Gute Praxis Datenlinkage (GPD)

Good Practice Data Linkage

Autoren

Stefanie March[1], Silke Andrich[2, 3], Johannes Drepper[4], Dirk Horenkamp-Sonntag[5], Andrea Icks[2, 3], Peter Ihle[6], Joachim Kieschke[7], Bianca Kollhorst[8], Birga Maier[9], Ingo Meyer[6], Gabriele Müller[10], Christoph Ohlmeier[11], Dirk Peschke[12, 13], Adrian Richter[14], Marie-Luise Rosenbusch[15], Nadine Scholten[16], Mandy Schulz[15], Christoph Stallmann[1], Enno Swart[1], Stefanie Wobbe-Ribinski[17], Antke Wolter[17], Jan Zeidler[18], Falk Hoffmann[19]

Institute

1 Institut für Sozialmedizin und Gesundheitsökonomie (ISMG), Medizinische Fakultät, Otto-von-Guericke-Universität Magdeburg, Magdeburg
2 Institut für Versorgungsforschung und Gesundheitsökonomie, Centre for Health and Society, Medizinische Fakultät, Heinrich-Heine-Universität Düsseldorf, Düsseldorf
3 Institut für Versorgungsforschung und Gesundheitsökonomie, Deutsches Diabetes-Zentrum (DDZ), Leibniz-Zentrum für Diabetes-Forschung an der Heinrich-Heine-Universität Düsseldorf, Düsseldorf
4 TMF – Technologie- und Methodenplattform für die vernetzte medizinische Forschung e. V., Berlin
5 Techniker Krankenkasse, Versorgungsmanagement, Hamburg
6 PMV forschungsgruppe, Universität zu Köln, Köln
7 Epidemiologisches Krebsregister Niedersachsen, Registerstelle, Oldenburg
8 Leibniz-Institut für Präventionsforschung und Epidemiologie – BIPS, Abteilung Biometrie und EDV, Bremen
9 Berlin-Brandenburger Herzinfarktregister e. V., Berlin-Brandenburger Herzinfarktregister, Berlin
10 Zentrum für Evidenzbasierte Gesundheitsversorgung (ZEGV), Universitätsklinikum und Medizinische Fakultät Carl Gustav Carus, TU Dresden, Dresden
11 IGES Institut GmbH, Berlin
12 Institut für Public Health und Pflegeforschung (IPP), Universität Bremen, Bremen
13 Department für Angewandte Gesundheitswissenschaften, Studienbereich Physiotherapie, Hochschule für Gesundheit Bochum, Bochum
14 Institut für Community Medicine, Universitätsmedizin Greifswald, Greifswald
15 Zentralinstitut für die kassenärztliche Versorgung in Deutschland (Zi), Fachbereich Versorgungsforschung, Systemanalyse und Data Science, Berlin
16 Institut für Medizinsoziologie, Versorgungsforschung und Rehabilitationswissenschaft (IMVR), Humanwissenschaftliche Fakultät und Medizinische Fakultät, Universität zu Köln, Köln
17 DAK Gesundheit, Vorstandsreferat Versorgungsforschung, Hamburg
18 Center for Health Economics Research Hannover (CHERH), Leibniz Universität Hannover, Hannover
19 Fakultät für Medizin und Gesundheitswissenschaften, Department für Versorgungsforschung, Carl von Ossietzky Universität Oldenburg, Oldenburg

Schlüsselwörter

Record Linkage, Leitlinie, Standard, personenbezogene Daten, Versorgungsforschung, Epidemiologie

Key words

record linkage, guidelines, standard, personal data, health services research, epidemiology

Bibliografie

DOI https://doi.org/10.1055/a-0962-9933
Online-Publikation: 8.8.2019
Gesundheitswesen 2019; 81: 636–650

ISSN 0941-3790

Korrespondenzadresse

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

ZUSAMMENFASSUNG

Das personenbezogene Verknüpfen verschiedener Datenquellen (Datenlinkage) für Forschungszwecke findet in den letzten Jahren in Deutschland zunehmend Anwendung. Jedoch fehlen hier-

für konsentierte methodische Standards. Ziel dieses Beitrages ist es, solche Standards für Forschungsvorhaben zu definieren. Eine weitere Intention ist es, dem Lesenden eine Checkliste zur Bewertung geplanter Forschungsvorhaben und Artikel bereitzustellen. Zu diesem Zweck hat eine aus Mitgliedern verschiedener Fachgesellschaften zusammengesetzte Expertengruppe seit 2016 insgesamt 7 Leitlinien mit 27 konkreten Empfehlungen erstellt. Die Gute Praxis Datenlinkage beinhaltet die folgenden Leitlinien: (1) Forschungsziele, Fragestellung, Datenquellen und Ressourcen, (2) Dateninfrastruktur und Datenfluss, (3) Datenschutz, (4) Ethik, (5) Schlüsselvariablen und Linkageverfahren, (6) Datenprüfung/Qualitätssicherung sowie (7) Langfristige Datennutzung für noch festzulegende Fragestellungen. Jede Leitlinie wird ausführlich diskutiert. Zukünftige Aktualisierungen werden wissenschaftliche und datenschutzrechtliche Entwicklungen berücksichtigen.

ABSTRACT

Individual data linkage of different data sources for research purposes is being increasingly used in Germany in recent years. However, generally accepted methodological guidance is missing. The aim of this article is to define such methodological standards for research projects. Another aim is to provide readers with a checklist for critical appraisal of research proposals and articles. Since 2016, an expert panel of members of different German scientific societies have worked together and developed 7 guidelines with a total of 27 practical recommendations. These recommendations include (1) research aims, questions, data sources and resources, (2) infrastructure and data flow, (3) data privacy, (4) ethics, (5) key variables and type of linkage, (6) data validation/quality assurance and (7) long-term use for future research questions. The authors provide a rationale for each recommendation. Future revisions will include any new developments in science and data privacy.

Für die Arbeitsgruppe Erhebung und Nutzung von Sekundärdaten (AGENS) der Deutschen Gesellschaft für Sozialmedizin und Prävention (DGSMP) und der Deutschen Gesellschaft für Epidemiologie (DGEpi), für die Arbeitsgruppe Validierung und Linkage von Sekundärdaten des Deutschen Netzwerks Versorgungsforschung (DNVF) sowie für die Arbeitsgruppe Datenschutz und die Arbeitsgruppe IT-Infrastruktur und Qualitätsmanagement der TMF – Technologie- und Methodenplattform für die vernetzte medizinische Forschung e. V.

Ziele und Zielgruppe der GPD

Mit der Guten Praxis Datenlinkage (GPD) wird ein Standard für die Durchführung von Forschungsvorhaben in der Gesundheits- und Sozialforschung formuliert, die ein Datenlinkage personenbezogener Daten[1] nach wissenschaftlichen Grundsätzen anstreben (▶ **Tab.1**).

Die GPD ergänzt die Leitlinien und Empfehlungen der Guten Praxis Sekundärdatenanalyse (GPS) der Deutschen Gesellschaft für Sozialmedizin und Prävention (DGSMP) und der Deutschen Gesellschaft für Epidemiologie (DGEpi) [1] sowie die Gute Epidemiologische Praxis (GEP) der DGEpi, der Deutschen Gesellschaft für Medizinische Informatik, Biometrie und Epidemiologie (GMDS), der DGSMP, des Deutschen Netzwerks Versorgungsforschung (DNVF) und der TMF – Technologie- und Methodenplattform für die vernetzte medizinische Forschung e. V. (TMF) [2]. Die GPD richtet sich insbesondere an die folgenden Zielgruppen:

- Dateneigner,
- Wissenschaftlerinnen und Wissenschaftler,
- Gutachterinnen und Gutachter wissenschaftlicher Projekte und Publikationen,
- Aufsichtsbehörden, Ethikkommissionen, Datenschutzverantwortliche.

Die GPD verfolgt vorrangig 2 Zielsetzungen. Zum einen stellt sie eine Handlungsanleitung im Sinne einer Leitlinie für Wissenschaftlerinnen und Wissenschaftler dar, die ein Forschungsvorhaben mit Datenlinkage planen und durchführen. Zwar werden nicht in allen Forschungsvorhaben die nachstehend aufgeführten Leitlinien und Empfehlungen vollständig umzusetzen sein, jedoch erfordert dieser methodische Standard eine kritische Reflexion des geplanten Vorgehens und eine Begründung für Fälle, in denen von den Empfehlungen dieser Guten Praxis abgewichen wird. Zum anderen dient die GPD den oben genannten Zielgruppen als Checkliste bei der Bewertung geplanter Vorhaben und bei der Bewertung von Publikationen über solche. Angesichts der steigenden Anzahl von Vorhaben mit Datenlinkage und den damit verbundenen methodischen Herausforderungen wird damit den genannten Institutionen ein Instrument für eine fundierte Rückmeldung an Forschergruppen an die Hand gegeben. Die GPD soll zu einer Verbesserung der Qualität von Forschungsvorhaben beitragen.

Anwendungsfelder

Unter Datenlinkage wird die Verknüpfung verschiedener Datenquellen verstanden. Im Kern wird hierbei ein umfassender Prozess betrachtet, der von der Planung eines Forschungsvorhabens über die eigentliche Zusammenführung unterschiedlicher Datenquellen bis hin zur Auswertung und Nutzung durch weitere Personen inklusive der Löschung der Forschungsdaten oder Anonymisierung reicht; kurz: alle Prozessschritte, die in der Anwendung von Datenlinkage zu beachten sind. Die Betrachtung der Zusammenführung gleicher Datenquellen im Rahmen eines Follow-ups steht nicht im Fokus.

Der Begriff „Datenlinkage“ steht im Verständnis des Autorenteams dabei als übergeordneter Begriff für den oben beschriebenen Gesamtprozess. Im Vergleich dazu wird „Record Linkage“ als Teil des Gesamtprozesses verstanden, nämlich als Instrumentarium, um Datensätze zusammenzuführen, und bezeichnet damit den technischen Aspekt der Datenverknüpfung.

1 Die wichtigsten Begriffe sind im Glossar in Tab.1 zu finden.

▶ **Tab.1** Glossar.

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." [26], S. 504.
Anonymisierung
Anonymisierung ist in der Datenschutzgrundverordnung (DSGVO) nur indirekt als Gegensatz zur Identifizierbarkeit definiert: „Die Grundsätze des Datenschutzes sollten daher nicht für anonyme Informationen gelten, d. h. für Informationen, die sich nicht auf eine identifizierte oder identifizierbare natürliche Person beziehen, oder personenbezogene Daten, die in einer Weise anonymisiert worden sind, dass die betroffene Person nicht oder nicht mehr identifiziert werden kann". (Erwägungsgrund 26 Satz 5 DSGVO [10]). „Um festzustellen, ob eine Person identifizierbar ist, sollten alle Mittel berücksichtigt werden, die hierzu nach allgemeinem Ermessen wahrscheinlich genutzt werden". (Erwägungsgrund 26 Satz 3 DSGVO [10]). Was nach allgemeinem Ermessen wahrscheinlich zur Identifizierung genutzt wird, soll nach objektiven Faktoren, wie den Kosten der Identifizierung und des dafür erforderlichen Zeitaufwands, ermittelt werden. Dabei sind die aktuell verfügbare Technologie und technologische Entwicklungen zu berücksichtigen (vgl. Erwägungsgrund 26 Satz 4 DSGVO [10]).
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 [54], S. 69." [3], S. e29
Bloom-Filter
Bloom-Filter sind Bit-Arrays definierter Länge (Ketten von Nullen und Einsen), wobei zunächst alle Positionen auf Null gesetzt sind. Bei einer Verarbeitung von Original-Identifikatoren bestimmen Hashfunktionen, an welcher Stelle Nullen im Bit-Array durch Einsen ersetzt werden. Eine abweichende Abfolge von Nullen und Einsen weist auf Unterschiede bei den Original-Identifikatoren hin. Es besteht die Möglichkeit, Ähnlichkeiten zu erkennen, um z. B. eine Fehlertoleranz zuzulassen. Beispiel für Record Linkage mit einem Bloom-Filter bei [55].
Data Dictionaries (Datensatzbeschreibung)
In einer Datensatzbeschreibung sollten alle Variablen eines Datensatzes aufgeführt und deren Eigenschaften beschrieben werden. Zu diesen Eigenschaften zählen der Name, evtl. ein Label, der Datentyp (z. B. numerisch, Zeichenkette, Datum, Datum-Zeitstempel) und eine Erklärung zum Inhalt der Variablen.
Dateneigner
Datenhaltende Stellen, die Daten primär erheben bzw. diese verwalten und weitere Rechte der Nutzung innehaben (aber keine Persönlichkeitsrechte), insbesondere auch das Recht der Sekundärnutzung, sofern keine gesetzlichen Bestimmungen dagegen sprechen.
Datenherkunft (Data provenance)
Die Beschreibung der Herkunft von Daten umfasst neben der Angabe der Quelle den Prozess der Entstehung eines Datenbestandes. Hierunter fallen insbesondere Bedingungen einer Selektion und datenkurierende Maßnahmen. Die Genauigkeit der Beschreibung kann bis auf Ebene einzelner Variablen erfolgen. Beispielsweise kann bei Ergebnissen eines MRT-Readings zur Häufigkeit von Ödemen beschrieben werden, ob diese aus einem Einfach- oder Mehrfachreading entstanden sind. Bei letzterem sollte dann bspw. der Umgang bei Dissens der Readingergebnisse beschrieben werden.
Datenqualität
Die Datenqualität ist spezifisch für die jeweilige Datenquelle und die Forschungsfrage zu beurteilen [30, 31]. Vollständigkeit und Fehlerfreiheit von Daten gelten dabei als besonders relevante Dimensionen von Datenqualität [29]. Mit variierendem Kontext einer wissenschaftlichen Fragestellung kann auch die Bewertung von Datenqualität desselben Datenbestandes unterschiedlich erfolgen. Beispielsweise könnten 10 % eines Gesamtdatenbestandes mit unsystematisch fehlenden Angaben auch als unproblematisch beurteilt werden. Dagegen kann dieser Anteil fehlender Angaben in einem psychologischen Messinstrument als problematisch beurteilt werden, wenn diese häufiger bei Probanden/Teilnehmenden mit psychologischer Beeinträchtigung beobachtet werden.
Datenschutz-Folgenabschätzung (DSFA)
DSFA ist ein neu mit der DSGVO in Art. 35 [10] eingeführtes Verfahren zur Dokumentation und Bewertung spezifischer Datenschutzrisiken, die durch eine Verarbeitung personenbezogener Daten entstehen. Die Durchführung einer DSFA ist u. a. bei einer umfangreichen Verarbeitung besonderer Kategorien personenbezogener Daten erforderlich, zu denen auch Gesundheitsdaten gehören. Bei der Durchführung ist der Rat eines Datenschutzbeauftragten, sofern benannt, einzuholen. Die DSFA enthält zumindest eine systematische Beschreibung der geplanten Verarbeitungsvorgänge und der Zwecke der Verarbeitung sowie eine Bewertung der Notwendigkeit und Verhältnismäßigkeit der Verarbeitung in Bezug auf die Zwecke. Zudem sind die Risiken für die Rechte und Freiheiten betroffener Personen sowie die zur Bewältigung der Risiken geplanten technischen und organisatorischen Maßnahmen zu beschreiben und zu bewerten. Ergibt eine DSFA, dass trotz der geplanten Maßnahmen die Risiken für die betroffenen Personen hoch sind, ist vor dem Beginn der Verarbeitung die zuständige Aufsichtsbehörde gemäß Art. 36 DSGVO [10] zu konsultieren. Die DSFA kann Teil eines Datenschutzkonzepts sein, wenn dies entsprechend kenntlich gemacht wird.
Datenminimierung
Synonym für Datensparsamkeit, Begriff aus der DSGVO [10]
Einwilligung
Synonym für consent/informed consent (informierte Einwilligung)
Entity-Relationship-Diagrams
Unter Berücksichtigung der Datensatzbeschreibung (data dictionaries) der einzelnen Datenquellen können für deren Verknüpfung die unterschiedlichen Inhalte (Entities) mit Namen und Datentyp in einem Diagramm dokumentiert werden. Die Beziehung der Schlüsselvariablen (Relationships) ist dabei anzugeben. Idealerweise werden hierbei Angaben zur Kardinalität der Beziehungen gemacht (1:1, 1:n, n:m) [37].
Juristische Personen
Eine juristische Person ist gemäß Buch 1, Abschnitt 1, Titel 2 des Bürgerlichen Gesetzbuches (BGB) eine Personenvereinigung (z. B. Verein, Stiftung) oder Zweckvermögen mit eigener Rechtsfähigkeit. Eine juristische Person ist damit Träger von Rechten und Pflichten, hat Vermögen, kann Verträge abschließen und kann in eigenem Namen klagen oder verklagt werden. Es wird unterschieden zwischen juristischen Personen des privaten Rechts und juristischen Personen des öffentlichen Rechts.

▶ **Tab.1** Fortsetzung.

Hashfunktion
Eine Hashfunktion ist eine Funktion, die eine Eingabemenge auf eine kleinere Zielmenge (die sogenannten Hashwerte) abbildet. Ein einfaches Beispiel wäre die Berechnung der Quersumme von zweistelligen Zahlen. Hashfunktionen können genutzt werden, um einen Inhalt nahezu eindeutig zu identifizieren, ohne etwas über den Inhalt auszusagen. Daher können Hashwerte auch als Pseudonyme für Namen und andere Identifikationsmerkmale im Rahmen eines Record Linkage verwendet werden.
Homonymfehler
Eine falsche Datenzusammenführung = falsch positiv, d. h. Daten werden verlinkt, die zu unterschiedlichen Personen/Entitäten gehören
Personenbezogene Daten
Nach Art. 4 Abs. 1 DSGVO sind „personenbezogene Daten“ alle Informationen, die sich auf eine identifizierte oder identifizierbare natürliche Person (im Folgenden „betroffene Person“) beziehen; als identifizierbar wird eine natürliche Person angesehen, die direkt oder indirekt, insbesondere mittels Zuordnung zu einer Kennung wie einem Namen, zu einer Kennnummer, zu Standortdaten, zu einer Online-Kennung oder zu einem oder mehreren besonderen Merkmalen identifiziert werden kann, die Ausdruck der physischen, physiologischen, genetischen, psychischen, wirtschaftlichen, kulturellen oder sozialen Identität dieser natürlichen Person sind“ [10].
Primärdaten
„Primärdaten sind Daten, die im Rahmen ihres originär vorgesehenen Verwendungszwecks aufbereitet und analysiert werden“ [1], S. 125.
Prüfziffer
Prüfziffern dienen der Eingabeprüfung einer Ziffern- oder Zeichenfolge. Sie werden nach einem definierten Algorithmus aus den zu prüfenden Ziffern bzw. Zeichen berechnet. Eingabefehler bzw. Fehlübermittlungen führen zu Fehlermeldungen. Für die Prüfzifferngenerierung gibt es verschiedene Verfahren. Beispiele siehe https://pub.uni-bielefeld.de/record/1775228; Zugriff am 19.07.2019.
Pseudonymisierung
„[…] die Verarbeitung personenbezogener Daten in einer Weise, dass die personenbezogenen Daten ohne Hinzuziehung zusätzlicher Informationen nicht mehr einer spezifischen betroffenen Person zugeordnet werden können, sofern diese zusätzlichen Informationen gesondert aufbewahrt werden und technischen und organisatorischen Maßnahmen unterliegen, die gewährleisten, dass die personenbezogenen Daten nicht einer identifizierten oder identifizierbaren natürlichen Person zugewiesen werden“ (Art. 4 Abs. 5 DSGVO [10]).
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“ [3], S. e29 Direkte Identifikatoren, oft auch als *personenbezogene Daten* bezeichnet, verweisen direkt auf eine natürliche Person (z. B. Kombination aus Name, Adresse und Geburtsdatum oder die Krankenversichertennummer). Indirekte Identifikatoren verweisen nicht eindeutig auf eine Person. Kann man durch Kontextinformationen jedoch davon ausgehen, dass sich in zwei (anonymisierten bzw. mittels unterschiedlicher Verfahren pseudonymisierten) Datensätzen Daten zu identischen Personen finden, so kann über die Kombination von Merkmalen eine Zuordnung erfolgen (z. B. Alter, Geschlecht, Klinikaufnahmedatum und Klinikaufnahmeuhrzeit) [20].
Selektionsbias
„Selektionsbias ist eine Verzerrung, die aus Verfahren zur Selektion von Studienteilnehmenden und aus Faktoren, die die Studienteilnahme beeinflussen, resultiert (= Stichprobenverzerrung)“ [56], S. 134. Diese Art von Verzerrung kann z. B. entstehen, wenn systematische Unterschiede zwischen Studienteilnehmenden und Nichtstudienteilnehmenden bestehen, sodass Studienrepräsentativität und Generalisierbarkeit von Aussagen nicht mehr gewährleistet sind. Selektion kann bei Datenlinkage dadurch entstehen, dass Studienteilnehmende mit erfolgreich verlinkten Daten sich von solchen ohne Verlinkung nennenswert unterscheiden.
Sekundärdaten
„Sekundärdaten sind Daten, die einer Auswertung über ihren originären, vorrangigen Verwendungszweck hinaus zugeführt 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 bspw. 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“ [1], 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“ [3], S. e29. Zudem zählen auch Daten, die z. B. im Rahmen von Surveys erhoben und unter neuer Fragestellung ausgewertet werden, zu den Sekundärdaten.
Sozialdaten
„Sozialdaten sind *personenbezogene Daten* (Artikel 4 Nummer 1 der Verordnung (EU) 679/2016), die von einer in § 35 des Ersten Buches genannten Stelle im Hinblick auf ihre Aufgaben nach diesem Gesetzbuch verarbeitet werden“ (§ 67 Abs. 2 S. 1 SGB X [33]).
Synonymfehler
Eine fehlende Datenzusammenführung = falsch negativ, d. h. Daten werden nicht verlinkt, obwohl sie die gleiche Person/Entität betreffen
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 v. a. die Speicherung der Personenidentifikatoren sowie der Schlüsselvariablen, die eine Zusammenspielung von Teildatensätzen erlauben“ [40], S. 14.

Bei der Art der zu verlinkenden Daten kann es sich einerseits um Informationen handeln, die sich auf natürliche Personen beziehen und diese direkt oder indirekt identifizieren können (im Weiteren als personenbezogene Daten bezeichnet). Andererseits können auch einrichtungsbezogene juristische Personen gemeint sein, wie bspw. Krankenhäuser.

Die GPD thematisiert ausschließlich die Verknüpfung personenbezogener Daten (Primär- mit Primär- bzw. Sekundärdaten sowie verschiedenen Sekundärdatenquellen untereinander). Das Linkage personenbezogener Daten mit aggregierten Forschungsdaten (Aggregatdaten) wie bspw. Klassifikationssystemen, Regionaldaten usw. wird im Rahmen der GPD nicht betrachtet. Überdies behandelt die GPD alle Prozessschritte nur insoweit, als sie das Thema Datenlinkage betreffen, und gibt auch nur hierfür entsprechende Empfehlungen[2]. Der primäre Fokus der GPD ist auf das Gesundheitssystem mit seinen spezifischen (Daten-)Strukturen gerichtet. Dennoch kann die GPD auch auf andere Felder angewendet werden.

Methodik

Im Jahr 2016 traf sich erstmalig eine Projektgruppe Datenlinkage, die zu Beginn aus elf Expertinnen und Experten der Arbeitsgruppe Erhebung und Nutzung von Sekundärdaten (AGENS) der DGSMP und der DGEpi sowie der Arbeitsgruppe Validierung und Linkage von Sekundärdaten des DNVF bestand. Im Rahmen einer Bestandsaufnahme zu deutschen Forschungsvorhaben mit Datenlinkage und deren Vorgehensweisen wurde Anfang 2018 der Status Quo Datenlinkage veröffentlicht [3]. Im Mai 2018 fand das Kick-off-Treffen für die Erstellung einer GPD in Hannover statt, bei dem die Projektgruppe auf 23 Mitglieder erweitert wurde.

Die GPD baut inhaltlich auf dem „Status Quo Datenlinkage“[3] [3] auf und stellt dessen Weiterentwicklung dar. ▶ **Abb. 1** visualisiert den Gesamtprozess des Datenlinkage in Form eines Flussdiagrammes und dient als strukturiertes, prozessorientiertes Inhaltsverzeichnis der GPD. Es wurden sieben Leitlinien definiert, die konkrete Empfehlungen enthalten und auf weiterführende Literatur verweisen. Damit orientiert sich die GPD in ihrer Struktur an den genannten anderen „Guten Praxen“ [1, 2]. Im Einzelnen umfasst die GPD die folgenden Leitlinien:

1 Forschungsziele, Fragestellung, Datenquellen und Ressourcen
2 Dateninfrastruktur und Datenfluss
3 Datenschutz
4 Ethik
5 Schlüsselvariablen und Linkageverfahren
6 Datenprüfung/Qualitätssicherung
7 Langfristige Datennutzung für noch festzulegende Fragestellungen

Die Aktualität der Guten Praxis Datenlinkage wird durch das Autorenteam unter Mitarbeit der o. g. Arbeitsgruppen und weiterer Gremien im Zuge eines regelmäßigen wissenschaftlichen Austauschs geprüft. Bei Bedarf erfolgt eine Revision.

Leitlinie 1: Forschungsziele, Fragestellungen, Datenquellen und Ressourcen

Bei der Formulierung von Forschungsfragen und der Ausarbeitung des Forschungsdesigns sollen mögliche geeignete Datenquellen bzgl. ihrer Potenziale und Limitationen benannt und der erwartete Erkenntnisgewinn durch ein Datenlinkage dargestellt werden.

Das Forschungsdesign muss sich an den Forschungszielen oder a priori formulierten Forschungsfragen orientieren. Insofern muss eine Bewertung von z. B. zu erhebenden Primärdaten und zu verlinkenden Sekundärdaten in jedem geplanten Forschungsvorhaben fragestellungsspezifisch vorgenommen werden. Potenziale und Limitationen der einzelnen Datenquellen sollten dabei separat gegenübergestellt werden. In der zusammenfassenden Bewertung aller potenziell geeigneten Datenquellen sollte anschließend in Abwägung des damit verbundenen Aufwands eine Abschätzung des zu erwartenden Erkenntnisgewinns durch ein Datenlinkage von 2 oder mehr Datenquellen dargestellt werden.

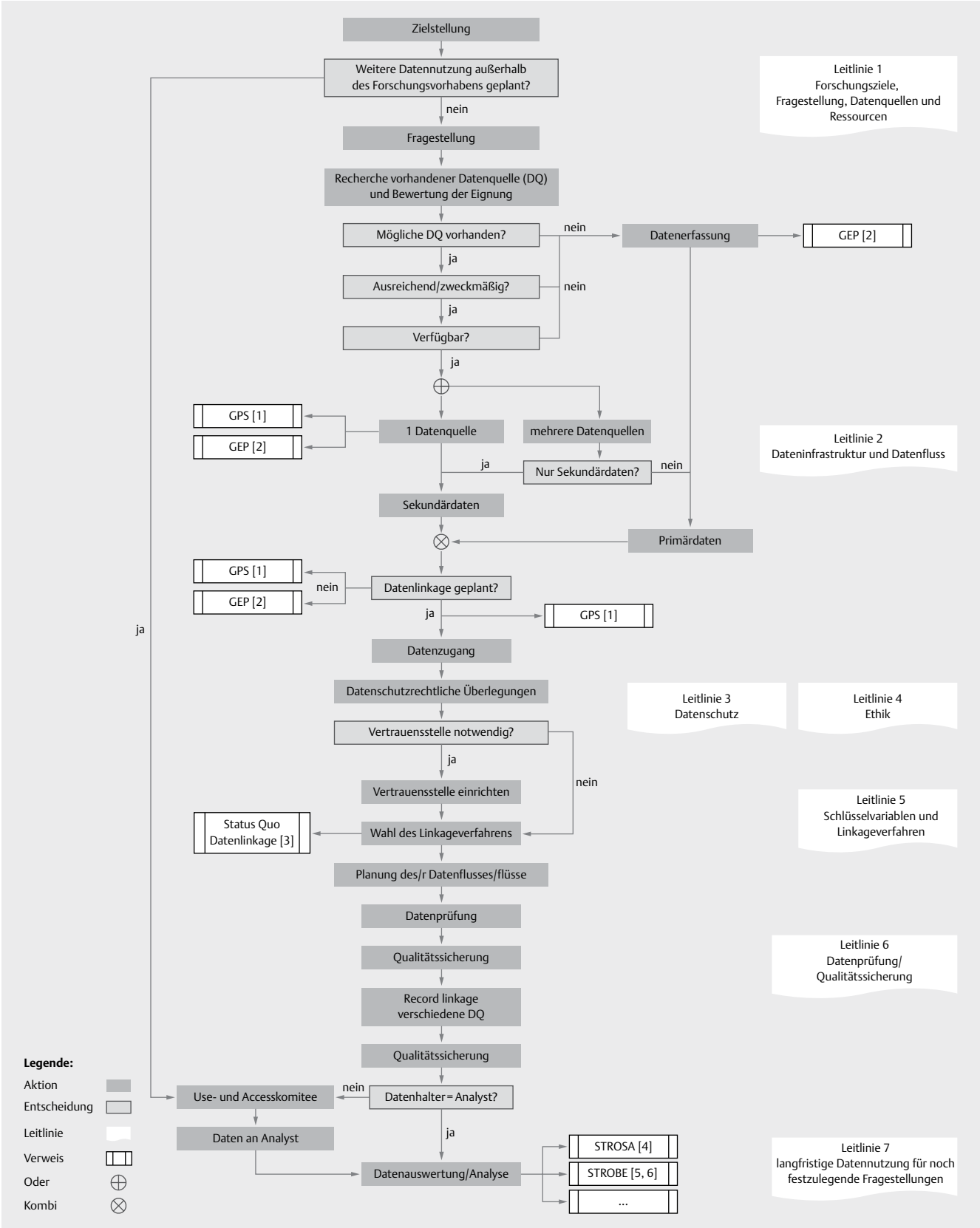
Empfehlung 1.1: Forschungsziele, Fragestellungen und ggf. Hypothesen müssen so präzise wie möglich formuliert werden, um darauf aufbauend ein detailliertes Anforderungsprofil für die zu nutzenden Daten zu erstellen

Forschungsziele und Fragestellungen müssen bei Forschungsvorhaben, in denen ein Datenlinkage vorgesehen ist, mit hoher Präzision formuliert werden, weil sie Anhaltspunkte für die Anforderungen an unterschiedliche Datenquellen liefern müssen. Daher sollte zunächst dargestellt werden, ob es nur Forschungsziele (wie z. B. bei der NAKO Gesundheitsstudie [7–9]) oder auch konkrete Forschungsfragen bzw. Hypothesen bei deduktiven Vorgehensweisen gibt (siehe GPS [1] und GEP [2]).

Damit Forschungsziele und Fragestellungen eine ausreichende Grundlage für die Auswahl passender Datenquellen oder für die Erhebung neuer Daten liefern, sind diese kleinschrittig und präzise zu operationalisieren. Erst dadurch wird es möglich, die Anforderungen an die zu nutzenden Datenquellen bezüglich ihrer enthaltenen Informationen und bezüglich der notwendigen Güte dieser Informationen zu spezifizieren (siehe Empfehlungen 1.2 und 1.5). Diese sollen in einem Daten-Anforderungsprofil („Was sollen die Daten können“) festgehalten werden, welches die Entscheidungsgrundlage für die Auswahl, ggf. Erhebung und die Art der Zusammenführung der zu nutzenden Daten bildet. Ein Anforderungsprofil ist somit eine Voraussetzung, um zielgerichtet passende Datenquellen aussuchen oder ggf. Neuerhebungen konzipieren zu können. Hierdurch wird auch erkennbar, ob ein Datenlinkage überhaupt notwendig ist (siehe Empfehlung 1.5). Zur Formulierung der Forschungsfrage und zur Konkretisierung des Datenlinkage soll zudem ein Entity-Relationship-Diagram erstellt werden.

2 Allgemeine Empfehlungen, die über das Thema Datenlinkage hinausgehen (z. B. zur allgemeinen Einhaltung datenschutzrechtlicher Bestimmungen) werden nicht gegeben. Hier wird auf die entsprechenden rechtlichen Vorgaben bzw. die GPS [1] verwiesen.

3 Im „Status Quo Datenlinkage“ [3] finden sich u. a. praktische Beispiele deutscher Studien, konkrete Verfahren und Arten des Datenlinkage, eine Auflistung verschiedener Softwaretools, konkrete Hinweise zur Qualitätssicherung sowie eine Checkliste der wichtigsten Fragen zum Datenlinkage.



▶ **Abb. 1** Gesamtprozess eines Forschungsvorhabens mit Datenlinkage.

Empfehlung 1.2: Es muss frühzeitig geprüft werden, ob die zur Verfolgung der Forschungsziele und Beantwortung der Forschungsfragen notwendigen Informationen verfügbar sind. Darüber hinaus müssen das Forschungsdesign, relevante Beobachtungszeiträume und die zu untersuchende Population spezifiziert werden

Ein wichtiger Schritt der Planung betrifft neben der konkreten inhaltlichen Ausgestaltung der Forschungsziele und Forschungsfragen die Überlegung, welche Daten zur Beantwortung der Fragestellung untersucht werden sollen. Damit gehen insbesondere eine Prüfung der Verfügbarkeit von benötigten Informationen sowie eine Abwägung der möglicherweise notwendigen Erhebung bislang fehlender Primärdaten einher. Bei der Verarbeitung personenbezogener Daten ist es gesetzlich vorgeschrieben zu überprüfen, ob diese erforderlich ist. Dabei muss der Grundsatz der Datenminimierung eingehalten werden (Artikel 5 Absatz 1 c Datenschutzgrundverordnung (DSGVO) [10]). Es muss demnach geprüft werden, welche Informationen und welche Datenquellen zwingend benötigt werden, um die Fragestellung sinnvoll und im Sinne der gesetzlich geforderten Datenminimierung ausreichend zu bearbeiten (siehe Empfehlung 1.1). Dem Vorhandensein und der Qualität von Schlüsselvariablen in den verschiedenen Datenquellen, als Voraussetzung für das Datenlinkage, kommt dabei ein besonderes Augenmerk zu (siehe Leitlinie 5). Ein weiterer wesentlicher Schritt betrifft die Festlegung und Benennung des Forschungsdesigns (z. B. Querschnitts-, Fall-Kontroll-, Kohorten-, Validierungsstudie), der relevanten Beobachtungszeiträume und der zu untersuchenden Studienpopulation (anhand von Ein- und Ausschlusskriterien) (siehe GPS [1] und GEP [2]).

Empfehlung 1.3: Unter dem Aspekt Datenherkunft (Data provenance) müssen die Dateneigner Entwicklungen der Datenquellen offenlegen und kommunizieren. Insbesondere sind historische/systematische Veränderungen an den Datenquellen, wie die Einführung neuer Kodierungssysteme oder von Schlüsselvariablen, zu kommunizieren

Datenkörper unterliegen häufig zeitlichen Veränderungen und können damit systematische Veränderungen beinhalten [11]; bspw. kann ein neues Kodierungssystem eingeführt worden sein. Solche systematischen Veränderungen sollten recherchiert, kommuniziert und für das Datenlinkage dokumentiert werden [12]. Idealerweise wird dies ebenfalls in dem Entity-Relationship-Diagram dokumentiert.

Empfehlung 1.4: Die potenziell zu verlinkenden Datenquellen müssen hinsichtlich ihrer Entstehung, ihres ursprünglichen Verwendungszwecks, ihres Dateneigners und ihrer Vor- und Nachteile beschrieben werden

Grundsätzlich lassen sich für ein Datenlinkage 2 Typen von Daten unterscheiden, die hinsichtlich ihrer Entstehung, ihres ursprünglichen Verwendungszwecks, ihres Dateneigners und ihrer Vor- und Nachteile beschrieben werden sollten:

1. Primärdaten, die im Rahmen ihres originär vorgesehenen Verwendungszwecks aufbereitet und analysiert werden, z. B. Befragungsdaten.
2. Sekundärdaten, die einer Auswertung über ihren originären, primären Verwendungszweck hinaus zugeführt werden. Hierzu zählen bspw. eine Vielzahl an Daten der Sozialversicherungsträger (z. B. Kranken- und Rentenversicherung), aber auch andere Leistungsdaten der gesundheitlichen Versorgung (z. B. aus der Arztpraxissoftware oder Krankenhausinformationssystemen) oder Daten von (klinischen) Forschungsvorhaben, die im Nachgang zur Beantwortung weiterer Fragestellungen genutzt werden. Eine ausführlichere Darstellung und Beispiele für verschiedene Arten des Datenlinkage finden sich bei Jacobs et al. [13], March et al. [3] und Swart et al. [14].

Darüber hinaus können weitere Datenquellen für ein Datenlinkage hinzugezogen werden: bspw. Daten des Instituts für Arbeitsmarkt- und Berufsforschung (IAB), Registerdaten wie z. B. Krebs- oder Mortalitätsregister oder Daten der Qualitätssicherung wie z. B. die der Disease Management Programme (DMP) oder der Landesärztekammern. Details zu diesen und weiteren Datenquellen finden sich u. a. bei Antoni et al. [15], Antoni & Seth [16], Czaplicki & Korbmacher [17], Kajüter et al. [18], Korbmacher & Czaplicki [19], Maier et al. [20], March et al. [21], March [22], Ohlmeier et al. [23, 24], Ohmann et al. [25], Swart et al. [26] und Stallmann et al. [27].

Die Beschreibung der Datenquellen bzw. die ausführliche Auseinandersetzung mit u. a. dem Prozess der Entstehung der zu verlinkenden Datenquellen wird als notwendig erachtet, da dieser z. B. einen Einfluss auf die Validität der enthaltenen Daten und somit auf die potenziell zu berücksichtigenden Linkage-Variablen haben kann. Kenntnisse über die Entstehung der Daten sollten somit in die Entscheidung über das Linkageverfahren (z. B. fehlertolerant ja/nein, siehe auch Leitlinie 5, Empfehlung 5.3) einfließen.

Bei der Bewertung der einzelnen Datenquellen und ihrer Inhalte sollten diesbezügliche Qualitätskriterien wie bspw. Vollständigkeit und Fehlerfreiheit [28] berücksichtigt werden [29]. Aus epidemiologischer Sicht sollten zudem Gütekriterien wie Objektivität, Reliabilität und Validität Beachtung finden. Je nach der Struktur der Daten [30] kann die Auswahl ergänzender Qualitätskriterien wie der Kontext der Erfassung, Aktualität, Zeit- und Personenbezug sowie ggf. Möglichkeiten der internen Validierung, dazu dienen, den erwarteten Erkenntniszugewinn durch ein individuelles Datenlinkage zu verdeutlichen (siehe Empfehlung 1.1) [29, 31].

Empfehlung 1.5: Die Durchführung eines Datenlinkage muss durch die Fragestellung begründet sein

Die Durchführung eines Datenlinkage begründet sich durch die Fragestellung (siehe Empfehlung 1.1), die ohne Datenlinkage nicht adäquat zu beantworten ist. Ist kein weiterer Erkenntnisgewinn durch das Datenlinkage zu erwarten, ist aus Gründen der Datenminimierung von einem Datenlinkage abzusehen (siehe Leitlinie 4). Ein Datenlinkage ist dann sinnvoll, wenn bspw. in einer Datenquelle alleine nicht alle benötigten Informationen enthalten sind oder die Information durch eine andere Datenquelle validiert werden soll. Somit können z. B. die Schwachstellen der einen Datenquelle durch Hinzunahme einer weiteren Datenquelle ausgeglichen werden.

Erfolgt das Datenlinkage mit (partieller) Einwilligungserklärung, dann sind mögliche Verzerrungen durch Non-Response zu beachten, denn mit jeder weiteren Datenquelle, mit der verknüpft werden soll, steigt das Risiko des Selektionsbias, da ggf. nur eine bestimmte Auswahl an Studienteilnehmenden allen Zusammenführungen zustimmen wird. Da dadurch zunehmend die Repräsentativität gefährdet wird, muss hier eine sorgfältige Abwägung der Vor- und Nachteile des Datenlinkage erfolgen. Einschränkungen können zudem auftreten, wenn nicht alle avisierten Dateneigner trotz vorliegender Einwilligung der Teilnehmenden Daten zur Verfügung stellen (Beispiel siehe March et al. [32]).

Empfehlung 1.6: Aufgrund der Komplexität eines Forschungsvorhabens mit Datenlinkage müssen bei der Planung und Ausarbeitung des Designs frühzeitig ausreichende zeitliche, finanzielle und personelle Ressourcen vorgesehen werden

Die Planung eines Forschungsvorhabens mit Datenlinkage ist vielschichtig. Ausreichende zeitliche, finanzielle und personelle Ressourcen müssen bereitgestellt werden, um neben den inhaltlich-theoretischen (Vor-)Überlegungen ebenso allgemein-praktische Erfordernisse an die Durchführung der Forschungsvorhaben zu definieren. Hierzu gehört die möglichst frühzeitige Einbindung der Dateneigner, deren Daten zum Linkage herangezogen werden sollen. Oft lassen sich etwaige Probleme bezüglich Dateninfrastruktur und Datenfluss (siehe Leitlinie 2) oder Datenschutz (siehe Leitlinie 3) bereits im Vorfeld klären bzw. zumindest reduzieren. Ferner ist mit den Datenschutzbeauftragten der Dateneigner zu prüfen, wie die datenschutzrechtlichen Anforderungen eingehalten werden können. Ebenso ist ggf. das Einholen eines Ethikvotums einzuplanen (siehe GPS [1] und GEP [2]) und den Dateneignern zur Verfügung zu stellen.

Bei der Planung muss berücksichtigt werden, ob das Forschungsvorhaben mit Datenlinkage mit oder ohne Einwilligungserklärung der Teilnehmenden durchgeführt werden soll bzw. ob eine (ergänzende) Einholung einer Einwilligungserklärung erwogen wird (siehe Leitlinie 3, Empfehlung 3.2). Schließlich kann es bei der Verwendung von Sozialdaten nach § 67 Absatz 2 Sozialgesetzbuch (SGB) X [33] im Rahmen eines Forschungsvorhabens notwendig sein, für eine Antragstellung nach § 75 SGB X [33] weitere Ressourcen (siehe Leitlinie 3, Empfehlung 3.2) einzuplanen.

In der Studienplanung sollten die technischen Voraussetzungen für das Datenmanagement/Datenlinkage abgeschätzt und mit den vorhandenen Gegebenheiten und Ressourcen abgeglichen werden. Die bei Bedarf rechtzeitige Anschaffung und Inbetriebnahme erforderlicher Hard- und/oder Software kann zum reibungslosen Studienablauf beitragen.

Die vorzeitige Klärung aller organisatorischen, technischen und rechtlichen Aspekte hilft abschließend bei der Entscheidung für, aber auch möglicherweise gegen ein Datenlinkage, wenn die notwendigen Voraussetzungen zum Datenlinkage nicht erfüllt werden können oder der zu erwartende Erkenntnisgewinn durch ein Datenlinkage den Aufwand nicht rechtfertigt.

Weiterführende Hinweise zu beispielhaften Studien finden sich im Status Quo Datenlinkage [3], bei Stallmann et al. [27] (NAKO Gesundheitsstudie [34]), bei March et al. [21, 22] (lidA-Studie) und bei Swart et al. [35] (AGil-Studie).

Leitlinie 2: Dateninfrastruktur und Datenfluss

Datenübermittlung und -verarbeitung in einem Forschungsvorhaben mit Datenlinkage finden auf der Basis einer technischen Infrastruktur statt, die hinsichtlich ihrer Zusammensetzung und der Konfiguration der einzelnen Bestandteile bestimmte Voraussetzungen erfüllen muss. Dies betrifft die an der Datenverarbeitung beteiligten Institutionen, die Datenübermittlung und die Datenhaltung.

Für die an der Datenverarbeitung beteiligten Institutionen müssen die spezifischen Rollen im Forschungsvorhaben sowie die Beziehungen untereinander definiert werden, insbesondere mit Blick auf:

1) die auszutauschenden Daten und weitere Informationen (z. B. Datensatzbeschreibung),
2) ggf. notwendige wechselseitige Verpflichtungen (z. B. zur fristgemäßen Datenlöschung) und
3) die verwendeten technischen Verfahren.

Hinsichtlich der Datenübermittlung bedeuten Vorhaben mit Datenlinkage, dass Übermittlungswege mit mehr als einem Dateneigner sowie mit ggf. zusätzlichen Stellen (Vertrauensstelle, zusätzliche Stelle zur Qualitätssicherung des Linkage usw., siehe Leitlinie 6) geplant und beschrieben werden müssen. Bei der Datenhaltung muss die Verwendung von Linkage-Software ebenso berücksichtigt werden wie der potenziell erhöhte Schutzbedarf der neu entstandenen gelinkten Daten, der sich auch in den Maßnahmen zum Datenschutz (bspw. in einer entsprechenden Sicherheitsinfrastruktur) niederschlagen kann.

Formal kann die Beschreibung der Infrastruktur sowie der Festlegungen und Prozesse, die zur Erfüllung der genannten Voraussetzungen nötig sind, im Rahmen eines Datenflusskonzeptes dargestellt werden. Hier werden alle Elemente in einem Flussdiagramm aufgeführt und weitergehend beschrieben.

Eine Beschreibung des Datenlinkage sollte auf Ebene der Variablen erfolgen und idealerweise von Data Dictionaries [36] ausgehen. Für die schematische Darstellung des Datenlinkage eignen sich bspw. Entity-Relationship-Diagrams, welche zur Dokumentation von Datenbankmodellen verwendet werden [37]. Die schematische Darstellung ermöglicht zudem den Abgleich von Schlüsselvariablen (siehe Leitlinie 5) sowie die Darstellung von Variableneigenschaften (z. B. des Datentyps).

Empfehlung 2.1: Der Datenfluss und die Verantwortlichkeiten müssen eindeutig festgelegt werden

Hinsichtlich der datenverarbeitenden Stellen empfiehlt sich zunächst eine Trennung zwischen den am Forschungsvorhaben beteiligten Institutionen und den im Vorhaben wahrzunehmenden Rollen (z. B. datenerhebende vs. datenverarbeitende vs. datenauswertende Stelle). So kann sich bspw. eine Institution A um die Erhebung/Bereitstellung personenidentifizierender Daten und deren Pseudonymisierung (z. B. in der Funktion einer Vertrauensstelle) kümmern, während Institution B die Verknüpfung der unterschiedlichen Datenquellen übernimmt. Eine derart getrennte Betrachtung kann dabei helfen, konzeptionelle Klarheit zu gewinnen und noch bestehende Lücken im Ablauf sichtbar zu machen.

In Vorhaben mit Datenlinkage sind typischerweise folgende Rollen vorzufinden:

- datenerhebende Stellen, die für die Datenerfassung zuständig sind,
- verschiedene Dateneigner (abhängig von der Art der gelieferten Daten),
- eine Vertrauensstelle, die inbesondere personenbezogene Daten pseudonymisiert,
- Stellen, die personenbezogene Daten anonymisieren,
- Stellen, die das Datenlinkage selbst durchführen,
- Stellen, die die Qualitätskontrolle des Linkage durchführen,
- Stellen, die die Datenauswertung durchführen.

Der Datenzugriff hat entsprechend rollengebunden zu erfolgen. Die Rollen können sich, je nach Forschungsvorhaben, auf eine oder mehrere Institutionen verteilen.

Datenflusskonzepte aus Vorhaben mit unterschiedlicher Komplexität finden sich bspw. bei Hassenpflug & Liebs [38], Jacobs et al. [13], Pommerening et al. [39] und Swart et al. [35].

Empfehlung 2.2: Die allgemeinen technischen und organisatorischen Anforderungen an die Datenübermittlung (siehe Leitlinie 6.1 der GPS) müssen beachtet und Besonderheiten bei Vorhaben mit Datenlinkage berücksichtigt werden

Grundsätzlich ist zu klären, ob die verwendeten Daten als Datensätze von Stelle zu Stelle übermittelt werden oder ob auf zumindest einige Daten lediglich ein Datenzugriff (z. B. über Virtual Private Network [VPN] oder Remote Desktop Protocol [RDP]) erfolgt. Bei Projekten mit Datenlinkage handelt es sich häufig um komplexe Vorhaben, bei denen dem Schutz der Daten besondere Aufmerksamkeit zu widmen ist (siehe Leitlinie 3). So muss durch geeignete Verschlüsselungs- und Pseudonymisierungsverfahren sichergestellt werden, dass zu keinem Zeitpunkt unzulässige Datenzugriffe (wie die Zusammenführung von direkt identifizierenden und medizinischen Daten) erfolgen können. Dies kann z. B. durch multiple Pseudonymisierung, durch die getrennte Verschlüsselung von Schlüsselvariablen und medizinischen Daten oder durch den Einsatz verschlüsselter Identifikatoren (z. B. Bloom-Filter, siehe Leitlinie 5) erreicht werden.

Empfehlung 2.3: Je nach Record Linkage-Verfahren muss eine geeignete Software eingesetzt werden

Ein direktes Record Linkage anhand vorhandener, eindeutiger Schlüsselvariablen kann in der Regel in der Datenverarbeitungssoftware selbst stattfinden. Für komplexere Verfahren geben March et al. [3] einen detaillierten Überblick über derzeitig verfügbare Software für sowohl exaktes als auch fehlertolerantes Linkage.

Empfehlung 2.4: Für die Löschung von Daten und das Widerspruchsmanagement muss ein geeigneter Prozess definiert werden

Zum einen ist der Umgang mit dem vollständigen Datensatz am Ende des Forschungsvorhabens zu regeln. Je nach Planung kann dies entweder die Löschung oder die Anonymisierung des Datenbestandes bedeuten. Dies betrifft den vollständigen Datenbestand bei allen beteiligten Institutionen und Rollen im Datenfluss, auch bei den Institutionen, die verschlüsseln oder pseudonymisieren. Zum anderen sind Besonderheiten der verwendeten IT-Infrastruktur zu berücksichtigen. Letzteres ist bspw. relevant, wenn Daten auch aus Datensicherungen entweder direkt oder durch zyklisches Überschreiben gelöscht werden müssen.

Für das Widerspruchsmanagement bedeutet dies, eindeutig definierte Prozesse für die Löschung individueller Datensätze bzw. für die Löschung bestimmter Merkmale aus individuellen Datensätzen vorzusehen (siehe auch Empfehlung 3.3).

Weiterführende Hinweise zu beispielhaften Studien finden sich im Status Quo Datenlinkage [3] sowie u. a. bei Jacobs et al. [13] und Swart et al. [35].

Leitlinie 3 Datenschutz

Wie bereits in den Leitlinien 1 und 2 ausführlich erläutert, kann schon allein die Verwendung von Primär- oder Sekundärdatenquellen die besondere Beachtung von Datenschutzvorgaben bedingen. Bei der Verknüpfung solcher Datenquellen ist mit einem nochmals erhöhten Schutzbedarf zu rechnen. Dabei müssen alle relevanten Personen frühzeitig in die Planung einbezogen werden. Dazu zählen u. a. neben den eigentlichen Dateneignern auch die internen oder externen Datenschutzbeauftragten der beteiligten Stellen und ggf. die zuständigen Aufsichtsbehörden [3, 21, 22, 39–41]. Je nach Datenquelle gibt es unterschiedliche Vorgaben oder Anträge auf deren Nutzung [42].

Empfehlung 3.1: Datenschutzrechtliche Vorgaben müssen bereits in der Planung und bis zum Abschluss des Vorhabens berücksichtigt und eine Deanonymisierung/Reidentifizierung einzelner Personen durch das Linkage verhindert werden

Je mehr Daten miteinander verknüpft werden, desto höher ist das Risiko einer Reidentifikation einer natürlichen Person. Über den Datenschutz hinaus kann die Identifizierung juristischer Personen auch andere Schutzinteressen betreffen, bspw. wettbewerbsrelevante Informationen über Arztpraxen, Krankenhäuser oder Krankenkassen.

Es muss sichergestellt sein, dass die Umsetzung angemessener Datenschutzmaßnahmen an jeder Stelle der Übermittlung und Verarbeitung gewährleistet ist. Aus diesem Grunde kommt datenschutzkonformen Verfahren sowie der Festlegung von Verantwortungen und Zuständigkeiten zur Verarbeitung, Speicherung und dem Transport der Daten eine besondere Rolle zu. Idealerweise werden Standard Operating Procedures (SOPs) zu folgenden Themen formuliert:

- Datenschutz und Datensicherheit
- Ethische und rechtliche Regelungen zu Datenzugang und -nutzung/sofern vorgesehen Use- und Access-Regelungen
- Aufbau und Pflege der Datenbank(en)
- Datenübermittlung und Datenlöschung.

Empfehlung 3.2: Es muss geprüft werden, ob eine Einwilligungserklärung notwendig ist

Forschungsvorhaben, die personenbezogene Daten verwenden, benötigen im Regelfall eine Einwilligungserklärung, einen sogenannten informed consent. Für Sozialdaten regelt dies der § 67b

SGB X bzw. § 75 SGB X [33]. In der Einwilligungserklärung müssen die Studienteilnehmenden nach angemessener Information explizit dem geplanten und beschriebenen Datenlinkage zustimmen [3, 40].

Zudem muss beachtet werden, dass die Verwendung von Angaben wie der Rentenversicherungsnummer oder der lebenslangen individuellen Krankenversichertennummer in der Regel einer Einwilligung durch den Betroffenen bedarf [27]. Es kann Ausnahmen von der Notwendigkeit einer Einwilligungserklärung geben (siehe hierzu [3]).

Zusätzlich muss in die Einwilligungserklärung mit aufgenommen werden, unter welchen Bedingungen oder wann Daten gelöscht werden. Es muss sichergestellt sein, dass die Einwilligung jederzeit widerrufen werden kann, was im Regelfall zur vollständigen Löschung aller noch nicht anonymisierten Daten der widerrufenden Person führen muss.

Diese Ausführungen für eine wissenschaftliche Nutzung von personenbezogenen Daten und deren Verlinkung gelten gleichermaßen für Forschungsvorhaben, die einen direkten Kontakt mit den Studienteilnehmenden vorsehen. Auch bei diesen ist im Regelfall eine Einwilligung zur Verlinkung verschiedener personenbezogener Datensätze einzuholen.

Ferner sollten die unter 4.1. aufgeführten Empfehlungen berücksichtigt werden.

Empfehlung 3.3: Ein Datenschutzkonzept muss erarbeitet werden

Zusätzlich muss zu jedem Projekt ein separates Datenschutzkonzept erarbeitet werden. In diesem sollen die Datenflüsse und Aufgaben, Pflichten und Verantwortlichkeiten aller am Projekt Beteiligten schriftlich geregelt werden [3]. Im Detail müssen die folgenden Angaben enthalten sein:

- Beschreibung des Projekts (Hintergrund, Ziel, Datenbasis und Methodik)
- Verantwortlichkeiten (welche öffentlichen und nichtöffentlichen Stellen sind involviert)
- Nennung der Personen, die Zugang zu den Daten haben (Vertrauensstelle/Treuhandstelle, Forschende)
- Nennung der betroffenen Personen und deren verwendeten Daten und/oder Datenkategorien (insbesondere die Schlüsselvariablen)
- Rechtsgrundlagen
- Datenbezogene Prozesse und die dadurch entstehenden Risiken bzw. Schutzbedarf und Vertraulichkeit
- Organisatorische und technische Maßnahmen oder Verfahren
- Fristen
- Konkretes Vorgehen beim Löschen der Daten, inklusive der Klärung, ab wann Daten nicht mehr gelöscht werden können/sollen z. B. aufgrund von Anonymität der Daten (siehe auch Empfehlung 2.4)
- Widerrufsmanagement: Festlegung eines Verfahrens zur Löschung einzelner Datensätze bei entsprechender Aufforderung durch einen Teilnehmenden (siehe auch Empfehlung 2.4).

Die GPS [1], GEP [2] und Deutsche Forschungsgemeinschaft (DFG) [43] empfehlen, die Daten für 10 Jahre nach Beendigung der Studie aufzubewahren. Entsprechend sind Daten, wenn möglich, erst nach dieser Frist zu löschen[4]. Für diese Dauer ist eine geeignete Speicherung der Daten sicherzustellen und ggf. auch vertraglich zu regeln [1].

Das Projekt MOSAIC des Instituts für Community Medicine der Universitätsmedizin Greifswald stellt eine Mustervorlage zur Erstellung eines Datenschutzkonzeptes bereit[5] [44]. Als Teil des Datenschutzkonzeptes ist bei umfangreicher Verarbeitung von Gesundheitsdaten eine Datenschutz-Folgenabschätzung nach DSGVO [10] notwendig. Generische konzeptuelle Vorgaben, die zudem mit allen relevanten Datenschutzbehörden abgestimmt wurden, finden sich bei Pommerening et al. [39].

Empfehlung 3.4: Wird ein Linkage erst im Nachhinein in einem Forschungsvorhaben geplant, müssen die sich ggf. ergebenen datenschutzrechtlichen Vorgaben sorgfältig geprüft werden

Wird erst im Laufe eines Forschungsvorhabens angestrebt, zusätzliche Daten mit den bereits bestehenden zu verknüpfen, sollten ebenfalls die Datenschutzbeauftragten der beteiligten Einrichtungen kontaktiert werden. Zudem ist zu prüfen, ob eine Verknüpfung unter Beachtung der vorliegenden Einwilligung möglich ist, da ggf. eine spätere Verknüpfung von der Einwilligung nicht abgedeckt oder gar von vornherein ausgeschlossen worden ist. Dies kann z. B. auch ein indirektes Datenlinkage zu einem späteren Zeitpunkt betreffen [42]. Es kann notwendig sein, dass Teilnehmende im Nachgang (erneut) um ihre Einwilligung in das Linkage ihrer Daten gebeten werden müssen. In diesem Fall muss zusätzlich geklärt werden, ob eine Einwilligung vorliegt, die Teilnehmenden überhaupt erneut kontaktieren zu dürfen, um ihre Einwilligung zum Linkage einzuholen.

Leitlinie 4 Ethik

Über den Datenschutz hinaus hat die Verwendung verlinkter Datenquellen in der Regel auch Auswirkungen auf die ethische Bewertung des Forschungsvorhabens sowie auf die ggf. berufsrechtlich erforderliche Beratung durch die zuständige Ethikkommission oder ein Use- und Access-Komitee, welches auf Basis anderer Regularien eingebunden ist (siehe Leitlinie 7). Dies betrifft insbesondere den wissenschaftlichen Wert, die Qualität und ggf. auch die Originalität des Vorhabens, wenn durch die Zusammenführung bisher nur getrennt oder gar nicht analysierter Daten neue Erkenntnisse geschaffen werden. Mit denselben Argumenten kann unter Umständen auch eine erhöhte Praxisrelevanz begründet werden. Die Zusammenführung verschiedener Datenquellen kann damit Auswirkungen auf das Nutzen-Schadenpotenzial des Vorhabens haben, wobei dem potenziellen Schaden durch ein ggf. höheres Reidentifizierungspotenzial infolge der Zusammenführung der Daten auch ein potenziell erhöhter Nutzen durch die neuen Erkenntnisse entgegenstehen kann.

4 Für das "Löschkonzept" kann man sich an der DIN 66398 (https://www.datenschutzbeauftragter-info.de/din-norm-66398-die-entwicklung-eines-loeschkonzepts/; Zugriff am 04.06.2019) orientieren, in der die abzuhandelnden Punkte aufgeführt werden.

5 https://www.toolpool-gesundheitsforschung.de/produkte/vorlage-datenschutzkonzept; Zugriff am 19.07.2019

Empfehlung 4.1: Mögliche Auswirkungen des Datenlinkage auf das Nutzen-Schadenpotenzial des Forschungsvorhabens müssen geprüft werden

Bei der Durchführung eines Datenlinkage sind aus ethischer Sicht die folgenden Aspekte zu beachten:

- Minimierung von fehlerhaftem Linkage und daraus resultierenden falschen Ergebnissen,
- Minimierung des Risikos der Reidentifizierung natürlicher Personen (siehe Leitlinie 3).

Im Hinblick auf die Gefahr der Reidentifizierung sollte unter ethischen Gesichtspunkten diskutiert werden, inwiefern Mechanismen in der Datenhaltung genutzt werden können, die eine Zusammenführung verschiedener Datenquellen nur zum Zeitpunkt der Analyse und nur mit den zur Analyse zwingend erforderlichen Merkmalen gestatten (siehe Leitlinie 7). Bei der Erstellung von Informationen für die Teilnehmenden und der Einwilligung muss darauf geachtet werden, dass jede Datenquelle (Datenkategorie) separat beschrieben wird. Die Information der Teilnehmenden muss dabei so gestaltet sein, dass jede Person in der Lage ist, die Art und Auswirkungen des Linkages zu verstehen.

Leitlinie 5: Schlüsselvariablen und Linkageverfahren

Beim Datenlinkage spielen Schlüsselvariablen und die verschiedenen Verfahren, mit denen Datensätze verknüpft werden sollen, eine zentrale Rolle. Auswahl und Anwendung von Schlüsselvariablen und Linkageverfahren stehen in wechselseitiger Beziehung zueinander und sind daher aufeinander abzustimmen.

Unter Schlüsselvariablen werden Variablen verstanden, die in allen zu verknüpfenden Datensätzen vorkommen und damit die Zuordnung ermöglichen. Linkageverfahren sind technische Verfahren, die dazu dienen, die Datenquellen über Schlüsselvariablen zu verknüpfen. Nach Verknüpfung der Datensätze ist zu prüfen, ob die Schlüsselvariablen weiterhin erforderlich sind oder in dem neuen Datensatz gelöscht werden können (Prinzip der Datenminimierung).

Empfehlung 5.1: Vor Definition und Nutzung von Schlüsselvariablen müssen die bestehenden Rahmenbedingungen für ihren Einsatz für das Linkage geklärt werden

Die Auswahl eines geeigneten Linkageverfahrens inklusive der dabei zu nutzenden Schlüsselvariablen hängt von verschiedenen Rahmenbedingungen ab:

1) Die zu beachtenden (datenschutz-)rechtlichen Vorgaben müssen geklärt sein. Besondere Bedeutung kommt dabei der Frage zu, ob das Linkage auf Basis einer Einwilligungserklärung erfolgt, weil damit der Rahmen für die zur Verfügung stehenden Schlüsselvariablen gegeben ist. Im Sinne einer Risikobewertung ist zu untersuchen, ob durch das Linkage die Gefahr der Reidentifizierbarkeit von Personen steigt (siehe Leitlinie 3).
2) Die Art der Datenquelle beeinflusst die Qualität der Schlüsselvariablen. Im Gegensatz zu retrospektiv erhobenen Daten können in prospektiven Daten ggf. Variablen ergänzt (erhoben) werden, die ein anschließendes Linkage erst ermöglichen oder vereinfachen.
3) Es ist zu klären, zu welchen Zeitpunkten das Record Linkage erfolgen soll: automatisiert zum Zeitpunkt der Datenerhebung, in regelmäßigen zeitlichen Intervallen (z. B. pro Quartal) oder nach der für die Forschungsfrage abschließenden Erfassung aller Datenquellen. Insbesondere bei langfristigen Projekten wie Registern oder dem Aufbau von Forschungsdatenbanken ist der zeitliche Ablauf der Datenerhebung und Datenzusammenführung für alle Quellen zu beschreiben. Zudem sind die zu erhebenden Variablen inkl. der Schlüsselvariablen zu definieren.

Empfehlung 5.2: Alle Schlüsselvariablen müssen präzise definiert und hinsichtlich ihrer Fehleranfälligkeit und Vollständigkeit überprüft werden

Die Qualität und Vollständigkeit des Datenlinkage wird maßgeblich von den zur Verfügung stehenden Schlüsselvariablen bestimmt. Daher sollte deren Auswahl und Beschreibung besonderes Augenmerk geschenkt werden.

1. Automatisiert erfassten Variablen, die als Schlüssel verwendet werden können (z. B. spezifische Versichertennummer), ist dabei (sofern möglich) Vorrang zu geben. Der Einsatz und der Abgleich der Prüfziffern reduzieren Linkagefehler, die aufgrund fehlerhaft erhobener oder fehlerhaft übermittelter Schlüsselvariablen entstehen können.
2. Es ist zu prüfen, inwieweit direkt identifizierende Identifikatoren (z. B. Name oder Versichertennummer) als Schlüsselvariable in der Klartextform genutzt oder durch geeignete Verfahren zu maskieren sind (Pseudonymisierung, Hashfunktion, Bloom-Filter). Dabei ist zu berücksichtigen, inwieweit das gewählte Maskierungsverfahren für jede Datenquelle genutzt werden kann. Abhängig von den bestehenden Rahmenbedingungen und datenschutzrechtlichen Vorgaben kann diese Maskierung durch jeden Dateneigner, dessen Daten verknüpft werden sollen, in gleicher Weise (ggf. unter Nutzung eines Pseudonymisierungsdienstes wie der Mainzelliste [3, 45]) oder unter Einbeziehung einer/s Treuhandstelle/Datentreuhänders erfolgen [39]. Weiterführende Informationen hierzu sind im Status Quo Datenlinkage zu finden [3].
3. Um falsch negative (Synonymfehler) oder falsch positive Klassifikationen (Homonymfehler) zu minimieren, sind geeignete Verfahren einzusetzen. So können verschiedene Schreibweisen der Schlüsselvariablen harmonisiert (z. B. phonetische Kodierungsverfahren, Teilstrings, Bloom-Filter [46–49]) sowie Zuordnungsfehler durch Einbeziehung weiterer Merkmale reduziert werden. Können sich Schlüsselvariablen über die Zeit ändern (z. B. Name bei Heirat), so sind entsprechende Vorkehrungen für eine weiterführende Zuordnung zu treffen (z. B. Übersetzungstabelle von alter auf neue ID oder Einbeziehung des Geburtsnamens – siehe auch Empfehlung 6.2).

Empfehlung 5.3: Für das Datenlinkage muss ein geeignetes technisches Verfahren gewählt werden

Für das Linkage sollen in Abhängigkeit von den Möglichkeiten und Zielsetzungen des Vorhabens angepasste Verfahren eingesetzt

werden. Diese werden im Folgenden nur kurz erwähnt, werden aber im Status Quo Datenlinkage [3] ausführlich beschrieben.

1. Beim Linkage wird unterschieden zwischen direktem vs. indirektem Linkage und zwischen probabilistischem vs. deterministischem Linkage. Zudem gibt es exakte und fehlertolerante Verfahren, die wiederum in regelbasierte und distanzbasierte fehlertolerante Verfahren unterschieden werden. Auch können Blocking-Verfahren, bei denen nur Datensätze mit den gleichen Ausprägungen spezifischer Merkmale verglichen werden, eine wichtige Rolle spielen, denn sie können die Performanz des Linkage-Prozesses steigern, jedoch auch die Qualität des Linkages negativ beeinflussen.
2. Alle Verfahren haben Vor- und Nachteile, die vor ihrem Einsatz berücksichtigt werden müssen. Sie können unter bestimmten Bedingungen miteinander verknüpft werden. Es wird empfohlen, frühzeitig IT-Verantwortliche und Experten für die Wahl und Umsetzung des Linkageverfahrens einzubinden.
3. Enthalten die zu verknüpfenden Datensätze Klartextidentifikatoren oder pseudonymisierte Klartextidentifikatoren als Schlüsselvariablen, können direkte Linkageverfahren eingesetzt werden. Die Nutzung pseudonymisierter Identifikatoren ist jedoch nur möglich, wenn die zu linkenden Datensätze Schlüsselvariablen enthalten, die nach dem gleichen Verfahren pseudonymisiert wurden.
4. Sollte das Ergebnis des direkten Linkage nicht zufriedenstellend sein, z. B. weil durch fehlerhafte Schlüsselvariablen Datensätze nicht verlinkt werden konnten, kann zusätzlich ein indirektes Verfahren erfolgen (siehe Leitlinie 6).
   Enthalten die zu verlinkenden Datensätze keine Schlüsselvariablen, so kann geprüft werden, inwieweit die Datensätze identische Variablen enthalten, die als Schlüsselvariablen herangezogen werden können und somit ein indirektes Linkage ermöglichen.

Weiterführende Hinweise und beispielhafte Studien finden sich im Status Quo Datenlinkage [3].

Leitlinie 6: Datenprüfung/ Qualitätssicherung

Im Rahmen der Planung und Vorbereitung eines Datenlinkage sind Aspekte zur Sicherung der Datenqualität zu berücksichtigen. Die Notwendigkeit entsprechender Aktivitäten beschränkt sich dabei nicht auf einen separaten Prozessschritt des Datenlinkage, sondern gilt von der Qualitätssicherung der einzelnen Datensätze über den eigentlichen Prozess des Record Linkage bis hin zur Plausibilisierung bzw. Qualitätssicherung des gelinkten Datensatzes, also für alle zu durchlaufenden Schritte des Datenlinkage.

Neben den technischen Voraussetzungen sollten ausreichend personelle Ressourcen für die Prüfung des erfolgreichen Datenlinkage sowie die Aufbereitung der Daten eingeplant werden. Beispielsweise kann nach erfolgtem Datenlinkage ein Abgleich mit den übergebenen Daten [50] (siehe zudem Empfehlung 6.3) notwendig sein bzw. Klartextangaben klinischer Befunde müssen, vorab zu einer Auswertung, noch standardisiert kodiert werden.

Empfehlung 6.1: Eine Beschreibung der Qualität der Schlüsselvariablen muss im Ergebnisbericht enthalten sein

Der Linkage-Erfolg hängt entscheidend von der Ausgangsqualität der Schlüsselvariablen ab. Die Validität der Schlüsselvariablen kann sich in verschiedenen Datenquellen deutlich unterscheiden, z. B. manuell erfasste Personenangaben von handschriftlich ausgefüllten Todesbescheinigungen vs. elektronisch übermittelte Sterbefallangaben aus Einwohnermeldeämtern. Darüber hinaus kann es notwendig sein, dass in die Plausibilitätskontrollen auch Datenfelder einfließen, die nicht Bestandteil der Schlüsselvariablen sind (z. B. Alter, Geschlecht, Gewicht oder Körpergröße), wenn damit fehlerhafte Zuordnungen aufgedeckt werden (Überblick siehe [51]). Wenn möglich, sollte eine Validierungsstudie zur Abschätzung der Qualität der Schlüsselvariablen durchgeführt werden. Das Vorgehen beim Datenlinkage und das Ergebnis des Datenlinkage sollten abschließend explizit im Bericht des Forschungsvorhabens beschrieben und bewertet werden.

Empfehlung 6.2: Es muss geprüft werden, ob ein iteratives Vorgehen zu einer besseren Linkagequalität führt

Der Anteil und die Art falscher Record Linkage-Ergebnisse (Synonym- oder Homonymfehler) hängen u. a. ab von der Schnittmenge derselben Personen in den zu verlinkenden Datensätzen, der Wahrscheinlichkeit zufällig übereinstimmender Schlüsselvariablen und der Wahrscheinlichkeit abweichender Angaben bei derselben Person (außer Fehlern z. B. auch Namensänderung nach Heirat, Wohnortwechsel nach Umzug u. a.). Abschätzungen zu diesen Parametern sollten erfolgen. Da diese Informationen bei der Planung einer Untersuchung oft noch nicht bekannt sind, können eventuell Einzelheiten zum Datenlinkage zu diesem Zeitpunkt noch nicht endgültig festgelegt werden. Das iterative Vorgehen sowie das vorgesehene Prozedere und die Höhe einer maximal noch zu tolerierenden Rate an Fehlzuordnungen sollten dennoch vorher, z. B. im Studienprotokoll, festgelegt werden.

Empfehlung 6.3: Im Rahmen der Qualitätssicherung müssen von der auswertenden bzw. der das Datenlinkage durchführenden Stelle Rückfragen an den Dateneigner möglich sein. Implausibilitäten müssen mit dem Dateneigner abgeklärt werden, um inkonsistente Daten bzw. eine fälschliche Interpretation der Daten zu vermeiden

Eine direkte und unmittelbare Einsichtnahme in die unverschlüsselten Daten der Dateneigner kann im Regelfall aus Datenschutzgründen nicht vorgenommen werden, insbesondere bei Sozialdaten. Aus diesem Grund sollten im Rahmen vertraglicher Regelungen zwischen der auswertenden bzw. der das Datenlinkage durchführenden Stelle Absprachen getroffen werden, in welcher Form eine Unterstützung durch den Dateneigner bei der Klärung von Rückfragen zu den Originaldaten erfolgen kann. So sollte z. B. im Falle einer Unklarheit die auswertende bzw. die das Datenlinkage durchführende Stelle die entsprechenden (Detail-)Fragen zum Datensatz an den Dateneigner richten, der dann innerhalb seiner (nicht pseudonymisierten) Originaldaten eine Prüfung durchfüh-

ren kann. In Abhängigkeit vom Prüfergebnis können weitere Validierungsschritte geplant oder ggf. ein modifizierter neuer Datensatz zur Verfügung gestellt werden. Solche Ergebnisse und entsprechende Anpassungsmaßnahmen könnten z. B. auch das Ergebnis einer vorab geplanten Validierungsstudie zur Bestimmung der Synonym- und Homonymfehlerrate sein.

Es ist zu berücksichtigen, dass auch Dateneigner (z. B. Krankenkassen) teilweise keine Auskünfte über Datensätze bekommen, die sie nicht direkt vom Leistungserbringer (z. B. dem Arzt im Rahmen der vertragsärztlichen Versorgung) erhalten, sondern indirekt über „Dritte" (z. B. Kassenärztliche Vereinigungen) geliefert bekommen und deshalb ihrerseits ggf. nur begrenzt auskunftsfähig sind.

Empfehlung 6.4: Nach jedem Datenlinkage muss die Zahl der zusammengeführten und der nicht zusammenführbaren Datensätze auf Basis der Ausgangsdateien überprüft werden

Hierfür ist im Vorfeld eine Abschätzung vorzunehmen, wie häufig erfolgreiche Verknüpfungen für die einzelnen Dateien auftreten müssten. Zudem sollten beobachtete Häufigkeitsverteilungen nach erfolgtem Linkage auf Plausibilität überprüft werden. Es ist zu überprüfen, ob die nicht verlinkten Datensätze eine spezifische Struktur aufweisen, die entweder für die fehlgeschlagene Verknüpfung verantwortlich ist (z. B. Quelle A ist aktuell und in Quelle B liegt der Datensatz noch nicht vor) oder die eine systematische Stichprobenverzerrung in dem verlinkten Datensatz bewirken könnte. Aus diesem Grund sollten die wesentlichen Merkmale im verlinkten und den zugrundeliegenden nicht verlinkten Datensätzen untersucht werden, um strukturelle Unterschiede zu verifizieren.

Empfehlung 6.5: Nach jedem Datenlinkage muss ein Abgleich zwischen den übergebenden und zusammengeführten Daten erfolgen

Die Zusammenführung von Datenquellen aus unterschiedlichen Betriebssystemen oder Software-Applikationen kann zu unbemerkten Fehlern oder Verzerrungen oder sogar zu Verlust von Daten führen. So können (i) Sonderzeichen durch eine falsche Kodierung falsch dargestellt werden, z. B. Abspeichern und Einlesen von Zeichenkodierungen, (ii) Datentypen/ -formate im Zielsystem entweder nicht existent oder falsch interpretiert werden oder (iii), aufgrund begrenzter Feldlängen, Trunkierungen der Zeichenlänge in Variablen erfolgen. Ein Abgleich zwischen den übergebenden und zusammengeführten Daten zur Prüfung solcher Verzerrungen oder Verluste kann stichprobenartig erfolgen. Insofern an einer Datenquelle systematische Veränderungen über die Zeit durchgeführt wurden, sollte dieser Datenabgleich diese unterschiedlichen Episoden berücksichtigen [50].

Empfehlung 6.6: Die tatsächlich erreichte Fehlerrate muss gemessen werden und im Ergebnisbericht enthalten sein. Bei mehrmals stattfindendem Linkage ist die Fehlerrate kontinuierlich zu prüfen und mit vorherigen Ergebnissen abzugleichen

Ist ein Linkage mehrmals vorgesehen, z. B. wenn jährlich neue Daten vorliegen, kann sich im Laufe der Zeit auch die Fehlerrate verändern. Dies kann einerseits an dynamische Daten, wie beispielsweise Daten von einem Krebsregister, welche immer wieder aktualisiert werden, liegen. Andererseits kann es Hinweise auf mögliche Implausibilitäten in den Daten liefern oder auf einen bisher noch nicht aufgetretenen Fehler hindeuten. Bei jedem Linkage ist dieser Prozess der Qualitätssicherung deshalb erneut durchzuführen und zu dokumentieren (siehe Empfehlung 6.1).

Empfehlung 6.7: Nach jedem Datenlinkage muss die Beschreibung der Eigenschaften des entstandenen Forschungsdatensatzes mit Bezug auf die Originaldaten erfolgen

Die zu verknüpfenden Datenquellen könnten hinsichtlich der Anzahl von Beobachtungen (z. B. Probanden, Studienteilnehmenden, Versicherte) nicht übereinstimmen. Das erfolgreiche Datenlinkage stellt damit implizit auch einen Selektionsprozess dar, der zu verzerrten Ergebnissen führen kann [52]. Jede Form der Selektion von Beobachtungen sollte beschrieben und die damit einhergehenden Veränderungen der Eigenschaften der Originaldaten abgeschätzt werden.

Leitlinie 7: Langfristige Datennutzung für noch festzulegende Fragestellungen

Datenlinkage ist ein aufwendiger Prozess und generiert komplexe und informationsreiche Datenkörper. Diese weisen oftmals Analysepotenziale auf, die über die Auswertungsziele in zeitlich begrenzten Forschungsvorhaben hinausgehen. Zudem zielen manche Forschungsvorhaben von vornherein auf den Aufbau einer langfristig verfügbaren Forschungsdatenbank mit umfangreichen Datensätzen ab. Insofern kann eine weiterführende Datennutzung durch autorisierte Dritte im Rahmen a priori festgelegter Bedingungen erwogen werden, wenn dies im Einklang mit den datenschutzrechtlichen Bestimmungen umsetzbar ist.

Empfehlung 7.1: Ist eine weiterführende Nutzung der zusammengeführten Daten durch den Datenhalter über die primäre Fragestellung hinaus vorgesehen oder soll diese Möglichkeit prinzipiell bestehen, so müssen die dafür vorgesehenen Regularien bereits bei der Konzeption eines Forschungsvorhabens berücksichtigt werden

Mit dem Linkage verschiedener Datenquellen entsteht zumeist ein Datenpool, der deutlich mehr Forschungspotenzial birgt, als nur die Beantwortung der primären Fragestellung zu ermöglichen. Daher sollten bereits während der Konzeptionsphase Überlegungen erfolgen, ob und in welcher Form die Daten genutzt werden sollen. Wenn diese Überlegungen dazu führen, dass eine spätere und offenere Datennutzung möglich sein soll, so muss das sowohl im Ethikantrag als auch im Datenschutzkonzept sowie in den vertraglichen Regelungen mit den Dateneignern explizit Erwähnung finden und mit den Dateneignern vertraglich geregelt sein. Erfolgt ein einwilligungsbasiertes Datenlinkage, so muss die Einwilligungserklärung eine Öffnungsklausel enthalten, welche die Nutzung der Daten für mögliche weitere Fragestellungen legitimiert und darlegt, unter welchen Voraussetzungen eine weitergehende Nutzung geplant ist

Empfehlung 7.2: Sollen die zusammengeführten Daten im Rahmen einer Forschungsdatenbank der wissenschaftlichen Nutzung durch Dritte zugänglich gemacht werden, muss diese Nutzung durch ein standardisiertes Zugangsverfahren reglementiert werden

Zahlreiche epidemiologische und sozialwissenschaftliche Forschungsvorhaben zur Beantwortung inhaltlich und zeitlich nicht abschließend eingegrenzter Forschungsfragen sind von vornherein auf die Etablierung einer auf diesen Zweck ausgerichteten Forschungsdatenbank angelegt. Diese Datenbanken sollen die wissenschaftliche Nutzung der erhobenen Daten durch nicht unmittelbar an den Forschungsvorhaben Beteiligte ermöglichen.

Diese geplante Nutzung aufwendig gewonnener Forschungsdaten bedarf eines standardisierten und transparenten Antrags- und Bewilligungsverfahrens, bspw. im Rahmen einer sog. Use- und-Access-Ordnung, bei dem die Entscheidung über einen Nutzungsantrag in der Verantwortlichkeit eines Use- und-Access-Komitees liegt [39, 53]. Dabei ist von vornherein festzulegen, unter welchen Voraussetzungen eine Datennutzung möglich ist und welches Gremium diese Voraussetzungen prüft und entsprechende Freigaben erteilt oder verweigert.

Beispiele für Use- und-Access-Ordnungen dieser Art liefern die bevölkerungsbezogenen Forschungsvorhaben wie die bundesweite NAKO Gesundheitsstudie bzw. regionale Surveys und Kohortenstudien (z. B. SHIP-Studie)[6] sowie die Medizininformatik-Initiative des Bundesministeriums für Bildung und Forschung (BMBF)[7].

Aus den Forschungsvorhaben heraus können auch vorgefertigte Datenkörper aus den bereits zusammengeführten Primär- und Sekundärdaten in Form von Scientific Use Files angeboten werden. Zur Wahrung des Datenschutzes, der ethischen wie auch der fachlichen und technischen Belange [39, 53] ist für den Zugriff auf Daten hier ebenso ein standardisiertes und transparentes Zugangsverfahren für forschende Dritte zu etablieren. Auch bei anonymisierten Daten ist zu bedenken, dass diese ggf. ein Linkagepotenzial behalten und dann ein indirektes Linkage weiterhin möglich ist.

Das ist insbesondere von Bedeutung, wenn die herauszugebenen Daten mit weiteren Datenkörpern verknüpft werden sollen und so durch erhöhte Informationstiefe das Re-identifizierungsrisiko von Studienteilnehmenden steigt. Risikovariablen, die potenziell ein solches indirektes Linkage mit anderen verfügbaren Datensätzen als Schlüsselvariablen unterstützen, sollten daher nach Möglichkeit gelöscht oder ihre Inhalte ausreichend vergröbert werden.

Danksagung

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

6 Die Use- und-Access-Ordnungen sind in der jeweils aktuellen Fassung auf den Projektwebseiten zu finden: https://nako.de/allgemeines/der-verein-nako-e-v/rechtliche-grundlagen/; http://www2.medizin.uni-greifswald.de/cm/fv/ship/datennutzung/; Zugriff am 30.07.2019

7 https://www.medizininformatik-initiative.de/sites/default/files/inline-files/MII_03_Use_and_Access_Policy_Key_Issues_Paper_1-0.pdf; Zugriff am 19.07.2019

Interessenkonflikt

Die Autoren geben an, dass kein Interessenkonflikt besteht.

Literatur

[1] Swart E, Gothe H, Geyer S et al. Gute Praxis Sekundärdatenanalyse (GPS). Leitlinien und Empfehlungen. Gesundheitswesen 2015; 77: 120–126

[2] Hoffmann W, Latza U, Baumeister SE et al. Guidelines and recommendations for ensuring Good Epidemiological Practice (GEP). A guideline developed by the German Society for Epidemiology. Eur J Epidemiol 2019; 34: 301–317

[3] March S, Antoni M, Kieschke J et al. Quo vadis Datenlinkage in Deutschland? Eine erste Bestandsaufnahme. Gesundheitswesen 2018; 80: e20–e31

[4] Swart E, Bitzer EM, Gothe H et al. STROSA-STandardisierte Berichts-ROutine fur Sekundärdaten Analysen (STROSA) – ein konsentierter Berichtsstandard für Deutschland, Version 2. Gesundheitswesen 2016; 78: e145–e160

[5] Elm von E, Altman DG, Egger M et al. Das Strengthening the Reporting of Observational Studies in Epidemiology (STROBE-) Statement. Internist 2008; 49: 688–693

[6] Vandenbroucke JP, Elm von E, Altman DG et al. Strengthening the Reporting of Observational Studies in Epidemiology (STROBE). Explanation and elaboration. PLoS Med 2007; 4: e297

[7] Wichmann H-E, Kaaks R, Hoffmann W et al. Die Nationale Kohorte. Bundesgesundheitsbl 2012; 55: 781–787

[8] Ahrens W, Jöckel K-H. Der Nutzen großer Kohortenstudien für die Gesundheitsforschung am Beispiel der Nationalen Kohorte. Bundesgesundheitsbl 2015; 58: 813–821

[9] German National Cohort. The German National Cohort. Aims, study design and organization. Eur J Epidemiol 2014; 29: 371–382

[10] Datenschutz-Grundverordnung. Verordnung (EU) 2016/679 des Europäische Parlaments und des Rates vom 27. April 2016 zum Schutz natürlicher Personen bei der Verarbeitung personenbezogener Daten, zum freien Datenverkehr und zur Aufhebung der Richtlinie 95/46/EG (04.05.2016). Online https://publications.europa.eu/de/publication-detail/-/publication/3e485e15-11bd-11e6-ba9a-01aa75ed71a1/language-de letzter Zugriff: 19.07.2019

[11] Buneman P, Chapman A, Cheney J et al. A Provenance Model for Manually Curated Data. In: Moreau L, Foster I, (Hrsg.). Provenance and Annotation of Data. International Provenance and Annotation Workshop. IPAW 2006; Chicago, IL, USA: May 3-5, 2006: Revised Selected Papers, 2006: 162–170

[12] Bohensky MA, Jolley D, Sundararajan V et al. Development and validation of reporting guidelines for studies involving data linkage. Aust N Z J Public Health 2011; 35: 486–489

[13] Jacobs S, Stallmann C, Pigeot I. Verknüpfung großer Sekundär- und Registerdatenquellen mit Daten aus Kohortenstudien. Doppeltes Potenzial nutzen. Bundesgesundheitsbl 2015; 58: 822–828

[14] Swart E, Stallmann C, Powietzka J et al. Datenlinkage von Primär- und Sekundärdaten. Ein Zugewinn auch für die kleinräumige Versorgungsforschung in Deutschland? Bundesgesundheitsbl 2014; 57: 180–187

[15] Antoni M, Jacobebbinghaus P, Seth S. ALWA-Befragungsdaten verknüpft mit administrativen Daten des IAB (ALWA-ADIAB) 1975-2009. Aktualisierte Version vom 25.05.2012. FDZ Datenreport 05/2011. Nürnberg: Bundesagentur für Arbeit 2011

[16] Antoni M, Seth S. ALWA-ADIAB – Linked Individual Survey and Administrative Data for Substantive and Methodological Research. Schmollers Jahrbuch 2012; 132: 141–146

[17] 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., (Hrsg.). Gesundheit, Migration und Einkommensungleichheit. 2010. Aufl. Berlin: Deutsche Rentenversicherung Bund; 2010: 28–37

[18] Kajüter H, Geier AS, Wellmann J et al. Kohortenstudie zur Krebsinzidenz bei Patienten mit Diabetes mellitus Typ 2. Record Linkage von kryptografierten Daten einer externen Kohorte mit Daten des Epidemiologischen Krebsregisters Nordrhein-Westfalen. Bundesgesundheitsbl 2014; 57: 52–59

[19] Korbmacher JM, Czaplicki C. Linking SHARE survey data with administrative records: First experiences from SHARE-Germany. In: Malter F, Börsch-Supan A, (Hrsg.). SHARE wave 4. Innovations & methodology. München: Munich center for the economics of aging; 2013: 47–52

[20] Maier B, Wagner K, Behrens S et al. Deterministisches Record Linkage mit indirekten Identifikatoren. Daten des Berliner Herzinfarktregisters und der AOK Nordost zum Herzinfarkt. Gesundheitswesen 2015; 77: e15–e19

[21] March S, Rauch A, Thomas D et al. Datenschutzrechtliche Vorgehensweise bei der Verknüpfung von Primär- und Sekundärdaten in einer Kohortenstudie. Die lidA-Studie. Gesundheitswesen 2012; 74: e122–e129

[22] March S. Individual Data Linkage of Survey Data with Claims Data in Germany-An Overview Based on a Cohort Study. Int J Environ Res Public Health 2017; 14: 1543

[23] Ohlmeier C, Hoffmann F, Giersiepen K et al. Verknüpfung von Routinedaten der Gesetzlichen Krankenversicherung mit Daten eines Krankenhausinformationssystems. Machbar, aber auch "nützlich"? Gesundheitswesen 2015; 77: e8–e14

[24] 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

[25] Ohmann C, Smektala R, Pientka L et al. A new model of comprehensive data linkage – Evaluation of its application in femoral neck fracture. Z Evid Fortbild Qual Gesundhwes 2005; 99: 547–554

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

[27] Stallmann C, Ahrens W, Kaaks R et al. Individuelle Datenverknüpfung von Primärdaten mit Sekundär- und Registerdaten in Kohortenstudien. Potenziale und Verfahrensvorschläge. Gesundheitswesen 2015; 77: e37–e42

[28] Stang A, Jöckel K-H. Avoidance of representativeness in presence of effect modification. Int J Epidemiol 2014; 43: 630–631

[29] Weiskopf NG, Weng C. Methods and dimensions of electronic health record data quality assessment. Enabling reuse for clinical research. J Am Med Inform Assoc 2013; 20: 144–151

[30] Keller S, Korkmaz G, Orr M et al. The Evolution of Data Quality. Understanding the Transdisciplinary Origins of Data Quality Concepts and Approaches. Annu Rev Stat Appl 2017; 4: 85–108

[31] Watts S, Shankaranarayanan G, Even A. Data quality assessment in context. A cognitive perspective. Decision Support Syst 2009; 48: 202–211

[32] March S, Swart E, Robra B-P. Können Krankenkassendaten Primärdaten verzerrungsfrei ergänzen? – Selektivitätsanalysen im Rahmen der lidA-Studie. Gesundh ökon Qual manag 2017; 104–115

[33] Sozialgesetzbuch Zehntes Buch (SGB X) - Sozialverwaltungsverfahren und Sozialdatenschutz – (in der Fassung der Bekanntmachung vom 18. Januar 2001 (BGBl. I S. 130), das zuletzt durch Artikel 16 des Gesetzes vom 18. Dezember 2018 (BGBl. I S. 2639) geändert worden ist). Online http://www2.medizin.uni-greifswald.de/cm/fv/ship/datennutzung/ letzter Zugriff: 30.07.2019

[34] NAKO Gesundheitsstudie. Online https://nako.de letzter Zugriff: 04.06.2019

[35] Swart E, Thomas D, March S et al. Erfahrungen mit der Datenverknüpfung von Primär – und Sekundärdaten in einer Interventionsstudie. Gesundheitswesen 2011; 73: e126–e132

[36] Brown JS, Kahn M, Toh S. Data quality assessment for comparative effectiveness research in distributed data networks. Medical care 2013; 51: S22–S29

[37] Purchase HC, Welland R, McGill M et al. Comprehension of diagram syntax. An empirical study of entity relationship notations. Int J Hum Comput Stud 2004; 61: 187–203

[38] Hassenpflug J, Liebs TR. Register als Werkzeug für mehr Endoprothesensicherheit. Erfahrungen aus anderen Ländern und dem Aufbau des Endoprothesenregisters Deutschland. Bundesgesundheitsbl 2014; 57: 1376–1383

[39] Pommerening K, Drepper J, Helbing K et al. Leitfaden zum Datenschutz in medizinischen Forschungsprojekten. Generische Lösungen der TMF 2.0; 1. Aufl. 2014

[40] March S, Rauch A, Bender S et al. Data protection aspects concerning the use of social or routine data. FDZ-Methodenreport 12/2015. Nürnberg: Bundesagentur für Arbeit 2015

[41] Ihle P. Datenschutzrechtliche und methodische Aspekte beim Aufbau einer Routinedatenbasis aus der Gesetzlichen Krankenversicherung zu Forschungszwecken. Bundesgesundheitsbl 2008; 51: 1127–1134

[42] Swart E, Stallmann C, Schimmelpfennig M et al. Gutachten zum Einsatz von Sekundärdaten für die Forschung zu Arbeit und Gesundheit. 1. Aufl. Dortmund: Bundesanstalt für Arbeitsschutz und Arbeitsmedizin (BAuA); 2018

[43] Deutsche Forschungsgemeinschaft. Denkschrift zur Sicherung guter wissenschaftlicher Praxis. Weinheim: Wiley-VCH; 2013

[44] Bialke M, Bahls T, Havemann C et al. MOSAIC – A Modular Approach to Data Management in Epidemiological Studies. Methods Inf Med 2015; 54: 364–371

[45] Lablans M, Borg A, Ückert F. A RESTful interface to pseudonymization services in modern web applications. BMC Med Inform Decis Mak 2015; 15: 2

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

[47] Boyd J, Randall S, Ferrante AM. Application of Privacy-Preserving Techniques in Operational Record Linkage Centres. In Gkoulalas-Divanis A, Loukides G, (Hrsg.). Medical Data Privacy Handbook. Springer; 2015: 267–287

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

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

[50] Nonnemacher M, Nasseh D, Stausberg J. Datenqualität in der medizinischen Forschung. Leitlinie zum adaptiven Management von Datenqualität in Kohortenstudien und Registern. 2. Aufl. 2014

[51] Sakshaug J, Antoni M, Errors in Linking Survey and Administrative Data. In Biemer PP, Leeuw EDd, Eckman S et al., (Hrsg.). Total survey error in practice. Hoboken, New Jersey: John Wiley & Sons; 2017

[52] Baldi I, Ponti A, Zanetti R et al. The impact of record-linkage bias in the Cox model. J Eval Clin Pract 2010; 16: 92–96

[53] Krawczak M, Weichert T. Vorschlag einer modernen Dateninfrastruktur für die medizinische Forschung in Deutschland. Kiel 2017

[54] Christen P. Data Matching. Berlin, Heidelberg: Springer; 2012

[55] Schnell R, Bachteler T, Reiher J. Private Record linkage with Bloom filters. Proceedings of Statistics Canada Symposium 2010. https://www.uni-due.de/~hq0215/documents/2010/Schnell_2010_Private_Record_Linkage_With_Bloom_Filters.pdf Zugriff am 04.06.2019

[56] Rothman KJ, Greenland S, Lash TL. Modern Epidemiology. 3. Aufl. Philadelphia: Lippincott Williams & Wilkins; 2008