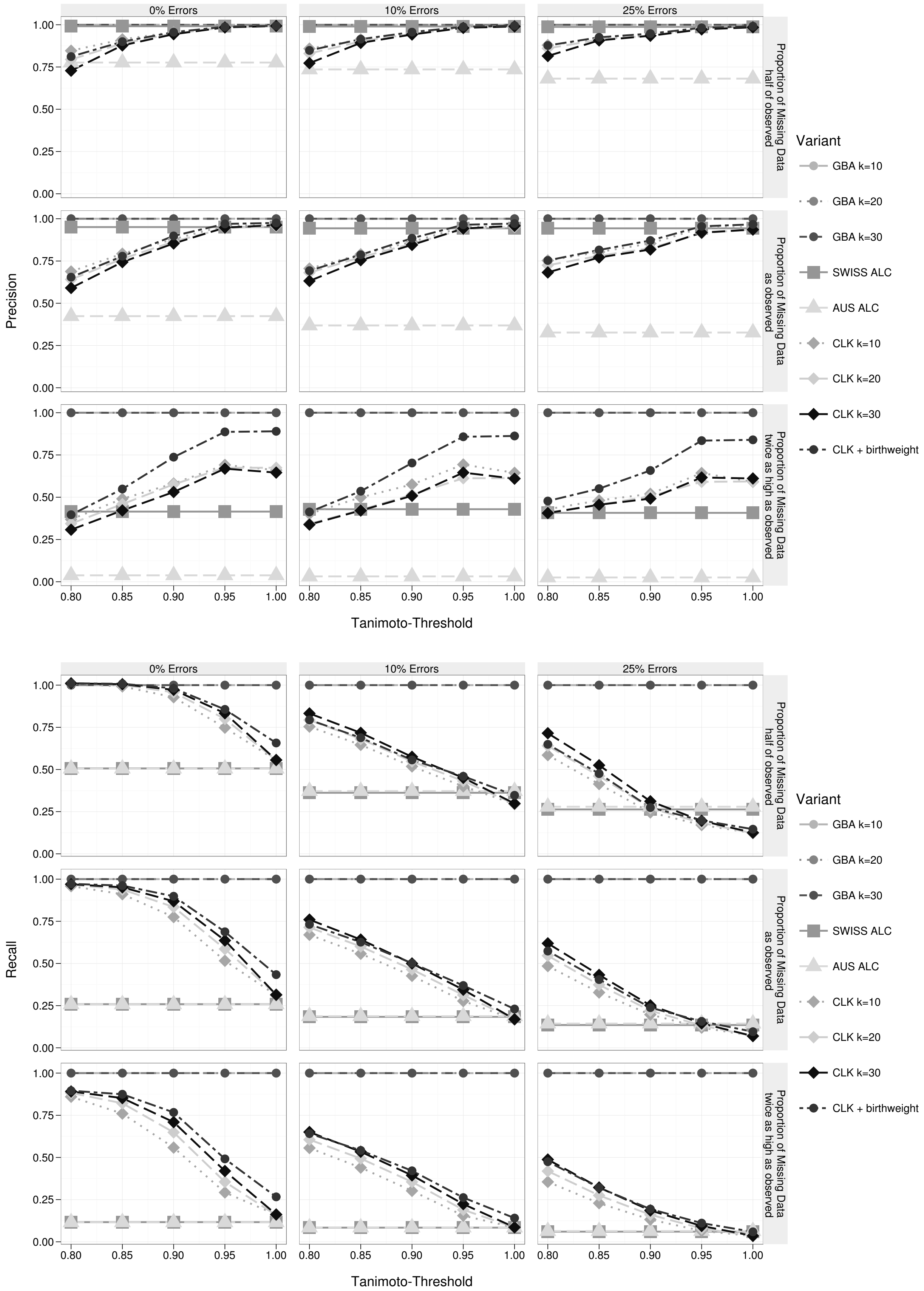

Fig. 2. Precision and recall of different linkage procedures depending on proportion of missing identifiers and errors in identifiers. Because the Tanimoto threshold is only relevant for CLKs using Multibit trees, all other encryption methods are shown as constants. Note that the differences between the GBA variants are too small to be visible.

REFERENCES

[1] T. Bachteler and J. Reiher, "Verknüpfung der Geburtshilfe- und Neonatalogiedaten des Jahres 2011 im Auftrag des AQUA-Instituts," German Record Linkage Center, Tech. Rep., 2012.

[2] X.-H. Zhou, C. Zhou, D. Liu, and X. Ding, *Applied Missing Data Analysis in the Health Sciences*. Hoboken: Wiley, 2014.

[3] J. K. Leiss, D. Giles, K. M. Sullivan, R. Mathews, G. Sentelle, and K. M. Tomashek, "U.S. maternally linked birth records may be biased for Hispanics and other population groups," *Annals of Epidemiology*, vol. 20, no. 1, pp. 23–31, Jan. 2010.

[4] J. T. Lariscy, "Differential record linkage by Hispanic ethnicity and age in linked mortality studies: implications for the epidemiologic paradox," *Journal of Aging and Health*, vol. 23, no. 8, pp. 1263–1284, Dec. 2011.

[5] I. Baldi, A. Ponti, R. Zanetti, G. Ciccone, F. Merletti, and D. Gregori, "The impact of record-linkage bias in the Cox model," *Journal of Evaluation in Clinical Practice*, vol. 16, no. 1, pp. 92–96, 2010.

[6] K. Harron, A. Wade, R. Gilbert, B. Muller-Pebody, and H. Goldstein, "Evaluating bias due to data linkage error in electronic healthcare records," *BMC Medical Research Methodology*, vol. 14, no. 1, p. 36, 2014.

[7] J. B. Ford, C. L. Roberts, and L. K. Taylor, "Characteristics of unmatched maternal and baby records in linked birth records and hospital discharge data," *Paediatric and Perinatal Epidemiology*, vol. 20, no. 4, pp. 329–337, 2006.

[8] J. P. Bentley, J. B. Ford, L. K. Taylor, K. A. Irvine, and C. L. Roberts, "Investigating linkage rates among probabilistically linked birth and hospitalization records," *BMC Medical Research Methodology*, vol. 12, p. 149, 2012.

[9] R. Schnell, C. Borgs, G. Heller, and F. Niedermeyer, "Pseudonymisierte Verknüpfung der Perinatal- und Neonatalerhebung mit Bloom-Filtern," German Record Linkage Center, Duisburg, unpublished Technical Report, 2015.

[10] C. Quantin, C. Binquet, K. Bourquard, R. Pattisina, B. Gouyon-Cornet, C. Ferdynus, J.-B. Gouyon, and F.-A. Allaert, "Which are the best identifiers for record linkage?" *Medical Informatics and the Internet in Medicine*, vol. 29, no. 3–4, pp. 221–227, 2004.

[11] W. E. Winkler, "Record linkage," in *Handbook of Statistics 29A, Sample Surveys: Design, Methods and Applications*, D. Pfeffermann and C. Rao, Eds. Amsterdam: Elsevier, North-Holland, 2009, pp. 351–380.

[12] M. A. Hernandez and S. S. Stolfo, "Real-world data is dirty: data cleansing and the merge/purge problem," *Data Mining and Knowledge Discovery*, vol. 2, no. 1, pp. 9–37, 1998.

[13] W. Stallings, *Cryptography and Network Security: Principles and Practice*, 6th ed. New Jersey: Pearson, 2014.

[14] D. Vatsalan, P. Christen, and V. S. Verykios, "A taxonomy of privacy-preserving record linkage techniques," *Information Systems*, vol. 38, no. 6, pp. 946–969, 2013.

[15] T. N. Herzog, F. J. Scheuren, and W. E. Winkler, *Data Quality and Record Linkage Techniques*. New York: Springer, 2007.

[16] F. Borst, F.-A. Allaert, and C. Quantin, "The Swiss solution for anonymous chaining patient files," in *Proceedings of the 10th World Congress on Medical Informatics: 2–5 September 2001; London*, V. Patel, R. Rogers, and R. Haux, Eds. Amsterdam: IOS Press, 2001, pp. 1239–1241.

[17] R. Karmel, P. Anderson, D. Gibson, A. Peut, S. Duckett, and Y. Wells, "Empirical aspects of record linkage across multiple data sets using statistical linkage keys: the experience of the PIAC cohort study," *BMC Health Services Research*, vol. 10, no. 41, 2010.

[18] R. Schnell, A. Richter, and C. Borgs, "Performance of different methods for privacy preserving record linkage with large scale medical data sets," University of Duisburg-Essen, 2014.

[19] R. Schnell, "Privacy preserving record linkage," in *Methodological Developments in Data Linkage*, K. Harron, H. Goldstein, and C. Dibben, Eds. Wiley, in print.

[20] B. H. Bloom, "Space/time trade-offs in hash coding with allowable errors," *Communications of the ACM*, vol. 13, no. 7, pp. 422–426, 1970.

[21] R. Schnell, T. Bachteler, and J. Reiher, "Privacy-preserving record linkage using Bloom filters," *BMC Medical Informatics and Decision Making*, vol. 9, no. 41, 2009.

[22] E. Durham, Y. Xue, M. Kantarcioglu, and B. Malin, "Private medical record linkage with approximate matching," in *Proceedings of the 2010 American Medical Informatics Association Annual Symposium*, 2010, pp. 182–186.

[23] C. E. Kühni, C. S. Rueegg, G. Michel, C. E. Rebholz, M.-P. F. Strippoli, F. K. Niggli, M. Egger, and N. X. von der Weid, "Cohort profile: the swiss childhood cancer survivor study," *International Journal of Epidemiology*, pp. 1–12, 2011.

[24] L. M. P. Santos, F. Guanais, D. L. Porto, O. L. de Morais Neto, A. Stevens, J. J. C. Escalante, L. B. de Oliveira, and L. Modesto, "Peso Ao Nascer Entre Crianas De Famílias De Baixa Renda Benefici árias E Não Beneficiárias Do Programa Bolsa Família Da Região Nordeste (Brasil): Pareamento Entre Cadúnico E Sinasc," in *Saúde Brasil 2010: Uma Análise Da Situaão De Saúde E De Evidências Selecionadas De Impacto De Aes De Vigilância Em Saúde*, Ministério da Saúde, Ed. Brasilia: Ministério da Saúde, 2011, pp. 271–293.

[25] S. M. Randall, A. M. Ferrante, J. H. Boyd, J. K. Bauer, and J. B. Semmens, "Privacy-preserving record linkage on large real world datasets," *Journal of Biomedical Informatics*, Dec. 2013.

[26] E. A. Durham, "A framework for accurate, efficient private record linkage," Dissertation. Vanderbilt University, 2012.

[27] E. Durham, Y. Xue, M. Kantarcioglu, and B. Malin, "Quantifying the correctness, computational complexity, and security of privacy-preserving string comparators for record linkage," *Information Fusion*, vol. 13, no. 4, pp. 245–259, 2012.

[28] D. Vatsalan, "Scalable and approximate privacy-preserving record linkage," Ph.D. dissertation, Australian National University, 2014.

[29] P. Christen, *Data Matching: Concepts and Techniques for Record Linkage, Entity Resolution, and Duplicate Detection*. Berlin: Springer, 2012.

[30] V. S. Verykios and P. Christen, "Privacy-preserving record linkage," *WIREs Data Mining and Knowledge Discovery*, vol. 3, no. 5, pp. 321–332, 2013.

[31] H. J. Appelrath, J. Michaelis, I. Schmidtmann, and W. Thoben, "Empfehlung an die Bundesländer zur technischen Umsetzung der Verfahrensweisen gemäß Gesetz über Krebsregister (KRG)," http://www.krebsregister-niedersachsen.de/registerstelle/dateien/veroeffentlichungen/Paper/Empfehlungen.html, 1996.

[32] K. Hentschel and A. Katalinic, "Das Krebsregister-Manual der Gesellschaft der epidemiologischen Krebsregister in Deutschland e.V." http://www.ekr.med.uni-erlangen.de/GEKID/Doc/Krebsregister-Manual%202008.pdf, 2008.

[33] R. L. Rivest, A. Shamir, and L. Adleman, "A method for obtaining digital signatures and public-key cryptosystems," *Communications of the ACM*, vol. 21, no. 2, pp. 120–126, 1978.

[34] X. Lai and J. L. Massey, "A proposal for a new block encryption standard," in *Advances in Cryptology - EUROCRYPT '90*. Springer-Verlag, 1991, pp. 389–404.

[35] P. Christen, "Febrl: a freely available record linkage system with a graphical user interface," in *HDKM '08 Proceedings of the second Australasian workshop on Health data and knowledge management*. Darlinghurst, Australia: Australian Computer Society, 2008, pp. 17–25.

[36] R. Schnell, T. Bachteler, and J. Reiher, "A novel error-tolerant anonymous linking code," German Record Linkage Center, Duisburg, Working Paper WP-GRLC-2011-02, 2011.

[37] T. G. Kristensen, J. Nielsen, and C. N. Pedersen, "A tree-based method for the rapid screening of chemical fingerprints," *Algorithms for Molecular Biology*, vol. 5, no. 1, pp. 9–20, 2010.

[38] T. Bachteler, J. Reiher, and R. Schnell, "Similarity filtering with multibit trees for record linkage," German Record Linkage Center, Nuremberg, Working Paper WP-GRLC-2013-02, 2013.

[39] R. Schnell, "An efficient privacy-preserving record linkage technique for administrative data and censuses," *Journal of the International Association for Official Statistics*, vol. 30, no. 3, pp. 263–270, 2014.

[40] J. M. Hancock and M. J. Zvelebil, Eds., *Concise Encyclopaedia of Bioinformatics and Computational Biology*, 2nd ed. Chichester: Wiley, 2014.

[41] R Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2015. [Online]. Available: http://www.R-project.org/

[42] M. Kuzu, M. Kantarcioglu, E. Durham, and B. Malin, "A constraint satisfaction cryptanalysis of Bloom filters in private record linkage," in *Privacy Enhancing Technologies, 11th International Symposium, PETS 2011*. Waterloo, Canada: Springer Berlin Heidelberg, 2011, pp. 226–245.

[43] M. Kuzu, M. Kantarcioglu, E. A. Durham, C. Toth, and B. Malin, "A practical approach to achieve private medical record linkage in light of public resources," *Journal of the American Medical Informatics Association*, vol. 20, no. 2, pp. 285–292, 2013.

[44] F. Niedermeyer, S. Steinmetzer, M. Kroll, and R. Schnell, "Cryptanalysis of basic bloom filters used for privacy preserving record linkage," *Journal of Privacy and Confidentiality*, vol. 6, no. 2, pp. 59–79, 2014.

[45] A. Kirsch and M. Mitzenmacher, "Less hashing same performance: building a better Bloom filter," in *Algorithms-ESA 2006. Proceedings of the 14th Annual European Symposium: 11-13 September 2006; Zürich, Switzerland*, Y. Azar and T. Erlebach, Eds. Berlin: Springer, 2006, pp. 456–467.

[46] M. Kroll and S. Steinmetzer, "Automated cryptanalysis of bloom filter encryptions of health records," in *8th International Conference on Health Informatics*, 2015.