Thieme

Quo vadis Datenlinkage in Deutschland? Eine erste Bestandsaufnahme

Quo Vadis Data Linkage in Germany? An Initial Inventory

Autoren
Stefanie March[1], Manfred Antoni[2], Joachim Kieschke[3], Bianca Kollhorst[4], Birga Maier[5], Gabriele Müller[6], Murat Sariyar[7, 8], Mandy Schulz[9], Swart Enno[1], Jan Zeidler[10], Falk Hoffmann[11]

Institute
1 Medizinische Fakultät, Institut für Sozialmedizin und Gesundheitsökonomie, Otto-von-Guericke-Universität Magdeburg, Magdeburg
2 Institut für Arbeitsmarkt- und Berufsforschung der Bundesagentur für Arbeit (IAB), Nürnberg
3 Registerstelle, Epidemiologisches Krebsregister Niedersachsen, Oldenburg
4 Abteilung Biometrie und EDV, Leibniz-Institut für Präventionsforschung und Epidemiologie – BIPS, Bremen
5 Berlin-Brandenburger Herzinfarktregister e.V., Berlin
6 Universitätsklinikum und Medizinische Fakultät Carl Gustav Carus, Zentrum für Evidenzbasierte Gesundheitsversorgung (ZEGV), TU Dresden, Dresden
7 TMF – Technologie- und Methodenplattform für die vernetzte medizinische Forschung e.V., Berlin
8 University of Applied Sciences Bern, Department of Medical Informatics, Bern
9 Zentralinstitut für die kassenärztliche Versorgung in Deutschland (Zi), Fachbereich Versorgungsforschung und Risikostruktur, Berlin
10 Center for Health Economics Research Hannover (CHERH), Leibniz Universität Hannover, Hannover
11 Department für Versorgungsforschung, Carl von Ossietzky Universität Oldenburg, Oldenburg

Schlüsselwörter
Datenlinkage, Record Linkage, Primärdaten, Routinedaten, administrative Daten

Key words
data linkage, record linkage, primary data, routine data, administrative data

Bibliografie
DOI https://doi.org/10.1055/s-0043-125070
Online-Publikation: 20.2.2018
Gesundheitswesen 2018; 80: e20–e31

ISSN 0941-3790

Korrespondenzadresse
Dr. Stefanie March, MA
Medizinische Fakultät
Institut für Sozialmedizin und Gesundheitsökonomie
Otto-von-Guericke-Universität Magdeburg
Leipziger Straße 44
39120 Magdeburg
stefanie.march@med.ovgu.de

ZUSAMMENFASSUNG

Die Verknüpfung verschiedener Datenquellen, genannt Datenlinkage oder auch Record Linkage, zur Beantwortung von wissenschaftlichen Fragestellungen findet in den letzten Jahren in Deutschland vermehrt Anwendung. Jedoch mangelt es bisher an publizierten Erfahrungen. Neue Projekte erarbeiten sich in der Regel autark voneinander das notwendige Handwerkszeug. Daher hat sich eine Gruppe von Forschern zusammengefunden, um ihre Erfahrungen zum Datenlinkage in Deutschland als mögliche Hilfestellung bzw. Anregung für Projekte, Gutachter sowie Datenschützer und Ethikkommissionen zusammenzustellen. Ziel dieser ersten Bestandsaufnahme zum Datenlinkage ist es deshalb, eine Unterstützung für zukünftige Projekte zu liefern, die Daten aus Deutschland auf individueller Ebene verknüpfen möchten. Neben den (datenschutz-)rechtlichen Rahmenbedingungen werden dabei auch praxisorientiert die Arten des Datenlinkage, deren Anwendungsfelder und Ansätze zur Vermeidung von Fehlern anhand von Beispielen dargestellt.

ABSTRACT

In recent years, linking different data sources, also called data linkage or record linkage, to address scientific questions, is being increasingly used in Germany. However, there are very few published reports and new projects develop the necessary tools independently of each other. Therefore, a team of researchers joined together to exchange their experiences on data linkage and to give suggestions on how linkage could be done for scientists, reviewers as well as members of data privacy boards and ethics committees. It is the aim of this article to assist future projects that want to link German data on an individual level. In addition to the legal framework conditions (data privacy), also examples of types of data linkage, their fields of application und potential pitfalls as well as the methods of preventing them will be described in an application-oriented fashion.

Einleitung

Für die Gesundheitsforschung werden in Deutschland mittlerweile zahlreiche und teils sehr unterschiedliche Daten verwendet [1]. Diese lassen sich grundsätzlich in Primär- und Sekundärdaten unterscheiden [2]. Primärdaten[1] werden im Rahmen ihres originär vorgesehenen Verwendungszwecks aufbereitet und analysiert. Dazu zählen regionale oder bundesweite Erhebungen, wie bspw. die Gesundheitsstudien des Robert Koch-Instituts (RKI), die Daten sowohl im (wiederholten) Quer- als auch im Längsschnitt durch Befragungen und/oder medizinische Untersuchungen erfassen [3]. Auch die NAKO Gesundheitsstudie, die größte epidemiologische Langzeitstudie in Deutschland bei der etwa 200 000 Menschen über 20–30 Jahre hinweg nachbeobachtet werden sollen, zählt dazu [4]. Solche Studien werden zwar oftmals mit sehr vielen Modulen, Untersuchungen und Erhebungsinstrumenten durchgeführt, dies ist jedoch mit erheblichen Belastungen für die Teilnehmer sowie hohen Kosten und dem Risiko von Kohortenausfällen (loss-to-follow-up) verbunden. Bei Längsschnittstudien finden die Erhebungswellen deshalb oft nur in mehrjährigen Abständen statt. Dies bedeutet jedoch, dass gerade bei sehr langen auseinanderliegenden Follow-Up-Erhebungen bestimmte Faktoren nur unvollständig oder mit methodischen Schwierigkeiten erfasst werden können (z. B. Inanspruchnahme medizinischer Leistungen oder Angaben zur Erwerbsbiografie) [5].

Sekundärdaten hingegen werden einer Auswertung über ihren originären primären Verwendungszweck hinaus zugeführt. Hierzu zählen eine Vielzahl an Daten der Sozialversicherungsträger (z. B. Kranken- und Rentenversicherung), aber auch andere Leistungsdaten der gesundheitlichen Versorgung (z. B. aus Arzt- oder Krankenhausinformationssystemen) oder Daten von (klinischen) Studien, die im Nachgang für andere Fragestellungen genutzt werden [2]. Aufgrund der zunehmenden elektronischen Erfassung von Gesundheitsdaten stehen mittlerweile Sekundärdaten mit einem hohen Informationsgehalt für Forschungsvorhaben zur Verfügung und werden dafür auch zunehmend genutzt [6–8]. Allerdings ist ihr Merkmals- und Erhebungsumfang durch den ursprünglichen Erhebungsgrund determiniert und eine Nutzung für Forschungszwecke aufgrund gesetzlicher Restriktionen oder fehlender Einwilligung nicht immer ohne Probleme möglich [5, 9].

Die verschiedenen verfügbaren Daten – Primärdaten ebenso wie Sekundärdaten – weisen also allesamt spezifische Vorteile auf, jedoch auch den Nachteil, dass sie ausschließlich für bestimmte Zwecke oder Forschungsfragen generiert wurden und weitere für die Beantwortung der Fragstellung relevante Informationen fehlen, die jedoch möglicherweise in anderen Daten vorhanden sind. Eine Verknüpfung dieser verschiedenen Daten ermöglicht dann einen Informationsgewinn, durch den ein breiteres Spektrum an zusätzlichen Fragen beantwortet werden kann, ohne dafür neue Daten erheben zu müssen [10]. Aus diesen Gründen wird in vielen aktuellen Studien in/aus Deutschland eine Verknüpfung verschiedener Primär- und/oder Sekundärdaten angestrebt. Erste Erfahrungen liegen dazu mittlerweile vor [4, 9, 11–13].

Diese Verknüpfung verschiedener Datenquellen, auch als Datenlinkage (oder weitgehend synonym Record Linkage) bezeichnet, erfolgt anhand geeigneter Schlüsselvariablen (Identifikatoren), die in allen zu verknüpfenden Daten vorhanden sein müssen [14]. Da die verschiedenen in der Gesundheitsforschung verwendeten Daten jedoch in aller Regel unabhängig voneinander erhoben und oftmals pseudonymisiert vorliegen, stellt deren Verknüpfung eine erhebliche methodisch-technische sowie datenschutzrechtliche Herausforderung dar, insbesondere im Zusammenhang mit Sozialdaten. Der fehlenden Integration von Daten im Gesundheitswesen wird mittlerweile auch förderpolitisch Rechnung getragen. Das vom Bundesministerium für Bildung und Forschung (BMBF) im November 2015 initiierte Förderkonzept Medizininformatik („Daten vernetzen – Gesundheitsversorgung verbessern“)[2] oder die erste Ausschreibung des Innovationsfonds im April 2016 („Einsatz und Verknüpfung von Routinedaten zur Verbesserung der Versorgung“ als eines von 5 Themenfeldern im Bereich Versorgungsforschung)[3] greifen explizit diese Thematik auf. Dies gilt zudem auch vermehrt für Gutachten wie bspw. das Gutachten des Normenkontrollrats vom Oktober 2017 „Mehr Leistung für Bürger und Unternehmen: Verwaltung digitalisieren. Register modernisieren“[4]. Dies unterstreicht die Wichtigkeit der Verknüpfung von Daten im Gesundheitswesen. Bisher sind hauptsächlich sehr spezifische Erfahrungen zum Datenlinkage, wie im Rahmen der Krebsregistrierung, vorhanden und neue Projekte müssen sich oftmals in einem aufwändigen Prozess den entsprechenden Herausforderungen stellen. Neben den Projektbeteiligten mangelt es oft auch Gutachtern von Zeitschriftenbeiträgen oder Anträgen ebenso wie Datenschützern oder Ethikkommissionen an publizierten Erfahrungen.

In Anlehnung sowohl an die Gute Epidemiologische Praxis (GEP) [15], die Gute Praxis Sekundärdatenanalyse (GPS) [2] als auch die Berichtsstandards RECORD (Reporting of studies Conducted using Observational Routinely-collected health Data; internationaler Standard) [16] und STROSA (STandardisierte BerichtsROutine für Sekundärdaten Analysen; deutscher Standard) [17] haben wir im Rahmen dieses Beitrages eine erste Bestandsaufnahme zum Datenlinkage in Deutschland erarbeitet. Hierfür hat sich eine Projektgruppe Datenlinkage, bestehend aus 11 Experten und Expertinnen der Arbeitsgruppe Erhebung und Nutzung von Sekundärdaten (AGENS) der Deutschen Gesellschaft für Sozialmedizin und Prävention (DGSMP) und der Deutschen Gesellschaft für Epidemiologie (DGEpi) sowie der Arbeitsgruppe Validierung und Linkage von Sekundärdaten des Deutschen Netzwerks für Versorgungsforschung (DNVF) gebildet. Das Kickoff-Treffen fand im Februar 2016 statt, gefolgt von einem weiteren Treffen und mehreren Telefonkonferenzen. Eine erste Entwurfsfassung wurde den Mitgliedern der beiden o. g. Arbeitsgruppen im September 2016 zur Kommentierung gegeben, und deren Rückmeldungen wurden im Anschluss eingearbeitet.

Ziel dieses Artikels ist es, eine Unterstützung für wissenschaftliche Projekte zu liefern, die Daten aus Deutschland auf individueller Ebene verknüpfen möchten (keine Aggregatdaten, d. h. bspw. die Verlinkung von regionalen Indikatoren über die Postleitzahl

1 Die wichtigtsen Begriffe sind im Glossar in Box 2 zu finden.

2 https://www.bmbf.de/pub/Medizininformatik.pdf; Zugriff am 19.11.2017

3 https://innovationsfonds.g-ba.de/versorgungsforschung/foerderbekanntmachung-versorgungsforschung-zum-themenspezifischen-bereich.1; Zugriff am 19.11.2017

4 https://www.normenkontrollrat.bund.de/Webs/NKR/Content/DE/Artikel/2017-11-07_gutachten_register.html?nn=1660354; Zugriff am 19.11.2017

oder Kreiskennziffer). Neben den (datenschutz-)rechtlichen Rahmenbedingungen werden dabei auch praxisorientiert die Arten des Datenlinkage, deren Anwendungsfelder und mögliche Fallstricke sowie Ansätze zu deren Vermeidung anhand von Beispielen aus Deutschland dargestellt[5].

Verfahren und Arten des Datenlinkage

Im folgenden Abschnitt werden die in der deutschen Gesundheitsforschung am häufigsten praktizierten Verfahren vorgestellt und die verschiedenen Arten des Datenlinkage dargestellt. In Abhängigkeit von den Forschungsfragen und Studienzielen empfiehlt es sich, bereits bei der Planung einer Studie, in der verschiedene Datensätze verknüpft werden sollen, die Art des Datenlinkage festzulegen. Dem Forscher stehen dabei eine Reihe von Möglichkeiten zur Verfügung, die unterschiedliche rechtliche, organisatorische und technische Anforderungen an das geplante Studiendesign stellen und unterschiedlich kombinierbar sind.

Wenn möglich sollten bereits im Vorfeld für Variablen, die für das Linkage herangezogen werden, geeignete Verfahren zur Fehlervermeidung bzw. zur Standardisierung der Erfassung installiert werden (z. B. Doppelnamen mit und ohne Bindestrich, Namen mit und ohne Umlaut, ggf. Prüfziffern).

Allen sich direkt mit dem Linkage von Daten befassenden Verfahren ist gemein, dass zum Finden von zusammengehörigen Daten klar definierte Paare von Merkmalskombinationen gebildet werden müssen, z. B. Müller, Stefan, 07-09-1988 sowie Mueller, Stefan, 09-07-1988. Die Verfahren entscheiden, welche Paare zu einer Person und welche zu verschiedenen Personen gehören.

Eine wichtige Unterscheidung ist dabei zwischen überwachten Lernverfahren und unüberwachten vorzunehmen [18–20]. Bei dem erstgenannten Verfahren gibt es in den Trainingsdaten Werte für die Zielgröße „Match“ (mit den möglichen Ausprägungen: ja, nein, vielleicht), bei letzterem nicht. Zu den überwachten Verfahren gehören u. a. das probabilistische Record Linkage [21] (s.u.), Entscheidungsbaum-Methoden [22], Künstliche Neuronale Netze, Support Vector Machines und die Diskriminanzanalyse [23]. Das unüberwachte Lernen nutzt z. B. Assoziationsregeln, Cluster-Methoden und einige Bayessche Analysemethoden. Weitere Anmerkungen finden sich bei Christen [24].

Unterschieden wird in der Regel zwischen den nachfolgend aufgeführten Verfahren, deren Bedeutung für das Datenlinkage im Anschluss diskutiert werden soll.

Exaktes Linkage

Das Linkage kann exakt oder fehlertolerant erfolgen. Beim exakten Linkage führt man die Daten mehrerer Datenquellen nur bei exakter Übereinstimmung eines eindeutigen Verknüpfungsschlüssels (z. B. Sozialversicherungsnummer) oder mehrerer Linkage-Variablen zusammen (Match–Merge Linkage). Gibt es Fehler oder unterschiedliche Schreibweisen im Schlüssel oder in den Linkage-Variablen, können die Daten mit exaktem Linkage nicht zusammengebracht werden [25, 26].

Fehlertolerantes Linkage

Ist kein exaktes Linkage möglich, sind fehlertolerante Verfahren erforderlich, welche die Zahl der verknüpften Beobachtungseinheiten erhöhen können, wie z. B. durch die Nutzung von String-Metriken (s. Distanzbasiertes Linkage). Innerhalb der fehlertoleranten Linkage-Methoden können wiederum mehrere Arten unterschieden werden: regelbasiertes, distanzbasiertes und probabilistisches Linkage.

Das Finden und Festlegen von Schwellenwerten für die Entscheidung, ob Datensätze zusammengeführt werden sollen, ist eine der wesentlichen Hausforderungen des Datenlinkage, bei der auch bestimmte „Sonderfälle“ zu berücksichtigen sind, die je nach verwendeten Daten unterschiedlich sein können (z. B. Zahlendreher, gleichgeschlechtliche Zwillinge, (Nicht-)Berücksichtigung zweiter Vornamen).

Regelbasiertes Linkage

Regelbasiertes Linkage weicht die Forderung des exakten Linkage dahingehend auf, dass hier durch Regeln definiert wird, welche Identifikatoren komplett übereinstimmen müssen, und bei welchen eine teilweise Übereinstimmung ausreichend ist. Eine Regel könnte beispielweise lauten, dass der Nachname und beim Vornamen die ersten 3 Buchstaben übereinstimmen müssen, sowie dass das Geburtsjahr höchstens um 3 Jahre abweichen darf (siehe [24], S. 139–142 für Beispiele hochentwickelter Regelsätze). Eine solche Regel ist leicht zu evaluieren und beinhaltet eine gewisse Fehlertoleranz, sie birgt jedoch eine erhöhte Gefahr einer falsch positiven Klassifikation (s. u.).

Eine spezielle und in der Forschungspraxis sehr gängige Variante des regelbasierten Linkage wird als deterministisches Linkage bezeichnet. Dabei werden Identifikatoren oder Transformationen von diesen (bspw. durch Anwendung des phonetischen Codes) weiterhin auf exakte Übereinstimmung hin überprüft. Die Entscheidungsregel kann jedoch vorsehen, dass nur ein bestimmter Anteil davon übereinstimmen muss (z. B. „5 von 7 Identifikatoren müssen übereinstimmen, darunter mind. Nachname und Geburtsjahr“), um ein Datenpaar als einen Match zu klassifizieren. Falls es die Ressourcen eines Projekts und die gewählten technischen Rahmenbedingungen erlauben, ist ein iteratives Vorgehen ratsam. In einem stufenweisen deterministischen Linkage [27] sollte zunächst mit der Regel begonnen werden, die Übereinstimmung auf allen Identifikatoren fordert. Anschließend können weitere Regeln gewählt werden, welche die Ansprüche an den Grad der Übereinstimmung stufenweise reduzieren. Im jeweils nächsten Schritt werden nur solche Einheiten übernommen, die im vorherigen Schritt nicht erfolgreich verknüpft wurden, wodurch der Aufwand für den Abgleich mit jedem Schritt sinkt. Ganz offensichtlich steigt das Risiko einer falsch positiven Klassifikation mit jeder weiteren Stufe, da diese jeweils eine Lockerung der Ähnlichkeits-Anforderung darstellt. Um diesem Umstand Rechnung zu tragen, sollte der finale verknüpfte Datensatz Informationen darüber enthalten, in welcher der Stufen die Zuordnung einer Beobachtungseinheit erfolgt ist. Damit wird Transparenz geschaffen, und so bleibt es den Datennutzern überlassen, welches Risiko einer falsch positiven Klassifikation sie für ihre spezifische Analyse in Kauf nehmen und welche Beobachtungen sie in ihre Analysen einbeziehen wollen.

5 Diese Publikation fokussiert lediglich auf alle Belange, die mit einem Datenlinkage verbunden sind. Es gelten darüber hinaus andere Standards sowie andere gesetzliche Regelungen usw.

Distanzbasiertes Linkage

Eine weitere fehlertolerante Art des Datenlinkage verwendet distanzbasierte Methoden. Hierbei erlaubt der Einsatz von String-Metriken (Stringähnlichkeitsfunktionen), die Ähnlichkeit der Merkmalsausprägungen einzelner Identifikatoren zu berechnen. Als Maß für die Ähnlichkeit kann z. B. die Anzahl an Veränderungen gezählt werden, die notwendig ist, um einen String in den zu vergleichenden String umzuformen. Um „Meyer“ in „Meier“ umzuformen ist nur ein Buchstabe auszutauschen, um „Waldemar“ in „Naldo“ umzuformen sind 5 Veränderungen notwendig. Durch distanzbasierte Methoden kann die Anzahl erfolgreich verknüpfter Beobachtungseinheiten weiter erhöht werden und, sofern die Klassifikationsschwelle nicht zu niedrig gewählt wird, ein Anstieg einer falsch positiven Klassifikation vermieden werden. Gängige String-Metriken basieren auf so genannten N-Grammen, dem Edit-Distanz-Maß oder den Maßen von Jaro und Winkler [28–32]. Die Summe der Stringähnlichkeiten über verschiedene Identifikatoren hinweg wird hierbei idealerweise auf das Intervall von 0 bis 1 normiert. Eine geeignete Schwelle für die Klassifikation eines Datenpaares als Match liegt erfahrungsgemäß im Bereich von 0,7 bis 0,8. Ein Programm zur Ermittlung haben Schnell et al. entwickelt [33]. Eingesetzt wird es u. a. für die gesetzlich vorgeschriebene stationäre Qualitätssicherung beim AQUA Institut[6].

Probabilistisches Linkage

Der in der medizinischen Forschung hauptsächlich genutzte Ansatz ist das probabilistische Record-Linkage-Verfahren. Eingeführt durch Newcombe [34], wurden die theoretische Fundierung und der Erfolg des Verfahrens durch Fellegi und Sunter [21] begründet. Wesentlich an diesem Modell ist die Annahme von Wahrscheinlichkeiten, also einem probabilistischen Modell, für die Funktionswerte von Vergleichsfunktionen (z. B. als Resultat des Vergleichs von Nachnamen in 2 Datensätzen) jeweils unter der Bedingung, dass die dem zu vergleichenden Datenpaar zugrunde liegenden Personen identisch/nicht-identisch sind. Dies nutzt die Tatsache, dass die Übereinstimmung mancher Identifikatoren mehr Aussagekraft hinsichtlich der Zusammengehörigkeit zweier Beobachtungseinheiten aufweisen als die Übereinstimmung anderer. Konkret kann z. B. der Nachname deutlich mehr verschiedene Ausprägungen annehmen als das Geschlecht. Die Wahrscheinlichkeit, dass bei 2 nicht identischen Personen der Nachname zufällig übereinstimmt, ist daher deutlich geringer als die der zufälligen Übereinstimmung des Geschlechts. Die Wahrscheinlichkeit, dass ein Nachname in unterschiedlichen Quellen abweicht, obwohl es sich um dieselbe Person handelt, ist dagegen höher als beim Geschlecht, da es beim Nachnamen leichter Abweichungen geben kann, z. B. durch Schreibfehler („Meyer“ vs. „Meier“). Dieser Unterschied führt zu verschiedenen Übereinstimmungs- und Nicht-Übereinstimmungsgewichten für diese beiden Identifikatoren, die entsprechend in das Gesamtgewicht (die Ähnlichkeit über alle Identifikatoren hinweg) einfließen. In Abhängigkeit von der Festsetzung der Wahrscheinlichkeitswerte als Kriterien der (Nicht-)Übereinstimmung zwischen den Verknüpfungsmerkmalen ist jedoch auch mit einem unterschiedlichen Anteil falsch positiv und falsch negativ gematchter Fälle zu rechnen [35]. Bei den Vergleichen kann und sollte man Häufigkeiten von Namen in Rechnung stellen (frequentistischer Ansatz [36, 37]). Beispiele finden sich u. a. bei Giersiepen et al. [38] sowie Kajüter et al. [39].

Linkage mit direkten und indirekten Identifikatoren

Sowohl beim Linkage mit direkten als auch beim Linkage mit indirekten Identifikatoren können alle der erläuterten Verfahren zum Einsatz kommen.

Beim Linkage mit direkten Identifikatoren können Datensätze über ein eindeutiges Einzelmerkmal (z. B. Sozialversicherungsnummer) oder über mehrere eindeutig identifizierende Merkmale, wie z. B. Name, Geburtsdatum, Anschrift oder Versicherungsnummer, verknüpft werden. Das Linkage mit indirekten Identifikatoren kann nicht auf solche immer eindeutig identifizierenden Attribute zurückgreifen. Daher muss es in den beiden zu verknüpfenden Datensätzen Linkage-Variablen bestimmen, die in ihrer Kombination dazu führen, dass die durch das Linkage zusammengeführten Daten mit hoher Wahrscheinlichkeit ein- und derselben Person entsprechen [11, 40–42]. So wurden bspw. bei der Evaluation eines Strukturvertrages der AOK Plus zum Gestationsdiabetes in Sachsen Abrechnungsdaten und Daten aus dem Perinataldatensatz anhand von Institutskennzeichen des Krankenhauses, Datum der Aufnahme in die Klinik, Datum der Entlassung aus der Klinik, Geburtsdatum der Mutter, Geburtsgewicht des Kindes und Postleitzahl der Mutter miteinander verlinkt. Mittels dieser Angaben konnten 97 % der Abrechnungsdaten einem Datensatz aus der Perinatalerhebung zugeordnet werden [43].

Linkage mit Klartextangaben und mit verschlüsselten Identifikatoren

Die bisherigen Ausführungen sind von der Nutzung von unverschlüsselten Identifikatoren ausgegangen, d. h. unter Verwendung der Originalausprägungen von Namen, Adressen, Versicherungsnummern usw. Aus Datenschutzgründen kann jedoch der Einsatz von verschlüsselten Identifikatoren zwingend geboten sein. In solchen Fällen werden häufig Kontrollnummern eingesetzt, die mittels Verschlüsselungsverfahren aus persönlichen Daten wie Name, Vorname, Geburtsname und Geburtsdatum gebildet werden. Identische Kontrollnummern entstehen dabei jedoch nur bei exakt gleicher Schreibweise der Original-Identifikatoren. Kleinste Abweichungen führen bei Verschlüsselungsverfahren zumeist zu großen Unterschieden der Kontrollnummern. Der Einsatz phonetischer Kodierungsverfahren der Namen kann Fehler wegen unterschiedlicher Schreibweisen gleich klingender Namen reduzieren, Tippfehler aber nicht ausgleichen.

International werden zunehmend Bloom-Filter empfohlen [26, 44–46], wenn Identifikatoren für ein Datenlinkage nicht im Klartext verwendet werden können. Auf Bloom-Filtern basierende Verfahren weisen gegenüber dem Einsatz von Kontrollnummern den Vorteil auf, ein fehlertolerantes Datenlinkage auf verschlüsselten Daten durchführen zu können (Stichwort: privacy-preserving record linkage). Dabei kann eine Verschlüsselung eingesetzt werden, die eine Rückkehr zu den ursprünglichen Klartextangaben und damit ein Re-Identifizieren der Beobachtungseinheiten praktisch unmöglich macht [47, 48].

6 https://www.dgou.de/uploads/media/AQUA_Qualitaetsreport_2013.pdf; Zugriff am 09.11.2017

Linkage mit und ohne individuelle Einwilligungserklärung

In der Regel erfolgt ein Linkage mit Einwilligung. Dafür müssen schon in der Einwilligungserklärung die zu verknüpfenden Daten beschrieben und die Mitteilung erfolgen, dass ein Datenlinkage geplant ist und die i.d.R. schriftliche Einwilligung der betroffenen Studienteilnehmer nach einer umfassenden Aufklärung dafür eingeholt werden [14]. Wenn die Einwilligung vorliegt, können die Daten über direkte Identifikatoren, wie z. B. Name und Geburtsdatum, und unter Wahrung der datenschutzrechtlichen Vorgaben unmittelbar zusammengeführt werden. Es handelt sich somit um ein Datenlinkage auf Basis von personenbezogenen Daten. Im Rahmen der Studienplanung ist daher auch zu prüfen, ob die zuständigen Aufsichtsbehörden und Ethikkommissionen zur Prüfung der Gültigkeit der Formulierung der Einwilligungserklärung einzubinden sind. Die Verknüpfung der Daten setzt zudem seitens der Dateneigner ein schlüssiges Datenschutzkonzept voraus.

Liegt keine Einwilligung vor, wie z. B. bei retrospektiven Analysen schon erhobener Daten, kann i.d.R. kein Linkage mit direkten Identifikatoren vorgenommen werden, da diese u. a. aus Datenschutzgründen nicht bekannt sind oder nicht erhoben wurden bzw. werden. Die Verknüpfung der Daten kann nach Prüfung datenschutzrechtlicher Aspekte dann eventuell über indirekte Identifikatoren erfolgen, die erlauben, die Singularität des Einzelnen abzubilden, ohne dass seine Identität erfahrbar wäre [49]. So kann bei großen Sozialdatensätzen ein Datenlinkage ohne vorliegende Einwilligung auch auf Grundlage des § 75 SGB X nach Genehmigung der zuständigen Bundes- oder Landesaufsichtsbehörde durchgeführt werden. Wenn eine entsprechende Genehmigung vorliegt, ist auch ohne Einwilligung ein Linkage über direkte Identifikatoren möglich, sofern diese vorhanden sind (wie z. B. der Rentenversicherungsnummer) [50][7].

An dieser Stelle soll zudem darauf hingewiesen werden, dass eine Einwilligung selektiv sein und von verschiedenen Merkmalen der Befragten, Interviewer usw. abhängen kann. Dies sollte bei der Planung der Studie berücksichtigt werden, indem z. B. Interviewer explizit auf die Wichtigkeit und die datenschutzrechtliche Unbedenklichkeit des Linkage geschult werden oder für bestimmte Befragtengruppen nur besonders erfahrene Interviewer eingesetzt werden [51, 52].

Linkage mit unterschiedlichen Formen des Blockings

Um die Laufzeit von Verfahren zu reduzieren, wird bei allen bisher beschriebenen Arten des Linkage häufig Blocking verwendet [53]. Blocking bedeutet, dass ein oder mehrere Merkmale ausgewählt und die Datensätze gemäß den – möglicherweise hierfür vorher transformierten – Ausprägungen dieser Merkmale gruppiert werden, damit nur Datensätze mit den gleichen Ausprägungen dieser Merkmale miteinander verglichen werden. Aufgrund von Fehlern in den Datensätzen ist oft mehrmaliges Blocking nötig, um das Ausmaß falsch negativer Klassifikationen zu minimieren. Bspw. würde man eine falsch negative Klassifikation erhalten, wenn für eine Person 2 Datensätze existieren, die für den Nachnamen jeweils den Wert „Kayser“ bzw. „Kaiser“ enthalten und Blocking die Gleichheit der ersten 3 Buchstaben des Nachnamens verlangt. Als Blocking-Variablen eignen sich Merkmale mit einer hohen Anzahl unterschiedlicher Werte (Discriminating Power), da somit eine gleichmäßige Aufteilung der Datensätze in relativ kleine Gruppen ermöglicht wird. Blocking ist relevant, wenn man es mit großen Datenmengen zu tun hat. Moderne Verfahren ermöglichen ein Blocking sogar bei mit Bloom-Filtern verschlüsselten Identifikatoren [54].

Auswahl praktischer Beispiele

In ▶ **Tab. 1** sind exemplarisch einige Studien aus Deutschland zusammengestellt, die unterschiedliche Arten des Datenlinkage repräsentieren bzw. verschiedene Datenquellen kombinieren. Diese Aufzählung stellt lediglich eine Auswahl an Beispielen dar, die dem Leser einen besseren praktischen Zugang zur Thematik ermöglichen sollen. Über die referenzierten Quellen sind weitere Detailinformationen zu diesen Studien und dem Vorgehen beim Datenlinkage verfügbar.

Rechtlicher Rahmen und Datenschutz

Neben der Art des Datenlinkage müssen auch von Beginn an die rechtlichen und hier insbesondere die datenschutzrechtlichen Vorgaben geprüft und berücksichtigt werden. Dafür sind ausreichende personelle und v. a. zeitliche Ressourcen einzuplanen. Zudem sollte ein Datenschutzkonzept erarbeitet werden und die Aufgaben, Pflichten und Verantwortlichkeiten aller am Projekt Beteiligter schriftlich fixiert werden [9].

Welche Datenquellen sollen verknüpft werden?

Handelt es sich um personenbezogene Daten und insbesondere auch um Sozialdaten gelten hohe datenschutzrechtliche Anforderungen [9, 55, 56]. Es muss vorab geprüft werden, ob anonymisierte oder pseudonymisierte Daten verwendet werden können. Hierbei sind sowohl die einzelnen Datenquellen relevant als auch der final entstandene gelinkte Datensatz, da dieser durch die Kombination von Informationen weitaus sensiblere Daten enthalten kann bzw. eine datenschutzrechtliche Relevanz erst durch das Linkage entstehen kann (s. hierzu ebenfalls die Ausführungen zum Linkage mit und ohne individuelle Einwilligungserklärung). Datenlinkage muss daher im geschützten Raum und unter bestimmten Auflagen erfolgen, um eine Person nicht versehentlich zu „deanonymisieren“.

Was ist bei personenidentifizierenden Merkmalen zu beachten?

Das Datenlinkage mithilfe von personenidentifizierbaren Merkmalen als Schlüsselvariablen bedingt in aller Regel die Implementierung einer Treuhandstelle, Vertrauensstelle oder ähnlichem. Diese Stelle ist verantwortlich für die Verwaltung der Schlüsselvariablen bzw. der Schlüsseltabellen und erzeugt die notwendigen pseudonymisierten bzw. anonymisierten Daten [9]. Konkrete Beispiele finden sich u. a. bei Ihle et al. [57], March et al. [55] oder Ohlmeier et al. [58].

Brauche ich eine Einwilligungserklärung für ein Datenlinkage?

Werden Sozialdaten im Rahmen eines Forschungsvorhabens verwendet, ist laut § 67b SGB X eine schriftliche Einwilligung (informed consent) einzuholen [9]. Ausnahmen bei Unzumutbarkeit der

7 Weitere Informationen über das hier zitierte BASiD-Projekt finden auch unter: http://fdz.iab.de/de/FDZ_Projects/BASID.aspx, Zugriff am 19.11.2017

▶ **Tab. 1** Ausgewählte Beispiele für verschiedene Arten des Datenlinkage.

Studientitel	Welche Daten wurden verknüpft?	Welche Arten des Datenlinkage wurden verwendet?	Referenzen
SHARE RV: Verknüpfung von Befragungsdaten des Survey of Health, Ageing and Retirement in Europe (SHARE) mit administrativen Daten der Rentenversicherung	Befragungsdaten mit Rentenversicherungsdaten	deterministisches Linkage, Linkage mit direkten Identifikatoren, Linkage mit Einwilligungserklärung	Czaplicki & Korbmacher (2010) [64] Korbmacher & Czaplicki (2013) [65] http://www.share-project.org/
lidA - leben in der Arbeit - eine Kohortenstudie zu Gesundheit und Älterwerden in der Arbeit	Befragungsdaten mit Daten der Gesetzlichen Krankenversicherung (GKV) und der Bundesagentur für Arbeit (BA)	deterministisches Linkage, Linkage mit direkten Identifikatoren, Linkage mit Einwilligungserklärung	March et al. (2012) [55] March et al. (2015) [12] www.lidA-studie.de
	Befragungsdaten mit aggregierten Daten der GKV	Linkage mit indirekten Identifikatoren, Linkage ohne Einwilligungserklärung (Antrag nach § 75 SGB X)	
Vergleich unterschiedlicher Linkageverfahren sowie Vollständigkeit klinischer Angaben	Daten der GKV mit Daten eines Krankenhausinformationssystems	Linkage mit direkten und indirekten Identifikatoren	Ohlmeier et al. (2015) [42]
QS-AMI Studie	Daten der GKV mit Daten eines klinischen Registers	deterministisches Linkage, Linkage mit indirekten Identifikatoren	Maier et al. (2015) [11]
ALWA-ADIAB – ALWA-Befragungsdaten verknüpft mit administrativen Daten des IAB	Befragungsdaten mit Daten der BA	deterministisches und probabilistisches Linkage	Antoni et al. (2011) [66] Antoni & Seth (2012) [67]
Validierungsstudie	Daten der GKV mit Daten eines Sterberegisters	probabilistisches Linkage, deterministisches Linkage, exaktes Linkage, fehlertolerantes Linkage, Linkage mit direkten Identifikatoren	Ohlmeier et al. 2016 [58]
Sektorenübergreifende Datenzusammenführung und Evaluation am Beispiel der Schenkelhalsfrakturen	Daten der GKV mit Daten der Externen Qualitätssicherung und Pflegegutachten des Medizinischen Dienstes der Krankenkassen	exaktes Linkage, Linkage mit direkten Identifikatoren, fehlertolerantes Linkage	Ohmann et al. (2005) [68]
Kohortenstudie zur Krebsinzidenz bei Patienten mit Diabetes mellitus Typ 2	Daten aus Disease-Management (DMP)-Programmen (GKV) mit Daten eines Epidemiologischen Krebsregisters	probabilistisches Linkage	Kajüter et al. (2014) [39]

Einholung regelt der § 75 SGB X. Diese Vorgaben betreffen auch ein geplantes Datenlinkage. In der Einwilligungserklärung muss deutlich erkennbar sein, welche Daten miteinander verknüpft werden sollen [55] (s. hierzu ebenfalls die Ausführungen zum Linkage mit und ohne individuelle Einwilligungserklärung).

Softwaretools

Es existieren mittlerweile eine Reihe von Tools für das Datenlinkage, in denen sowohl exakte als auch fehlertolerante Verfahren zum Zusammenführen von Daten implementiert sind. Derzeit auf dem Markt befindliche relevante kostenpflichtige Produkte stammen alle aus Nordamerika: LinkageWiz, das von Statistics Canada entwickelte und u. a. von der Swiss National Cohort genutzte G-Link, das auf MS Access aufsetzende LinkSolv und DataMatch. Die Tools unterscheiden sich in ihrem Funktionsumfang nicht wesentlich voneinander. Eine Differenzierung gibt es bei der Unterstützung der Datenbereinigung vor der Durchführung des Linkage: Bis auf G-Link bieten alle Tools eine Standardisierungs- und Bereinigungskomponente, bspw. um Adressen in ein einheitliches Format zu bringen. Dass diese Tools v. a. mit den Zusatzfunktionalitäten locken, ist ein Indiz dafür, dass die Linkage-Verfahren v. a. mit Blick auf probabilistisches Datenlinkage auf Basis des Fellegi-Sunter-Modells relativ standardisiert und in ihren Resultaten vergleichbar sind [59, 60].

Im Bereich der kostenlos erhältlichen Softwareprodukte gibt es neben sich rein auf das Linkage beschränkenden Tools auch solche, die das Datenlinkage in ein Patienten-Identifikatoren-Management-System (PIMS) integriert haben. Zur ersten Kategorie gehören das auf SAS aufsetzende Produkt The Link King, das Machine-Learning-Methoden nutzende und auf Eclipse aufsetzende Tool ChoiceMaker, Febrl, Link Plus vom Centers for Disease Control, das R-Paket RecordLinkage, die Merge ToolBox, Oyster, BigMatch und FRIL. Zur zweiten Kategorie gehören OpenEMPI, E-PIX, Mainzelliste und der PID-Generator. Charakteristisch für die Tools dieser zweiten Kategorie ist, dass sie eher einfache Verfahren umsetzen. Dies liegt zu einem großen Teil daran, dass PIMS v. a. für den Aufbau von Datenbanken genutzt werden und beim iterativen Aufbau einer Datenbank das Linkage einen höheren Grad an robustem Automatismus bedarf als bei einem Projekt, das manuelle Be- und Nachbearbeitung vorsieht.

▶ **Tab. 2** liefert eine zusammenfassende Übersicht der Tools mit wichtigen charakterisierenden Fragen. Detailliertere Eigenschaften wie Export/Import-Möglichkeiten, Behandlung fehlender Werte, Blocking, Big-Data, usw. werden nicht behandelt, dafür sei auf die jeweiligen Links verwiesen.

Qualitätssicherung

Im Rahmen der Planung und Vorbereitung eines Datenlinkage sollten Aspekte zur Sicherung der Qualität der zu verknüpfenden Daten berücksichtigt werden. In **Box 1** sind alle im Text aufgeführten Fragen in Form einer Checkliste zusammengestellt, die dem Anwender eine schnelle Übersicht über die wichtigsten Aspekte des Datenlinkage bieten soll.

Überprüfung auf mögliche Fehler

Welche Fehler können in Identifikatoren vorkommen und wie können diese minimiert werden?

Jede Datenerhebung birgt die Gefahr von Schreibfehlern oder „Zahlendrehern", die bei Identifikatoren zu fehlerhaften Zuordnungen führen können. Darüber hinaus können Felder Abkürzungen von Merkmalsausprägungen enthalten, die im entsprechenden Feld des anderen Datensatzes nicht oder anders abgekürzt werden. Schließlich können Felder komplett leer sein, z. B. falls Befragte die Angabe ihres tagesgenauen Geburtstags verweigern. Auch wenn fehlertolerante Verfahren des Datenlinkage gerade beim Auftreten solcher Fehler gegenüber exakten Verfahren deutlich überlegen sind, so muss dennoch betont werden, dass der Linkage-Erfolg stark von der Ausgangsqualität der Identifikatoren abhängt. Für einen Überblick über Maßnahmen, die während der Studienplanung oder -durchführung zu einer hohen Datenqualität beitragen können, siehe Sakshaug und Antoni [61].

BOX 1 CHECKLISTE DER WICHTIGSTEN FRAGEN BEIM DATENLINKAGE

Frage

- Welche Datenquellen sollen verknüpft werden?
- Kommen personenidentifizierende Merkmale vor?
- Brauche ich eine Einwilligungserklärung für das Datenlinkage?
- Welche Fehler können in Identifikatoren vorkommen und wie können diese minimiert werden?
- Gibt es Datenfelder, die nicht Bestandteil des Identifikators sind, aber auf eine korrekte/fehlerhafte Zuordnung hinweisen können?
- Wie hoch ist die Anzahl verlinkter Datensätze?
- Wie kann die Güte des Linkage-Verfahrens eingeschätzt werden?
- Gibt es strukturelle Unterschiede zwischen verlinkten und nicht verlinkten Datensätzen?
- Wie gut ist die Datenqualität der Identifikatoren?
- Wie groß sind falsch negative bzw. falsch positive Klassifikationen?
- Wie gehe ich mit Zeitaspekten um?

Gibt es Datenfelder, die nicht Bestandteil des Identifikators sind, aber auf eine korrekte/fehlerhafte Zuordnung hinweisen können?

Der zu verlinkende Datensatz sollte auf Felder untersucht werden, die Hinweise geben, ob 2 Datensätze mit höherer Wahrscheinlichkeit zusammen gehören oder eher nicht. Beispiele dafür sind:

- Enthält der eine Datensatz das Alter eines Kindes und der andere eine Größenangabe, so kann anhand der Größenperzentile die Passgenauigkeit der Zuordnung abgeschätzt werden.
- Diagnosen wie Hodenkrebs oder Schwangerschaft sind geschlechtsspezifische Angaben und können für Aussagen über die Zuordnungsgenauigkeit herangezogen werden.
- Werden Datensätze im zeitlichen Verlauf verlinkt, so kann überprüft werden, inwiefern Parameter über die Zeit variieren. So sollte die Körpergröße eines Kindes nur ansteigen, die eines Erwachsenen im mittleren Lebensalter hingegen in etwa konstant bleiben, im höheren Alter eher fallen. Auch Dauerdiagnosen könnten herangezogen werden.

Plausibilitätskontrollen

Wie hoch ist die Anzahl verlinkter Datensätze?

Nach jedem Datenlinkage ist die Zahl der zusammengeführten und der nicht zusammenführbaren Datensätze auf Basis der Ausgangsdateien zu überprüfen. Hierfür ist im Vorfeld eine Abschätzung vorzunehmen, wie häufig Verknüpfungen für die einzelnen Dateien auftreten müssten. Zudem sollten beobachtete Häufigkeitsverteilungen nach erfolgtem Linkage auf Plausibilität überprüft werden.

Wie kann die Güte des Linkage-Verfahrens eingeschätzt werden?

Zur Überprüfung des gewählten Verfahrens kann auch eine Validierungsstudie als Referenzlösung eingesetzt werden [38]. Die Referenzlösung würde in diesem Fall die Zuordnung von Datenpaaren beinhalten, z. B. basierend auf Klartextdaten, die als zusammengehörig eingestuft wurden. Dadurch kann dann eine Bewertung der Güte des Datenlinkages mittels Sensitivität und Spezifität erfolgen, die den Anteil der korrekt verknüpften Personen bzw. den Anteil der korrekt nicht-verknüpften Personen angibt und eine Aussage über die Größe der falsch negativen bzw. falsch positiven Klassifikation (s. u.) erlaubt.

Gibt es systematische Unterschiede zwischen verlinkten und nicht verlinkten Datensätzen?

Nach Abschluss des Datenlinkage werden i.d.R. die zu beantwortenden Fragestellungen nur mit den Datensätzen bearbeitet, bei denen ein erfolgreiches Linkage stattgefunden hat. Datensätze, die nicht verlinkt wurden, bleiben bei diesen Analysen unberücksichtigt.

Es ist jedoch zu überprüfen, ob die nicht verlinkten Datensätze eine spezifische Struktur aufweisen, die entweder für die fehlgeschlagene Verknüpfung verantwortlich zeichnet (z. B. könnten die Daten einer Quelle aufgrund der Aktualität der zweiten Quelle noch nicht vorliegen) oder die einen systematischen Bias in dem verlinkten Datensatz bewirken könnte. Aus diesem Grunde sollten die wesentlichen Merkmale im verlinkten und nicht verlinkten Datensatz untersucht werden, um strukturelle Unterschiede zu verifizieren [52].

▶ **Tab. 2** Zusammenfassende Übersicht über die Tools des Datenlinkage.

Tools	Links	Ist die Software kostenfrei verfügbar?	Jahr der letzten Änderung	Wird deterministisches Linkage umgesetzt?	Wird probabilistisches Linkage umgesetzt?	Gibt es Datenbereinigungsprozeduren?	Gibt es eine grafische Oberfläche?	Ist die Software eingebettet in ein PIMS?
BigMatch	https://github.com/chapinhall/bigmatch_utilities	Ja	2014	Nein	Ja	Nein	Nein	Nein
ChoiceMaker	https://sourceforge.net/projects/oscmt	Ja	2016	Ja	Nein	Nein	Ja	Nein
DataMatch	https://dataladder.com	Nein	2016	Ja	Ja	Ja	Ja	Nein
E-PIX	https://mosaic-greifswald.de/werkzeuge-und-vorlagen/id-management-e-pix.html	Ja	2015	Ja	Nein	Nein	Ja	Ja
Febrl	https://sourceforge.net/projects/febrl	Ja	2013	Nein	Ja	Ja	Ja	Nein
FRIL	http://fril.sourceforge.net	Ja	2011	Ja	Ja	Nein	Ja	Nein
G-Link	http://www5.statcan.gc.ca/olc-cel/olc.action?lang=en&ObjId=10H0036&ObjType=22	Nein	2011	Nein	Ja	Nein	Ja	Nein
Link Plus	http://www.cdc.gov/cancer/npcr/tools/registryplus/lp.htm	Ja	2015	Ja	Nein	Nein	Ja	Nein
LinkageWiz	http://linkagewiz.net	Nein	2015	Ja	Ja	Ja	Ja	Nein
LinkSolv	http://strategicmatching.com	Nein	2012	Nein	Ja	Ja	Ja	Nein
Mainzelliste	www.mainzelliste.de	Ja	2016	Ja	Nein	Nein	Ja	Ja
Merge ToolBox	http://www.uni-due.de/~hq0215/mtb	Ja	2016	Nein	Ja	Nein	Ja	Nein
OpenEMPI	www.openempi.org/	Ja	2016	Ja	Ja	Nein	Ja	Ja
Oyster	https://sourceforge.net/projects/oysterer/	Ja	2013	Ja	Nein	Nein	Nein	Nein
PID-Genator	https://www.toolpool-gesundheitsforschung.de/produkte/pid-generator	Ja	2008	Ja	Nein	Nein	Ja	Ja
PPRL	http://record-linkage.de/-Downloads--software.htm	Ja	2017	Ja	Ja	Ja	Nein	Nein
RecordLinkage	https://cran.r-project.org/web/packages/RecordLinkage/index.html	Ja	2016	Ja	Ja	Nein	Nein	Nein
The Link King	http://the-link-king.com	Ja	2016	Ja	Ja	Nein	Ja	Nein

PIMS: Patienten-Identifikatoren-Management-System; Links Zugriff am 20.11.2017

Qualität der Identifikatoren

Im Rahmen der Studienplanung sollte eruiert werden, inwiefern die zum Datenlinkage vorgesehenen Identifikatoren eine eindeutige Identifizierung des tragenden Objektes, also in der Regel einer Person, ermöglichen. Es könnte aber auch ein spezieller Klinikaufenthalt einer Person sein, an den Angaben von Krankenkassen oder Rententrägern gelinkt werden sollen. Folgenden Fragen sollten in diesem Zusammenhang nachgegangen werden:

Wie gut ist die Datenqualität der Identifikatoren?

Ist zu Beginn der Studie bekannt, dass die Datenqualität der vorhandenen Identifikatoren schlecht ist oder diese nur teilweise vorhanden sind, liefern indirekte Linkage-Verfahren bessere Ergebnisse als direkte Verfahren, verbrauchen aber mehr Ressourcen und Zeit. Vor- und Nachteile der einzelnen Verfahren sollten daher im Kontext der Fragestellung vor Beginn des Datenlinkages abgewogen werden.

Wie groß sind falsch negative bzw. falsch positive Klassifikationen?

Der zu verlinkende Datensatz sollte auf Felder untersucht werden, die Hinweise geben, ob 2 Datensätze mit höherer Wahrscheinlichkeit zusammen gehören oder eher nicht. Beispiele dafür sind:

- Falsch negative Klassifikation oder Synonymfehler: Zusammengehörende Datensätze können wegen unterschiedlicher Identifikatoren nicht zusammengeführt werden. Falsch negative Klassifikationen entstehen, wenn Schreibfehler oder zu viele Änderungen in den Identifikatoren auftreten.
- Falsch positive Klassifikation oder Homonymfehler: Nicht zusammengehörende Datensätze werden wegen scheinbar identischen Identifikatoren fälschlicherweise als zusammengehörend ausgewiesen. Ursachen für falsch positive Klassifikationen sind zufällig identische Merkmale, die z. B. durch einen häufigen Namen entsteht, oder eine zu geringe Trennschärfe des Linkage-Verfahrens.

Erfolgt das Datenlinkage aufgrund von Klartextangaben, so kann der Einsatz von Ähnlichkeitsfunktionen oder die Mitführung von Kontrollnummern bei numerischen Identifikatoren Fehler reduzieren helfen.

Sofern möglich, sollten stichprobenhaft mittels eines anderen Verfahrens der Umfang von Fehlern überprüft werden (s. Punkt zur Validierungsstudie [38]).

Datenlinkage im zeitlichen Verlauf

Wie gehe ich mit Zeitaspekten um?

- Falsch negative Klassifikation bei einmaliger Zusammenführung: Wie lange liegen die Erhebungszeiträume der unterschiedlichen Datenquellen auseinander und wie wahrscheinlich ist es, dass sich in dieser Zeit identifizierende Merkmale geändert haben?
- Falsch negative Klassifikation bei mehrmaliger Zusammenführung: Sind die gewählten Identifikatoren persistent oder könnten sich Merkmale, die in den Identifikator einfließen, über die Zeit ändern? (Bsp. 1: Vor Einführung der lebenslangen Versichertennummer in der Gesetzlichen Krankenversicherung (GKV) änderte sich die Versichertennummer einer Person bei jedem Versicherungswechsel, aber auch bei Krankenkassenfusionen; Bsp. 2: Nachnamen, Wohnort, aber auch das Geschlecht können sich im Laufe des Lebens ändern)

Hat man zum Zeitpunkt des Linkage Einfluss auf die Bildung der Identifikatoren, so kann man im Falle von Änderungen mit „Übersetzungstabellen“ arbeiten, die alte und neue Identifikatoren und ihre Zuordnung enthalten.

Ausblick

Mit dieser ersten Bestandsaufnahme haben wir versucht, eine mögliche Hilfestellung bzw. Anregungen für Projekte, Gutachter sowie Datenschützer und Ethikkommissionen zu geben und konkrete Ansatzpunkte für ein erfolgreiches Datenlinkage zu benennen. Im März 2017 wurde mit GUILD (GUidance for Information about Linking Data sets) auch eine erste internationale Richtlinie für das Datenlinkage veröffentlicht [62], welche parallel zu unserem Projekt entstand. Bei unserer Publikation handelt es sich in Abgrenzung zu GUILD um eine erste Bestandsaufnahme im Kontext der deutschen Rahmenbedingungen, die im Zuge neuer relevanter Entwicklungen praxisnah und in strukturierter Form durch die Projektgruppe überarbeitet werden sollte.

Generell lässt sich festhalten, dass die Rahmenbedingungen für ein Linkage verschiedener Datenquellen in Deutschland im Vergleich zu anderen Ländern schwieriger sind. Dies ist sicherlich ein Grund dafür, dass (vorhandene) Daten hierzulande bislang vergleichsweise selten verknüpft werden. So gibt es in Kanada, Schweden oder Dänemark bspw. ab Geburt eine eindeutige Sozialversicherungsnummer, die in allen gesundheitsrelevanten Daten vorhanden ist und auch für Forschungszwecke die Verknüpfung verschiedener Daten ermöglicht. International findet Deutschland mittlerweile langsam Anschluss an den Stand in anderen Ländern, in denen Datenlinkage bereits seit diversen Jahren erfolgreich praktiziert wird [63].

BOX 2 RELEVANTE DEFINITIONEN

Aggregatdaten

„Im Sinne des Nutzens für Sekundärdatenforschung sind unter Aggregatdaten zusammenfassende Darstellungen von statistischen Auswertungen in Form von Häufigkeits- und Kreuztabellen zu verstehen. Sie können als Vergleichswerte für Repräsentativitätsprüfungen einer Stichprobe verwendet werden oder als eine Quelle von Makrodaten für Mehrebenenanalysen dienen.“ [69], S. 504

Anonymisierung

„Definition nach BDSG §3 (6) BDSG: Anonymisieren ist das Verändern personenbezogener Daten derart, dass die Einzelangaben über persönliche oder sachliche Verhältnisse nicht mehr oder nur mit einem unverhältnismäßig großen Aufwand an Zeit, Kosten und Arbeitskraft einer bestimmten oder bestimmbaren natürlichen Person zugeordnet werden können.“ [2], S. 126

Blocking
Blocking, das in der Literatur alternativ auch als Filtering oder Indexing bezeichnet wird, dient einer Effizienzsteigerung des Verknüpfungsprozesses. Statt alle Beobachtungseinheiten eines Datenbestandes mit allen Beobachtungseinheiten eines anderen Datenbestandes abzugleichen, werden nur jene Datenpaare abgeglichen, die bei einem oder mehreren Identifikatoren identisch sind (klassisches Blocking) oder eine sehr hohe Ähnlichkeit aufweisen. [24] (S. 69)

Bloom-Filter
Ganz allgemein sind Bloom-Filter Bit-Arrays (Ketten von Nullen und Einsen) deren Länge vorab zu definieren sind, wobei zunächst alle Positionen auf Null gesetzt sind. Im Rahmen des privacy-preserving record linkage verschlüsseln nun Hash-Funktionen einen Original-Identifikator und bestimmen so, welche der Nullen im Bit-Array durch Einsen ersetzt werden. Diese Bit-Arrays lassen keinen Rückschluss auf die Original-Identifikatoren zu, können aber für ein fehlertolerantes Record Linkage herangezogen werden. [26]

Dateneigner
„Unter diesem Begriff werden im Rahmen der GPS diejenigen Institutionen verstanden, die die Daten (primär) erheben, speichern und nutzen. Dateneigner und Primärnutzer sind als Synonym zu verstehen. Der Begriff Dateneigner hebt jedoch zusätzlich hervor, dass der Primärnutzer auch die rechtliche Verfügungsgewalt über die Daten besitzt. Im Bereich der Gesetzlichen Sozialversicherung sind Dateneigner beispielsweise Krankenkassen oder Rentenversicherungsträger, die versichertenbezogene (medizinische) Daten für administrative Aufgaben speichern, ebenso wie (Krebs-)Registerstellen, arbeitsmedizinische Untersuchungsstellen oder epidemiologische Einrichtungen.“ [2], S. 126

Datenlinkage / Record Linkage
„Datenlinkage (in der Informatik „record linkage“) bezeichnet die Verknüpfung verschiedener Datenquellen mittels geeigneter Schlüsselvariablen.“ [5], S. 180

Personenbezogene Daten
„Unter personenbezogenen Daten sind im epidemiologischen Sinne solche Informationen zu verstehen, die einer einzelnen bestimmten oder bestimmbaren natürlichen Person als Beobachtungseinheit zugeordnet werden können.“ [2], S. 126

Primärdaten
„Primärdaten sind Daten, die im Rahmen ihres originär vorgesehenen Verwendungszwecks aufbereitet und analysiert werden.“ [2], S. 125

Privacy-preserving record linkage
Unter dem Überbegriff privacy-preserving record linkage lassen sich Verfahren zusammenfassen, die eine Verknüpfung von Datensätzen unterschiedlicher Dateneigner ermöglichen, ohne dass zwischen den Dateneignern personenidentifizierende Daten ausgetauscht werden. Zu diesem Zweck kann eine Treuhandstelle herangezogen werden. Alternativ können die Dateneigner verschlüsselte Identifikatoren (z. B. Kontrollnummern, Bloom Filter) austauschen und für die Verknüpfung heranziehen, die keinen oder nur mit unverhältnismäßig großem Aufwand Rückschluss auf die Identität der Beobachtungseinheiten erlauben. [24]

Pseudonymisierung
„Pseudonymisieren ist das Ersetzen des Namens und anderer Identifikationsmerkmale durch ein Kennzeichen zu dem Zweck, die Bestimmung des Betroffenen auszuschließen oder wesentlich zu erschweren (‚faktische Anonymisierung‘; Bundesdatenschutzgesetz (BDSG) §3 (6a)). Dabei werden die direkt personenidentifizierenden Daten (z. B. Name, Vorname, Telefonnummer, Sozialversicherungsnummer, Personalausweisnummer) aus den Daten entfernt und durch eindeutige Kennzeichen (z. B. eine Identifikationsnummer) ersetzt. Pseudonymisierte Daten sind weiterhin personenbezogen. Eine Pseudonymisierung ist insbesondere dann notwendig, wenn personenbezogene Daten über ein bekanntes Pseudonym bereits pseudonymisierten Daten zugeordnet werden sollen.“ [2], S. 126

Schlüsselvariable / Identifikator
Eine Schlüsselvariable (Identifikator) dient der eindeutigen Identifizierung des zu verlinkenden Objektes. Sie kann aus einem oder mehreren Merkmalen bestehen. Man unterscheidet zwischen direkten und indirekten Identifikatoren.

Sekundärdaten
„Sekundärdaten sind Daten, die einer Auswertung über ihren originären, vorrangigen Verwendungszweck hinauszugeführt werden. Maßgeblich für die Einstufung als Sekundärdaten sind Unterschiede zwischen dem primären Erhebungsanlass und der nachfolgenden Nutzung. Für die Einstufung ist es unerheblich, ob die weitergehende Nutzung durch den Dateneigner selbst oder durch Dritte erfolgt. Demnach sind beispielsweise Routinedaten einer Krankenkasse nicht nur Sekundärdaten, wenn sie für wissenschaftliche Fragestellungen genutzt werden, sondern z. B. auch dann, wenn sie durch die Krankenkasse für Zwecke der Versorgungsplanung herangezogen werden.“ [2], S. 125 f.

Der Begriff der Sekundärdaten wird oftmals umgangssprachlich synonym mit anderen Begriffen wie claims data, administrativen Daten, Abrechnungs- oder Routinedaten verwendet. Bei den genannten Begriffen handelt es sich zweifelsohne um Sekundärdaten, sie sind allerdings nur Teile davon.

Sozialdaten
Nach SGB X §67 Abs. 1 Satz 1 handelt es sich dabei um „... Einzelangaben über persönliche oder sachliche Verhältnisse einer bestimmten oder bestimmbaren natürlichen Person (Betroffener), die von einer in § 35 des Ersten Buches genannten Stelle im Hinblick auf ihre Aufgaben nach diesem Gesetzbuch erhoben, verarbeitet oder genutzt werden.“ Sozialdaten unterliegen besonderen datenschutzrechtlichen Auflagen, die nach Anonymisierung nicht mehr gelten.

Treuhandstelle/Vertrauensstelle
„Sollen in einem Forschungsprojekt die Daten unterschiedlicher Dateneigner zusammengefügt oder Datensätze mit Personenidentifikatoren gespeichert werden, ist die Einrichtung einer Vertrauensstelle (oft als Treuhänderstelle bezeichnet) notwendig. Ihre Aufgabe ist neben der Weitergabe von pseudonymisierten/anonymisierten Daten vor allem die Speicherung der Personenidentifikatoren sowie der Schlüsselvariablen, die eine Zusammenspielung von Teildatensätzen erlauben." [9], S. 14

Danksagung

Die Erarbeitung dieses Quo Vadis Datenlinkage wurde ohne externe finanzielle Unterstützung durchgeführt. Wir danken allen Mitgliedern der beteiligten Arbeitsgruppen, die uns durch ihre Hinweise unterstützt haben.

Interessenkonflikt

Die Autoren geben an, dass kein Interessenkonflikt besteht.

Literatur

[1] Glaeske G, Augustin M, Abholz H et al. Epidemiological methods for health services research. Gesundheitswesen 2009; 71: 685–693

[2] Swart E, Gothe H, Geyer S et al. Good practice of secondary data analysis (GPS): guidelines and recommendations. Third Revision 2012/2014. Gesundheitswesen 2015; 77: 120–126

[3] Kurth BM. Monitoring and no end in sight: after the survey is before the survey. Bundesgesundheitsblatt Gesundheitsforschung Gesundheitsschutz 2013; 56: 607–608

[4] German National Cohort (GNC) Consortium. The German National Cohort: aims, study design and organization. Eur J Epidemiol 2014; 29: 371–382

[5] Swart E, Stallmann C, Powietzka J et al. Data linkage of primary and secondary data: a gain for small-area health-care analysis? Bundesgesundheitsblatt Gesundheitsforschung Gesundheitsschutz 2014; 57: 180–187

[6] Kreis K, Neubauer S, Klora M et al. Status and perspectives of claims data analyses in Germany-A systematic review. Health Policy 2016; 120: 213–226

[7] Hoffmann F. Review on use of German health insurance medication claims data for epidemiological research. Pharmacoepidemiol Drug Saf 2009; 18: 349–356

[8] Swart E, Ihle P, Gothe H et al. Routinedaten im Gesundheitswesen. Handbuch Sekundärdatenanalyse Grundlagen, Methoden und Perspektiven. 2. Aufl 2014 Bern: Huber

[9] March S, Rauch A, Bender S et al. Data protection aspects concerning the use of social or routine data. FDZ Methodenreport 2015; 12: 1–22

[10] Hoffmann F, Abbas S. Gut gelinkt ist halb gewonnen: Es könnte alles so einfach sein, ist es aber nicht. Gesundheitswesen 2015; 77: 72–73

[11] Maier B, Wagner K, Behrens S et al. Deterministic record linkage with indirect identifiers: data of the Berlin Myocardial Infarction Registry and the AOK Nordost for patients with myocardial infarction. Gesundheitswesen 2015; 77: e15–e19

[12] March S, Powietzka J, Stallmann C et al. The significance of a large number of health insurance funds and fusions for health services research with statutory health insurance data in Germany - Experiences of the lidA Study. Gesundheitswesen 2015; 77: e32–e36

[13] Schmidt CO, Reber K, Baumeister SE et al. Integration of primary and secondary data in the Study of Health in Pomerania and description of clinical outcomes using stroke as an example. Gesundheitswesen 2015; 77: e20–e25

[14] March S, Stallmann C, Swart E. Datenlinkage. In: Swart E, Ihle P, Gothe H, Matusiewicz D.(eds) Routinedaten im Gesundheitswesen. Handb. Sekundärdatenanalyse Grundlagen, Methoden und Perspekt. 2. Aufl. 2014:Bern: Huber; pp 347–355

[15] Hoffmann W, Latza U, Terschüren C. Deutsche Arbeitsgemeinschaft für Epidemiologie (DAE), Deutsche Gesellschaft für Medizinische Informatik, Biometrie und Epidemiologie (GMDS), Deutsche Gesellschaft für Sozialmedizin und Prävention (DGSMP) DR der IBG (DR-I. Guidelines and Recommendations for Ensuring Good Epidemiological Practice (GEP) - Revised Version after Evaluation. Gesundheitswesen 2005; 67: 217–225

[16] Benchimol EI, Smeeth L, Guttmann A et al. The REporting of studies Conducted using Observational Routinely-collected health Data (RECORD) statement. Z Evid Fortbild Qual Gesundhwes 2016; 115–116: 33–48

[17] Swart E, Bitzer E, Gothe H et al. A Consensus German Reporting Standard for Secondary Data Analyses, Version 2 (STROSA-STandardisierte BerichtsROutine für SekundärdatenAnalysen). Gesundheitswesen 2016; 78: e145–e160

[18] Sariyar M, Borg A. The Record Linkage package: Detecting errors in data. R J 2010; 2: 61–67

[19] Harron K, Goldstein H, Dibben C. Methodological developments in data linkage. 2015 John Wiley & Sons

[20] Christen P, Winkler WE. Record Linkage. Encycl. Mach. Learn. Data Min.2016:Boston, MA: Springer US; pp 1–10

[21] Fellegi IP, Sunter AB. A Theory for Record Linkage. J Am Stat Assoc 1969; 64: 1183–1210

[22] Cochinwala M, Kurien V, Lalk G et al. Efficient data reconciliation. Inf Sci (Ny) 2001; 137: 1–15

[23] Elmagarmid AK, Ipeirotis PG, Verykios VS. Duplicate record detection: A survey. IEEE Trans Knowl Data Eng 2007; 19: 1–16

[24] Christen P. Data matching: concepts and techniques for record linkage, entity resolution, and duplicate detection. 2012 Springer Science & Business Media

[25] Schnell R, Bachteler T, Reiher J. Die Anwendung statistischer Record-Linkage-Methoden auf selbst-generierte Codes bei Längsschnitterhebungen. ZA-Information 2006; 128–152

[26] Schnell R, Bachteler T, Reiher J. Entwicklung einer neuen fehlertoleranten Methode bei der Verknüpfung von personenbezogenen Datenbanken unter Gewährleistung des Datenschutzes. Methoden, Daten, Anal 2009; 3: 203–217

[27] Gomatam S, Carter R, Ariet M et al. An empirical comparison of record linkage procedures. Stat Med 2002; 21: 1485–1496

[28] Bilenko M, Mooney RJ. Adaptive duplicate detection using learnable string similarity measures. Proc. ninth ACM SIGKDD Int. Conf. Knowl. Discov. data Min. 2003; pp 39–48

[29] Navarro G. A guided tour to approximate string matching. ACM Comput Surv 2001; 33: 31–88

[30] Ristad ES, Yianilos PN. Learning string-edit distance. IEEE Trans Pattern Anal Mach Intell 1998; 20: 522–532

[31] Winkler WE. String comparator metrics and enhanced decision rules in the Fellegi-Sunter model of record linkage. 1990;

[32] Jaro MA. Advances in record-linkage methodology as applied to matching the 1985 census of Tampa, Florida. J Am Stat Assoc 1989; 84: 414–420

[33] Schnell R, Bachteler T, Reiher J. MTB: ein Record-Linkage-Programm für die empirische Sozialforschung. ZA-Information 2005; 56: 93–103

[34] Newcombe HB, Kennedy JM, Axford SJ et al. Automatic linkage of vital records. Science 1959; 130: 954–959

[35] Tromp M, Ravelli AC, Bonsel GJ et al. Results from simulated data sets: probabilistic record linkage outperforms deterministic record linkage. J Clin Epidemiol 2011; 64: 565–572

[36] Yancey WE. Evaluating string comparator performance for record linkage. Stat Res Div Res Rep 2005

[37] Zhu VJ, Overhage MJ, Egg J et al. An empiric modification to the probabilistic record linkage algorithm using frequency-based weight scaling. J Am Med Informatics Assoc 2009; 16: 738–745

[38] Giersiepen K, Bachteler T, Gramlich T et al. Performance of record linkage for cancer registry data linked with mammography screening data. Bundesgesundheitsblatt Gesundheitsforschung Gesundheitsschutz 2010; 53: 740–747

[39] Kajüter H, Geier A, Wellmann I et al. Cohort study of cancer incidence in patients with type 2 diabetes : Record linkage of encrypted data from an external cohort with data from the Epidemiological Cancer Registry of North Rhine-Westphalia. Bundesgesundheitsblatt Gesundheitsforschung Gesundheitsschutz 2014; 57: 52–59

[40] Hammill BG, Hernandez AF, Peterson ED et al. Linking inpatient clinical registry data to Medicare claims data using indirect identifiers. Am Heart J 2009; 157: 995–1000

[41] Brennan JM, Peterson ED, Messenger JC et al. Linking the National Cardiovascular Data Registry CathPCI Registry with Medicare claims data: validation of a longitudinal cohort of elderly patients undergoing cardiac catheterization. Circ Cardiovasc Qual Outcomes 2012; 5: 134–140

[42] Ohlmeier C, Hoffmann F, Giersiepen K et al. Linkage of Statutory Health Insurance Data with those of a Hospital Information System: Feasible, but also "Useful"?. Gesundheitswesen 2015; 77: e8–e14

[43] Rothe U, Müller G. Evaluation eines Strukturvertrages zur Inzidenz des Gestationsdiabetes auf der Basis von Sekundärdaten. Diabetol und Stoffwechsel 2013; 8: FV65

[44] Boyd JH, Randall SM, Ferrante AM. Application of privacy-preserving techniques in operational record linkage centres. Med. Data Priv. Handb 2015; Springer pp 267–287

[45] Randall SM, Ferrante AM, Boyd JH et al. Privacy-preserving record linkage on large real world datasets. J Biomed Inform 2014; 50: 205–212

[46] Vatsalan D, Christen P. Privacy-preserving matching of similar patients. J Biomed Inform 2016; 59: 285–298

[47] Niedermeyer F, Steinmetzer S, Kroll M et al. Cryptanalysis of basic bloom filters used for privacy preserving record linkage. J Priv Confidentiality 2014; 6: 3

[48] Randall SM, Ferrante AM, Boyd JH et al. Limited privacy protection and poor sensitivity. Is it time to move on from the statistical linkage key-581? Heal Inf Manag J 2016; 45: 71–79

[49] Weber SC, Lowe H, Das A et al. A simple heuristic for blindfolded record linkage. J Am Med Inform Assoc 2012; 19: e157–e161

[50] Hochfellner D, Voigt A, Budzak U et al. Das Projekt BASiD: Biografiedaten ausgewählter Sozialversicherungsträger in Deutschland. Projektinhalte, aktueller Stand der Arbeiten und Analysemöglichkeiten. Dtsch Rentenversicherung Bund (Hrsg). FDZ-RV-Daten zur Rehabil über Versicherte und Rentner 2010; 74–86

[51] Korbmacher JM, Schroeder M. Consent when linking survey data with administrative records: the role of the interviewer. Surv Res Methods 2013; pp 115–131

[52] March S, Swart E, Robra B-P. Können Krankenkassendaten Primärdaten verzerrungsfrei ergänzen? – Selektivitätsanalysen im Rahmen der lidA-Studie. Gesundheitsökonomie Qual 2017; 22: 104–115

[53] Christen P. A survey of indexing techniques for scalable record linkage and deduplication. IEEE Trans Knowl Data Eng 2012; 24: 1537–1555

[54] Schnell R. An efficient privacy-preserving record linkage technique for administrative data and censuses. Stat J IAOS 2014; 30: 263–270

[55] March S, Rauch A, Thomas D et al. Procedures according to data protection laws for coupling primary and secondary data in a cohort study: the lidA study. Gesundheitswesen 2012; 74: e122–e129

[56] Ihle P. Data protection and methodological aspects in compiling a routine database from statutory health insurance data for research purposes. Bundesgesundheitsblatt Gesundheitsforschung Gesundheitsschutz 2008; 51: 1127–1134

[57] Ihle P, Köster I, Herholz H et al. Sample survey of persons insured in statutory health insurance institutions in Hessen–concept and realisation of person-related data base. Gesundheitswesen 2005; 67: 638–645

[58] Ohlmeier C, Langner I, Garbe E et al. Validating mortality in the German Pharmacoepidemiological Research Database (GePaRD) against a mortality registry. Pharmacoepidemiol Drug Saf 2016; 25: 778–784

[59] Sariyar M, Borg A. Deterministic linkage as a preceding filter for other record linkage methods. Int J Inf Technol Decis Mak 2015; 14: 521–533

[60] Sariyar M, Borg A, Pommerening K. Evaluation of record linkage methods for iterative insertions. Methods Inf Med 2009; 48: 429–437

[61] Sakshaug JW, Antoni M. Errors in linking survey and administrative data. In: Biemer PP, De Leeuw ED, Eckman S et al. (eds) Total Surv. Error Pract. Improv. Qual. Era Big Data. 2016 John Wiley & Sons

[62] Gilbert R, Lafferty R, Hagger-Johnson G et al. GUILD: GUidance for Information about Linking Data sets. J Public Health (Oxf) 2017; 1–8

[63] Ferrie JE. IJE series old and new. Int J Epidemiol 2014; 43: 1689–1690

[64] Czaplicki C, Korbmacher J. SHARE-RV: Verknüpfung von Befragungsdaten des Survey of Health, Ageing and Retirement in Europe mit administrativen Daten der Rentenversicherung. In: Deutsche Rentenversicherung Bund (ed) Gesundheit, Migr. und Einkommensungleichheit. 2010: DRV-Schriften Band 55/2010 pp 28–37

[65] Korbmacher J, Czaplicki C. Linking SHARE survey data with administrative records: First experiences from SHARE-Germany. In: Malter F, Börsch-Supan A. (eds) SHARE Wave 4 Innov. Methodol. 2013:Munich: Max Planck Institute for Social Law and Social Policy; pp 47–52

[66] Antoni M, Jacobebbinghaus P, Seth S. ALWA-Befragungsdaten verknüpft mit administrativen Daten des IAB (ALWA-ADIAB) 1975-2009. FDZ Methodenreport 2011; 5: 1–64

[67] Antoni M, Seth S. ALWA-ADIAB-linked individual survey and administrative data for substantive and methodological research. Schmollers Jahrb 2012; 132: 141–146

[68] Ohmann C, Smektala R, Pientka L et al. A new model of comprehensive data linkage–evaluation of its application in femoral neck fracture. Zeitschrift für ärztliche Fortbildung und Qual 2005; 99: 547–554

[69] Swart E, Ihle P, Gothe H et al. Glossar. In: Swart E, Ihle P, Gothe H, Matusiewicz D. (eds) Routinedaten im Gesundheitswesen. Handb. Sekundärdatenanalyse Grundlagen, Methoden und Perspekt. 2. Aufl. 2014: Bern: Huber; pp 504–515