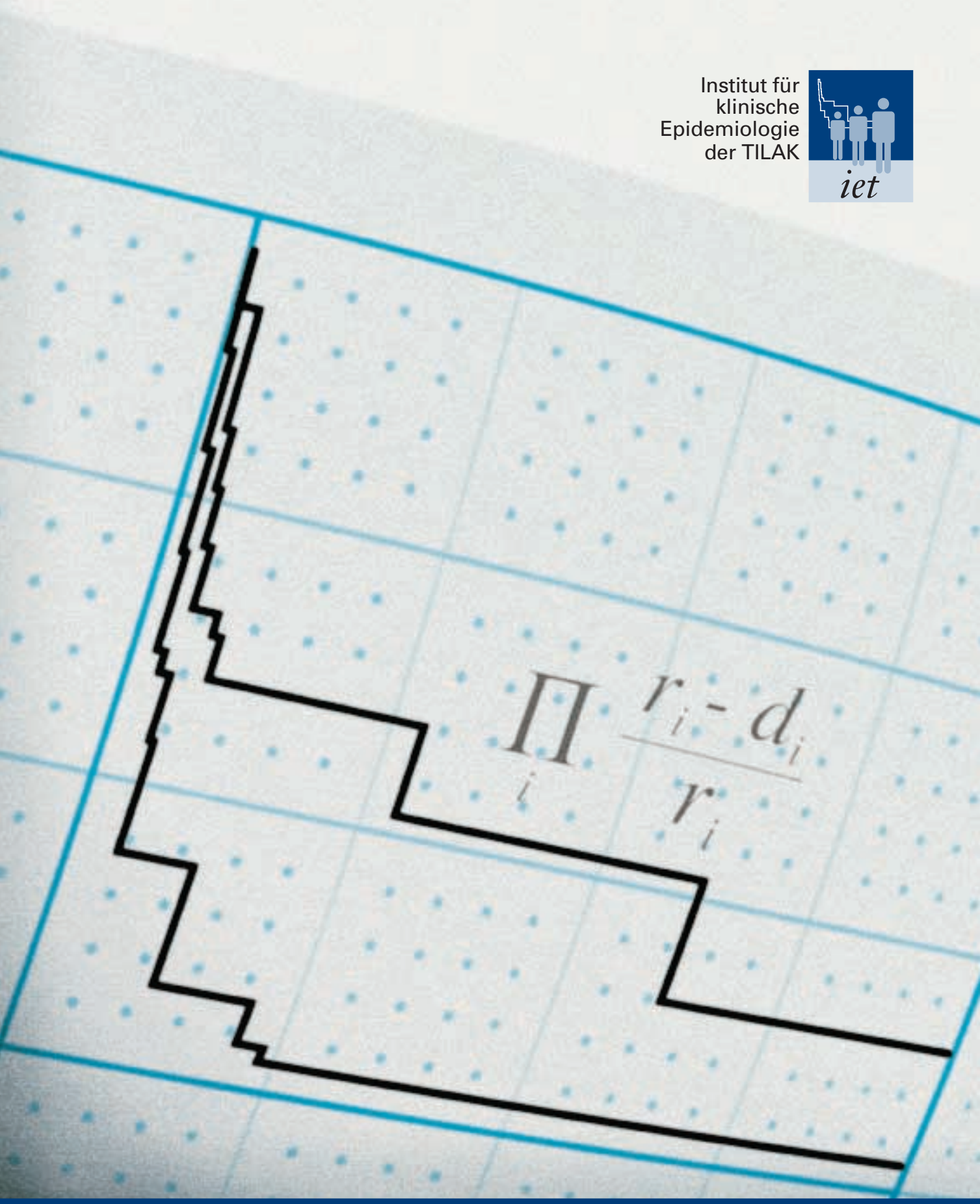
Institut für
klinische
Epidemiologie
der TILAK

iet



Publikationen des Instituts für klinische Epidemiologie der TILAK

Record-Linkage im Tumorregister Tirol

```
set heading off;
set pagesize 10000;
set newpage 1;
set termout off;
set feedback off;
set verify off;
set timing off;
set echo off;
spool pat1.dat;
select nr,
       ||'\'||
       lower(replace(
        replace(
         replace(
          replace(
           replace(
            replace(
             replace(
              rtrim(ltrim(surname)),
                    'ä','ae'),
                    'ö','oe'),
                    'ü','ue'),
                  'Ä','AE'),
                  'Ö','OE'),
                  'Ü','UE'),
         'ß','ss'))
       ||'\'||
       lower(replace(
        replace(
         replace(
          replace(
           replace(
            replace(
             replace(
              rtrim(ltrim(first_name)),
                    'ä','ae'),
                    'ö','oe'),
                    'ü','ue'),
                  'Ä','AE'),
                  'Ö','OE'),
                  'Ü','UE'),
         'ß','ss'))
```

Record-Linkage im Tumorregister Tirol

Wilhelm Oberaigner

Institut für
klinische Epidemiologie
der TILAK GmbH

IMPRESSUM

iet – Institut für
klinische Epidemiologie
der TILAK GmbH
Anichstraße 35
A-6020 Innsbruck
www.iet.at

Dr. Wilhelm Oberaigner

Innsbruck, November 2001

INHALT

Record Linkage im Tumorregister Tirol

Zusammenfassung

Record-Linkage für die Verbindung von Patientendaten aus unterschiedlichen Datenquellen ist eine zentrale Aufgabe jedes Krebsregisters. Weil es in Österreich keine einheitliche Identifizierung von Personen gibt, war es notwendig, eine Strategie für Record-Linkage zu entwickeln. Wir haben die Methode des probabilistischen Record-Linkage an unsere speziellen Gegebenheiten adaptiert; dabei verwenden wir neben üblichen Komponenten auch Tippfehler in Namen und Geburtsdatum. Verglichen mit früher eingesetzten ad hoc-Methoden sind wir in der Lage, etwa 15 bis 20% zusätzlicher Datensätze zu verbinden.

1 Hintergrund

Es ist generelles Ziel eines jeden bevölkerungsbezogenen Krebsregisters, jeden Tumorfall in einer Bevölkerung zu dokumentieren[1 2 3 4]. Nach international gültigen Empfehlungen soll man alle verfügbaren Datenquellen nutzen, die valide Informationen über Tumordiagnosen enthalten. Dies bedeutet, dass neben den Informationen, die direkt von den behandelnden Spitalsärzten kommen, andere Datenquellen wie Pathologiebefunde, Daten aus Abteilungssystemen, Daten aus Therapieabteilungen oder auch Krankenhausinformationssystemen verwendet werden müssen. Um valide Überlebensdaten berechnen zu können, ist weiters ein zuverlässiges Record-Linkage mit den Mortalitätsdaten notwendig[5].

Zusammenfassend ist also eine gute Methode für Record-Linkage eine zentrale Aufgabe jedes Krebsregisters. In Österreich wird im medizinischen System keine allgemeine Identifizierung von Patientendaten verwendet (wie z.B. in skandinavischen Ländern). Die Sozialversicherungsnummer ist wohl in den meisten Fällen eindeutig (dies ist aber nicht garantiert), wird aber nur zum Teil verwendet. Daher ist die Entscheidung, ob 2 Datensätze zu derselben Person gehören, im allgemeinen relativ kompliziert und zeitaufwendig und muss auf Komponenten wie Familienname, Vorname, Geburtsdatum etc. basieren. Natürlich muss dabei noch berücksichtigt werden, dass alle Komponenten durch Tippfehler oder andere Fehlerarten verfälscht sein können.

Um dieses Problem zu lösen, haben wir uns entschieden, basierend auf der Methode des probabilistischen Record-Linkage ein eigenes Programm zu entwickeln, das vor allem auch typische Fehler der deutschen Sprache berücksichtigt.

2 Methode

2.1 Grundlagen

Daten bestehen aus mehreren Komponenten. Ein Teil dieser Komponenten, oft Personenstammdaten genannt, beschreibt bzw. identifiziert die Person. Wir setzen voraus, dass keine einzelne Komponente die Person eindeutig identifiziert. Wenn nun eine Person durch n Komponenten k_1 bis k_n beschrieben wird, so weisen wir jeder Komponente k_i ein standardisiertes Gewicht w_i zu mit der Bedingung

$$w_1+...+ w_n=1.$$

Für das Record-Linkage Verfahren von 2 Datensätzen mit Komponenten k_i^1 und k_i^2 definieren wir p_i für jede Komponente k_i durch:

$$p_i = \begin{cases} 1 \; if \; k_i^1 = k_i^2 \\ 0 \; sonst \end{cases} \quad (1)$$

Dies führt zu folgender Definition einer Summenwahrscheinlichkeit

$$p= w_1 p_1+...+ w_n p_n \quad (2)$$

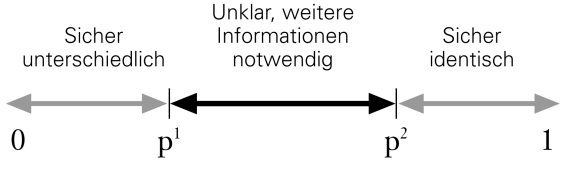
p (im folgenden oft bezeichnet als p-Wahrscheinlichkeit) kann interpretiert werden als Maß für die Übereinstimmung von 2 Personen, die durch die Komponenten k_i^1 und k_i^2 beschrieben werden. Im folgenden werden 2 Schranken p^1 und p^2 definiert mit folgender Interpretation:

- Falls p kleiner ist als p^1, so wird ohne weitere Überprüfungen angenommen, dass es sich um verschiedene Personen handelt.
- Falls p größer als p^2 ist, so wird wiederum ohne weitere Überprüfungen angenommen, dass es sich um dieselbe Person handelt.
- Falls p zwischen p^1 und p^2 liegt, so muss die Entscheidung, ob es sich um dieselbe Person handelt, auf individueller Basis getroffen werden.

Üblicherweise ist dieser Schritt mit Beschaffung weiterer Informationen verbunden.

Dieser Entscheidungsprozess ist in der folgenden Grafik dargestellt:

Abbildung 1



2.2 Wahl der Gewichte

Für die Wahl der Gewichte werden nach der Strategie des probabilistischen Record-Linkage 2 Wahrscheinlichkeiten berechnet, meist bezeichnet als m- und u-Wahrscheinlichkeit.

Für eine beliebige Komponente k_i, ist m_i definiert als die Wahrscheinlichkeit, dass die Komponente k_i für identische Personen identisch ist.
u_i ist definiert als die Wahrscheinlichkeit, dass die Komponente k_i für unterschiedliche Personen gleich ist. Dann werden die Gewichte w_i durch folgende Formel definiert:

$$w_i = log_2 \left(\frac{m_i}{u_i}\right) \tag{3}$$

Nach den Erfahrungen mit unseren Krebsregisterdaten haben wir die Komponenten folgendermaßen gewählt[9]:

Tabelle 1

Komponente k_i
Familienname
Phonetische Transformation des Familiennamens nur falls Familienname ungleich ist, siehe Tabelle 2
Geburtsname
Phonetische Transformation des Geburtsnamens nur falls Geburtsname ungleich, siehe Tabelle 2
Vorname
Geburtsdatum
Geschlecht
Postleitzahl (oder Gemeindenummer)

In der deutschen Sprache gibt es typische Transformationen von Namen, die bestimmten Regeln genügen. Daher haben wir das Konzept einer phonetischen Transformation berücksichtigt, die nach folgenden Regeln definiert ist (sogenannte Kölner-Transformation [6 10]):

Tabelle 2

Regel	**Beispiel**		
Eliminiere Doppellaute	Wimmer	→	WIMER
Transformiere deutsche Umlaute	Müller	→	MUELER
Transformiere c vor e,i in z	Cicero	→	ZIZERO
Transformiere c vor a,o,u in k	Cugel	→	KUGEL
andernfalls Transformiere c in z	Mucke	→	MUZKE
Transformiere v in f	Vogel	→	FOGEL
Transformiere j in i	Deljc	→	DELIZ
Transformiere ie in i	Liederlich	→	LIDERLIZH
Transformiere ai in ei	Aigner	→	EIGNER
Transformiere ae in e	Jaeger	→	IEGER
Transformiere th in t	Thaler	→	TALER
Transformiere tz in z	Matzer	→	MAZER
Transformiere d in t	Danner	→	TANER
Lösche stummes h	Gehler	→	GELER
Transformiere qu in q	Qualler	→	QALER

Die Wahrscheinlichkeiten m_i und u_i wurden nach Resultaten berechnet, die vor der Einführung des hier beschriebenen Record-Linkage-Programmes entstanden sind. Datensätze wurden mit heuristischen Methoden und aufwendigen Überprüfungsmechanismen verbunden. Alle Resultate wurden in einer Meta-Relation gesammelt, die sowohl Informationen über die Komponenten als auch die Resultate enthalten. Basierend auf dieser Meta-Relation kann m_i wie folgt berechnet werden:

$$m_i = \frac{Anzahl\ Patienten\ mit\ identischer\ Komponente\ k_i}{Anzahl\ von\ Patienten} \tag{4}$$

Analogerweise kann die Wahrscheinlichkeit u_i berechnet werden (wir nehmen an, dass unsere Datenbank Pat keine Doppelpatienten enthält, daher enthält das kartesische Produkt Pat x Pat keine Paare identischer Patienten):

$$u_i = \frac{Anzahl\ Patienten\ mit\ identischer\ Komponente\ k_i}{Anzahl\ von\ Patienten\ insgesamt} \tag{5}$$

Diese Berechnung hat zu folgenden Gewichten geführt:

Tabelle 3

Komponenten k_i	**w_i (standardisiert)**
Phonetische Transformation Familienname	0.22
Phonetische Transformation Geburtsname	0.202
Vorname	0.139
Geburtsdatum	0.289
Geschlecht	0.075
Postleitzahl bzw. Gemeindenummer	0.075
Summe	1

Nach detaillierter Analyse unserer Datenbank[9] hat sich ergeben, dass
1. es häufige Tippfehler in Familienname und Geburtsname gibt,
2. es häufige Tippfehler im Geburtsdatum gibt.
Um diese Fehler berücksichtigen zu können, haben wir folgende Methoden ergänzt:

Tabelle 4

Methode	**Beispiel**
»Linker« oder »rechter« Teil des Namens identisch	Müller und Müller-Westernhagen
1 Zeichen falsch	Maier and Mayer
1 Zeichen fehlend	Maier und Mair
2 benachbarte Zeichen vertauscht	Maier und Miaer

Daher haben wir die Gewichte von *Tabelle 3* erweitert für die oben beschriebenen Methoden:

Tabelle 5

Komponenten bzw. Methode für Komponenten	**Gewicht**
Familienname: »linker« oder »rechter« Teil identisch	$w_{Familienname} * 0.9$
Familienname: 1 Zeichen falsch	$w_{Familienname} * 0.8$
Familienname: 1 Zeichen fehlend	$w_{Familienname} * 0.8$
Familienname: 2 Buchstaben vertauscht	$w_{Familienname} * 0.8$
Erste 3 Buchstaben des Familiennamens identisch	$w_{Familienname} * 0.4$
Vorname: »linker« oder »rechter« Teil identisch	$w_{Vorname} * 0.5$
Familienname and Geburtsname vertauscht	$w_{Familienname}$
Geburtsdatum: 1 Zeichen falsch	$w_{Geburtsdatum} * 0.8$
Geburtsdatum: 2 Zeichen vertauscht	$w_{Geburtsdatum} * 0.8$
Geburtsdatum: Tag und Monat vertauscht	$w_{Geburtsdatum} * 0.8$
Geburtsdatum: Tag identisch	$w_{Geburtsdatum} * 0.3$
Geburtsdatum: Monat identisch	$w_{Geburtsdatum} * 0.3$
Geburtsdatum: Jahr identisch	$w_{Geburtsdatum} * 0.3$

Für jede Komponenten k_i ist das maximale Gewicht beschränkt durch das Gewicht wie in *Tabelle 3* definiert, auch falls alle in *Tabelle 5* beschriebenen Methoden in Summe zu einem höheren Gewicht führen würden.

2.3 Wahl der kritischen Grenzen p^1 und p^2

Nach den Erfahrungen unseres Krebsregister ist $p^1=0.75$ und $p^2=0.99$ eine sinnvolle Wahl der kritischen Grenzen für die p-Wahrscheinlichkeit. Dies bedeutet konkret, dass wir alle Paare mit einer p-Wahrscheinlichkeit zwischen 0.75 und 0.99 individuell überprüfen und dass wir Paare mit p=1 ohne weitere Überprüfung als identisch ansehen.

Die Überprüfung von Paaren mit $p \in [0.75, 0.99]$ ist zeitaufwendig und verlangt viel Konzentration. Nach unserer Erfahrung gibt es aber immer wieder Paare mit kleiner p-Wahrscheinlichkeit, die dieselbe Person beschreiben (man denke etwa an Zwillinge, die im selben Ort wohnen und deren Vorname mit demselben Buchstaben beginnt). Wenn also die Homonym- und Synonymraten klein gehalten werden sollen (vergleiche auch die Diskussion über die Konsequenzen falscher Entscheidungen), so ist es notwendig, alle Teile der Ergebnislisten mit voller Konzentration zu bearbeiten.

2.4 Homonymrate

Falsch-positive Ergebnisse können in folgenden Situationen auftreten:

- $p \in (p^2,1]$: In diesem Intervall der p-Wahrscheinlichkeit wird ohne weitere Überprüfungen der Schluss gezogen, dass es sich um dieselbe Person handelt. Ein falsch-positives Ergebnis tritt also bei unserer Wahl von $p^2=0.99$ nur dann auf, wenn ein Paar mit p=1 unterschiedliche Personen beschreibt. p=1 wird aber nur dann als Ergebnis geliefert, falls alle Komponenten identisch sind. Würde man auch in diesem Fall annehmen, dass es sich eventuell um unterschiedliche Personen handeln kann, so hätte dies die Konsequenz, dass man alle Paare »händisch« überprüfen müsste.
- $p \in [p1,p2]$: In diesem Intervall werden alle Entscheidungen durch den Bearbeiter getroffen. Die von uns vorgeschlagene Methode kann falsch-positive Entscheidungen »provozieren«, falls es lange Listenteile mit unterschiedlichen Personenpaaren liefert, die nur durch einzelne identische Paare unterbrochen werden (vergleiche auch die Diskussion über Konsequenzen von falsch-positiven und falsch-negativen Entscheidungen).

2.5 Synonymrate

Falsch-negative Ergebnisse können in folgenden Fällen auftreten:

- $p \in [0,p^1)$: in diesem Intervall wird ein Paar überhaupt nicht in die Ergebnisliste aufgenommen. Um zu überprüfen, ob durch diese Festlegung falsch-negative Entscheidungen getroffen werden, führten wir unabhängig vom hier vorgestellten Programm ad hoc-Heuristiken durch, fanden aber kein falsch-negatives Paar mit $p \in [0, p^1)$.
- $p \in [p^1,p^2]$: In diesem Intervall werden alle Entscheidungen durch den Bearbeiter getroffen. Die Methode kann falsch-negative Entscheidungen »provozieren«, wenn lange Listen mit nicht-identischen Paaren erzeugt werden, die unterbrochen werden durch einzelne identische Paare (siehe Diskussion der Konsequenzen von falsch-positiven und falsch-negativen Entscheidungen).

2.6 Implementierung

Die oben beschriebene Methode wurde mit einem Programm implementiert, das mit der Programmiersprache Pascal formuliert wurde. Schnittstelle für die Daten sind zwei ASCII-Dateien mit identischer Struktur (siehe *Anhang 5.2*). Alle Resultate werden auf eine ASCII-Datei geschrieben: diese Datei enthält die Originaldaten, ergänzt mit der p-Wahrscheinlichkeit (siehe *Gleichung (2)*) und Information über die angewandten Methoden. Paare mit p-Wahrscheinlichkeit unter 0.70 werden nicht aufgenommen. Die Ergebnisse erlauben uns auch, retrospektive Analysen durchzuführen.

Das Programm transformiert alle Namen nach der oben beschriebenen Kölner-Transformation und führt die in den *Tabellen 4* und *5* beschriebenen rechenintensiven Methoden durch. Für den Vergleich eines Patienten gegen 50.000 Patienten benötigt das Programm auf einem Pentium II mit 400 MHz ca. 3 Sekunden.

Da die damit entstehenden Zeiten für unsere typischen Anwendungen akzeptabel sind, haben wir verzichtet, Techniken zur Reduktion der Rechenzeit wie z.B. Blockungstechniken einzuführen (es ist bekannt, dass Blockungsstrategien die Rechenzeit um einen quadratischen Faktor reduzieren können[11]).

Das Programm bewährt sich in der Praxis durch die Einfachheit der Bedienung und gute Interpretation der Resultate. Hauptvorteil für uns ist die Tatsache, dass das Programm auch die häufigen Tippfehler berücksichtigt (Fehler hängen ja auch von der Sprache ab, die im Register bzw. der Arbeitsumgebung verwendet wird).

3 Diskussion

Wir verwenden das Programm für zwei Anwendungsbereiche, nämlich für die Verbindung von zwei 2 Datenmengen und um Doppelerfassungen zu entdecken. Für die Wahl der kritischen Grenzen p^1 und p^2 muss man die Konsequenzen von falsch-positiven und falsch-negativen Entscheidungen bedenken [12 13 14 15 16 17].

Für medizinische Anwendungen haben falsch-positive Entscheidungen die Konsequenz, dass Daten einer falschen Person zugeordnet werden. Dies muss natürlich unter allen Umständen vermieden werden. Falsch-negative Entscheidungen (ein vorhandener Befund wird einer Person nicht zugeordnet) haben die Konsequenz, dass vorhanden Daten nicht für medizinische Entscheidungen zur Verfügung stehen.

Für epidemiologische Studien können falsch-positive Linkage-Ergebnisse zu Unterschätzung der wahren Rate führen, falsch-negative Entscheidungen zu Überschätzung der tatsächlichen Rate. Es ist bekannt, dass Linkage-Fehlerraten um 5% zu deutlichen Fehlern in den Resultaten führen können (z.B. Pukkala, Vortrag bei IARC-98 Konferenz in Atlanta).

Um Homonym- und Synonymraten zu verkleinern, können die kritischen Grenzen p^1 und p^2 geändert werden und man kann mehr Zeit in die Bearbeitung der Listen investieren, indem z.B. versucht wird, möglichst umfassende Informationen für die Entscheidungen zu beschaffen. Wir benötigen im Durchschnitt 3 – 5 Minuten für eine Entscheidung von unklaren Fällen (dazu zählen nicht Fälle, die rein auf Grund der vorhandenen Information und ohne Beschaffung von weiteren Details entschieden werden können).

»Gute« Entscheidungen basieren natürlich neben guten Programm-Resultaten auf guter Kenntnis der Datenquellen, Kenntnis typischer Tippfehler und guter Kenntnis von Häufigkeiten von Familien- und Vornamen sowie auch von Gemeindegrößen (falls Mobilität keine große Rolle spielt).

Es gibt kommerzielle Programme wie Automatch[6 10 18 17]. Gegenüber Automatch hat unser Programm den Vorteil, dass typische Transformation der deutschen Sprache (siehe Kölner Transformation) sowie typische Tippfehler berücksichtigt werden (siehe *Tabelle 5*). Unser Programm kann also Fehler berücksichtigen, die von Automatch nicht behandelbar sind.

Eines der großen Probleme mit Record-Linkage ist, dass man relativ genaue Informationen über die Personen benötigt, um Grenzfälle zu entscheiden. Andererseits ist jedes Register zu strikter Einhaltung von Datenschutzbestimmungen verpflichtet [19 20 21].

Obwohl wir auf Grund unserer rechtlichen Situation in der Lage sind, Individualdaten zu speichern, hoffen wir sehr, dass die Einführung einer Personenkennzahl in Österreich zumindest längerfristig die Verwendung von sophistizierten und zeitaufwendigen Record-Linkage-Verfahren überflüssig machen wird [22].

4 Reference List

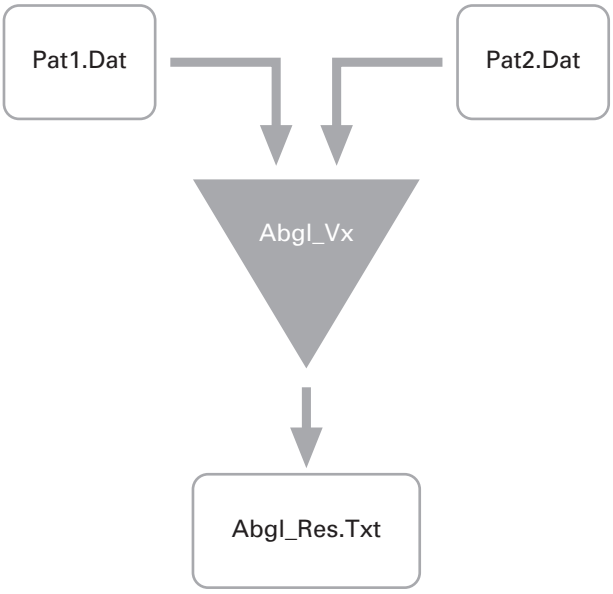
1. Parkin D.M., Whelan SL, Ferlay J, Raymond L, Yuen J. *Cancer Incidence in Five Continents Vol. VII.* Lyon: 1997;
2. Jensen OM, Parkin DM, MacLennan R, Muir C.S., Skeet R.G. *Cancer Registration: Principles And Methods.* IARC Scientific Publications No 95, 1991;
3. Schouten LJ, Hoppener P, van-den BP, Knottnerus JA, Jager JJ. *Completeness of cancer registration in Limburg, The Netherlands.* Int.J.Epidemiol. 1993; 22: 369-376.
4. Ziegler H. and Stegmaier, C. *Bevölkerungsbezogene Krebsregistrierung in Deutschland.* Onkologie 19, 268-277. 1996. Freiburg, Karger,S.
5. Calle, E. E. and Terrell, D. D. *Utility Of The National Death Index For Ascertainment Of Mortality Among Cancer Prevention Study.* Ii. Participants. Am.J.Epid. 137(2), 235-241. 1993.
6. Stegmaier C, Ziegler H. *Zusammenführung personenbezogener Informationen (Record-Linkage) - Probleme und Möglchkeiten am Beispiel des Krebsregisters Saarland.* Krebsregister Saarland Vierteljahresheft 1992; 4: 1-7.
7. Jaro, M. A. *Advances in Record-Linkage Methodology as Applied to Matching the 1985 Census of Tampa, Florida.* J.Am.Stat.Ass. 84(406), 414-420. 1989.
8. Fellegi IP, Sunter AB. *a theory for record linkage.* JASA 1969; 64: 1183-1210.
9. Leitner H. *Probleme und Lösungen beim Zusammenführen von Patientendaten unterschiedlicher Quellen.* In: Anonymous1995:
10. Schmidtmann, I. and Michaelis, J. *Untersuchungen zum Record Linkage für das Krebsregister Rheinland-Pfalz.* 1994.
11. Jaro, M. A. *Probabilistic Linkage Of Large Public Health Data Files.* Stat.Med. 14, 491-498. 1995.
12. Brenner H, Schmidtmann I. *Determinants of homonym and synonym rates of record linkage in disease registration.* Methods Inf.Med 1996; **35**: 19-24.
13. Brenner H, Schmidtmann I, Stegmaier C. *Effects of record linkage errors on registry-based follow-up studies.* Stat.Med 1997; 16: 2633-2643.
14. Newcombe H.B, Smith M.E., Howe G.R., Mingay J., Strugnell A., Abbatt J.D. *Reliability of Computerized versus manual death serches in a study of th health of eldorado uranium workers.* Comput.Biol.Med.Vol. 1983; Vol 13 No. 3.: 157-169.
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: 553-558.
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: 257-270.
17. Kendrick SW, Douglas MM, Gardner D, Hucker D. *Best-link matching of Scottish health data sets.* Methods Inf.Med 1998; 37: 64-68.
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: 233-237.
19. Pommerening K, Miller M, Schmidtmann I, Michaelis J. *Pseudonyms for cancer registries.* Methods Inf.Med 1996; 35: 112-121.
20. Meux E. *Encrypting personal identifiers.* Health Serv.Res 1994; 29: 247-256.
21. Rabinowitz J. *A method for preserving confidentiality when linking computerized registries [letter].* Am.J.Public Health 1998; 88: 836
22. Kelman-C, Smith-L. *It's time: record linkage—the vision and the reality.* Aust.N.Z.J.Public Health 2000; 24: 100-1.

5 Anlage: User's Guide

5.1 Datenfluss

Der Benutzer muss 2 Dateien vorbereiten, die die Patientendaten enthalten (*Pat1.Dat* und *Pat2.Dat*). Alle Resultate werden in eine Textdatei mit Namen *Abgl_Res.Txt* geschrieben, siehe folgende Abbildung:

Abbildung 2



5.2 Beschreibung der Dateiformate

Struktur der Eingabedateien (Pat1.Dat, Pat2.Dat)
Jede Zeile enthält einen Datensatz in folgendem Format:

```
Nummer\Familienname\Vorname\Geburtsname\Geburtsdatum\Postleitzahl\Geschlecht\Quelle
```

Die folgende Tabelle beschreibt die einzelnen Komponenten

Komponente	Max. Länge	Spezifikation der Inhalte
Nummer	10	Eindeutige Nummer des Datensatzes; nur für Organisation, damit man in der Ausgabedatei den entsprechenden Datensatz identifizieren kann
Familienname	20	Familienname in Kleinbuchstaben; keine deutschen Umlaute
Vorname	15	Vorname in Kleinbuchstaben; keine deutschen Umlaute
Geburtsname	20	Geburtsname in Kleinbuchstaben; keine deutschen Umlaute
Geburtsdatum	10	Geburtsdatum im Format dd.mm.yyyy
Postleitzahl	8	Postleitzahl oder Gemeindenummer (wird vom Programm nicht inhaltlich analysiert)
Geschlecht	1	Geschlecht, vorzugsweise in Form m für Männer und w für Frauen (man kann auch eine andere Codierung verwenden, aber dann natürlich dieselbe in beiden Dateien
Quelle	250	Text, der den Ursprung der Daten beschreibt; nur für Organisation, damit man in der Ausgabedatei den entsprechenden Datensatz identifizieren kann

Beispiel:

```
23524\jaeger\gerhilde\\02.06.1927\70508\w\TRT-All
```

5.2.1 Beispielprogramme für Erzeugen der Daten

5.2.1.1 SQL

```
set heading off;
set pagesize 10000;
set newpage 1;
set termout off;
set feedback off;
set verify off;
set timing off;
set echo off;
spool pat1.dat;
select nr,
       ||'\'||
       lower(replace(
        replace(
         replace(
          replace(
           replace(
            replace(
             replace(
              rtrim(ltrim(surname)),
                 'ä','ae'),
                 'ö','oe'),
                 'ü','ue'),
              'Ä','AE'),
              'Ö','OE'),
              'Ü','UE'),
        'ß','ss'))
       ||'\'||
       lower(replace(
        replace(
         replace(
          replace(
           replace(
            replace(
             rtrim(ltrim(first_name)),
                 'ä','ae'),
                 'ö','oe'),
                 'ü','ue'),
              'Ä','AE'),
              'Ö','OE'),
              'Ü','UE'),
        'ß','ss'))
       ||'\'||
       lower(replace(
        replace(
         replace(
          replace(
           replace(
            replace(
             rtrim(ltrim(maiden_name)),
                 'ä','ae'),
                 'ö','oe'),
                 'ü','ue'),
              'Ä','AE'),
              'Ö','OE'),
              'Ü','UE'),
        'ß','ss'))
||'\'||date_of_birth
||'\'||zip_code
||'\'||sex
||'\'||'my-rel'
from my-rel;

spool off;
```

5.2.1.2 SPSS

Wir setzen voraus, dass die Daten in folgendem SPSS-Datenformat zur Verfügung stehen:

Nummer	pnr
Familienname	fn
Vorname	vn
Geburtsname	gn
Geburtsdatum	gd
Postleitzahl	plz
Geschlecht	sex

```
* Voraussetzung: Felder enthalten keine deutschen Umlaute,
* daher genügt Umwandlung in Kleinbuchstaben.
Compute FN=Lower(FN).
Compute VN=Lower(VN).
Compute GN=Lower(GN).
* Voraussetzung: GD ist schon im Format dd.mm.yyyy.
Compute Sex=Lower(Sex).
Compute PLZ=Lower(PLZ).
* Postleitzahl: keine Veränderung notwendig.
String StrAbgl (A100).
Compute StrAbgl=Concat(
    RTrim(PNr),'\',
    RTrim(FN) ,'\',
    RTrim(VN) ,'\',
    RTrim(GN) ,'\',
    RTrim(GD) ,'\',
    RTrim(PLZ),'\',
    RTrim(Sex),'\',
    'mysource').
Execute.
WRITE OUTFILE='mydir\Pat1.dat'
TABLE /StrAbgl.
EXECUTE.
```

5.2.1.3. Struktur der Ausgabedatei (Abgl_Res.Txt)

Die Ausgabedatei enthält die Daten in einer Form, die leicht in SPSS oder Excel importierbar ist.

Aufbau:
1. Spalten getrennt durch |
2. Reihenfolge der Spalten: siehe folgende Tabelle bzw. SPSS-Programm für Datenimport:

Nr1	Nummer Datei #1
FN1	Familienname Datei #1
VN1	Vorname Datei #1
GN1	Geburtsname Datei #1
GD1	Geburtsdatum Datei #1
PLZ1	Postleitzahl Datei #1
Sex1	Geschlecht Datei #1
Nr2	Nummer Datei #2
FN2	Familienname Datei #2
VN2	Vorname Datei #2
GN2	Geburtsname Datei #2
GD2	Geburtsdatum Datei #2
PLZ2	Postleitzahl Datei #2
Sex2	Geschlecht Datei #2
Prot	Codierung der angewendeten Regeln
Ursprung	Ursprung (Ursprung#1+Ursprung#2+DatumProgrammausführung)
Rate	Berechnete p-Wahrscheinlichkeit

Beispiel für Import und Liste in SPSS:

```
data list list('|')
file='E:\TRV\Bericht\DatenNeu\Abgl\Abgl_Res.Txt'
/
Nr1 (A10)
FN1 (A25)
VN1 (A25)
GN1 (A25)
GD1 (A10)
PLZ1 (A10)
Sex1 (A1)
Nr2 (A10)
FN2 (A25)
VN2 (A25)
GN2 (A25)
GD2 (A10)
PLZ2 (A10)
Sex2 (A1)
Prot (A50)
Ursprung (A100)
Rate (F3).
Execute.


********************* Example for printing (with Summarize).
Sort Cases By Rate (D).
Select If Nr1<Nr2.
SUMMARIZE
/TABLES=Rate Nr1 Nr2 FN1 FN2 VN1 VN2 GN2 GN2 GD1 GD2 PLZ1 PLZ2 Sex1 Sex2
/FORMAT=VALIDLIST NOCASENUM TOTAL
/TITLE='Record linkage results'
/MISSING=VARIABLE
```