German Medical Data Sciences: Bringing Data to Life
R. Röhrig et al. (Eds.)

doi:10.3233/SHTI210057

Record Linkage in Clinical Cancer Registration: Experiences and Findings from Lower Saxony

Kira Scheel[a], Thomas Franke[a], Alisha Weikert[a], Melanie Dick[a], Anna Walter[b], Jan Zeidler[b], Tobias Hartz[a]
a Clinical Cancer Registry of Lower Saxony (KKN)
b Center for Health Economics Research Hannover

Abstract In cancer registries, record linkage procedures are used to link records of the same patient from different health care providers. In the Clinical Cancer Registry of Lower Saxony, a multi-level combination of exact assignment using the statutory health insurance number and a probabilistic procedure with control numbers and address data is applied. The procedure implemented in the register application assigns the incoming messages in this way as far as possible automatically. The aim of the observation carried out was to check the efficiency of the match variables and threshold values used, above which manual assignment is required. Weak points were identified and approaches to solutions were developed.

1. Background

The epidemiological and clinical cancer registries in Germany regularly receive reports from different health care providers on the same affected person. In the Clinical Cancer Registry of Lower Saxony (KKN), reporting is only possible electronically via manual entry or upload of an XML file from the primary system of the notifier (interface) in the web-based reporting portal. Record linkage procedures are used to uniquely assign these data records and thus to be able to display the course of a tumor disease in relation to the patient [1]. Since there is no number in the German health care system that uniquely identifies each person for life, probabilistic record linkage procedures are often used [2]. These procedures work with weights and thresholds: below a certain threshold, an assignment is rejected, and a new patient is created; above another threshold, however, the data set is automatically assigned to an existing patient [3]. Usually there is a range between the lower and upper threshold values in which no automatic rejection or assignment is made. In this case, the record link must be resolved manually by an employee - if necessary, by consulting further information.

KKN uses a combination of different record linkage methods: In the first step it is checked whether an incoming report can be uniquely assigned to a patient already in the register using the statutory health insurance number in combination with the date of birth [1, 2, 4]. If this assignment attempt fails, for example, because the patient is member in a private health insurance and does not have a statutory health insurance number, the message then undergoes a probabilistic record linkage procedure in the next step.

The probabilistic method used first determines a set of assignment candidates in which at least three of the following seven characteristics match [1]:

- Last name
- First name
- Day of date of birth
- Month of date of birth
- Year of date of birth
- zip code
- Community code (GKZ)

This upstream restriction of the number of assignment candidates is carried out for performance reasons. With several thousand incoming messages per day and, in perspective, hundreds of thousands of patients in the registry, a complete comparison (each incoming message against each patient) would be too time-consuming.

Following the identification of the assignment candidates, the nationwide uniform control numbers (UNICON) in the cancer registration system are used in combination with address data (zip code, GKZ) and the date of birth to determine match weights [1, 2, 5]. Depending on the defined threshold values, the system then decides whether the incoming notification can be assigned to a patient existing in the register (match weight with exactly one patient is greater than the upper threshold value), whether a new patient is created (the match weights of all assignment candidates are less than the lower threshold value), or whether the case must be resolved manually (the match weights lie between the threshold values).

Probabilities of conformity *m* and *u* are used to calculate the weights of conformity [6].

- *m* is the probability that a characteristic of a person (e.g. the surname) is always reported identically. Due to spelling mistakes, missing information, name changes, changes of residence, etc., this probability is never 1 in reality. *m* must be estimated individually for each characteristic. Lower Saxony currently uses the estimates adopted from the Schleswig-Holstein Cancer Registry.
- *u* is the probability that two messages match in a characteristic even though they belong to different persons. The more frequently a characteristic value occurs, the higher the probability (for example, if first names occur very frequently). For each characteristic value *a*, *u* is calculated as follows:

$$u = \frac{number\ of\ records\ with\ characteristic\ value\ a}{total\ number\ of\ records}$$

For each assignment candidate the match weight *W* is calculated with the new message. To do this, individual weights are first formed for each characteristic using the following calculation rules:

- If the two messages match in the considered characteristic *i*:

$$w_i = log_{10}\left(\frac{m}{u}\right)$$

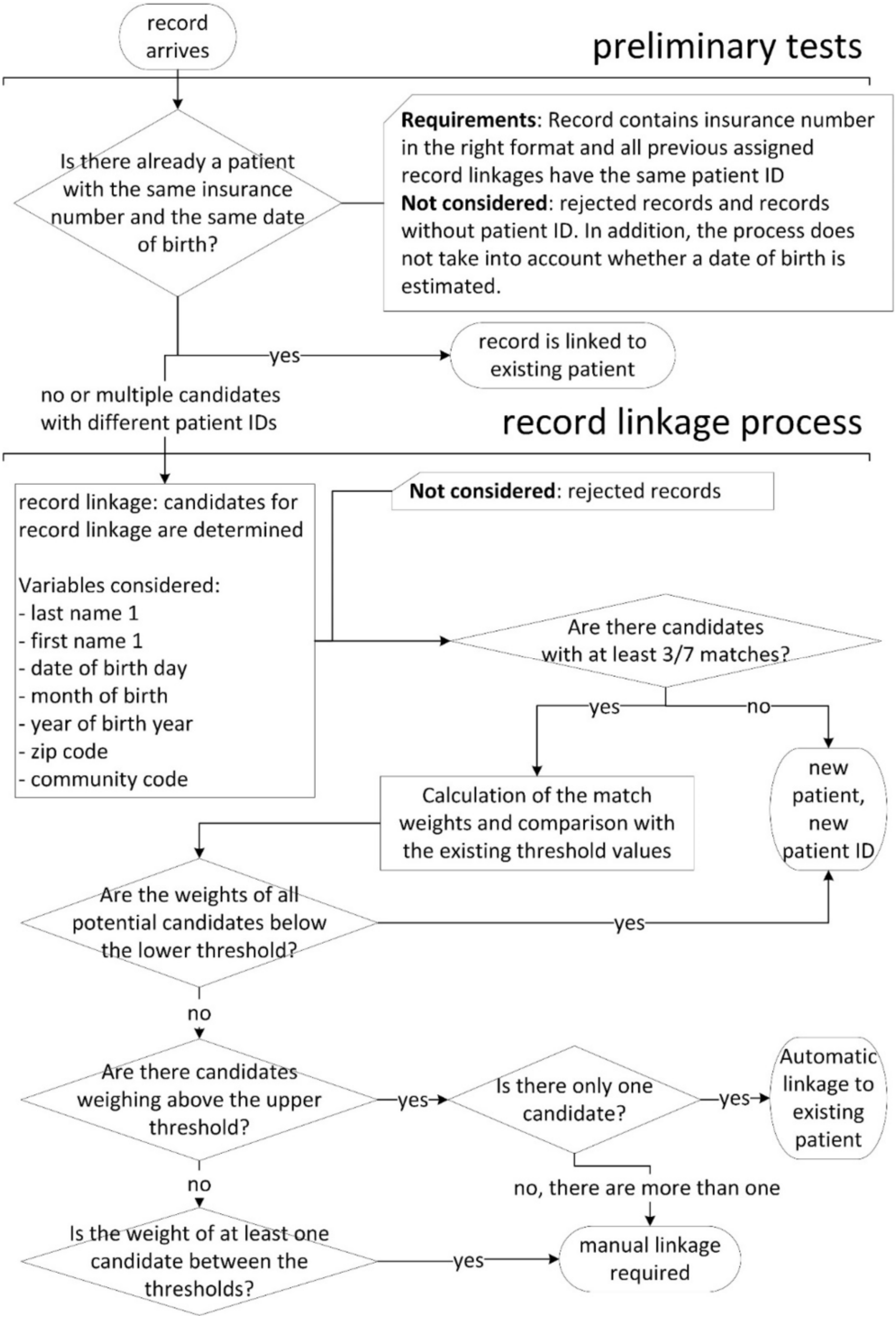
record arrives
preliminary tests
Is there already a patient with the same insurance number and the same date of birth?
Requirements: Record contains insurance number in the right format and all previous assigned record linkages have the same patient ID
Not considered: rejected records and records without patient ID. In addition, the process does not take into account whether a date of birth is estimated.
yes
record is linked to existing patient
no or multiple candidates with different patient IDs
record linkage process
record linkage: candidates for record linkage are determined
Variables considered:
- last name 1
- first name 1
- date of birth day
- month of birth
- year of birth year
- zip code
- community code
Not considered: rejected records
Are there candidates with at least 3/7 matches?
yes
no
new patient, new patient ID
Calculation of the match weights and comparison with the existing threshold values
Are the weights of all potential candidates below the lower threshold?
yes
no
Are there candidates weighing above the upper threshold?
yes
Is there only one candidate?
yes
Automatic linkage to existing patient
no, there are more than one
no
Is the weight of at least one candidate between the thresholds?
yes
manual linkage required

Figure 1 Schematic representation of the multi-stage record linkage procedure in KKN

- If the two messages in characteristic i do not match, the following applies:

$$w_i = log_{10}\left(\frac{(1-m)}{(1-u)}\right)$$

- If characteristic i is not present, the following applies:

$$w_i = 0$$

The sum of all individual weights w_i then gives the matching weight W:

$$W := \sum_{i=1}^{n} w_i$$

After one year of productive operation of the KKN, the implemented record linkage process was evaluated regarding the quality of the results. Two questions were considered:

1. Are the thresholds for the match weights of 4 (lower threshold) and 14 (upper threshold) appropriate to keep the number of manual assignments low while at the same time making as few errors (incorrect assignments) as possible? For this purpose, the hypothesis was made that the match weights of the assignment candidates in the two possible cases (creating a new patient or assigning a message to an existing patient) differ significantly from each other.
2. Where are the weaknesses of the method used? Possible problem areas were derived from the experience with the manual allocation procedure.

2. Methods

As of August 30, 2019, the data collection point of the KKN had about 3,600 reports that had been received in 2018 and required manual allocation. Between September and November 2019, this subset of reports was compared with the assignment candidates determined by the system and the assignment decision was made manually. A table was used to record whether the message could be assigned to a person already registered in the register or whether a new patient ID was assigned. The highest match weight of the proposed assignment candidates was also noted. The weights of the other assignment candidates were not recorded. The persons in charge were free to add additional observations as comments.

3. Results

The 3,600 reports were assigned to a total of 1,515 patients. Of these, new patients were created in 1,130 cases (75 percent). In 385 cases (25 percent), the notifications could be assigned to persons already registered. In 1,397 cases (92 percent), several notifications existed for one person, which were assigned within the procedure. In 20 cases, the register application could not propose an assignment candidate at the time of assignment. In 52 cases, the assignment decision could only be made with the help of a request to the residents' registration office.

The highest weight of the proposed assignment candidates was between 4.01 and 25.72 (n = 1,082, values reduced to two decimal places), whereby the highest weight of assignment candidates was only recorded for 1,082 of the 1,515 persons assigned. In the group of newly assigned patient IDs (n = 790 weights known) the highest weight was between 4.01 and 20.96. If the person already existed in the register and the new messages could be assigned, the weight of the assignment candidates was between 4.01 and 25.72 (n = 292 weights known). The distribution of the weights is shown in Figure 2 and Figure 3.

Table 1 Characteristics of the sample

Group	**Person already exists (assignment to existing patient ID)**	**Person newly registered (assignment of a new patient ID)**	**Total**
Number of patients in the sample	385	1130	**1515**
Percentage	25.41%	74.59%	**100.00%**
Number of cases where the highest weight of allocation candidates was recorded	292	790	**1082**
Minimum of the highest weight of the assignment candidates	4.01	4.01	**4.01**
Maximum of the highest weight of the assignment candidates	25.72	20.96	**25.72**

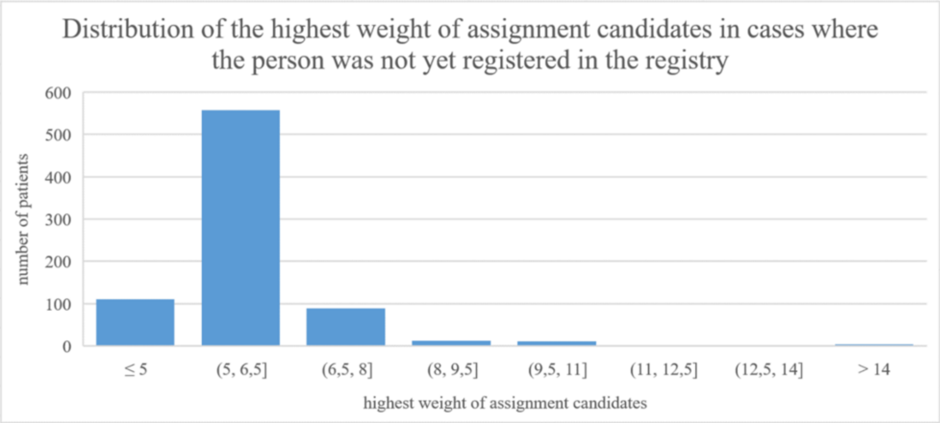

Figure 2 Distribution of the highest weight of assignment candidates in cases where the person was not yet registered in the registry and a new Patient ID was assigned

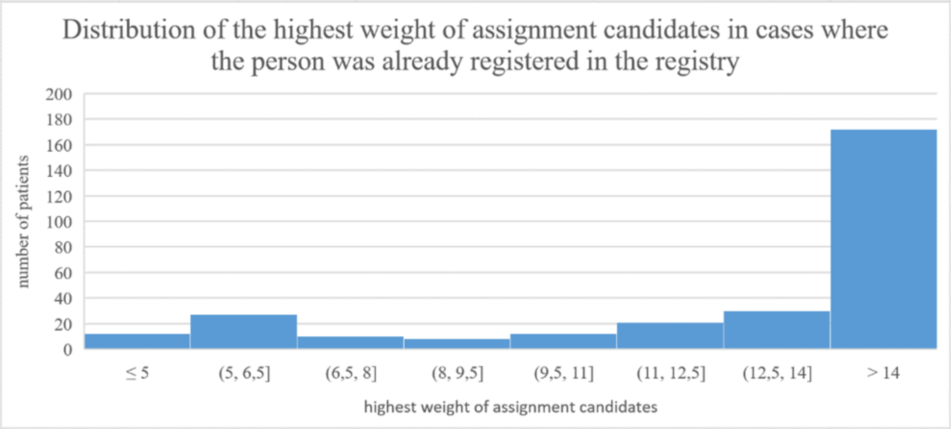

Figure 3 Distribution of the highest weight of assignment candidates in cases where the person was already registered in the registry and an existing patient ID could be assigned

3.1. Findings on the thresholds

The systematic recording of the highest weight of allocation candidates allows an initial assessment of the efficiency of the thresholds.

It is noticeable that assignment candidates with a weight above the upper threshold value were proposed for many messages, but no automatic assignment took place. Possible reasons for this situation are:

- Identification of multiple assignment candidates with a match weight above the upper threshold: A secure assignment is only possible if only one candidate with a match weight above the upper threshold is identified. As soon as several candidates with a match weight above the upper threshold value exist, a manual decision is required.
- Time difference between the date of receipt and the processing date: The information system executes the record link only when the personal data is saved; that is, when the notification is received and when personal data is changed later. If further notifications are received in the database in the meantime, this influences the relative frequencies of the proficiencies and thus the match weights but does not lead to automatic assignment. Since the sample in 2019 included messages received in 2018, it is not possible to reconstruct the original weight distribution at the time of receipt of the message and thus the first record linkage.

With the help of the histograms of the distribution of the highest matching weight of the assignment candidates, the hypothesis put forward at the beginning was tested. It was found that in cases where the person was not yet registered in the KKN and a new patient ID was issued, the majority (n = 759, 96.1 percent) of the assignment candidates had a match weight below 8. Only in six cases did the potential assignment candidate have a weight above 12 (weights from 13.98 to 20.96).

In cases where the report could be assigned to a person who was already registered, the distribution was two-sided. As formulated in the hypothesis, it was found that the majority of the assignment candidates had a weight above 12 (n = 208, 71.2 percent). In addition, a second accumulation was evident with a weight between 5 and 8. From the experience of the employees it can be deduced that the identical combination of first name, gender and date of birth resulted in many cases in a match weight of 5.39. In the case of frequently assigned first names, the relative frequency of this characteristic used to calculate the match weight increases in the database of the register application and a correspondingly large number of potential candidates for assignment are possible.

In 208 reports (71.2%) that could be assigned to an already registered person, the highest weight of assignment candidates was above 12; in the reports for which a new patient ID was assigned, only 6 (0.7%) had a candidate weight above 12. In the information system, the upper threshold value was therefore changed from 14 to 12 to reduce the effort required for manual assignments. The hypothesis that the potential assignment candidates have a lower weight when a patient ID has to be reassigned than for messages that can be assigned to a person who is already registered cannot be confirmed from the data collected due to the second peak. Therefore, it is not possible to raise the lower threshold value without incorrectly assigning new patient IDs although the persons could have been assigned.

3.2. Findings on the Record Linkage

Through the observations noted by the authors, aspects that repeatedly impaired the assignment could be identified and solution approaches derived.

1. The zip code and the community code (GKZ) are used as variables for the candidate identification. The following issues are therefore problematic:
    a. Several city names for the same zip code (city districts)
    b. Several zip codes for one place (districts)
    c. Changes to the GKZ (and zip code, if applicable) through incorporations without an actual change of residence (e.g. Kreiensen zu Einbeck)
    d. "Unknown" places (e.g. Ilsede), which are due to a lack of updating of the list of places stored in the system
    e. Location not filled in (often for messages transmitted via specific interfaces)
    f. Places with unofficial additions to place names (e.g. Winsen (Luhe), Neustadt am Rübenberge, Laatzen bei Hannover, often in the case of reports transmitted via specific interfaces)
    g. Patients resident abroad for whom no GKZ can be identified

    Approach: The list of places stored in the register application must be updated regularly. Synonymous place names and frequently used place additions should be added. In the case of interface detectors, the respective software manufacturer must ensure that the patient's place of residence is transmitted in the corresponding field and in the official designation.
2. The probability *u* is dependent on the number and diversity of the data sets in the registry and thus inherently dependent on the catchment area of the cancer registry. The same applies to the estimates of *m*. City-states have different requirements (e.g. for the characteristic place of residence, zip code, GKZ) than states with a large catchment area [1]. The cancer registries have different local and social conditions to which the estimates of *m* must be adapted. An individual analysis and adjustment per registry is necessary.

3. Estimated dates of birth: If the health care provider does not know the full date of birth of the patient, it is possible to enter it without day or without day and month. In these cases, the system will assign the month 07 and the day 15 [6]. In addition, a variable "Estimated Date of Birth" indicates which entry was estimated. This variable is not considered when determining the assignment candidates, so that assignment candidates with an "identical" date of birth are also proposed, although the date of birth was only estimated. If the variable "Estimated Date of Birth" is not considered, this may lead to incorrect assignments if employees are not familiar with the facts.
   Approach: The variable "Estimated Date of Birth" should be considered in the record linkage and highlighted in the assignment candidates. The KKN has decided to remove the option to estimate the date of birth in the reporting portal.
4. Confusion of first and last names: From time to time first and last name are mixed up.
   Approach: This circumstance should be considered in cross comparisons. During message processing, a plausibility check is carried out to alert the employees to unknown first names. The stored list should always be kept up to date.
5. Birth name corresponds to last name: Especially via interfaces the field birth name is often filled, even if it is identical with the current last name. The existence of a birth name often results in a weighting between the threshold values for the report without birth name compared to the report with birth name.
   Approach: The system could delete birth names that are identical to the current last name. At the same time, the software manufacturers of the interfaces should be informed that the birth name only needs to be transmitted if it differs from the current last name.

4. Discussion

The figures presented here are taken from an observation made during regular work in the cancer registry. They represent only a sample and were systematically recorded by only a small selection of employees. It was only during the course of the sample that the highest weight of the identified allocation candidates was begun to be recorded, which is why it is not available for all the reports considered. In most cases, several potential assignment candidates with further lower weights were also identified, which were not taken into account in this investigation.

The accumulation of messages without patient assignment is due to the processing process in the KKN: In the data collection point, messages are essentially processed in chronological order after their receipt. If a person could not be automatically assigned by the record linkage, manual assignment is carried out first for the oldest messages and with a time delay for the more recent messages. The more recent messages are not automatically assigned by the record linkage, since the record linkage process is only executed when the personal data is saved, that is, when the message is received and processed.

The 20 messages for which manual assignment was necessary although no assignment candidate was proposed can also be explained by the time interval between message receipt and message processing: Presumably, there was an assignment candidate when the message was received, whose message was deleted, rejected, or changed by the time the message was processed, so that it was no longer available when the message was processed.

In clinical cancer registries, it is important to avoid homonym errors in patient allocation [1, 4]: If information from different persons is allocated to the same patient ID, incorrect conclusions are drawn when compiling the clinical best of. Thus, incorrect information, for example, therapies not carried out on this patient, would be reported back to the treating health care providers. This can, for example, in the case of tumor board support, have a negative influence on therapy decisions. For this reason, record linkage procedures must be configured in such a way that automatic assignment only takes place with a high degree of certainty (high match weight). Inevitably, however, this also results in more cases that have to be resolved manually.

5. Conclusion

The observation contributed to a better understanding of the record linkage procedure in the KKN and revealed weaknesses. It became clear that an evaluation of the assignment weights is only possible in a meaningful way at the time the message is saved, since the record linkage is only executed when the message is saved (message receipt, message processing).

References

[1] Stegmaier C, Hentschel S, Hofstädter F, Katalinic A, Tillack A, Klinkhammer-Schalke M. Das Manual der Krebsregistrierung. 1. Auflage, W. Zuckschwerdt Verlag, München, 2019.

[2] March S, Antoni M, Kieschke J, Kollhorst B, Maier B, Müller G, Sariyar M, Schulz M, Enno S, Zeidler J, Hoffmann F. Quo vadis Datenlinkage in Deutschland? Eine erste Bestandsaufnahme. *Das Gesundheitswesen.* 80 (2018) e20-e31. doi: 10.1055/s-0043-125070.

[3] 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.* **12** (2016) Doc02. doi: 10.3205/mibe000164.

[4] March S, Andrich S, Drepper J, Horenkamp-Sonntag D, Icks A, Ihle P, Kieschke J, Kollhorst B, Maier B, Meyer I, Müller G, Ohlmeier C, Peschke D, Richter A, Rosenbusch ML, Scholten N, Schulz M, Stallmann C, Enno S, Wobbe-Ribinski S, Wolter A, Zeidler J, Hoffmann F. Gute Praxis Datenlinkage (GPD) Good Practice Data Linkage. *Das Gesundheitswesen.* **81** (2019). doi: https://doi.org/10.1055/a-0962-9933

[5] Hinrichs H. Bundesweite Einführung eines einheitlichen Record Linkage-Verfahrens in den Krebsregistern der Bundesländern nach dem KRG. Abschlussbericht, Projekt gefördert von der Deutschen Krebshilfe, Antragsnummer 70-2043-AP I, OFFIS, Oldenburg, Juni 1999

[6] Hentschel S, Katalinic A. Das Manual der epidemiologischen Krebsregistrierung. 1. Auflage, W. Zuckschwerdt Verlag, München, 2008.