Bundesgesundheitsblatt
Gesundheitsforschung | Gesundheitsschutz

Indexed in Medline, SCIE and SCOPUS

Bundesgesundheitsbl 2014 · 57:180–187 · DOI 10.1007/s00103-013-1892-1

E. Swart · C. Stallmann · J. Powietzka · S. March

Datenlinkage von Primär- und Sekundärdaten

Ein Zugewinn auch für die kleinräumige Versorgungsforschung in Deutschland?

www.Bundesgesundheitsblatt.de

Springer Medizin

Bundesgesundheitsbl 2014 · 57:180–187
DOI 10.1007/s00103-013-1892-1
Online publiziert: 23. Januar 2014

E. Swart · C. Stallmann · J. Powietzka · S. March
Institut für Sozialmedizin und Gesundheitsökonomie (ISMG), Medizinische Fakultät, Otto-von-Guericke-Universität Magdeburg

Datenlinkage von Primär- und Sekundärdaten

Ein Zugewinn auch für die kleinräumige Versorgungsforschung in Deutschland?

Hintergrund

In der Versorgungsforschung werden unterschiedliche Methoden und Datenzugänge zur Erfassung von sowohl bundesweiten als auch regionalen Unterschieden im Inanspruchnahmeverhalten von Gesundheitsleistungen sowie der daraus resultierenden ökonomischen und logistischen Auswirkungen auf das Gesundheitssystem genutzt. Neben der Erhebung von Primärdaten durch Interviews und schriftliche Befragungen kommt der Analyse von Sekundärdaten eine besondere und stetig zunehmende Bedeutung zu. Bei diesen Sekundärdaten handelt es sich in der Regel um Routinedaten, also um Verwaltungs- und Abrechnungsdaten, z. B. von gesetzlichen Krankenversicherungen (GKV) oder anderen Sozialversicherungsträgern. Andere Daten wie geografische Raumdaten sind für ergänzende Analysen ebenfalls nutzbar. Wennbergs grundlegende Arbeiten zur regionalen Versorgungsforschung in Amerika („small area analysis“) [1, 2, 3] beziehen sich auf lediglich eine Datenquelle. Dies trifft auch für vergleichbare Arbeiten neueren Datums in Deutschland zu [4, 5]. Sowohl für Primär- als auch für Sekundärdaten gilt aber, dass sich die Verwendung einer einzigen Datenquelle aufgrund methodischer Einschränkungen des Ansatzes limitierend auf deren Aussagekraft auswirkt. Einerseits werden Sekundärdaten für ein anderes primäres Ziel erhoben und müssen entsprechend aufbereitet werden. Andererseits weisen auch Primärdaten, wie z. B. Surveydaten, bezüglich der Inanspruchnahme medizinischer Leistungen grundsätzliche Limitationen auf [6].

In der Versorgungsforschung werden derzeit verschiedene Ansätze diskutiert, um die jeweiligen inhaltlichen und methodischen Grenzen von Primär- und Sekundärdaten über eine Verknüpfung verschiedener Datensätze zu überwinden. **Tab. 1 und 2** stellen – für das Beispiel der Erfassung der Inanspruchnahme medizinischer Leistungen – abweichende und damit sich synergistisch ergänzende Inhalte von Primär- und (GKV-)Sekundärdaten gegenüber. Es ist zu erkennen, dass sowohl Angaben aus persönlichen oder fragebogengestützten Befragungen als auch Sekundärdaten unabweisbar Lücken aufweisen, die durch den jeweils anderen Datenzugang geschlossen werden können (**Tab. 1**).

Gleichzeitig gibt es bei der Analyse von Primär- und Sekundärdaten unvermeidbare methodische Probleme (**Tab. 2**), die sich ebenfalls tendenziell bei der Nutzung des jeweils anderen Datenzugangs vermeiden oder zumindest verringern lassen.

Zielstellung des vorliegenden Beitrages soll es daher sein, verschiedene Formen des Datenlinkage (auf aggregiertem bzw. individuellem Niveau), deren Potenziale und Limitationen für die kleinräumige Versorgungsforschung herauszuarbeiten und anhand einiger Beispiele zu erläutern. Der potenzielle Zugewinn eines Datenlinkage für eine kleinräumige Versorgungsforschung wird dadurch hergestellt, dass abschließend auf einzelne versorgungsrelevante Beiträge aus diesem Schwerpunktheft Bezug genommen wird.

Was bedeutet eigentlich Datenlinkage?

Datenlinkage (in der Informatik „record linkage“) bezeichnet die Verknüpfung verschiedener Datenquellen mittels geeigneter Schlüsselvariablen. Diese müssen in beiden Datenquellen vorkommen und sollten eine eineindeutige Zuordnung ermöglichen und somit eine fehlerfreie Kombination erlauben. Beim Datenlinkage wird zwischen dem Linkage auf aggregiertem Niveau und auf individueller Ebene unterschieden. Hierfür bieten sich verschiedene Datenquellen mit regionalem Bezug an. Diese sollen im Folgenden kurz vorgestellt werden, bevor auf die beiden Formen des Datenlinkage detaillierter eingegangen wird.

Datenquellen mit regionalem Bezug

Für ein Linkage von Sekundärdaten mit räumlichem Bezug bietet sich eine Reihe gut zugänglicher Datenquellen an, die

Tab. 1 Abweichende Inhalte und Primär- und Sekundärdaten (bezüglich des Themenfeldes Inanspruchnahme)

Primärdaten	Sekundärdaten (GKV-Abrechnungsdaten)
Anlass der Inanspruchnahme	Vollständiger, längsschnittlicher Verlauf
Häufigkeit ambulanter Kontakte	Einzelleistungen (EBM-, OPS-Leistungen; Medikamente)
Determinanten der Inanspruchnahme	Inanspruchnahme als relevante Outcomes (z. B. stationäre Aufenthalte)
Selbst finanzierte Leistungen [OTC (over the counter)-Medikamente, Prävention, zahnmedizinische Leistungen]	Zeit unter Risiko

Tab. 2 Methodische Probleme bei Nutzung von Primär- und Sekundärdaten (bezüglich des Themenfeldes Inanspruchnahme)

Primärdaten	Sekundärdaten (GKV-Abrechnungsdaten)
Validität der Angaben (z. B. recall bias, Verständlichkeit)	Nur Leistungen zulasten der GKV, keine Angaben zu Privatversicherten, keine individuellen Gesundheitsleistungen
Enges nutzbares Zeitfenster	Begrenzte soziodemografische Merkmale
Umfang der zu erfassenden Merkmale	Verwaltungsdaten, keine klinischen Daten
(Noch) keine Standards	Systembedingte Rahmenbedingungen (z. B. Änderungen in Klassifikationen, Vergütungen)

Informationen auf der Ebene von Landkreisen oder kreisfreien Städten, Gemeinden oder Raumordnungstypen für die wissenschaftliche Forschung bereitstellen. Auf 2 der am weitesten verbreiteten Datenkörper soll hier kurz eingegangen werden; nähere Informationen finden sich auf den nachfolgend aufgeführten Webseiten.

I) Das Bundesinstitut für Bau-, Stadt- und Raumforschung (BBSR; http://www.bbsr.de) stellt jährlich ca. 600 Indikatoren und Karten zur Raum- und Stadtentwicklung in Deutschland und Europa (INKAR) aus verschiedenen Themenbereichen bereit. Gesundheitsbezogene Indikatoren werden allerdings nur für die Bereiche medizinische Versorgung und Infrastruktur angeboten. Die Indikatoren werden aus amtlichen Statistiken des Bundes und der Länder abgeleitet. Raumbezogene Daten werden auf Basis administrativer (Länder, Kreise, Gemeindeverbände) und nicht-administrativer Gebiete (z. B. Siedlungsstrukturtypen) mit indirektem Raumbezug erstellt. Je nach regionaler Auflösung der Primär- und Sekundärdaten können so Informationen aus dem INKAR-Datensatz auf diesen beiden Ebenen zugespielt werden.

II) Über die Webseite des Statistischen Bundesamtes (http://www.destatis.de) werden, mit wenigen Ausnahmen, nur Daten auf Regierungsbezirks- oder Kreisebene zur Verfügung gestellt. Zusätzliche feingliedrige Daten der amtlichen Statistik sind jedoch über die „Regionaldatenbank Deutschland“ (http://www.regionalstatistik.de) abrufbar. Insgesamt 80 Indikatoren aus allen Themenbereichen der amtlichen Statistik werden hier, aufbereitet in Form thematischer Karten, für alle Landkreise und kreisfreien Städte Deutschlands bereitgehalten. Aspekte des Gesundheitswesens beziehen sich allerdings lediglich auf Strukturinformationen zu Krankenhäusern sowie Vorsorge- und Rehabilitationseinrichtungen.

Das amtliche Gemeindeverzeichnis (GV-ISys) enthält ergänzend für jede politisch selbstständige Gemeinde Deutschlands Angaben zur Fläche, Einwohnerzahl und siedlungsstrukturellen Typisierung. Auch diese Daten stehen, wie die der Regionaldatenbank, kostenfrei über die Webseite des Statistischen Bundesamtes (http://www.destatis.de) zur Verfügung.

Datenlinkage mit aggregierten Daten

Im Weiteren wird das Datenlinkage mit aggregierten Daten (Primär- und Sekundärdaten bzw. mehrere Sekundärdaten) näher dargestellt. Bei der wohl häufigsten Form des Datenlinkage werden zu Primärdaten, die ein Merkmal mit regionalem Bezug aufweisen (z. B. Landkreis, kreisfreie Stadt, Postleitzahl), räumlich aggregierte Daten zugespielt. In einer Studie des Robert Koch-Instituts (RKI) wurden Daten von rund 21.000 Teilnehmern des telefonischen Gesundheitssurveys 2009 verknüpft [6]. Beispielhaft sollten mithilfe dieser Daten die regionale Verteilung von 3 Gesundheitsindikatoren (subjektiver Gesundheitszustand, Rauchen und Adipositas) und ihre kleinräumigen Einflussfaktoren geschätzt werden. Die Studie ergab eine beträchtliche kleinräumige Variabilität und konnte eine Reihe von soziodemografischen und sozioökonomischen Determinanten dieser 3 Indikatoren identifizieren.

Mit einem ähnlichen Vorgehen lässt sich grundsätzlich auch der Einfluss räumlich aggregierter Variablen auf individuelle Gesundheitsfaktoren untersuchen. Damit können Daten aus epidemiologischen Studien mit regionalem Bezug um aggregierte Merkmale auf der Ebene kleiner räumlicher Einheiten ergänzt werden. So wurde in einem Linkage von Daten des Bundesgesundheitssurveys 1998 und INKAR-Daten (Sekundärdaten) versucht, das Inanspruchnahmemodell nach Andersen in einer Mehrebenenanalyse zu überprüfen [7].

Methodisch kaum anders gestaltet sich der Fall, wenn fall- oder versichertenbezogene Sekundärdaten (z. B. Abrechnungsdaten der gesetzlichen Krankenversicherung) mit räumlich aggregiert vorliegenden Sekundärdaten verknüpft werden, beispielsweise zur Untersuchung regionaler Unterschiede in der Kodierqualität [8] bzw. der Häufigkeit von Krankenhausaufnahmen bei Diabetes [9].

Diese Form des Datenlinkage ist datentechnisch und organisatorisch vergleichsweise unkritisch umzusetzen, da die zuzuspielenden Daten leicht und kostengünstig verfügbar sind. Für den vorliegenden Beitrag ist bedeutsam, dass sich die regionale Auflösung der Daten in gut zugänglichen Datenbanken an den politischen Gebietskörperschaften bis hinunter zur Gemeindeebene am amtlichen Gemeindeschlüssel orientiert. Die Kompatibilität mit anderen räumlichen Auflösungen, z. B. nach Postleitzahlbereichen,

muss dagegen im Einzelfall geprüft und ggf. durch Überleitungsalgorithmen entwickelt werden [10]. So sind z. B. nicht alle Postleitzahlbereiche eindeutig einem Bundesland oder einem Landkreis zuzuordnen (**Abb. 1**). Hinzu kommt, dass verbreitet nicht nur bezogen auf ihre Fläche und/oder Einwohnerzahl große Landkreise, sondern auch große (Großstadt-)Gemeinden in diesen Datenbanken als Entitäten erscheinen, bei denen u. U. eine feinere Auflösung wünschenswert wäre. Beim Datenlinkage spielt die gewählte Schlüsselvariable also eine entscheidende Rolle.

Primärdaten können im Übrigen auch mit aggregierten Sekundärdaten zusammengespielt werden, ohne dass die Verknüpfung notwendigerweise über eine Raumvariable erfolgen muss. So wurde in einer kürzlich veröffentlichten Studie ein Linkage von Primärdaten aus der BIBB/BAuA-Erwerbstätigenbefragung 2005/06 mit aggregierten Arbeitsunfähigkeitsdaten durchgeführt. Als Schlüsselvariable diente die „berufliche Tätigkeit“ basierend auf der Klassifizierung der Berufe 1988 (KldB88) [11]. Unter regionalen Gesichtspunkten könnten mit dieser Analyse auch Erkenntnisse über die gesundheitlichen Auswirkungen regionaler Arbeits(markt)bedingungen gewonnen werden.

Individuelles Datenlinkage

Das individuelle Datenlinkage von Primär- und Sekundärdaten ist mit erheblichen rechtlichen und methodischen Problemen verbunden. Dabei können 2 Ansätze unterschieden werden:

Zuspielung individueller Primärdaten zu personenbezogenen Sekundärdaten

Eine Untersuchung von Bitzer et al. [12] dient als Beispiel für diesen Ansatz. Ausgangspunkt waren Routinedaten von Versicherten der AOK Niedersachsen, die anhand definierter Ein- und Ausschlusskriterien ausgewählt wurden. In der Studie wurden Versicherte ein halbes Jahr nach einer endoprothetischen Hüftgelenksoperation zu ihrem aktuellen Schmerzniveau und zu Einschränkungen in der Alltagsfunktionalität befragt. Die Ergebnisqualität wurde anhand von Abrechnungsdaten zusätzlich über Revisionseingriffe und die Häufigkeit von Krankenhausaufnahmen wegen Komplikationen nach endoprothetischen Eingriffen abgebildet. Identifiziert wurden die Fälle über die entsprechenden OPS- und ICD-Codes. Der Fragebogen zu den patientenbezogenen Ergebnisindikatoren wurde den Versicherten durch die Krankenkasse zugestellt und von ihnen an das für die Auswertung zuständige wissenschaftliche Institut zurückgeschickt. Dort erfolgte eine pseudonymisierte Zusammenführung und Analyse von Primär- und Sekundärdaten.

Auf diese Weise wurden in der Studie die patientenbezogene Ergebnisindikatoren erfasst, die in keiner Sekundärdatenquelle zu finden sind. Bei einer hohen Response von 83% der angeschriebenen Versicherten gelang es damit, die

Bundesgesundheitsbl 2014 · 57:180–187 DOI 10.1007/s00103-013-1892-1

E. Swart · C. Stallmann · J. Powietzka · S. March

Datenlinkage von Primär- und Sekundärdaten. Ein Zugewinn auch für die kleinräumige Versorgungsforschung in Deutschland?

Zusammenfassung

Die Versorgungsforschung in Deutschland behandelt eine Vielzahl von Themen im regionalen Kontext und nutzt dafür überwiegend eine (in der Regel: Sekundär-)Datenquelle. Deren spezifische Nachteile und methodischen Einschränkungen können sich limitierend auf eine Analyse auswirken. Zur regionalen Aufgliederung existieren vielfältige Datenquellen, die für die regionale Versorgungsforschung von Interesse sein könnten. Eine Verknüpfung verschiedener Datenquellen (Datenlinkage) könnte somit die Analysemöglichkeiten erweitern. In der Versorgungsforschung selbst werden derzeit verschiedene Ansätze diskutiert, um die jeweiligen Schwächen von Primär- und Sekundärdaten über ein Datenlinkage zu überwinden. Der vorliegende Beitrag thematisiert die verschiedenen Formen des Datenlinkage (auf aggregiertem bzw. individuellem Niveau) sowie deren Potenziale und Restriktionen für die kleinräumige Versorgungsforschung. Der Fokus liegt auf dem individuellen Datenlinkage, das ein schriftliches Einverständnis voraussetzt (informed consent). Unter Berücksichtigung der methodischen und insbesondere datenschutzrechtlichen Herausforderungen werden Schlussfolgerungen über zukünftige Anwendungsfelder und -möglichkeiten der kleinräumigen Versorgungsforschung gezogen und an Beispielen konkretisiert.

Schlüsselwörter

Datenlinkage · Primärdaten · Sekundärdaten · Kleinräumige Versorgungsforschung · Datenschutz

Data linkage of primary and secondary data. A gain for small-area health-care analysis?

Abstract

In Germany, research on health-care services addresses many topics within a regional context, and it predominantly uses a single (typically secondary) data source for this purpose. The specific disadvantages and methodological challenges associated with these data sources may limit analysis. Various data sources break the data down by region and may be of interest in regional health-care research. Linking multiple data sources (data linkage) could therefore expand analysis options in this area. Researchers in this field are currently discussing various approaches for using data linkage to overcome the respective weaknesses of primary and secondary data. This contribution covers the various types of data linkage (on an aggregate or individual level) and their potentials and limitations in small area health services research. The focus lies on individual data linkage, which requires written informed consent. Taking into account methodological and particularly data protection challenges, conclusions are drawn regarding future application areas and options of small area health services research and specific examples are provided.

Keywords

Data linkage · Primary data · Secondary data · Small-area health-care services research · Data protection

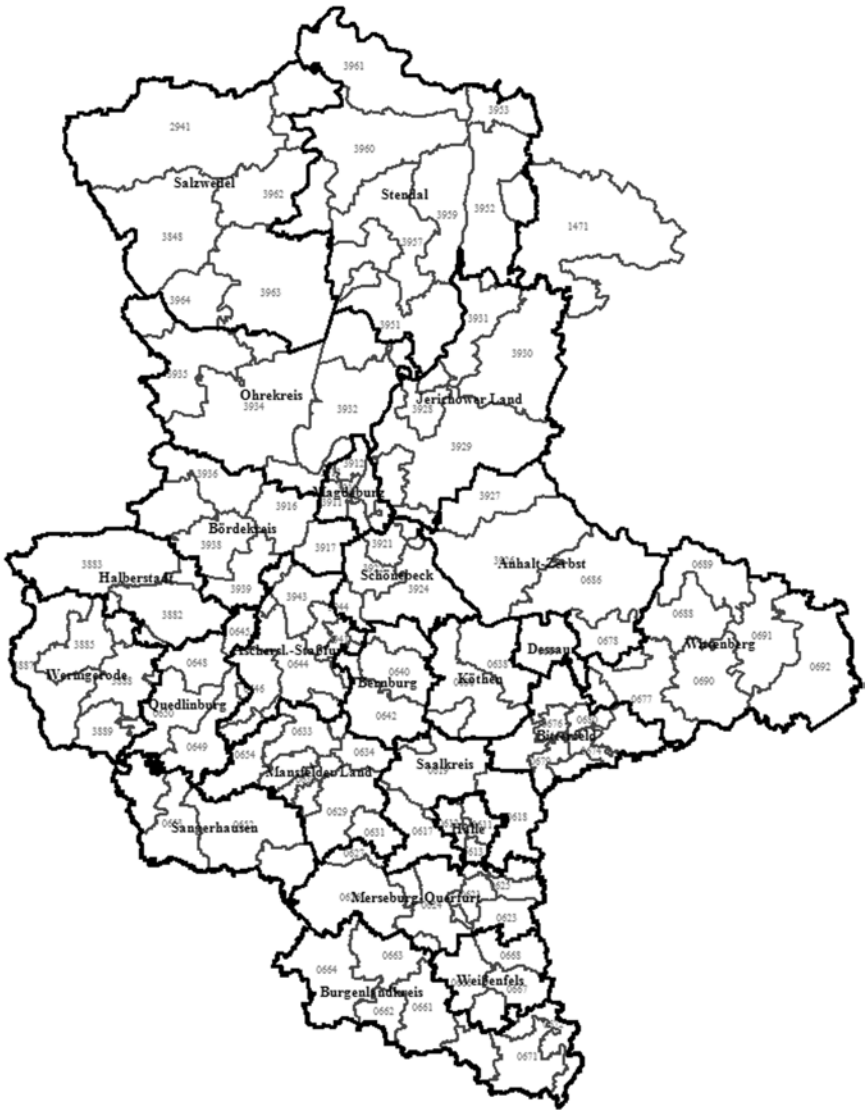

Abb. 1 ▲ Vierstelliger Postleitzahlbereich und Landkreise bzw. kreisfreie Städte in Sachsen-Anhalt (Stand: 2006 vor der Gebietsreform 2007)

versorgungsbezogene Sicht auf die Ergebnisqualität um die wichtige versichertenbezogene Perspektive zu erweitern. Für eine regionalisierte Versorgungsforschung bietet sich dieser Ansatz an. Untersuchungen zum Versorgungsbedarf auf der Ebene von Landkreisen oder Postleitzahlbereichen können um patientenbezogene Angaben zur Erreichbarkeit und Akzeptanz von Versorgungseinrichtungen oder zu Motiven der Inanspruchnahme ergänzt werden [6, 13]. Datenschutzrechtlichen Anforderungen wird Genüge getan, wenn zum einen bei der Pseudonymisierung auf ein einheitliches, sicheres Verfahren gesetzt und das Linkage bei einer neutralen dritten Institution (Vertrauensstelle) durchgeführt wird. Zum anderen sorgt ein Anschreiben der jeweiligen Krankenversicherung an die beteiligten Versicherten für ein hohes Maß an Transparenz [14].

Zuspielung individueller Sekundärdaten zu Primärdaten

Hier werden den Primärdaten von Teilnehmern einer (versorgungs-)epidemiologischen Studie bei Vorliegen eines entsprechenden individuellen Einverständnisses (informed consent) Sekundärdaten zugespielt. Dies können Daten der gesetzlichen Krankenversicherung, aber auch anderer Sozialversicherungsträger (z. B. der gesetzlichen Rentenversicherung) sein. Epidemiologische oder klinische Krankheitsregister wären ebenfalls geeignet.

In Deutschland wurde und wird dieser Ansatz in kürzlich beendeten bzw. laufenden Studien erprobt und umgesetzt. In der AGil-Studie, eingebettet in das (regionale) IV-(Integrierte Versorgung)-Projekt „Gesundes Kinzigtal", konnte mit Sekundärdaten der AOK Baden-Württemberg die fragebogengestützte Evaluation eines Gesundheitsförderungsprogramms bei älteren Menschen gezielt um Sekundärdaten ergänzt werden. So konnte überprüft werden, ob die Intervention eine Veränderung bei der Inanspruchnahme von Gesundheitsleistungen bewirkt hat [15].

In der arbeitsepidemiologischen lidA-(leben in der Arbeit)-Studie (http://www.lida-studie.de)[1] werden den Primärdaten der rund 6600 Studienteilnehmer sowohl Prozessdaten des Instituts für Arbeitsmarkt- und Berufsforschung (IAB) als auch Routinedaten der gesetzlichen Krankenversicherung zugespielt. Dadurch kann einerseits die Berufsbiografie als wesentliche Determinante des gegenwärtigen und zukünftigen Gesundheitszustands umfassender als in einer persönlichen Befragung erfasst werden. Andererseits können objektive Resultatvariablen teilnehmerbezogen multivariat untersucht werden (Erwerbslosigkeit, Verrentung, akute und chronische Erkrankungen, Krankenhausaufenthalte). Zudem besteht die Möglichkeit, den Primärdaten regionale Daten der amtlichen Statistik der Bundesagentur für Arbeit (BA) zuzuspielen [14, 16, 17].

In der 2014 beginnenden Nationalen Kohorte[2] (NaKo) [18] wird schließlich der bislang umfassendste Versuch eines individuellen Linkage von Primär- mit ergänzenden Sekundär- und Registerdaten unternommen (http://www.nationale-kohorte.de). Alle 200.000 Studien-

[1] Die Studie wird gefördert durch das Bundesministerium für Bildung und Forschung (BMBF) (Förderkennzeichen der am Verbund beteiligten Vorhaben: 01ER0806, 01ER0825, 01ER0826, 01ER0827).

[2] Die Studie wird durch das BMBF mit dem Förderkennzeichen 01ER1001A-I finanziert und durch die Helmholtz Gemeinschaft sowie die beteiligten Universitäten und Institute der Leibniz-Gemeinschaft unterstützt.

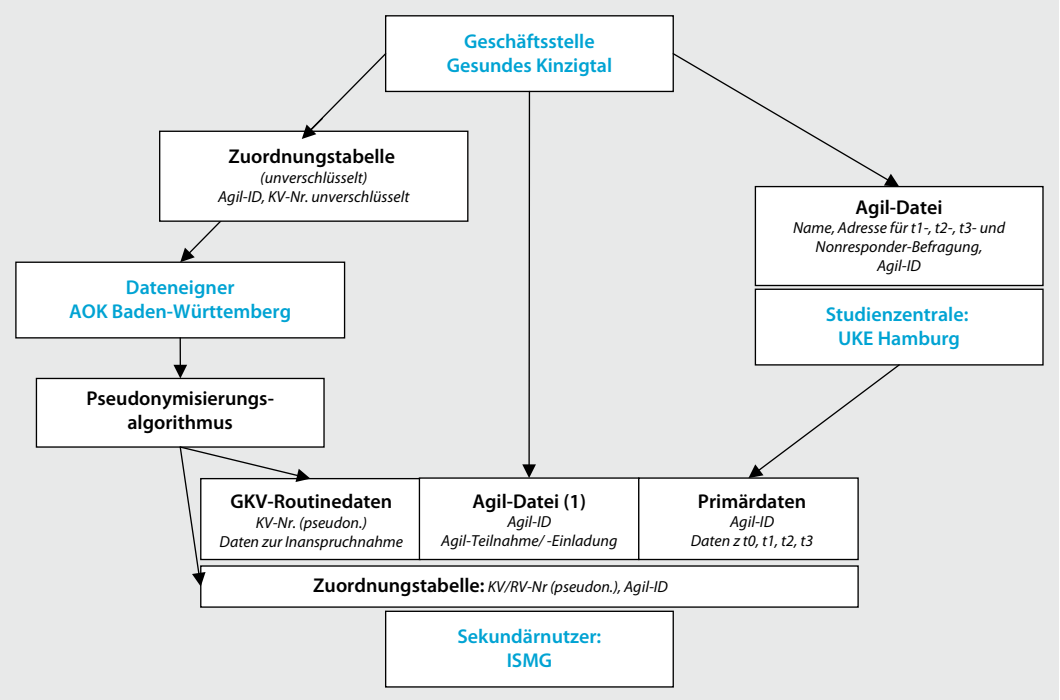

Abb. 2 ◀ Datenfluss und Datenzusammenführung der verschiedenen Datenkörper in der AGil-Studie. (Nach [15])

teilnehmer sollen um ihr Einverständnis für diese Zuspielung gebeten werden. Dazu zählen u. a. Daten der Krankenversicherungen (gesetzliche und erstmals auch private), kassenübergreifende Daten des Zentralinstituts für die Kassenärztliche Versorgung in Deutschland (ZI), Daten epidemiologischer und klinischer Krebsregister sowie aus Sterberegistern. Ergebnisse aus dem NaKo-Prätest lassen bei den Studienteilnehmern eine hohe Bereitschaft erkennen, der Zuspielung zuzustimmen [19].

Ein derartiges Datenlinkage bietet sich für jede regional verankerte epidemiologische Studie an. So wurde z. B. in der Heinz-Nixdorf-Recall-Studie [20] ein Datenlinkage mit GKV-Daten für eine gesundheitsökonomische Begleitevaluation unternommen [21]. In der SHIP-Studie (Leben und Gesundheit in Vorpommern, [22]) wird derzeit ein Datenlinkage mit Daten der Kassenärztlichen Vereinigung Mecklenburg-Vorpommern vorbereitet. Große epidemiologische Studien wie die NaKo haben in diesem Zusammenhang auch für kleinräumige Analysen größere statistische Power. In der lidA-Studie werden jedoch auch die Grenzen deutlich: So entfallen bei einer Auflösung nach den insgesamt 222 Erhebungspunkten rechnerisch nur noch 30 Teilnehmer auf jeden einzelnen Punkt [16]. Abnehmende Fallzahlen bei stärkerer regionaler Auflösung und eine damit erhöhte Anfälligkeit kleinerer regionaler Einheiten gegenüber zufälligen Einflüssen stellen ein Grundproblem kleinräumiger Analysen dar [4, 5]. Insofern muss bei kleinräumigen epidemiologischen Analysen eine Abwägung zwischen der Feinheit der regionalen Auflösung und der statistischen Power vorgenommen werden.

Datenschutzrechtliche Anforderungen beim individuellen Datenlinkage

Handelt es sich bei den Sekundärdaten um personenbezogene Daten oder um Sozialdaten, greifen umfassende datenschutzrechtliche Vorgaben. Die Nutzung solcher Daten im Rahmen der wissenschaftlichen Forschung regelt § 75 Sozialgesetzbuch (SGB) X. In diesem werden auch die Bedingungen definiert, unter denen ein bzw. kein individuelles Einverständnis für die Verknüpfung verschiedener Datenquellen etc. erforderlich wird. Der § 67 SGB X legt weitere unabweisbare Details fest, etwa die Aufklärung der Studienteilnehmer, das Einfordern eines informierten Einverständnisses in Schriftform (informed consent) oder die Möglichkeit seines Widerrufs [23].

Anhand der 3 Studien (AGil, lidA, NaKo) sollen die datenschutzrechtlichen Aspekte eines individuellen Datenlinkage skizziert werden. Für die Evaluation des AGil-Projekts erwies es sich als vorteilhaft, dass die Zustimmung zur wissenschaftlichen Nutzung der (pseudonymisierten) GKV-Abrechnungsdaten durch die Projektträgergesellschaft Voraussetzung für die Teilnahme am IV-Projekt „Gesundes Kinzigtal“ war. Alle Teilnehmer stimmten dieser Nutzung bereits mit ihrer Einschreibung explizit zu.[3] Den-

[3] Auszüge aus der Einverständniserklärung „Ich bin damit einverstanden, dass meine personenbezogenen Daten bei Bedarf und nur im unbedingt erforderlichen Umfang […] an Universitäten und Forschungseinrichtungen nur in pseudonymisierter Form, d. h. ohne Bezug zu meiner Person, zu Zwecken der Auswertung und Erfolgskontrolle übermittelt werden“. In einem separaten Merkblatt wurden die Dateninhalte der GKV-Daten präzisiert: „Daten zu früheren Erkrankungen, Daten zu Krankenhausaufenthalten und ambulanten Operationen, Daten zu Arbeitsunfähigkeitszeiten mit Diagnosen, Daten zu in der Vergangenheit erfolgten Therapien, Vorsorge- und Rehabilitationsmaßnahmen, Angaben zu Art und Kosten von verordneten Medikamenten, Heil- und Hilfsmitteln, Informationen über Ihre persönliche und familiäre

noch bedurfte es für die Datenverknüpfung von Primär- und Sekundärdaten einer zusätzlichen Datenschutzvereinbarung zwischen den Projektbeteiligten (Krankenkasse, Projektträger, Universität Magdeburg). Über eine Überleitungstabelle zwischen AGil-Teilnehmernummer und Versichertenpseudonym konnten Primär- und Sekundärdaten zusammengespielt werden (**Abb. 2,** [15]).

Aufwendiger gestaltet sich das individuelle Datenlinkage bei größeren epidemiologischen Studien. Es müssen u. a. umfassende Aufklärungsmaterialien für die Teilnehmer erstellt werden, in denen detailliert über das Forschungsvorhaben (wer, wie, was, wo, wofür) informiert wird. Kohortenstudien benötigen aufgrund ihres prospektiven Charakters eine Einwilligung zum Verbleib (und somit Erhalt) des Teilnehmers in der Studie für die nächsten Befragungszeitpunkte. Zudem kann kein universelles Einverständnis für die verschiedenen Datenkörper eingeholt werden. Daher wurden z. B. im Rahmen der lidA-Studie mehrere Einverständniserklärungen, eine je Datenkörper, abgefordert. Die lidA-Studie hat sowohl für die Nutzung der IAB-Daten einen Antrag beim Bundesministerium für Arbeit und Soziales als auch für die Nutzung der Daten der gesetzlichen Krankenversicherung bei den jeweiligen Aufsichtsbehörden auf Bundes- und Länderebene gestellt [14]. In solchen Konstellationen muss in Absprache mit den Datenschützern in der Formulierung von Teilnehmerinformationen und Einverständniserklärungen eine Balance gefunden werden zwischen Allgemeinverständlichkeit und Knappheit der Formulierungen einerseits und den datenschutzrechtlichen Auflagen bezüglich einer umfassenden und datenkörperspezifischen Aufklärung über die zur Nutzung vorgesehenen Daten.

In der NaKo wird, im Gegensatz zur lidA-Studie, die rechtliche Zulässigkeit lediglich einer einzelnen umfassenden Einverständniserklärung geprüft. In dieser wird stufenweise über jeden Datenkörper informiert. Ergänzt wird die Einverständniserklärung durch eine umfangreiche Informationsbroschüre zu den Sekundär- und Registerdaten. Das Einverständnis wird, analog einer Checkliste, je Datenkörper gegeben oder verweigert und final über eine Unterschrift vom Teilnehmer bestätigt. Die Einverständniserklärung und Aufklärungsmaterialien werden im Vorfeld mit Datenschützern auf Bundes- und Landesebene sowie mit denen der Dateneigner abgestimmt. Alle datenschutzrechtlichen Verpflichtungen und Vorgaben werden in einem Datenschutzkonzept festgeschrieben.

Methodische Herausforderungen beim individuellen Datenlinkage

Zu den oben kurz angerissenen methodischen Problemen beim Datenlinkage mit aggregierten Daten kommen spezifische Herausforderungen hinzu, wenn ein individuelles Datenlinkage vorgenommen werden soll. Bei längsschnittlicher Nutzung von Sekundärdaten und deren wiederholter Zuspielung zu Primärdaten muss darauf geachtet werden, dass der Pseudonymisierungsalgorithmus unverändert bleibt bzw. dass bei seiner Änderung ein eindeutiger Umsteigeschlüssel existiert. In der AGil-Studie kam es zu einem solchen Wechsel des Algorithmus. Davon betroffen waren jedoch nur 0,3 Promille aller AOK-Versicherten [15].

Weiterhin muss beachtet werden, dass beim individuellen Datenlinkage nur die Daten zugespielt werden dürfen, für die ein schriftliches Einverständnis vorliegt. Es bedarf daher spezifischer Non-Responder-Analysen, die der Frage nachgehen, inwieweit sich Studienteilnehmer mit erteiltem informed consent von allen Studienteilnehmern unterscheiden. Insofern sind beim individuellen Datenlinkage gegenüber Studien, die allein auf Primärdaten beruhen, auf einer weiteren Ebene Selektionseffekte möglich [24]. Hier sei jedoch der Hinweis erlaubt, dass auch Primärdaten Verzerrungen durch Selektionsbias etc. unterliegen können.

Bei einer angestrebten individuellen Zuspielung von GKV-Daten kommt hinzu, dass der logistische Aufwand zur Kontaktierung aller, derzeit noch mehr als 100 gesetzlichen Krankenversicherungen häufig nicht tragbar ist. So wurden in der Heinz-Nixdorf-Recall-Studie nur Krankenversicherungen, bei denen eine große Zahl von Studienteilnehmern versichert war, um die Bereitstellung ihrer Daten gebeten [21]. In der lidA-Studie wurden hingegen alle 103 gesetzlichen Krankenversicherungen angeschrieben, auf die sich die Teilnehmer mit informed consent verteilten. Es ist jedoch absehbar, dass letztendlich nicht zu allen Teilnehmern mit informed consent auch individuelle Daten zur Inanspruchnahme genutzt werden können. Daher muss auch geprüft werden, inwieweit sich diese dritte Selektionsebene auf die interne und externe Validität der Studienergebnisse auswirkt.

Erkenntnisgewinn durch Datenlinkage

Der konkrete Erkenntnisgewinn durch individuelles Datenlinkage soll abschließend anhand der in anderen Beiträgen in diesem Schwerpunktheft thematisierten typischen Fragestellungen der kleinräumigen Versorgungsforschung näher benannt werden. Stock und Danner [25] bewerten die kleinräumige Versorgungsforschung als wichtige empirische Grundlage der Gesundheitspolitik, die die lokalen bzw. regionalen Versorgungsbedarfe zur Deckung mit den Angebotsstrukturen und Versorgungsmustern bringen muss. Sie nennen vielfältige für diesen Zweck nutzbare Primär- und Sekundärdatenquellen, ohne deren Verknüpfung zu thematisieren. Allerdings sind der subjektive und objektive Versorgungsbedarf nicht mit Inanspruchnahme gleichzusetzen. Gerade hier könnte die individuelle Verlinkung von Primärdaten, in denen der subjektive und objektive Bedarf der Patienten (z. B. operationalisiert über den subjektiven Gesundheitszustand bzw. über somatische und psychische Beschwerden sowie Einschränkungen und Behinderungen) und die Präferenzen der Leistungserbringer (operationalisiert über Einstellungen zu verschiedenen Behandlungsoptionen) erfasst werden können, mit GKV-Sekundärdaten zur realisierten Inanspruchnahme eine Möglichkeit eröffnen, zwischen Bedarf und Inanspruchnahme zu differenzieren. Damit wären vertiefte empirische

Situation, Familienstand, Beruf und eventuell bestehende besondere Belastungen".

Erkenntnisse zur Problematik von Über-, Unter- und Fehlversorgung zu gewinnen. Auch eine bessere Prognose des zukünftigen Versorgungsbedarfs würde damit möglich [6].

Der Beitrag von Zorn [26] berührt zunächst einen vergleichbaren Verwertungszusammenhang in der Problematisierung der Disease-Management-Programme (DMP). Seit einiger Zeit ist es möglich, die spezifische teilnehmerbezogene DMP-Programmevaluation mit GKV-Daten zu verlinken und damit eine Datenbasis zu generieren, die längsschnittlich sowohl klinische als auch Inanspruchnahmeinformationen enthält [27]. Bei einem regionalen Fokus auf die Evaluation derartiger Programme könnten durch Befragung der Programmteilnehmer noch patientenbezogene Bewertungen und regionale Versorgungsstrukturen einbezogen werden. Zorn geht in seinem Beitrag ebenfalls auf die regionalisierte Bedarfsplanung ein und fordert eine breite Verfügbarkeit von Routinedaten zur Inanspruchnahme für alle Akteure im Gesundheitswesen und deren Ergänzung durch spezielle Surveys. Der synergistische Gewinn – bei zugegebenermaßen höherem Aufwand – könnte dabei, wie gerade skizziert, durch eine direkte Verlinkung gesteigert werden.

Robra [28] wirft die Frage auf, welches Außenkriterium bei offenkundiger, auch in Deutschland flächendeckend auftretender Versorgungsheterogenität zur Beantwortung der Frage „Which rate is right?“ herangezogen werden kann. Bei beiden von ihm diskutierten Kriterien, den Patientenpräferenzen und den patientenbezogenen Versorgungsergebnissen, kann eine Verlinkung von Sekundärdaten zur Inanspruchnahme mit Primärdaten wertvolle Erkenntnisse erbringen. Dies gilt sowohl, wie oben skizziert, für den subjektiven und objektiven Bedarf als auch für die Entscheidung für oder gegen eine medizinische Intervention sowie zur wahrgenommenen Erreichbarkeit von Versorgungsangeboten.

Zur Abbildung der Versorgungsqualität durch Routinedaten gibt es methodisch weit gediehene Verfahren (QSR – Qualitätssicherung mit Routinedaten [29]), doch könnten die dort verwendeten Ergebnisindikatoren wie Revisionsraten oder Wiederaufnahmeraten wie im Beispiel von Bitzer et al. [12] durch eine Anreicherung durch patientenbezogene (in Routinedaten nicht abbildbare) Qualitätsindikatoren erfahren.

Mangiapanes Beschreibung des Versorgungsatlasses [30] bietet schließlich ein Beispiel für die Verlinkung von bereits aggregierten Sekundärdatenkörpern. Die über den Versorgungsatlas bereitgestellten und für die weitere Nutzung exportierbaren Informationen zum Versorgungsgeschehen auf Ebene von Landkreisen und kreisfreien Städten können für multivariate ökologische Analysen ergänzt werden. Regionale soziodemografische und sozioökonomische Informationen über potenzielle Determinanten der Versorgung können so miteinander verknüpft werden.

Fazit

Regionale Daten sind auf aggregiertem Niveau im Rahmen der amtlichen Statistiken in Deutschland bereits gut erhältlich. Lediglich die Verfügbarkeit von regionalen Gesundheitsindikatoren und -daten muss noch bemängelt werden. Daher bietet das individuelle Datenlinkage künftig die Chance, auch diesem Defizit zu begegnen. Weitere, bislang weitgehend ungenutzte Datenquellen, wie beispielsweise zu Umweltexpositionen oder Herzinfarkt-, Krebs- und Mortalitätsregister, können und sollten künftig für die Versorgungsforschung erschlossen werden.

Neben den in diesem Beitrag thematisierten methodischen und datenschutzrechtlichen Aspekten des (individuellen) Datenlinkage sei im Übrigen auf ausländische Empfehlungen zum Datenlinkage hingewiesen [31, 32], deren Übertragbarkeit auf die spezifischen deutschen Bedingungen bezüglich Datenverfügbarkeit, Datenstruktur und Datenschutz geprüft werden müssen.

Abschließend sei noch erwähnt, dass bei Studien, die die Methode des Datenlinkage verwenden, mit einem sehr hohen zeitlichen Aufwand insbesondere in der Vorbereitungsphase zu rechnen ist. Aufgrund des Pilotcharakters des individuellen Datenlinkage mit mehreren Sekundärdatenquellen müssen vereinbarte Konzepte auch in der Feldphase immer wieder auf ihre Praxistauglichkeit hin überprüft und ggf. angepasst werden. Weiterhin wird empfohlen, umfassende Beratungsgespräche mit allen beteiligten Akteuren (Forscher, Dateneigner, Datenschützer etc.) zu führen, um das Verfahren des Datenlinkage datenschutzkonform und funktionsfähig zu gestalten.

Das Datenlinkage kann zusammenfassend, unter Beachtung aller methodischen sowie rechtlichen Anforderungen und Vorgaben, als Zugewinn für die regionale Versorgungsforschung betrachtet werden. Zudem bietet die stetig steigende Zahl an verfügbaren Datenquellen, insbesondere von Geodaten etc., vielfältige und teilweise noch nicht vollständig abschätzbare zukünftige Analysemöglichkeiten.

Der mögliche Zugewinn muss allerdings unter den spezifischen Fragestellungen einer wissenschaftlichen Studie unter Berücksichtigung des damit verbundenen Aufwands abgeschätzt werden. Die hier und in anderen Beiträgen des Schwerpunkthefts geschilderten konkreten Themen der kleinräumigen Versorgungsforschung geben darauf vielfältige Hinweise.

Korrespondenzadresse

Dr. E. Swart
Institut für Sozialmedizin und Gesundheitsökonomie (ISMG), Medizinische Fakultät, Otto-von-Guericke-Universität Magdeburg
Leipziger Str. 44, 39120 Magdeburg
enno.swart@med.ovgu.de

Einhaltung ethischer Richtlinien

Interessenkonflikt. E. Swart, C. Stallmann, J. Powietzka und S. March geben an, dass kein Interessenkonflikt besteht.

Dieser Beitrag beinhaltet keine Studien an Menschen oder Tieren.

Literatur

1. Wennberg JE, Gittelsohn A (1973) Small area variation in health care delivery. Science 182:1102–1108
2. Wennberg JE (1984) Dealing with medical practice variations: a proposal for action. Health Aff 3:6–32

3. Wennberg JE (2010) Tracking medicine. A researcher's quest to understand health care. Oxford Univ. Press, New York
4. Swart E (2005) Kleinräumige Versorgungsforschung mit GKV-Routinedaten. In: Swart E, Ihle P (Hrsg) Routinedaten im Gesundheitswesen. Handbuch Sekundärdatenanalyse: Grundlagen, Methoden und Perspektiven. Huber, Bern, S 243–252
5. Swart E, Deh U, Robra B-P (2008) Die Nutzung der GKV-Daten für die kleinräumige Analyse und Steuerung der stationären Versorgung. Bundesgesundheitsbl Gesundheitsforsch Gesundheitsschutz 51:1183–1192
6. Nowossadeck E, Kroll LE, Nowossadeck S et al (2011) Kleinräumige Bedarfsprognosen – Eine Machbarkeitsstudie für Deutschland. Robert Koch-Institut, Berlin
7. Thode N, Bergmann E, Kamtsiuris P et al (2005) Einflussfaktoren auf die ambulante Inanspruchnahme in Deutschland. Bundesgesundheitsbl Gesundheitsforsch Gesundheitsschutz 48:296–306
8. Ozegowski S (2013) Regionale Kodierqualität ambulanter Diagnosen. G+G Wissenschaft 13(1):23–34
9. Robra B-P, Swart E, Schlichthaar H et al (1999) Reduktion der Krankenhaushäufigkeit des Diabetes mellitus nach Diabetes-Vereinbarung im ambulanten Sektor – kleinräumige Evaluation anhand regionaler Krankenhausdaten. Diabetes Stoffwechs 8:107–112
10. Hering R, Augustin J (o J) Methoden zur Zusammenführung inkompatibler räumlicher Ebenen im Gesundheitswesen am Beispiel Hamburgs. Gesundheitswesen (eingereicht)
11. March S, Iskenius M, Hardt J et al (2013) Methodische Überlegungen für das Datenlinkage von Primär- und Sekundärdaten im Rahmen arbeitsepidemiologischer Studien. Bundesgesundheitsbl Gesundheitsforsch Gesundheitsschutz 56:571–578
12. Bitzer EM, Neusser S, Lorenz C et al (2007) Krankenhaus-Rangfolgen nach Ergebnisqualität in der Hüftendoprothetik – Routinedaten mit oder ohne Patientenbefragungen? – Teil 2: Patientenbefragung in Kombination mit Routinedaten. GMS Med Inform Biom Epidemiol 3(1):Doc07
13. Hoffmann W, Bobrowski C, Fendrich K (2008) Sekundärdatenanalyse in der Versorgungsepidemiologie – Potenzial und Limitationen. Bundesgesundheitsbl Gesundheitsforsch Gesundheitsschutz 51:1193–1201
14. March S, Rauch A, Thomas D et al (2012) Datenschutzrechtliche Vorgehensweise bei der Verknüpfung von Primär- und Sekundärdaten in einer Kohortenstudie: die lidA-Studie. Gesundheitswesen 74:e122–e129
15. Swart E, March S, Thomas D et al (2011) Erfahrungen mit der Datenverknüpfung von Primär- und Sekundärdaten in einer Interventionsstudie. Gesundheitswesen 73:e126–e132
16. Schröder H, Kersting A, Gilberg R et al (2013) Methodenbericht zur Haupterhebung lidA – leben in der Arbeit. FDZ-Methodenreport 01/2013. Institut für Angewandte Sozialwissenschaft, Bonn
17. Hasselhorn HM, Peter R, Rauch A et al (o J) Cohort profile: the lidA study – a German Cohort Study on work, age, health and work participation. Int J Epidemiol (eingereicht)
18. Wichmann H-E, Kaaks R, Hoffmann W et al (2012) Die Nationale Kohorte. Bundesgesundheitsbl Gesundheitsforsch Gesundheitsschutz 55:781–789
19. Swart E, Stallmann C, Powietzka J et al (2013) Nationale Kohorte – Prätestprojekt 8. Erschließung von Sekundärdaten und Prüfung ihrer Nutzungsmöglichkeiten (PP8). Endbericht. Otto-von-Guericke-Universität, Magdeburg
20. Erbel R, Eisele L, Moebus S et al (2012) Die Heinz Nixdorf Recall Studie. Bundesgesundheitsbl Gesundheitsforsch Gesundheitsschutz 55:809–815
21. Moebus S, Andrich S (2010) Gesundheitsökonomische Begleitevaluation der Heinz Nixdorf Recall Studie. Fachlicher Abschlussbericht Teil A. Institut für Medizinische Informatik, Biometrie und Epidemiologie, Essen
22. Völzke H (2012) Study of Health in Pomerania (SHIP). Konzept, Kohortendesign und ausgewählte Ergebnisse. Bundesgesundheitsbl Gesundheitsforsch Gesundheitsschutz 55:790–794
23. Sozialgesetzbuch Zehntes Buch (SGB X) – Sozialverwaltungsverfahren und Sozialdatenschutz – in der Fassung der Bekanntmachung vom 18. Januar 2001 (BGBl. I S. 130), zuletzt durch Artikel 8 des Gesetzes vom 21. Juli 2012 (BGBl. I S. 1566) geändert. http://www.gesetze-im-internet.de/bundesrecht/sgb_10/gesamt.pdf
24. Swart E, March S, Thomas D et al (2011) Die Eignung von Sekundärdaten zur Evaluation eines Interventionsprojekts – Erfahrungen aus der AGil-Studie. Präv Gesundheitsf 6:305–311
25. Stock S, Danner M (2014) Kann die Erhebung von Einstellungen und Präferenzen die kleinräumige Versorgungsanalyse sinnvoll ergänzen? Eine gesundheitspolitische Perspektive. Bundesgesundheitsbl Gesundheitsforsch Gesundheitsschutz 57:188–196
26. Zorn U (2014) Versorgungsforschung aus Sicht der Bundessärztekammer unter Berücksichtigung kleinräumiger Analysen. Bundesgesundheitsbl Gesundheitsforsch Gesundheitsschutz 57:169–173
27. Linder R, Ahrens S, Köppel D et al (2011) Nutzen und Effizienz des Disease-Management-Programms Diabetes Mellitus Typ 2. Dtsch Ärztebl 108:155–162
28. Robra B-P (2014) John E. Wennberg, Pionier der regionalen Versorgungsforschung. Was kann eine deutsche Versorgungswissenschaft von ihm lernen? Bundesgesundheitsbl Gesundheitsforsch Gesundheitsschutz 57:164–168
29. AOK-Bundesverband, Forschungs- und Entwicklungsinstitut für das Sozial- und Gesundheitswesen Sachsen-Anhalt (FEISA), HELIOS Kliniken, Wissenschaftliches Institut der AOK (WIdO) (Hrsg) (2007) Qualitätssicherung der stationären Versorgung mit Routinedaten (QSR). Abschlussbericht. Bonn. http://wido.de/fileadmin/wido/downloads/pdf_krankenhaus/wido_kra_qsr-abschlussbericht_0407.pdf
30. Mangiapane S (2014) Lernen aus regionalen Unterschieden: Die Webplattform http://www.versorgungsatlas.de. Bundesgesundheitsbl Gesundheitsforsch Gesundheitsschutz 57:215–223
31. Kelman C, Smith L (1999) It's time: record linkage – the vision and the reality. Aust N Z J Public Health 24:100–101
32. Kelman CW, Bass AJ, Holman CDJ (2002) Research use of linked health data – a best practice protocol. Aust N Z J Public Health 26:251–255

Gute Gesundheit bei Kindern in Deutschland

Den meisten Kindern in Deutschland geht es gut oder sehr gut. Das zeigen erste Daten der Studie zur Gesundheit von Kindern und Jugendlichen in Deutschland, die Teil des langfristigen Gesundheitsmonitorings des Robert Koch-Instituts ist. 94 % der befragten Eltern schätzten den allgemeinen Gesundheitszustand ihrer Kinder als gut oder sehr gut ein. 88 % der 11-bis 17-Jährigen kamen zur selben Einschätzung. Bei der Frage zu Allergien stellte sich heraus, dass 9% der Kinder und Jugendlichen in den zwölf Monaten vor der Befragung von Heuschnupfen betroffen waren, 6 % von Neurodermitis und 4 % von Asthma. Weitere Ergebnisse der Studie zeigen, dass fast die Hälfte der 14- bis 17-Jährigen vollständig gegen HPV immunisiert ist.

In der Studie beantworteten Teilnehmer von 2009 – 2012 telefonisch Fragen zum gesundheitlichen Wohlergehen, sowie zu weiteren Themengebieten wie HPV-Impfquote und Allergien. Bei der Befragung gaben die Eltern Auskunft zur Gesundheit ihrer Kinder, ab 11 Jahren beantworteten die Kinder zusätzlich einen Teil der Fragen selbst. Untersuchungen zur psychischen Gesundheit, Motorik und körperlichen Aktivität wurden stichprobenartig durchgeführt. Die vollständigen Ergebnisse der KiGGS-Studie werden Mitte 2014 erwartet.

Weitere Informationen zur Studie zur Gesundheit von Kindern und Jugendlichen in Deutschland (KiGGS-Studie) finden Sie unter: www.kiggs-studie.de.

Quelle:
Robert Koch-Institut,
www.rki.de