© 2005 Schattauer GmbH

Record Linkage in the Cancer Registry of Tyrol, Austria

W. Oberaigner[1, 2], W. Stühlinger[2]
[1]Cancer Registry of Tyrol, Department of Clinical Epidemiology of the Tyrolean State Hospitals Ltd., Innsbruck, Austria
[2]Department of Quality Science, Medical Planning and Information Management, University for Health Informatics and Technology Tyrol, Innsbruck, Austria

Summary

Objective: Record linkage of patient data originating from various data sources and record linkage for checking uniqueness of patient registration are common tasks for every cancer registry. In Austria, there is no unique person identifier in use in the medical system. Hence, it was necessary and the goal of this work to develop an efficient means of record linkage for use in cancer registries in Austria.

Methods: We adapted the method of probabilistic record linkage to the situation of cancer registries in Austria. In addition to the customary components of this method, we also took into consideration typing errors commonly occurring in names and dates of birth. The method was implemented in a program written in DELPHI™ with interfaces optimised for cancer registries.

Results: Applying our record linkage method to 130,509 linkages results in 105,272 (80.7%) identical pairs. For these identical pairs, 88.9% of decisions were performed automatically and 11.1% semi-automatically. For results decided automatically, 6.9% did not have simultaneous identity of last name, first name and date of birth. For results decided semi-automatically, 48.4% did not have an identical last name, 25.6% did not have an identical date of birth and 83.1% did not have simultaneous identity of last name and date of birth.

Conclusions: The method implemented in our cancer registry solves all record linkage problems in Austria with sufficient precision.

Keywords

Probabilistic record linkage, cancer registry, homonym rate, synonym rate

Methods Inf Med 2005; 44: ■ - ■

Introduction

The prime objective of population-based cancer registries is to document every incident of cancer cases diagnosed in the target population [1-5]. According to international guidelines, a cancer registry should take into account various data sources containing valid information on cancer cases. Consequently, in addition to data sent to the registry by treating physicians, data sources like pathology reports, department information systems (i.e. radiotherapy) and hospital information systems must be included in the registration process. Many cancer registries analyze survival rates as the most important outcome measure, and for this analysis patient life status has to be assessed. Most registries apply a passive method, meaning record linkage between incidence data and mortality data [6].

Summing up, record linkage is a central task to be solved by cancer registries. In Austria, there is no general use of unique person identifiers as, for example, in Scandinavian countries. There is a social insurance number that is known to not be unique in all cases and it is not widely used in medical information systems. Therefore, the decision on whether data describe the same person must be based on information like last name, first name, date of birth etc. and can be time-consuming when a high degree of precision is involved. All registries aim to obtain complete and reliable information needed for patient identification, but it must be remembered that in actual practice all the components mentioned above can be distorted by (registration as well as typing) errors.

Administrative workflow in cancer registries differs in some respect from that in administrative units in hospital departments. In contrast to hospital administration, in cancer registries there is no need to register patient data immediately. Since cancer registries collect data mostly on the basis of year of diagnosis, their data collection efforts are more thorough and generally ensure good quality of data needed for record linkage.

In order to develop an efficient, scientifically founded method for record linkage, we decided some years ago to implement a method based on the theory of probabilistic record linkage and taking into account common types of error sources in the German language.

Methods

Basics

This chapter presents the basics of the theory of probabilistic record linkage to the extent needed to understand the method developed for our cancer registry. Detailed descriptions of the theory can be found for example in [7, 8].

Data in a cancer registry consist of several components describing an individual person or cancer case. One part of these components, often called person data, identifies the person. We assume that no single component uniquely identifies a person.

If a person is described by n components k_1 to k_n, we assign standardized weights to each component, i.e. w_1 to w_n, where $w_1 + ... + w_n = 1$.

Received: February 9, 2004; accepted: December 20, 2004

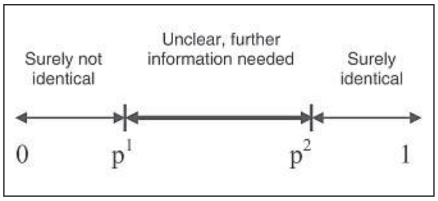

Fig. 1

For linkage of two records with components k_i^1 and k_i^2 we define p_i for each component k_i as follows:

$$p_i = \begin{cases} 1 \text{ if } k_i^1 = k_i^2 \\ 0 \text{ otherwise} \end{cases} \qquad (1)$$

This gives a sum probability defined as

$$p = w_1\, p_1 + \ldots + w_n\, p_n \qquad (2)$$

p (in the following often denoted by p probability) can be interpreted as a measure of whether two records describe the same person. Then, two cut points p^1 and p^2 are introduced with the following consequences:

- If p is smaller than p^1, it is assumed (without further checks) that the records describe different persons.
- If p is greater than p^2, it is assumed (again without further checks) that the records describe the same person.
- If p lies between p^1 and p^2, it must be decided on an individual basis whether the two records describe the same person or different persons. Usually, this means further information must be obtained.

The decision process is shown in Figure 1.

Choice of Weights

In order to choose weights according to the theory of probabilistic record linkage, two probabilities are computed, usually denoted as m and u probability.

For any component k_i, m_i is defined as the probability that k_i is equal for identical persons. ui describes the probability that k_i is equal for non-identical persons. The weight w_i is then defined by the following formula:

$$w_i = \log_2(\frac{m_i}{u_i}) \qquad (3)$$

Table 1 Transformations according to the "Kölner Transformation"

Rule	Example
Eliminate diphthongs	Wimmer → WIMER
Transform German "Umlaute"	Müller → MUELER
Transform "c" in front of "e,l" to "z"	Cicero → ZIZERO
Transform "c" in front of "a,o,u" to "k"	Cugel → KUGEL
Otherwise transform „c" to „z"	Mucke → MUZKE
Transform "v" to "f"	Vogel → FOGEL
Transform "j" to "i"	Deljc → DELIZ
Transform "ie" to "i"	Liederlich → LIDERLIZH
Transform "ai" to "ei"	Aigner → EIGNER
Transform "ae" to "e"	Jaeger → IEGER
Transform "th" to "t"	Thaler → TALER
Transform "tz" to "z"	Matzer → MAZER
Transform "d" to "t"	Danner → TANER
Delete silent "h"	Gehler → GELER
Transform "qu" to "q"	Qualler → QALER

Table 2 Standardized weight for components

Component k_i	w_i (standardized)
Phonetic transformation last name	0.22
Phonetic transformation birth name	0.202
First name	0.139
Date of birth	0.289
Sex	0.075
Zip code (or municipality code)	0.075

From the experience in our cancer registry the components were chosen as follows [9]:

- Last name
- Phonetic transformation of last name (used only if last name is not identical for the two persons under investigation), see Table 1.
- Birth name
- Phonetic transformation of birth name (used only if birth name is not identical for the two persons under investigation), see Table 1.
- First name
- Date of birth
- Sex
- Zip code (or municipality code)

The German language contains typical transformations of names following certain rules. We thus introduced the concept of phonetic transformation defined by the rules given in Table 1 (derived from the so-called Kölner Transformation, see [10, 11]).

The probabilities m_i and u_i were calculated based on results obtained before introducing the method described here, when we performed record linkage by heuristic methods and individual checks. All results were stored in a meta-relation describing pairs of data to be linked as well as linkage results. Based on this relation, it is straightforward to compute the probability m_i as follows:

$$m_i = \frac{\textit{number of patients with identical component } k_i}{\textit{number of patients}} \qquad (4)$$

In the same way, we can compute the probability ui as follows (we assume that every patient in our database is unique, hence the Cartesian product Pat × Pat (denoting all possible combinations of patients) does not contain pairs of equal patients):

$$u_i = \frac{\textit{number of patients with identical component } k_i}{\textit{number of different pairs of patients}} \qquad (5)$$

These computations gave the weights shown in Table 2.

Table 3 Additional methods

Method	Example
Left part or right part of name identical	Müller and Müller-Westernhagen
1 character wrong	Maier and Mayer
1 character missing	Maier and Mair
2 neighboring characters exchanged	Maier and Miaer

Table 4 Correction factors for weights

Component or method for component	Weight
Last name: left or right part identical	$w_{last\ name}$*0.9
Last name: 1 character wrong	$w_{last\ name}$*0.8
Last name: 1 character missing	$w_{last\ name}$*0.8
Last name: 2 characters exchanged	$w_{last\ name}$*0.8
First three digits of last name identical	$w_{last\ name}$*0.4
First name: left or right part identical	$w_{first\ name}$*0.5
Last name and birth name exchanged	$w_{last\ name}$
Date of birth: 1 character wrong	$w_{date\ of\ birth}$*0.8
Date of birth: 2 characters exchanged	$w_{date\ of\ birth}$*0.8
Date of birth: day and month exchanged	$w_{date\ of\ birth}$*0.8
Date of birth: day identical	$w_{date\ of\ birth}$*0.3
Date of birth: month identical	$w_{date\ of\ birth}$*0.3
Date of birth: year identical	$w_{date\ of\ birth}$*0.3

After detailed analysis of our database and after investigating typical errors occurring in our registry, we found that our registry contains [9] common typing errors in last name and birth name and common typing errors in date of birth.

In order to properly deal with these errors, we added the methods described in Table 3 to the components described above and consequently extended the weights given in Table 2 by the weights defined in Table 4.

For every component k_i the maximum weight is limited by the weight for this component as defined in Table 2, even if all methods defined in Table 4 add up to a greater weight.

Table 5 Results of evaluation for years 1999-2003

Number of linkages	130509	
Identical pairs	105272 (80.7%)	
Decision automatic	93627 (88.9%)	
semiautomatic	11645 (11.1%)	
Applied rules	**Decision automatic**	**Decision semiautomatic**
Last name identical	91835 (98.1%)	6014 (51.6%)
Phonetic transformation of last name identical	1692 (1.8%)	157 (1.3%
First name identical	88949 (95%)	9926 (85.2%)
Date of birth identical	93627 (100%)	8662 (74.4%)
Sex identical	91750 (98%)	10872 (93.3%)
Last name AND date of birth identical	91835 (98.1%)	3159 (27.1%)
Last name AND date of birth AND first name identical	87160 (93.1%)	1970 (16.9%)
One character rules (see Table 3) apply for last name	0	4701 (40.4%)

Choice of Critical Bounds p^1 and p^2

Our experience shows that $p^1 = 75$ and $p^2 = 95$ are good choices for cancer registries in Austria. This means that we inspect all cases with a p probability between 75 and 95 and assume without further inspection that pairs with $p \in (95, 100)$ describe the same person.

Inspection of all pairs with $p \in (75, 95)$ is a very time-consuming and tedious job. Scanning through the lists requires a great deal of concentration. However, there are usually some pairs describing the same person but with a smaller p probability (think, for example, of twins living in the same residence, perhaps with similar first names). Hence, in order to keep homonym and synonym rates low (see also the discussion on the consequences of wrong decisions) it is necessary to run through all parts of the resulting list with full concentration.

Implementation

The method described above was implemented as a program written in DELPHI. Interfaces for input are either plain text files with fields separated by "\", or Oracle tables (our cancer registry database is implemented in Oracle^{TM}). Results are written both in a plain text file and in an Oracle table. Output in either format can be imported for further analysis to any statistical package and contains original data as well as p probability (see equation (2)) and information on the rules applied. Pairs of data with p probability less than 70 are not included in the output. This information allows us to also do detailed analyses of the method.

The DELPHI program first transforms all names according to the Kölner Transformation and implements the methods defined in Tables 3 and 4. When comparing one person against 100,000 persons the program needs about two seconds on a common PC. The resulting computing times are acceptable for our typical projects. Therefore, we did not implement blocking techniques, which are known to reduce computing time by a quadratic factor [8].

The program runs well in practice and has proven advantages with regard to simplicity of interface and interpretation of results. From the point of view of our cancer registry its main advantage is that it takes into account typing errors that derive from the language used, thus here restricted to the German language.

Results

The program described above is applied in the Cancer Registry of Tyrol to join various data sources and check duplicates in the incidence database. Table 5 describes the main results for all linkages done in the years 1999 to 2003. A total of 130,509 linkages were conducted, of which 105,272 (80.7%) were identical pairs. Of these identical pairs, 88.9% of decisions were performed automatically and 11.1% semi-automatically (meaning they were made by the clerical staff).

For results decided automatically, 98.1% had identical last name and 1.8% had identical phonetic transformation of the last name; 95% of cases had identical first name and all cases had identical date of birth. Simultaneous identity of last name and date of birth and first name was observed for 93.1%.

For results decided semi-automatically, 51.6% had identical last name and 1.3% identical phonetic transformation of the last name. First name was identical in 85.2% of cases and date of birth was identical in 74.4%. Simultaneous identity of last name and date of birth and first name was observed for 16.9%. One-character rules (defined in Table 3) applied to last name for 40.4%.

Discussion

Choice of Critical Bounds

We use this program for two main purposes, namely for linking two different data sources and for identification of persons registered more than once in the database.

One of the key decisions during implementation was to choose specific values for the critical bounds p^1 and p^2. In order to evaluate this decision, one must bear in mind the consequences of false-positive and false-negative decisions [12-17].

For medical applications, false-positive linkages cause wrong medical information to be assigned to a person. This must be avoided in all cases. The consequences of false-negative linkage (not assigning, for example, diagnoses or results to a patient) would mean that data available for a person are not recognized. Of course, this should also be avoided, but the consequences are not as dramatic as for false-positive linkage.

In epidemiological studies, false-positive linkages generally result in underestimating true rates, whereas false-negative linkages result in overestimating rates. It is well known that small errors in record linkage (5%) can yield a substantial error in the estimated rates (see e.g. Pukkala, lecture at the IARC 1998 conference in Atlanta).

When applying our method, false-positive record linkage results (homonyms) can occur in the following situations based on p probability: For $p \in (p^2, 100)$ the decision is based only on the p probability. Based on our choice of $p^2 = 95$, a false-positive decision occurs only when there are minimal differences in a single component and all other components have identical values. For $p \in (p^1, p^{2)}$ all decisions are made by the user. The method can prompt false-positive decisions if the resulting list contains long parts with identical pairs interspersed by a few pairs describing different persons.

False-negative record linkage results (synonyms) can occur in the following situations based on the p probability: For $p \in (0, p^1)$, the pair is not included in the output file. For $p \in (p^1, p^{2)}$, all decisions are made by the user. The method can provoke false-negative decisions if the resulting list contains long parts with non-identical pairs interrupted by a few pairs describing the same person.

In order to reduce false-positive and false-negative results, the critical bounds p^1 and p^2 can be changed. It should be noted that every change in the critical bounds has consequences for the time needed to decide the unclear cases and in some respect also for the overall result, bearing in mind the potentially longer lists with unclear cases which can also provoke additional errors. Many decisions can be made just by taking a close look at the components. Other decisions require further information and in general a few minutes of time. Good decisions are based on proper knowledge of data origins, on knowledge of typical registration errors and on good knowledge of frequent last names and first names.

Validation of Method

The correctness of the method presented depends on three factors, namely the correct implementation of the probabilistic record linkage method, the proper choice of critical bounds and the thoroughness of the clerical staff working on the list of unclear cases.

Implementation of the method by writing a software program was checked and carefully tested by proper cross-reading of the code and by applying the program to suitable test data. The proper choice of critical bounds was discussed in the previous chapter.

By implicit assumption, the method also depends on the availability of the key information needed for the method. As described in the Introduction, the cancer registries usually collect these data accurately.

In order to check the overall result of the method, we reanalyzed two typical applications of the record linkage method. As mentioned, we use the program for two purposes, namely to detect persons registered multiple times and to combine two databases. Both functions were checked systematically.

Checking for persons registered multiple times was done for all incident cancer cases of the year of diagnosis 1996. Checking for errors when combining two databases was performed by linking the incidence data of the year 1996 and the mortality data for the years 1996 to 2001. We searched for false-positive and false-negative pairs. This was done by means of a long list of heuristic checks, for example persons for whom the first three letters of their last name and their complete date of birth are identical, or persons for whom the first five letters of their last name and the month and year of birth are identical. In total, we could not find any false-positive or false-negative combination. Also, we could not find any person registered multiple times. It should be mentioned that one possible bias within this check is the fact that the re-evaluation was done by the same clerical staff, who therefore could make the same wrong decision a second time.

Practical Considerations

Our method needs additional time as compared to deterministic procedures. This is the case for every probabilistic record linkage procedure, because they result in cases that cannot be decided automatically per definition. Thus, when applying a probabilistic method, one has to decide how much time to spend on deciding the status of unclear cases. Both our main applications, namely detecting persons registered multiple times and assessing patient life status, have direct impact on main results and we therefore decided to invest the extra time in order to obtain reliable incidence and survival rates.

Table 5 shows that 11% of identical pairs were not decided automatically and that of those cases decided automatically 6.9% did not have simultaneous identity of last name and first name and date of birth. This means that around 15% of cases would not have been linked by the widely used rules of deterministic record linkage procedures.

One further aspect should be mentioned that is specific for our region: residential mobility is low. We know from studies that patients have on average only about three residences throughout their lifetime [20]. This means that change of patient address is rather unlikely to occur and so the component municipality code or zip code is very stable.

Commercial programs are available for record linkage, Automatch [10, 11, 18, 19] being one of the main programs used in this area. Automatch offers very good implementation of the methodology of probabilistic record linkage. The main difference between Automatch and our solution is the consideration of what we call additional methods defined in Table 3. In addition, our implementation is adapted to cancer registry data structure, and all decisions concerning choice of parameters are fix-coded so that all user interactions are minimized, resulting in a very time-efficient operation. A further reason was the rather high price of Automatch.

One of the problems encountered in practical record linkage is that more or less precise information is needed to identify a person while every registry must observe strict data privacy laws [21-23]. The legal basis for our cancer registry allows us to store all data on identification of patients, of course in compliance with strict guidelines to safeguard confidentiality. We hope that in future a unique person identifier will be introduced in our country, which would overcome record linkage problems and all data privacy concerns [24].

Conclusions

We have developed a record linkage method for cancer registries in Austria based on the theory of probabilistic record linkage adjusted for special conditions in the German language. The method serves two main purposes, namely record linkage of various data sources and identification of persons registered more than once in the database. Both goals were reached with adequate precision. The time needed to decide unclear cases is justifiable.

References

1. Parkin DM, Whelan SL, Ferlay J, Raymond L, Yuen J. Cancer Incidence in Five Continents. Volume VII. IARC Scientific Publications No. 143. Lyon: IARC, 1997.
2. Parkin DM, Whelan SL, Ferlay J, Teppo L, Thomas B. Cancer Incidence in Five Continents. Volume VIII. IARC Scientific Publications No. 155. Lyon: 2002.
3. Jensen OM, Parkin DM, MacLennan R, Muir CS, Skeet RG. Cancer Registration. Principles and Methods. Lyon: IARC, 1991.
4. Schouten LJ, Hoppener P, van den Brandt P, Knottnerus JA, Jager JJ. Completeness of Cancer Registration in Limburg, The Netherlands. Int J Epidemiol 1993; 22 (3): 369-76.
5. Ziegler H, Stegmaier C. Bevölkerungsbezogene Krebsregistrierung in Deutschland. Onkologie 1996; 19: 268-77.
6. Calle EE, Terrell DD. Utility of the National Death Index for Ascertainment of Mortality among Cancer Prevention Study II Participants. Am J Epid 1993; 137 (2): 235-41.
7. Fellegi IP, Sunter AB. A Theory for Record Linkage. JASA 1969; 64: 1183-1210.
8. Jaro MA. Probabilistic Linkage of Large Public Health Data Files. Stat Med 1995; 14: 491-8.
9. Leitner H. Probleme und Lösungen beim Zusammenführen von Patientendaten unterschiedlicher Quellen. Innsbruck: 1995.
10. Stegmaier C, Ziegler H. Zusammenführung personenbezogener Informationen (Record-Linkage) – Probleme und Möglichkeiten am Beispiel des Krebsregisters Saarland. Krebsregister Saarland, Vierteljahresheft 1992; 4:1-7.
11. Schmidtmann I, Michaelis J. Untersuchungen zum Record Linkage für das Krebsregister Rheinland-Pfalz. 1994.
12. Brenner H, Schmidtmann I. Determinants of Homonym and Synonym Rates of Record Linkage in Disease Registration. Methods Inf Med 1996; 35 (1): 19-24.
13. Brenner H, Schmidtmann I, Stegmaier C. Effects of Record Linkage Errors on Registry-based Follow-up Studies. Stat Med 1997; 16 (23): 2633-43.
14. Newcombe HB, Smith ME, Howe GR, Mingay J, Strugnell A, Abbatt JD. Reliability of Computerized versus Manual Death Searches in a Study of Health of Eldorado Uranium Workers. Comput Biol Med 1983; 13 (3): 157-69.
15. van den Brandt P, Schouten LJ. Development of a Record Linkage Protocol for Use in the Dutch Cancer Registry for Epidemiological Research. Int J Epid 1990; 19 (3): 553-8.
16. MacLeod MC, Bray CA, Kendrick SW, Cobe S. Enhancing the Power of Record Linkage involving low quality Personal Identifiers: Use of the Best Link Principle and Cause of Death prior Likelihoods. Comput Biomed Res 1998; 31 (4): 257-70.
17. Gomatam S, Carter R, Ariet M, Mitchell G. An empirical Comparison of Record Linkage Procedures. Stat Med 2002; 21 (10): 1485-96.
18. Thomas B, Newman M. Use of Commercial Record Linkage Software and Vital Statistics to Identify Patient Deaths. J Am Med Inform Assoc 1997; 4 (3): 233-7.
19. Kendrick SW, Douglas MM, Gardner D, Hucker D. Best-link Matching of Scottish Health Data Sets. Methods Inf Med 1998; 37 (1):64-8.
20. Oberaigner W, Kreienbrock L, Schaffrath-Rosario A, Kreuzer M, Wellmann J, Keller G, Gerken M, Langer B, Wichmann HE. Radon und Lungenkrebs im Bezirk Imst/Österreich. In: Wichmann HE, Schlipkoeter HW, Fuelgraff G (eds). Fortschritte in der Umweltmedizin. Landsberg/Lech: Ecomed Verlag; 2002.
21. Pommering K, Miller M, Schmidtmann I, Michaelis J. Pseudonyms for Cancer Registries. Methods Inf Med 1996; 35 (2): 112-21.
22. Meux E. Encrypting Personal Identifiers. Health Serv Res 1994; 29(2):247–256.
23. Rabinowitz J. A Method for Preserving Confidentiality when Linking Computerized Registries. Am J Public Health 1998; 88 (5): 836.
24. Kelman C, Smith L. It's time: Record Linkage – The Vision and the Reality. Aust N Z J Public Health 2000; 24 (1):100-1.

Correspondence to:
Dr. Willi Oberaigner
Anichstrasse 35
Innsbruck
6020 Austria
E-Mail willi.oberaigner@iet.at