OPEN ACCESS

Research Article

Quality of record linkage in a highly automated cancer registry that relies on encrypted identity data

Zur Güte des Record Linkage mit verschlüsselten Identitätsdaten in einem weitgehend automatisierten Krebsregister

Abstract

Objectives: In the absence of unique ID numbers, cancer and other registries in Germany and elsewhere rely on identity data to link records pertaining to the same patient. These data are often encrypted to ensure privacy. Some record linkage errors unavoidably occur. These errors were quantified for the cancer registry of North Rhine Westphalia which uses encrypted identity data.
Methods: A sample of records was drawn from the registry, record linkage information was included. In parallel, plain text data for these records were retrieved to generate a gold standard. Record linkage error frequencies in the cancer registry were determined by comparison of the results of the routine linkage with the gold standard. Error rates were projected to larger registries.
Results: In the sample studied, the homonym error rate was 0.015%; the synonym error rate was 0.2%. The F-measure was 0.9921. Projection to larger databases indicated that for a realistic development the homonym error rate will be around 1%, the synonym error rate around 2%.
Conclusion: Observed error rates are low. This shows that effective methods to standardize and improve the quality of the input data have been implemented. This is crucial to keep error rates low when the registry's database grows. The planned inclusion of unique health insurance numbers is likely to further improve record linkage quality. Cancer registration entirely based on electronic notification of records can process large amounts of data with high quality of record linkage.

Keywords: record linkage, cancer registry, evaluation, encrypted data, data quality

Zusammenfassung

Ziel der Arbeit: Krankheitsregister in Deutschland und einigen anderen Ländern sind darauf angewiesen, mehrere Meldungen zu einem Patienten anhand von Identitätsdaten zusammenzuführen, da keine eindeutige Personenkennung verfügbar ist. Diese Identitätsdaten werden häufig aus Datenschutzgründen pseudonymisiert. Einige Fehler beim Record Linkage sind unvermeidbar. In der vorliegenden Arbeit werden sie für das Epidemiologische Krebsregister Nordrhein-Westfalen, das chiffrierte Identitätsdaten zum Record Linkage verwendet, quantifiziert.
Methoden: Eine Stichprobe von Meldungen an das Epidemiologische Krebsregister Nordrhein-Westfalen, die mit der Information über die Zuordnung zu einer Person versehen war, wurde gezogen. Parallel dazu wurden Klartextidentitätsdaten für diese Meldungen durch Dechiffrierung wiedergewonnen, um einen Gold-Standard zu erzeugen. Die Zuordnungsfehlerhäufigkeiten wurden durch Vergleich der Ergebnisse des Routine-Record Linkage mit dem Gold-Standard bestimmt. Die Fehlerraten wurden für umfangreichere Register hochgerechnet.

Irene Schmidtmann[1]
Murat Sariyar[2]
Andreas Borg[1]
Aslihan Gerold-Ay[1]
Oliver Heidinger[3]
Hans-Werner Hense[3]
Volker Krieg[3]
Gaël Paul Hammer[4]

1 Institut für Medizinische Biometrie, Epidemiologie und Informatik (IMBEI), Universitätsmedizin der Johannes Gutenberg-Universität Mainz, Germany

2 TMF – Technologie- und Methodenplattform e. V., Berlin, Germany

3 Epidemiologisches Krebsregister Nordrhein-Westfalen (EKR NRW), Münster, Germany

4 Registre Morphologique des Tumeurs au Grand-Duché de Luxembourg, Laboratoire National de Santé, Luxembourg, Luxembourg

Ergebnisse: In der untersuchten Stichprobe betrug die Homonymfehlerrate 0.015%, die Synonymfehlerrate 0.2%, das F-Maß ergab 0.9921. Eine Hochrechnung auf umfangreichere Datenbanken ergab, dass bei realistischen Annahmen über die Zunahme des Umfangs eine Homonymfehlerrate von etwa 1% und eine Synonymfehlerrate von etwa 2% zu erwarten sind.
Schlussfolgerung: Die beobachteten Fehlerraten sind niedrig, darin zeigt sich, dass effektive Methoden zur Standardisierung und Sicherung der Datenqualität implementiert wurden. Dies ist essenziell, damit die Fehlerraten niedrig bleiben, wenn der Umfang des Registers zunimmt. Durch den geplanten Einschluss der eindeutigen Krankenversicherungsnummer ist eine weitere Verbesserung der Güte des Record Linkage zu erwarten. Krebsregistrierung, die allein auf elektronische Meldungen basiert, kann große Datenmengen verarbeiten bei hoher Qualität des Record Linkage.

Schlüsselwörter: Record Linkage, Krebsregister, Evaluierung, verschlüsselte Daten, Datenqualität

Introduction

In Germany, no unique identity number can be used for linking records in cancer registries. Personal identity data must be used instead. To protect identity data, registries store them only in encrypted form and record linkage is performed exclusively using encrypted identifiers. The basic principles of this cancer registration model have been described before [1].
The population based cancer registry of North Rhine Westphalia (Epidemiologisches Krebsregister NRW, short EKR NRW) is one of the most recent German registries having been in operation since July 2005. By law, it registers all cases of cancer in the German federal state of North Rhine Westphalia (NRW), covering a population of almost 18 million [2]. The registration scheme builds on previous experience gained locally in the former cancer registry for the Münster region and in Germany in general since the 1990s. The Münster cancer registry was incorporated into the new registry in 2005. One distinctive feature of the EKR NRW is that cases are notified to the cancer registry exclusively electronically, using secure data transmission. Another distinctive feature is that the software for stochastic record linkage, based on the Fellegi-Sunter model [3] and implementing the German standard for record linkage in cancer registries [4], is integrated in the registry's database system developed by the EKR NRW itself. In this study, we investigate the actual accuracy of the record linkage procedures used in this cancer registry.

Scientific background

The most common methods used for record linkage are stochastic methods; their main feature is to assign probability-based weights to pairs of records. These weights are derived from the probabilities of record agreement given that the records in a pair pertain to one person or to two persons. A high weight corresponds to a high probability that the two underlying records represent the same person. Two thresholds are chosen such that record pairs with weights exceeding the first threshold are considered to pertain to one person; record pairs with weights below the second threshold are assumed to represent two persons [3]. Pairs of records with weights between the two thresholds are marked for a clerical review. If both thresholds are equal, record linkage can be fully automatized.
One important issue when performing record linkage is to minimize both homonym and synonym errors. Homonym errors occur if records are erroneously linked (false positives); this leads to underestimation of the true number of cases [5]. When linking cancer registry records with death certificates, homonym errors cause overestimation of mortality [6]. Synonym errors occur if records are erroneously not matched although they pertain to one person (false negatives) resulting in overestimation of the true number of cases and underestimation of mortality. Therefore, it is necessary that record linkage in a cancer registry minimizes the amount of errors, thereby enabling unbiased estimation of epidemiological measures.
Previous investigations on record linkage with encrypted data in the former Münster cancer registry found homonym errors to be 0.5% and synonym errors to be 2% [7]. This is generally considered as being sufficiently low (<<5%), i.e. incidence and survival estimates of such a registry are approximately unbiased.

Objectives of study

The laws governing the EKR NRW stipulate that it has to evaluate its procedures [2]. We report on the evaluation of the record linkage procedures, as conducted by an independent evaluation group.
The objective of this study was to determine the frequency of record linkage errors in the EKR NRW, thereby providing insights into the quality of stochastic record linkage in a real-world application with iterative insertions into a large database. We project the results to an ever-increasing

database because homonym error rates increase with the size of the cancer registry and synonym error rates increase with the number of records per person [8].

Notification process and record linkage in the EKR NRW

At the time of evaluation, the EKR NRW received notifications of incident cancer cases from physicians directly and from the oncological quality assurance databases. Further, pathology laboratories provided pathology reports, and death records were obtained for mortality follow-up from population registers. Specialized software has been made available for physicians to notify cancer cases to the EKR NRW.

1. Each new case entered by a physician receives a unique record ID derived from the sequence number, an ID unique to the notifying person or body (e.g. physician or hospital department), and a time-stamp. This unique record ID (URID) is used to link separately transmitted portions of data relating to the same notification. Records are processed as follows (see Figure 1).

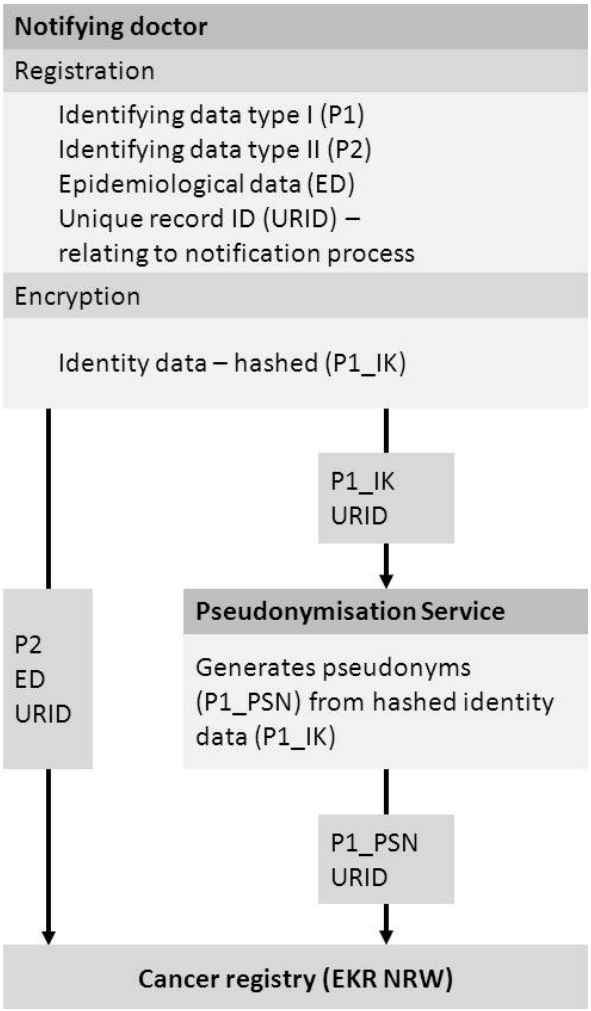

Figure 1: Encryption data flow [29]

2. Part of the patient identity information to be used for record linkage is not allowed to be stored in plaintext (P1: first name, family name, former names, day of birth; street and house number of residence), another part must be kept in plain text (P2: sex, month and year of birth, zip code and place of residence).
   Part P1 is encrypted in two ways, first as a single string using an asymmetric encryption algorithm. This encrypted information, the plain text identifiers (P2), the epidemiological tumor data (ED), and the unique record ID are transmitted to the cancer registry directly. The asymmetric key pair was generated by an external trusted authority, namely the Medical Association (Ärztekammer Westfalen-Lippe) which also safeguards the private key. Data are encrypted using the "public" key and can be only decrypted using the private key. The encrypted information is stored to allow for possible future decryption (such as the present study), but is not used for record linkage.
3. In the second encryption, pseudonyms (P1_PSN) for each individual component of the identity information P1 are generated; the principle has been described in [9]: Names are partitioned into components (up to three per name), a phonetic algorithm [10] is applied to the components. Using the MD5 hash algorithm hashed versions of each ID variable (P1_IK) are obtained. Following this step, the ID hashes and the unique record ID are transmitted to a pseudonymization service which is provided by the regional Association of Statutory Health Insurance Physicians (Kassenärztliche Vereinigung Westfalen-Lippe). There, they are encrypted into a pseudonyms (P1_PSN) using a symmetric key, which is only known to the pseudonymization service. This symmetrical encryption largely prevents "dictionary based" decryption attempts.
4. The pseudonymization service transmits the unique record ID and pseudonyms (P1_PSN) to the EKR NRW.
5. EKR NRW merges plain text personal identifiers P2, the epidemiological data ED and pseudonyms P1_PSN using the unique record ID, thereby re-assembling the separately transmitted portions of each notification.

Notifications from oncological quality assurance databases are processed in batch. However, the data fields transmitted to the cancer registry and encryption procedure are identical.
Every night, new notifications are linked to the registry database. Stochastic record linkage is performed using software developed within the EKR NRW using SQL Windows on a SQLBase database. The algorithm described in [4] which is the standard for German cancer registries has been implemented. Matching variables are: sex, administrational code of place of residence, month and year of birth in plain text and pseudonyms based on the variables surname, given name, and name at birth, and day of birth. Frequency attacks using publicly available information about the frequency distribution of identifiers as described in [11] are unlikely to be successful here be-

cause the frequency distribution of identifiers of cancer patients is different from that of the general population but unknown to the cancer registry staff. Furthermore, strong organizational measures are in place to prevent attacks that are conceivable in theory.
In order to avoid comparing every new record with every other record in the database, which would result in many unnecessary comparisons, "blocking" is applied, where a set of "blocking variables" must match perfectly between two records in order that pairs of records are compared. Seven passes with different sets of blocking variables are performed. Pseudonyms (denoted by PSN(·)) and plain text blocking variables in each pass are

1. PSN(phonetic code of surname), PSN(phonetic code of given name), PSN(day of birth), month of birth, year of birth;
2. PSN(phonetic code of surname), sex, administrational code of place of residence;
3. PSN(phonetic code of given name), PSN(day of birth);
4. PSN(phonetic code of given name), month of birth;
5. PSN(phonetic code of given name), year of birth;
6. PSN(day of birth), month of birth, year of birth;
7. PSN(phonetic code of surname).

Pairs of records above an upper threshold are linked automatically; pairs of records below a lower threshold are not linked. Matching thresholds have been adapted over time. The remaining records are subjected to clerical review. This is based on tables containing the pseudonyms for first name, family name, former names, day of birth, and the identifiers available in plain text (sex, month and year of birth, zip code and place of residence). For each pseudonym a single letter is displayed such that identical pseudonyms have the same letters. Further, all diagnostic information from the compared records and the matching weights is shown.
Besides blocking, qualitative linkage rules are applied: new records referring to a living patient are never linked to records that refer to deceased patients who died at least three months before the diagnosis of the new case. Further, records referring to patients whose death occurred at least five years before the current data are excluded from record linkage with new notifications.

Materials and methods

Study design

To ensure confidentiality and neutrality, the evaluation was performed by an independent evaluation group at the University Medical Center in Mainz. No EKR NRW personnel had access to the study datasets. A sample of cancer registry records was drawn from the EKR NRW. A dataset containing non-encrypted identity data was obtained and used to generate a gold standard to which the results of the record linkage of the EKR NRW were compared.

Data acquisition

In September 2008 the EKR NRW drew a random sample of 100,000 records of patients notified to the registry from 2006 onwards, subject to the constraint that 50% of the records should be notifications from pathologists, 35% from oncological quality assurance and 15% general notifications of incident cases. Further, the EKR NRW provided 50,000 death notifications which it had received from population registers. 11.2% of these death notifications had been linked in the EKR NRW to cases in the random sample of notifications. This reflects the contributions of the different sources to the cancer registry and the proportion of death notifications linked to registered cases at the time of the evaluation.
The EKR NRW generated a unique sequence number for each record in the sample. It forwarded epidemiological data and sequence numbers to the evaluation group and sent the sequence numbers and encrypted ID data to the external trusted authority. The trusted authority decrypted patient identifiers and transmitted plain text patient identity data and sequence number to the evaluation group, which recombined epidemiological data and patient identity using the sequence number.
Permission for this decryption and data transmission had been obtained from data protection agencies. All data transmissions either took place within a restricted safe network environment or using additional transport encryption. At no time plain text patient identity data were available to anybody outside the evaluation group. Persons involved with plain text data in the clerical review of the record linkage were obliged to ensure strict confidentiality.
The variables contained in the database were as follows:

- unique sequence number
- identity data: surname, first name, title, name at birth (maiden name), other previous name, sex, date of birth (day, month, year), address (street, number, post code, place of residence), nationality
- in the case of deceased patients: date of death, place of death, cause of death, vital records office registering the death
- medical information: date of diagnosis, tumor diagnosis (ICD-10 code), morphology and site (ICD-0-3 codes)
- a unique Person ID assigned to all records pertaining to one person as identified by EKR NRW

Generating the gold standard

The data then were subjected to data cleansing and preprocessing. Inconsistent spelling of places of residence was detected and corrected, additions to names and to places of residence were removed, obvious errors typical for automatically scanned documents (e. g. "5" instead of "S") were corrected. Similar to the routines described above, names were separated in up to three parts. Further, phonetic codes (Kölner Phonetik, similar to the

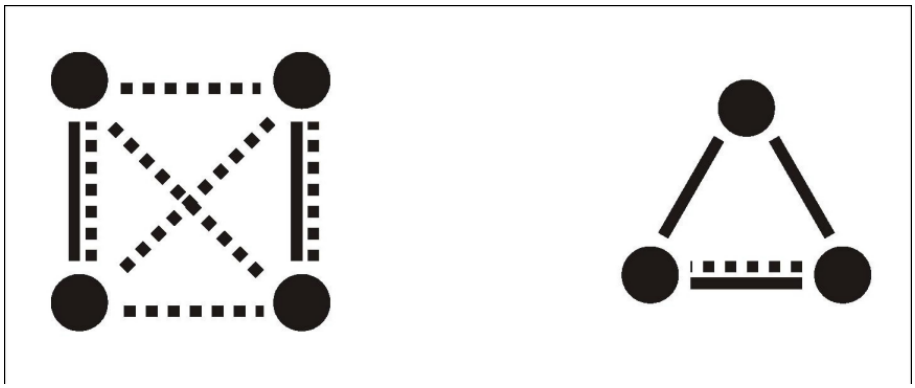

Figure 2: Record linkage results represented as graphs. Nodes represent records, matches correspond to edges. Solid edges represent matches in the gold standard; dotted edges represent matches determined in the EKR NRW. Left: four records pertaining to two persons that were considered pertaining to one person in the cancer registry (homonym error). Right: three records for one person that were considered as two records for one person and one record for another person (synonym error).

Soundex algorithm, but more suitable for German names) [9] were generated for all name components.

As a next step, a gold standard of links was created using three stochastic record linkage software packages based on the Fellegi-Sunter model [3], i.e. Automatch 4.3 [12], MTB 0.61 [13], and a stochastic record linkage based on an EM algorithm as implemented in the R-package RecordLinkage [14], [15], [16]. Each method was applied to find duplicates among the cancer notifications (deduplication) and pairs of matching incident cases and death certificates (matching).

The primary purpose of using several record linkage programs and several string distance measures was to obtain different decision boundaries. These were combined in order to safeguard decisions concerning matches, non-matches and a comprehensive dataset for the clerical review of possible matches.

A blocking strategy was devised using the different combinations of "blocking variables". The subsets of blocking variables were chosen such that no single variable was contained in all sets of blocking variables, thus allowing for errors in each variable. For the linkage with Automatch and MTB, ten passes with different sets of blocking variables were used, i.e.

1. Phonetic code surname, Phonetic code given name, Day of birth, Month of birth, Year of birth;
2. Phonetic code surname, Day of birth, Month of birth, Year of birth;
3. Phonetic code surname, Sex, Post code;
4. Phonetic code surname, Sex, Place of residence;
5. Day of birth, Month of birth, Year of birth, Place of residence;
6. Phonetic code given name, Day of birth;
7. Phonetic code given name, Month of birth;
8. Phonetic code given name, Year of birth;
9. Day of birth, Month of birth, Year of birth;
10. Phonetic code surname, Sex.

The blocking strategies for the method using the EM-algorithm had been formulated independently; therefore only passes 1, 5–8, 10 were applied. The matching variables were first name, family name, sex, day, month and year of birth, zip code. They were used in all passes and with all methods. When generating the gold standard, the blocking strategy differed from the strategy applied in the EKR NRW. The administrational code was not contained in the transmitted data; therefore, two passes were used, replacing the administrational code with post code and place of residence respectively. Pass 2 was added and sex was added in the last pass to reduce the number of overly large blocks that could not be processed by Automatch.

When linking the data with Automatch, a positive weight was only assigned when exact agreement with respect to a matching variable was observed. When linking data with MTB we used exact agreement, Levenshtein Damerau distance [17], [18] and bigrams [19]; in the method using an EM-algorithm fuzzy matches were also used [20].

Each program's output consisted of pairs of record numbers and a weight that was classified as "certain match", "potential match" or "non-match". The weights computed by the record linkage software were based on the Fellegi-Sunter model [3]; thresholds were determined after inspection of the results. The single records can be represented as nodes of an undirected graph; and matches correspond to edges connecting these nodes. The connected components of this graph are the records that possibly pertain to the same person (Figure 2).

All 15,146 groups of records that were identified as possibly pertaining to one person by at least one of the programs were included in a clerical review, amounting to a total of 34,633 records. There were few differences between the three programs in the pairs of records that were identified as possible matches. The EM-algorithm with Levenshtein distance found at most 5 records more per pass – amounting to 0.0007% overall [21].

The identity data were supplemented by date of diagnosis, date of death and diagnosis (ICD code). Difficult decisions were discussed in the group of reviewers. The decisions resulting from the clerical review were entered into a database and compared with the cancer registry decision; in some cases the decision was revised. A sample of 4,000 pairs of records that had received low matching

weights and had automatically been classified as non-matches for the gold standard also was reviewed.
We then compared the results of EKR NRW's record linkage to our gold standard by comparing the number of persons and the number of records (notifications) identified as pertaining to each person.
In order to assess the quality of the record linkage and to compare the results of the current study to previous investigations on record linkage quality, we computed homonym and synonym error rates. Denote by N_H the number of homonyms, N_S the number of synonyms, N_G the number of persons in the gold standard, and N_{GP} the number of persons in the gold standard with more than one record. The homonym error rate is given by $H=N_H/N_G$. The synonym error rates are $S_1=N_S/N_G$ and $S_2=N_S/N_{GP}$. Whereas S_1 measures the effect of synonym errors on the cancer registry, S_2 measures the quality of the record linkage procedure. We further computed precision (positive predictive value) prec=(true matches/(true matches + homonyms)) and recall (sensitivity) rec=(true matches/(true matches + synonyms)) and the F-measure F=2(prec*rec)/(prec+rec) [22]. The relative net error $ne=(N_S-N_H)/N_G$ measures the impact of record linkage errors on epidemiological measures.

Projection to larger databases

The homonym and synonym error rates depend on the number of records in the registry, the number of records per case, the probability of erroneously linking unrelated records and the probability of erroneously not linking records pertaining to the same person. As the cancer registry database increases over time it was desirable to project error rates to larger databases. One step in this projection was to obtain estimates for the probabilities *h* and *s* where
h = P(pair of records is linked | records pertain to two different persons) and
s = P(pair of records is not linked | records pertain to one person).
The formulae are given in the appendix.

Results

After establishing the gold standard, we found that the 150,000 records pertained to 132,267 persons, for 118,218 persons there was exactly one record, 14,049 persons had more than one record in the sample; this resulted in 31,782 records having a least one matching record. So with 34,633 records entering in the clerical review a proportion of 91.8% was found to have matching records. None of the 4,000 groups of records with low weights actually contained any matches. The record linkage performed by the EKR NRW yielded 132,515 persons of which 13,841 had more than one record. Hence the net error was 248 cases resulting in 0.19% overestimation of cases. Table 1 displays the distribution of multiple records in detail.

Synonym errors and homonym errors

While it is possible that synonym and homonym errors occur simultaneously, i.e. a record that should be linked to record(s) of person A is instead linked to record(s) of person B, this was not observed in this study. The differences in linkage decisions between the routine in the EKR NRW and the study are shown in Table 2 and Table 3. Synonym errors lead to 268 additional cases whereas 20 cases were missed due to homonym errors
Missing a link of two records for one person gave rise to 223 synonyms. Synonym errors were observed for 38 persons with three records; in 37 instances two records had been linked and the link of the third had been missed (2+1), in one instance three records had been allocated to three different persons (1+1+1) giving rise to 2 synonyms.
Hence the synonym error rates were $S_1=N_S/N_G=268/132,267=0.2\%$; i.e. due to synonym errors the number of cases was overestimated by 0.2%. $S_2=268/14,049=1.91\%$, i.e. the proportion of synonym errors among the cases with more than one record was 1.91%. The recall was rec=22,252/22,571=98.59%, i.e. 98.59% of the record pairs that should have been linked, were indeed linked.
When records of different persons were linked erroneously one homonym error occurred in 20 cases, i.e. in these cases a single record or a pair of correctly linked records was linked to one or more records referring to one different person. No more complex homonym errors were found. The resulting homonym error rate was $H=N_H/N_G=20/132,267=0.015\%$, i.e. due to homonym errors the number of cases was underestimated by 0.015%. The precision was prec=22,252/22,287= 99.84%; i. e. 99.84% of the matches were correct.
From precision and recall the F-measure was computed as F=0.9921.

Projection of results to a larger registry

Currently approximately 140,000 new cases of cancer occur in NRW per year. On average 2.65 notifications per case are received by the cancer registry. Assuming constant incidence and constant notification behavior, after 20 years approximately 7,420,000 notifications are to be expected.
We estimated the average probability that a homonym error occurs for an arbitrary pair of records pertaining to two different persons to be $h=3.111*10^{-9}$. The homonym error rate increases linearly with the number of records, leading to a projected homonym rate of 1.15% in a registry with 7,420,000 records.
We estimated the probability of an arbitrary pair of records for one person not to be linked to be s=0.008458. We projected the synonym error rate to various scenarios for the distribution of multiple notifications; the results are given in Table 4.

Table 1: Linkage results – distribution of multiple records per person

	Linkage results gold standard			Linkage results EKR NRW		
Number of records for one person (i)	Number of persons with i records (n_{Gi})	Proportion of persons with i records (n_{Gi} /N_G)	Number of records = $i \cdot n_{Gi}$	Number of persons with i records (n_{Mi})	Proportion of persons with i records (n_{Mi} /N_{Mi})	Number of records = $i \cdot n_{Mi}$
1	118,218	89.38%	118,218	118,674	89.56%	118,674
2	11,243	8.50%	22,486	11,073	8.36%	22,146
3	2,131	1.61%	6,393	2,098	1.58%	6,294
4	524	0.40%	2,096	518	0.39%	2,072
5	115	0.09%	575	115	0.09%	575
6	24	0.02%	144	24	0.02%	144
7	9	0.01%	63	10	0.01%	70
8	2	0.00%	16	2	0.00%	16
9	1	0.00%	9	1	0.00%	9
Total	132,267		150	132,515		150,000

Table 2: Synonym errors

Number of records per person in gold standard	*Allocation of records to person(s) by EKR NRW*	*Number of persons concerned*	*Number of records concerned*	*Number of synonym errors*
2	1+1	223	446	223
3	2+1	37	111	37
3	1+1+1	1	3	2
4	3+1	5	20	5
4	2+2	1	4	1
Total			854	268

Table 3: Homonym errors

Allocation of records to person(s) in gold standard	*Number of records per person in EKR NRW*	*Number of groups of linked records*[1]	*Number of records concerned*	Number of homonym errors
1+1	2	16	32	16
2+1	3	1	3	1
3+1	4	1	4	1
4+1	5	1	5	1
5+2	7	1	7	1
Total			51	20

[1] group of linked records = records deemed to pertain to one person by EKR NRW

Table 4: Projection of synonym error rates to various distributions of multiple notifications

s		μ	q_1	q_2	q_3	q_4	q_5	S_1	S_2
0.008458		1.13	89.38%	8.50%	1.61%	0.40%	0.09%	0.20%	1.91%
0.008458		1.26	78.76%	17.00%	3.22%	0.79%	0.17%	0.41%	1.91%
0.008458		1.40	68.13%	25.50%	4.83%	1.19%	0.26%	0.61%	1.91%
0.004229		1.13	89.38%	8.50%	1.61%	0.40%	0.09%	0.10%	0.96%
0.004229		1.26	78.76%	17.00%	3.22%	0.79%	0.17%	0.20%	0.96%
0.004229		1.40	68.13%	25.50%	4.83%	1.19%	0.26%	0.30%	0.96%
Poisson distribution of number of notifications per patient									
s	λ	μ	q_1	q_2	q_3	q_4	q_5	S_1	S_2
0.008458	1.0	1.6	58.2%	29.1%	9.7%	2.4%	0.5%	0.84%	2.02%
0.008458	1.6	2.0	40.5%	32.4%	17.3%	6.9%	2.2%	1.35%	2.27%
0.008458	2.0	2.3	31.3%	31.3%	20.9%	10.4%	4.2%	1.69%	2.46%
0.008458	2.2	2.5	27.4%	30.2%	22.1%	12.2%	5.4%	1.86%	2.56%
0.008458	2.8	3.0	18.1%	25.4%	23.7%	16.6%	9.3%	2.36%	2.89%
0.008458	3.0	3.2	15.7%	23.6%	23.6%	17.7%	10.6%	2.53%	3.00%
0.004229	1.0	1.6	58.2%	29.1%	9.7%	2.4%	0.5%	0.42%	1.01%
0.004229	1.6	2.0	40.5%	32.4%	17.3%	6.9%	2.2%	0.68%	1.14%
0.004229	2.0	2.3	31.3%	31.3%	20.9%	10.4%	4.2%	0.85%	1.23%
0.004229	2.2	2.5	27.4%	30.2%	22.1%	12.2%	5.4%	0.93%	1.28%
0.004229	2.8	3.0	18.1%	25.4%	23.7%	16.6%	9.3%	1.18%	1.44%
0.004229	3.0	3.2	15.7%	23.6%	23.6%	17.7%	10.6%	1.26%	1.50%

q_i=Number of patients with i notifications; for definitions of the parameters λ and μ, see text.

In some scenarios we simply assumed that the proportions of cases with more than one record were doubled or tripled. In other scenarios we assumed that the distribution of the number of records per person followed a zero-truncated Poisson distribution with parameter λ and mean μ=λ/(1-exp(-λ)).

Apart from the s estimated from the data, we also projected error rates for the case where s can be halved, e.g. by introducing more rigorous procedures for standardizing data entry.

With an increasing proportion of persons with multiple records the synonym error rate increases. If on average there are three records per person S_1 increases to 2.36%. If the error probability can be halved and the average number of notifications is 2.5 S_1 will be less than 1.0%.

Discussion

There are only a few studies evaluating the performance of a record linkage approach based on a sophisticated process of creating a gold standard for a large database. Duvall et al. [23] linked data from a health care provider to the Utah population database. They found a minimum accuracy of 96.3%. Joffe et al. [24] performed a benchmark comparison in which samples of records from a large clinical data base where de-duplicated. They found that optimized deterministic record linkage procedures might be superior to probabilistic record linkage, particularly in the amount of record pairs left for clerical review. Giersiepen et al. [25] studied the linkage of cancer registry data, which were encrypted as usual in Germany, with data from a mammography screening program. Their

procedure to obtain the gold standard was similar to ours. While they found the F-measure to be 96.7% when record linkage thresholds were chosen optimally, we obtained F=99.2%. This is due to the fact that both precision (99.8% versus 98%) and recall (98.5% versus 95.9%) were somewhat higher in the current study than in the study by Giersiepen et al. [25]. Errors in case numbers of 1% or less are generally deemed acceptable in a cancer registry context. Both, the homonym and synonym error rates observed here (H=0.015% ; S_2=1.91%) were somewhat lower than in a previous study using data of the Münster cancer registry (H<0.5%, S_2=2% [7]). The previous study was considerably smaller, covering the Münster area which is only part of the area now covered by the EKR NRW. Furthermore, only data of patients who had agreed to be registered were used. This result indicated that care had been taken to improve record linkage procedures when implementing the new cancer registration scheme.

The projection to larger databases yielded homonym error rates close to 1% when the number of records is about 7.5 million, a size of the database that will take approximately 20 years to reach. Most cancer patients die within 15 years after their diagnosis; patients who died five or more years before record linkage are excluded from record linkage. Therefore it is unlikely that more than 7.5 million records are included in any record linkage and hence the implemented record linkage procedure is adequate in terms of record linkage errors.

Projection of synonym error rates showed that the synonym rate remained below 1% (in relation to the total number of cases) while the number of notifications per case was 1.6 or less. If the number of notifications per case increases the synonym error rate is more likely to be around 2% given the current data quality. In order to obtain synonym rates below 1% even with increasing numbers of records per case, some improvement regarding the standardization of data entry is indicated.

In the present study, the number and nature of corrections to data from each source of notifications to the registry was documented. There is a clear relationship between the quality of the initial data and the number of synonym errors. Therefore, it is very important to ensure good quality of data (standardized data input, e.g. using patient cards, coded diagnosis etc.) in order to obtain good record linkage quality on the output side, all the more with encrypted data. The small number of differences between the gold standard and the results of EKR NRW shows that a thorough preprocessing was implemented.

This study has some limitations: First, as in many other record linkage studies, even the gold standard does not necessarily give the correct linkage decision. Second, this study did not compare software solutions and it did not subject the EKR NRW software to a benchmark comparison. While this would be an interesting exercise and might help the cancer registry to improve its record linkage performance further, the focus of this investigation was to evaluate the actual performance of the implemented record linkage procedure. Further, this study did not investigate errors due to incompleteness, i.e. some cases of cancer are never reported to the registry. Errors in case numbers due to incompleteness are typically up to 5% or 10% even in good cancer registries. Compared to the size of these deficits, the errors in case numbers caused by record linkage inaccuracy were found to be small. However, the strength of this study is that a large stratified random sample of cancer registry data could be analyzed and that plain text data were available to generate the gold standard.

Since the inception of the registry privacy-preserving record linkage using other methods has been studied by several authors. Schnell et al. and Randall et al. [26], [27] have shown that privacy preserving record linkage using Bloom filters gives good record linkage results, which are even better than those obtained from phonetic encoding.

Multiparty strategies have been suggested e.g. by Pal et al. and Kum et al. [28], [29]. However, given the already complex notification process and the necessity to name all participating parties in the cancer registration law, an implementation currently is not feasible.

However, these results may be considered in future refinement of registry procedures.

A minor disadvantage is that only death notifications, but no death certificates were available. The data quality of death certificates is not always good and synonym error rates might be higher if death certificates were included.

With the introduction of the electronic health insurance card in 2012, every German citizen obtains a unique lifelong health insurance ID. An amendment to the cancer registry legislation in North Rhine Westphalia was passed in 2013 [30] that allows to include this ID in encrypted form in the cancer registry data. The automatic transfer of ID data from the health insurance card to the registration system and the use of the health insurance ID will tremendously reduce the number of record linkage errors. To the best of the authors' knowledge, North Rhine Westphalia is the first federal state to include the health insurance ID in the cancer registry.

Conclusion

While many cancer registries are still – in part or completely – based on paper notifications, the EKR NRW has established a notification system that is entirely based on electronic notification of records, involving also existing databases like resident registries, hospital information systems, tumor documentation systems, and alike. The results of this study show that these techniques are able to process large amounts of data with very high quality of record linkage.

The results of this study may be generalized to other cancer registries using similar record linkage techniques as well as epidemiological studies linking cohort data to cancer registries, given that the privacy requirements are

met for the application at hand and the data quality is sufficient.

Notes

Competing interests

The authors declare that they have no competing interests.

Authors' contribution

The study was planned by IS, HWH and GPH, IS and GPH analyzed the record linkage results. IS drafted the original manuscript to which all others made edits and improvements. VK was involved in designing the cancer registry database, pre-processed the data and made them available for further analysis. AGA and GPH standardized the received data for record linkage. GPH, MS and AB applied the different record linkage methods to obtain the gold standard. IS devised the formulae to compute error rates. IS and OH performed the projection to larger registries. All authors read and approved the final manuscript.

Acknowledgement

We thank the colleagues who performed the clerical review: Monika Decher-Neff, Martina Hick, Susanna Siebert, Manuel Sudhof, Claudia Trübenbach, Franziska Wandtner.

Attachments

Available from
http://www.egms.de/en/journals/mibe/2016-12/mibe000164.shtml

1. mibe000164_Appendix.pdf (151 KB)
   Appendix: Derivation of formulae for projection to larger database

References

1. Michaelis J, Miller M, Pommerening K, Schmidtmann I. A new concept to ensure data privacy and data security in cancer registries. Medinfo. 1995;8 Pt 1:661-5.
2. Gesetz zur Einrichtung eines flächendeckenden bevölkerungsbezogenen Krebsregisters in Nordrhein-Westfalen (EKR-NRW). Vom 5. April 2005. Gesetz- und Verordnungsblatt für das Land Nordrhein-Westfalen. 2005 May 4;(19):372-426.
3. Fellegi IP, Sunter AB. A Theory for Record Linkage. J Am Stat Assoc. 1969;64(328):1183-210. DOI: 10.1080/01621459.1969.10501049
4. Meyer M. Kontrollnummern und Record-Linkage. In: Hentschel SK, Alexander, editors. Das Manual der epidemiologischen Krebsregistrierung. München: Zuckschwerdt; 2008. p. 57-68.
5. Brenner H, Schmidtmann I. Effects of record linkage errors on disease registration. Methods Inf Med. 1998 Jan;37(1):69-74.
6. Brenner H, Schmidtmann I, Stegmaier C. Effects of record linkage errors on registry-based follow-up studies. Stat Med. 1997;16(23):2633-43. DOI: 10.1002/(sici)1097-0258(19971215)16:23<2633::aid-sim702>3.0.co;2-1
7. Krieg V, Hense HW, Lehnert M, Mattauch V. Record Linkage mit kryptografierten Identitätsdaten in einem bevölkerungsbezogenen Krebsregister. Entwicklung, Umsetzung und Fehlerraten [Record linkage with cryptographic identification data in a population-based cancer registry. Development, implementation and error rates]. Gesundheitswesen. 2001 Jun;63(6):376-82. DOI: 10.1055/s-2001-15686
8. Brenner H, Schmidtmann I. Determinants of homonym and synonym rates of record linkage in disease registration. Methods Inf Med. 1996 Mar;35(1):19-24.
9. Appelrath HJ, Michaelis J, Schmidtmann I, Thoben W. Empfehlungen an die Bundesländer zur technischen Umsetzung der Verfahrensweisen gemäß Gesetz über Krebsregister (KRG). Informatik, Biometrie und Epidemiologie in Medizin und Biologie. 1996;27:101-10.
10. Postel HJ. Die Kölner Phonetik. Ein Verfahren zur Identifizierung von Personennamen auf der Grundlage der Gestaltanalyse. IBM Nachrichten. 1969;19:925-31.
11. Kuzu M, Kantarcioglu M, Durham EA, Toth C, Malin B. A practical approach to achieve private medical record linkage in light of public resources. J Am Med Inform Assoc. 2013 Mar-Apr;20(2):285-92. DOI: 10.1136/amiajnl-2012-000917
12. Jaro MA. Advances in Record-Linkage Methodology As Applied to Matching the 1985 Census of Tampa, Florida. J Am Stat Assoc. 1989;84(406):414-20. DOI: 10.1080/01621459.1989.10478785
13. Schnell R, Bachteler T, Reiher J. MTB: Ein Record-Linkage-Programm für die empirische Sozialforschung. ZA-Information / Zentralarchiv für empirische Sozialforschung. 2005;56:93-103.
14. Borg A, Sariyar M. Record Linkage. Record Linkage in R. 0.4-1. ed. 2013.
15. Espeland M, Odoroff C. Algorithms for computing maximum likelihood estimates from incomplete discrete data. Rochester: University of Rochester, Statistical Department; 1984.
16. Sariyar M, Borg A. The Record Linkage Package: Detecting Errors in Data. The R Journal. 2010;2(2):61-7.
17. Damerau FJ. A Technique for Computer Detection and Correction of Spelling Errors. Commun ACM. 1964;7(3):171-6. DOI: 10.1145/363958.363994
18. Levenshtein VI. Binary Codes for Correcting Deletion Insertion and Substitution Errors. Sov Phys Dokl. 1966;10(8):707-10.
19. Collins MJ. A New Statistical Parser Based on Bigram Lexical Dependencies. Proceedings of the 34th Annual Meeting of the Association for Computational Linguistics; 1996 June 24-27; Santa Cruz, California, USA. p. 184-91. DOI: 10.3115/981863.981888
20. Sariyar M, Borg A, Pommerening K. Evaluation of record linkage methods for iterative insertions. Methods Inf Med. 2009;48(5):429-37. DOI: 10.3414/ME9238
21. Schmidtmann I, Hammer GP, Sariyar M, Gerhold-Ay A. Evaluation des Krebsregisters NRW. Schwerpunkt Record Linkage. Abschlussbericht. 2009 [cited 2016 Sept 03]. Available from: http://www.krebsregister.nrw.de/fileadmin/user_upload/dokumente/Evaluation/EKR_NRW_Evaluation_Abschlussbericht_2009-06-11.pdf
22. Christen P, Goiser K. Quality and Complexity Measures for Data Linkage and Deduplication. In: Guillet F, Hamilton HJ, editors. Quality Measures in Data Mining. Berlin, Heidelberg: Springer; 2007. (Studies in Computational Intelligence; 43). p. 127-51. DOI: 10.1007/978-3-540-44918-8_6

23. DuVall SL, Fraser AM, Rowe K, Thomas A, Mineau GP. Evaluation of record linkage between a large healthcare provider and the Utah Population Database. J Am Med Inform Assoc. 2012 Jun;19(e1):e54-9. DOI: 10.1136/amiajnl-2011-000335

24. Joffe E, Byrne MJ, Reeder P, Herskovic JR, Johnson CW, McCoy AB, Sittig DF, Bernstam EV. A benchmark comparison of deterministic and probabilistic methods for defining manual review datasets in duplicate records reconciliation. J Am Med Inform Assoc. 2014 Jan-Feb;21(1):97-104. DOI: 10.1136/amiajnl-2013-001744

25. Giersiepen K, Bachteler T, Gramlich T, Reiher J, Schubert B, Novopashenny I, Schnell R. Zur Leistungsfähigkeit des Record-Linkage zwischen epidemiologischen Krebsregistern und dem Mammographie-Screening [Performance of record linkage for cancer registry data linked with mammography screening data]. Bundesgesundheitsblatt Gesundheitsforschung Gesundheitsschutz. 2010 Jul;53(7):740-7. DOI: 10.1007/s00103-010-1084-1

26. Schnell R, Bachteler T, Reiher J. Privacy-preserving record linkage using Bloom filters. BMC Med Inform Decis Mak. 2009 Aug 25;9:41. DOI: 10.1186/1472-6947-9-41

27. Randall SM, Ferrante AM, Boyd JH, Bauer JK, Semmens JB. Privacy-preserving record linkage on large real world datasets. J Biomed Inform. 2014 Aug;50:205-12. DOI: 10.1016/j.jbi.2013.12.003

28. Kum HC, Krishnamurthy A, Machanavajjhala A, Reiter MK, Ahalt S. Privacy preserving interactive record linkage (PPIRL). J Am Med Inform Assoc. 2014 Mar-Apr;21(2):212-20. DOI: 10.1136/amiajnl-2013-002165

29. Pal D, Chen T, Zhong S, Khethavath P. Designing an algorithm to preserve privacy for medical record linkage with error-prone data. JMIR Med Inform. 2014 Jan 20;2(1):e2. DOI: 10.2196/medinform.3090

30. Gesetz zur Änderung des Krebsregistergesetzes. Vom 5. November 2013. Gesetz- und Verordnungsblatt für das Land Nordrhein-Westfalen. 2013 Nov 22;(35):624-6.

Corresponding author:
Irene Schmidtmann
Institut für Medizinische Biometrie, Epidemiologie und Informatik (IMBEI), Universitätsmedizin der Johannes Gutenberg-Universität Mainz, Langenbeckstr. 1, 55101 Mainz, Germany, Phone: +49 6131 173951, Fax: +49 6131 172968
Irene.Schmidtmann@uni-mainz.de

Please cite as
Schmidtmann I, Sariyar M, Borg A, Gerold-Ay A, Heidinger O, Hense HW, Krieg V, Hammer GP. Quality of record linkage in a highly automated cancer registry that relies on encrypted identity data. GMS Med Inform Biom Epidemiol. 2016;12(1):Doc02.
DOI: 10.3205/mibe000164, URN: urn:nbn:de:0183-mibe0001640

This article is freely available from
http://www.egms.de/en/journals/mibe/2016-12/mibe000164.shtml

Published: *2016-06-13*