Nutzungsmöglichkeiten des Kontrollnummernkonzepts epidemiologischer Krebsregister für Kohortenstudien in Deutschland

J. Kieschke, I. Wellmann, H. Hinrichs
Registerstelle des Epidemiologischen Krebsregisters Niedersachsen (EKN)
OFFIS, Escherweg 2, 26121 Oldenburg, e-mail: <name>@krebsregister-niedersachsen.de

Einleitung

Die hohen datenschutzrechtlichen Anforderungen in Deutschland müssen auch bei Kohortenstudien berücksichtigt werden. Zwar ist eine Forschung mit personenbezogenen Daten ohne Einwilligung der Betroffenen möglich [1], doch muß dabei der Grundsatz der Verhältnismäßigkeit zwischen informationellem Selbstbestimmungsrecht und epidemiologischen Forschungszielen gewahrt bleiben. Hierbei kann die Anwendung des in den epidemiologischen Krebsregistern in Deutschland entwickelten Kontrollnummernkonzeptes helfen und die Nutzung der dort erfaßten Daten erleichtern. Dabei nimmt ein Treuhänder, die "Vertrauensstelle", eine Einwegverschlüsselung der personenidentifizierenden Angaben vor, die Speicherung und Auswertung der Daten erfolgt in einer räumlich und personell getrennten "Registerstelle".

Methode

Zur Erhöhung der Fehlertoleranz werden die personenidentifizierenden Angaben in ihrer Schreibweise standardisiert, wozu auch die Bildung phonetischer Codes für die Namensangaben gehört. Anschließend werden die Angaben in eine bestimmte Anzahl von Einzelattributwerten zerlegt und diese jeweils per Einwegverschlüsselung (MD5) und anschließender symmetrischer Verschlüsselung (IDEA) in nicht dechiffrierbare "Pseudonyme", die sog. Kontrollnummern, umgewandelt (näheres in [2]).
Im EKN wird ein probabilistisches Record Linkage-Verfahren angewandt. Dabei wird geprüft, ob ein ermitteltes Übereinstimmungsmuster zweier Datensätze eher für oder gegen die Zugehörigkeit der Datensätze zu einer Person spricht. Würden z.B. die Kontrollnummern für Name, Vorname und Geburtstag verglichen, könnte als Übereinstimmungmuster entstehen, daß der Name unterschiedlich ist, der Vorname und der Geburtstag aber übereinstimmen. Anhand des jeweiligen Übereinstimmungsmusters wird ein Gewicht als Maß für die Wahrscheinlichkeit der Zusammengehörigkeit zweier Datensätze vergeben.

Zur Berechnung der Gewichte im EKN werden als „Matchvariablen" herangezogen:

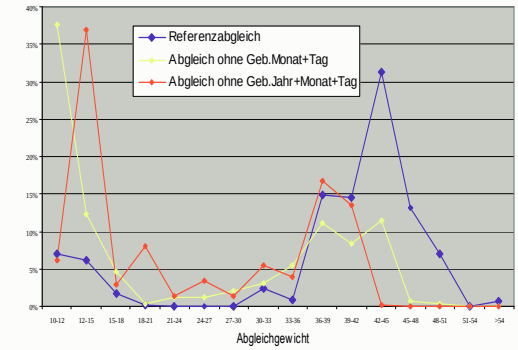

Abb. 1: Darstellung der prozentualen Verteilung von Abgleichgewichten > 10

jeweils ein Array über 3 Namens-, 3 Vornamens- und 3 Geburtsnamenskomponenten , Geburtstag, -monat, -jahr, Geschlecht und die Gemeindekennziffer.
Mit dem Gewicht wird versucht, die Gruppe zusammengehörender Datensätze von der Gruppe nicht zusammengehörender Datensätze zu trennen. Abbildung 1 zeigt die prozentuale Verteilung von ermittelten Gewichten. Es ergeben sich zwei Gipfel der Verteilungskurven, wobei der linke nur zum Teil dargestellt wird. Die blaue Kurve zeigt das Ergebnis eines durchschnittlichen Abgleichs. In einem Bereich zwischen einem Gewicht von 12 und 30 liegen nur wenige Abgleichgewichte. Aufgrund von Erfahrungswerten ist dies der derzeit festgelegte Bereich, in welchem Personenzuordnungen interaktiv von den Dokumentationskräften überprüft werden müssen. Unterhalb eines Abgleichgewichts von 12 wird keine Personenübereinstimmung angenommen, oberhalb von 30 werden Datensätze automatisch zusammengeführt. Die gelbe und rote Kurve zeigen Ergebnisse, bei denen nur Teile oder gar keine Informationen zum Geburtsdatum einbezogen wurden. Mit geringerem Informationsgehalt wird erwartungsgemäß die Trennung der beiden Gipfel unschärfer.

Fehlertoleranz und Suchstrategien

Tab. 1: Suchstrategien mit 7 Blockvariablen

Saarländisches Krebsregister	bisherige Vorgehensweise im EKN	neuer Vorschlag mit erhöhter Fehlertoleranz
pN, GT, GM, GJ, Ges	pN, pV, GT, GM, GJ	pN, pV, GT, GM, GJ
GT, GM, GJ, Ges	GT, GM, GJ	GT, GM, GJ
pN, GKZ, Ges	pN, pV, GKZ, Ges	pN, GKZ, Ges
pN, Ges	pN, pV	pN, pV
	pV, GT	pV, GT, GKZ
	pV, GM	pV, GM, GKZ
	pV, GJ	pV, GJ
		pN, GJ, GKZ
		pN, GT, GKZ

Abkürzungen der Blockvariablen:
GT: Geburtstag,
GM: Geburtsmonat,
GJ: Geburtsjahr,
Ges: Geschlecht,
pV: phonetischer Code Vorname,
pN: phonetischer Code Name,
GKZ: Gemeindekennziffer

Tab. 2: Vergleich der Fehlertoleranz der Suchstrategien von Tabelle 1

Abweichende Angaben in n der 7 Blockvariablen	Saarland	EKN (derzeit)	Neuer Vorschlag
	keine Fehlertoleranz bei Abweichung von:	keine Fehlertoleranz bei Abweichung von:	keine Fehlertoleranz bei Abweichung von:
n = 1	[Ges]	----	----
n = 2	[pN ∧ (GT ∨ GM ∨ GJ)] [Ges ∧ x]	[pV ∧ (GT ∨ GM ∨ GJ)]	----
n = 3	[pN ∧ (GT ∨ GM ∨ GJ) ∧ x] [Ges ∧ x ∧ x]	[pV ∧ (GT ∨ GM ∨ GJ) ∧ x]	[pV ∧ pN ∧ (GT ∨ GM ∨ GJ)], [pV ∧ (GT ∨ GM ∨ GJ) ∧ GKZ] [pN ∧ GJ ∧ GKZ]
n = 4	[pN ∧ (GT ∨ GM ∨ GJ) ∧ x ∧ x] [Ges ∧ x ∧ x ∧ x]	[pV ∧ (GT ∨ GM ∨ GJ) ∧ x ∧ x] [pN ∧ GT ∧ GM ∧ GJ]	[pV ∧ pN ∧ (GT ∨ GM ∨ GJ) ∧ x] [pN ∧ GT ∧ GM ∧ GJ] [pN ∧ GJ ∧ GKZ ∧ (GT ∨ GM ∨ Ges)] [pV ∧ GM ∧ GKZ (GT ∨ GJ)] [pV ∧ (GT ∨ GM ∨ GJ) ∧ Ges ∧ GKZ] [pV ∧ GT ∧ GJ ∧ (GKZ ∨ Ges)]

x = beliebige Blockvariable

Bei großen Datenmengen ist es nicht praktikabel, alle möglichen paarweisen Kombinationen von Kontrollnummern auf Gleichheit ihrer Ausprägungen zu überprüfen.
Deshalb wird die Strategie des Blocking verwandt, wobei durch Auswahl von einzelnen Merkmalen die Anzahl der zu vergleichenden Paare stark reduziert wird, für die dann jeweils mit den vollständigen Matchvariablen das Abgleichgewicht ermittelt wird.

Um trotzdem eine gute Fehlertoleranz zu erreichen, werden phonetische Codes statt der Namensangaben verwandt, und es werden in verschiedenen Umläufen unterschiedliche Merkmale zur Blockbildung herangezogen. Die gewählte Kombination von Blockvariablen ist entscheidend für die Effizienz und Fehlerrate beim Abgleich. Vergleichbare Verfahren können auch unabhängig von einer Verschlüsselung oder Kontrollnummernbildung eingesetzt werden.
Tabelle 1 zeigt in der ersten Spalte z.B. die im Saarländischen Krebsregister benutzten „Suchstrategien" zur Vermeidung von Doppelerfassungen. In Tabelle 2 wird aufgeführt, bei welchen Blockvariablenkombinationen abweichende Angaben zu keinem Match führen können.
So wird z.B. im Saarland das Geschlecht in allen 4 Suchstrategien einbezogen. Weiterhin würde dort bei jeder Abweichung des phonetischen Codes des Nachnamens in Verbindung mit abweichendem Geburtsdatum (Tag, Monat oder Jahr) kein Match möglich sein. In der dritten Spalte wird ein neuer Vorschlag unterbreitet, der eine hohe Trennschärfe bei verbesserter Fehlertoleranz bietet.

Einsatz von Kontrollnummern epidemiologischer Krebsregister in Kohortenstudien

Kohortenstudien mit personenbezogenen Daten ohne Einwilligung der Betroffenen bedürfen besonderer Maßnahmen, um dem Grundsatz der Verhältnismäßigkeit bei der Einschränkung des informationellen Selbstbestimmungsrechtes zu genügen. Mittels des in den epidemiologischen Krebsregistern Deutschlands angewandten Kontrollnummernkonzeptes kann für faktisch anonymisierte Datensätze ein Personenbezug hergestellt werden, ohne daß personenidentifizierende Angaben im Klartext vorliegen. Folgendes Vorgehen ermöglicht einen weitestgehenden Datenschutz (siehe Abbildung 2):

Bereits bei der Datenquelle, z.B. einer Datenbank mit arbeitsmedizinischen Angaben, werden die epidemiologisch relevanten Angaben von den personenidentifizierenden Angaben getrennt und beide Teildatensätze mit einer Referenznummer versehen. Mit dieser Referenznummer wird der epidemiologische Teildatensatz direkt zum auswertenden Studienzentrum übermittelt, während die personenidentifizierenden Angaben zur Vertrauensstelle des zuständigen Krebsregisters geschickt werden. Dort erfolgt nach Standardisierung der

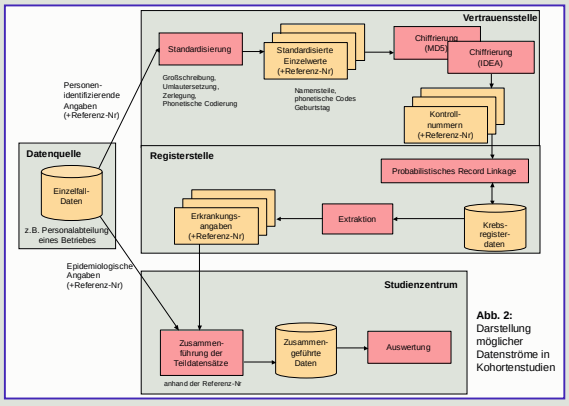

Abb. 2: Darstellung möglicher Datenströme in Kohortenstudien

Angaben eine Einwegverschlüsselung mit MD5, gefolgt von einer zweiten Verschlüsselung mittels eines registerspezifischen Schlüssels (IDEA). Die so gebildeten Kontrollnummern werden zusammen mit den Referenznummern zur Registerstelle übermittelt, wo ein probabilistisches Record Linkage durchgeführt wird. Im Registerdatenbestand befindliche Angaben zu Krebserkrankungen in der untersuchten Kohorten- bzw. Kontrollgruppe werden zusammen mit den Referenznummern an das Studienzentrum zur Auswertung geschickt.

Bei dieser Vorgehensweise werden personenidentifizierende Angaben nur kurzfristig in der Vertrauensstelle verarbeitet, ohne daß die Sachbearbeiter dabei wissen müssen, ob die Personen zur Studien- oder zur Kontrollgruppe gehören. Die beiden ersten Arbeitsschritte (Standardisierung der Angaben und Einwegverschlüsselung mit MD5) sind auch bei der Datenquelle durchführbar, so daß nur bereits verschlüsselte personenidentifizierende Angaben weitergegeben werden. Dies impliziert jedoch einen erhöhten Arbeitsaufwand vor Ort, also z.B. in einer Personalabteilung.

Diskussion und Zusammenfassung

Das Konzept der Kontrollnummern der epidemiologischen Krebsregister in Deutschland bietet nicht nur sehr gute Möglichkeiten des Datenschutzes, sondern erreicht in Verbindung mit einem probabilistischen Record Linkage auch effiziente Abgleichergebnisse. Natürlich besteht die Möglichkeit, daß aufgrund der Verschlüsselung Ähnlichkeiten von Angaben, wie Vornamen und ihre Kurzformen, nicht erkannt werden und eine Zusammenführung von Datensätzen unterbleibt. Dies dürfte jedoch sehr selten der Fall sein. Wesentlicher erscheint, daß bei der Vorauswahl potentieller Matchpartner stärker auf eine hohe Fehlertoleranz geachtet wird. Dies läßt sich durch mehrere Umläufe mit unterschiedlichen Blockvariablen erreichen.

Zusammenfassend erscheint dieses Konzept ideal für die Durchführung von Kohortenstudien in Zusammenarbeit mit den epidemiologischen Krebsregistern in Deutschland geeignet zu sein. Die einheitliche Definition von Kontrollnummern und der Einsatz eines einheitlichen Softwareproduktes zu ihrer Generierung in den Krebsregistern erleichtert die Einbeziehung unterschiedlicher Bundesländer in eine Studie. Derzeit wird mittels dieses Kontrollnummernkonzeptes auch der länderübergreifende Abgleich zwischen den Krebsregistern erprobt.

Literatur
[1] Wichmann H.-E.: Epidemiologie und Datenschutz - Wege zur partnerschaftlichen Zusammenarbeit. Informatik, Biometrie und Epidemiologie in Medizin und Biologie 30 (1) 1999.
[2] Appelrath H.-J., Michaelis J., Schmidtmann I., Thoben W.: Empfehlung an die Bundesländer zur technischen Umsetzung der Verfahrensweisen gemäß Gesetz über Krebsregister (KRG). Informatik, Biometrie und Epidemiologie in Medizin und Biologie 27 (2),1996