German Record Linkage Center
WORKING PAPER SERIES
NO. WP-GRLC-2014-03 | MARCH 28, 2014

‘STROKES’ – Record Linkage der Schlaganfälle in Hessen 2007-2010

Tobias Gramlich

Inhaltsverzeichnis

Zusammenfassung

Dieser Bericht beschreibt die Verknüpfung verschiedener Behandlungsstufen der Behandlung und Rehabilitation von Schlaganfallpatienten in Hessen. Verknüpft wurden alle Akutfälle der Jahre 2007/2008 bzw. 2009/2010 mit den Fällen der Frührehabilitation sowie der Rehabilitationsbehandlung in hessischen Kliniken. Insgesamt wurden fünf Verknüpfungen durchgeführt. Dieser Bericht beschreibt die Aufbereitung der Daten, das Vorgehen bei der Verknüpfung sowie die Ergebnisse der Verknüpfung.

Keywords: Record Linkage, Qualitätssicherung, Schlaganfall, Hessen

1 Datenquellen

Verknüpft wurden 4 verschiedene Datenquellen mit unterschiedlicher bzw. unbekannter Überlappung:

1. Akutbehandlung aller Schlaganfallpatienten in den Jahren 2007 und 2008
2. Akutbehandlung aller Schlaganfallpatienten in den Jahren 2009 und 2010
3. Patienten der Frührehabilitationsprogramme
4. Patienten der Rehabilitationsbehandlung

Verknüpft wurden die Schlaganfallpatienten mit den Patienten in der Frührehabilitation, d.h.

- 1) und 3) sowie 2) und 3),

die Schlaganfallpatienten mit den Patienten in der Rehabilitation, d.h.

- 1) und 4) sowie 2) und 4),

sowie die Patienten in der Frührehabilitation und Rehabilitation, d.h. 3) und 4). Nicht verknüpft wurden die verschiedenen Schlaganfallkohorten 2007/2008 und 2009/2010.

Für alle Verknüpfungen sind Überlappungen unterschiedlich groß bzw. unbekannt: Ein Teil der akuten Schlaganfallpatienten verstirbt, ein Teil verlässt – evtl. zur weiteren Behandlung – das Bundesland. Nicht alle Schlaganfallpatienten nehmen an Frührehabilitationsprogramme oder Rehabilitationsprogrammen (überhaupt oder in hessischen Kliniken) teil oder erst nach großer Zwischenzeit. Das bedeutet, dass nicht notwendigerweise alle Schlaganfallpatienten aus den Akut-Datensätzen in den Datensätzen der Frührehabilitation und/oder der Rehabilitation wieder erscheinen.

1.1 Akutfälle 2007/2008 und 2009/2010

Aufbereitet werden zwei Datensätze: alle Akutfälle aus den Jahren 2007/2008 (42301 Zeilen) sowie aus den Jahren 2009/2010 (49147 Zeilen). Die Aufbereitung der Datensätze verläuft für beide Kohorten nahezu identisch (auf Unterschiede wird ggf. explizit hingewiesen).

Als Block- und Matchvariablen für die verschiedenen Verknüpfungen können aus den Akutdatensätzen verwendet werden: Die Namensinitialen (jeweils erster Buchstabe Vor- und Nachname), das Geschlecht, das Geburtsdatum, das Aufnahmedatum in die Akutklinik, das Entlassungsdatum aus der Akutklinik, die PLZ des Wohnorts, die Krankenhausnummer der Akutklinik sowie die Nummer der Folgeeinrichtung, der ICD-Code sowie prinzipiell auch verschiedene Indizes bei Entlassung (die allesamt aber sehr wenig Entropie besitzen und deren Datenqualität unbekannt ist), z.B. der Rankin-Scale-Index, verschiedene Indizes (zu Lagewechsel, Fortbewegung oder Blasenkontrolle), sowie verschiedene Indizes verschiedener Behinderungen (Dysphagie, Aphasie, Dysarthrie und Paresie) bei Entlassung aus der Akutbehandlung.

1.1.1 Aufbereitung

Zunächst werden alle verstorbenen Patienten (`entverl==11`) aus dem Datensatz entfernt; für den Datensatz der Akutfälle 2007/2008 betrifft das 2658 Fälle, für den Datensatz 2009/2010 2795 Zeilen. Damit verbleiben für eine Verknüpfung 39643 bzw. 46352 Zeilen.

Die nächsten Schritte umfassen nun die Aufbereitung der Initialen (einheitliche Großschreibung, Auflösen von Umlauten in einstellige Vokale (z.B. Ö–>O), Trennung von Vor- und Nachnameninital; Erzeugung einer Variante mit „gedrehten“ Initialen (VN–>NV))

Aus allen Datumsangaben (Geburtsdatum, Aufnahme- und Entlassdatum) werden die einzelnen Komponenten – Jahr, Monat und Tag – extrahiert; bei der Postleitzahl wird die Länge kontrolliert und Teilstücke der Postleitzahl gebildet (5-stellig, die ersten vier Stellen, die ersten drei Stellen der Postleitzahl).

Bei den ICD-Klassifikationen treten nur Codes der Kapitel `VI` (Krankheiten des Nervensystems) und `IX` (Krankheiten des Kreislaufsystems) der Gruppen `G45` (zerebrale transitorische Ischämie und verwandte Syndrome) und `I60–I64` (Subarachnoidalblutung; intrazerebrale Blutung; sonstige nichttraumatische intrakranielle Blutung; Hirninfarkt; Schlaganfall, nicht als Blutung oder Infarkt bezeichnet) auf. Verwendet für die Verknüpfung werden nur diese Kapitel und Gruppen, keine weiteren ICD-Untergruppen (da die Datenqualität und die Stabilität der Einordnung in die Untergruppen unklar ist und ICD-Codes z.T. auch nur bis zur Gruppe vorliegen); d.h. für die Verknüpfung werden keine Krankheitsklassen (Kategorien/Subkategorien) verwendet sondern nur die allgemeine dreistellige Systematik.

Tabelle 1 und 2 zeigen die potentiell verfügbaren bzw. zur Verknüpfung verwendeten Merkmale (nach Bereinigung und Aufbereitung). Zur Beurteilung, ob ein Merkmal als Block- oder Matchvariable verwendet werden kann ist die Beurteilung des Füllgrads (d.h. der Anteil fehlender Werte) und die Differenzierungskraft (hier: Entropie) von Bedeutung. Deutlich ist aus den Tabellen 1 und 2 zu erkennen, dass einige Merkmale zwar prinzipiell verfügbar sind (da sie auch in den zu verknüpfenden Datensätzen der Frührehabilitation oder Rehabilitation erfasst sind), aber aufgrund eines geringen Füllgrades nicht als Merkmal zum Blocken oder Verknüpfen eignen, oder aber trotz hohem Füllgrad ungeeignet sind, weil sie nur eine geringe Differenzierungskraft aufweisen (z.B. weil Merkmale nur wenige Ausprägungen annehmen können – z.B. deutlich beim Geschlecht oder dem Jahr der Aufnahme (jeweils nur zwei Ausprägungen) – oder weil sich viele Beobachtungen dann stark auf einzelne wenige Ausprägungen konzentrieren (z.b. beim ICD-Code).

Schließlich werden Datensätze dedupliziert nach Aufnahmedatum, Geschlecht, Initialen, Geburtsdatum und Entlassdatum innerhalb einer Akutklinik. 35 Dubletten sind im Datensatz 2007/2008 enthalten (und werden entfernt), so dass 39608 Zeilen verbleiben; im Datensatz 2009/2010 sind 314 Dubletten enthalten und werden entfernt, so dass schließlich 46308 Zeilen verbleiben.

Tabelle 1: Schlaganfälle 2007/2008: Füllgrade und Entropie verfügbarer/verwendeter Merkmale

Merkmale	% Missing	Entropie
Aufnahmedatum – Jahr	0.0	1.0
Aufnahmedatum – Monat	0.0	3.6
Aufnahmedatum – Tag	0.0	5.0
Entlassdatum – Jahr	0.0	1.1
Entlassdatum – Monat	0.0	3.6
Entlassdatum – Tag	0.0	5.0
Geburtsjahr	0.0	5.6
Geburtsmonat	0.0	3.6
Geburtstag	0.0	5.0
Geschlecht	0.0	1.0
ICD-Code	0.0	1.4
Initial Vorname	0.0	4.1
Initial Nachname	0.0	4.1
PLZ5	0.2	7.9
scbibl_e	0.1	1.1
scbifo_e	0.1	1.7
scbitr_e	0.1	1.6
scrank_e	0.1	2.5
sypar_e	0.1	1.3
syschl_e	88.7	0.8
syspra_e	83.7	0.0
syspre_e	87.1	0.0
KHNr Folgeeinrichtung	82.4	5.0

Datensatz und Merkmale aufbereitet, 39608 Zeilen

Tabelle 2: Schlaganfälle 2009/2010: Füllgrade und Entropie verfügbarer/verwendeter Merkmale

Merkmale	% Missing	Entropie
Aufnahmedatum – Jahr	0.0	1.0
Aufnahmedatum – Monat	0.0	3.6
Aufnahmedatum – Tag	0.0	5.0
Entlassdatum – Jahr	5.7	1.1
Entlassdatum – Monat	5.7	3.6
Entlassdatum – Tag	5.7	4.9
Geburtsjahr	0.0	5.6
Geburtsmonat	0.0	3.6
Geburtstag	0.0	5.7
Geschlecht	0.0	1.0
ICD-Code	0.0	1.4
Initial Vorname	0.1	4.1
Initial Nachname	0.2	4.1
PLZ5	0.3	8.9
scbibl_e	6.4	1.1
scbifo_e	6.4	1.7
scbitr_e	6.4	1.6
scrank_e	5.8	2.5
sypar_e	0.1	1.3
syschl_e	92.2	1.3
syspra_e	83.6	0.0
syspre_e	88.2	0.0
KHNr Folgeeinrichtung	–	–

Datensatz und Merkmale aufbereitet, 46038 Zeilen

Tabelle 3: Anzahl Zeilen im Datensatz der Frührehabilitationspatienten nach (verzeichneten) Jahr des Schlaganfalls

Jahr	Anzahl Zeilen
2006	113
2007	778
2008	955
2009	1010
2010	709
2011	2
unbekannt	186
Insgesamt	3753

1.2 Frührehabilitation

Als Block und Matchvariablen können verwendet werden: Die Namensinitialen (jeweils erster Buchstabe Vor- und Nachname), das Geschlecht, das Geburtsdatum, Aufnahmedatum in Frührehaklinik, Entlassungsdatum aus Frührehaklinik, PLZ des Wohnorts, die Krankenhausnummer der Klinik, der ICD-Code sowie verschiedene Indizes bei Aufnahme (die allesamt aber sehr wenig Entropie besitzen und deren Datenqualität unbekannt ist).

Der Datensatz der Patienten der Frührehabilitationsprogramme enthält Patienten aus verschiedenen Jahren. Alle Zeilen mit verzeichnetem Ereignisjahr bzw. Aufnahme in eine Frührehabilitationsklinik vor 2007 werden entfernt, ebenso wie alle Zeilen mit verzeichnetem Schlaganfall nach 2010. Damit verbleiben 3452 (von 3753) Zeilen. Für die Verknüpfung mit den Datensätzen der jeweiligen Schlaganfallkohorten werden entsprechend weitere Fälle aus dem Datensatz entfernt.

Die Aufbereitung der verfügbaren und zur Verknüpfung verwendeten Block- und Matchvariablen verläuft analog zu den Datensätzen der Schlaganfallkohorten (Abschnitt 1.1): Initialen von Vor- und Nachname, Aufspaltung der Datumsangaben, Aufbereitung des ICD-Codes.

Tabelle 4 zeigt Füllgrad und Entropie der verfügbaren bzw. zur Verknüpfung verwendbaren Merkmale nach Aufbereitung.[1]

Der Datensatz der Frührehabilitationspatienten wird ebenfalls dedupliziert nach Aufnahmedatum, Geschlecht, Namen und Geburtsdatum, enthält aber keine Dubletten (innerhalb einer Krankenhausnummer). Somit verbleiben insgesamt 3452 Zeilen zur Verknüpfung mit den Schlaganfallkohorten bzw. dem Datensatz mit den Rehabilitationspatienten.

1.3 Rehabilitation

Im Datensatz der Rehabilitationspatienten liegen als mögliche Block- und Matchvariablen vor: Die Namensinitialen (jeweils erster Buchstabe Vor- und Nachname), das Geschlecht,

[1] Die Tabelle unterscheidet nicht nach dem Jahr des Schlaganfalls 2007/2008 – 2009/2010. Die Entropie der Merkmale für die getrennten Schlaganfallkohorten wird daher niedriger als berichtet liegen

Tabelle 4: Frührehabilitation: Füllgrade und Entropie verfügbarer/verwendeter Merkmale

Merkmale	% Missing	Entropie
Aufnahmedatum – Jahr	0.0	2.0
Aufnahmedatum – Monat	0.0	3.6
Aufnahmedatum – Tag	0.0	5.0
Schlaganfalldatum – Jahr	0.0	2.0
Schlaganfalldatum – Monat	0.0	3.6
Schlaganfalldatum – Tag	0.0	4.9
Geburtsjahr	0.0	5.7
Geburtsmonat	0.0	3.6
Geburtstag	0.0	4.9
Geschlecht	0.0	1.0
ICD-Code	0.0	1.3
Initial Vorname	0.0	4.1
Initial Nachname	0.0	4.1
PLZ5	48.5	8.8
aphasie_a	9.1	1.0
dysarthrie_a	10.1	1.0
dysphagie_a	3.0	0.9
scbi_06_a	0.0	0.4
scbi_08_a	0.0	1.0
scbi_09_a	0.0	0.2
scrank_a	0.0	1.1
KHNr der zuweisende Einrichtung	–	–
KHNr der Folgeeinrichtung	–	–

Datensatz und Merkmale aufbereitet, 3452 Zeilen

Zahlen enthalten alle berücksichtigten Jahre; Entropie daher eher überschätzt

Tabelle 5: Anzahl Zeilen im Datensatz der Rehabilitationspatienten nach (verzeichneten) Jahr des Schlaganfalls

Jahr	Anzahl Zeilen
2006	3863
2007	3214
2008	2830
2009	3442
2010	3076
2011	234
Insgesamt	16659

das Geburtsdatum, Aufnahmedatum in Rehaklinik, Datum des Schlaganfalls, Entlassungsdatum aus Akuteinrichtung, Nummer der zuweisenden Einrichtung, die Krankenhausnummer der Klinik, der ICD-Code sowie verschiedene Indizes bei Aufnahme (die allesamt aber sehr wenig Entropie besitzen und deren Datenqualität unbekannt ist, s.o.).

Ausgeschlossen von einer weiteren Aufbereitung und Verknüpfung werden Zeilen, bei denen offenbar ungültige Datumskombinationen erfasst sind (z.B. Datum Schlaganfall *nach* Aufnahme in Rehaklinik; oder Aufnahme in Rehabilitationsklinik *vor* dem Entlassdatum aus der Akutklinik.[2] Dadurch werden von ursprünglich 16721 Zeilen 53 Beobachtungen aus dem Datensatz entfernt.

Dieser Datensatz umfasst die Patienten in Rehabilitationsprogrammen mehrerer Jahre, d.h. für die Verknüpfung mit den verschiedenen Schlaganfallkohorten werden Zeilen über das im Rehabilitationsdatensatz erfasste Jahr des Schlaganfalls ausgeschlossen.

Die Aufbereitung der verfügbaren und zur Verknüpfung verwendeten Block- und Matchvariablen verläuft analog zu den Datensätzen der Schlaganfallkohorten (Abschnitt 1.1.1) bzw. dem Datensatz zur Frührehabilitation: Initialen von Vor- und Nachname, Aufspaltung der Datumsangaben, Aufbereitung des ICD-Codes. Schließlich wird der Datensatz dedupliziert und 9 Dubletten entfernt, so dass insgesamt 16659 Zeilen verbleiben.

2 Probleme und Auffälligkeiten

Problematisch für eine Verknüpfung der Datensätze ist die Anzahl und Qualität der zur Verknüpfung zur Verfügung stehenden Merkmale: das betrifft zum Einen zunächst den Füllgrad, d.h. der Grad, zu dem das Merkmal in den Datensätzen tatsächlich auch vorliegt. Insbesondere das Merkmal der Krankenhausnummern, *in die* ein Patient/eine Patientin verlegt wird bzw. *aus der* ein Patient/eine Patientin verlegt wird ist schlecht gefüllt und kann dann nur für einen (kleinen) Teil der Patienten als zusätzliches Merkmal zur Verknüpfung verwendet werden. Zudem ist bei einigen Merkmalen die Datenqualität unklar, d.h. das Ausmaß, wie *akkurat und korrekt* Merkmale erfasst sind. Auch das betrifft u.a. die Nummern des Krankenhauses, *in die* bzw. *aus denen* Patienten verlegt werden, aber auch Datumsangaben (Datum Schlaganfall, Aufnahme- und Entlassdaten, Geburtsdatum) und

[2] Erlaubte Abweichung lediglich 1 Tag.

Tabelle 6: Rehabilitation: Füllgrade und Entropie verfügbarer/verwendeter Merkmale

Merkmale	% Missing	Entropie
Aufnahmedatum – Jahr	0.0	2.4
Aufnahmedatum – Monat	0.0	3.6
Aufnahmedatum – Tag	0.0	4.9
Schlaganfalldatum – Jahr	4.7	2.5
Schlaganfalldatum – Monat	4.7	3.6
Schlaganfalldatum – Tag	4.7	5.0
Entlassdatum Akut – Jahr	68.2	1.0
Entlassdatum Akut – Monat	68.2	3.8
Entlassdatum Akut – Tag	68.2	4.9
Geburtsjahr	0.9	5.5
Geburtsmonat	0.9	3.6
Geburtstag	0.9	4.9
Geschlecht	1.1	1.0
ICD-Code	1.4	1.0
Initial Vorname	2.0	4.1
Initial Nachname	2.0	4.1
PLZ5	–	–
aphasie_a	68.5	1.3
ba_06_a	0.7	1.4
ba_08_a	0.7	1.4
ba_10_a	0.7	1.9
ra_aufn	0.1	2.3
KHNr zuweisende Einrichtung	8.1	4.4

Datensatz und Merkmale aufbereitet, 16659 Zeilen

Zahlen enthalten alle berücksichtigten Jahre; Entropie daher eher überschätzt

Namensangaben (z.B. verdrehte Initialen). Ein weiteres Problem stellen potentiell verfügbare Merkmale dar, die aber nur eine geringe Differenzierung erlauben, z.B. weil sie nur wenige Ausprägungen haben oder sich die Verteilung auf einzelne wenige Ausprägung konzentrieren, d.h. bei denen die Entropie insgesamt gering ist. Zudem bereitet Probleme, dass nicht alle Merkmale in den dann jeweils zu verknüpfenden Datensätzen verfügbar sind (das betrifft z.B. die Nummern der Krankenhäuser in die verlegt wird oder aus denen Patienten verlegt werden oder die PLZ der Wohnort der Patienten), so dass für verschiedene Verknüpfungen auf unterschiedliche Sets von Merkmalen zurückgegriffen werden muss (d.h. die Qualität der Verknüpfung ist nicht notwendigerweise vergleichbar über verschiedene Verknüpfungen oder Datensätze hinweg). Tabellen 1, 2, 4 und 6 zeigen eine Übersicht über die verfügbaren potentiellen Block- und Matchvariablen. Deutlich zeigt sich, dass das Set geeigneter Merkmale beschränkt ist: Entweder ist die Differenzierungskraft gering (niedrige Entropie) und/oder der Anteil fehlender Werte ist hoch. Einige Merkmale eigneten sich sehr gut als Matchvariable (insbesondere PLZ oder Krankenhausnummern), liegen dann aber nicht auf *beiden* Seiten in den Datensätzen vor. Andere Variablen eignen sich gut als Blockvariable, weil sie als vergleichsweise fehlerfrei angenommen werden können und vergleichsweise vollständig in beiden Datensätze sind; zudem müssen sie als stabil zwischen den verschiedenen Zeitpunkten angenommen werden. Insbesondere die „medizinischen Merkmale“, d.h. verschiedene Indizes, die bei Aufnahme oder Entlassung erfasst werden eignen sich aber kaum als Block- oder Matchvariable: zwar liegen sie jeweils in den zu verknüpfenden Datensätzen vor (und auch mit jeweils hohem Füllgrad), allerdings ist die Datenqualität unklar, insbesondere, wie stabil diese Merkmale zwischen den beiden Zeitpunkten sind; als Matchvariable eignen sie sich wegen ihrer geringen Differenzierungskraft (deutlich durch die geringe Entropie) kaum.[3]

Bei einigen Datumsangaben kommen (selten) offensichtlich Übertragungs- oder Tippfehler vor (Jahr „3001“ anstatt vermutlich „2001“ o.ä.), die dann entsprechend korrigiert wurden. Durch diese Korrekturen kann es v.a. zu falsch-positiven (und weniger zu falsch negativen) Verknüpfungen kommen, wenn z.B. nicht nur ein Tippfehler sondern mehrere Tippfehler gleichzeitig auftreten, d.h. z.B. nicht nur das Jahrtausend fehlerhaft erfasst ist, sondern auch das Jahrzehnt oder Jahr selbst („3001“ anstatt z.B. korrekt „2002“ oder „2010“).

Insgesamt ist eine Bewertung der Links, d.h. der potentiell zusammengehörigen Paare, sehr schwierig; (weniger) die Anzahl und (mehr) die Art und Qualität der Matchvariablen erlaubt keinen fehlertoleranten Abgleich unter Zuhilfenahme von Ähnlichkeitsfunktionen: bei Unigrammen wie den Initialen ist ein Einsatz von Ähnlichkeitsfunktionen kaum sinnvoll, ebenfalls bei den Datumsangaben; hier sind allenfalls „Dreher“ (Vorname/Nachname – Nachname/Vorname bzw. Tag/Monat – Monat/Tag) in einigen Fällen als plausibel annehmbar und regelbasiert zu berücksichtigen. Die Entscheidung beim manuellen Review, ob ein Link tatsächlich auch ein Match darstellt, ist daher allein regelbasiert und stützt sich darauf, ob Übereinstimmungen bzw. Abweichungen von Merkmalen ein Match noch als “plausibel“ erscheinen lassen.

[3] Das Merkmal „Geschlecht“ weist zwar auch eine niedrige Entropie auf, ist aber jeweils vollständig(er) und vor allem als stabiler anzunehmen und eignet sich daher besser als Block- oder Matchvariable.

3 Record-Linkage

Verknüpft werden die Datensätze jeweils deterministisch, d.h. für jede zu vergleichende Datenzeile wird als Matchkriterium die Summe S berechnet, indem für jede Datenzeile i die Anzahl der exakten Übereinstimmungen über alle Merkmale y^k gezählt wird.

$$S_i = \sum_{j=1}^{k} x_i \text{ mit } x_i = \begin{cases} 1, & y_a^k = y_b^k \\ 0, & y_a^k \neq y_b^k \end{cases}$$

Abgleiche erfolgen alle exakt, d.h. ohne weitere Ähnlichkeitsfunktionen.[4] Durch das begrenzte Set an verfügbaren und geeigneten Block- und Matchvariablen ist ein aufwändiges manuelles Review unerlässlich, um auch Datenzeilen verknüpfen zu können, bei denen es keine exakte Übereinstimmung von Merkmalen gibt (z.B. Dreher in den Namensinitialen oder im Geburtsdatum, Abweichungen um einen oder wenige Tage in verschiedenen Datumsangaben). Diese Regeln umfassen bspw. Entscheidungen zu gedrehten Namensinitialen (z.B. $\texttt{Vorname}_{(A)}$ == $\texttt{Nachname}_{(B)}$ *und* $\texttt{Nachname}_{(A)}$==$\texttt{Vorname}_{(B)}$) oder Datumsangaben (z.B $\texttt{Tag}_{(A)}$==$\texttt{Monat}_{(B)}$ *und* $\texttt{Monat}_{(A)}$==$\texttt{Tag}_{(B)}$ oder auch Drehern in einzelnen Datumsangaben, z.B. dem Tag), dem Umgang mit fehlenden Werten bei Matchvariablen.

Alle Datensätze enthalten Datumsangaben (insbesondere natürlich das Datum der jeweiligen Aufnahme bzw. Entlassung). Z.T. sind in den verschiedenen Datenquellen Datumsangaben derselben Ereignisse erfasst. Zumindest lassen sich oft erfasste Datumsangaben verschiedener Ereignisse in eine Abfolge bringen und wenn schon nicht Datumsangaben aus verschiedenen Datensätzen über dasselbe Ereignis miteinander vergleichen, so doch die jeweils erfassten Ereignisse in eine Abfolge bringen und plausible oder ungültige Differenzen zwischen verschiedenen Datumsangaben beim manuellen Review und der Entscheidung, ob ein Link ein Match oder Nonmatch ist berücksichtigen.

Diese zeitlichen Bezüge sind zu berücksichtigen, wenn Links im manuellen Review als Match oder Nonmatch klassifiziert werden: beim exakten Abgleich der Datumsangaben werden viele Datumskomponenten daher nicht exakt übereinstimmen, obwohl es sich um einen potentiellen Match handelt, weil zumindest die Abfolge der Ereignisse bzw. Datumsangaben korrekt und plausibel ist; dieser Umstand macht das manuelle Review einerseits aufwendig und unverzichtbar, erlaubt aber andererseits die Formulierung und Anwendung einiger Regeln, die die (Un-)Gültigkeit verschiedener zeitlicher Bezüge berücksichtigen. Aus dem Umstand, dass –per Design– einige der Matchvariablen nicht übereinstimmen (müssen) folgt, dass viele “gute” Matches nicht das höchste mögliche Matchgewicht (die maximale Summe aller möglichen –exakten– Übereinstimmungen) erreichen werden: z.B. muss nicht notwendigerweise davon ausgegangen werden, dass bei allen Matches Entlassdatum aus der Akutklinik und Aufnahme in Frührehaklinik (exakt) übereinstimmen, zu-

[4] Das liegt darin begründet, dass die Verwendung von Ähnlichkeitsfunktionen bei den verfügbaren und verwendeten Merkmalen wenig sinnvoll ist; die für das Record-Linkage verwendete Software MTB beinhaltete hingegen eine Reihe verschiedener Ähnlichkeitsfunktionen. „Fehlertoleranz“ und eine Verknüpfung bei nicht exakter Übereinstimmungen von Merkmalen ist hier allein durch das anschließende regelbasierte manuelle Review der Links möglich.

Abbildung 1: zeitliche Bezüge zwischen Datumsangaben der verschiedenen Datenquellen

```
Datenquellen:
=============
A) Akutdatensätze 2007/2008 bzw. 2009/2010
B) Frührehadatensatz
C) Rehadatensatz


zeitliche Abfolge der Ereignisse:
=================================
                  (1)        (2)           (3)           (4)              (5)              (6)
                             Aufnahme      Entlassung    Aufnahme         Entlassung       Aufnahme
Ereignis        Anfall     Akutklinik    Akutklinik    Frührehaklinik   Frührehaklinik   Rehaklinik
               ---|-----------|--------------|---------------|-----------------|----------------|--------- ...
enthalten in:   B) C)          A)            A) C)            B)                B)               C)


zeitliche Bezüge:
=================
1B == 1C (Anfallsdatum in beiden Datensätzen gleich)
3A == 3C (Entlassdatum in beiden Datensätzen gleich)

1B <= 2A (Anfallsdatum VOR Aufnahme in Akutklinik)
1C < =2A (Anfallsdatum VOR Aufnahme in Akutklinik)

3A <= 4B (Entlassung aus Akutklinik VOR Aufnahme in Frühreha)
3C <= 4B (Entlassung aus Akutklinik VOR Aufnahme in Frühreha)

3A <= 6C (Entlassung aus Akutklinik VOR Aufnahme in Reha)
5B <= 6C (Entlassung aus Frühreha VOR Aufnahme in Reha)
```

mindest muss die Entlassung aus der Akutklinik aber *vor* der Aufnahme in die Frührehabilitationsklinik liegen, d.h. die Datumskomponente „Tag“ muss nicht notwendigerweise übereinstimmen, noch nicht einmal „Monat“ oder sogar „Jahr“. Z.B. bei Entlassung vor Jahresende und Aufnahme in eine Rehabilitationsklinik am Anfang des nächsten Jahres gäbe es weder beim Tag, noch beim Monat oder Jahr eine Übereinstimmung, und da alle vier Datumsangaben (vollständiges Datum, Tag, Monat und Jahr getrennt) als Matchvariable verwendet werden, kann als Matchgewicht schon nur noch $S = S_{max} - 4$ erreicht werden. Die Trennung der Datumsangaben und die Verwendung aller Komponenten als Matchvariable erlaubt dafür eine gewisse Fehlertoleranz für Datumsangaben: selbst wenn eine Datumsangabe nicht vollständig übereinstimmt, sondern z.B um einen Tag abweicht, tragen die verbliebenen – dann übereinstimmenden – Datumskomponenten zu einem höheren Matchgewicht bei. Das erlaubt eine deutlichere Trennung von „ähnlichen“ und (sehr) „unähnlichen“ Datumsangaben.

4 Ergebnisse

Dieser Abschnitt beschreibt die jeweils für die verschiedenen Verknüpfungen verwendeten Block- und Matchvariablen und Ergebnisse der Verknüpfung. Als Blockvariablen werden solche Merkmale gewählt, die in den zu verknüpfenden Datensätzen möglichst vollständig und fehlerfrei vorliegen.[5] Nur Datenzeilen, bei denen Ausprägungen der Blockvariable(n)

[5] Blockvariablen müssen zumindest fehlerfrei sein; Beobachtungen mit fehlenden Werten bei den Blockvariablen können – je nach Einstellung – Teil jedes Blocks sein (oder alternativ einen eigenen Block bilden); das ermöglicht die Verknüpfung auch bei fehlenden Werten auf einer oder mehreren Blockvariablen, verringert

Tabelle 7: Akutpatienten 2007/2008 – Frührehabilitation: Anzahl Matches nach Gesamtgewicht

Anzahl Übereinstimmungen	Anzahl Matches	in %
11	2	0.7
13	8	2.6
14	20	6.6
15	13	4.3
16	38	12.5
17	30	9.9
18	118	38.9
19	74	24.4
Matches insgesamt	303	100.0

übereinstimmen werden miteinander verglichen und diese Links einem manuellen Review unterzogen.

4.1 Akutpatienten 2007/2008 – Frührehabilitation

Als Blockvariablen wurden die 3-stellige Postleitzahl sowie das Geschlecht verwendet. Das Merkmal `Geschlecht` zeichnet sich durch einen hohen Füllgrad (und vermutlich auch hohe Datenqualität) aus; die Postleitzahl weist zwar einen niedrigen Füllgrad auf (daher sind Beobachtungen mit fehlenden Werten in allen Blocks erhalten), durch die Verwendung der 3-stelligen Postleitzahl als zusätzliche Blockvariable kann einerseits die Anzahl der durchzuführenden Vergleiche deutlich reduziert werden, durch das Abschneiden der letzten beiden Ziffern kann auch eine gewisse Fehlertoleranz erreicht werden.

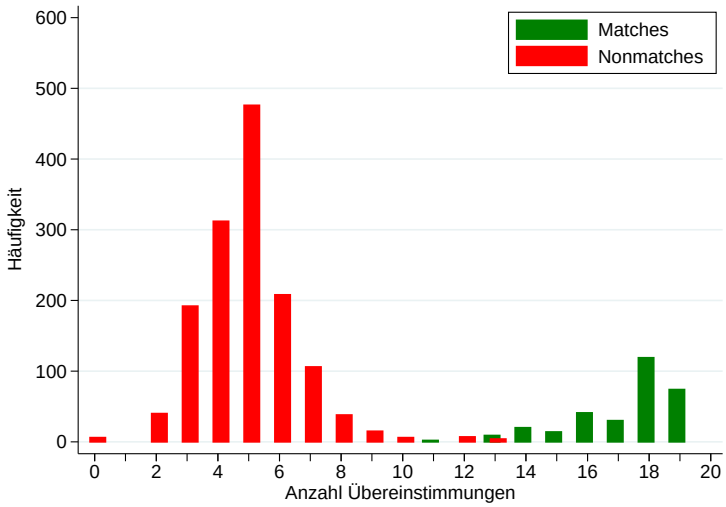
Als Matchvariable wurden verwendet: die Initialen, Initial Vorname, Initial Nachname, 5-stellige PLZ, 4-stellige PLZ, Geburtsdatum, Geburtsdatum (Jahr, Monat, Tag einzeln), Entlassdatum, Entlassdatum (Jahr, Monat, Tag einzeln), Aufnahmedatum, Aufnahmedatum (Jahr, Monat, Tag einzeln), ICD-Code, Krankenhausnummer (Patient wird verlegt nach ...). Damit ist eine maximale Summe von $S = 19$ bei vollständiger Übereinstimmung aller Merkmale möglich.

Insgesamt 1919 Paare liegen innerhalb der Blocks, nach manuellem Review aller Paare und regelbasierter Entscheidung Match–Nonmatch verbleiben 303 Matches (das entspräche einer Matchrate von 15.8% bei unbekannter wahrer Überlappung der Datensätze, die aber deutlich unter 100% liegen dürfte).

Dass selbst bei „nur“ 15 oder weniger Übereinstimmungen (von maximal hier 19 Übereinstimmungen) Matches möglich sind ist v.a. auf den Umstand zurückzuführen, dass Matchvariablen fehlende Werte aufweisen können (insbesondere z.B. die Krankenhausnummer, in die ein Patient verlegt wird) und bspw. durch das Aufspalten der Datumsangaben und Verwendung der einzelnen Datumskomponenten als Matchvariable: fehlt z.B. eine

aber die Effizienz des Verfahrens, da eine Beobachtung mit fehlendem Wert bei einer Blockvariablen dann Teil aller durch die Ausprägungen einer Blockvariablen gebildeten Blocks ist.

Abbildung 2: Akutpatienten 2007/2008 – Frührehabilitation: Verteilung Matchgewicht (Anzahl Übereinstimmungen) Matches – Nonmatches nach manuellem Review



Datumsangabe und stimmen sonst alle Merkmale überein, ist das dann maximale Matchgewicht bereits nur noch 15, fehlt dann noch die Krankenhausnummer und die PLZ, beträgt das maximal mögliche Matchgewicht bereits nur noch 13. Die regelbasierten Entscheidungen Match-Nonmatch berücksichtigen nun die verschiedenen möglichen Kombinationen von fehlenden Werten bei den Matchvariablen und berücksichtigen nur in wenigen Fällen tatsächlich „ähnliche“ Werte, z.B. wenn Datumsangaben um einen Tag abweichen (oder um genau einen Monat, bei sonst aber übereinstimmenden Merkmalen, etc.).

Abbildung 2 zeigt die Verteilung der Matchgewichte (Anzahl Übereinstimmung) der Links nach manuellem Review und regelbasierter Entscheidung, ob es sich bei einem Link um ein Match- oder Nonmatch handelt. Dabei zeigt sich ein für Record-Linkage-Anwendungen typisches Bild: u-förmige Verteilung der Matchgewichte, wobei bei hohen Matchgewichten (fast) ausschließlich Matches, bei niedrigen Matchgewichten (fast) ausschließlich Nonmatches, beide Kurven gehen ineinander über.

4.2 Akutpatienten 2009/2010 – Frührehabilitation

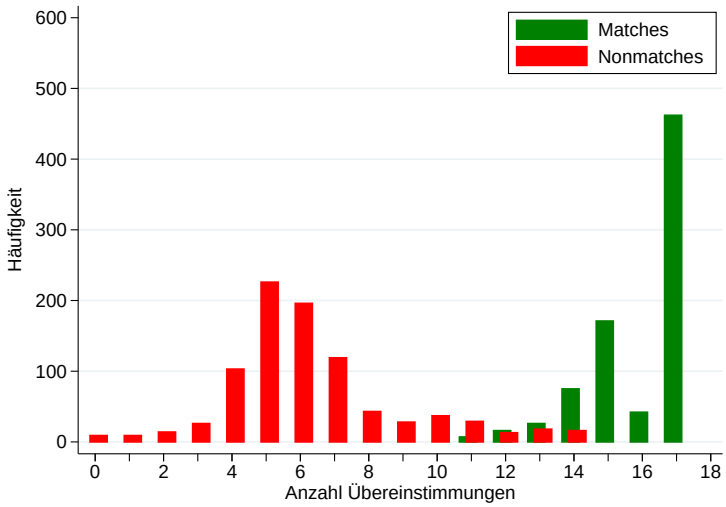
Die Verknüpfung der Akutpatienten 2009/2010 mit den Rehabilitationspatienten verläuft analog. Lediglich kleinere Unterschiede ergeben sich dadurch, dass eine Matchvariable (Krankenhausnummer, *in die* ein Patient nach der Akutbehandlung verlegt wird) nicht im Datensatz der Akutpatienten enthalten ist: das Geschlecht und die 3-stellige Postleitzahl bilden Blöcke, innerhalb derer Merkmale verglichen werden; als Matchvariable wurden verwendet: die Initialen, Initial Vorname, Initial Nachname, 5-stellige PLZ, 4-stellige PLZ, Geburtsdatum, Geburtsdatum (Jahr, Monat, Tag einzeln), Entlassdatum, Entlassdatum (Jahr, Monat, Tag einzeln), Aufnahmedatum, Aufnahmedatum (Jahr, Monat, Tag einzeln), ICD-Code. Damit ist eine maximale Summe von $S = 18$ bei vollständiger Übereinstimmung aller Merkmale möglich.

1685 Paare (Links) liegen innerhalb der Blöcke (Geschlecht und 3-stellige PLZ), nach einem manuellen und regelbasierten Review der Links lassen sich davon 702 Matches bil-

Tabelle 8: Akutpatienten 2009/2010 – Frührehabilitation: Anzahl Matches nach Gesamtgewicht

Anzahl Übereinstimmungen	Anzahl Matches	in %
11	7	1.0
12	13	1.9
13	21	3.0
14	61	8.7
15	149	21.2
16	36	5.1
17	415	59.1
Insgesamt	702	100.0

Abbildung 3: Akutpatienten der Kohorte 2009/2010 – Frührehabilitation: Verteilung Matchgewicht (Anzahl Übereinstimmungen) Matches – Nonmatches nach manuellem Review



den. Bei 1905 Zeilen aus dem Datensatz der Rehabilitationspatienten mit verzeichnetem Datum des Schlaganfalls im Jahr 2009 oder 2010 (oder fehlendem Datum) entspräche das einer Matchrate von 36.9% (wiederum bei unbekannter wahrer Überlappung der beiden Datensätze, die aber wieder deutlich unter 100% liegen dürfte).

Abbildung 3 zeigt wieder die Verteilung der Matchgewicht nach dem manuellen Sichten und regelbasierter Entscheidung Match – Nonmatch. auch hier zeigt sich wieder das typische u-förmige Bild der Verteilung der Matchgewicht.

4.3 Akutpatienten 2007/2008 – Rehabilitation

Für die Verknüpfung der Akutpatienten mit den Patienten aus dem Rehabilitationsdatensatz werden Blöcke nach Geschlecht, Geburtsjahr und dem Jahr des Schlaganfalls (bzw. Aufnahme in Akutklinik)[6] gebildet. Innerhalb der so gebildeten Blöcke werden Beobachtungen verglichen nach Initialen, Initial Vorname, Initial Nachname, Geburtsdatum, Geburts-

[6] Beobachtungen mit fehlenden Werten auf einer Blockvariable bilden wieder keinen eigenen Block, sondern sind Teil aller Blöcke.

datum (Monat und Tag), Entlassdatum (Jahr, Monat und Tag), Rankin-Skalen-Index bei Aufnahme/Entlassung, Krankenhausnummer der verlegenden Klinik, Krankenhausnummer der zuweisenden Klinik, ICD-Code, Anfallsdatum bzw. Aufnahmedatum, Anfallsdatum bzw. Aufnahmedatum (Jahr, Monat und Tag). Daraus ergibt sich ein maximales Matchgewicht von $S = 18$ bei exakter Übereinstimmung aller Matchvariablen.

Als Grundlage der Verknüpfung dienen 39606 Patienten des Akutdatensatzes, die zwischen 2007 und 2009 aus der Akutklinik entlassen wurden, und 12751 Rehabilitationspatienten mit verzeichnetem Entlassdatum zwischen 2007 und 2009 (oder fehlenden Entlassdatum). Von diesen 12571 Patienten können 2621 Patienten mit Einträgen im Akutdatensatz verknüpft werden, das entspräche einer Matchrate von 20.6%. Auch hier muss aber wieder berücksichtigt werden, dass die Überlappung der beiden Datensätze unbekannt ist und nicht alle 12751 Rehabilitationspatienten im Datensatz der akuten Schlaganfallpatienten enthalten sind; d.h. auch hier stellt eine Rate von 20.6% eine Untergrenze der wahren Matchrate dar.

Beim manuellen Review der Links können hier nicht nur exakte Übereinstimmungen zwischen Merkmalen einer Entscheidung Match–Nonmatch zugrunde gelegt werden, sondern da auch mehrere in Zusammenhang stehende Datumsangaben in den beiden Datensätzen enthalten sind, auch nicht übereinstimmende Datumsangaben besser Berücksichtigung finden, so dass auch bei insgesamt niedrigeren Gesamtgewichten (durch fehlende exakte Übereinstimmung) noch Matches identifiziert werden, z.B. können Abweichungen zwischen Datumsangaben berechnet und zur Beurteilung der Links und Einschätzung Match–Nonmatch herangezogen werden:

- Datum Schlaganfall (Reha) – Aufnahmedatum (Akut): Differenz klein (0, 1, ... Tag); Schlaganfall *vor* Aufnahme in Klinik
- Entlassdatum (Akut) – Entlassdatum Akutklinik (Reha): keine Differenz, Abweichung um 1 Tag/wenige Tage;
- Entlassdatum (Akut) – Aufnahmedatum (Reha): kleine Differenz, maximale Differenz (180 Tage); Entlassung aus Akutklinik i.d.R. *vor* Aufnahme in Rehaklinik

Dadurch lassen sich auch bei niedrigem Gesamtgewicht (Summe der *exakten* Übereinstimmung noch Matches identifizieren (z.B. reduziert sich das Matchgewicht bei Abweichung von zwei verschiedenen Datumsangaben am Ende/zu Beginn eines Monats auch nur um einen Tag bereits um 6 Punkte, fehlen dann noch noch die beiden Krankenhausnummern und der ICD-Code in einem der beiden Datensatz beträgt das Matchgewicht bei sonst exakten Übereinstimmungen bereits nur noch 9.

Das zeigt sich auch in der Verteilung der Matchgewichte: die beiden Kurven überlappen sich doch vergleichsweise weit, auch bei niedrigen Matchgewichten lassen sich Links noch als plausible Matches identifizieren. Andererseits: auch wenn bei einem Link mehrere der Datumsangaben (zufällig) vollständig übereinstimmen, handelt es sich eben nicht notwendigerweise auch um ein plausiblen Match, das Matchgewicht ist aber vergleichsweise hoch, so dass auch bei hohen Matchgewichten Links als Nonmatches klassifiziert wer-

Tabelle 9: Akutpatienten der Kohorte 2007/2008 – Rehabilitation: Anzahl Matches nach Gesamtgewicht

Anzahl Übereinstimmungen	Anzahl Matches	in %
6	3	0.1
7	5	0.2
8	26	1.0
9	79	3.0
10	229	8.7
11	291	11.1
12	867	33.1
13	452	17.3
14	205	7.8
15	78	3.0
16	201	7.7
17	118	4.5
18	67	2.6
Insgesamt	2621	100.0

den. Beim regelbasierten Review muss diesem Umstand Rechnung getragen werden, d.h. hier wiegen Übereinstimmungen bzw. Abweichungen bei anderen Identifikatoren stärker.

4.4 Akutpatienten 2009/2010– Rehabilitation

Die Verknüpfung der beiden Datensätze Akutpatienten 2009/2010 – Rehabilitationspatienten entspricht der oben beschrieben Verknüpfung der vorhergehenden Kohorte: Geschlecht, Geburtsjahr und Jahr des Schlaganfalls bilden die Blocks,[7] innerhalb der Blocks werden die Merkmale Initialen, Vorname, Nachname, Geburtsdatum, Geburtsdatum (Jahr, Monat, Tag), Entlassdatum aus Akutklinik, Entlassdatum (Jahr, Monat, Tag), Rankin-Index bei Entlassung/Aufnahme, ICD-Code, sowie die Krankenhausnummer der Akutklinik mit der Nummer der Akutklinik miteinander verglichen und exakte Übereinstimmungen gezählt; das maximale Matchgewicht bei exakter Übereinstimmung beträgt damit 11. Auch hier können nicht nur aus dem exakten Vergleich verschiedener Datumsangaben sondern auch der Differenz zwischen zusammenhängenden Datumsangaben zusätzliche Hinweise für eine Entscheidung Match-Nonmatch gewonnen werden, d.h. auch hier können noch Links als Matches klassifiziert werden, obwohl das Matchgewicht der exakten, deterministischen Verknüpfung niedrig ist.

46038 Patienten der Schlaganfallkohorte 2009/2010 haben ein verzeichnetes Entlassdatum in den Jahren 2009, 2010 oder 2011; im Rehabilitationsdatensatz sind 6745 Patienten mit fehlendem Entlassdatum der Akutklinik oder Entlassung in den Jahren 2009, 2010 verzeichnet. Von diesen 60745 Patienten aus dem Rehabilitationsdatensatz können 45.0% (3083) in der Schlaganfallkohorte 2009/2010 gefunden werden.

[7] Implizit bilden alle Schlaganfallpatienten, die 2009 oder später aus der Akutklinik entlassen werden (bzw.mit unbekanntem Entlassungsjahr) auch einen Block.

Abbildung 4: Akutpatienten der Kohorte 2007/2008– Rehabilitation: Verteilung Matchgewicht (Anzahl Übereinstimmungen) Matches – Nonmatches nach manuellem Review

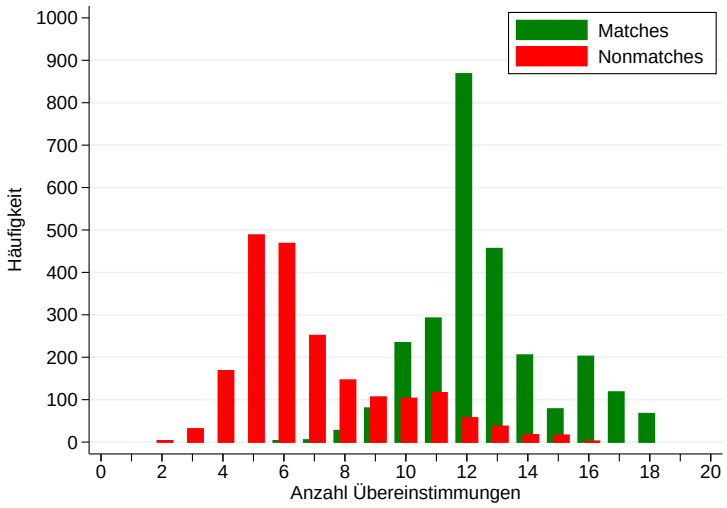


Tabelle 10: Akutpatienten 2009/2010 – Rehabilitation: Anzahl Matches nach Gesamtgewicht

Anzahl Übereinstimmungen	Anzahl Matches	in %
4	1	0.0
5	32	1.1
6	116	3.8
7	1700	56.0
8	1011	33.3
9	161	5.3
10	15	0.5
11	2	0.1
Insgesamt	3038	100.0

Abbildung 5: Schlaganfallkohorte 2009/2010 – Rehabilitation: Verteilung Matchgewicht (Anzahl Übereinstimmungen) Matches – Nonmatches nach manuellem Review

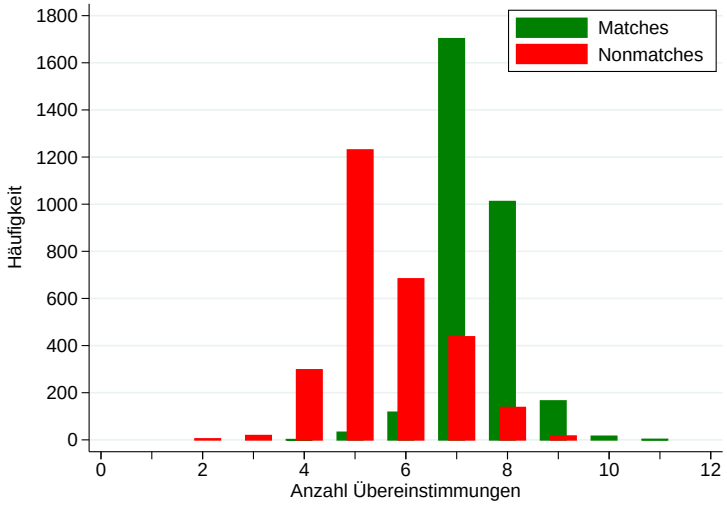


Tabelle 11: Frühreha – Rehabilitation: Anzahl Matches nach Gesamtgewicht

Anzahl Übereinstimmungen	Anzahl Matches	in %
11	9	3.2
12	2	0.7
13	18	6.5
14	44	15.8
15	22	7.9
16	151	54.3
17	14	5.0
18	10	3.6
20	8	2.9
Insgesamt	278	100.0

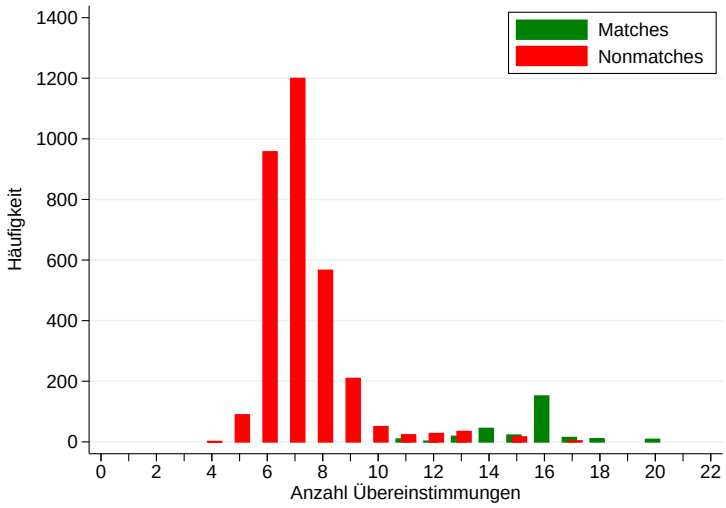
4.5 Frührehabilitation – Rehabilitation

Zuletzt werden die beiden Datensätze der Rehabilitationsstufen miteinander verknüpft: 3452 Zeilen aus dem Datensatz der Frührehabilitationspatienten und 16659 Zeilen im Datensatz der Rehapatienten.

Als Blockvariable wird wieder Geschlecht und das in beiden Datensätze vorhandene Jahr des Schlaganfalls verwendet. Als Matchvariablen wurden verwendet die Initialen, sowie einzeln Vor- und Nachname (3), das Geburtsdatum (4), das Datum des Schlaganfalls (4), das Aufnahmedatum in die Frührehabilitation bzw. das Datum der Entlassung aus der Akutklinik (4), das Datum der Entlassung aus der Frührehabilitation bzw. das Aufnahmedatum zur Rehabilitation (4), schließlich der ICD-Code bei Entlassung bzw. Aufnahme (1) und die Krankenhausnummer der Frührehabilitationsklinik bzw. der zuweisenden Klinik (1). Das maximale Gesamtgewicht S ergibt 21, wobei hier im besonderen gilt, dass erwartbar ist, dass auch bei Matches viele der Matchvariablen nicht exakt übereinstimmen, sondern lediglich eine korrekte Abfolge verschiedener Ereignisse abbilden: während das Datum des Schlaganfalls in beiden Datensätzen zwar übereinstimmen sollte, stimmen Entlassdatum aus der Akutklinik und Aufnahmedatum in der Frührehabilitation nicht notwendigerweise miteinander überein, ebenso wie das Entlassdatum aus der Frührehabilitation und das Aufnahmedatum in einer Rehaklinik. Damit verringerte sich die Summe der Übereinstimmungen entsprechend bereits um bis zu 6 Punkte, weitere Abweichungen bei weiteren Matchvariablen auch für Matches sind plausibel, so dass das Matchgewicht auch für Matches noch weiter sinken kann.

Nach manuellem Review aller Links, bei denen zumindest Geschlecht und Jahr des Schlaganfalls übereinstimmen (d.h. innerhalb der Blocks) und regelbasierter Entscheidung Match–Nonmatch konnten schließlich 278 Zeilen miteinander verknüpft werden. Bei 3452 (wie oben bereits beschrieben aufbereiteten) Frührehapatienten entspräche das einer Matchrate von 8.1%. Bei 16659 Zeilen des Rehadatensatzes (wie oben beschrieben aufbereitet) entspräche das einer Matchrate von 1.7% der Rehapatienten. Auch her gilt zu berücksichtigen, dass die wahre Überlappung der Datenätze nicht bekannt ist (und kaum 100% betragen dürfte) und zudem der Rehadatensatz weiter zurückreicht, d.h. auch viele Pa-

Abbildung 6: Frührehabilitation – Rehabilitation: Verteilung Matchgewicht (Anzahl Übereinstimmungen) Matches – Nonmatches nach manuellem Review



tienten mit deutlich länger zurückliegenden Schlaganfällen enthält als der Datensatz der Frührehaptienten.

5 Zusammenfassung

Dieser kurze Bericht beschreibt die Verknüpfung verschiedener Behandlungsstufen (akut – Frührehabilitation – Rehabilitation) zweier Schlaganfallkohorten, d.h. insgesamt fünf Verknüpfungen. Die Vorraussetzungen einer erfolgreichen und „guten“ (i.S. einer hohen Matchrate) Verknüpfung sind schlecht. Zum einen ist die Anzahl der Merkmale begrenzt, die als Block- oder Matchvariable verwendet werden können, zum anderen ist die Eignung der prinzipiell verfügbaren Merkmale als Block- oder Matchvariable schlecht (geringe Entropie durch wenige Ausprägungen oder starke Konzentration auf bestimmte Ausprägungen, unklare Datenqualität insbesondere Stabilität zwischen verschiedenen Zeitpunkten, z.T. geringer Füllgrad bei bestimmten Merkmalen). Als nachteilig wirkt sich auch aus, dass die verfügbaren Merkmale kaum fehlertoleranten Abgleich ermöglichen (z.B. durch den Einsatz von Ähnlichkeitsfunktionen), d.h. alle Abgleich müssen exakt erfolgen, erst durch ein intensives und aufwendiges manuelles Review (und entsprechend Regelformulierung) der verknüpften Paare können auch nicht exakte Übereinstimmungen als Match klassifiziert werden.

Eine Evaluation der Ergebnisse und Angabe oder Abschätzung richtig-positiver, falsch-positiver und falsch-negativer und richtig-negativer Zuordnungen ist nicht möglich. Tatsächlich sind sowohl falsch-negative Verknüpfungen möglich (d.h. zu wenige Matches), aber insgesamt weniger wahrscheinlich als falsch-positive Verknüpfungen (d.h. zu viele, falsche Zuordnungen zwischen den Datensätzen). Für eine Verbesserung der Verknüpfungsqualität würden dringend – nicht unbedingt mehr aber vollständigere und bessere Identifikatoren benötigt; so könnte der Füllgrad der Postleitzahlen und der Krankenhausnummern erhöht werden und womöglich könnten mehr Buchstaben des Vor- und Nachnamens verwendet werden, um einmal eine höhere Differenzierung als alleine durch den ersten Buchstaben

zu erreichen oder um Ähnlichkeiten oder häufigkeitsgewichtete Übereinstimmungen bestimmen zu können.

IMPRINT

Publisher

German Record-Linkage Center
Regensburger Str. 104
D-90478 Nuremberg

Editors

Stefan Bender, Rainer Schnell

Template layout

Christine Weidmann

www.record-linkage.de