Evaluation des Krebsregisters NRW
Schwerpunkt Record Linkage

Irene Schmidtmann, Gaël Hammer,
Murat Sariyar, Aslihan Gerhold-Ay

Institut für Medizinische Biometrie, Epidemiologie und Informatik

Universitätsmedizin der Johannes Gutenberg-Universität Mainz,
Körperschaft des öffentlichen Rechts

Abschlussbericht

11.6.2009

Inhaltsverzeichnis

1 Zusammenfassung

Ziel

Dieser Abschlussbericht beschreibt die im Krebsregistergesetz NRW vorgesehene Evaluation des Krebsregisters NRW durch eine Arbeitsgruppe des Instituts für Medizinische Biometrie, Epidemiologie und Informatik der Universitätsmedizin Mainz, des Deutschen Kinderkrebsregisters und des Krebsregisters Rheinland-Pfalz. Die Evaluation wurde durchgeführt von September 2008 bis Mai 2009. Schwerpunkt der Evaluation war – gemäß Gesetz und Ausschreibung – die Bewertung des Record Linkage im Krebsregister NRW.

Methoden

Durch die Arbeitsgruppe wurde aus einer für diesen Zweck zur Verfügung gestellten Stichprobe von 150.000 Meldungen des Krebsregisters NRW mit Klartext-Identitätsdaten ein Gold-Standard erstellt. Die dort erzielten Zuordnungen von Meldungen zu Personen wurden verglichen mit den im Krebsregister NRW vorgenommenen Zuordnungen.

Ergebnis

Dabei zeigte sich, dass die 150.000 Meldungen zu 132.267 Personen gehörten. Aufgrund von Synonymfehlern (mehrere Meldungen zu einer Person nicht als zusammengehörend erkannt) waren 268 Fälle zu viel gezählt worden, teilweise kompensiert dadurch, dass aufgrund von Homonymfehlern (Meldungen zu verschiedenen Personen fälschlich einer Person zugeordnet) 20 Fälle zu wenig gezählt wurden. Daraus ergibt sich ein Nettofehler von 248 Fällen, das entspricht einer Überschätzung der Fallzahl um 0,19%. Unter den 14.049 Fällen mit mehr als einer Meldung traten 1,91% Synonyme auf, bezogen auf alle 150.000 Meldungen waren es lediglich 0,18%. Die Homonymfehlerrate bezogen auf die 132.267 verschiedenen Personen betrug 0,015%.

Eine Hochrechnung auf größere Datenbestände ergab, dass die Homonymfehlerrate erst bei einem Datenbestand von über 5.000.000 ins Record Linkage eingehenden Meldungen auf 1% steigt.

Eine genauere Untersuchung der Synonymfehler ergab, dass die Inzidenzmeldungen aus der onkologischen Qualitätssicherung, die eine höhere Datenqualität als die Meldungen aus den anderen Meldequellen aufwiesen, besonders selten an Synonymfehlern beteiligt waren. Bei einer Zunahme der Mehrfachmeldungen auf durchschnittlich bis zu drei pro Person kann die Synonymfehlerrate bis auf etwa 2,4% steigen.

Der Zugewinn durch den Einsatz von Stringmetriken ist für das Record Linkage im Krebsregister zu vernachlässigen, die Verwendung von Kontrollnummern statt Klartextangaben liefert – eine gute Datenaufbereitung vorausgesetzt – ein weitestgehend gleiches Ergebnis.

Empfehlungen

Maßnahmen zur Verbesserung der Qualität der Identitätsdaten der Meldequellen Pathologiebefunde und direkte Inzidenzmeldungen lassen eine Reduktion der Synonymfehlerrate erwarten und sind daher zu empfehlen. Weiter ist sehr zu wünschen, dass die im Rahmen der onkologischen Qualitätssicherung aufgebaute Dokumentation von Tumorerkrankungen auf qualitativ hohem Niveau fortgeführt wird, auch wenn die Onkologischen Schwerpunkte geschlossen werden.

Die Überprüfung der Verfahrensweisen ergab, dass die Abläufe sinnvoll und durchdacht sind. Einzelne Fragen, etwa das genaue Vorgehen beim Abgleich von Sterbemeldungen mit den

Todesursachendaten, werden im weiteren Verlauf des Aufbaus des Krebsregisters noch zu klären sein.

Es ist anzuraten, dass die umfassende und informative Dokumentation der Arbeitsabläufe bei Einführung neuer Prozesse oder der Änderung von Verfahrensweisen aktualisiert wird.

2 Einleitung

Dieser Abschlussbericht beschreibt die Evaluation des Krebsregisters NRW durch eine Arbeitsgruppe des Instituts für Medizinische Biometrie, Epidemiologie und Informatik der Universitätsmedizin Mainz, des Deutschen Kinderkrebsregisters und des Krebsregisters Rheinland-Pfalz[1]. Diese Evaluation ist im Gesetz zur Einrichtung eines flächendeckenden bevölkerungsbezogenen Krebsregisters in Nordrhein-Westfalen (EKR-NRW) [1] vom 05. April 2005 in §11 (2) vorgesehen. Die Evaluation wurde durchgeführt von September 2008 bis Mai 2009. Schwerpunkt der Evaluation war – gemäß Gesetz und Ausschreibung – die Bewertung des Record Linkage im Krebsregister NRW. Das Record Linkage im Krebsregister NRW findet mit chiffrierten Daten statt. Zum Zwecke der Evaluierung wurden Identitätsdaten im Klartext verwendet, um bestmögliche Ergebnisse zu erzielen. Damit konnten durch die Arbeit mit verschlüsselten Daten eventuell entstandene Fehler erkannt und quantifiziert werden.

Auch mit Klartextdaten können Record Linkage-Fehler gemacht werden, jedoch ist eine sorgfältige Arbeit mit Klartextdaten, bei der möglichst viele potenziell zusammen gehörende Datensätze berücksichtigt und zusätzliche Informationen (Straßenangabe, Diagnosen, Telefonbuch) herangezogen werden, als Gold-Standard anzusehen. Für die Herleitung und Berechnung von Fehlerraten wird daher im folgenden davon ausgegangen, dass der Gold-Standard korrekt ist.

Ein weiterer Aspekt der Evaluation war die Überprüfung der Verfahrensweisen. Dazu wurden Dokumente, die die Arbeitsprozesse im Krebsregister NRW beschreiben, kritisch gesichtet und vor Ort die Arbeitsprozesse nachvollzogen. Diese Aufgabe wurde von aktuellen und ehemaligen Mitarbeiterinnen und Mitarbeitern des Krebsregisters Rheinland-Pfalz übernommen.

Wir danken den Kolleginnen und Kollegen im Krebsregister NRW für das uns entgegen gebrachte Vertrauen und die angenehme Kooperation. Wir haben alle Daten und Dokumente, um die wir gebeten haben, zügig erhalten. Alle Fragen, die wir stellten, wurden umfassend beantwortet.

[1] Folgende Personen haben an der Evaluation des Krebsregisters mitgearbeitet: Andreas Borg, Monika Decher-Neff, Katharina Emrich, Aslihan Gerhold-Ay, Gaël Hammer, Martina Hick, Gabriele Husmann, Sabine Rost, Murat Sariyar, Irene Schmidtmann, Gerhard Seebauer, Susanna Siebert, Manuel Sudhof, Ursula Sudhof, Claudia Trübenbach, Franziska Wandtner

3 Daten und Methoden

3.1 Record Linkage

3.1.1 Beschreibung der Stichprobe

Das IMBEI hat vom Krebsregister NRW am 30.09.2008 150.000 Datensätze erhalten. Diese Datensätze enthielten einen (unten beschriebenen) Auszug aus Meldungen an das Krebsregister NRW. Folgende Kriterien sollten für die Meldungen in der Stichprobe berücksichtigt werden:

- Jahre: 2006-2008
- 100.000 Meldungen aus den Quellen Inzidenzmeldungen (IMD), Inzidenzmeldungen aus der Onkologischen Qualitätssicherung (IMO) und Pathologiebefunden (PBF), dabei sollten 15% IMD-Meldungen, 35% IMO-Meldungen und 50% Pathologenmeldungen sein
- 50.000 Sterbefallmeldungen, gezogen aus allen Sterbefallmeldungen. Der Anteil der zugeordneten Meldungen darunter sollte auch dem Anteil der zugeordneten insgesamt entsprechen, d. h. von den 50.000 Sterbefallmeldungen sollten ca. 11,2% zu Inzidenzmeldungen oder Pathologiebefunden zugeordnet sein.

Die Dateien sollten folgende Merkmale (soweit vorhanden) enthalten:

- Personenidentifizierende Merkmale (***alle*** Daten im Klartext): Name, Vorname, Titel, Geburtsname, früherer Name, Geschlecht, Geburtstag, Geburtsmonat und -jahr, Anschrift: Straße und Hausnummer, Postleitzahl und Wohnort, Staatsangehörigkeit, Referenz, Sterbedatum, Sterbeort, Todesursache, beurkundendes Standesamt, Sterbebuchnummer
- Kontrollnummern
- Medizinisch-epidemiologische Merkmale: Diagnosedatum, verschlüsselte Tumordiagnose (ICD 10), ICD-03 Morphologie und Topographie.

Informationen über die Zuordnung mehrerer Datensätze zu einer Person (Personen-ID)

Zusätzlich wurde eine Laufnummer und eine Information über die Art der Meldung (IMD, IMO, PBF, MA) angefordert.

Es wurden übermittelt:

50.000 Datensätze von Pathologiebefunden (kurz: PBF)

15.000 Datensätze von direkten Inzidenzmeldungen (kurz: IMD)

35.000 Datensätze von Meldungen über die Onkologische Qualitätssicherung (kurz: IMO)

44.400 Datensätze von nicht zugeordneten Sterbemeldungen (kurz: MANZ)

5.600 Datensätze von zugeordneten Sterbemeldungen (kurz: MA)

Jeder Datensatz war in einer eigenen Datei gespeichert, Gruppen von Dateien in ZIP-Archiven zusammengefasst.

3.1.2 Sicherheitsmaßnahmen zur Übermittlung und im IMBEI

An das IMBEI übermittelte das Krebsregister NRW direkt die epidemiologischen Daten und Kontrollnummern der personen-identifizierenden Daten. Die asymmetrisch chiffrierten Identitätsdaten wurden vom Krebsregister NRW zunächst an die Ärztekammer Westfalen-Lippe übermittelt, dort wurden aus den Identitätschiffraten Klartexte erzeugt. Die so dechiffrierten Daten wurden dann an das IMBEI übermittelt. Für den Transport wurden sie wiederum asymmetrisch mit WinPT/GPG verschlüsselt. Dazu wurde ein spezieller Public Key des

IMBEI verwendet, so dass die übermittelten Daten nur durch am Projekt beteiligte Mitarbeiter des IMBEI dechiffriert werden konnten.

Am IMBEI waren die übermittelten Daten auf einem Notebook gespeichert, das während der Projektlaufzeit nur für die Evaluation verwendet wurde und keinen Zugang zum Internet hatte. Die Daten waren auf einer mit TrueCrypt chiffrierten Festplatten-Partition gespeichert, auf die nur Projektmitarbeiter zugreifen konnten. Weitere organisatorische Maßnahmen gewährleisteten, dass kein Unbefugter Zugriff auf die Daten erhalten konnte. Wenn nicht mit dem Notebook gearbeitet wurde, wurde es in Räumen des Deutschen Kinderkrebsregisters am IMBEI aufbewahrt.

Nach Beendigung des Projekts wurden die Daten in verschlüsselter Form auf CD gebrannt und im Safe des Krebsregister Rheinland-Pfalz bis 31.12.2009 aufbewahrt. Die verschlüsselte Partition auf dem Notebook wird noch im Juni 2009 sicher gelöscht. Nach Ablauf der Frist wird der Datenträger zerstört.

3.1.3 Datenaufbereitung

Vor dem Datenabgleich waren einige Schritte zur Aufbereitung der Daten nötig. Zuerst wurden die Datensätze mit den Klartexten der personen-identifizierenden Daten, die in je einer XML-Datei abgelegt waren, in fünf Dateien (IMD, IMO, PBF, MA und MANZ) zusammengefasst. Anschließend wurden die Daten standardisiert und bereinigt. Schließlich wurden die bearbeiteten Daten zu zwei Dateien mit Inzidenzmeldungen (IMD, IMO und PBF) bzw. Sterbefallmeldungen (MA und MANZ) zusammengefasst.

Das Prozedere der Datenaufbereitung im einzelnen wurde bereits im Zwischenbericht beschrieben.

3.1.4 Erzeugung des Gold-Standards aus den Klartextdaten

Potenziell zusammengehörige Paare von Datensätzen (Matches) aus den Dateien mit Inzidenz- und Sterbefallmeldungen wurden mit mehreren Verfahren identifiziert:

1. Stochastisches Record Linkage mit Automatch 4.3 [2]
2. Stochastisches Record Linkage mit MTB 0.61 [3]
3. Stochastisches Record Linkage auf Basis eines EM-Verfahrens (Expectation-Maximization-Algorithmus) [4]

Mit jedem dieser Verfahren, die im Folgenden genauer beschrieben werden, wurden gesucht:

1. zusammengehörige Datensätze in den Inzidenzmeldungen (Deduplizierung) und
2. Datensatz-Paare aus Inzidenz- und Sterbefallmeldungen (Matching).

Das Ergebnis dieser Suche sind Listen von Paaren von Zuordnungsnummern, versehen mit einem Indikator für „sicherer Match“ bzw. „potenzieller Match“. Anhand dieser Listen wurden jeweils alle möglicherweise zu einer Person gehörenden Datensätze zu „Matchgruppen“ zusammengefasst.

Die Matchgruppen wurden in folgender Weise identifiziert: In graphentheoretischer Terminologie kann ein Datensatz als Knoten eines Graphen bezeichnet werden. Identifiziert mindestens eines der obigen Verfahren ein Paar von Datensätzen als potenziell oder sicher zusammengehörend, so ist zwischen den Knoten eine Kante gegeben. Die Aufgabe, jeweils alle möglicherweise zu einer Person gehörenden Datensätze zu finden, bedeutet, die Zusammenhangskomponenten des Graphen zu finden. Dies wurde in SAS mit einer Breitensuche realisiert.

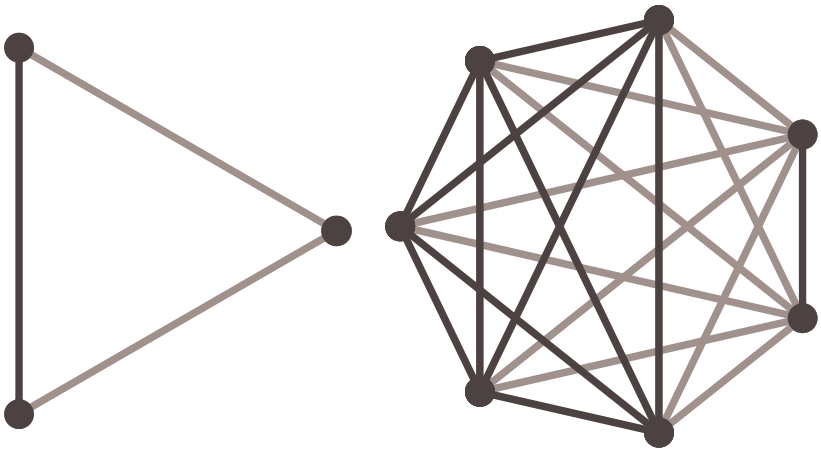

Abbildung 1: Beispiel für das Record-Linkage. Hier sind Meldungen als Ecken eines Graphen dargestellt. Im Krebsregister als zusammengehörend entschiedene Meldungen sind durch eine schwarze Kante verbunden. Grau ist das Ergebnis im Gold-Standard. Im linken Beispiel kennt das Krebsregister zwei Personen, darunter eine mit 2 Meldungen, während im Gold-Standard als eine Person mit 3 Meldungen zusammengeführt wurden. Im rechten Beispiel hat das Krebsregister die Meldungen zwei Personen zugeordnet, einer 5 Meldungen, der anderen zwei Meldungen. Der Gold-Standard hat alle 7 Meldungen einer Person zugeordnet.

Diese Matchgruppen wurden – abweichend vom Studienprotokoll, das bei den perfekten Matches nur Stichproben vorsah, – sämtlich der Durchsicht („Clerical Review“) durch Dokumentationspersonal des Deutschen Kinderkrebsregisters und wissenschaftliche Hilfskräfte des IMBEI zugeführt. Dazu wurden die Datensätze um Diagnose- und Sterbedatum, sowie die ICD-Codes zu den Tumoren ergänzt. Zusätzlich wurden die Namensfelder in (bis zu drei) Namenskomponenten zerlegt und die Kölner Phonetik daraus errechnet. Das Clerical Review wird in Abschnitt 3.1.7 beschrieben.

3.1.5 Stochastisches Record Linkage auf Basis eines EM-Verfahrens

Folgende Daten gehen in die Prozedur ein:

- Vorname (erste und zweite Komponente)
- Nachname (erste und zweite Komponente)
- Geschlecht
- Geburtstag
- Geburtsmonat
- Geburtsjahr
- PLZ und Wohnort

Zunächst sind aus den einzelnen Datensätzen Datensatzpaare zu bilden. Die Bildung von Datensatzpaaren und Blöcken wird im Datenbanksystem PostgreSQL realisiert. Aus diesen Datensatzpaaren werden Vergleichsmuster auf Basis von binären Vergleichen und der Verwendung der Jaro-Winkler-String-Metrik [5;6] erzeugt. Folgende Blöcke werden gebildet, wobei Vor- und Nachnamen (VN und NN) jeweils in Form der Kölner Phonetik in die Blockbildung eingehen:

Tabelle 1: Blockbildung

Stufe	Nachname	Vorname	Geschlecht	Geburts-tag	Geburts-monat	Geburts-jahr
1	X	X		X	X	X
2		X		X		
3		X			X	
4		X				X
5				X	X	X
6	X		X			

Darauf aufbauend erfolgt die Berechnung der u- und m-Gewichte[2] pro Datensatzpaar im Rahmen des Fellegi-Sunter-Modells auf Grundlage eines EM-Verfahrens, welches die Wahrscheinlichkeit für die Zellen einer Kontingenztafel schätzt. EM steht dabei für Expectation-Maximization und stellt eine Erweiterung der klassischen Überlegungen von Maximum-Likelihood-Verfahren dar. Konkret bedeutet das in diesem Fall, dass aufgrund der beobachtbaren Kontingenztabelle bei Vorliegen einer binären latenten Variable (mit Ausprägungen Duplikat und Non-Duplikat) iterativ mit Hilfe einer Erwartungswertbildung gemäß der Prozedur von Espeland und Odoroff [4] die vollständige doppelt so lange Kontingenztafel geschätzt wird, um anschließend eine Maximierung der Wahrscheinlichkeiten gemäß der iterativen Scaling-Methode von Darroch und Ratcliff [7] vorzunehmen. Zum Schluss werden manuell obere und untere Schranke für die endgültige Klassifikation der Datensatzpaare als Non-Duplikat, potenzielles und sicheres Duplikat bestimmt.

3.1.6 Stochastisches Record Linkage: Abgleich-Strategie

Das stochastische Record Linkage wurde mit Software durchgeführt, die nach dem Modell von Fellegi und Sunter [8] arbeitet. Das Ergebnis des Vergleichs eines Paares von Datensätzen ist dabei ein Score, der umso höher ist, je ähnlicher die Datensätze sind. Dieser Score ist nicht normiert. Seine Höhe hängt von der Anzahl zu seiner Berechnung verwendeter Variablen und der Häufigkeit der einzelnen Ausprägungen der Variablen ab.

Nach einer groben Sichtung der Ergebnisse wurden der Schwellenwert 17 für „potenzieller Match“ und 40 für „sicherer Match“ festgelegt. Die untere Schwelle ist dabei bewusst niedrig gesetzt. Es wurde außerdem noch eine Stichprobe von 4000 Datensatzpaaren mit 0< Score < 17 gesichtet.

Die Deduplizierung der Inzidenzmeldungen und das Matching von Sterbefall- zu Inzidenzmeldungen verwenden dasselbe Schema. Es wurde in 10 Stufen vorgegangen, wobei jede Stufe unabhängig von den Ergebnissen der vorangegangenen war. In jeder Stufe wurden dieselben Matchvariablen zur Bildung des Scores herangezogen. Je Stufe wurden unterschiedliche Blockvariablen festgelegt, die steuern, welche Variablen übereinstimmen *müssen,* damit Datensätze verglichen werden. Sukzessive wurden diese Kriterien gelockert (siehe Tabelle 4).

Zum Vergleich der Ausprägungen der Matchvariablen wurde in Automach die exakte Übereinstimmung gefordert (analog zur Arbeit mit Kontrollnummern). MTB bietet darüber hinaus die Möglichkeit, Stringvergleiche zu verwenden. Daher wurden die Verfahren mit MTB in drei Varianten durchgerechnet: a) mit exaktem Vergleich (als Vergleich zu Automatch), b) mit der Levenshtein-Damerau-Distanz [9;10], c) mit Bigrammen [11].

[2] Mit *m* wird die Wahrscheinlichkeit bezeichnet, Übereinstimmung bezüglich eines Merkmals zu beobachten, wenn zwei Datensätze sich auf eine Person beziehen, mit *u* die Wahrscheinlichkeit, Übereinstimmung bezüglich eines Merkmals zu beobachten, wenn zwei Datensätze sich auf zwei verschiedene Personen beziehen.

Das primäre Ziel des Einsatzes verschiedener Stringmetriken war es, einen möglichst vollständigen Datensatz für das Clerical Review zusammenzustellen. Darüber hinaus konnte in diesem Projekt auch verglichen werden, welche Vorteile diese Klartext-basierten Verfahren gegenüber dem in Krebsregister eingesetzten Kontrollnummern-basierten Verfahren bieten. Dies geschah durch die Gegenüberstellung der Treffermengen (potenzielle und sichere Matches) der Verfahren und der Berechnung der Anzahl der Meldungspaare, die ausschließlich in einem der Verfahren gefunden wurden.

Tabelle 2: Variablenliste der Dateien mit Sterbefall- und Inzidenzmeldungen

Variable	**Erläuterung**
Datei	*Dateiname IMD IMO PBF MA MANZ*
Nummer	*lfd. Nummer innerhalb der Datei*
	Bildet zusammen mit Datei die Datensatz-ID
Geschlecht	Geschlecht
Vorname	[3]
Nachname	[3]
Titel	Titel
GeburtsTag	Geburtstag
GeburtsMonat	Geburtsmonat
GeburtsJahr	Geburtsjahr
Strasse	
Nr	
PLZ	Postleitzahl
Wohnort	
Vorname_C1	Vorname, 1. Komponente
Vorname_C2	Vorname, 2. Komponente
Vorname_C3	Vorname, 3. Komponente
Vorname_PC	Vorname, Kölner Phonetik
Nachname_C1	Name, 1. Komponente
Nachname_C2	Name, 2. Komponente
Nachname_C3	Name, 3. Komponente
Nachname_PC	Name, Kölner Phonetik

[3] Aus den Variablen Vorname und Nachname sind neue Variablen abgeleitet worden: 1) Zerlegung in bis zu 3 Komponenten, 2) die Kölner Phonetik.

Tabelle 3: Matchvariablen

Variable	m[4]	u[4]
Vornamen (alle 3 Komponenten)	0,95	0,01
Nachnamen (alle 3 Komponenten)	0,95	0,01
Geschlecht	0,98	0,50
Geburtstag	0,98	0,03
Geburtsmonat	0,98	0,08
Geburtsjahr	0,98	0,02
PLZ	0,95	0,01

In den ersten Stufen des Abgleichs wird die exakte Übereinstimmung der meisten wichtigen Merkmale gefordert, wobei in der Regel je ein Merkmal ausgelassen wird. In späteren Stufen werden diese Regeln relaxiert. Das hat zur Folge, dass in jeder dieser Stufen eine wesentlich größere Anzahl Meldungen verglichen wird, mit einer potenziell entsprechend größeren Treffermenge. Es ist also zu vermuten, dass Treffermengen der späteren Stufen diejenigen der frühen Stufen beinhalten und somit die ersten Stufen ausgelassen werden könnten. Dieses wurde durch Gegenüberstellen der Treffermengen evaluiert.

Tabelle 4: Blockvariablen

Stufe	Nachname, phonetisch	Vorname, phonetisch	Geschlecht	Geburtstag	Geburtsmonat	Geburtsjahr	PLZ	Wohnort
1	X	X	-	X	X	X	-	-
2	X	-	-	X	X	X	-	-
3	X	-	X	-	-	-	X	-
4	X	-	X	-	-	-	-	X
5	-	-	-	X	X	X	-	X
6	-	X	-	X	-	-	-	-
7	-	X	-	-	X	-	-	-
8	-	X	-	-	-	X	-	-
9	-	-	-	X	X	X	-	-
10	X	-	X	-	-	-	-	-

3.1.7 Clerical Review

Die Ergebnisse aller Ansätze aus 3.1.4 des Record Linkage wurden für das Clerical Review wie folgt aufbereitet: Es wurde mit SAS eine Excel-Datei erzeugt und formatiert, in der die zu einer Matchgruppe zusammengehörigen Datensätze untereinander zusammengefasst waren. Matchgruppen wurden durch eine Leerzeile und zusätzlich durch einen Strich getrennt. Abweichungen der Variablen-Ausprägungen innerhalb einer Matchgruppe wurden durch ihre Formatierung (Fett- bzw. Kursivschrift) gekennzeichnet, so dass Unterschiede optisch auffäl-

[4] m- und u-Werte des Fellegi-Sunter Modells. Mit *m* wird die Wahrscheinlichkeit bezeichnet, Übereinstimmung bezüglich eines Merkmals zu beobachten, wenn zwei Datensätze sich auf eine Person beziehen, mit *u* die Wahrscheinlichkeit Übereinstimmung bezüglich eines Merkmals zu beobachten, wenn zwei Datensätze sich auf zwei verschiedene Personen beziehen. Die angegebenen Voreinstellungen für u-Werte wurden vom Programm durch die inversen Häufigkeiten der Variablen-Ausprägungen ersetzt

lig waren und leicht wahrgenommen werden konnten. Für den Ausdruck wurden die Wiederholung der Spaltenköpfe und eine Seitennummerierung eingestellt.

Um zu ermöglichen, dass sich mehrere Personen gleichzeitig am Clerical Review beteiligen, wurde die Excel-Datei ausgedruckt. Die Ausdrucke im Format DIN-A4-Quer wurden auf DIN-A3 vergrößert, so dass ein Schriftgrad von ca. 10 Punkt zustande kam. Die Ausdrucke und Kopien wurden in einem abgeschlossenen Container im Kinderkrebsregister aufbewahrt und nach Abschluss des Clerical Review der datenschutzgerechten Vernichtung zugeführt.

Mehrere Mitarbeiterinnen und Mitarbeiter (Dokumentationspersonal des Kinderkrebsregisters und wissenschaftliche Hilfskräfte) sichteten diese Gruppen und trafen die Entscheidung, ob sich die Datensätze auf eine oder mehrere Personen beziehen. Zweifelsfälle wurden in der Gruppe diskutiert und dann entschieden.

Tabelle 5:Fiktives Datenbeispiel für das Clerical Review

Gruppe	Geschlecht	Vorname	Nachname	Titel	GeburtsDatum	Strasse	Nr	PLZ	Wohnort	TodesDatum	DiagnoseDatum	ICD_10	Topo	Morph
9876	M	FRANZ	MUSTERMANN		13.02.1939	MUSTERWEG	16	12345	MUSTERSTADT		11.2004	C34.9	C349	80423
9876	**W**	**LUDWIG FRANZ**	MUSTERMANN		13.02.1939	MUSTERWEG	16	12345	MUSTERSTADT		11.2004	C34.9	C349	80423
9876	*U*	FRANZ	**MUSERMANN**		**13.02.1938**	MUSTERWEG	16	12345	MUSTERSTADT	18.11.2008				

Die Zuordnungen wurden in die Datenbank eingegeben. Danach wurde überprüft, wo Abweichungen zu den in Münster getroffenen Entscheidungen auftraten. Die abweichend entschiedenen Gruppen von Datensätzen wurden nochmals überprüft. In einigen Fällen wurde die Entscheidung revidiert. Außerdem war zwischenzeitlich sowohl in Mainz als auch in Münster aufgefallen, dass es auch unter den Meldeamtsdaten Dubletten gab, was ursprünglich nicht erwartet worden war, so dass einige zusammengehörige Datensatzpaare zunächst überhaupt nicht verglichen worden waren. Die revidierten Entscheidungen wurden in die Datenbank eingegeben, so dass es jetzt möglich war, detaillierte Auswertungen der Record Linkage Ergebnisse vorzunehmen.

3.1.8 Methodische Überlegungen zur Beurteilung von Record Linkage Fehlern

Wie eingangs erwähnt, wird bei der Beurteilung des Record Linkage und der Wahl der Bezeichnungen davon ausgegangen, dass die Entscheidungen im Gold-Standard korrekt sind. Die Entscheidungen des EKR NRW werden damit verglichen. Wie oben beschrieben, werden die Meldungen als Ecken eines ungerichteten Graphen aufgefasst, Kanten sind gegeben, wenn zwei Meldungen als zu einer Person gehörend klassifiziert werden.

Gehört zu einer Person nur eine Meldung, bleibt die entsprechende Ecke im Graphen ohne Kanten. Existieren zwei Meldungen zu einer Person, so ergibt sich genau eine Kante zwischen den beiden entsprechenden Ecken. Allgemein gilt, dass sich beim Vorliegen von m Meldungen zu einer Person $\frac{m(m-1)}{2}$ Kanten ergeben, d. h. wir nehmen – unabhängig davon, wie die Zuordnung gefunden wurde – an, dass die Meldungen zu einer Person einen vollständigen Teilgraphen bilden.

Bezeichnungen:

Es seien

n_{Gi} die Anzahl der Personen im Gold-Standard, zu denen genau *i* Meldungen vorliegen,

I_G die maximale Anzahl Meldungen für eine Person im Gold-Standard,

n_{Mi} die Anzahl der Gruppen von *i* Meldungen, die im EKR NRW zu einer Person zusammengeführt wurden, d. h. die Anzahl der Matchgruppen aus *i* Meldungen.

I_M die maximale Größe einer Meldungsgruppe im EKR NRW.

Es gilt:

$N_G = \sum_{i=1}^{I_G} n_{Gi}$ für die Anzahl der Personen im Gold-Standard,

$N_M = \sum_{i=1}^{I_M} n_{Mi}$ für die Anzahl der Meldungsgruppen im EKR NRW,

$N = \sum_{i=1}^{I_G} i \cdot n_{Gi} = \sum_{i=1}^{I_M} i \cdot n_{Mi}$ für die Anzahl der Meldungen,

$K_G = \sum_{i=1}^{I_G} \frac{i(i-1)}{2} \cdot n_{Gi}$ für die Anzahl Kanten im Gold-Standard,

$K_M = \sum_{i=1}^{I_M} \frac{i(i-1)}{2} \cdot n_{Mi}$ für die Anzahl Kanten im EKR NRW,

$K = \frac{N(N-1)}{2}$ für die Anzahl aller möglichen Kanten unter N Meldungen

Synonyme – Zuordnungen, die im Gold-Standard erkannt wurden, aber nicht im EKR NRW

Wie viele Personen zu viel wurden aufgrund von Synonymfehlern gezählt?

Es sei $s_{ij}\,(1 \le j \le i)$ die Häufigkeit mit der *j* Gruppen aus *i* Meldungen zu einer Person des Gold-Standards im EKR NRW gebildet wurden. (Also: *i* Meldungen zu einer Person liegen „tatsächlich“ im Gold-Standard vor, die Meldungen werden aber im EKR NRW *j* Personen zugeordnet d. h. sie erscheinen in *j* Meldungsgruppen.) Es gilt $n_{Gi} = \sum_{j=1}^{i} s_{ij}$, d. h. die Anzahl der Personen mit *i* Meldungen setzt sich zusammen aus s_{i1} Gruppen, die korrekterweise als zusammengehörig erkannt wurden, aus s_{i2} Gruppen, von denen angenommen wird, dass sie sich auf 2 Personen beziehen, aus s_{i3} Gruppen, von denen angenommen wird, dass sie sich auf 3 Personen beziehen usw. Wenn Synonymfehler auftreten, hat der entstehende Graph zu wenige Kanten. Für eine detaillierte Darstellung siehe Tabelle 17 im Anhang.

Die Zahl der Synonyme insgesamt ergibt sich als $N_S = \sum_{i=2}^{I_G} \sum_{j=2}^{i} (j-1) s_{ij}$

Zur Quantifizierung der Synonymfehlerrate kommen mehrere Ausdrücke in Frage:

1. Zahl der Synonyme bezogen auf alle Fälle. Diese Zahl gibt an, um welchen Anteil die Fallzahl zu hoch geschätzt wird dadurch, dass Synonymfehler auftreten. Diese Zahl gibt die Auswirkung der Synonymfehler auf das Register an.

$$S_1 = \frac{N_S}{N_G} = \frac{\sum_{i=2}^{I_G} \sum_{j=2}^{i} (j-1) s_{ij}}{\sum_{i=1}^{I_G} n_{Gi}}$$

2. Zahl der Synonyme bezogen auf die Fälle mit mindestens zwei Meldungen. Diese Zahl gibt die Zahl der Synonyme an, bezogen auf die Zahl der Personen, bei denen Synonymfehler auftreten können. Diese Zahl sagt etwas über die Güte des Record Linkage-Verfahrens aus.

$$S_2 = \frac{N_S}{\sum_{i=2}^{I_G} n_{Gi}} = \frac{\sum_{i=2}^{I_G} \sum_{j=2}^{i} (j-1) s_{ij}}{\sum_{i=2}^{I_G} n_{Gi}}$$

3. Anteil der Fälle mit Mehrfachmeldungen, bei denen tatsächlich Synonyme auftreten – egal ob eines oder mehrere:

$$S_3 = \frac{\sum_{i=2}^{I_G} \sum_{j=2}^{i} s_{ij}}{\sum_{i=2}^{I_G} n_{Gi}} = \frac{\sum_{i=2}^{I_G} (n_{Gi} - s_{i1})}{\sum_{i=2}^{I_G} n_{Gi}}$$

4. Anteil der Meldungen, die Synonyme sind:

$$S_4 = \frac{N_S}{N} = \frac{\sum_{i=2}^{I_G} \sum_{j=2}^{i} (j-1) s_{ij}}{\sum_{i=1}^{I_G} i \cdot n_{Gi}}$$

Außerdem kann man noch angeben: $\frac{\text{Anzahl fehlende Kanten}}{\mathrm{K_G}}$

Folgende Relationen gelten:

$S_4 < S_1 < S_2$ und $S_3 < S_2$.

Homonyme –Meldungen, die im EKR NRW zusammengeführt wurden, aber nicht im Gold-Standard

Wie viele Personen zu wenig wurden aufgrund von Homonymfehlern gezählt?

Es sei $h_{ij}\,(1 \le j \le i)$ die Anzahl der Meldungsgruppen der Größe *i* im EKR NRW, die aus Meldungen zu *j* verschiedenen Personen gebildet wurden. (Also: *i* Meldungen werden im EKR NRW zu einer Person zusammengeführt, sie gehören aber „tatsächlich" im Gold-Standard zu *j* Personen.) Es gilt $n_{Mi} = \sum_{j=1}^{i} h_{ij}$, d. h. die Anzahl der Personen mit *i* Meldungen setzt sich zusammen aus h_{i1} Gruppen, die zu Recht zu einer Person zusammengeführt wurden, aus h_{i2} Gruppen, die aus 2 Personen bestehen, aus h_{i3} Gruppen, die aus 3 Personen bestehen usw.

Die Überlegungen sind wie oben, nur dass Gold-Standard und EKR NRW ihre Rollen tauschen. Wenn Homonymfehler auftreten, hat der entstehende Graph zu viele Kanten. Für eine detaillierte Darstellung siehe Tabelle 18 im Anhang.

Die Zahl der Homonyme insgesamt ergibt sich als $N_H = \sum_{i=2}^{I_M} \sum_{j=2}^{i} (j-1) h_{ij}$

Zur Quantifizierung der Homonymfehlerrate kommen ebenfalls mehrere Ausdrücke in Frage:

1. Zahl der Homonyme bezogen auf alle Fälle. Diese Zahl gibt an, um welchen Anteil die Fallzahl zu niedrig geschätzt wird dadurch, dass Homonymfehler auftreten. Diese Zahl gibt die Auswirkung der Homonymfehler auf das Register an. Sie sagt zugleich auch etwas über die Güte des Record Linkage-Verfahrens aus.

$$H_1 = \frac{N_H}{N_G} = \frac{\sum_{i=2}^{I_M} \sum_{j=2}^{i} (j-1) h_{ij}}{\sum_{i=1}^{I_G} n_{Gi}}$$

2. Anteil der Meldungsgruppen mit mindestens zwei Meldungen, bei denen tatsächlich Homonyme auftreten – egal ob eines oder mehrere:

$$H_3 = \frac{\sum_{i=2}^{I_M} \sum_{j=2}^{i} h_{ij}}{\sum_{i=2}^{I_M} n_{Mi}} = \frac{\sum_{i=2}^{I_M} (n_{Mi} - h_{i1})}{\sum_{i=2}^{I_M} n_{Mi}}$$

3. Anteil der Meldungen, die Homonyme sind:

$$H_4 = \frac{N_H}{N} = \frac{\sum_{i=2}^{I_M} \sum_{j=2}^{i} (j-1) h_{ij}}{\sum_{i=1}^{I_G} i \cdot n_{Gi}} = \frac{\sum_{i=2}^{I_M} \sum_{j=2}^{i} (j-1) h_{ij}}{\sum_{i=1}^{I_M} i \cdot n_{Mi}}$$

Außerdem kann man noch angeben, um wie viel die Zahl der Kanten überschätzt wurde:

$$\frac{\text{Anzahl überzählige Kanten}}{K_G}$$

und den Anteil der fälschlich identifizierten Kanten an allen nicht zu identifizierenden Kanten

$$\frac{\text{Anzahl überzählige Kanten}}{K - K_G}$$

Folgende Relation gilt:

$H_4 < H_1$

Zu S_2 gibt es kein sinnvolles Analogon bei der Bewertung von Homonymfehlern.

Konsequenzen von Homonym– und Synonymfehlern

Verzerrung der Fallzahlschätzung durch Homonyme und Synonyme

Homonymfehler und Synonymfehler kompensieren sich im Hinblick auf die Fallzahlschätzung teilweise, je nachdem, ob Homonyme oder Synonyme überwiegen, wird die Fallzahl zu hoch oder zu niedrig angegeben. Der absolute Nettofehler ergibt sich als $\text{Fehler}_{\text{abs}} = N_S - N_H$, bzw. der relative Nettofehler bezogen auf die tatsächliche Fallzahl N_G als $\text{Fehler}_{\text{rel}} = \frac{N_S - N_H}{N_G}$. Ist der Fehler positiv, wird die Fallzahl überschätzt, ist er negativ, wird die Fallzahl unterschätzt [12].

Verzerrung der Überlebenszeitschätzung durch Homonyme und Synonyme

Beim Abgleich von Registerdaten mit Mortalitätsdaten führen Synonymfehler zu einer Überschätzung der Überlebenszeit, da Sterbeinformationen nicht den im Krebsregister registrierten Patienten zugeordnet werden [13]. Sofern die Sterbeinformation auch noch eine Tumorinformation enthält, führt dies außerdem zu zusätzlichen DCO-Fällen[5], wobei die Gefahr besteht, dass typische Metastasensitze wie Lunge, Leber oder Gehirn als Primärtumoren erfasst werden und damit die Fallzahl speziell für diese Entitäten überschätzt wird.

Beim Abgleich von Registerdaten mit Mortalitätsdaten führen Homonymfehler zu einer Unterschätzung der Überlebenszeit, da Sterbeinformationen vorzeitig zugeordnet werden. Sofern die Sterbeinformation auch noch eine Tumorinformation enthält, ergeben sich zu wenige DCO-Fälle. Dieser Fehler kann u. U. korrigiert werden, wenn das Register einige Zeit später die richtige Sterbeinformation erhält. Sobald aber als verstorben gekennzeichnete Patienten

[5] DCO = "death certificate only", DCO-Fälle = Krebsfälle, die einem Krebsregister nur über Todesbescheinigung bekannt sind.

nicht mehr in das Record Linkage einbezogen werden, ist diese Korrektur nicht mehr möglich.

Hochrechnung der Homonymfehlerraten auf größere Datenbestände

Untersuchungen an realen Datenbeständen haben gezeigt [14], dass die Homonymfehlerrate mit der Größe des Datenbestands zunimmt. Dies liegt daran, dass – bei gleich bleibender Wahrscheinlichkeit, dass zwei beliebige nicht zusammengehörende Meldungen zufällig Übereinstimmung aufweisen – die Fehlermöglichkeiten quadratisch mit der Größe des Datenbestands anwachsen.

Es sei h die Wahrscheinlichkeit, ein zufällig ausgewähltes Paar von Meldungen fälschlich zuzuordnen. In einem Datenbestand mit N Datensätzen gibt es $\frac{N(N-1)}{2}$ mögliche Kanten. Eine obere Abschätzung für die Zahl fälschlich getroffener Zuordnungen ist gegeben durch $h \cdot \frac{N(N-1)}{2}$ und dementsprechend für die Homonymfehlerraten durch $\frac{h(N-1)}{2}$.

Hochrechnung der Synonymfehlerraten auf Datenbestände mit anderer Verteilung der Mehrfachmeldungen

Es sei s die Wahrscheinlichkeit, dass ein Fehler in den Daten auftritt, der das Zusammenführen zweier Meldungen zur selben Person verhindert. Unter der Annahme, dass Fehler bei verschiedenen Meldungen unabhängig voneinander auftreten und dass die Wahrscheinlichkeit, dass zweimal derselbe Fehler auftritt, vernachlässigbar klein ist, ergibt sich für die Wahrscheinlichkeit $p_i(k)$, dass bei i Meldungen zu einer Person k verschiedene Meldungen, d. h. k-1 Synonyme auftreten $p_i(k) = \begin{cases} \binom{i}{k} s^{k-1}(1-s)^{i-k+1} & \text{falls } 1 \le k < i \\ i s^{i-1}(1-s) + s^i, & \text{falls } k = i \end{cases}$ [14].

Weiter sei n_i die Anzahl der Personen im Datenbestand, zu denen genau i Meldungen vorliegen. Dann ist

$N_G = \sum_{i \ge 1} n_i$ die Anzahl der Personen im Datenbestand,

$N = \sum_{i \ge 1} i \cdot n_i$ die Anzahl der Meldungen im Datenbestand,

$q_i = \frac{n_i}{\sum_{i \ge 1} n_i} = \frac{n_i}{N_G}$ der Anteil von Personen mit genau i Meldungen

Für die erwartete Anzahl Synonyme S gilt: $\mathrm{E}(N_S) = sN - \sum_{i \ge 1} n_i s^i$ [14]. Daraus ergibt sich

$$\begin{aligned}\mathrm{E}(N_S) &= s\sum_{i\geq 1} i\cdot n_i - \sum_{i\geq 1} n_i s^i \\ &= s\sum_{i\geq 1} n_i\left(i - s^{i-1}\right) \\ &= sN_G\sum_{i\geq 1} q_i\left(i - s^{i-1}\right)\end{aligned} \quad .$$

Damit lassen sich die oben definierten Synonymraten S_1 und S_2 folgendermaßen ausdrücken:

$$S_1 = \frac{N_S}{N_G} = \frac{sN - \sum_{i\geq 1} n_i s^i}{N_G} = \frac{sN_G\sum_{i\geq 1} q_i\left(i - s^{i-1}\right)}{N_G} = s\sum_{i\geq 1} q_i\left(i - s^{i-1}\right) = s\sum_{i\geq 2} q_i\left(i - s^{i-1}\right),$$

denn für i=1 ist auch s^{i-1}=1

$$S_2 = \frac{N_S}{\sum_{i\geq 2} n_i} = \frac{sN_G\sum_{i\geq 1} q_i\left(i - s^{i-1}\right)}{N_G\sum_{i\geq 2} q_i} = \frac{s\sum_{i\geq 2} q_i\left(i - s^{i-1}\right)}{\sum_{i\geq 2} q_i} = \frac{s\sum_{i\geq 2} q_i\left(i - s^{i-1}\right)}{1 - q_1}.$$

Aus den beobachteten Synonymen kann man auf s schließen, wenn man für N_s die beobachtete Anzahl Synonyme einsetzt und die Gleichung nach s auflöst.

Wenn die Zahl der Meldungen an das Krebsregister je Fall einer Poissonverteilung mit Parameter λ folgt, ergibt sich – da Fälle mit 0 Meldungen nicht registriert werden – dass die Häufigkeitsverteilung der *registrierten* Meldungen pro Fall eine gestutzte Poissonverteilung mit $q_i = \frac{1}{1-e^{-\lambda}}\frac{e^{-\lambda}\lambda^i}{i!}$ und Erwartungswert $\mu = \frac{\lambda}{1-e^{-\lambda}}$ ist.

Bei λ=2,3 ist die Wahrscheinlichkeit für Nicht-Registrierung 10%, was einer Vollzähligkeit von 90% entspricht, bei einer Vollzähligkeit von 95% wäre λ=3.

Damit ergibt sich für die erwartete Anzahl Synonyme

$$\begin{aligned}\mathrm{E}(N_S) &= \frac{sN_G}{1-e^{-\lambda}}\sum_{i\geq 1}\frac{e^{-\lambda}\lambda^i}{i!}\left(i - s^{i-1}\right) = \frac{sN_G}{1-e^{-\lambda}}\left(\sum_{i\geq 1}\frac{e^{-\lambda}\lambda^i}{i!}i - \sum_{i\geq 1}\frac{e^{-\lambda}\lambda^i}{i!}s^{i-1}\right) \\ &= \frac{sN_G}{1-e^{-\lambda}}\left(\sum_{i\geq 0}\frac{e^{-\lambda}\lambda^i}{i!}i - \frac{e^{-\lambda}}{s}\left(\sum_{i\geq 0}\frac{(\lambda s)^i}{i!} - 1\right)\right) = \frac{sN_G}{1-e^{-\lambda}}\left(\lambda - \frac{e^{-\lambda}}{s}\left(e^{\lambda s} - 1\right)\right) \\ &= \frac{N_G\left(\lambda s + e^{-\lambda}\left(1 - e^{\lambda s}\right)\right)}{1-e^{-\lambda}}\end{aligned}$$

Für die Synonymraten S_1 und S_2 erhält man dann

$$S_1 = \frac{\lambda s + e^{-\lambda}\left(1 - e^{\lambda s}\right)}{1-e^{-\lambda}} \text{ und}$$

$$S_2 = \frac{s\sum_{i\geq 2} q_i\left(i - s^{i-1}\right)}{1 - q_1} = \frac{\frac{s}{1-e^{-\lambda}}\sum_{i\geq 1}\frac{e^{-\lambda}\lambda^i}{i!}\left(i - s^{i-1}\right)}{1 - \frac{\lambda e^{-\lambda}}{1-e^{-\lambda}}} = \frac{\lambda s + e^{-\lambda}\left(1 - e^{\lambda s}\right)}{1 - e^{-\lambda} - \lambda e^{-\lambda}} = \frac{\lambda s + e^{-\lambda}\left(1 - e^{\lambda s}\right)}{1 - e^{-\lambda}\left(1 + \lambda\right)}$$

Aus der Wahrscheinlichkeit *s*, einen Synonymfehler zu begehen und der mittleren Anzahl Meldungen je Fall, λ, ergibt sich damit die Synonymfehlerrate bezogen auf den gesamten Datenbestand (S_1) bzw. auf die Fälle mit Mehrfachmeldungen (S_2).

Wir haben aus der beobachteten Häufigkeitsverteilung Meldungsverteilungen mit einem größeren Anteil an Mehrfachmeldungen abgeleitet und außerdem einige gestutzte Poissonverteilungen untersucht. Für die Fehlerwahrscheinlichkeit *s* haben wir zum einen die beobachtete Fehlerwahrscheinlichkeit zugrunde gelegt und zum anderen angenommen, dass sie halbiert werden kann.

3.2 Überprüfung der Verfahrensweisen

Zur Überprüfung der Verfahrensweisen wurden am 16.12.08 Dokumente zu folgenden Themenbereichen aus dem Krebsregister Münster angefordert.

- Eingesetzte Chiffrierverfahren Datenflüsse und Datenspeicherung
- Record Linkage, speziell zu den Entscheidungsstrategien bei der Zuordnung mehrerer Meldungen zu einem Tumor bzw. zu einem Patienten
- Qualitätssicherung und Rückfragen bei Meldern
- Datenschutz und Datensicherheit

Diese Dokumente wurden am 16.12.2008 und 07.01.2009 an das IMBEI übermittelt. Sie wurden von Mitarbeiterinnen und Mitarbeitern des Krebsregisters Rheinland-Pfalz und des IMBEI durchgearbeitet, dabei ergaben sich eine Reihe von Fragen, die beim Vor-Ort-Besuch am 26./27.02.2009 angesprochen und beantwortet wurden.

3.2.1 Vor-Ort-Besuch

Am 26./27.02.2009 besuchten Herr Seebauer (Informatiker in der Vertrauensstelle des Krebsregisters Rheinland-Pfalz), Frau Sudhof (Medizinische Dokumentarin in der Registerstelle des Krebsregisters Rheinland-Pfalz) und Frau Schmidtmann das Krebsregister NRW. Dr. Krieg und die übrigen Mitarbeiter und Mitarbeiterinnen des Krebsregisters NRW erläuterten sehr detailliert die Datenerhebung und -verarbeitung im Krebsregister NRW.

Die folgenden Themenbereiche wurden beim Vor-Ort-Besuch angesprochen: Datenflüsse und Datenspeicherung, eingesetzte Chiffrierverfahren, Datenschutz und Datensicherheit, Record Linkage, SOPs, Qualitätssicherung und Rückfragen bei Meldern.

4 Ergebnisse

4.1 Record Linkage

Für das Clerical Review ergaben sich 15.146 Gruppen mit 2 bis 26 Datensätzen je Gruppe (siehe Tabelle 6).

Tabelle 6: Anzahl und Größe der gefundenen Matchgruppen

Größe	Anzahl Gruppen
2	11.955
3	2.377
4	608
5	144
6	39
7	11
8	4
9	3
10	2
11	1
13	1
26	1
Summe	15.146

Unter den 4.000 Paaren von Datensätzen mit Übereinstimmungsgewicht zwischen 0 und 17 wurde keines gefunden, das hätte zusammengeführt werden müssen.

Nach Erstellung des Gold-Standards ergab sich, dass sich die 150.000 Meldungen auf 132.267 Personen beziehen. Für 118.218 Personen lag genau ein Datensatz vor, für 14.049 Personen lag mehr als ein Datensatz vor. Dagegen hatte die Zuordnung im Krebsregister NRW 132.515 Personen, davon 13.841 mit Mehrfachmeldungen ergeben. Daraus ergibt sich ein Nettofehler von 248, das entspricht 0,19%, die Fallzahl wurde überschätzt.

Die Zuordnung der Mehrfachmeldungen durch das Krebsregister NRW ist in Tabelle 7 dargestellt.

Tabelle 7: Mehrfachmeldungen im Krebsregister NRW in der untersuchten Stichprobe

Anzahl zu einer Person zusammengeführter Meldungen (n)	**Entsprechende Anzahl Kanten[6] $m = n(n-1)/2$**	**Personen mit dieser Anzahl Meldungen (p)**	**Meldungen $= n \times p$**	**Kanten $= m \times p$**
1	0	118.674	118.674	0
2	1	11.073	22.146	11.073
3	3	2.098	6.294	6.294
4	6	518	2.072	3.108
5	10	115	575	1.150
6	15	24	144	360
7	21	10	70	210
8	28	2	16	56
9	36	1	9	36
Summe		132.515	150.000	
davon (mit) Mehrfachmeldungen		13.841	31.326	22.287

Die Verteilung der Mehrfachmeldungen im Gold-Standard ist aus Tabelle 8 zu entnehmen.

Tabelle 8: Mehrfachmeldungen im Gold-Standard in der untersuchten Stichprobe

Anzahl zu einer Person zusammengeführter Meldungen (n)	**Entsprechende Anzahl Kanten[7] $m = n(n-1)/2$**	**Personen mit dieser Anzahl Meldungen (p)**	**Anteil Personen**	**Meldungen $= n \times p$**	**Kanten $= m \times p$**
1	0	118.218	89,38%	118.218	0
2	1	11.243	8,50%	22.486	11.243
3	3	2.131	1,61%	6.393	6.393
4	6	524	0,40%	2.096	3.144
5	10	115	0,09%	575	1.150
6	15	24	0,02%	144	360
7	21	9	0,01%	63	189
8	28	2	0,00%	16	56
9	36	1	0,00%	9	36
Summe		132.267		150.000	
davon (mit) Mehrfachmeldungen		14.049		31.782	22.571

[6] Anzahl paarweiser Vergleiche von Meldungen pro identifizierter Person

4.1.1 Homonym- und Synonymfehler

Die zwischen dem Krebsregister NRW und dem Gold-Standard abweichenden Entscheidungen sind in Tabellen 9 und 10 dargestellt. Es zeigte sich, dass aufgrund von Synonymfehlern 268 Fälle zu viel gezählt wurden, dies wird teilweise kompensiert dadurch, dass aufgrund von Homonymfehlern 20 Fälle zu wenig gezählt wurden. Wenn Synonymfehler auftraten, wurde meistens eine Zuordnung übersehen. Jeweils einmal kam es dazu, dass drei Meldungen einer Person im Krebsregister als drei verschiedene Personen (1+1+1 statt 3) und dass vier Meldungen als je zwei Meldungen zu zwei Personen aufgefasst wurden (2+2 statt 4).

Synonymfehler

Tabelle 9: Abweichungen der Zuordnung von Mehrfachmeldungen zwischen Krebsregister und Gold-Standard in der untersuchten Stichprobe – Synonymfehler

Meldungen je Person im Gold-Standard (*k*)	**Zuordnung Meldungen zu Personen Krebsregister NRW**	**Personen (*p*)**	**Meldungen (*k*×*p*)**	**Anzahl Synonymfehler[7]**	**fehlende Kanten[8]**
2	1+1	223	446	223	223
3	2+1	37	111	37	74
3	1+1+1	1	3	2	3
4	3+1	5	20	5	15
4	2+2	1	4	1	4
Summe				268	319

Damit ergibt sich für die Synonymfehlerraten:

$S_1 = \frac{N_S}{N_G} = \frac{268}{132267} = 0,0020 = 0,2\%$, d. h. es aufgrund von Synonymfehlern werden 0,2% zu viele Fälle gezählt.

$S_2 = \frac{N_S}{\sum_{i=2}^{I_G} n_{Gi}} = \frac{268}{14049} = 0,0191 = 1,91\%$, d. h. bezogen auf die mehrfach gemeldeten Fälle ergeben sich 1,91% Synonyme.

$S_3 = \frac{267}{14049} = 0,0190 = 1,90\%$, d. h. bei 1,90% der Fälle, bei denen Synonymfehler überhaupt möglich sind, kamen tatsächlich Synonymfehler vor, dabei gab es nur bei einer Person mit drei und mehr Meldungen zwei zu viel gezählte Fälle.

$S_4 = \frac{N_S}{N} = \frac{268}{150000} = 0,0018 = 0,18\%$, d. h. 0,18% Synonyme unter 150.000 Meldungen.

Von 22.571 tatsächlich vorhandenen Kanten wurden 319 nicht erkannt, das entspricht 1,41%.

[7] Wenn nur die Meldungen zu einer Person auf zwei Personen zugeordnet werden, entsteht ein Synonymfehler. Werden die Meldungen zu einer Person *k* Personen zugeordnet, so entstehen *k*-1 Synonymfehler.

[8] Die Zahl der fehlenden Kanten ergibt sich aus der tatsächlichen Anzahl Meldungen zu einer Person und der jeweiligen Zuordnung zu mehreren Personen (siehe auch Abbildung 1).

Homonymfehler

Tabelle 10: Abweichungen der Zuordnung von Mehrfachmeldungen zwischen Krebsregister und Gold-Standard in der untersuchten Stichprobe - Homonymfehler

Meldungen je Person im Gold-Standard	Zuordnung Meldungen je Person EKR NRW (*k'*)	Meldungsgruppen Krebsregister NRW (*p'*)	Meldungen (*k'*×*p'*)	Anzahl Homonymfehler[9]	überzählige Kanten[10]
1+1	2	16	32	16	16
2+1	3	1	3	1	2
3+1	4	1	4	1	3
4+1	5	1	5	1	4
5+2	7	1	7	1	10
Summe				20	35

Damit ergibt sich für die Homonymfehlerraten:

$H_1 = \frac{N_H}{N_G} = \frac{20}{132267} = 0,00015 = 0,015\%$, d. h. aufgrund von Homonymfehlern werden 0,015% zu wenige Fälle gezählt.

$H_3 = \frac{\sum_{i=2}^{I_M} \sum_{j=2}^{i} h_{ij}}{\sum_{i=2}^{I_M} n_{Mi}} = \frac{20}{13841} = 0,00144 = 1,44\%$, d. h. unter den 13.841 Meldungsgruppen, die nach Zuordnung im Krebsregister NRW mehr als eine Meldung enthalten, sind 1,44% mit Homonymen.

$H_4 = \frac{N_H}{N} = \frac{20}{150000} = 0,00013 = 0,013\%$, d. h. unter den 150.000 untersuchten Meldungen sind 0,013% Homonyme.

Schließlich wurden 35 Kanten mehr als die tatsächlich vorhandenen 22.571 gezählt, das entspricht 0,16%. Von $\frac{150.000 \cdot 149.999}{2} - 22571 = 11.249.902.429 = 1,125 \cdot 10^{10}$ nicht auszuwählenden Kanten wurden 35 ausgewählt. Daraus ergibt sich die Wahrscheinlichkeit, dass ein beliebiges Paar nicht zusammengehöriger Meldungen zu einer Person zusammengeführt wird mit $h = 3,111 \cdot 10^{-9}$

Der Nettofehler bei den Kanten beträgt 284, das entspricht 1,26%.

4.1.2 Record Linkage-Fehler nach Meldequellen

Im Gold-Standard gab es 11.243 Personen mit zwei Meldungen und 2.131 Personen mit drei Meldungen und zusammen 17.636 Kanten. Die Verteilung dieser Kanten auf die Kombinationen von Meldequellen ist in Tabelle 11 dargestellt. Dort ist auch angegeben, wie sich die

[9] Wenn nur die Meldungen zu zwei Personen einer Person zugeordnet werden, entsteht ein Homonymfehler. Werden die Meldungen von *k'* Personen einer Personen zugeordnet, so entstehen *k-1* Homonymfehler.

[10] Die Zahl der überzähligen Kanten ergibt sich aus der tatsächlichen Anzahl Personen, deren Meldungen einer Person zugeordnet werden und aus der Anzahl der zu diesen Personen vorliegenden Meldungen (siehe auch Abbildung 1).

Synonyme auf die Kombinationen von Meldequellen verteilen. Auffällig ist, dass bei der Zuordnung von Meldungen aus der onkologischen Qualitätssicherung zu einem Fall, besonders wenige Synonyme auftraten. Diese Kombination von Meldequellen macht 37,2% aller Kanten aus, aber nur 4,3% der übersehenen Kanten, alle anderen Meldequellen waren in größerem Umfang an den Synonymen beteiligt als an den Zuordnungen im Gold-Standard. D. h. die schon im Zwischenbericht beschriebene höhere Datenqualität der Meldungen aus der onkologischen Qualitätssicherung wirkt sich auch positiv auf die korrekten Zuordnungen aus.

Die Verteilung der Homonyme auf die Meldequellen wird hier nicht dargestellt, da hier nur 35 Kanten beteiligt sind, beschränkt auf die Meldungsgruppen mit zwei oder drei Meldungen, sogar nur 18 Meldungen. Zudem sind Homonymfehler kaum von der Datenqualität abhängig, sondern im Wesentlichen vom Umfang der Merkmalsausprägungen der zum Record Linkage verwendeten Merkmale.

Tabelle 11: Kombination der Meldequellen im Gold-Standard und bei den Synonymen, nur Personen mit zwei oder drei Meldungen

Meldequellen-Kombination		Gold-Standard				Synonymfehler			
		Anzahl Kanten	Anteil	95%-Konfidenzintervall für den Anteil		Anzahl Kanten	Anteil	95%-Konfidenzintervall für den Anteil	
IMD	IMD	1199	6,8%	6,4%	7,2%	20	6,7%	3,9%	9,5%
IMD	IMO	276	1,6%	1,4%	1,8%	17	5,7%	3,1%	8,3%
IMD	MA	115	0,7%	0,6%	0,8%	9	3,0%	1,1%	4,9%
IMD	PBF	947	5,4%	5,1%	5,7%	22	7,3%	4,4%	10,2%
IMO	IMO	6568	37,2%	36,5%	37,9%	13	4,3%	2,0%	6,6%
IMO	MA	652	3,7%	3,4%	4,0%	44	14,7%	10,7%	18,7%
IMO	PBF	3416	19,4%	18,8%	20,0%	61	20,3%	15,7%	24,9%
MA	MA	41	0,2%	0,1%	0,3%	0	0,0%		
MA	PBF	698	4,0%	3,7%	4,3%	87	29,0%	23,9%	34,1%
PBF	PBF	3724	21,1%	20,5%	21,7%	27	9,0%	5,8%	12,2%
IMD	beliebig	2537	14,4%	13,9%	14,9%	68	22,7%	17,9%	27,4%
IMO	beliebig	10912	61,9%	61,2%	62,6%	135	45,0%	39,4%	50,6%
MA	beliebig	1506	8,5%	8,1%	9,0%	140	46,7%	41,0%	52,3%
PBF	beliebig	8785	49,8%	49,1%	50,6%	197	65,7%	60,3%	71,0%
Summe		17636				300			

4.2 Evaluation der Abgleich-Strategie

4.2.1 Unterschiede zwischen den Stufen der Abgleich-Strategie

Bei der Überprüfung der Abgleich-Strategie stellte sich heraus, dass die Ergebnismengen (potenzielle oder sicher zusammengehörende Meldungen) der restriktiven ersten 5 Stufen – wie es vermutet werden konnte – in den Ergebnismengen der nachfolgenden 5 Stufen enthalten sind. Die Ergebnismengen der letzten 5 Stufen überschneiden und ergänzen sich, aber jede liefert einen Beitrag zur Gesamt-Ergebnismenge. Tabelle 12 fasst dieses Ergebnis knapp zusammen ohne die beschriebenen Befunde exakt wiederzugeben. Die genaue Auflistung der Trefferzahlen nach Art des Abgleichs findet sich im Anhang (siehe Tabelle 19). Inwiefern Stufe 10, die sehr viele Treffer generiert, für das Ergebnis des Abgleichs relevant ist, wurde nicht überprüft. Der – verglichen an der Gesamtzahl – geringe Zugewinn an potenziell zusammengehörenden Meldungen erlaubt die Vermutung, dass auf Stufe 10 hätte verzichtet werden können.

Tabelle 12: Exklusiv in der jeweiligen Stufe gefundene Treffer

	Stufe	Exklusiv in der jeweiligen Stufe gemachte Vergleiche	Exklusiv in der jeweiligen Stufe gefundene Treffer		
			Exakter Vergleich	Bigramme	Levenshtein-Distanz
Abgleich der Inzidenz- und Mortalitäts-Meldungen	01	0	0	0	0
	02	0	0	0	0
	03	0	0	0	0
	04	0	0	0	0
	05	0	0	0	0
	06	356.052	9	9	9
	07	969.340	0	0	0
	08	293.089	6	6	6
	09	245.069	4	4	4
	10	1.453.728	35	35	35
Deduplizieren der Inzidenz-Meldungen	01	0	0	0	0
	02	0	0	0	0
	03	0	0	0	0
	04	0	0	0	0
	05	0	0	0	0
	06	810.576	2	2	2
	07	2.212.379	0	0	0
	08	813.425	1	1	1
	09	303.882	17	17	17
	10	1.379.904	12	13	12

4.2.2 Informations-Zugewinn durch Einsatz von Stringmetriken

Gemessen an der Anzahl durchgeführter Vergleiche von Meldungen wurden durch Einsatz von Bigrammen insgesamt nur ein und durch Einsatz der Levenshtein-Distanz als Stringmetrik im Durchschnitt 0,0007 % zusätzliche *potenziell* zusammengehörende Meldungen identifiziert (Tabelle 13).

Tabelle 13: Exklusiv mit den unterschiedlichen Stringmetriken gefundene Treffer

			Exklusiv mit der jeweiligen Methode gefundene Treffer		
	Stufe	**Vergleiche insgesamt**	**Exakter Vergleich**	**Bigramme**	**Levenshtein-Distanz**
Abgleich der Inzidenz- und Mortalitäts-Meldungen	01	758	0	0	1
	02	1.701	0	0	2
	03	4.487	0	0	0
	04	15.320	0	0	0
	05	3488	0	0	3
	06	401.565	0	0	2
	07	1.031.949	0	0	2
	08	331.683	0	0	2
	09	249.864	0	0	5
	10	1.470.656	0	0	2
Deduplizieren der Inzidenz-Meldungen	01	21.084	0	0	2
	02	20.824	0	0	2
	03	23.416	0	0	1
	04	41.078	0	0	2
	05	24.424	0	0	5
	06	937.298	0	0	8
	07	2.387.985	0	0	8
	08	940.257	0	0	8
	09	331.868	0	0	9
	10	1.424.205	0	1	2

4.3 Überprüfung der Verfahrensweisen

4.3.1 Datenflüsse und Datenspeicherung

Anhand der Dokumente „Rechnernetze 200781030.doc“ und „datenbänke 20071031.doc“ wurden Datenhaltung und Datenflüsse erläutert. Die Datenhaltung ist durchdacht und erlaubt nachzuvollziehen, wie der jeweilige Datenbestand zustande gekommen ist. Im Falle von Fehlern lässt sich der vorherige Zustand wiederherstellen. Der Eingang und die Verarbeitung von Datensätzen wurden uns an Testfällen und realen Fällen gezeigt. Bei den realen Fällen waren naturgemäß keine Klartext-Identitätsdaten sichtbar.

4.3.2 Eingesetzte Chiffrierverfahren

Die Kontrollnummernerzeugung entspricht den von der GEKID übernommenen Empfehlungen [15] und dem Quasistandard der deutschen Krebsregister, UNICON. Es wird lediglich auf die inzwischen obsoleten Baden-Württemberg-Kontrollnummern sowie auf den DDR-Namenscode verzichtet.

Zusätzlich zu den empfohlenen Kontrollnummern werden auch aus Straße und Hausnummer Kontrollnummern gebildet. Dabei werden Straßennamen ähnlich wie Namen in bis zu fünf Teile zerlegt und standardisiert.

Identitätschiffrate werden mit asymmetrischen Chiffrierverfahren erzeugt. Sowohl die asymmetrische Chiffrierung als auch der erste Schritt der Kontrollnummernerzeugung, nämlich die Erstellung der Kryptogramme, findet bei den Meldern statt und ist – ohne dass diese sich darum kümmern müssten – in die Meldesoftware integriert. Nur die Identitätschiffrate werden direkt an das EKR NRW übermittelt. Die Kryptogramme gehen an die Kassenärztliche Vereinigung Westfalen-Lippe, wo ein Pseudonymisierungsserver – gewissermaßen als eine virtuelle Vertrauensstelle – die zweite Stufe der Kontrollnummernerzeugung übernimmt. Die Kryptogramme werden über eine https-Verbindung verschlüsselt an die Kassenärztlichen Vereinigung übertragen.

4.3.3 Datenschutz und Datensicherheit

Dem Datenschutz wird grundsätzlich Rechnung getragen durch das Konzept der pseudonymisierten Datenspeicherung. Darüber hinaus sind die verwendeten Server und Arbeitsplatzrechner nur den Mitarbeitern des Krebsregisters zugänglich, das Krebsregister-Rechnernetz ist von der Außenwelt abgeschottet. Die Mitarbeiter haben von ihren Arbeitsplätzen Zugriff auf das KV-SafeNet, über das auch die Meldungen übermittelt werden. Die Rechnernetze sind näher im Dokument „Rechnernetze 20071030.doc“ beschrieben. Der Zugang zu den Räumen des Krebsregisters ist ebenfalls gesichert, eine Zugangskontrolle findet statt. Weiterhin gibt es ein Rollenmodell, das den Mitarbeitern nur die für ihre Aufgaben notwendigen Rechte einräumt.

Die Datensicherheit wird durch regelmäßige nächtliche Bandsicherungen gewährleistet. Während der Sicherung kann nicht an der Datenbank gearbeitet werden, so dass die Integrität der gesicherten Datenbanken gewährleistet ist. Durch Vorhalten von Historien ist auch die Wiederherstellung früherer Zustände der Datenbank möglich.

4.3.4 Record Linkage

Die Mitarbeiter des EKR NRW gaben uns anhand einiger aktueller Fälle Einblick in die Durchführung des Record Linkage. Das EKR NRW erhält die neuen Meldungen in Teilen: die Identitätschiffrate, die bei der Kassenärztlichen Vereinigung Westfalen-Lippe gebildeten Kontrollnummern und den medizinischen Teil der Meldung. Jede Meldung ist durch eine Meldernummer, eine laufende Nummer und einen Zeitstempel eindeutig bestimmt.

Sollte einmal ein Meldungsbestandteil nicht korrekt übermittelt worden sein, so fällt dies beim Zusammensetzen der Meldungsteile auf. Das sendende Programm erhält Rückmeldung, wenn ein Meldungsteil angekommen ist. Bleibt die Rückmeldung aus, wird die komplette Meldung beim nächsten Sendeversuch erneut gesendet. Durch „Report Linkage“ werden die Teile zusammengeführt, im Krebsregister sind unvollständige Meldungsteile und neue Sendeversuche sichtbar. Wenn Meldungsteile länger unvollständig bleiben, wird Kontakt mit der Meldestelle aufgenommen. Wenn nötig, arbeitet auch ein technischer Mitarbeiter des Krebsregisters beim Melder an der Lösung des Problems.

Jeweils nachts werden die neuen Meldungen mit dem Registerbestand abgeglichen. Die Abgleichprotokolle werden archiviert. Außerdem wird regelmäßig der Datenbestand auf mögliche Duplikate oder fälschlich zusammengeführte Meldungen überprüft.

Die Record Linkage Software ist eine Eigenentwicklung des EKR NRW, die das Verfahren von Fellegi und Sunter [8] umsetzt. Die Matchvariablen und Blockvariablen sind fest, es werden sieben Läufe mit unterschiedlichen Blockvariablen durchgeführt. Meldungspaare mit Scores oberhalb einer oberen Schranke werden automatisch zusammengeführt, Meldungspaare mit Scores unterhalb einer unteren Schranke bleiben automatisch getrennt. Die Meldungspaare mit Scores im Graubereich zwischen den beiden Schranken werden nachbearbeitet. Dabei kommen einige Regeln zum Einsatz, die eine automatische Zuordnung bzw. Nichtzuordnung in weiteren Fällen erlauben. Diese Regeln sind im Dokument "Interne Arbeitsabläufe – krnrw-intern.doc“ in Abschnitt 15.3 ausgeführt. Bei der Durchsicht erschienen uns die Regeln 3 und 4 problematisch. Beim Vor-Ort-Besuch erfuhren wir, dass diese Regeln nicht mehr angewandt werden, da auch die Mitarbeiter im EKR NRW zu dieser Einschätzung gelangt waren.

Eine Reduktion der abzugleichenden Meldungen wird auch dadurch erreicht, dass Meldungen von Personen, die seit mehr als fünf Jahren verstorben sind, in eine separate Datenbank („Closed“) übertragen werden. Für die manuelle Nachbearbeitung des Record Linkage stehen alle Informationen aus den Meldungen zur Verfügung, auch die medizinischen. Die Übereinstimmungsscores sind ebenfalls sichtbar. Für die manuelle Nachbearbeitung gibt es noch keine schriftliche Fixierung des Vorgehens. Regelmäßige Schulungen gewährleisten jedoch eine einheitliche Bearbeitung auch durch verschiedene Personen.

4.3.5 SOPs

Die Arbeitsabläufe im Krebsregister NRW sind in dem umfangreichen Dokument „Interne Arbeitsabläufe“ nachvollziehbar beschrieben. Weitere Dokumente beschreiben den Ablauf der Diagnosecodierung („ablauf dc.doc“), die „Best-Of-Merkmale-Generierung“ („best-of-merkmlae 20080617.doc“). In den Codierhilfen werden Detailfragen der Diaganosecodierung beantwortet.

4.3.6 Qualitätssicherung und Rückfragen bei Meldern

Diagnosecodierung

Für die Diagnosecodierung gibt es für jeden Meldungstyp Ablaufdiagramme. Bei direkten Inzidenzmeldungen werden Diagnosen doppelt codiert. Diskrepanzen werden überprüft, die entsprechenden Meldungen werden ggf. zurückgestellt und in einer gemeinsamen Besprechung geklärt. Falls in dieser Besprechung noch keine Klärung möglich ist, sind weitere Schritte vorgesehen, wie Rückfrage beim Melder oder Hinzuziehen eines ärztlichen Mitarbeiters. In besonders schwierigen Fällen kann eine Pathologin konsultiert werden.

Bei Meldungen über die onkologische Qualitätssicherung wird die im Krebsregister vergebene Codierung mit der des Krankenhauses verglichen. Pathologiemeldungen werden – aufgrund des Zeitaufwandes – ebenfalls nur einmal im Krebsregister NRW codiert.

Plausibilitätsprüfungen und Best-of-Regeln

Umfangreiche Plausibilitätsprüfungen gewährleisten, dass Diagnosen möglichst fehlerfrei codiert werden. Plausibilitätsprüfungen sind in Form von Tabellen hinterlegt. Auch für die Meldestellenprogramme sind Plausibilitätsprüfungen hinterlegt, etwa zulässige Postleitzahlen, Ortsnamen, Gemeindekennziffern, ICD und ICD-O. Beim Export aus bestehenden Datenbanken von Pathologen, klinischen Registern und Brustzentren kommen bislang keine Plausibilitätsprüfungen zum Einsatz. Für die Datenübernahme aus klinischen Registern wird aber derzeit ein elektronisches System für Rückfragen etabliert.

Die beste Information für einen Tumor („Best of Tumor“) aus allen Meldungen zu einem Tumor kann in etwa 80% der Fälle automatisch extrahiert werden. Ob eine Entscheidung automatisch getroffen werden kann, ist in Entscheidungsmatrizen festgelegt. Die Regeln sind teilweise in Datenbanken abgelegt, teilweise ausprogrammiert, ferner werden Entscheidungsmatrizen verwendet. Die hier verwendeten Regeln beruhen auf nationalen und internationalen Standards. Die Regeln und ihre Quellen sind in den „Internen Arbeitsabläufen“ dargestellt.

Nach Durchsicht der Unterlagen wurden einige wenige Ergänzungen zu den Listen benachbarter Topographien in Kapitel 18.5.1 der „Internen Arbeitsabläufe“ zusammengestellt und den Mitarbeitern des EKR NRW beim Vor-Ort-Besuch zur Verwendung übergeben.

Auch für die Ergebnisse der Best-of-Algorithmen greifen Plausibilitäten. Widersprüche werden manuell aufgelöst.

Löschung von Meldungen

Meldungen können gelöscht werden. Allerdings haben nur wenige Mitarbeiter das Recht, Meldungen zu löschen, sie müssen zwei Abfragen bejahen und einen Löschgrund angeben. Die übrigen Mitarbeiter müssen Meldungen zum Löschen zurückstellen, erst nach Freigabe durch einen zum Löschen berechtigten Mitarbeiter wird die Löschung vollzogen. Gelöschte Meldungen werden mit Löschgrund in einer Archivdatenbank gespeichert. Die komplette Entfernung von Meldungen inklusive aller Zwischensicherungen ist nicht vorgesehen, da es in Nordrhein-Westfalen kein Widerspruchsrecht der Patienten gibt. Wenn eine Meldung zu Unrecht erfolgt, müsste der meldende Arzt die Löschung verlangen, dieser Fall ist aber bisher nicht aufgetreten. Falls ein Pathologe aufgrund einer verdächtigen Probeexzision meldet und sich dann doch herausstellt, dass kein Tumor vorliegt, kann er die Löschung des Datensatzes verlangen, das EKR löscht dann diesen Datensatz aus der Datenbank – eine Archivkopie bleibt aber erhalten.

Fehlerkorrekturen und Änderungen

Fehlerkorrekturen sind während der manuellen Nachbearbeitung möglich. Es ist festgelegt, welche Merkmale geändert werden dürfen und wer welche Merkmale ändern darf. Änderungen werden protokolliert, so dass nachvollziehbar ist, wer was wann geändert hat.

Nachfragen bei Meldern

Bei unplausiblen Meldungen oder Widersprüchen zwischen mehreren Meldungen zu einem Tumor kann bei den Meldern nachgefragt werden. Dies geschieht schriftlich in Form von (Kurz-)Briefen. Verschiedene Mitarbeiter sind für verschiedene Meldergruppen zuständig. Nachgefragt wird hauptsächlich bei den Onkologischen Schwerpunkten und bei niedergelassenen Ärzten. Die meisten Nachfragen bei den Onkologischen Schwerpunkten beziehen sich

auf medizinische Fragen. Außerdem werden Unklarheiten bei Geschlecht und Geburtsmonat nachgefragt. Aus Zeitgründen kann aber nicht bei jeder Meldung mit Problemen nachgefragt werden. Rückfragen werden mit dem Datum der Anfrage und dem Datum der Antwort protokolliert.

5 Diskussion und Schlussfolgerungen

5.1 Record Linkage

5.1.1 Bewertung der Record Linkage-Ergebnisse

Homonymfehler

Aus den oben dargestellten Fehlerraten ergibt sich, dass die Wahrscheinlichkeit, dass ein beliebiges Paar nicht zusammengehöriger Meldungen zu einer Person zusammengeführt wird, durchschnittlich etwa $3{,}111 \cdot 10^{-9}$ beträgt, bei einer Kombination häufiger Merkmalsausprägungen, etwa häufiger Nachname, häufiger Vorname und Großstadt, ist diese Wahrscheinlichkeit größer, bei seltenen Merkmalsausprägungen kleiner.

Unter der Annahme, dass $h = 3{,}111 \cdot 10^{-9}$, ergeben sich für die Homonymfehlerrate in Abhängigkeit vom Umfang des Datenbestands die in Tabelle 14 aufgeführten Abschätzungen. D. h. erst bei Record Linkage mit Datenbeständen von über 5 Millionen Meldungen ist mit einer Homonymfehlerrate von mehr als 1% zu rechnen.

Tabelle 14: Hochrechnung Homonymfehlerrate in Abhängigkeit vom Umfang des Datenbestands

Umfang Datenbestand	obere Abschätzung für die Homonymfehlerrate
100.000	0,02%
150.000	0,02%
200.000	0,03%
500.000	0,08%
1.000.000	0,16%
2.000.000	0,31%
5.000.000	0,78%
10.000.000	1,56%

Synonymfehler

Die Wahrscheinlichkeit, dass ein Paar von Meldungen, das zu einer Person gehört, fälschlicherweise nicht zusammengeführt wird, beträgt s=0,008458. Dies wurde wie in Abschnitt 3.1.8 beschrieben errechnet. In den Tabellen 15 und 16 ist dargestellt, wie sich bei unterschiedlichen Verteilungen der Anzahl Fälle pro Person die Synonymraten bezogen auf die Zahl der registrierten Personen (S_1) bzw. bezogen auf die Zahl der Fälle mit mehr als einer Meldung (S_2) verhalten. Neben der hier beobachteten Fehlerwahrscheinlichkeit s=0,008458 wird auch dargestellt, wie sich die Synonymfehlerraten entwickeln, wenn die Fehlerwahrscheinlichkeit halbiert werden kann.

Nimmt man an, dass sich die Anteile der Fälle mit mehr als einer Meldungen jeweils verdoppeln bzw. verdreifachen, so erhält man – in Abhängigkeit der Häufigkeitsverteilung der Meldungen je Person die in Tabelle 15 angegeben Synonymfehlerraten.

Tabelle 15: Hochrechnung Synonymfehlerrate in Abhängigkeit von der Häufigkeitsverteilung der Meldungen je Person

s	q_1	q_2	q_3	q_4	q_5	S_1	S_2
0,008,458	89,38%	8,50%	1,61%	0,40%	0,09%	0,20%	1,91%
0,008458	78,76%	17,00%	3,22%	0,79%	0,17%	0,41%	1,91%
0,008458	68,13%	25,50%	4,83%	1,19%	0,26%	0,61%	1,91%
0,004229	89,38%	8,50%	1,61%	0,40%	0,09%	0,10%	0,96%
0,004229	78,76%	17,00%	3,22%	0,79%	0,17%	0,20%	0,96%
0,004229	68,13%	25,50%	4,83%	1,19%	0,26%	0,30%	0,96%

Nimmt man an, dass die Zahl der Meldungen je Fall einer Poissonverteilung mit Erwartungswert λ folgt, so beobachtet man im Krebsregister als Häufigkeitsverteilung eine gestutzte Poissonverteilung, da Fälle mit 0 Meldungen nicht registriert werden. Wenn man nun annimmt, dass die Häufigkeitsverteilung der Meldungen je Person einer gestutzten Poissonverteilung mit Parameter λ folgt, so erhält man die in Tabelle 16 angegebenen Synonymfehlerraten.

Tabelle 16: Hochrechnung Synonymfehlerrate in Abhängigkeit von der Häufigkeitsverteilung der Meldungen je Person (gestutzte Poissonverteilung mit Parameter λ und Mittelwert $\mu = \frac{\lambda}{1-e^{-\lambda}}$)

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%

Es zeigt sich, dass mit Zunehmen von Mehrfachmeldungen auch die Synonymfehlerrate zunimmt. Bei im Mittel drei Meldungen pro Person erreicht die Synonymfehlerrate S_1, die den Grad der Überschätzung von Fällen misst, bis zu 2,36%. Dies wäre aber für ein Krebsregister bereits eine sehr hohe Quote von Mehrfachmeldungen. Gelingt es, die Fehlerwahrscheinlichkeit zu halbieren, ist bei einer mittleren Meldungszahl von 2,5 je Fall eine Synonymfehlerrate von knapp 1% zu erreichen.

5.1.2 Vorschläge zur Optimierung des Verfahrens

Datengewinnung

Es hat sich gezeigt, dass bei hoher Datenqualität weniger Synonyme auftreten, daher lohnt es sich, Maßnahmen zu treffen, die eine hohe Datenqualität sicherstellen. Es ist zu prüfen, ob es bei den direkten Inzidenzmeldungen und den Pathologiebefunden noch Möglichkeiten der Verbesserung bei der Erfassung der Identitätsdaten gibt. Weiter ist sehr zu wünschen, dass die im Rahmen der onkologischen Qualitätssicherung aufgebaute Dokumentation von Tumorerkrankungen auf qualitativ hohem Niveau fortgeführt wird, auch wenn die Onkologischen Schwerpunkte geschlossen werden. Die Qualität der Daten von den Einwohnermeldeämtern ist wahrscheinlich durch das Krebsregister nicht zu beeinflussen.

Generell sollten angesichts der Dateneingabe durch die Melder möglichst viele Plausibilitätsprüfungen und Formatprüfungen möglichst früh im Verarbeitungsprozess stattfinden. Dabei muss allerdings abgewogen werden zwischen den erreichbaren Qualitätssteigerungen und eventueller Demotivation von Meldern, wenn sie Meldungen nur mit mehrfachen Korrekturen abgeben können.

Verfahrensmodifikationen beim Record Linkage

Überkreuzvergleiche im Hinblick auf Vor- und Nachnamen sollten auch bei Anwendung des Fellegi-Sunter-Modells in Betracht gezogen werden, vor allem wenn ausländische Namen häufig sind. So wurden in unserer Evaluation hierdurch zusätzlich 124 Synonyme gefunden. Möglicherweise kann also durch Hinzunehmen von Überkreuzvergleichen die Synonymfehlerrate gesenkt werden.

Die Schranken für das Fellegi-Sunter-Modell können mittels strukturgleicher Trainingsdaten verbessert werden. Dabei sind Nutz- und Kostenpotentiale gegeneinander abzuwägen, wenn die manuell bestimmten Schranken ohnehin gute Ergebnisse zeigen.

Der Zugewinn durch den Einsatz von Stringmetriken ist für das Record-Linkage im Krebsregister zu vernachlässigen (siehe 4.2.2), die Verwendung von Kontrollnummern statt Klartextangaben liefert – eine gute Datenaufbereitung vorausgesetzt – ein weitgehend gleiches Ergebnis. Der Einsatz von Kontrollnummern statt Klartext führt also nicht zu einem schlechteren Record Linkage-Ergebnis.

Wenn im Record Linkage die Ergebnisse aller Stufen vereinigt werden und in ein Clerical Review eingehen, kann auf die ersten Stufen mit restriktiven Blockvariablen verzichtet werden, was zu leichten Laufzeitvorteilen führen dürfte.

5.2 Überprüfung der Verfahrensweisen

5.2.1 Eingesetzte Chiffrierverfahren

Die im EKR NRW entfallenen Kontrollnummern sind verzichtbar. Die Baden-Württemberg-Kontrollnummern sind mit der Einstellung des Krebsregisters Baden-Württemberg obsolet, der DDR-Namenscode würde lediglich beim Abgleich mit Altbeständen (vor 1989) des Gemeinsamen Krebsregisters der Länder Berlin, Brandenburg, Mecklenburg-Vorpommern, Sachsen-Anhalt und der Freistaaten Sachsen und Thüringen (GKR) benötigt. Aufgrund der zeitlichen und räumlichen Distanz wären hier höchstens Einzelfälle zu erwarten, die sowohl in den Altbeständen des GKR als auch im EKR NRW vorkommen.

Die zusätzlich gebildeten Kontrollnummern sind im Hinblick auf das Record Linkage hilfreich. Ihre Erzeugung und Speicherung erschien jedoch zum Zeitpunkt der Erstellung der Empfehlungen unter Datenschutzgesichtspunkten nicht machbar.

Die Chiffrierung beim Melder war bereits in dem Konzept aus Baden-Württemberg [16-18] vorgesehen, was den Vorteil hat, dass keine Klartext-Identitätsdaten übermittelt werden müssen. Allerdings muss sichergestellt sein, dass weder die Schlüssel korrumpiert werden noch eine Dechiffrierung durch Abhandenkommen der Zuordnungslisten unmöglich wird. Dies ist durch die im EKR NRW eingesetzten Verfahren der zentralen Verwaltung und Vergabe der Schlüssel gegeben. Durch das Verfahren ist sichergestellt, dass das EKR NRW – wie sonst die Registerstellen der Krebsregister – nur Identitätschiffrate und Kontrollnummern erhält und nicht in der Lage ist, Identitätsdaten zu dechiffrieren.

Die eingesetzten Chiffrierverfahren sind aus unserer Sicht sinnvoll und wo nötig kompatibel mit denen der übrigen deutschen Krebsregister.

5.2.2 Record Linkage

Durch die Reduktion der abzugleichenden Meldungen wird verhindert, dass der ins Record Linkage eingehende Datenbestand zu groß wird. Dies hat einerseits Vorteile für die Laufzeit des Record Linkage Programms und zum anderen wird damit verhindert, dass die Homonymfehlerrate mit zunehmendem Datenbestand zu stark anwächst. Falls allerdings eine Sterbemeldung fälschlich einer registrierten Person zugeordnet wurde, kann bei diesem Vorgehen der Fehler nicht mehr korrigiert werden, wenn die tatsächliche Sterbeinformation erst viel später eingeht.

Die Frage, wie verfahren werden soll, wenn zu einer Sterbemeldung aus einem Einwohnermeldeamt kein passender Datensatz mit Todesursache unter den Meldungen des LDS[11] zu finden ist, ist noch offen. Die Zusammenführung von Sterbemeldungen und Todesursachendatensätzen aus dem LDS ist noch nicht realisiert. Klar ist, dass die Zusammenführung nicht über das ansonsten angewandte Record Linkage-Verfahren stattfindet, sondern über Merkmale, die speziell für diese Datensätze verfügbar sind.

5.2.3 Qualitätssicherung und Rückfragen bei Meldern

Plausibilitätsprüfungen und Best-of-Regeln

Es ist zu erwägen, ob beim Export von Pathologiedaten Plausibilitätsprüfungen für die Identitätsdaten ergänzt werden können.

Löschung von Meldungen

Aufgrund des Vorgehens im Register wird versehentliches Löschen von Meldungen sehr unwahrscheinlich und ist zudem reversibel. Beabsichtigte Löschvorgänge sind nachvollziehbar. Es ist nicht möglich, Datensätze zu löschen, ohne Spuren zu hinterlassen.

5.2.4 Schlussfolgerungen aus der Überprüfung der Verfahrensweisen

Zusammenfassend lässt sich sagen, dass die Abläufe sinnvoll und durchdacht sind. Einzelne Fragen, etwa das genaue Vorgehen beim Abgleich von Sterbemeldungen mit den Todesursachendaten, werden im weiteren Verlauf des Aufbaus des Krebsregisters noch zu klären sein. Die Fragen, die sich beim Durcharbeiten der zur Verfügung gestellten Dokumente ergaben, konnten umfassend beantwortet werden.

Die Dokumentation der Arbeitsabläufe ist umfassend. Es sollte darauf geachtet werden, dass die Dokumentation bei Einführung neuer Prozesse oder der Änderung von Verfahrensweisen aktualisiert wird. Auch wenn im Kontext des weiteren Registeraufbaus zunächst die Gewinnung und Einbindung weiterer Melder Priorität haben wird, sollte ein Mitführen der Doku-

[11] Landesamt für Datenverarbeitung und Statistik

mentation nicht unterbleiben. Eine aktuelle Dokumentation ist insbesondere bei der Einarbeitung neuer Mitarbeiter hilfreich.

6 Literatur

[1] Gesetz zur Einrichtung eines flächendeckenden bevölkerungsbezogenen Krebsregisters in Nordrhein-Westfalen (EKR-NRW). GV.NRW.2005, 414. 29-5-2005.

[2] Jaro MA. Advances in Record-Linkage Methodology As Applied to Matching the 1985 Census of Tampa, Florida. Journal of the American Statistical Association 1989;84(406):414-20.

[3] Schnell R, Bachteler T, Reiher J. MTB: Ein Record-Linkage-Programm für die empirische Sozialforschung. In: Zentralarchiv für empirische Sozialforschung, Universität zu Köln, editor. ZA-Information. Köln: 2005. p. 93-103.

[4] Espeland M, Odoroff C. Algorithms for computing maximum likelihood estimates from incomplete discrete data. Rochester: University of Rochester, Stat. Dep.; 1984.

[5] Winkler WE. String Comparator Metrics and Enhanced Decision Rules in the Fellegi-Sunter Model of Record Linkage. 1990 p. 354-9.

[6] Winkler WE, Thibaudeau Y. An application of the Fellegi-Sunter model of record linkage to the 1990 U.S. decennial census. U. S. Census Bureau; 1990.

[7] Darroch JN, Ratcliff D. Generalized Iterative Scaling for Log-Linear Models. Annals of Mathematical Statistics 1972;43(5):1470-80.

[8] Fellegi IP, Sunter AB. A Theory for Record Linkage. Journal of the American Statistical Association 1969;64(328):1183-210.

[9] Damerau FJ. A Technique for Computer Detection and Correction of Spelling Errors. Communications of the Acm 1964;7(3):171-6.

[10] Levenshtein VI. Binary Codes for Correcting Deletion Insertion and Substitution Errors. Sov Phys Dokl 1966;10(8):707-10.

[11] Collins MJ. A New Statistical Parser Based on Bigram Lexical Dependencies. 1996 p. 184-1991.

[12] Brenner H, Schmidtmann I. Effects of record linkage errors on disease registration. Methods Inf Med 1998;37(1):69-74.

[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] Brenner H, Schmidtmann I. Determinants of homonym and synonym rates of record linkage in disease registration. Methods Inf Med 1996;35(1):19-24.

[15] Appelrath H-J, 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.

[16] Gruner G, Hartmann S, Meisner C, Pietsch-Breitfeld B, Selbmann HK. Epidemiologisches Krebsregister Baden-Württemberg. Abschlussbericht Langfassung. Tübingen: Institut für Medizinische Informationsverarbeitung der Universität Tübingen; 1989.

[17] Gruner G, Hartmann S, Meisner C, Pietsch-Breitfeld B, Selbmann HK. Epidemiologisches Krebsregister Baden-Württemberg. Stuttgart; 1989.

[18] Schrage R. Krebsregister-Melderechtsmodell mit peripherer Teilanonymisierung. Stuttgart: Krebsverband Baden-Württemberg; 1991.

7 Anhang

Tabelle 17: Übersicht über die Anzahl Synonyme und übersehenen Kanten

Anzahl Meldungen zu Person in Gold-Standard	Anzahl Kanten	Anzahl Personen		Gruppenaufteilung			Anzahl Synonyme	Anzahl übersehene Kanten
i		n_{Gi}						
1	0	n_{G1}	Aufteilung	1			0	0
			Anzahl	s_{11}				
2	1	n_{G2}	Aufteilung	2	1:1		s_{22}	s_{22}
			Anzahl	s_{21}	s_{22}			
			Anzahl Kanten in Gruppen	1	0			
3	3	n_{G3}	Aufteilung	3	2:1	1:1:1	$s_{32} + 2 \cdot s_{33}$	$2 \cdot s_{32} + 3 \cdot s_{33}$
			Anzahl	s_{31}	s_{32}	s_{33}		
			Anzahl Kanten in Gruppen	3	1	0		

Anzahl Meldungen zu Person in Gold-Standard	Anzahl Kanten	Anzahl Personen		Gruppenaufteilung					Anzahl Synonyme	Anzahl übersehene Kanten
4	6	n_{G4}	Aufteilung	4	3:1 2:2		2:1:1	1:1:1:1	$s_{42} + 2 \cdot s_{43}$ $+3 \cdot s_{44}$	$3 \cdot s_{3:1} + 4 \cdot s_{2:2}$ $+5 \cdot s_{43} + 6 \cdot s_{44}$
			Anzahl	s_{41}	$s_{42} = s_{3:1} + s_{2:2}$		s_{43}	s_{44}		
			Anzahl Kanten in Gruppen	6	3 2		1	0		
5	10	n_{G5}	Aufteilung	5	4:1 3:2	3:1:1 2:2:1	2:1:1:1	1:1:1:1:1	$s_{52} + 2 \cdot s_{53}$ $+3 \cdot s_{54} + 4 \cdot s_{55}$	$4 \cdot s_{4:1} + 6 \cdot s_{3:2}$ $+7 \cdot s_{3:1:1} + 8 \cdot s_{2:2:1}$ $+9 \cdot s_{54} + 10 \cdot s_{55}$
			Anzahl	s_{51}	$s_{52} = s_{4:1} + s_{3:2}$	$s_{53} = s_{3:1:1} + s_{2:2:1}$	s_{54}	s_{55}		
			Anzahl Kanten in Gruppen	10	6 4	3 2	1	0		

Anzahl Meldungen zu Person in Gold-Standard	Anzahl Kanten	Anzahl Personen		Gruppenaufteilung						Anzahl Synonyme	Anzahl übersehene Kanten
6	15	n_{G6}	Aufteilung	6	5:1 4:2 3:3	4:1:1 3:2:1 2:2:2	3:1:1:1 2:2:1:1	2:1:1:1:1	1:1:1:1:1:1	$s_{62}+2\cdot s_{63}$ $+3\cdot s_{64}+4\cdot s_{65}$ $+5\cdot s_{66}$	$5\cdot s_{5:1}+8\cdot s_{4:2}$ $+9\cdot s_{3:3}+9\cdot s_{4:1:1}$ $+11\cdot s_{3:2:1}+12\cdot s_{2:2:2}$ $+12\cdot s_{3:1:1:1}+13\cdot s_{2:2:1:1}$ $+14\cdot s_{65}+15\cdot s_{66}$
			Anzahl	s_{61}	$s_{62}=s_{5:1}+s_{4:2}+s_{3:3}$	$s_{63}=s_{4:1:1}+s_{3:2:1}+s_{2:2:2}$	$s_{64}=s_{3:1:1:1}+s_{2:2:1:1}$	s_{65}	s_{66}		
			Anzahl Kanten in Gruppen	15	10 7 6	6 4 3	3 2	1	0		
...	...			...						...	...
Summe	K_G	N_G								N_S	

Tabelle 18: Übersicht über die Anzahl Homonyme und überzählige Kanten

Anzahl zu einer Person zusammengeführte Meldungen im EKR NRW	Anzahl Kanten	Anzahl Matchgruppen		Zusammensetzung: Wie viele Personen sind tatsächlich in Meldungsgruppe, Aufteilung der Meldungen je Person			Anzahl Homonyme	Anzahl überwzähliger Kanten
i		n_{Mi}						
1	0	n_{M1}	Zusammensetzung	1 Person mit 1 Meldung			0	0
			Anzahl	s_{11}				
2	1	n_{M2}	Zusammensetzung	1 Person mit 2 Meldungen	2 Personen mit je 1 Meldung (1:1)		h_{22}	h_{22}
			Anzahl	h_{21}	h_{22}			
			Anzahl „echter“ Kanten	1	0			
3	3	n_{M3}	Zusammenfassung	1 Person mit 3 Meldungen	2 Personen mit 3 Meldungen, Aufteilung 2:1	3 Personen mit je 1 Meldung (1:1:1)	$h_{32} + 2 \cdot h_{33}$	$2 \cdot h_{32} + 3 \cdot h_{33}$
			Anzahl	h_{31}	h_{32}	h_{33}		
			Anzahl „echter“ Kanten	3	1	0		

Anzahl zu einer Person zusammengeführte Meldungen im EKR NRW	Anzahl Kanten	Anzahl Matchgruppen		Zusammensetzung: Wie viele Personen sind tatsächlich in Meldungsgruppe, Aufteilung der Meldungen je Person					Anzahl Homonyme	Anzahl überwzähliger Kanten
4	6	n_{M4}	Zusammenfassung	1 Person mit 4 Meldungen	2 Personen mit 4 Meldungen, Aufteilung 3:1 oder 2:2		3 Personen mit 4 Meldungen, Aufteilung 2:1:1	4 Personen mit je 1 Meldung (11:1:1)	$h_{42} + 2 \cdot h_{43}$ $+3 \cdot h_{44}$	$3 \cdot h_{3:1} + 4 \cdot h_{2:2}$ $+5 \cdot h_{43} + 6 \cdot h_{44}$
			Anzahl	h_{41}	$h_{42} = h_{3:1} + h_{2:2}$		h_{43}	h_{44}		
			Anzahl „echter“ Kanten	6	3 oder2		1	0		
5	10	n_{M5}	Zusammenfassung	1 Person mit 5 Meldungen	2 Personen mit 5 Meldungen, Aufteilung 4:1 oder 3:2	3 Personen mit 5 Meldungen, Aufteilung 3:1:1 oder 2:2:1	3 Personen mit 5 Meldungen, Aufteilung 2:1:1:1	5 Personen mit je 1 Meldung (1:1:1:1:1)	$h_{52} + 2 \cdot h_{53}$ $+3 \cdot h_{54} + 4 \cdot h_{55}$	$4 \cdot h_{4:1} + 6 \cdot h_{3:2}$ $+7 \cdot h_{3:1:1} + 8 \cdot h_{2:2:1}$ $+9 \cdot h_{54} + 10 \cdot h_{55}$
			Anzahl	h_{51}	$h_{52} = h_{4:1} + h_{3:2}$	$h_{53} = h_{3:1:1} + h_{2:2:1}$	h_{54}	h_{55}		
			Anzahl „echter“ Kanten	10	6 oder 4	3 oder 2	1	0		

Anzahl zu einer Person zusammengeführte Meldungen im EKR NRW	Anzahl Kanten	Anzahl Matchgruppen		Zusammensetzung: Wie viele Personen sind tatsächlich in Meldungsgruppe, Aufteilung der Meldungen je Person						Anzahl Homonyme	Anzahl überwzähliger Kanten
6	15	n_{M6}	Zusammenfassung	1 Person mit 6 Meldungen	2 Personen mit 6 Meldungen, Aufteilung 5:1, 4:2 oder 3:3	3 Personen mit 6 Meldungen, Aufteilung 4:1:1, 3:2:1 oder 2:2:2	4 Personen mit 6 Meldungen, Aufteilung 3:1:1:1 oder 2:2:1:1	5 Personen mit 6 Meldungen, Aufteilung 2:1:1:1:1	6 Personen mit je 1 Meldung (1:1:1:1:1:1)	$h_{62} + 2 \cdot h_{63}$ $+3 \cdot h_{64} + 4 \cdot h_{65}$ $+5 \cdot h_{66}$	$5 \cdot h_{5:1} + 8 \cdot h_{4:2}$ $+9 \cdot h_{3:3} + 9 \cdot h_{4:1:1}$ $+11 \cdot h_{3:2:1} + 12 \cdot h_{2:2:2}$ $+12 \cdot h_{3:1:1:1} + 13 \cdot h_{2:2:1:1}$ $+14 \cdot h_{65} + 15 \cdot h_{66}$
			Anzahl	h_{61}	$h_{62} = h_{5:1} + h_{4:2} + h_{3:3}$	$h_{63} = h_{4:1:1} + h_{3:2:1} + h_{2:2:2}$	$h_{64} = h_{3:1:1:1} + h_{2:2:1:1}$	h_{65}	h_{66}		
			Anzahl „echter“ Kanten	15	10, 7 oder 6	6, 4 oder 3	3 oder 2	1	0		
...	...			...						...	...
Summe	K_M	N_M								N_H	

Tabelle 19: Kombinationen von Stufen, in denen die selben Kanten verglichen werden

Art des Abgleichs	Methode	Stufe										Anzahl Vergleiche
		1	2	3	4	5	6	7	8	9	10	
Match Inzidenz und Mortalitätsdaten	Bigram	0	0	0	0	0	0	0	0	0	1	1.453.728
Match Inzidenz und Mortalitätsdaten	Bigram	0	0	0	0	0	0	0	0	1	0	245.069
Match Inzidenz und Mortalitätsdaten	Bigram	0	0	0	0	0	0	0	1	0	0	293.089
Match Inzidenz und Mortalitätsdaten	Bigram	0	0	0	0	0	0	0	1	0	1	182
Match Inzidenz und Mortalitätsdaten	Bigram	0	0	0	0	0	0	1	0	0	0	969.340
Match Inzidenz und Mortalitätsdaten	Bigram	0	0	0	0	0	0	1	0	0	1	610
Match Inzidenz und Mortalitätsdaten	Bigram	0	0	0	0	0	0	1	1	0	0	26.653
Match Inzidenz und Mortalitätsdaten	Bigram	0	0	0	0	0	0	1	1	0	1	19
Match Inzidenz und Mortalitätsdaten	Bigram	0	0	0	0	0	1	0	0	0	0	356.052
Match Inzidenz und Mortalitätsdaten	Bigram	0	0	0	0	0	1	0	0	0	1	236
Match Inzidenz und Mortalitätsdaten	Bigram	0	0	0	0	0	1	0	1	0	0	9.953
Match Inzidenz und Mortalitätsdaten	Bigram	0	0	0	0	0	1	0	1	0	1	4
Match Inzidenz und Mortalitätsdaten	Bigram	0	0	0	0	0	1	1	0	0	0	33.516
Match Inzidenz und Mortalitätsdaten	Bigram	0	0	0	0	0	1	1	0	0	1	24
Match Inzidenz und Mortalitätsdaten	Bigram	0	0	0	0	0	1	1	1	1	0	997
Match Inzidenz und Mortalitätsdaten	Bigram	0	0	0	0	1	0	0	0	1	0	2.075
Match Inzidenz und Mortalitätsdaten	Bigram	0	0	0	0	1	1	1	1	1	0	22
Match Inzidenz und Mortalitätsdaten	Bigram	0	0	0	1	0	0	0	0	0	1	11.152
Match Inzidenz und Mortalitätsdaten	Bigram	0	0	0	1	0	0	0	1	0	1	3
Match Inzidenz und Mortalitätsdaten	Bigram	0	0	0	1	0	0	1	0	0	1	8
Match Inzidenz und Mortalitätsdaten	Bigram	0	0	0	1	0	1	0	0	0	1	3
Match Inzidenz und Mortalitätsdaten	Bigram	0	0	1	0	0	0	0	0	0	1	291
Match Inzidenz und Mortalitätsdaten	Bigram	0	0	1	1	0	0	0	0	0	1	2.786
Match Inzidenz und Mortalitätsdaten	Bigram	0	0	1	1	0	0	0	1	0	1	1
Match Inzidenz und Mortalitätsdaten	Bigram	0	0	1	1	0	0	1	1	0	1	2
Match Inzidenz und Mortalitätsdaten	Bigram	0	1	0	0	0	0	0	0	1	0	65
Match Inzidenz und Mortalitätsdaten	Bigram	0	1	0	0	0	0	0	0	1	1	109
Match Inzidenz und Mortalitätsdaten	Bigram	0	1	0	0	1	0	0	0	1	0	17
Match Inzidenz und Mortalitätsdaten	Bigram	0	1	0	1	1	0	0	0	1	1	23
Match Inzidenz und Mortalitätsdaten	Bigram	0	1	1	0	0	0	0	0	1	1	50
Match Inzidenz und Mortalitätsdaten	Bigram	0	1	1	1	1	0	0	0	1	1	679
Match Inzidenz und Mortalitätsdaten	Bigram	1	1	0	0	0	1	1	1	1	0	3
Match Inzidenz und Mortalitätsdaten	Bigram	1	1	0	0	0	1	1	1	1	1	44
Match Inzidenz und Mortalitätsdaten	Bigram	1	1	0	0	1	1	1	1	1	0	9
Match Inzidenz und Mortalitätsdaten	Bigram	1	1	0	1	1	1	1	1	1	1	24
Match Inzidenz und Mortalitätsdaten	Bigram	1	1	1	0	0	1	1	1	1	1	39
Match Inzidenz und Mortalitätsdaten	Bigram	1	1	1	1	1	1	1	1	1	1	639

Art des Abgleichs	Methode	Stufe										Anzahl Vergleiche
		1	2	3	4	5	6	7	8	9	10	
Match Inzidenz und Mortalitätsdaten	Exakt	0	0	0	0	0	0	0	0	0	1	1.453.728
Match Inzidenz und Mortalitätsdaten	Exakt	0	0	0	0	0	0	0	0	1	0	245.069
Match Inzidenz und Mortalitätsdaten	Exakt	0	0	0	0	0	0	0	1	0	0	293.089
Match Inzidenz und Mortalitätsdaten	Exakt	0	0	0	0	0	0	0	1	0	1	182
Match Inzidenz und Mortalitätsdaten	Exakt	0	0	0	0	0	0	1	0	0	0	969.340
Match Inzidenz und Mortalitätsdaten	Exakt	0	0	0	0	0	0	1	0	0	1	610
Match Inzidenz und Mortalitätsdaten	Exakt	0	0	0	0	0	0	1	1	0	0	26.653
Match Inzidenz und Mortalitätsdaten	Exakt	0	0	0	0	0	0	1	1	0	1	19
Match Inzidenz und Mortalitätsdaten	Exakt	0	0	0	0	0	1	0	0	0	0	356.052
Match Inzidenz und Mortalitätsdaten	Exakt	0	0	0	0	0	1	0	0	0	1	236
Match Inzidenz und Mortalitätsdaten	Exakt	0	0	0	0	0	1	0	1	0	0	9.953
Match Inzidenz und Mortalitätsdaten	Exakt	0	0	0	0	0	1	0	1	0	1	4
Match Inzidenz und Mortalitätsdaten	Exakt	0	0	0	0	0	1	1	0	0	0	33.516
Match Inzidenz und Mortalitätsdaten	Exakt	0	0	0	0	0	1	1	0	0	1	24
Match Inzidenz und Mortalitätsdaten	Exakt	0	0	0	0	0	1	1	1	1	0	997
Match Inzidenz und Mortalitätsdaten	Exakt	0	0	0	0	1	0	0	0	1	0	2.075
Match Inzidenz und Mortalitätsdaten	Exakt	0	0	0	0	1	1	1	1	1	0	22
Match Inzidenz und Mortalitätsdaten	Exakt	0	0	0	1	0	0	0	0	0	1	11.152
Match Inzidenz und Mortalitätsdaten	Exakt	0	0	0	1	0	0	0	1	0	1	3
Match Inzidenz und Mortalitätsdaten	Exakt	0	0	0	1	0	0	1	0	0	1	8
Match Inzidenz und Mortalitätsdaten	Exakt	0	0	0	1	0	1	0	0	0	1	3
Match Inzidenz und Mortalitätsdaten	Exakt	0	0	1	0	0	0	0	0	0	1	291
Match Inzidenz und Mortalitätsdaten	Exakt	0	0	1	1	0	0	0	0	0	1	2.786
Match Inzidenz und Mortalitätsdaten	Exakt	0	0	1	1	0	0	0	1	0	1	1
Match Inzidenz und Mortalitätsdaten	Exakt	0	0	1	1	0	0	1	1	0	1	2
Match Inzidenz und Mortalitätsdaten	Exakt	0	1	0	0	0	0	0	0	1	0	65
Match Inzidenz und Mortalitätsdaten	Exakt	0	1	0	0	0	0	0	0	1	1	109
Match Inzidenz und Mortalitätsdaten	Exakt	0	1	0	0	1	0	0	0	1	0	17
Match Inzidenz und Mortalitätsdaten	Exakt	0	1	0	1	1	0	0	0	1	1	23
Match Inzidenz und Mortalitätsdaten	Exakt	0	1	1	0	0	0	0	0	1	1	50
Match Inzidenz und Mortalitätsdaten	Exakt	0	1	1	1	1	0	0	0	1	1	679
Match Inzidenz und Mortalitätsdaten	Exakt	1	1	0	0	0	1	1	1	1	0	3
Match Inzidenz und Mortalitätsdaten	Exakt	1	1	0	0	0	1	1	1	1	1	44
Match Inzidenz und Mortalitätsdaten	Exakt	1	1	0	0	1	1	1	1	1	0	9
Match Inzidenz und Mortalitätsdaten	Exakt	1	1	0	1	1	1	1	1	1	1	24
Match Inzidenz und Mortalitätsdaten	Exakt	1	1	1	0	0	1	1	1	1	1	39
Match Inzidenz und Mortalitätsdaten	Exakt	1	1	1	1	1	1	1	1	1	1	639
Match Inzidenz und Mortalitätsdaten	Levenshtein	0	0	0	0	0	0	0	0	0	1	1.453.728
Match Inzidenz und Mortalitätsdaten	Levenshtein	0	0	0	0	0	0	0	0	1	0	245.069

Art des Abgleichs	Methode	Stufe										Anzahl Vergleiche
		1	2	3	4	5	6	7	8	9	10	
Match Inzidenz und Mortalitätsdaten	Levenshtein	0	0	0	0	0	0	0	1	0	0	293.089
Match Inzidenz und Mortalitätsdaten	Levenshtein	0	0	0	0	0	0	0	1	0	1	182
Match Inzidenz und Mortalitätsdaten	Levenshtein	0	0	0	0	0	0	1	0	0	0	969.340
Match Inzidenz und Mortalitätsdaten	Levenshtein	0	0	0	0	0	0	1	0	0	1	610
Match Inzidenz und Mortalitätsdaten	Levenshtein	0	0	0	0	0	0	1	1	0	0	26.653
Match Inzidenz und Mortalitätsdaten	Levenshtein	0	0	0	0	0	0	1	1	0	1	19
Match Inzidenz und Mortalitätsdaten	Levenshtein	0	0	0	0	0	1	0	0	0	0	356.052
Match Inzidenz und Mortalitätsdaten	Levenshtein	0	0	0	0	0	1	0	0	0	1	236
Match Inzidenz und Mortalitätsdaten	Levenshtein	0	0	0	0	0	1	0	1	0	0	9.953
Match Inzidenz und Mortalitätsdaten	Levenshtein	0	0	0	0	0	1	0	1	0	1	4
Match Inzidenz und Mortalitätsdaten	Levenshtein	0	0	0	0	0	1	1	0	0	0	33.516
Match Inzidenz und Mortalitätsdaten	Levenshtein	0	0	0	0	0	1	1	0	0	1	24
Match Inzidenz und Mortalitätsdaten	Levenshtein	0	0	0	0	0	1	1	1	1	0	997
Match Inzidenz und Mortalitätsdaten	Levenshtein	0	0	0	0	1	0	0	0	1	0	2.075
Match Inzidenz und Mortalitätsdaten	Levenshtein	0	0	0	0	1	1	1	1	1	0	22
Match Inzidenz und Mortalitätsdaten	Levenshtein	0	0	0	1	0	0	0	0	0	1	11.152
Match Inzidenz und Mortalitätsdaten	Levenshtein	0	0	0	1	0	0	0	1	0	1	3
Match Inzidenz und Mortalitätsdaten	Levenshtein	0	0	0	1	0	0	1	0	0	1	8
Match Inzidenz und Mortalitätsdaten	Levenshtein	0	0	0	1	0	1	0	0	0	1	3
Match Inzidenz und Mortalitätsdaten	Levenshtein	0	0	1	0	0	0	0	0	0	1	291
Match Inzidenz und Mortalitätsdaten	Levenshtein	0	0	1	1	0	0	0	0	0	1	2.786
Match Inzidenz und Mortalitätsdaten	Levenshtein	0	0	1	1	0	0	0	1	0	1	1
Match Inzidenz und Mortalitätsdaten	Levenshtein	0	0	1	1	0	0	1	1	0	1	2
Match Inzidenz und Mortalitätsdaten	Levenshtein	0	1	0	0	0	0	0	0	1	0	65
Match Inzidenz und Mortalitätsdaten	Levenshtein	0	1	0	0	0	0	0	0	1	1	109
Match Inzidenz und Mortalitätsdaten	Levenshtein	0	1	0	0	1	0	0	0	1	0	17
Match Inzidenz und Mortalitätsdaten	Levenshtein	0	1	0	1	1	0	0	0	1	1	23
Match Inzidenz und Mortalitätsdaten	Levenshtein	0	1	1	0	0	0	0	0	1	1	50
Match Inzidenz und Mortalitätsdaten	Levenshtein	0	1	1	1	1	0	0	0	1	1	679
Match Inzidenz und Mortalitätsdaten	Levenshtein	1	1	0	0	0	1	1	1	1	0	3
Match Inzidenz und Mortalitätsdaten	Levenshtein	1	1	0	0	0	1	1	1	1	1	44
Match Inzidenz und Mortalitätsdaten	Levenshtein	1	1	0	0	1	1	1	1	1	0	9
Match Inzidenz und Mortalitätsdaten	Levenshtein	1	1	0	1	1	1	1	1	1	1	24
Match Inzidenz und Mortalitätsdaten	Levenshtein	1	1	1	0	0	1	1	1	1	1	39
Match Inzidenz und Mortalitätsdaten	Levenshtein	1	1	1	1	1	1	1	1	1	1	639
Deduplizieren Inzidenz-Meldungen	Bigram	0	0	0	0	0	0	0	0	0	1	1.379.904
Deduplizieren Inzidenz-Meldungen	Bigram	0	0	0	0	0	0	0	0	1	0	303.882
Deduplizieren Inzidenz-Meldungen	Bigram	0	0	0	0	0	0	0	1	0	0	813.425
Deduplizieren Inzidenz-Meldungen	Bigram	0	0	0	0	0	0	0	1	0	1	529

Art des Abgleichs	Methode	Stufe										Anzahl Vergleiche
		1	2	3	4	5	6	7	8	9	10	
Deduplizieren Inzidenz-Meldungen	Bigram	0	0	0	0	0	0	1	0	0	0	2.212.379
Deduplizieren Inzidenz-Meldungen	Bigram	0	0	0	0	0	0	1	0	0	1	1.283
Deduplizieren Inzidenz-Meldungen	Bigram	0	0	0	0	0	0	1	1	0	0	75.121
Deduplizieren Inzidenz-Meldungen	Bigram	0	0	0	0	0	0	1	1	0	1	58
Deduplizieren Inzidenz-Meldungen	Bigram	0	0	0	0	0	1	0	0	0	0	810.576
Deduplizieren Inzidenz-Meldungen	Bigram	0	0	0	0	0	1	0	0	0	1	436
Deduplizieren Inzidenz-Meldungen	Bigram	0	0	0	0	0	1	0	1	0	0	27.148
Deduplizieren Inzidenz-Meldungen	Bigram	0	0	0	0	0	1	0	1	0	1	19
Deduplizieren Inzidenz-Meldungen	Bigram	0	0	0	0	0	1	1	0	0	0	75.108
Deduplizieren Inzidenz-Meldungen	Bigram	0	0	0	0	0	1	1	0	0	1	52
Deduplizieren Inzidenz-Meldungen	Bigram	0	0	0	0	0	1	1	1	1	0	2.647
Deduplizieren Inzidenz-Meldungen	Bigram	0	0	0	0	1	0	0	0	1	0	4.372
Deduplizieren Inzidenz-Meldungen	Bigram	0	0	0	0	1	1	1	1	1	0	143
Deduplizieren Inzidenz-Meldungen	Bigram	0	0	0	1	0	0	0	0	0	1	17.547
Deduplizieren Inzidenz-Meldungen	Bigram	0	0	0	1	0	0	0	1	0	1	3
Deduplizieren Inzidenz-Meldungen	Bigram	0	0	0	1	0	0	1	0	0	1	9
Deduplizieren Inzidenz-Meldungen	Bigram	0	0	0	1	0	0	1	1	0	1	2
Deduplizieren Inzidenz-Meldungen	Bigram	0	0	0	1	0	1	0	0	0	1	1
Deduplizieren Inzidenz-Meldungen	Bigram	0	0	0	1	0	1	1	0	0	1	4
Deduplizieren Inzidenz-Meldungen	Bigram	0	0	1	0	0	0	0	0	0	1	49
Deduplizieren Inzidenz-Meldungen	Bigram	0	0	1	1	0	0	0	0	0	1	3.701
Deduplizieren Inzidenz-Meldungen	Bigram	0	0	1	1	0	0	0	1	0	1	13
Deduplizieren Inzidenz-Meldungen	Bigram	0	0	1	1	0	0	1	0	0	1	6
Deduplizieren Inzidenz-Meldungen	Bigram	0	0	1	1	0	0	1	1	0	1	39
Deduplizieren Inzidenz-Meldungen	Bigram	0	0	1	1	0	1	0	0	0	1	4
Deduplizieren Inzidenz-Meldungen	Bigram	0	0	1	1	0	1	0	1	0	1	26
Deduplizieren Inzidenz-Meldungen	Bigram	0	0	1	1	0	1	1	0	0	1	50
Deduplizieren Inzidenz-Meldungen	Bigram	0	1	0	0	0	0	0	0	1	0	89
Deduplizieren Inzidenz-Meldungen	Bigram	0	1	0	0	0	0	0	0	1	1	113
Deduplizieren Inzidenz-Meldungen	Bigram	0	1	0	0	1	0	0	0	1	0	14
Deduplizieren Inzidenz-Meldungen	Bigram	0	1	0	1	1	0	0	0	1	1	18
Deduplizieren Inzidenz-Meldungen	Bigram	0	1	1	1	1	0	0	0	1	1	130
Deduplizieren Inzidenz-Meldungen	Bigram	1	0	0	0	0	1	1	1	0	0	624
Deduplizieren Inzidenz-Meldungen	Bigram	1	1	0	0	0	1	1	1	1	0	29
Deduplizieren Inzidenz-Meldungen	Bigram	1	1	0	0	0	1	1	1	1	1	609
Deduplizieren Inzidenz-Meldungen	Bigram	1	1	0	0	1	1	1	1	1	0	222
Deduplizieren Inzidenz-Meldungen	Bigram	1	1	0	1	1	1	1	1	1	1	202
Deduplizieren Inzidenz-Meldungen	Bigram	1	1	1	0	0	1	1	1	1	1	75
Deduplizieren Inzidenz-Meldungen	Bigram	1	1	1	1	1	1	1	1	1	1	19.323

Art des Abgleichs	Methode	Stufe										Anzahl Vergleiche
		1	2	3	4	5	6	7	8	9	10	
Deduplizieren Inzidenz-Meldungen	Exakt	0	0	0	0	0	0	0	0	0	1	1.379.904
Deduplizieren Inzidenz-Meldungen	Exakt	0	0	0	0	0	0	0	0	1	0	303.882
Deduplizieren Inzidenz-Meldungen	Exakt	0	0	0	0	0	0	0	1	0	0	813.425
Deduplizieren Inzidenz-Meldungen	Exakt	0	0	0	0	0	0	0	1	0	1	529
Deduplizieren Inzidenz-Meldungen	Exakt	0	0	0	0	0	0	1	0	0	0	2.212.379
Deduplizieren Inzidenz-Meldungen	Exakt	0	0	0	0	0	0	1	0	0	1	1.283
Deduplizieren Inzidenz-Meldungen	Exakt	0	0	0	0	0	0	1	1	0	0	75.121
Deduplizieren Inzidenz-Meldungen	Exakt	0	0	0	0	0	0	1	1	0	1	58
Deduplizieren Inzidenz-Meldungen	Exakt	0	0	0	0	0	1	0	0	0	0	810.576
Deduplizieren Inzidenz-Meldungen	Exakt	0	0	0	0	0	1	0	0	0	1	436
Deduplizieren Inzidenz-Meldungen	Exakt	0	0	0	0	0	1	0	1	0	0	27.148
Deduplizieren Inzidenz-Meldungen	Exakt	0	0	0	0	0	1	0	1	0	1	19
Deduplizieren Inzidenz-Meldungen	Exakt	0	0	0	0	0	1	1	0	0	0	75.108
Deduplizieren Inzidenz-Meldungen	Exakt	0	0	0	0	0	1	1	0	0	1	52
Deduplizieren Inzidenz-Meldungen	Exakt	0	0	0	0	0	1	1	1	0	0	594
Deduplizieren Inzidenz-Meldungen	Exakt	0	0	0	0	0	1	1	1	1	0	2.647
Deduplizieren Inzidenz-Meldungen	Exakt	0	0	0	0	1	0	0	0	1	0	4.372
Deduplizieren Inzidenz-Meldungen	Exakt	0	0	0	0	1	1	1	1	1	0	143
Deduplizieren Inzidenz-Meldungen	Exakt	0	0	0	1	0	0	0	0	0	1	17.547
Deduplizieren Inzidenz-Meldungen	Exakt	0	0	0	1	0	0	0	1	0	1	3
Deduplizieren Inzidenz-Meldungen	Exakt	0	0	0	1	0	0	1	0	0	1	9
Deduplizieren Inzidenz-Meldungen	Exakt	0	0	0	1	0	0	1	1	0	1	2
Deduplizieren Inzidenz-Meldungen	Exakt	0	0	0	1	0	1	0	0	0	1	1
Deduplizieren Inzidenz-Meldungen	Exakt	0	0	0	1	0	1	1	0	0	1	4
Deduplizieren Inzidenz-Meldungen	Exakt	0	0	1	0	0	0	0	0	0	1	49
Deduplizieren Inzidenz-Meldungen	Exakt	0	0	1	1	0	0	0	0	0	1	3.701
Deduplizieren Inzidenz-Meldungen	Exakt	0	0	1	1	0	0	0	1	0	1	13
Deduplizieren Inzidenz-Meldungen	Exakt	0	0	1	1	0	0	1	0	0	1	6
Deduplizieren Inzidenz-Meldungen	Exakt	0	0	1	1	0	0	1	1	0	1	39
Deduplizieren Inzidenz-Meldungen	Exakt	0	0	1	1	0	1	0	0	0	1	4
Deduplizieren Inzidenz-Meldungen	Exakt	0	0	1	1	0	1	0	1	0	1	26
Deduplizieren Inzidenz-Meldungen	Exakt	0	0	1	1	0	1	1	0	0	1	50
Deduplizieren Inzidenz-Meldungen	Exakt	0	1	0	0	0	0	0	0	1	0	89
Deduplizieren Inzidenz-Meldungen	Exakt	0	1	0	0	0	0	0	0	1	1	113
Deduplizieren Inzidenz-Meldungen	Exakt	0	1	0	0	0	1	1	1	1	0	7
Deduplizieren Inzidenz-Meldungen	Exakt	0	1	0	0	0	1	1	1	1	1	219
Deduplizieren Inzidenz-Meldungen	Exakt	0	1	0	0	1	0	0	0	1	0	14
Deduplizieren Inzidenz-Meldungen	Exakt	0	1	0	0	1	1	1	1	1	0	62
Deduplizieren Inzidenz-Meldungen	Exakt	0	1	0	1	1	0	0	0	1	1	18

Art des Abgleichs	Methode	Stufe										Anzahl Vergleiche
		1	2	3	4	5	6	7	8	9	10	
Deduplizieren Inzidenz-Meldungen	Exakt	0	1	0	1	1	1	1	1	1	1	53
Deduplizieren Inzidenz-Meldungen	Exakt	0	1	1	0	0	1	1	1	1	1	22
Deduplizieren Inzidenz-Meldungen	Exakt	0	1	1	1	1	0	0	0	1	1	130
Deduplizieren Inzidenz-Meldungen	Exakt	0	1	1	1	1	1	1	1	1	1	4.069
Deduplizieren Inzidenz-Meldungen	Exakt	1	0	0	0	0	1	1	1	0	0	30
Deduplizieren Inzidenz-Meldungen	Exakt	1	1	0	0	0	1	1	1	1	0	22
Deduplizieren Inzidenz-Meldungen	Exakt	1	1	0	0	0	1	1	1	1	1	390
Deduplizieren Inzidenz-Meldungen	Exakt	1	1	0	0	1	1	1	1	1	0	160
Deduplizieren Inzidenz-Meldungen	Exakt	1	1	0	1	1	1	1	1	1	1	149
Deduplizieren Inzidenz-Meldungen	Exakt	1	1	1	0	0	1	1	1	1	1	53
Deduplizieren Inzidenz-Meldungen	Exakt	1	1	1	1	1	1	1	1	1	1	15.254
Deduplizieren Inzidenz-Meldungen	Levenshtein	0	0	0	0	0	0	0	0	0	1	1.379.904
Deduplizieren Inzidenz-Meldungen	Levenshtein	0	0	0	0	0	0	0	0	1	0	303.882
Deduplizieren Inzidenz-Meldungen	Levenshtein	0	0	0	0	0	0	0	1	0	0	813.425
Deduplizieren Inzidenz-Meldungen	Levenshtein	0	0	0	0	0	0	0	1	0	1	529
Deduplizieren Inzidenz-Meldungen	Levenshtein	0	0	0	0	0	0	1	0	0	0	2.212.379
Deduplizieren Inzidenz-Meldungen	Levenshtein	0	0	0	0	0	0	1	0	0	1	1.283
Deduplizieren Inzidenz-Meldungen	Levenshtein	0	0	0	0	0	0	1	1	0	0	75.121
Deduplizieren Inzidenz-Meldungen	Levenshtein	0	0	0	0	0	0	1	1	0	1	58
Deduplizieren Inzidenz-Meldungen	Levenshtein	0	0	0	0	0	1	0	0	0	0	810.576
Deduplizieren Inzidenz-Meldungen	Levenshtein	0	0	0	0	0	1	0	0	0	1	436
Deduplizieren Inzidenz-Meldungen	Levenshtein	0	0	0	0	0	1	0	1	0	0	27.148
Deduplizieren Inzidenz-Meldungen	Levenshtein	0	0	0	0	0	1	0	1	0	1	19
Deduplizieren Inzidenz-Meldungen	Levenshtein	0	0	0	0	0	1	1	0	0	0	75.108
Deduplizieren Inzidenz-Meldungen	Levenshtein	0	0	0	0	0	1	1	0	0	1	52
Deduplizieren Inzidenz-Meldungen	Levenshtein	0	0	0	0	0	1	1	1	1	0	2.647
Deduplizieren Inzidenz-Meldungen	Levenshtein	0	0	0	0	1	0	0	0	1	0	4.372
Deduplizieren Inzidenz-Meldungen	Levenshtein	0	0	0	0	1	1	1	1	1	0	143
Deduplizieren Inzidenz-Meldungen	Levenshtein	0	0	0	1	0	0	0	0	0	1	17.547
Deduplizieren Inzidenz-Meldungen	Levenshtein	0	0	0	1	0	0	0	1	0	1	3
Deduplizieren Inzidenz-Meldungen	Levenshtein	0	0	0	1	0	0	1	0	0	1	9
Deduplizieren Inzidenz-Meldungen	Levenshtein	0	0	0	1	0	0	1	1	0	1	2
Deduplizieren Inzidenz-Meldungen	Levenshtein	0	0	0	1	0	1	0	0	0	1	1
Deduplizieren Inzidenz-Meldungen	Levenshtein	0	0	0	1	0	1	1	0	0	1	4
Deduplizieren Inzidenz-Meldungen	Levenshtein	0	0	1	0	0	0	0	0	0	1	49
Deduplizieren Inzidenz-Meldungen	Levenshtein	0	0	1	1	0	0	0	0	0	1	3.701
Deduplizieren Inzidenz-Meldungen	Levenshtein	0	0	1	1	0	0	0	1	0	1	13
Deduplizieren Inzidenz-Meldungen	Levenshtein	0	0	1	1	0	0	1	0	0	1	6
Deduplizieren Inzidenz-Meldungen	Levenshtein	0	0	1	1	0	0	1	1	0	1	39

Art des Abgleichs	Methode	Stufe										Anzahl Vergleiche
		1	2	3	4	5	6	7	8	9	10	
Deduplizieren Inzidenz-Meldungen	Levenshtein	0	0	1	1	0	1	0	0	0	1	4
Deduplizieren Inzidenz-Meldungen	Levenshtein	0	0	1	1	0	1	0	1	0	1	26
Deduplizieren Inzidenz-Meldungen	Levenshtein	0	0	1	1	0	1	1	0	0	1	50
Deduplizieren Inzidenz-Meldungen	Levenshtein	0	1	0	0	0	0	0	0	1	0	89
Deduplizieren Inzidenz-Meldungen	Levenshtein	0	1	0	0	0	0	0	0	1	1	113
Deduplizieren Inzidenz-Meldungen	Levenshtein	0	1	0	0	1	0	0	0	1	0	14
Deduplizieren Inzidenz-Meldungen	Levenshtein	0	1	0	1	1	0	0	0	1	1	18
Deduplizieren Inzidenz-Meldungen	Levenshtein	0	1	1	1	1	0	0	0	1	1	130
Deduplizieren Inzidenz-Meldungen	Levenshtein	1	0	0	0	0	1	1	1	0	0	624
Deduplizieren Inzidenz-Meldungen	Levenshtein	1	1	0	0	0	1	1	1	1	0	29
Deduplizieren Inzidenz-Meldungen	Levenshtein	1	1	0	0	0	1	1	1	1	1	609
Deduplizieren Inzidenz-Meldungen	Levenshtein	1	1	0	0	1	1	1	1	1	0	222
Deduplizieren Inzidenz-Meldungen	Levenshtein	1	1	0	1	1	1	1	1	1	1	202
Deduplizieren Inzidenz-Meldungen	Levenshtein	1	1	1	0	0	1	1	1	1	1	75
Deduplizieren Inzidenz-Meldungen	Levenshtein	1	1	1	1	1	1	1	1	1	1	19.323