International Journal of Population Data Science

Journal Website: www.ijpds.org

Microsimulation of an educational attainment register to predict future record linkage quality

Rainer Schnell[1,*] and Severin Weiand[1]

Submission History

Submitted:	18/12/2022
Accepted:	19/02/2023
Published:	03/04/2023

[1]Research Methodology Group, University of Duisburg-Essen, 47057 Duisburg, Germany

Abstract

Introduction
Population wide educational attainment registers are necessary for educational planning and research. Regular linking of databases is needed to build and update such a register. Without availability of unique national identification numbers, record linkage must be based on quasi-identifiers such as name, date of birth and sex. However, the data protection principle of data minimization aims to minimize the set of identifiers in databases.

Objectives
Therefore, the German Federal Ministry of Research and Education commissioned a study to inform legislation on the minimum set of identifiers required for a national educational register.

Methods
To justify our recommendations empirically, we implemented a microsimulation of about 20 million people. The simulated register accumulates changes and errors in identifiers due to migration, regional mobility, marriage, school career and mortality, thereby allowing the study of errors on longitudinal datasets. Updated records were linked yearly to the simulated register using several linkage methods. Clear-text methods as well as privacy-preserving (PPRL) methods were compared.

Results
The results indicate linkage bias if only the primary identifiers are available in the register. More detailed identifiers, including place of birth, are required to minimize linkage bias. The amount of information available to identify a person for matching is more critical for linkage quality than the record linkage method applied. Differences in linkage quality between the best procedures (probabilistic linkage and multiple matchkeys) are minor.

Conclusions
Microsimulation is a valuable tool for designing record linkage procedures. By modelling the processes resulting in changes or errors in quasi-identifiers, predicting data quality to be expected after the implementation of a register seems possible.

Keywords
simulation; register; errors; record linkage; identifiers; longitudinal studies

*Corresponding Author:
Email Address: rainer.schnell@uni-due.de (Rainer Schnell)

https://doi.org/10.23889/ijpds.v8i1.2122

Introduction

Evaluation of educational policies is facilitated by population wide educational attainment registers [20]. These kinds of data sets foster research as the example of the UK Longitudinal Educational Outcomes study has shown [36]. Building an educational register requires regular linking of updates on educational attainment to the register. If no unique national identifier is available, record linkage (RL) must be based on quasi-identifiers (QIDs) such as first name and date of birth [8, p. 8]. For linkage, the set of QIDs should be large to yield unique identification of a person despite occasional missing data. Furthermore, the probability of at least a few exact duplicates (the same identifiers for different persons) increases in population wide datasets.

QIDs are personally identifiable information. Their usage is usually restricted by laws such as the European General Data Protection Regulation (GDPR). Most regulations aim to replace personally identifiable data (pseudonymization) or their deletion (anonymization). Since the combination of indirect identifiers such as place of birth, occupation and place of residence can be used for unlawful reidentification, avoiding additional indirect identifiers in databases ('data minimization') is standard practice in data protection. Therefore, the set of QIDs should be as large as needed for linkage and as small as possible for data protection.

Because of this, the German Federal Ministry for Education and Research (BMBF) commissioned a study to inform legislation on the minimum identifiers for a national register. We considered the first name, surname, sex and full date of birth as available QIDs for an educational register without debate. However, a linkage based on these identifiers might not be sufficient to achieve acceptable linkage quality in a national register.

In large datasets, exact duplicates of identifiers denoting different persons will occur. Furthermore, due to data entry problems identifiers will be missing for some records. For these records, linking will be challenging. How many records in the database will be affected depends on the data quality of the processes generating the database. Due to the decentralized federal structure of Germany population wide datasets are rare, and nothing on their data quality has been published.

Due to the lack of information necessary for a rational decision on the set of required QIDs we used a microsimulation to study the effect of additional identifiers on record linkage quality.

Implementation of the microsimulation

In demographical applications, microsimulation models simulate social processes at the level of persons. In most cases, transitions between states of interest are modelled with conditional probabilities given a set of covariates. Due to the complexity of microsimulations in general [23] only the main modules and design decisions can be presented here. More details are discussed in two technical reports [30, 35].

Four types of individuals were simulated: schoolchildren, students, apprentices and persons who have finished primary, secondary or tertiary education. The simulated educational system uses one school type at the primary level and different school types at the secondary level. University education is available after secondary education and modelled in detail. To account for migration different procedures for entering the educational system were implemented for autochthones and migrants.

In the German dual vocational training system, trainees acquire practical knowledge in a company and theoretical knowledge in a state vocational school [34]. Therefore, vocational training is simulated accordingly.

Finally, persons beyond primary, secondary or tertiary education can enter the educational register by adult education.

The structure of a microsimulation is always similar [19, 22, 23]:

1. An initial dataset has to be created, modelling the characteristics of the real-world population as close as possible.

2. The initial dataset is updated regularly (most often: yearly) to reflect changes in the population.

Usually, a census or a large survey is needed for modelling the initial dataset. Since no dataset containing all information required was available, the initial dataset was also simulated. For this initial dataset, a start dataset for the year 2000 was generated and updated for ten simulated years. The start dataset consists of about 18.9 million records; the final dataset has about 27 million records.

Data generation

The basic QIDs (first name, surname, sex, full date of birth, place of birth) are simulated in detail. QIDs, in general, do not follow simple, known distributions. This fact is one of the main reasons why a microsimulation seemed necessary for the given problem.

The distribution of personal names is far from uniform. For example, Hanks and Tucker [13] reported that 90% of the population of the United States have one of only 67,000 surnames. Furthermore, the birthday is not uniformly distributed across a year or a week. For example, in modern societies more births are recorded during the week than during the weekend [18]. Due to these skew distributions some combinations of QIDs are more common than others. In addition, QIDs are not independent, and their dependencies are challenging to model. Consequently, the joint distribution of QIDs is hard to describe analytically and empirical studies seem to be rare.

Names are essential QIDs for identifying people (if no unique identification number or proxies given by reliable and stable attributes are available). To model changing fashions concerning first names an empirical conditional joint distribution of first names and surnames conditioned on a specific year was implemented in the simulation. To get access to a joint distribution of encrypted identifiers, a formal request, including a detailed description of the project to the data guardian of a local educational register was necessary. For name generation, the joint distribution of the encrypted identifiers was aligned to the marginal frequency distributions of an unencrypted commercial database. The alignment process for first names was stratified for sex (two strata). Additionally, the process for first names and the process for surnames was stratified for migration status (two strata: foreign birth or not). The alignment used the frequency distributions of each token of first names (up to six tokens) and the frequency distribution of each token of surnames (up to four tokens) was grouped into percentiles. Within each stratum of the encrypted distribution, the hash code of the token was replaced by a randomly sampled token from the same percentile in the corresponding stratum of the unencrypted distribution. This process was repeated for each token of a full name in the encrypted database. Since the encrypted database contained about 400,000 names for 600,000 records, the repeated sampling from the same percentile resulted in 10 million names for 27 million records.

The simulated day and month of birth were sampled from an empirical distribution of births in German hospitals. The distribution of the year of birth was taken from official statistics. However, due to data protection regulations, no empirical distribution of the day and month of birth for migrants has been available. Therefore, for this subpopulation, a uniform distribution was used. If this assumption should be wrong, more QIDs for the unique identification of persons are required.

The place of birth was randomly assigned proportionally to the population size according to official statistics of about 10.760 communities. Due to German data protection regulations no empirical distribution of places of birth for migrants are available. Therefore, simulated migrants, stratified by country of origin, were assigned randomly to lists of country-specific communities.

Error generation

Usually, the data in a microsimulation is considered as free of errors [19, 23]. In contrast, in the application given here, the errors in a subset of the data (the QIDs) are of central interest. Since the data quality of the register is unknown, the simulation covers a range of plausible error rates.

Error rates are the percentage of QID values entered with errors. Four error rates (0.1%, 0.3%, 0.7% and 1%) were simulated. Since foreign names may be unfamiliar to data entry personnel, we assumed a doubling of the error rates for foreign names. Every time an individual record is affected by an event (for example, moving, marriage or finishing school), an error-generating process is triggered.

The errors are produced by error-generating functions specific to each type of QID. For all integer QIDs (day, month, year, sex), an error is the replacement of the integer by a different integer. Errors in strings (first name, surname, place of birth) are more challenging to model. Letters can be substituted, inserted, deleted or transposed. It is also known that errors are more common at the end than at the beginning of a string. Christen and Pudjijono [7] reported 8% of string errors in the first letter. Due to the lack of additional information, we assumed a uniform distribution for the remaining errors across the other positions. Although multiple errors are possible, one error per string is more common [7]. Peterson [25] notes that 7.1% of string fields contain multiple errors. We limited the number of multiple errors to two. The type of error was simulated using the frequencies given by Peterson [25]. Since we simulated an educational register whose purpose is surveillance of awarded certificates, incomplete data – common in other administrative databases – was not simulated.

Simulated events

The most common events resulting in an update of the register had to be simulated. The probabilities for transitions are based on official statistics and a large number of different special purpose surveys. A simplified flowchart of the simulation is shown in Figure 1.

The model simulates at least ten years of compulsory schooling. The simulated secondary education reflects at least 13 years of school attendance (classes can be repeated). For schoolchildren, demographic processes such as death or out-migration are modelled as a separate process resulting in a termination of the updating processes for these cases. In contrast, for students and apprentices the length of education was implemented as a random variable (fixed at the start of tertiary education). For an attainment register no records are ever deleted. This might increase the error rate in RL for a long-running register if no information of death or out-migration is considered.

An essential marker for changes in identifiers is marriage. Marriages often result in the change of surname of one partner and sometimes in the change of address. In the simulation, everyone over eighteen years of age is eligible for marriage. The probabilities for marriage conditioned on age are based on official statistics. The kind of name change and the sex of the partner that changes the name were modelled using data published by a government-sponsored language society.[1]

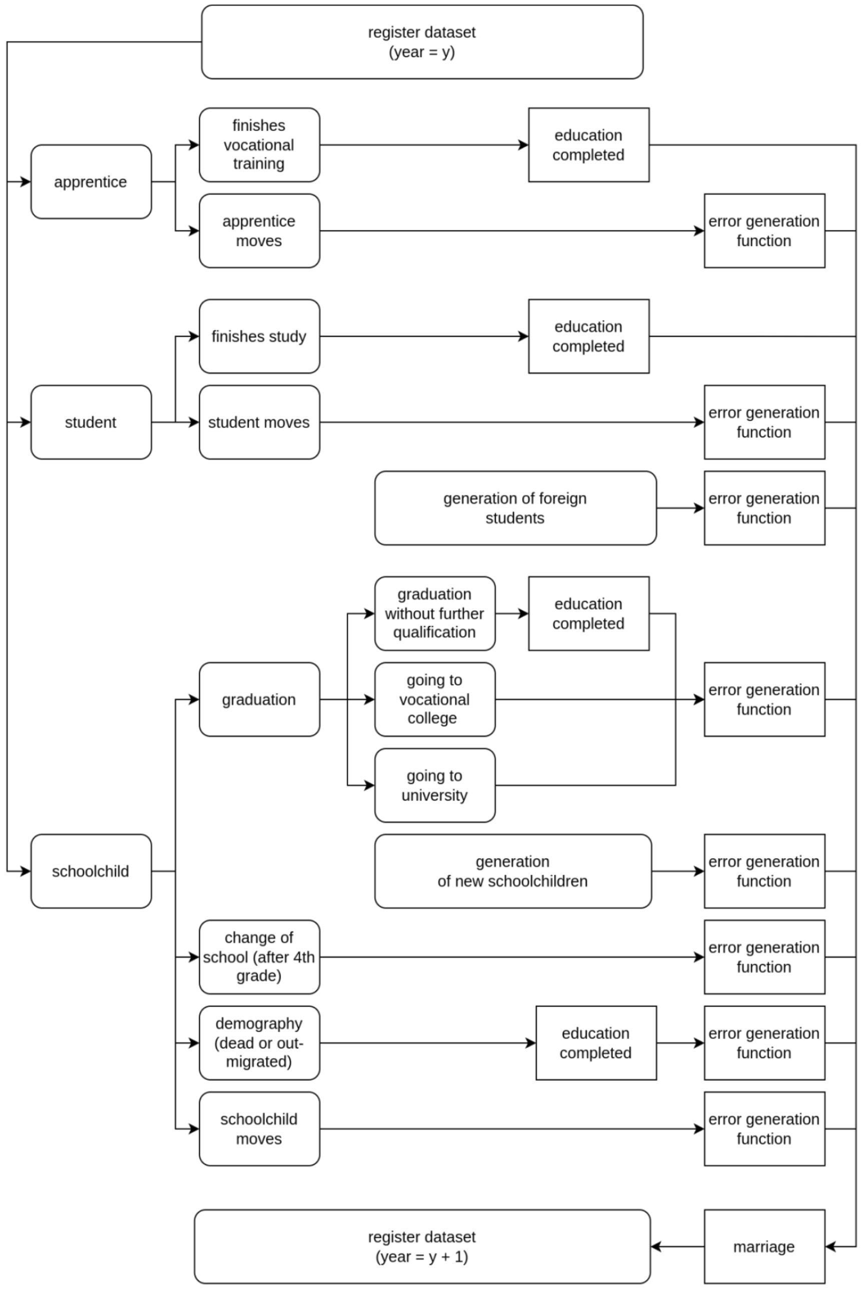
Record linkage methods

RL is a classification problem of potential pairs of records as match or non-match. For a given record, the best matching record has to be found. If a probability for a match is calculated, the linkage is probabilistic [6, 15]. If only exact matches are used, the linkage is deterministic. Many different RL methods have been proposed [6]. If the data quality of QIDs is high and the number of missing QIDs low, the performance of modern RL methods should be similar, at least in small datasets. In population wide datasets these assumptions are usually not given. Therefore, many national statistical institutes seem to rely on probabilistic record linkage. In medical settings with smaller datasets other RL methods are also popular. Studies comparing the performance of different RL methods on population wide datasets seem to be rare.

However, the available options for real-world applications that exceed about 10 million records per dataset are limited, since most elaborate methods do not scale well in theory. In addition, many suggested algorithms have no working implementation tested on datasets of this size. Therefore, we selected the most widely used methods for large datasets (exact matching, phonetic matching, Fellegi-Sunter, Bloom-filters) and a recently suggested promising method (multiple matchkeys). Each method will be described briefly. The most

[1] https://gfds.de/familiennamen-bei-der-heirat-und-vornamenprognose-2018/, last retrieved 15.06.2022.

Figure 1: Simplified scheme of the main generating processes within a simulated year. In the model, more than one event can happen per year.


register dataset (year = y)
apprentice
finishes vocational training
education completed
apprentice moves
error generation function
student
finishes study
education completed
student moves
error generation function
generation of foreign students
error generation function
graduation without further qualification
education completed
graduation
going to vocational college
error generation function
going to university
schoolchild
generation of new schoolchildren
error generation function
change of school (after 4th grade)
error generation function
demography (dead or out-migrated)
education completed
error generation function
schoolchild moves
error generation function
register dataset (year = y + 1)
marriage

basic form of RL is deterministic RL using a single matchkey. A matchkey is a set of QIDs on which two records must agree to be classified as a match [15]. Several combinations of QIDs for matchkeys have been proposed in the respective literature, for example, the Australian Statistical Linkage Key (SLK-581; [16]) and the Swiss Anonymous Linkage Code (Swiss ALC; [4]). In the simulation, additional variants, including phonetic codes such as Soundex [6], place of birth and surname, were tested to cover informal suggestions of data protection officers involved in discussions of the educational register.

Probabilistic RL is based on the Fellegi-Sunter-Model [11] and later developments [10, 15]. The model computes agreement vectors for all given record pairs. These vectors consist of the agreements and non-agreements of each QID. Based on previously estimated weights, the agreement vectors are transformed into match probabilities. If the match probability for a record pair exceeds a certain threshold, the pair is classified as a match. Based on previous experience in linking simulated German census data, we use a threshold of 0.8 for all probabilistic linkages. The weights are calculated using the Estimation Conditional Maximization (ECM) algorithm [15, 21].

Although probabilistic RL usually yields excellent linkage quality, the ECM is computationally demanding. To reduce the number of comparisons pairwise computations are limited to subsets of records sharing common attributes, for example, year of birth. These subsets are denoted as blocks. Most RL systems apply multiple blocking variables sequentially [6]. After each blocking step, matches were removed from the database, thereby reducing the number of records and thus enabling more relaxed blocking rules. The following blocking rules were used:

- If place of birth is available: (1) day & month & year, (2) Soundex(first name) & Soundex(surname), (3) Soundex(first name), (4) place of birth.
- If place of birth is not available: (1) day & month & year, (2) Soundex(first name) & Soundex(surname), (3) Soundex(first name) & year.

For each blocking step, parameter estimations are based on a random sample of 200,000 records of the total dataset.

We tested different popular ECM implementations. The simulation was run on a Supermicro server with an AMD EPYC 7702P CPU (64 cores, 128 threads, max. 2.1 GHz) and 1 TB RAM operating under Ubuntu 20.04.4 using R 4.1.2 and Python 3.8.10. Most linkage programs became unstable with large datasets or exceeded the available memory. Promising implementations seemed to be *Splink* (version 1.0.6) [24], *FastLink* (version 0.6.0) [10] and *recordlinkage* (version 0.14) [9], which is based on *Febrl* [5]. After many experiments, *recordlinkage* proved to be the only stable implementation with large datasets.

The idea of using a set of separate encrypted QIDs for matching has been suggested several times in the respective literature [1, 27, 33]. A recent suggestion is due to Randall et al. [26]. In this multiple matchkey approach, the matchkeys are combinations of all non-missing QID fields. The QIDs are concatenated and hashed. Since every field combination yields a matchkey, a large number of matchkeys results. This number of matchkeys is reduced by removing matchkeys containing a superset of QIDs if a subset of these QIDs would be accepted as a match [26]. For example, if two matchkeys differ only in their surname usage, they can be reduced to the matchkey that does not contain the surname. The selected matchkeys are based on weights estimated by an EM algorithm. Only matchkeys exceeding a predefined threshold are used.

The multiple matchkey approach considers a sum of the weights in the Fellegi-Sunter-model exceeding a threshold as indicating a suitable matchkey. To select this threshold, we linked the first datasets with integer thresholds within the observed range from 1 to 19 (see Figure 2).

The threshold with the highest linkage quality for a given set of available QIDs was selected. Linkage quality is almost always measured by precision and recall [6]. If a single measure has to be reported, the recent literature reports the arithmetic mean (denoted as F_2) instead of the traditional harmonic mean (F_1) of precision and recall [12].

To identify matches among record pairs, we started with record pairs agreeing on all matchkeys. These records are removed from the datasets. Then we stepwise decremented the number of identical matchkeys required down to one. At every step, exact matches are removed, and the next step operates only on the remaining records. In the end, records are classified as matches if they agree on at least one matchkey.

This approach links two years of the simulated register in less than two minutes.

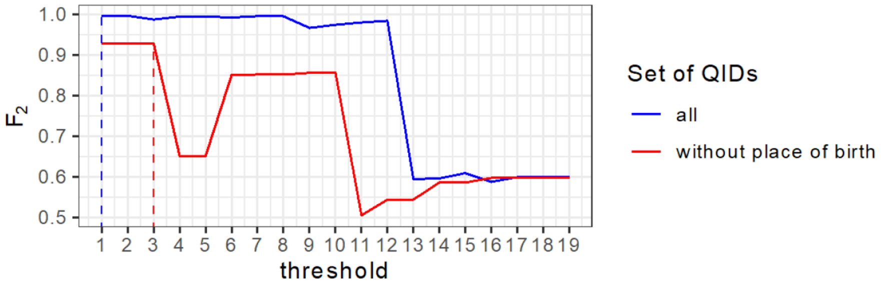
From the point of view of a data guardian, a linkage based on encoded identifiers might be preferable over a linkage based on clear text. Record linkage using encoded identifiers is called privacy-preserving record linkage (PPRL; [8]). Therefore, PPRL methods were tested in the simulation to study if a PPRL solution would be feasible for an educational attainment register. Single and multiple matchkeys can be encrypted using keyed hash functions. Thus, they can be considered PPRL methods. Besides linkage keys, we included another popular PPRL method in the tests.

This method is based on Bloom Filters (BF). A QID is split into consecutive letters (q-grams) to create a BF. Every q-gram is mapped by k different hash functions (such as SHA-1) into several bit positions on a bit vector with length l. By setting all selected bits to one, a record pair can be compared using q-gram similarity functions like the Dice coefficient [31]. Different modifications of the initial suggestion aim to protect BF against cryptographic attacks which have been engineered in the last decade to attack BF [8]. One such modification is a Cryptographic Longterm Key (CLK) where all QIDs are encoded into a single large BF [32]. Record pairs can be compared according to the similarity of their CLKs. In the simulation, CLKs of length $l = 1000$ with $k = 10$ hash functions and bigrams ($q = 2$) for string QIDs (first name, surname, place of birth) and unigrams ($q = 1$) for integer QIDs (sex, date of birth) were employed.

During the initial tests, we noted that German administrative data contain unusual long strings in the name fields. These long strings are due to compound surnames and multiple first names. For example, the full name of a former minister of defence is a string with 108 characters.

Long strings containing many q-grams will set many bits to one in BFs and CLKs. For the similarity calculation, long strings will result in higher weights of these strings. We,

Figure 2: Threshold selection for the multiple matchkey method. The plots show the linkage quality (F_2) dependent on the tested thresholds (integer of the sum of weights in the Fellegi-Sunter-model) of the multiple matchkey approach. The maximum of each combination of QIDs is indicated by a dashed line.



therefore, tried to limit the effect of long strings by taking only the first token of a string into account.

Multibit trees [17, 28] were used to match records with CLKs. A Multibit tree is a binary tree structure generated from an existing list of bit vectors. Each vector gets sorted into the leaves of the tree. To find similar vectors, the tree is searched, and matches above a predefined similarity threshold are returned. Similarity is estimated by the Tanimoto coefficient [17]. In previous applications of Multibit trees for PPRL, Tanimoto thresholds between 0.8 and 0.85 were successful [3, 28]; consequently, these two values were tested. To reduce the size of the Multibit tree, we used two independent blockings (year of birth, Soundex of surnames). Finally, for each record in the previous year, the highest-scoring potentially matching record of the following year among the two blocking strategies was selected.

Results

Precision for error rates above 0.1% are shown in Figures 3 and 4 (complete results are available in the technical reports [30, 35]). The columns are the levels of error simulated, and the rows are the different RL methods tested. Figure 3 shows the precision for single matchkey techniques, Figure 4 the precision for all other RL methods.

Even under the worst simulated conditions (high error rates, foreign names, no place of birth, ten years of register activity), precision for exact matchkeys exceeds 0.995. Among the more elaborate methods, ECM and CLK achieve precision above 0.990. However, four tendencies are obvious for all methods. Precision decreases

1. with increasing error rates,
2. with the duration of register activity,
3. if the place of birth is not available and
4. for foreign names.

Figure 5 shows the recall for single matchkey techniques, Figure 6 the recall for all other RL methods. Like precision, recall decreases with increasing error rates, duration of register activity, foreign names and missing place of birth. However, the recall of all single matchkeys beyond the lowest error rates is below 0.9 and approaches nearly 0.7 under unfavourable conditions. The plots show large differences between migrants and non-migrants, indicating linkage bias. However, in a complex linkage strategy single linkage keys might be employed to filter records with few errors quickly.

Overall, multiple matchkeys using all identifiers yield the highest recall of all methods considered. The second-best recall was observed for ECM. If the place of birth is unavailable, recall of both methods declines, but the loss is higher for ECM.

This difference might be due to a specific feature of the RL problem given here: the simulated dataset of the previous year is always a perfect subset of the following year. Since no records are ever deleted, the dataset will contain more records each year. In most other RL settings, the overlap between datasets will be smaller, and rarely will any dataset be an exact subset of another dataset. Although the difference in recall between multiple matchkeys and ECM cannot be generalized to other datasets, the advantage of multiple matchkeys is evident for the given application.

Among the PPRL methods, CLKs show the best recall. In comparison to single matchkeys, CLKs are less affected by increasing error rates. This might be due to the usage of tokens for string QIDs. Since the use of just the first token

Figure 3: Precision for single matchkey RL methods.

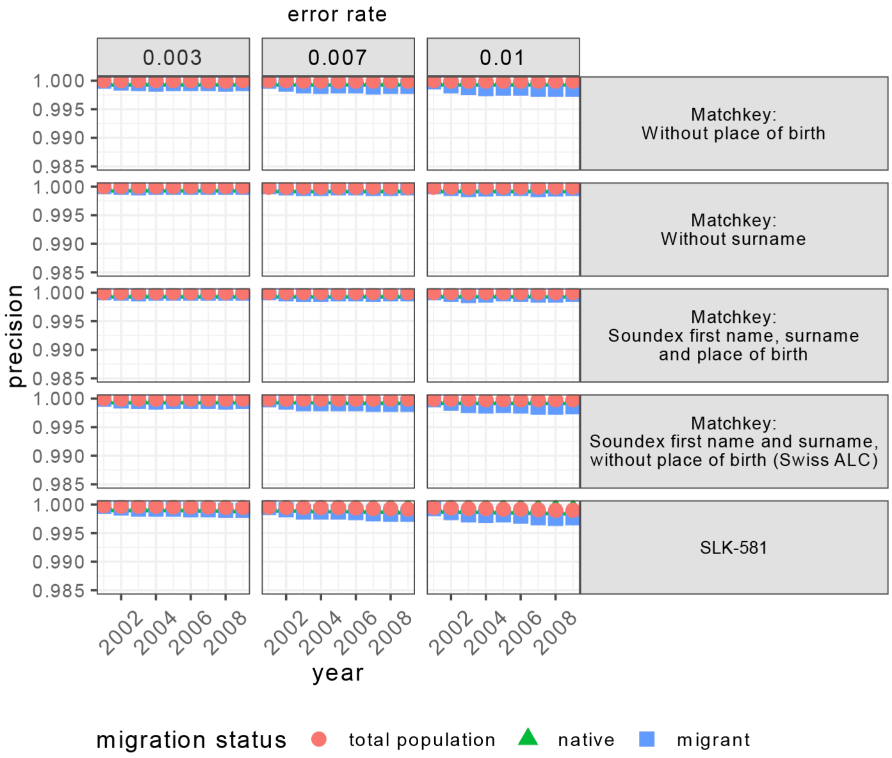


Figure 4: Precision for multiple matchkey, ECM and CLK methods. The symbol *T* in the facet labels on the right is the Tanimoto threshold of the Multibit trees

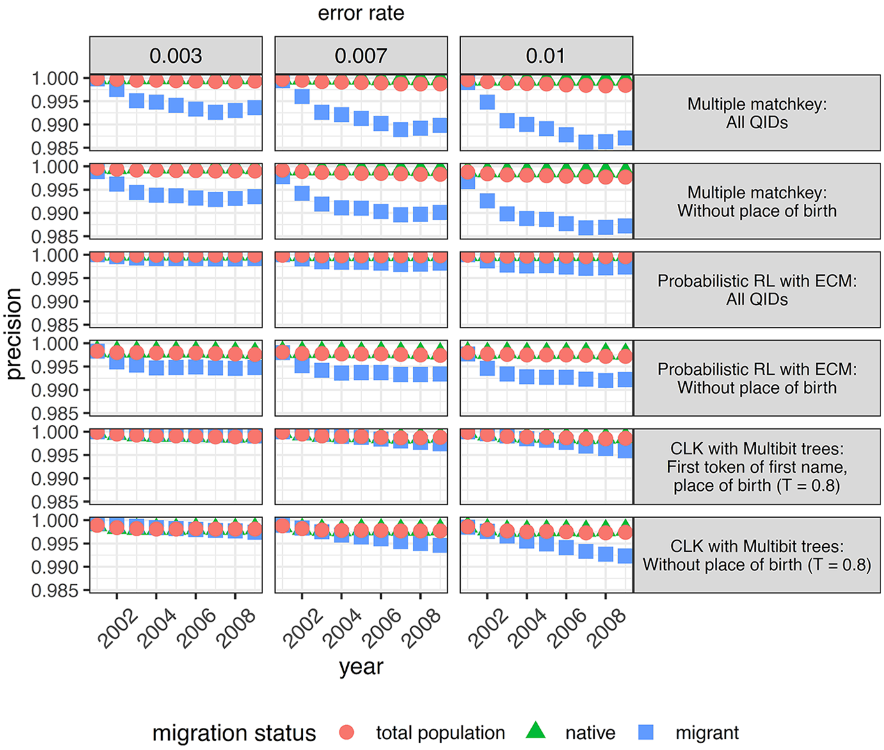
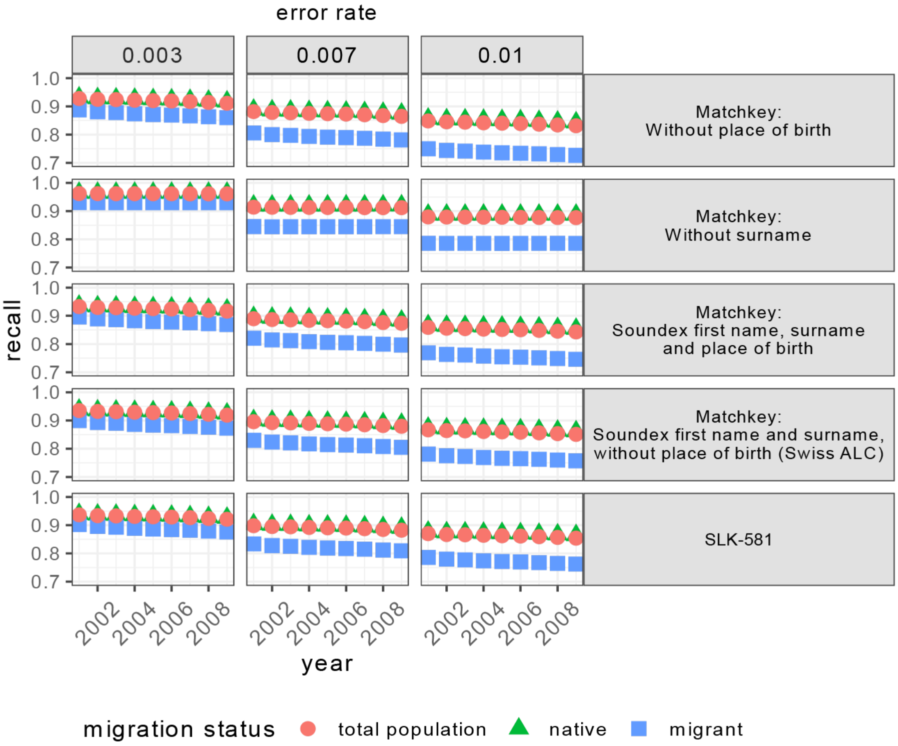


Figure 5: Recall for single matchkey RL methods.

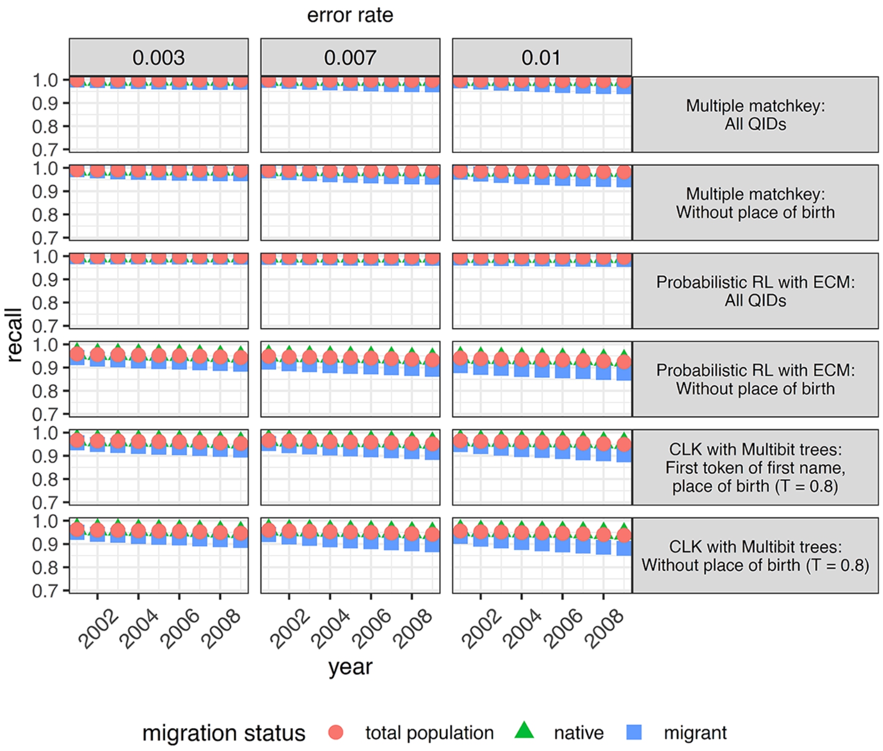


excludes all errors in the remaining tokens, this result seems plausible. Therefore, applying tokens might improve CLKs. Without place of birth, recall for CLKs decreases, but is still comparable to ECM. However, in this application multiple matchkeys show better recall than CLKs.

Discussion

The novel microsimulation approach to systematially evaluate the use of QIDs of a register showed that the main problem of a register would be linkage bias caused by insufficient

Figure 6: Recall for multiple matchkey, ECM and CLK methods. The symbol T in the facet labels on the right is the Tanimoto threshold of the Multibit trees.



information to identify a person uniquely. Therefore, additional identifiers beyond the basic set of QIDs (first name, surname, sex, date of birth) are required in the register. Many administrators see the costs resulting from administrative actions based on wrong positive links as exceeding the costs of missed links. In law enforcement or medical contexts, the loss function is often different. However, an educational attainment register is used for evaluating policies, not to decide on the handling of individuals. Therefore, losing cases over time will happen with RL methods yielding low recall in longitudinal studies. The resulting attrition might be related to individual characteristics and cause linkage bias: dependent variables of interest differ between linked and unlinked records. For an educational attainment register, recall is at least as important as precision. Therefore, RL methods suited to yield high precision at the cost of a low recall, such as single matchkeys, are unsuitable for national educational attainment registers.

RL methods suited for population-size datasets are limited. Currently, multiple matchkeys and ECM seem to be acceptable choices. However, a final decision on the RL method is not necessary at the start of a register. Since an attainment register is a long-term project that might operate for decades using QIDs, it seems reasonable to store plaintext QIDs. In general, plaintext allows clerical review and, in this way, estimations of error rates in subsamples. Furthermore, unexpected encodings or unknown variants of QIDs could be detected and used for improving preprocessing. Finally, plaintext will guarantee that future developments in RL can use existing datasets. In addition, also historical linkages could be updated using future methods. These advantages of plaintext are further arguments against using PPRL methods in official statistics [29].

Conclusion

To answer the question of the required set of QIDs for a national educational attainment register, we built a microsimulation of the educational processes underlying such a registry. Since the backbone of this kind of register is a record linkage method, the performance of different RL methods was tested. Beyond the principle feasibility of a QID-based register, the central problem is the potential of linkage bias. Linkage bias in this simulation is caused by two processes: Different error rates between subpopulations and marriages. It could be shown that linkage bias will occur regardless of the RL method applied. However, the availability of additional identifiers (in the given case: the place of birth) will yield higher precision and recall despite errors in identifiers. Selecting a suitable RL method will mitigate the RL problems. Simple matchkeys are unsuited for a national register. Multiple matchkeys and ECM are the best available options for the given problem. PPRL methods should only be employed if the clear text is unavailable due to legal constraints.

A microsimulation was not used before to simulate the processes generating QIDs for a register. We showed that microsimulation of QIDs can be a valuable tool to inform policies on the required set of QIDs for a register. Without microsimulations, the potential of linkage bias for vulnerable subgroups (here: female migrants) in registers can hardly be shown empirically before a register exists. Therefore, we hope this approach will contribute to design national registers beyond this educational attainment register.

The limitations of the conclusions are given by the design of the simulation study. Since the data quality is unknown, the range of errors covered by the simulation could be too low or too high. If more errors are likely, the main effects described above will be observed earlier. Less than 0.1% errors seem unlikely, but in this case the same effects will be seen after more years of register effect. The assumption of a doubled error rate for migrants might also be erroneous. However, linkage bias is likely as long as the error rates between subpopulations differ. Furthermore, ethnicity is consistently associated with higher identifier error rates and lower linkage rates [2, 14]. Therefore, we consider the conclusions stable regarding the assumed error rates.

Only four principal RL methods (matchkeys, multiple matchkeys, ECM, CLK) have been studied here. The main conclusions are independent of the methods applied. The main problem for RL in such contexts is the lack of information to identify a person uniquely. Since the lack of information is due to the design of the databases and not a consequence of any features of persons, computational procedures cannot generate additional information required to identify a person. Therefore, it seems unlikely that applying other RL methods would yield other conclusions.

Some RL methods match multiple records of a dataset to the same entity in another dataset (many-to-one mappings). In real-world applications, such multiple assignments are often resolved by manually selecting the most likely match [6, p. 174]. In a simulation, such a clerical review is not possible. Therefore, all many-to-one matches in the simulation are regarded as non-matches. Linkage quality might be better due to clerical review in a real-world register. However, since clerical review could, in principle, be applied independently from the RL method, this additional step most likely will not impact the relative performance of the RL methods. Furthermore, a clerical review is impossible if a blindfolded PPRL is required.

Two problems were identified during the preparation of the simulation. There is no stable, proven publicly available open-source software for RL using QIDs of datasets with more than 10 million records and large blocks. Furthermore, there is little published research on the quality of identifiers in administrative databases. Both problems need attention for improving research using population-covering databases.

Acknowledgements

We are indebted to Tobias Brändle for his help accessing and analysing data of a regional educational registry. Sean Randall gratefully provided his source code for the multiple matchkey approach.

Ethics statement

Since only simulated data was used, ethical approval was not required.

Statement on conflicts of interest

The authors declare no conflicts of interest.

Funding

The study was commissioned to the first author by the German Federal Ministry of Education and Research. The funding body had no role in the design of the study, simulation, analysis or interpretation, nor in writing the manuscript.

References

1. Abbott O, Jones P, and Ralphs M. Large-scale linkage for total populations in official statistics. Ed. by Harron K, Goldstein H, and Dibben C. Chichester: Wiley, 2016. Chap. 8:170–200. https://doi.org/10.1002/9781119072454.ch8

2. Bohensky M. Bias in Data Linkage Studies. *Methodological Developments in Data Linkage*. Ed. by Harron K, Goldstein H, and Dibben C. Chichester: Wiley, 2016. Chap. 4:63–82. https://doi.org/10.1002/9781119072454.ch4

3. Borgs C. Optimal Parameter Choice for Bloom Filter-Based Privacy-Preserving Record Linkage. PhD thesis. University of Duisburg-Essen, 2019. https://doi.org/10.17185/duepublico/70274

4. Borst F, Allaert FA, and Quantin C. The Swiss Solution for Anonymously Chaining Patient Files. *Proceedings of the 10th World Congress on Medical Informatics: Held in London, United Kingdom*. Ed. by Patel VL, Rogers R, and Haux R. Studies in Health Technology and Informatics 84. Amsterdam: IOS Press, 2001.

5. Christen P. Febrl – a Freely Available Record Linkage System with a Graphical User Interface. *HDKM '08 Proceedings of the Second Australasian Workshop on Health Data and Knowledge Management*. Ed. by Warren JR, Yu P, Yearwood J, and Patrick JD. Wollongong: ACS, 2008:17–25.

6. Christen P. Data Matching: Concepts and Techniques for Record Linkage, Entity Resolution, and Duplicate Detection. Data-Centric Systems and Applications. Berlin, New York: Springer, 2012.

7. Christen P and Pudjijono A. Accurate Synthetic Generation of Realistic Personal Information. *Advances in Knowledge Discovery and Data Mining*. Ed. by Theeramunkong T, Kijsirikul B, Cercone N, and Ho TB. Vol. 5476. Lecture Notes in Computer Science. Berlin, Heidelberg: Springer, 2009:507–14. https://doi.org/10.1007/978-3-642-01307-247

8. Christen P, Ranbaduge T, and Schnell R. Linking Sensitive Data: Methods and Techniques for Practical Privacy-Preserving Information Sharing. Cham: Springer, 2020.

9. De Bruin J. Probabilistic Record Linkage with the Fellegi and Sunter Framework: Using Probabilistic Record Linkage to Link Privacy Preserved Police and Hospital Road Accident Records. PhD thesis. Delft University of Technology, 2015.

10. Enamorado T, Fifield B, and Imai K. Using a Probabilistic Model to Assist Merging of Large-Scale Administrative Records. American Political Science Review. 2019 May; 113(2):353–71. https://doi.org/10.1017/S0003055418000783

11. Fellegi IP and Sunter AB. A Theory for Record Linkage. Journal of the American Statistical Association. 1969; 64(328):1183–210.

12. Hand D and Christen P. A Note on Using the F-Measure for Evaluating Record Linkage Algorithms. Statistics and Computing. 2018; 28(3):539–47. https://doi.org/10.1007/s11222-017-9746-6

13. Hanks P and Tucker DK. A Diagnostic Database of American Personal Names. Names 2000; 48:59–69. https://doi.org/10.1179/nam.2000.48.1.59

14. Harron K, Hagger-Johnson G, Gilbert R, and Goldstein H. Utilising Identifier Error Variation in Linkage of Large Administrative Data Sources. BMC Medical Research Methodology 2017; 17. https://doi.org/10.1186/s12874-017-0306-8

15. Herzog TN, Scheuren F, and Winkler WE. Data Quality and Record Linkage Techniques. New York, London: Springer, 2007.

16. Karmel R. Data Linkage Protocols Using a Statistical Linkage Key. Tech. rep. Canberra: Australian Institute of Health and Welfare, 2005. Available from: https://www.aihw.gov.au/reports/aged-care/data-linkage-protocols-statistical-linkage-key/contents/table-of-contents [Accessed on: 2022 Oct 4].

17. Kristensen TG, Nielsen J, and Pedersen CN. A Tree-Based Method for the Rapid Screening of Chemical Fingerprints. Algorithms for Molecular Biology. 2010; 5(1):9–20. https://doi.org/10.1186/1748-7188-5-9

18. Lerchl A. Where are the Sunday babies? Observations on a marked decline in weekend births in Germany. Naturwissenschaften 2005; 92:592–4. https://doi.org/10.1007/s00114-005-0049-y

19. Li J and O'Donoghue C. A Survey of Dynamic Microsimulation Models: Uses, Model Structure and Methodology. International Journal of Microsimulation. 2013; 6(2):3–55. https://doi.org/10.34196/ijm.00082

20. Mellander E. On the Use of Register Data in Educational Science Research. Nord-STEP 2017; 3(1):106–18. https://doi.org/10.1080/20020317.2017.1313680

21. Meng XL and Rubin DB. Maximum Likelihood Estimation via the ECM Algorithm: A General Framework. Biometrika. 1993; 80(2):267–78.

22. Münnich R, Schnell R, Brenzel H, Diekmann H, Dräger S, Emmenegger J, Höcker P, Kopp J, Merkle H, Neufang K, Obersneider M, Reinhold J, Schaller J, Schmaus S, and Stein P. A Population Based Regional Dynamic Microsimulation of Germany: The MikroSim Model. methods, data, analyses. 15 (2):241–64. https://doi.org/10.12758/MDA.2021.03

23. O'Donoghue C. Practical Microsimulation Modelling. Oxford: Oxford University Press, 2021.

24. Office for National Statistics. Splink: MoJ's Open Source Library for Probabilistic Record Linkage at Scale. 2020. Available from: https://www.gov.uk/government/publications/joined-up-data-in-government-the-future-of-data-linking-methods/splink-mojs-open-source-library-for-probabilistic-record-linkage-at-scale [Accessed on: 2022 May 4].

25. Peterson JL. A Note on Undetected Typing Errors. Communications of the ACM. 1986; 29(7):633–7.

26. Randall S, Brown AP, Ferrante AM, and Boyd JH. Privacy Preserving Linkage Using Multiple Dynamic Match Keys. International Journal of Population Data Science. 2019; 4(1). https://doi.org/10.23889/ijpds.v4i1.1094

27. Schmidtmann I, Sariyar M, Borg A, Gerold-Ay A, Heidinger O, Hense HW, Krieg V, and Hammer GP. Quality of Record Linkage in a Highly Automated Cancer Registry that Relies on Encrypted Identity Data. GMS Medizinische Informatik, Biometrie und Epidemiologie 2016; 12.

28. Schnell R. An Efficient Privacy-Preserving Record Linkage Technique For Administrative Data and Censuses. Statistical Journal of the IAOS 2014; 30:263–70. https://doi.org/10.3233/SJI-140833

29. Schnell R. Privacy Preserving Record Linkage in the Context of a National Statistical Institute. Office of National Statistics, 2021. Available from: www.gov.uk/government/publications/joined-up-data-in-government-the-future-of-data-linking-methods/privacy-preserving-record-linkage-in-the-context-of-a-national-statistics-institute [Accessed on: 2022 Oct 8].

30. Schnell R. Verknüpfung von Bildungsdaten in einem Bildungsregister mittels Record-Linkage auf Basis von Personenmerkmalen. GRLC Working Paper Series 2022; 03/2022. https://doi.org/10.17185/duepublico/76331

31. Schnell R, Bachteler T, and Reiher J. Privacy-Preserving Record Linkage Using Bloom Filters. BMC Medical Informatics and Decision Making. 2009; 9(41):1–11. https://doi.org/10.1186/1472-6947-9-41

32. Schnell R, Bachteler T, and Reiher J. A Novel Error-Tolerant Anonymous Linking Code. SSRN Electronic Journal. 2011. https://doi.org/10.2139/ssrn.3549247

33. Shipsey R and Plachta J. Linking with Anonymised Data – How not to Make a Hash of it. Tech. rep. ONS, 2020. Available from: www.gov.uk/

government/publications/joined-up-data-in-government-the-future-of-data-linking-methods/linking-with-anonymised-data-how-not-to-make-a-hash-of-it [Accessed on: 2022 Oct 5].

34. Solga H, Protsch P, Ebner C, and Brzinsky-Fay C. The German vocational education and training system: Its institutional configuration, strengths, and challenges. WZB Discussion Paper SP I 2014-502. Berlin, 2014. Available from: http://hdl.handle.net/10419/104536 [Accessed on: 2022 Oct 3].

35. Weiand SV. Vergleich von Record-Linkage Methoden anhand der Mikro-Simulation eines bundesweiten Schülerregisters. Working Paper 04/2022. University of Duisburg-Essen: German Record-Linkage Center, 2022. https://doi.org/10.17185/duepublico/76361

36. Zhu Y and Xu L. Returns to Higher Education – Graduate and Discipline Premiums. Discussion Paper 15229. IZA, 2022. https://doi.org/10.2139/ssrn.4114886

Abbreviations

BF:	Bloom-Filter
CLK:	Cryptographic Longterm Key
ECM:	Expectation Conditional Maximization
EM:	Expectation Maximization
GDPR:	General Data Protection Regulation
PPRL:	Privacy-Preserving Record Linkage
QID:	Quasi-Identifier
RL:	Record Linkage

DOI: 10.23889/ijpds.v8i1.2122
URN: urn:nbn:de:hbz:465-20230821-170642-2