CHERH Center for Health Economics Research Hannover

Leibniz Universität Hannover

Grundlagen und Methoden von GKV-Routinedatenstudien

Dipl.-Ök. Sarah Neubauer, Dr. Jan Zeidler, Dipl.-Ök. Ansgar Lange,
Prof. Dr. J.-Matthias Graf von der Schulenburg

Leibniz Universität Hannover, Center for Health Economics Research Hannover (CHERH)

Diskussionspapier Nr. 534

August 2014

Kontakt
Dipl.-Ök. Sarah Neubauer
Leibniz Universität Hannover
Center for Health Economics Research Hannover (CHERH)
Otto-Brenner-Str. 1
D-30159 Hannover
Tel.: +49 (0)511 | 762-14242
E-Mail: sn@ivbl.uni-hannover.de

Interessenkonflikte
Die vorliegende Studie wurde durch das Bundesministerium für Bildung und Forschung (BMBF) gefördert.

Zusammenfassung

Routinedatenstudien können den Akteuren des Gesundheitswesens nützliche Informationen liefern. Infolgedessen hat die Bedeutung und wissenschaftliche Nutzung von Routinedaten der gesetzlichen Krankenversicherung in den letzten Jahren stetig an Relevanz gewonnen. Bisher liegen nur sehr allgemeine Leitlinien in Bezug auf einzelne Prozessschritte einer GKV-Routinedatenstudie vor. Ziel dieses Diskussionspapier ist es daher, eine detaillierte Übersicht über relevante konzeptionelle und methodische Aspekte bei der Durchführung von GKV-Routinedatenstudien zu entwickeln, um qualitativ hochwertigere, transparentere und vergleichbarere Studien zu erhalten und den methodischen Austausch weiter zu fördern. Dabei werden die vielfältigen Publikationen auf dem Gebiet der GKV-Routinedaten systematisch aufgearbeitet sowie die Vor- und Nachteile unterschiedlicher methodischer Herangehensweisen diskutiert.

Abstract

Claims data studies are becoming an increasingly important source of information for healthcare stakeholders. The importance and scientific use of claims data of the statutory health funds has further increased. We saw a general lack of elaborated recommendations for best practices in this field and a need for a thorough overview of published methods used in each step of conducting a claims data study. This discussion paper examines and compares conceptual and methodological approaches used in claims data studies aiming to stimulate discussion on quality of the studies and to promote creation of standards and guidelines for consistent and transparent claims data studies and reports. Our results emphasize the importance guidelines in the field of claims data analyses and discussed the advantages and disadvantages of different methodological approaches.

Keywords: GKV-Routinedaten, Datenkategorien, Studiendesigns, Validierung, Datenaufbereitung, claims data, data categories, study design, validation, data processing, best practices

JEL-Classification: I13 Health Insurance, Public and Private

Inhaltsverzeichnis

Abbildungsverzeichnis

Tabellenverzeichnis

Abkürzungsverzeichnis

A	Ausschlussdiagnose
ADHS	Aufmerksamkeitsdefizit-/Hyperaktivitätsstörung
AHB	Anschlussheilbehandlung
AKR	Ambulante Kodierrichtlinien
AOK	Allgemeine Ortskrankenkasse
AU	Arbeitsunfähigkeit
Aufl.	Auflage
BBSR	Bau-, Stadt- und Raumforschung
BDSG	Bundesdatenschutzgesetz
BKK	Betriebskrankenkassen
BMG	Bundesministerium für Gesundheit
BSHG	Bundessozialhilfegesetz
BTMG	Betäubungsmittelgesetz
CD	Compact Disc
COPD	Chronic Obstructive Pulmonary Disease
CRT	Cardiac Resynchronization Therapy
DAK	Deutsche Angestellten-Krankenkasse
DaTraV	Datentransparenzverordnung
DDD	Defined Daily Dose
DEGAM	Deutsche Gesellschaft für Allgemeinmedizin und Familienmedizin
DIMDI	Deutsches Institut für Medizinische Dokumentation und Information
DMP	Disease-Management-Programm
DRG	Diagnosis Related Groups
EBM	Einheitlicher Bewertungsmaßstab
ed.	Edition
FA	Facharzt
G	Gesicherte Diagnose
G-BA	Gemeinsamer Bundesausschuss
GEK	Gmünder Ersatzkasse
GG	Grundgesetz
GKV	Gesetzliche Krankenversicherung
GKV-WSG	GKV-Wettbewerbsstärkungsgesetz

GPS	Gute Praxis Sekundärdatenanalyse
HA	Hausarzt
HDIA	Hauptdiagnose
i. V. m.	In Verbindung mit
ICD	International Classification of Diseases
IGeL	Individuelle Gesundheitsleistungen
IGES	Institut für Gesundheits- und Sozialforschung
IKK	Innungskrankenkassen
IQWiG	Institut für Qualität und Wirtschaftlichkeit im Gesundheitswesen
KBS	Knappschaft-Bahn-See
KBV	Kassenärztliche Bundesvereinigung
KV	Kassenärztliche Vereinigung
KVDT	Kassenärztliche Vereinigung-Datentransfer
MAR	Missing at random
MCAR	Missing completely at random
Morbi-RSA	Morbiditätsorientierter Risikostrukturausgleich
MPR	Medication possession ratio
No.	Number
OAR	Observed at random
OLS	Ordinary least squares / Methode der kleinsten Quadrate
OPS	Operationen- und Prozedurenschlüssel
OTC	Over the counter
PIA	Psychiatrische Institutsambulanzen
PKV	Private Krankenversicherung
PZN	Pharmazentralnummer
Q	Quartal
RSA	Risikostrukturausgleich
SGB V	Fünftes Sozialgesetzbuch
SGB IX	Neuntes Sozialgesetzbuch
SGB X	Zehntes Sozialgesetzbuch
SQL	Structured Query Language
STROBE	Strengthening the Reporting of Observational Studies in Epidemiology
SVLFG	Sozialversicherung für Landwirtschaft, Forsten und Gartenbau
SVR	Sachverständigenrat

TK Techniker Krankenkasse

V Verdachtsdiagnose

vdek Verband der Ersatzkassen

VE Versicherter

WIdO Wissenschaftliches Institut der Ortskrankenkassen

WINEG Wissenschaftliches Institut der Techniker Krankenkasse für Nutzen und Effizienz im Gesundheitswesen

Z „Zustand nach“-Diagnose

1 Hintergrund und Motivation

Die Routinedaten der gesetzlichen Krankenversicherung (GKV) spielen seit einigen Jahren – beispielsweise in der Versorgungsforschung – eine große Rolle und werden immer häufiger für wissenschaftliche Zwecke genutzt. Diese steigende Bedeutung spiegelt sich unter anderem in der zunehmenden Anzahl der routinedatenbasierten Publikationen (Hoffmann 2009), jährlich veranstalteten Fachtagungen und Kongressen sowie öffentlichen Förderinitiativen wider. Darüber hinaus lässt sich die wachsende Anzahl routinedatenbasierter Forschungsprojekte anhand der Projektdatenbank „Versorgungsforschung Deutschland“ aufzeigen, die bereits heute ein breites Spektrum an GKV-Routinedatenstudien zu ganz unterschiedlichen Forschungsfragen umfasst (IMVR und WINEG). Der Bedarf an validen Datengrundlagen zur Beschreibung des Versorgungsgeschehens wird aufgrund der vielfältigen Herausforderungen bei der notwendigen Transformation des Gesundheitswesens und den steigenden technischen Möglichkeiten in Zukunft noch weiter wachsen.

Unter GKV-Routinedaten werden Abrechnungsdaten der Krankenkassen verstanden. Die umfassende elektronische Dokumentation des Versorgungsgeschehens durch die Krankenkassen ist für die Erfüllung administrativer Aufgaben erforderlich. Bei nahezu allen Kontakten des Patienten mit dem Gesundheitssystem werden relevante Informationen dokumentiert und an die Krankenkassen übermittelt. Ein weiter gefasster, dennoch auch häufig verwendeter Begriff ist der der sogenannten „Sekundärdaten“. Hierunter werden alle Daten subsumiert, „die einer Auswertung über ihren originären, vorrangigen Verwendungszweck hinaus zugeführt werden“ (AGENS 2012). Der primäre Erhebungsanlass ist abgekoppelt von der nachfolgenden Nutzung. Unter einer Sekundärdatenanalyse wird somit die Nutzung von Daten im Rahmen wissenschaftlicher oder praxisrelevanter Untersuchungen ohne direkten Bezug zum primären Erhebungsanlass verstanden (AGENS 2012).

Dem Informations- und Wissensmanagement kommt eine entscheidende Rolle bei der Verbesserung der Versorgung sowie bei der Erschließung von Wirtschaftlichkeitsreserven zu. Die Krankenkassen können ihre Daten daher zu Forschungszwecken, zur internen Bedarfsplanung sowie für die Entwicklung und Evaluation von Versorgungskonzepten selbst nutzen. Auch können Dritte im Auftrag der Krankenkassen mit der Forschung beauftragt werden. Insbesondere für die Versorgungsfor-

schung bieten sich GKV-Routinedaten aus verschiedenen Gründen an. So können die GKV-Routinedatenstudien den Akteuren des Gesundheitswesens wie der Politik, Leistungserbringern oder Krankenkassen nützliche Informationen zur Entscheidungsfindung, Evaluation verschiedener Versorgungsprogramme, Qualitätssicherung sowie Weiterentwicklung des Gesundheitswesens liefern (Mansky et al. 2012). Des Weiteren eignen sie sich zur Beschreibung komplexer Versorgungsprozesse im Gesundheitswesen, zur Optimierung des Leistungsgeschehens sowie für epidemiologische Analysen z. B. zur Inzidenz- und Prävalenzschätzung (Zeidler und Braun 2012; Schubert et al. 2008). Der Alltagsbezug ist eine weitere Stärke von GKV-Routinedaten. Sie spiegeln die in der Versorgungsrealität eingetretenen Leistungsverbräuche wider, ohne strenge Ein- und Ausschlusskriterien für Probanden oder Kontrollkriterien wie in Experimenten z. B. in klinischen Studien zu erheben (Swart und Ihle 2008). Des Weiteren sind der Aufwand und die Kosten bei der Gewinnung und Nutzung dieser Datenquelle gering, da abrechnungsrelevante Informationen routinemäßig von den Krankenkassen erhoben werden. Eine Verzerrung durch Nichtteilnahme (Non-Response) oder selektives Erinnern (Recall Bias) existiert – anders als bei der primären Datenerhebung – nicht. Im Gegensatz zu Sekundärdaten werden unter Primärdaten Daten verstanden, die für empirische Untersuchungen neu gewonnen und erhoben werden (Pirk und Schöffski 2012). Weiterhin lassen GKV-Routinedaten auch Analysen von Personengruppen zu, die sonst üblicherweise eher schwer durch eine Primärdatenerhebung erfasst werden können. Hierzu zählen beispielsweise Kinder, Schwerstkranke, Demente oder Bewohner von Altenheimen (Hoffmann und Icks 2012).

Der Datenzugang zu den GKV-Routinedaten wurde für Forschungseinrichtungen in den letzten Jahren durch verschiedene Gesetze vereinfacht. Insbesondere die Implementierung des Datenpools des Deutschen Institut für Medizinische Dokumentation und Information (DIMDI) ermöglicht nun einem breiten Nutzerkreis den Zugriff auf Abrechnungsdaten der Krankenkassen. In dem Gutachten des Sachverständigenrats (SVR) zur Begutachtung der Entwicklungen im Gesundheitswesen wird eine Ausweitung der Versorgungsforschung schon seit vielen Jahren gefordert (SVR 2002). Trotz dieses rechtlich-politischen Zuspruchs und der steigenden Popularität dieser Datenquelle liegen bisher für die konkrete Validierung, Auswertung und die einzelnen Prozessschritte einer GKV-Routinedatenstudie nur sehr allgemeine Leitlinien vor. Es mangelt heute immer noch an einem einheitlichen Methodenspektrum und an me-

thodischen Standards, wie sie in anderen Forschungsfeldern schon lange existieren. Nur durch einen systematischen Standardisierungsprozess kann jedoch die Vergleichbarkeit, Transparenz und Qualität von GKV-Routinedatenstudien weiter erhöht werden. Um ihr volles Potenzial entfalten zu können, muss diese Datenquelle daher durch eine noch intensivere Harmonisierung der Methoden und Kriterien für eine breitere wissenschaftliche Auswertung nutzbar gemacht werden.

Ziel dieser Studie ist es daher, eine detaillierte Übersicht über relevante konzeptionelle und methodische Aspekte bei der Durchführung von GKV-Routinedatenstudien zu entwickeln, um qualitativ hochwertigere, transparentere und vergleichbarere Studien zu erhalten. Hierbei sollen insbesondere konkrete Empfehlungen zur Lösung methodischer Herausforderungen gegeben werden. Mit dem vorliegenden Diskussionspapier sollen das Interesse und das Verständnis für dieses Forschungsgebiet geweckt und Wissenschaftler, die zum ersten Mal mit GKV-Routinedaten arbeiten, bei methodischen Fragen unterstützt werden. Die Begriffe „Leitfaden“ und „Handbuch“ wurden vermieden, da die Arbeit an diesem Diskussionspapier gezeigt hat, dass aufgrund der facettenreichen Fragestellungen nur begrenzt allgemeine Standards gesetzt werden können. Des Weiteren gibt dieses Diskussionspapier angesichts der Dynamik des Forschungsfeldes lediglich eine Momentaufnahme bezüglich der Methoden und technischen Aspekte der GKV-Routinedatenanalyse. Das Diskussionspapier erhebt auch angesichts der raschen Entwicklung des Forschungsgebietes nicht den Anspruch, einen dauerhaften allgemeinen Goldstandard zu präsentieren. Dennoch werden wichtige Gesichtspunkte der GKV-Routinedatenanalyse skizziert sowie die Vor- und Nachteile verschiedener methodischer Herangehensweisen kritisch diskutiert. Jedoch müssen diese Verfahren projekt- und fragestellungsspezifisch angepasst werden. Die in diesem Diskussionspapier vorgestellten Methoden und Studiendesigns sollen so weit wie möglich dazu beitragen, den notwendigen methodischen Standardisierungsprozess voranzutreiben. Nur so können die Potenziale von GKV-Routinedatenstudien in Zukunft vollumfänglich ausgeschöpft werden.

Der Fokus des Diskussionspapiers liegt dabei ausschließlich auf den Aspekten der GKV-Routinedatenanalyse. Andere Routinedatenquellen, wie beispielsweise Daten der Pflegekassen oder der Rentenversicherung, Routinedaten der privaten Krankenversicherung (PKV) und grundlegende Aspekte der kostenträgerunabhängigen Routinedatenanalyse, wie z. B. methodische Herausforderungen bei der Verknüpfung

von Primär- und Sekundärdaten, müssen zugunsten einer stringenten inhaltlichen Orientierung ausgeklammert werden. Die Gliederung dieses Diskussionspapiers orientiert sich primär an den klassischen Prozessschritten einer GKV-Routinedatenstudie. Im nachfolgenden Kapitel 2 werden daher zunächst die einzelnen Prozessschritte einer GKV-Routinedatenstudie erläutert und grundsätzliche Aspekte zum Datenzugang thematisiert. Auch die verfügbaren Datenkategorien werden aufgezeigt und die wissenschaftlich nutzbaren Variablen detailliert beschrieben. Des Weiteren wird auf datenschutzrechtliche Besonderheiten eingegangen. In Kapitel 3 werden dann ausgewählte Studiendesigns dargestellt. Mit den vorgeschlagenen Designs lassen sich viele Fragestellungen auf Basis der GKV-Routinedaten beantworten. Hierunter fallen beispielsweise gesundheitsökonomische Analysen, regionale Auswertungen sowie auch die Überprüfbarkeit von Compliance und Leitliniengerechtigkeit. Auf die Datenextraktion und unterschiedliche Validierungsstrategien wird in Kapitel 4 eingegangen. Dies umfasst sowohl die interne als auch die externe Validierung, aber auch die Plausibilitäts- und Vollständigkeitskontrolle der Datensätze. Der Aufwand der Datenaufbereitung sollte dabei nicht unterschätzt werden. Die GKV-Routinedaten werden zu Abrechnungszwecken erhoben und müssen für die wissenschaftliche Nutzung noch aufbereitet, gegebenenfalls umcodiert und nutzbar gemacht werden. Auf in diesem Zusammenhang auftretende spezifische methodische Schwierigkeiten und Herausforderungen soll daher in Kapitel 5 aufmerksam gemacht werden. Zum Abschluss wird in Kapitel 6 auf die Limitationen der Datenquelle hingewiesen. Am Ende eines jeden Kapitels finden sich Empfehlungen, zusammenfassende Tabellen und Grafiken.

Die Autoren möchten allen Personen und Institutionen danken, die zur Entstehung dieses Diskussionspapiers beigetragen haben. Besonderer Dank für die zahlreichen Diskussionen und Anregungen gilt dabei der Arbeitsgruppe „Routinedatenanalysen“, die sich aus Mitarbeitern der Universität Bielefeld, der Medizinischen Hochschule Hannover, der Leibniz Universität Hannover und der Herescon GmbH zusammensetzt. Bei der Erstellung dieses Diskussionspapiers wurde systematisch die vorhandene Literatur berücksichtigt und die bereits existierenden umfassenden Vorarbeiten herangezogen. Sollten dabei methodische Beiträge nicht angemessen gewürdigt worden sein, so bitten wir um Hinweise. Das Ziel dieses Diskussionspapiers ist es, eine methodisch orientierte Ergänzung zu der existierenden Standardliteratur zu implementieren. Dabei werden die vielfältigen Publikationen auf dem Gebiet der GKV-

Routinedaten systematisch aufgearbeitet sowie die Vor- und Nachteile unterschiedlicher methodischer Herangehensweisen diskutiert. Wir hoffen mit diesem Diskussionspapier den methodischen Austausch noch weiter zu fördern, die methodische Standardisierung voranzubringen sowie eine differenzierte methodische Orientierung und vielfältige Hilfestellungen für alle GKV-Routinedatennutzer anbieten zu können.

2 Prozessschritte und Datenverfügbarkeit

In diesem Abschnitt wird der Prozess einer GKV-Routinedatenstudie systematisch dargestellt. Zudem werden Möglichkeiten für den Datenzugang aufgezeigt und es wird ein Überblick über wissenschaftlich nutzbare Datenkategorien gegeben. Der Umfang der von den Leistungserbringern übermittelten Daten sowie der zu den Versicherten erhobenen Merkmalen ist zwischen den gesetzlichen Krankenkassen weitgehend vergleichbar. Jedoch bestehen Unterschiede bezüglich der Datenhaltung, Datenpflege, Datenverfügbarkeit, Datenstruktur sowie Datendarstellung (Grobe 2008). Aufgrund der zahlreichen Variablen und Ausprägungen, die teilweise durch die Krankenkassen lediglich für interne Betriebszwecke dokumentiert werden und für Wissenschaftler nur von begrenztem Interesse sind, werden ausschließlich Variablen dargestellt, die sich bisher als wissenschaftlich nutzbar erwiesen haben. Ein Anspruch auf Vollständigkeit kann daher nicht erhoben werden.

2.1 Prozessschritte

Am Anfang einer jeden GKV-Routinedatenstudie ist gemäß der Guten Praxis Sekundärdatenanalyse (GPS) ein Studienplan anzufertigen (AGENS 2012). Dieser sollte Informationen zum Studientyp, Studiendesign, Projektziel, zur Forschungsfrage, zu Kooperationspartnern sowie den Grundlagen und die Nennung der Indikation beinhalten (AGENS 2012; Scharnetzky et al. 2013). Damit der Datenhalter abschätzen kann, welche Daten für die Studie zur Verfügung gestellt werden sollen, sind eine Auflistung der relevanten Leistungsbereiche und eine Spezifikation der Variablen notwendig. Vorher müssen alle Rahmeninformationen, wie Studienpopulation bzw. Aufgreifkriterien, Analysezeitraum und Datenbasis, schriftlich festgelegt werden. Nach der Erstellung des Studienplans muss, wenn die Auswertung auf Einzelkassenbasis basieren soll, mindestens eine Krankenkasse als Kooperationspartner gefunden werden. Hierbei sind die Größe und Regionalität der jeweiligen Krankenkasse im Kontext der zu beantwortenden Forschungsfragen zu berücksichtigen (Näheres siehe Abschnitt 2.2.1). Falls bei der Studie ein Antrag auf Forschungsförderung gestellt wird, kann die geplante Kooperation vorab über einen Letter of Intent fixiert werden. Der Letter of Intent ermöglicht der Forschungseinrichtung eine frühzeitige Planbarkeit des Datenzugangs und kann als Nachweis für die Durchführbarkeit der Studie dem Formantrag beigelegt werden.

Häufig sind die Krankenkassen solchen Kooperationen gegenüber aufgeschlossen. Auch die zunehmende Implementierung von krankenkasseninternen Forschungseinrichtungen wie beispielsweise dem Wissenschaftlichen Institut der Techniker Krankenkasse für Nutzen und Effizienz im Gesundheitswesen (WINEG) und dem Wissenschaftlichen Institut der AOK (WIdO) spiegelt den Stellenwert von Krankenkassendaten zur Beantwortung wissenschaftlicher Fragestellungen wider. Aber auch unterschiedliche Berichte und Reports auf Basis dieser Datenquelle häufen sich in den letzten Jahren. Der Gesundheitsreport der Deutschen Angestellten-Krankenkasse (DAK) ist ein gutes Exempel für die Nutzung der Routinedaten durch Krankenkassen. Die DAK analysiert in Kooperation mit dem Institut für Gesundheits- und Sozialforschung (IGES) jährlich den Krankenstand ihrer erwerbstätigen Mitglieder. Die nach Alter und Geschlecht getrennten Auswertungen haben zum Ziel, relevante Hintergrundinformationen für Unternehmen bereitzustellen, um z. B. ein betriebliches Gesundheitsmanagement aufzubauen, das die Gesundheit der Arbeitnehmer fördert und systematisch Belastungen, beispielsweise durch Stress, abbaut (DAK Forschung und IGES Institut GmbH 2013). Auch die Barmer GEK veröffentlich seit einigen Jahren jährlich einen Krankenhausreport (Barmer GEK 2010-2014). Als aktuell größte Krankenkasse Deutschlands (Deutsches Ärzteblatt 2014) sucht die Techniker Krankenkasse auch mithilfe ihres angeschlossenen WINEG-Instituts schon seit einigen Jahren nach Antworten auf Fragen zur Verbesserung der gesundheitlichen Versorgung (WINEG). Auch die Allgemeinen Ortskrankenkassen stellen z. B. mit der „Versichertenstichprobe AOK Hessen/KV Hessen“ ihre GKV-Routinedaten für versorgungsepidemiologische Analysen zur Verfügung, ähnlich wie die Betriebskrankenkassen (Ihle et al. 2005; Hoffmann et al. 2004). Neben der Kooperation mit Einzelkassen besteht ein alternativer Datenzugang in der Nutzung des neu eingeführten DIMDI-Datenpools. Die Voraussetzungen und Möglichkeiten der Verwendung dieses Datenpools werden in Abschnitt 2.2.2 beschrieben.

Im nächsten Schritt sollte auf Basis der Projektskizze eine Datenanfrage bei der Krankenkasse gestellt werden. Willigt eine Krankenkasse ein, so ist ein Kooperationsvertrag von den Projektpartnern aufzusetzen. Dieser sollte die Projektleitung und Projektdurchführung klar regeln sowie vertraglich die Zuständigkeiten und Rahmenbedingungen, beispielsweise zur Datenschutzinfrastruktur, festhalten. Die Projektskizze kann dem Kooperationsvertrag beigefügt werden (AGENS 2012; Scharnetzky et al. 2013).

Nach Abschluss des Kooperationsvertrages folgen die Extraktion und Lieferung der Daten durch den Dateneigner. In der Regel extrahiert eine Fachabteilung der jeweiligen Krankenkasse die relevanten Variablen der einzelnen Leistungsbereiche. Dieses Vorgehen kann unter Umständen einige Zeit in Anspruch nehmen. Häufig treten viele interne und externe Datenanfragen parallel bei den Krankenkassen auf, was zu personellen Kapazitätsengpässen führen kann (Scharnetzky et al. 2013). Potenzielle zeitliche Verzögerungen sollten in die Zeitplanung des Projektes daher mit eingeplant werden. Wenn der Forscher die Datenstruktur und das Datawarehouse des Dateneigners (Definition siehe Hoffmann und Glaeske 2011) gut kennt, wäre auch die Zusendung eines SQL-Skripts zur Unterstützung des Extraktionsprozesses denkbar. Dieses Skript müsste dann lediglich vor Ort angestoßen werden und könnte automatisch alle relevanten Extraktionsschritte durchlaufen. Laut der GPS sollte vorher ein Probedatensatz zur Verfügung gestellt werden, um die Verwendbarkeit der Daten, insbesondere hinsichtlich der Datenformate und -struktur, beurteilen zu können (A-GENS 2012).

Eine enge Abstimmung und Zusammenarbeit zwischen Dateneigner und Forscher ist sinnvoll, um mehr über die individuelle kassenspezifische Datenverfügbarkeit zu erfahren. So sind Treffen zwischen den Projektpartnern und regelmäßige Statusberichte, in denen auch über Herausforderungen diskutiert werden kann, empfehlenswert.

Die Datenlieferung erfolgt anschließend über eine gesicherte Onlineplattform oder über einen physischen Datenträger. Auf der Onlineplattform werden die Daten zeitlich begrenzt zur Verfügung gestellt. Nach Ablauf einer Frist werden die Dateien gelöscht. Die Übertragung erfolgt verschlüsselt und für den Abruf der Dateien ist daher ein die Sicherheitsbestimmungen erfüllendes Kennwort erforderlich. Dieses Kennwort wird in geeigneter Form, getrennt von den anderen Zugangsdaten, vom Datenhalter übermittelt. Aus datenschutzrechtlichen Gründen wird der Datenträger ausschließlich durch einen Boten oder per Einschreiben verschickt und persönlich überreicht. Diese Maßnahmen verhindern den Zugriff unbefugter Dritter auf die Daten (Grobe und Ihle 2005). Bei dem Austausch der Informationen ist es wichtig, sich auf eine gemeinsame Datenschnittstelle zu verständigen und beim Datenim- und -export eine einheitliche oder kompatible Software zu verwenden. Übliche Programme für die Datenhaltung und die statistischen Auswertungen sind SQL-Server, SAS, STATA, SPSS und Datenbanksoftwarepakete, wie Oracle und Access (Grobe und Ihle 2005).

Die Qualität der Daten muss aufgrund des Sekundärdatencharakters vor jeder wissenschaftlichen Analyse überprüft werden. Grund dafür ist, dass die Daten primär zu einem anderen Zweck und von anderen Personen erhoben worden sind. Auf die primäre Datenerhebung und die Qualität der Dokumentation hat der Sekundärdatennutzer somit keinen Einfluss, sodass eine begleitende Qualitätssicherung immer erforderlich ist. Zur Qualitätssicherung sind daher unter anderem Validierungsverfahren einzusetzen. Unter diese fallen die interne und externe Validierung sowie Plausibilitäts- und Vollständigkeitsprüfungen (siehe Kapitel 4). Eine weitere Maßnahme ist die Datenaufbereitung. So müssen die im vorherigen Schritt identifizierten Datenauffälligkeiten entfernt, berichtigt oder mögliche Codierungen angepasst werden. Sämtliche Datenaufbereitungsschritte müssen dokumentiert werden, um sie im späteren Projektverlauf nachvollziehen zu können. Hilfreich kann hier je nach verwendeter Software ein Skript oder eine Syntax sein (AGENS 2012).

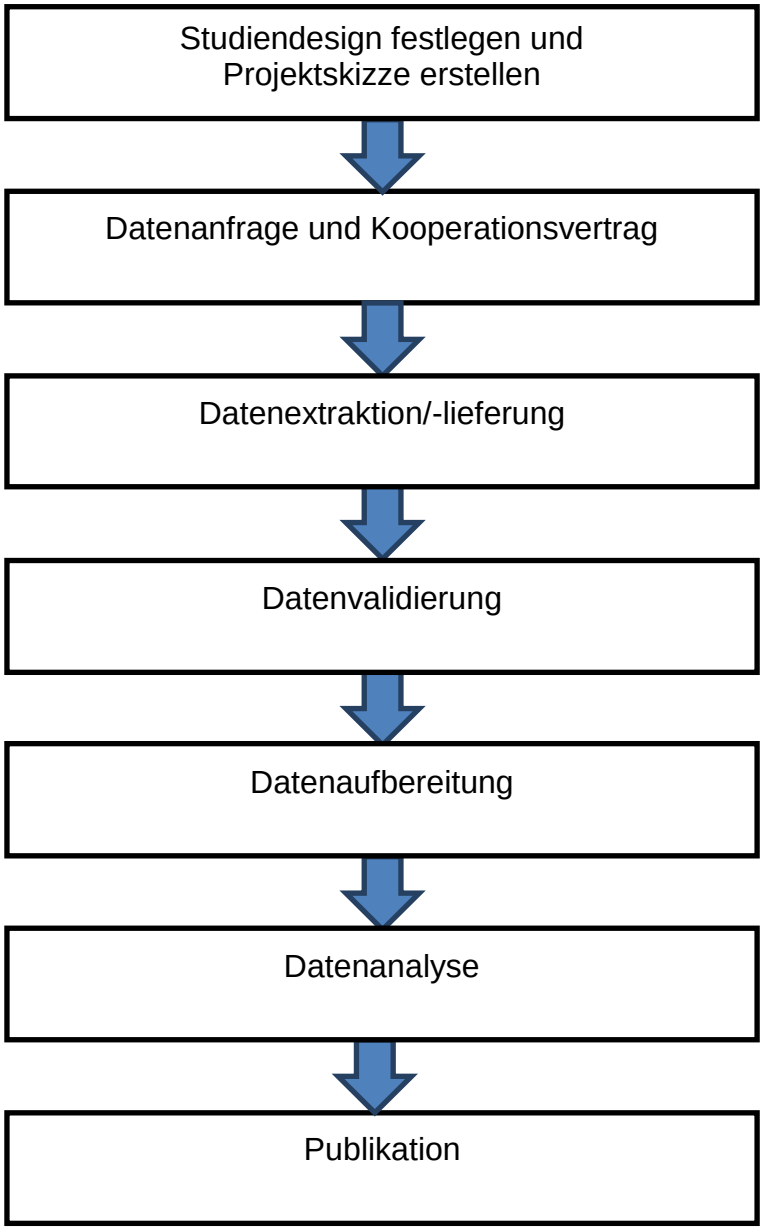
Laut der GPS soll die Datenauswertung mithilfe „adäquater Methoden erfolgen“ (AGENS 2012). Diese recht allgemein gehaltene Aussage wird in diesem Diskussionspapier im Kapitel 5 detailliert aufgearbeitet und es werden Empfehlungen zur Datenanalyse gegeben. Alle im Studienplan aufgeführten Auswertungsschritte, wie z. B. die Selektion der Studienpopulation, müssen nachvollziehbar und rekonstruierbar sein.

Als finaler Prozessschritt ist eine Publikation in einem Fachmedium anzustreben. In der Veröffentlichung sollten ausgewählte Ergebnisse der Studie systematisch und transparent zusammenfasst sowie die Ergebnisse interpretiert und kritisch diskutiert werden (AGENS 2012). Eine einheitliche Empfehlung für einen Berichtsstandard von GKV-Routinedatenstudien existiert bislang jedoch weder national noch international (Swart und Schmitt 2014). Im Jahr 2007 wurde als Berichtsstandard für epidemiologische Beobachtungsstudien das Strengthening the Reporting of Observational Studies in Epidemiology (STROBE-) Statement eingeführt. Das STROBE-Statement enthält eine Checkliste, die eine Hilfestellung geben soll, wie die Ergebnisse zu strukturieren und zu berichten sind (Elm et al. 2008). Aktuell gibt es von Swart und Schmitt Bestrebungen einen solchen Berichtsstandard für GKV-Routinedaten zu entwickeln (Swart und Schmitt 2014). Sie prüften die STROBE-Kriterien auf ihre Kompatibilität mit der GPS und auf die Anwendbarkeit auf Sekundärdaten. Wurden die

STROBE-Kriterien den Sekundärdatenanalysen nicht gerecht, wurden Ergänzungen zu den bisherigen Ausführungen formuliert.

Die nachfolgende Abbildung 1 fasst alle wesentlichen Prozessschritte noch einmal übersichtsweise zusammen.

Abbildung 1: Prozessschritte einer GKV-Routinedatenstudie



Quelle: eigene Darstellung

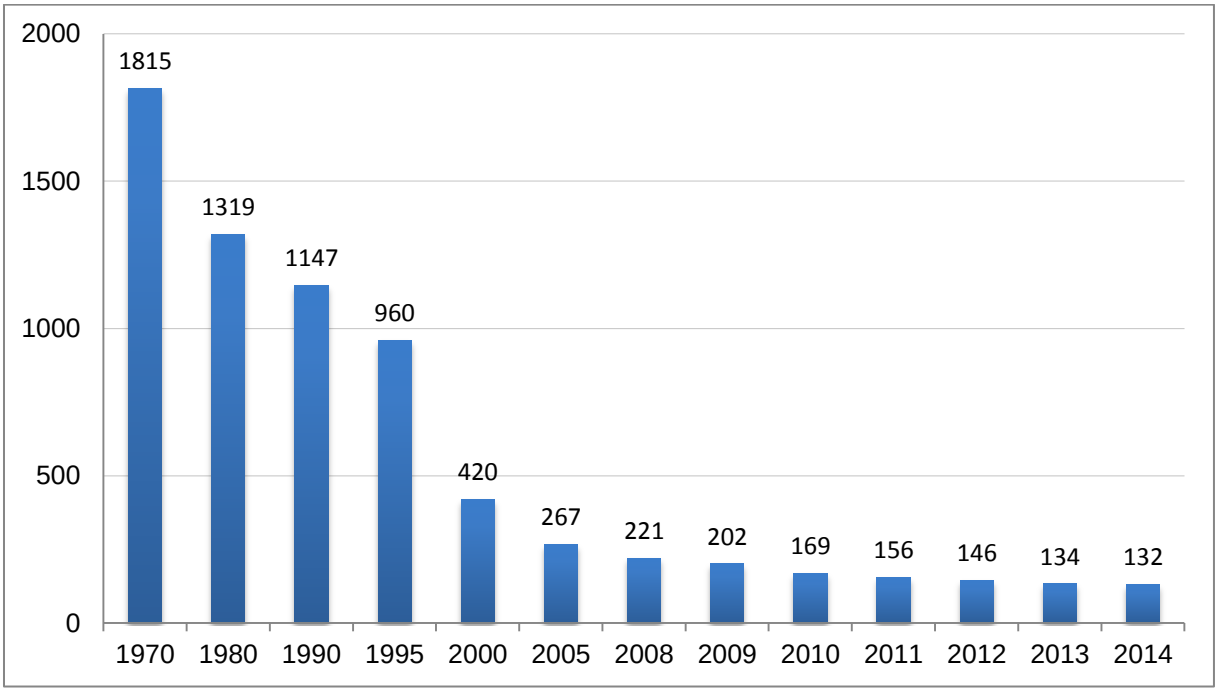
2.2 Zugang zu GKV-Routinedaten

Grundsätzlich existieren mehrere Möglichkeiten für einen Zugang zur wissenschaftlichen Nutzung von GKV-Routinedaten. Zum einen kann an eine oder mehrere einzelne Krankenkassen herangetreten werden. Zum anderen wurde durch die Datentransparenzverordnung ein neuer Zugang zu GKV-Routinedaten geschaffen – der sogenannte Datenpool des DIMDI.

2.2.1 Einzelkassen

Den Krankenkassen ist es laut § 299 SGB V erlaubt, Daten zu erheben, zu verarbeiten und z. B. zur Qualitätssicherung zu nutzen. Mit Einführung des Risikostrukturausgleichs (RSA) und des späteren morbiditätsorientierten Risikostrukturausgleichs (Morbi-RSA) wurde die elektronische Datenübermittlung immer häufiger eingesetzt (Vauth 2010; GKV-Spitzenverband 2012). So ist es wissenschaftlichen Institutionen möglich, einzelne Krankenkassen anzusprechen und mit ihnen Projekte zur Versorgungsforschung zu initiieren. Ein bedeutender Vorteil von GKV-Routinedatenstudien ist die große Datenbasis, da rund 70 Mio. Personen in Deutschland gesetzlich krankenversichert sind (Bundesministerium für Gesundheit 2013). In den letzten Jahren – insbesondere durch das GKV-Wettbewerbsstärkungsgesetz (GKV-WSG) – ist eine starke Konzentration des Krankenkassenmarktes z. B. durch Fusionen zu beobachten (GKV-Spitzenverband 2014b). So hat sich die Anzahl der Krankenkassen von ehemals 1815 im Jahre 1970 bis hin zu derzeit 132 gesetzlichen Krankenkassen (GKV-Spitzenverband 2014b) im Laufe der Jahre stetig reduziert (siehe Abbildung 2).

Abbildung 2: Anzahl der Krankenkassen im Zeitablauf seit 1970 (Angaben zum Stichtag 1. Januar)



Quelle: GKV-Spitzenverband (2014b)

Die 132 gesetzlichen Krankenkassen teilen sich in elf Allgemeine Ortskrankenkassen (AOK), sechs Krankenkassen, die zu dem Verband der Ersatzkassen (vdek) gehören, sechs Innungskrankenkassen (IKK), 106 Betriebskrankenkassen (BKK), eine Knappschaft (KBS) und eine Landwirtschaftliche Krankenkasse (SVLFG) auf (GKV-Spitzenverband 2014b; Bundesministerium für Gesundheit 2013). Bei der Auswahl einer oder mehrerer kooperierender Krankenkassen ist zu beachten, dass nicht alle Krankenkassen bundesweit tätig sind, was für einige Analysten eine Herausforderung darstellt.

Sollen aus GKV-Routinedaten repräsentative Aussagen generiert werden, so empfiehlt es sich größere Krankenkassen anzusprechen. Insbesondere um überregionale Aussagen treffen zu können oder das gesamte GKV-System valide abbilden zu können, ist es sinnvoll, eine möglichst große und repräsentative Stichprobe zu akquirieren. Seit Mitte der 1990er-Jahre, mit der Liberalisierung des Krankenkassenmarktes und der damit verbundenen freien Krankenkassenwahl (SGB V § 175), entstand mehr Wettbewerb und Dynamik zwischen den Krankenkassen. Dies wirkt sich auch auf die Mitgliederstruktur aus. Dennoch existieren heute noch einige Krankenkassen, die lediglich regional, z. B. nur innerhalb einzelner Bundesländer, tätig sind. Andere hingegen sind betriebsbezogen und damit ausschließlich für Mitarbeiter wählbar

(GKV-Spitzenverband 2014b). Viele Krankenkassen sind jedoch mittlerweile bundesweit tätig. Hoffmann und Icks untersuchten in einer Studie die Versichertenstrukturunterschiede und die Auswirkung dieser Divergenz auf die Versorgungsforschung (Hoffmann und Icks 2012). Sie kamen durch ein logistisches Regressionsmodell zu dem Schluss, dass trotz der freien Krankenkassenwahl und den damit einhergehenden Wechselmöglichkeiten der Versicherten zwischen den Krankenkassen in der Versicherten- und Morbiditätsstruktur erhebliche Unterschiede existieren. Diese lassen sich jedoch nicht ausschließlich durch das Alter und das Geschlecht erklären. Die mangelnde Repräsentativität ist jedoch nicht nur auf GKV-Routinedaten begrenzt, sondern gilt teilweise auch für Primärdatenerhebungen. So basieren beispielsweise die in der Region Augsburg durchgeführten KORA-Studien (Kooperative Gesundheitsforschung in der Region Augsburg) (Werner et al. 2005) oder die in Vorpommern realisierten SHIP-Studien (Study of Health in Pomerania) auf regionalen Stichproben (Völzke et al. 2011).

Die Generalisierbarkeit von Ergebnissen ist daher zu diskutieren und die Versicherten- sowie Morbiditätsstruktur der kooperierenden Krankenkasse zu prüfen. Es existieren bereits anerkannte Methoden zu einer bundesweiten Standardisierung, die in Kapitel 5.5 näher erläutert werden.

Da es sich bei den GKV-Routinedaten um personenbezogene Daten im Sinne des § 67 SGV X handelt, unterliegt deren Nutzung einer Reihe von datenschutzrechtlichen Aspekten, die genauer in Kapitel 2.3 beschrieben werden.

Empfehlungen

- Bei der Auswahl einer oder mehrerer kooperierender Krankenkassen ist zu beachten, dass nicht alle Krankenkassen bundesweit tätig sind
- Eine höchstmögliche Repräsentativität der Analyseergebnisse ist durch eine geeignete Krankenkassenwahl sicherzustellen
- Potenzielle zeitliche Verzögerungen sollten in die Zeitplanung des Projektes eingeplant werden
- Eine enge Abstimmung und Zusammenarbeit zwischen Dateneigner und Forscher ist zu empfehlen
- Die Kompatibilität der verwendeten Softwareanwendungen ist sicherzustellen

2.2.2 Datenpool

Eine umfassende, auf den Einzeldaten aller gesetzlichen Krankenkassen basierende Routinedatenquelle ist der neu eingeführte Datenpool, der seit Anfang des Jahres 2014 für Auswertungen zur Verfügung steht (DIMDI 2013a). Am 18.09.2012 wurde dem DIMDI mit Inkrafttreten der „Verordnung zur Umsetzung der Vorschriften über die Datentransparenz“ (Datentransparenzverordnung (DaTraV) nach §§ 303a bis 303e Sozialgesetzbuch V) die Verwaltung des Informationssystems Versorgungsdaten (DaTraV-Daten) übergeben (DaTraV 2012). Dieses Informationssystem beinhaltet zum einen eine Vertrauensstelle, die für die Verschlüsselung der Versichertenpseudonyme zuständig ist, und zum anderen eine Datenaufbereitungsstelle (Krüger-Brand 2013).

Der Datenpool besteht aus den beim Bundesversicherungsamt zusammenfließenden und für den morbiditätsorientierten Risikostrukturausgleich (Morbi-RSA) bestimmten gesetzlichen Krankenkassendaten (Müller 2012; GKV-Spitzenverband 2012). Die Morbi-RSA-Daten werden jährlich von den Krankenkassen an das Bundesversicherungsamt gemeldet und nach Plausibilitätsprüfungen und Korrekturmeldungen an das DIMDI übermittelt. Der Datenpool des DIMDI umfasst Versichertenstammdaten, Diagnosen sowie Leistungsausgaben der stationären und ambulanten Behandlung sowie Daten zur Arzneimittelversorgung, die mittels Pseudonym verknüpft werden können (Müller 2012; DIMDI 2014a).

Mit diesem, auf einer Oracle-Datenbank basierenden Informationssystem sind flächendeckende und sektorübergreifende Analysen aller gesetzlich Krankenversicherten über mehrere Jahre hinweg möglich. Das heißt, nicht nur krankenkassenspezifische Auswertungen, sondern auch deutschlandweite Analysen des Versorgungsgeschehens sind nun im Rahmen der Versorgungsforschung umsetzbar.

Per Gesetz wurde ein vorgegebener Nutzerkreis definiert, der diese Daten zu Forschungszwecken verwenden darf (Krüger-Brand 2013). Nutzungsberechtigt sind laut § 303e Sozialgesetzbuch V die Krankenkassen selbst und deren Verbände, die Kassenärztlichen Vereinigungen, zuständigen Landes- und Bundesbehörden, der G-BA, das IQWiG, Patientenvertretungen, Hochschulen und sonstige Einrichtungen mit der Aufgabe unabhängiger wissenschaftlicher Forschung (sofern die Daten wissenschaftlichen Vorhaben dienen) (SGB V 2014).

Neben diesem vorgegebenen Nutzerkreis wird auch der Nutzungszweck konkret vorgegeben (§ 303e SGB V):

- Wahrnehmung von Steuerungsaufgaben durch die Kollektivvertragspartner,
- Verbesserung der Qualität der Versorgung,
- Planung von Leistungsressourcen (z. B. Krankenhausplanung),
- Längsschnittanalysen über längere Zeiträume, Analysen von Behandlungsabläufen, Analysen des Versorgungsgeschehens zum Erkennen von Fehlentwicklungen und von Ansatzpunkten für Reformen (Über-, Unter- und Fehlversorgung),
- Unterstützung politischer Entscheidungsprozesse zur Weiterentwicklung der gesetzlichen Krankenversicherung,
- Analyse und Entwicklung von sektorenübergreifenden Versorgungsformen sowie von Einzelverträgen der Krankenkassen.

Die Auswertung soll zunächst ausschließlich mittels Datenfernverarbeitung durchgeführt werden. Hierbei wird eine Datenanfrage per SQL-Skript vom Datennutzer an die Datenverarbeitungsanlage des DIMDI gestellt, um sie dort zu verarbeiten. Das DIMDI versendet anschließend die Ergebnisse an die forschende Institution. Jedoch werden lediglich aggregierte Versorgungsdaten als Ergebnis übermittelt. Das Angebot von Analysen an Gastarbeitsplätzen ist künftig ebenfalls geplant, sodass auch Auswertungen pseudonymisierter Einzeldaten durchführbar sind.

Die Finanzierung des Datenpools stützt sich auf Geldmittel der gesetzlichen Krankenkassen sowie auf Mittel aus Nutzungsentgelten (DIMDI 2014c). Die Nutzungsgebühren gliedern sich in eine Grundgebühr in Höhe von 200 € für die Bearbeitung eines Antrags und einer Zusatzgebühr von 300 € pro ausgewerteten Jahrgang mit Hilfe von standardisierten Datensätzen. Des Weiteren fallen 100 € je Arbeitsstunde für die Anpassung von vorformulierten Abfrage (höchstens jedoch 400 €) bzw. für die Erstellung der Auswertungssyntax bei eingereichten Fragestellungen (höchstens jedoch 700 €) an. An einem wissenschaftlichen Gastarbeitsplatz in der Datenaufbereitungsstelle entstehen Kosten in Höhe von 50 € für jeden begonnenen Arbeitstag zuzüglich der Zusatzgebühr von 300 €. Bei Ablehnung eines Antrages aus formalen bzw. inhaltlichen Gründen ergeben sich weiterhin Ablehnungsgebühr in Höhe von 100 € bzw. 150 €. Erfordert eine Fragestellung oder eine vorformulierte Abfrage einen be-

trächtlich hohen Personal- und Sachaufwand, so können die vorgesehenen Gebühren von der Datenaufbereitungsstelle bis auf das Doppelte erhöht werden (DaTraGebV 2014).

Neben den vielen Vorteilen (z. B. kassenübergreifende und somit repräsentative Auswertungen für alle gesetzlich Versicherten, Auswertungsmöglichkeiten für die Leistungserbringerseite, ein zentraler Ansprechpartner, Möglichkeiten zur Berechnung der Behandlungsprävalenz) bringt der neue Datenpool auch Nachteile mit sich. Kritisch äußern sich beispielsweise Krüger-Brand und Mansky et al.. So existieren z. B. Einschränkungen bezüglich der Stammdaten und es fehlen Regionalmerkmale wie der Wohnort des Versicherten (z. B. Postleitzahl bzw. Kreis-/ Gemeindekennziffer); diese sind bereits in den Ausgangsdaten des Morbi-RSA nicht mehr enthalten (Krüger-Brand 2013; Mansky et al. 2012). Weiterhin wird kein genaues Eintritts-/ Austrittsdatum der Versicherten dokumentiert. Ebenso wenig ist die Versicherungsart (Berufstätiger/ Rentner etc.) codiert.

Des Weiteren finden sich im DIMDI-Datenpool keine Informationen zu Prozeduren und Leistungen (Operationen und Eingriffe) sowie zum Todestag – lediglich eine Ja-/ Nein-Aussage, ob der Versicherte im Berichtsjahr verstorben ist. Ebenfalls nicht vorhanden sind Entlassungs- und Verlegungsgrund, Aufnahmedatum ins Krankenhaus sowie die Arztgruppen der behandelnden Ärzte. Die Daten zu Heil- und Hilfsmitteln, zur Pflegeversicherung, hierbei im Speziellen die Angaben zur Pflegestufe und Daten zur Rehabilitation, sind im Datenpool ebenfalls nicht erfasst.

Zusätzlich zu den Informationseinschränkungen ist der hohe Zeitverzug von aktuell vier Jahren ein wesentlicher Nachteil des Datenpools. Aktuell liegen die Daten der Jahre 2009 und 2010 vor, im zweiten Quartal 2014 soll der Datenpool um das Jahr 2011 erweitert werden. Einige Variablen verlieren außerdem an Informationsgehalt, da sie für den Morbi-RSA aggregiert wurden. So sind manche Variablen nur jahres- oder monatsgenau dokumentiert. Ein Beispiel hierfür ist das Entlassungsdatum im Krankenhaus, das nur monatsgenau zur Verfügung gestellt wird. Des Weiteren beinhaltet der Datenpool ausschließlich Versicherte der gesetzlichen Krankenkassen; Auswertungen und Aussagen zu privatversicherten Personen können daher nicht getroffen werden.

Für beide Datenquellen gilt: Da Unterschiede in der Versichertenstruktur zwischen PKV und GKV existieren, kann keine Repräsentativität für die Gesamtbevölkerung Deutschlands sichergestellt werden, sondern lediglich die gesetzlich Versicherten mit einbezogen werden (Hoffmann und Icks 2012).

Die nachfolgende Tabelle fasst die Vor- und Nachteile der beiden Datenquellen zusammen und soll als Entscheidungshilfe dienen.

Tabelle 1: Vor- und Nachteile von GKV-Routinedaten einzelner Krankenkassen sowie des Datenpools

	GKV-Routinedaten einzelner Krankenkassen	Datenpool des DIMDI
Vorteile	• Hohe Flexibilität bezüglich der Datenbankstruktur • Variablen im höchsten Detaillierungsgrad verfügbar • Zeitnahe Verfügbarkeit, aktuelle Daten (max. Zeitverzug ca. 9 Monate) • Abbildung kassenindividueller Versorgungsprogramme • Ergänzung von Primärdaten möglich (z. B. Versichertenbefragungen)	• Ermöglicht flächen- und sektorübergreifende Analysen der gesamten gesetzlich Krankenversicherten • Kassenübergreifende und daher repräsentative Ergebnisse • Zentraler Ansprechpartner • Möglichkeiten zur umfassenden Berechnung der Behandlungsprävalenz • Beinhaltet bereits validierte Daten • Großes Studienkollektiv; daher Untersuchungen von seltenen Erkrankungen möglich
Nachteile	• Begleitende Qualitätssicherung und umfassende Validierung ist erforderlich • Regionale und betriebsbezogene Krankenkassen sind nicht für repräsentative Stichproben geeignet • Einschränkung der Repräsentativität durch Unterschiede in der Versicherten- und Morbiditätsstruktur der Krankenkassen	• Geringere Genauigkeit der Stammdaten (keine genaues Eintritts-/ Austrittsdatum der Versicherten, keine Informationen zur Versicherungsart) • Regionale Kennziffern fehlen • Teilweise geringer Informationsgehalt der Variablen aufgrund zu starker Aggregation (z. B. nur jahres- oder monatsgenau) • Kein Aufnahmedatum im Krankenhaussektor • Fehlende Angabe der Facharztgruppe der behandelnden Ärzte im ambulanten und stationären Bereich • Keine Informationen zu Prozeduren und Leistungen (Operationen und Eingriffen) • Kein Sterbedatum – lediglich eine Ja-/ Nein-Codierung • Hoher zeitlicher Verzug

Quelle: eigene Darstellung

2.3 Datenschutz

Die datenschutzrechtlichen Aspekte werden u. a. auch aufgrund der aufkommenden Verknüpfungsmöglichkeiten von Primär- und Sekundärdaten immer komplexer. Im Folgenden wird daher ein Überblick über die relevanten datenschutzrechtlichen Aspekte unter Berücksichtigung der verschiedenen Zugangswege gegeben. Zudem werden notwendigen Inhalte der Datenschutzkonzepte präsentiert.

In Deutschland existieren umfassende datenschutzrechtliche Voraussetzungen bezüglich des Zugangs und der wissenschaftlichen Nutzung von Sekundärdaten. Da es sich um personenbezogene Daten im Sinne des § 67 SGV X sowie des § 3 Abs. 9 Bundesdatenschutzgesetz (BDSG) handelt, müssen bei der Nutzung von GKV-Routinedaten zu Forschungszwecken zwei Grundprinzipien abgewogen werden (Ihle 2008): einerseits das Recht auf informationelle Selbstbestimmung, das sich aus Art. 2 Abs. 1 GG i. V. m. Art. 1 Abs. 1 GG ableitet, andererseits das Grundrecht auf Forschungsfreiheit (Art. 5 Abs. 3 GG) (GG 2012).

Bei Forschungsvorhaben mit GKV-Routinedaten ist zunächst zu prüfen, ob die jeweiligen zu übermittelnden Daten personenbezogene Sozialdaten im Sinne des § 67 ff. SGB X darstellen und daher dem Regelungsregime des BDSG bzw. SGB X zu unterstellen sind (BDSG 2099; SGB X 2013). „Anonymisieren ist das Verändern von Sozialdaten derart, dass die Einzelangaben über persönliche oder sachliche Verhältnisse nicht mehr oder nur mit einem unverhältnismäßig großen Aufwand an Zeit, Kosten und Arbeitskraft einer bestimmten oder bestimmbaren natürlichen Person zugeordnet werden können“ (§ 67 Abs. 8 SGB X). Anonymisierten Daten fehlt daher der Personalbezug und damit die Eigenschaft der personenbezogenen Daten im Sinne des § 67 Abs. 1 SGB X (Wulffen und Schütze 2014). Sie unterliegen nicht mehr den Bestimmungen der Datenschutzgesetze. Einschränkend ist hierbei anzumerken, dass keine Einigkeit darüber besteht, ob diese Schlussfolgerung auch für eine „unechte Anonymisierung“ gilt, also wenn der Wiederherstellungsaufwand unverhältnismäßig groß ist (Wulffen und Schütze 2014). Daher wird generell empfohlen, datenschutzrechtliche Fragestellungen mit den zuständigen Bundes- oder Landesbehörden zu klären. Zu beachten ist zusätzlich, dass der eigentliche Vorgang der Anonymisierung selbstverständlich unter die Regelungen der Datenschutzgesetze fällt, da dieser sich auf (noch) personenbezogene Daten bezieht. Die Anonymisie-

rung darf daher in der Regel nur durch die Krankenkassen oder im Auftrage der Krankenkasse, z. B. durch eine Vertrauensstelle, durchgeführt werden.

„Pseudonymisieren ist das Ersetzen des Namens und anderer Identifikationsmerkmale durch ein Kennzeichen zu dem Zweck, die Bestimmung des Betroffenen auszuschließen oder wesentlich zu erschweren." (§ 67 Abs. 8a SGB X) Hierbei bleiben personenbezogene Auswertungen möglich (Vauth 2010). Pseudonymisierte Daten unterliegen im Vergleich zu anonymisierten Daten daher eindeutig den Regelungen der Datenschutzgesetze (Scharnetzky et al. 2013). Die Abgrenzung zwischen den beiden Begriffen „Anonymisierung" und „Pseudonymisierung" ist mitunter nicht eindeutig und sollte, wie zuvor bereits empfohlen, mit den zuständigen Behörden für das jeweilige Vorhaben geklärt werden.

Unabhängig von der zuvor genannten Problematik ist die Verwendung von personenbezogenen (Sozial-)Daten für wissenschaftliche Zwecke grundsätzlich durch zwei Ansätze möglich: einerseits durch Rechtsvorschriften, andererseits durch die Zustimmung der Versicherten.

Prinzipiell ist die Nutzung von Sozialdaten nur zulässig, soweit eine Vorschrift des SGB X oder eine andere Rechtsvorschrift des SGB es erlaubt oder die Versicherten (Betroffenen) eingewilligt haben (§ 67b Abs. 1 SGB X). Für die Nutzung von GKV-Routinedaten durch externe Forschungseinrichtungen (z. B. Universitäten) auch ohne vorliegende Einwilligung kommt insbesondere die Vorschrift in § 75 SGB X „Übermittlung von Sozialdaten für die Forschung und Planung" als Rechtsvorschrift in Betracht (Hase 2011). Diese geht mit einer Reihe von verschiedenen Auflagen einher. Demnach ist die Datenübermittlung nur zulässig, wenn:

- sie erforderlich für eine bestimmtes Vorhaben der wissenschaftlichen Forschung im Sozialleistungsbereich ist,
- der Zweck der Forschung nicht auf andere Weise zu erreichen ist,
- die Einholung einer Einwilligung unzumutbar ist,
- die schutzwürdigen Interessen der Versicherten nicht beeinträchtigt werden oder das öffentliche Interesse an der Forschung das Geheimhaltungsinteresse des Versicherten erheblich überwiegt,
- die oberste Bundes- oder Landesbehörde die Übermittlung vorher genehmigt hat.

Sind die genannten Voraussetzungen und weitere Dokumentationspflichten (Art, Zweck und Dauer der Datennutzung) erfüllt, darf die jeweilige Krankenkasse die Daten für den beantragten Forschungszweck übermitteln.

Darüber hinaus gibt es weitere Rechtsvorschriften und Vorschriften z. B. für Krankenkassen, die ihre eigenen Daten zu Forschungszwecken nutzen wollen oder Dritte als Datenverarbeitung im Auftrag anweisen können. Weiterhin ist auch der Zugang zu GKV-Routinedaten im Rahmen der Evaluation von strukturierten Behandlungsprogrammen nach § 137 f SGB V (Disease-Management-Programme) für benannte externe Sachverständige möglich. Auf diese Aspekte wird an dieser Stelle nicht weiter eingegangen, da in dem vorliegenden Diskussionspapier vorrangig Aspekte der externen Datennutzung durch Forschungseinrichtungen zu wissenschaftlichen Zwecken im Vordergrund stehen.

Alternativ muss eine Einwilligung der Versicherten eingeholt werden, um die Nutzung von GKV-Routinedaten zu ermöglichen. Diese ist insbesondere dann einzuholen, wenn keine entsprechende Vorschrift oder Rechtsvorschrift die Nutzung zulässt, zusätzliche Daten erhoben werden sollen (z. B. durch Befragungen oder aus klinischen Dokumentationen) oder wenn eine Einwilligung, z. B. bei Modellvorhaben im Sinne des § 63 SGB V, explizit gefordert ist (Ihle 2008). Einen guten Überblick über den Ablauf und die Voraussetzung eines solchen Verfahrens wird von Scharnetzky et al. gegeben (Scharnetzky et al. 2013). Eine Einwilligung stellt eine vorherige Einverständniserklärung dar. Hierzu ist der Versicherte vor der Einwilligung umfänglich über den Zweck der Nutzung sowie über die Folgen der Verweigerung der Zustimmung aufzuklären. Der Versicherte muss im Anschluss frei entscheiden und schriftlich zustimmen können. Vom Zwang zur Schriftform der Einwilligung kann allerdings im Rahmen der wissenschaftlichen Forschung aufgrund besonderer Umstände eine Befreiung erteilt werden (Wulffen und Schütze 2014) – allerdings nicht davon, die Einwilligung prinzipiell einzuholen. Nähere Informationen zu dieser Thematik finden sich in Harnischmacher et al. (2006) und Majeed et al. (2007). Darüber hinaus können weitere rechtliche Aspekte bei der Einholung von Einwilligungen relevant sein, z. B. die Problematik, wie mit nicht einwilligungsfähigen Personengruppen, wie beispielsweise Kindern, Jugendlichen und dementen Patienten, umgegangen wird (Ihle 2008).

Unabhängig von den zuvor aufgezeigten Datenzugangsmöglichkeiten sollte im Sinne der GPS ein Datenschutzkonzept vor der Nutzung von GKV-Routinedaten entwickelt

werden. Dieses sollte unbedingt schriftlich fixiert und als bindender Vertrag zwischen Datenlieferant und -nutzer formuliert werden. Dabei ist es sinnvoll, den jeweilig zuständigen Datenschutzbeauftragten frühzeitig mit einzubinden, um etwaigen Problemen und Verzögerungen vorzugreifen (Ihle 2008). Maßgeblich sind die geltenden Datenschutzrichtlinien. Die technischen und organisatorischen Maßnahmen werden in § 9 BDSG und dessen Anlage konkretisiert. Hierzu zählen insbesondere Regelungen zur Zutrittskontrolle, Zugangskontrolle, Zugriffskontrolle, Weitergabekontrolle, Eingabekontrolle, Auftragskontrolle und Verfügbarkeitskontrolle (BDSG 2099). Darüber hinaus gibt die GPS konkrete Empfehlungen zur Ausgestaltung eines Datenschutzkonzeptes. Die Empfehlungen beziehen sich dabei auf folgende Aspekte:

- Zweck der Datenbereitstellung,
- Pseudonymisieren und Anonymisieren,
- De-Pseudonymisierung und Re-Identifikation,
- Weitergabe von personenbezogenen Daten an Dritte,
- Personenbezogenes Datenlinkage mit externen Datenquellen,
- Verantwortlicher für den Datenschutz,
- Löschfristen,
- Zusammenarbeit mit Datenschutzbeauftragen.

Weitere Informationen und Hinweise zur Einbeziehung einer Vertrauensstelle zur Pseudonymisierung der Daten gibt Ihle (2008). Die Vertrauensstelle pseudonymisiert die Primärdaten, beispielsweise Fragebögen oder Registerdaten, und verknüpft diese mit den Datensätzen der Krankenkasse, damit die forschende Institution keine Informationen über die natürliche Person erhält und lediglich mit pseudonymisierten Daten arbeiten kann (Scharnetzky et al. 2013). Der Einbezug kann relevant werden, wenn Daten aus unterschiedlichen Datenquellen zusammengeführt werden müssen – wie es beispielsweise bei der gleichzeitigen Nutzung von GKV-Routinedaten und Versichertenbefragungen der Fall ist.

Empfehlungen

- Die geltenden Datenschutzvorschriften zum Schutz der informationellen Selbstbestimmung sind bei der Planung und Durchführung zu beachten
- Ein Datenschutzkonzept im Sinne der GPS ist bereits zu Beginn der Studie zu erstellen
- Datenschutzbeauftragte und zuständige Behörden sollten frühzeitig in Projekte mit eingebunden werden
- Die Notwendigkeit von Versicherteneinwilligungen ist zu prüfen
- Es ist zu prüfen, ob eine Vertrauensstelle mit einbezogen werden muss

2.4 Datenkategorien

In Deutschland fließen jegliche Regelleistungen, d. h. alle Leistungen des Versorgungsgeschehens, die über die GKV abgerechnet werden, bei den gesetzlichen Krankenkassen zusammen. Diese GKV-Routinedaten gehören zu der Kategorie der administrativen Datenbankensysteme und spiegeln die Verwaltungsperspektive wider. Im Vergleich zu arztbasierten Datenbanken sind hierbei sektorübergreifende Kontakte des Versicherten mit dem Gesundheitssystem ersichtlich (Hennessy 2006). Mit der Einführung der Krankenversichertenkarte im Jahre 1995 respektive der Einführung der elektronischen Gesundheitskarte – sukzessive seit 2009 – wurde dieses elektronische Abrechnungsverfahren automatisiert und jedem Versicherten können die in Anspruch genommenen Leistungsdaten individuell zugeschlüsselt werden (Deutscher Bundestag 1995). Die einzelnen Datenkategorien lassen sich unterschiedlichen Sektoren der Versorgung zuordnen, welche die Grundlage für die Gliederung dieses Kapitels bilden. So finden sich in den Datawarehouses der gesetzlichen Krankenkassen unter anderem Informationen zu folgenden Leistungsbereichen:

- Daten der ambulanten Versorgung,
- der stationären Versorgung,
- zu Arzneimitteln,
- zu Heil- und Hilfsmitteln,
- zur Arbeitsunfähigkeit und zum Krankengeld,
- zur Rehabilitation,
- zu Disease-Management-Programmen (DMP),

- Institutsambulanzen,
- sowie Stammdaten der Versicherten.

In den jeweiligen Leistungsbereichen werden unter anderem abrechnungsrelevante Informationen zu den Zeiträumen der Inanspruchnahme, Kosten, Indikationen auf Basis der ICD-10-Codierung sowie Klassifikationsinstrumente bzw. Pauschalen wie Diagnosis Related Groups (DRGs), der Einheitliche Bewertungsmaßstab (EBM) und Operationen- und Prozedurenschlüssel (OPS) erfasst. Aufgrund der vielen Variablen und unterschiedlichen Datawarehouse-Strukturen kann in diesem Kapitel kein Anspruch auf Vollständigkeit erhoben werden. Dennoch werden alle wesentlichen Variablen dargestellt, die sich bisher als wissenschaftlich nutzbar erwiesen haben und von großem Interesse für Routinedatenforscher sind.

2.4.1 Stammdaten

Die Stammdaten bieten grundlegende personenbezogene Informationen zu den Versicherten, wie z. B. das Alter, das Geschlecht, Versichertenzeiten, und werden, anders als die Leistungsdaten, unabhängig von der Inanspruchnahme erfasst. Die Dokumentation dieser personenbezogenen Merkmale zählt zu den grundlegenden Aufgaben der Datenerfassung in der GKV (Grobe und Ihle 2005). Während diese Informationen aus Perspektive der Krankenkassen bei der Erfüllung ihrer Kernaufgaben unterstützen, sind sie aus wissenschaftlicher Sicht für soziodemografische und regionale Auswertungen sowie zur Abbildung der beruflichen Stellung unabdingbar.

Grundsätzlich erfolgt die versichertenbezogene Zuordnung der Leistungsinanspruchnahme bei den Krankenkassen durch eine individuelle Versichertennummer. Für kassenexterne Auswertungen werden bei pseudonymisierten Daten in der Regel Identifikationskennziffern bereitgestellt, die unabhängig von der Versichertennummer personenbezogen eindeutig generiert werden (Grobe und Ihle 2005). Die originale Versichertennummer wird somit für interne Auswertungszecke bzw. Auswertungen Dritter durch die Krankenkasse anonymisiert. Dieses Pseudonym dient als Primärschlüssel, um die Informationen aus den einzelnen Leitungsbereichen miteinander zu verknüpfen.

Grundsätzlich liegen den Krankenkassen Informationen zum Vor- und Nachnamen ihrer Versicherten vor. Aufgrund von Heirat ändert sich relativ häufig der Nachname. Bei ungewöhnlichen Namen kann es darüber hinaus zu Erfassungsfehlern kommen.

Als weitere persönliche Informationen sind bei den Krankenkassen Angaben zum Wohnort, der Postleitzahl sowie der genauen Anschrift inklusive Telefonnummer gespeichert. Diese Informationen werden häufig jedoch nicht im Sinne einer Historie vorgehalten, sondern meist ist nur der aktuelle Wohnort dokumentiert (Grobe und Ihle 2005). Aus Datenschutzgründen können diese persönlichen Daten lediglich mit ausdrücklicher Genehmigung oder mit vorheriger Zustimmung der Betroffenen weitergegeben werden (Grobe und Ihle 2005). Dies erfordert allerdings spezielle Genehmigungen und ein besonderes Datenschutzkonzept (siehe Kapitel 2.3). Die Telefonnummer steht darüber hinaus auch bei kasseninternen Auswertungen häufig nicht zur Verfügung, da sie bei der Krankenkasse nicht hinterlegt werden muss und zudem oftmals nicht in öffentlich zugänglichen Verzeichnissen genannt wird (Grobe und Ihle 2005).

Die Postleitzahl wird bei der Krankenkasse erfasst, jedoch bei kassenexternen Auswertungen nicht im vollen Umfang zur Verfügung gestellt, um Rückschlüsse auf Einzelpersonen zu verhindern. Vielmehr dienen entweder die ersten drei Ziffern des Postleitzahlengebietes oder die sogenannte Kreiskennziffer dazu eine, zumindest grobe, regionale Zuordnung zu ermöglichen (siehe Kapitel 3.2). Sofern aus Datenschutzgründen nur auf die dreistellige Postleitzahl zurückgegriffen werden kann, sollte bei Bedarf eine kasseninterne Zuordnung von Versicherten zu detaillierten räumlichen Gliederungen erfolgen (Grobe und Ihle 2005). Die ersten beiden Ziffern der Kreiskennziffer geben das Bundesland an. Die Ziffern 01-09 wurden den alten Bundesländern von Nord nach Süd zugeordnet, dem Saarland die 10 und Berlin die 11. Anschließend wurden die neuen Bundesländer in alphabetischer Reihenfolge nummeriert (12-16). Die Bundesländer Hamburg (02), Bremen (04) und Berlin (11) sind nicht in Landkreise unterteilt, da sie sogenannte kreisfreie Städte sind (Statistisches Bundesamt 2012). In Deutschland existieren derzeit 402 Landkreise und kreisfreie Städte (Stand 31.12.2011), wobei sich in der Vergangenheit einige Änderungen durch Gebietsreformen gerade in den neuen Ländern ergeben haben ((Bundesinstitut für Bau-, Stadt- und Raumforschung (BBSR) 2011a)). Auf Basis entsprechender Überleitungstabellen lassen sich auch andere regionale Zuordnungen, beispielsweise nach Gemeinden oder Bundesländern, sowie Versorgungsgebiete der Kassenärztlichen Vereinigungen herstellen (Grobe und Ihle 2005). Eine genauere Darstellung der Möglichkeiten erfolgt in 3.2.

Aus datenschutzrechtlichen Gründen steht oftmals lediglich das Geburtsjahr und nicht das tagesgenaue Geburtsdatum des Versicherten für wissenschaftliche Untersuchungen zur Verfügung. Grundsätzlich handelt es sich beim Geburtsdatum um eines der wenigen unveränderlichen Merkmale im engeren Sinne. In einzelnen Subgruppen, wie beispielsweise bei Migranten aus bestimmten Kulturkreisen, häufen sich jedoch bestimmte Geburtstage wie z. B. der Erste eines Monats bzw. eines Jahres (Grobe und Ihle 2005). Gründe hierfür sind die mitunter unzureichenden Meldeverhältnisse in ländlichen Regionen.

Auch beim Geschlecht handelt es sich in der Regel um ein unveränderliches Merkmal. In Einzelfällen kann es hier jedoch zu Veränderungen im Zeitablauf aufgrund einer Geschlechtsumwandlung kommen (Grobe und Ihle 2005). Bei Familienversicherten können gelegentliche Fehlerfassungen nicht ausgeschlossen werden (Grobe und Ihle 2005).

Um Versicherungsintervalle und mögliche Wechsel des Versicherungsstatus zu erfassen, ist das Wissen um die Versichertenzeiten notwendig. Anhand der Versichertenzeiten wird ersichtlich, in welchen Zeiträumen überhaupt mit einer Erfassung der Inanspruchnahme gesundheitsbezogener Leistungen zu rechnen ist. Die Dokumentation der Versichertenzeiten bildet die Grundlage für jegliche populations- bzw. nennerbezogenen Auswertungen (Grobe und Ihle 2005).

Der Beendigungsgrund des Versicherungsvertrages ist z. B. für Überlebenszeitanalysen eine wichtige Variable. So kann diese die Ausprägung „Tod“ annehmen und Aufschluss geben, ob der Versicherte im Berichtszeitraum verstorben ist und damit für Mortalitätsanalysen wichtige Informationen liefern. Die Todesursachen werden jedoch nicht dokumentiert (siehe Abschnitt 6). Außerdem kann bei Familienversicherten in manchen Fällen der Austrittsgrund „Tod“ codiert sein, obwohl ausschließlich das zugehörige Mitglied verstorben ist (Grobe und Ihle 2005).

In den meisten Datawarehouses werden Stammdaten nach Mitgliedern und Familienversicherten getrennt dargestellt. Als Mitglieder werden diejenigen Versicherten bezeichnet, die Versicherungsvertragsnehmer sind und somit auch Versicherungsbeiträge entrichten. In Deutschland gilt für die Krankenversicherung eine Versicherungspflicht (siehe § 5 SGB V). Für alle Arbeitnehmer, deren Jahresarbeitsentgelt unterhalb der Versicherungspflichtgrenze von derzeit 53.550 € (Stand: 01.01.2014)

liegt, sowie für viele weitere Personen gilt die verpflichtende Mitgliedschaft der gesetzlichen Krankenversicherung. Eine freiwillige GKV-Mitgliedschaft kann unter bestimmten Voraussetzungen auch erworben werden.

Familienversicherte sind Personen, die bei einem Mitglied mitversichert sind. In Deutschland können Ehegatten, Lebenspartner, Kinder von Mitgliedern sowie Kinder von familienversicherten Kindern beitragsfrei familienversichert sein (§ 10 SGB V). Dies gilt jedoch nur, wenn der Familienversicherte nicht hauptberuflich selbstständig oder versicherungsfrei nach § 6 SGB V ist oder wenn dieser kein monatliches regelmäßiges Gesamteinkommen bezieht.

Für die Familienversicherten werden häufig weniger Daten bzw. Variablen erfasst als für die Mitglieder (Grobe und Ihle 2005). Nur für die Mitglieder existieren primär Daten zur Beitragshöhe und gegebenenfalls zum ausgeübten Beruf, zur Ausbildung sowie zum Arbeitgeber. Der Versicherungsstatus gibt Auskunft, ob es sich um ein Mitglied der GKV handelt oder um einen Familienversicherten. Eine weitere Variable, der sogenannte Familienschlüssel, gibt für Familienversicherte an, in welchem Verhältnis der Familienversicherte zum originären Mitglied steht und kann unter anderem die Ausprägung Ehegatte oder Kind annehmen. Dies ermöglicht die Zuordnung zum jeweiligen Hauptversicherten, womit gewisse Informationen, beispielsweise zur finanziellen Situation, auch für Familienversicherte indirekt verfügbar sind (Grobe und Ihle 2005). Eine familienbezogene Zusammenfassung von Versicherten scheitert jedoch immer dann, wenn beide Ehepartner berufstätig und bei unterschiedlichen Krankenkassen versichert sind. Daher lassen sich beispielsweise Informationen zum Haushaltseinkommen nicht generell aus den Routinedaten der Krankenkassen ableiten.

Die Personengruppe bzw. Beitragsgruppe gibt an, ob es sich bei dem Versicherten um einen Angestellten/Arbeiter, Selbstständigen, Arbeitslosen, Sozialhilfeempfänger, Studenten/Fachschüler, Rentner/Pensionär, Aussiedler, Flüchtling etc. handelt. Da sich der Krankenkassenbeitrag nach dem sozialversicherungspflichtigen Einkommen und nicht nach dem Gesamteinkommen richtet, ist diese Variable ein wichtiger Indikator für den sozialen Status des Versicherten.

Eine detailliertere Aufgliederung über Ausbildungsstatus und ausgeübten Beruf bietet der sogenannte Tätigkeitsschlüssel. Dieser wird bei den Mitgliedern vom Arbeitgeber

an die Krankenkasse gemeldet und enthält Informationen zum Schulabschluss, zum sozialen Status und zur ausgeübten Tätigkeit des Versicherten. Dieser Tätigkeitsschlüssel wurde mit Wirkung zum 01.12.2011 aktualisiert, da sich in den vergangenen Jahren sowohl in der Berufs- und Beschäftigungslandschaft als auch der Ausbildungsstruktur Veränderung ergeben haben (Damm et al. 2012).

Die Angaben zu Beitragsgruppen und beruflichen Tätigkeiten können bei einer Betrachtung längerer Zeiträume sehr komplex und unübersichtlich sein, da sich die Merkmale in Einzelfällen ausgesprochen häufig verändern können (Grobe und Ihle 2005). Bei Daten zu Berufsintervallen wurden beispielsweise über einen Zeitraum von 15 Jahren mehr als 1.000 Statuswechsel berichtet (Grobe und Ihle 2005). Zudem können zu einem Zeitpunkt personenbezogen auch mehre Versicherungszustände, beispielsweise bei Mehrfachbeschäftigungen, relevant sein (Grobe und Ihle 2005).

In der nachstehenden Tabelle 2 sind ausgewählte, für wissenschaftliche Zwecke wichtige Variablen aufgeführt.

Tabelle 2: Variablenbeschreibung der Stammdaten

Variable	Mögliche Ausprägungen	Erläuterung
Identifikationsnummer	Ziffernfolge oder Buchstaben-/ Zahlenkombination	Anonymisierte Versicherten-ID; dient als Primärschlüssel und zur personenbezogenen Zuordnung der Leistungen
Name und Vorname	Buchstabenkombination	Name und Vorname des Versicherten
Staatsangehörigkeit	Zumeist dreistelliger Zahlenschlüssel	Angabe über die Nationalität
Wohnort und Anschrift	Straße, Straßennummer, Ort, Postleitzahl etc.	Wohnort und Anschrift des Versicherten
Familienstand	Ledig, verheiratet, verwitwet	Angabe über den Familienstatus
Geburtstag	TT.MM.JJJJ	Geburtstag des Mitglieds bzw. Familienversicherten
Geschlecht	Männlich/weiblich	Geschlecht des Mitglieds bzw. Familienversicherten
Kreiskennziffer	Fünfstellige Ziffernfolge	Kreis, in dem der Versicherte wohnt; die ersten beiden Ziffern kennzeichnen das Bundesland
Beginn und Ende des Versicherungsstatus	TT.MM.JJJJ	Datum der Versicherungsvertragslaufzeiten bzw. -status; bei laufenden Verträgen kann das Enddatum auf einen artifiziellen Wert, z. B. den 01.01.9999, gesetzt sein
Beendigungsgrund eines Versicherungsverhältnisses	Tod, Krankenkassenwechsel	Grund für die Beendigung des Versicherungsverhältnisses bzw. das Ausscheiden eines Versicherten aus der gesetzlichen Krankenversicherung bzw. der jeweiligen Krankenkasse
Familienversicherungsschlüssel	Ehegatte, Kind, Lebenspartner, Pflegekind, Kind des Kindes	Stellung/Beziehung zum Mitglied
Tätigkeitsschlüssel	Neunstellige Ziffernfolge	Der Tätigkeitsschlüssel gibt Auskunft über den letzten Schulabschluss, höchsten Ausbildungsabschluss und die aus-

Variable	Mögliche Ausprägungen	Erläuterung
		geübte Tätigkeit
Versicherungsstatus	Mitglied, familienversichert	Codierung, ob es sich um ein Mitglied oder um einen Familienversicherten handelt
Personengruppe/ Beitragsgruppe	Angestellte/Arbeiter, Selbstständiger, Arbeitsloser, Sozialhilfeempfänger, Student/ Fachschüler, Rentner/Pensionär, Aussiedler, Flüchtling etc., Sonstige	Angabe über die Personen- bzw. Beitragsgruppe des Versicherten
Arbeitgebernummer	Numerischer Ausdruck	Eindeutige Kennung des Arbeitgebers
Branche	Land- und Forstwirtschaft, Baugewerbe, Dienstleistungen, Energieversorgung etc.	Branchenzuordnung des Arbeitgebers
Befreiung von Zuzahlungen nach § 62 SGB V	ja/nein	Versicherte müssen bis zu einer bestimmten Höhe die Zuzahlung selbst tragen; in § 62 SGB V wird diese Belastungsgrenze genauer erläutert

Quelle: eigene Darstellung aus Grobe und Ihle (2005); Vauth (2010); Zeidler und Braun (2012); Reinhold et al. (2011a); GKV-Datenaustausch (a)

2.4.2 Ambulante Versorgung

Häufig stellt der ambulant-ärztliche Sektor den ersten Kontakt des Versicherten mit dem Gesundheitssystem dar. Grundsätzlich rechnet nicht jeder einzelne Arzt bzw. jede einzelne Praxisgemeinschaft mit den jeweiligen gesetzlichen Krankenkassen ab, sondern dieser Abrechnungsprozess geschieht über die zuständige Kassenärztliche Vereinigung (KV). Die Abrechnung und der damit verbundene Datentransfer erfolgt heute in der Regel per EDV-gestütztem Datentransfer (Kerek-Bodden et al. 2005). Die Abrechnungsunterlagen der kassenärztlichen Leistungserbringer werden am Quartalsende (viermal jährlich) der zuständigen KV vorgelegt. Die KV prüft die übermittelten Daten und zahlt das Honorar aus (Kerek-Bodden et al. 2005). Dieser Austausch der ambulanten Leistungs- und Diagnosedaten ist seit dem Inkrafttreten des GKV-Modernisierungsgesetzes im Jahr 2004 gesetzlich vorgeschrieben und in § 295 SGB V geregelt (Vauth 2010); GKV-Modernisierungsgesetz (GMG) 2003, Artikel 1 Nr. 167d zur Änderung von § 295 Abs. 2 Satz 1 SGB V).

In Deutschland existieren 17 KVen; jeweils eine KV je Bundesland, lediglich Nordrhein-Westfalen (NRW) gliedert sich in zwei KVen (KV Nordrhein und KV Westfalen-Lippe). Die gesetzliche Grundlage für das Bestehen von zwei Kassenärztlichen Vereinigungen in NRW findet sich im § 77 des SGB V. Im Abs. 1 Satz 2 stand bis zum 01.01.2012: „Soweit in einem Land mehrere Kassenärztliche Vereinigungen mit weniger als 10.000 Mitgliedern bestehen, werden diese zusammengelegt." Im Umkehrschluss bedeutet dies, dass die KV Westfalen-Lippe mit rund 13.000 Mitgliedern sowie die KV Nordrhein mit rund 18.500 Mitgliedern weiterhin eigenständig sein können und die Kassenärzte eine „eigene" Verwaltung unterhalten dürfen. Zum 01.01.2012 wurde der Satz 2 aus dem SGB V gestrichen, da dort die Mitglieder-Verhältnisse eindeutig und klar geregelt sind; insofern war der Satz 2 dann überflüssig.

Die Vertragsärzte rechnen auf Basis des Einheitlichen Bewertungsmaßstabes (EBM) ihre erbrachten Leistungen ab. Seit Einführung eines neuen EBM, Anfang des Jahres 2009, erfolgt die Bewertung dieser erbrachten Leistungen in Euro-Beträgen. Vor dem Januar 2009 erfolgte diese ausschließlich über Punkte (Institut des Bewertungsausschusses). Für eine monetäre Bewertung wird die Punktzahl, mit der die jeweilige Leistung abgerechnet wurde, mit einem Orientierungspunktwert multipliziert (Prenzler et al. 2010). Für 2009 hatte der Erweiterte Bewertungsausschuss diesen Orientierungspunktwert mit 3,5001 Cent und für 2010, 2011 und 2012 mit 3,5048 Cent festgelegt, 2013 stieg dieser auf 3,5363 Cent. Zum 4. Quartal 2013 wurden der Orientierungswert und der kalkulatorische Punktwert auf 10 Cent angehoben. Im Gegenzug dazu sanken die Punktzahlen je Leistung proportional, sodass die Erhöhung des Punktwerts kostenneutral blieb. Mit diesem Punktwert, der für alle Kassenarten und Fachgruppen bundesweit einheitlich ist, werden fast alle ambulant-ärztlichen Leistungen vergütet. Der Orientierungspunktwert für das Jahr 2014 beträgt 10,1300 Cent (KV Berlin). Bei der Ermittlung der Kosten im ambulanten Sektor sind weiterhin die arztgruppenspezifischen Regelleistungsvolumina zu berücksichtigen, welche die maximale Höhe der Honorare der Ärzte einschränken.

Zu dem Bereich der Kosten zählen Sachkosten wie Verbrauchsmaterialen (z. B. Briefmarken für Arztbriefe) und extrabudgetäre Leistungen. Die ambulanten Leistungen werden, wie bereits erwähnt, anhand der EBM-Ziffern und der jeweiligen Punkte abgerechnet. Die Punktesummen sind jeweils in einer separaten Variable aufgeführt

und sind das Produkt der Punkte multipliziert mit dem Faktor, d. h. der Anzahl an abgerechneten Leistungen.

Ambulante Abrechnungs- und Diagnosedaten bieten detaillierte Informationen zur Art und Anzahl der in Anspruch genommenen Leistungen (EBM-Gebührenziffern), dem Datum der Leistungsinanspruchnahme, pseudonymisierte Informationen zum behandelnden Arzt (Facharztgruppe), ICD-Diagnosen, Quartal der Diagnosestellung und eine Spezifikation der Diagnosesicherheit sowie Überweisungsfälle (§ 295 SGB V). Die Zusatzkennzeichen für die Diagnosesicherheit sind laut den ambulanten Codierrichtlinien (AKR) zwingend erforderlich. Die Ausprägungen können laut Kassenärztliche Vereinigung-Datentransfer (KVDT) wie folgt codiert werden: G: gesicherte Diagnose, V: Verdacht auf, A: ausgeschlossene Diagnose, Z: symptomloser Zustand (KBV 2011b). Eine Behandlungsdiagnose erhält das Zusatzkennzeichen „G“, wenn der Arzt sie nach den gültigen medizinisch-wissenschaftlichen Grundsätzen sichern konnte. So lange eine Behandlungsdiagnose weder gesichert noch ausgeschlossen werden kann, erhält der ICD-Code für die Behandlungsdiagnose das Zusatzkennzeichen „V“. Das „A“ steht für ausgeschlossene Diagnose und ist definiert als „Diagnose, für die es primär einen Verdacht gab, die aber ausgeschlossenen wird“. Eine Behandlungsdiagnose erhält das Zusatzkennzeichen „Z“, wenn die betreffende Diagnose nicht mehr besteht und auch keine krankheitsspezifische Diagnostik und/oder Therapie mehr erfolgt. Der Zustand nach dieser Diagnose hat eine Leistungserbringung verursacht, die zu einer Codierung führt (z. B. die Gabe von ASS nach einer abgeschlossenen Schlaganfallbehandlung). Die Zusatzkennzeichnung soll den medizinischen Entscheidungsprozess bei einer Diagnosefindung abbilden. Eine gesicherte Diagnose zu verschlüsseln, ist insbesondere bei einem Erstkontakt des Patienten oft nicht möglich, da die Abklärung und Diagnostik von Beschwerden längere Zeit in Anspruch nimmt (KBV 2011a). Auch nach austherapierter Krankheit können noch Leistungen erbracht werden. Ein Beispiel hierfür ist eine Dauermedikation bei Patienten nach einem Herzinfarkt, um einem erneuten Infarkt abzuwenden bzw. diesem vorzubeugen. Um diesen Leistungen eine Diagnose zuordnen zu können, werden diese als „Zustand nach“ einer Erkrankung oder Operation codiert. Neben dem „Zustand nach“ und der gesicherten Diagnose existiert die Möglichkeit, eine Verdachtsdiagnose zu codieren. Die Abgrenzung zwischen der Wahl einer gesicherten Diagnose und einer Verdachtsdiagnose ist fließend. Letztendlich obliegt es dem Arzt, diese Einteilung vorzunehmen. Auch bei noch ausstehenden Befunden oder beim

Warten auf eine spezifische Therapie sind Verdachtsdiagnosen anzugeben (Deutsches Ärzteblatt 2011). Zusätzlich zu der Diagnosesicherheit kann als Ergänzung bei paarigen Organen und Körperteilen eine Angabe der Seitenlokalisation (links (L), rechts (R), beidseitig (B)) sinnvoll sein (KBV 2011a). Diese Angabe wird als eigenständige Variable im ambulanten Datensatz gesondert übermittelt.

Die Diagnosen werden nach der gesetzlich vorgeschriebenen Klassifikation für Krankheiten und verwandte Gesundheitsprobleme nach ICD-10-GM verschlüsselt. Diese gliedert sich in eine dreistellige allgemeine Systematik, die vierte Stelle stellt eine ausführlichere Systematik dar und gelegentlich wird auch die fünfstellige Codierung als Verfeinerung verwendet (DIMDI). Diese Systematik wird gleichwertig sowohl im ambulanten als auch im stationären Versorgungsbereich angewendet.

Die Abrechnung über die KVen ist dem Umstand geschuldet, dass eine spezifische Diagnose lediglich einmal – und dies auch nur quartalsweise – an die Krankenkasse gemeldet wird. Die Leistungen, die in der Regel mittels EBM abgerechnet werden können, werden hingegen tagesgenau erfasst. Daher erscheinen in den ambulanten Daten sowohl ein tagesgenaues Datum für die erbrachten Leistungen als auch eine Quartalsangabe, die wiederum der ICD-10-Diagnose zugeordnet ist. Die tagesgenaue Abrechnung der EBM kann genutzt werden, um approximativ die Anzahl der Arztkontakte zu ermitteln. Aufgrund der Grundpauschale kann es jedoch hierbei zu einer Unterschätzung der tatsächlichen Arztkontakte eines Versicherten kommen, da nicht bei jedem Arzt-Patienten-Kontakt eine Leistung abgerechnet wird bzw. werden kann.

Die KV-Behandlernummer, auch lebenslange Arztnummer (LANR) genannt, ist eine neunstellige Kennzeichnung für jeden Vertragsarzt. Die ersten sechs der insgesamt neun Stellen gelten „lebenslang“ für die gesamte vertragsärztliche Tätigkeit. Sie sind KV-übergreifend, unabhängig vom Status, der Zugehörigkeit zu Berufsausübungsgemeinschaften und dem Tätigkeitsort (KBV 2008). An siebter Stelle ist eine Prüfziffer eingebaut, die aus den vorangehenden sechs Stellen berechnet wird. Die achte und neunte Ziffer gibt die Facharztrichtung an, womit es möglich ist, Subgruppenanalysen bezüglich unterschiedlicher Facharztgruppen durchzuführen. Gelegentlich wird die Facharztgruppe auch als separate Variable übermittelt. Sowohl die Arztnummer als auch die Systematik des Arztgruppenschlüssels legt die Kassenärztliche Bundesvereinigung fest (siehe Anlage 2 der Richtlinie der Kassenärztlichen Bundesvereini-

gung nach § 75 Abs. 7 SGB V zur Vergabe der Arzt- und Betriebsstättennummern). Diese lebenslange Arztnummer (LANR) wurde erst im zweiten Quartal 2009 eingeführt, da die vorangegangene Arztnummernsystematik die Untergliederung der Facharztrichtungen nicht differenziert genug abbilden konnte. Dies ist bei Analysen, die diesen Zeitraum inkludieren, zu beachten, da es andernfalls zu nicht vergleichbaren Facharztgruppenzuteilungen kommen könnte.

Tabelle 3 gibt einen Überblick über die relevanten Variablen des ambulanten Leistungssektors.

Tabelle 3: Variablenbeschreibung in der ambulanten Versorgung

Variable	Mögliche Ausprägungen	Erläuterung
Identifikationsnummer	Ziffernfolge oder Buchstaben- / Zahlenkombination	Anonymisierte Versicherten-ID; dient als Primärschlüssel und zur personenbezogenen Zuordnung der Leistungen
Beginn der Leistungserbringung	TT.MM.JJJJ	Behandlungsbeginn einer Arztleistung
Ende der Leistungserbringung	TT.MM.JJJJ	Behandlungsende einer Arztleistung
Quartal (Jahr)	Zahlenkombination: teilweise sind Quartal und Jahr durch ein „Q“ voneinander abgegrenzt	Quartal und Jahr der Diagnosestellung
Diagnose(n)	Drei- bis fünfstellige alphanumerische Systematik (z. B. I5014)	ICD-10-Diagnose
Diagnosesicherheit	V, G, A oder Z	ICD-Diagnosesicherheit z. B. V = Verdacht, G = Gesichert, A = Ausschluss oder Z = „Zustand nach“-Diagnose
Arztnummer	Numerischer Ausdruck	Anonymisierte Arztnr. (Vorgängerversion der Arztnummernsystematik gültig bis 2009)
KV-Behandlernummer	Numerischer Ausdruck	neunstellige KV-Behandlernr. Gültig ab dem 2. Quartal 2009
Facharztgruppe	Zweitstellige Nummer	Facharztgruppe

Variable	Mögliche Ausprägungen	Erläuterung
Gebührenordnungsziffer(n)/ (EBM) Leistungsziffern	Numerischer Ausdruck	Gebührenordnungsziffer nach EBM (z. B. 05230)
Euro-EBM	EBM in €	Orientierungswert für Honorarhöhe in Euro
Punktzahl (gemäß EBM)	Numerische Ziffer	Abgerechnete EBM-Punkte (z. B. 345)
Anzahl Leistungen (je Einzelzeile)	Numerischer Ausdruck	Anzahl (z. B. 1)
Kosten	Euro-Betrag	Extrabudgetäre Leistungen/ Sachkosten aus Perspektive der Krankenkasse in €
OPS-Ziffern	Numerischer Ausdruck	Klassifikation bei ambulanten Operationen
Zuzahlungen	Euro-Betrag	z. B. Praxisgebühr
Art der Inanspruchnahme	Original-, Sekundär-, Not- oder Vertretungsfall	Klassifikation eines Behandlungsfalls
Interne Fallnummer	Numerischer Ausdruck	Interne Fallnummer zur Verknüpfung der einzelnen Tabellen (Leistungs- und Diagnosetabellen werden häufig separat verwaltet)

Quelle: eigene Darstellung aus § 295 SGB V; Vauth (2010); Zeidler und Braun (2012); GKV-Datenaustausch (c)

2.4.3 Stationäre Versorgung

Seit dem Jahr 2004 basiert die Vergütung der voll- und teilstationären Leistungen auf einem Fallpauschalensystem, den sogenannten Diagnosis Related Groups (DRGs) (§ 17b KHG) (KHG 2013). Die Zuordnung zu einer diagnosebezogenen Fallgruppe erfolgt dabei vorrangig durch die Hauptdiagnose des Behandlungsfalles. Die Hauptdiagnose ist die ausschlaggebende ICD-10-codierte Indikation für den stationären Aufenthalt des Patienten. Zusätzlich fließen individuelle Patientendaten, wie z. B. Alter, Geschlecht, das Gewicht bei Neugeborenen sowie Fallcharakteristika wie Komplikationen und anhand OPS codierte Operationen und Prozeduren in die Zuordnung zu einer DRG ein (DIMDI 2014b; Grobe 2005).

Laut deutschen Kodierrichtlinien wird die Hauptdiagnose definiert als die Diagnose, „die hauptsächlich für die Veranlassung des stationären Krankenhausaufenthaltes der Patientin/des Patienten verantwortlich ist“ und ist entsprechend ICD-10-GM zu

codieren (gbe-bund 2012). „Als relevante Nebendiagnose (Komorbidität und Komplikation) gelten Krankheiten oder Beschwerden, die entweder gleichzeitig mit der Hauptdiagnose bestehen oder sich während des Krankenhausaufenthalts entwickeln“ (gbe-bund 2012). Diagnostische bzw. therapeutische Maßnahmen (Verfahren und/oder Prozedur) oder ein erhöhter Pflege- und/oder Überwachungsaufwand sind die Voraussetzungen für eine mögliche Codierung dieser Nebendiagnosen. Die Nebendiagnosen stehen gleichwertig nebeneinander, sodass hier keine Hierarchie erzeugt werden kann.

Im stationären Bereich wird neben Haupt- und Nebendiagnosen auch zwischen Einweisungs-, Aufnahme- und Entlassungsdiagnosen unterschieden. Die Einweisungsdiagnose zählt zu den Kannangaben nach § 301 und wird beispielsweise vom einweisenden Arzt in verschlüsselter Form (ICD-10) mitgeteilt. Kommt ein Patient ohne Einweisung in ein Krankenhaus, wird die Einweisungsdiagnose häufig nicht codiert. Nach § 39 SGB V entscheidet dann der Krankenhausarzt bei der Aufnahme über die Notwendigkeit einer stationären Behandlung. Wenn diese Notwendigkeit besteht, wird in einer ersten Einschätzung die Aufnahmediagnose dokumentiert. Die Entlassungsdiagnose entspricht dann der Hauptdiagnose nach dem Krankenhausaufenthalt; das heißt auch nach ausführlicher Untersuchung durch das Krankenhaus. Diese kann aufgrund der zahlreichen diagnostischen Maßnahmen während des Krankenhausaufenthalts erheblich von der Einweisungsdiagnose bzw. der Aufnahmediagnose abweichen.

Die von den Krankenhäusern an die Krankenkassen übermittelten stationären Abrechnungsdaten umfassen unter anderem Informationen zum Aufnahme- und Entlassungsdatum (exaktes Datum verfügbar), zum Entlassungsgrund, zu allen Haupt- und Nebendiagnosen gemäß ICD, zu Operationen und Prozeduren gemäß OPS, zur Art der stationären Behandlung sowie zur abrechnungsrelevanten DRG (siehe Tabelle 4). Die Übermittlung der Daten ist in § 301 SGB V geregelt.

Grundsätzlich sind durch die pauschalierte Abrechnung über die DRGs keine Daten zum Arzneimittelverbrauch während des Krankenhausaufenthalts verfügbar (siehe Kapitel 6). Dennoch können ausgewählte Arzneimittel, beispielsweise recht hochpreisige Arzneimittel wie TNF-α-Hemmer, über die OPS codiert werden.

Anders als beispielsweise bei der Rehabilitation existiert keine eigene Variable für die Dauer des jeweiligen Krankenhausaufenthalts (siehe Abschnitt 2.4.7). Die Behandlungsdauer kann jedoch manuell mittels Aufnahme- und Entlassungsdatum wie folgt berechnet werden (Grobe 2005):

Aufenthalts-/Behandlungsdauer = Entlassungsdatum - Aufnahmedatum + 1

Tabelle 4 gibt einen Überblick über die relevanten Variablen des stationären Leistungssektors.

Tabelle 4: Variablenbeschreibung in der stationären Versorgung

Variable	Mögliche Ausprägungen	Erläuterung
Identifikationsnummer	Ziffernfolge oder Buchstaben-/ Zahlenkombination	Anonymisierte Versicherten-ID; dient als Primärschlüssel und zur personenbezogenen Zuordnung der Leistungen
Tag der Aufnahme	TT.MM.JJJJ	Aufnahmedatum
Tag der Entlassung	TT.MM.JJJJ	Entlassungsdatum
Einweisungsdiagnose	Drei- bis fünfstellige alphanumerische Systematik	ICD-10-Diagnosen
Aufnahmediagnose	Drei- bis fünfstellige alphanumerische Systematik	ICD-10-Diagnosen
Hauptdiagnose (bei Entlassung)	Drei- bis fünfstellige alphanumerische Systematik	ICD-10-Diagnosen
Hauptdiagnose eines Krankenhausfalls	Drei- bis fünfstellige alphanumerische Systematik	ICD-10-Diagnosen
Nebendiagnosen	Drei- bis fünfstellige alphanumerische Systematik	Weitere abrechnungsrelevante Diagnosen und Komplikationen; Sternchendiagnosen werden häufig in den Nebendiagnosen codiert
OPS	Fünfstellige Nummer	OPS-Schlüssel: Im KHS-Fall wird der Hauptoperationsschlüssel gespeichert
Operationsdatum	TT.MM.JJJJ	Tag der Operation
DRG	Drei- bis fünfstellige alphanumerische Systematik	Abgerechnete DRG
Kosten DRG-Betrag	Euro-Betrag	Kosten aus Perspektive der Krankenversicherung in €

Variable	Mögliche Ausprägungen	Erläuterung
ICD Typ	Aufnahme-, Entlassungs-Haupt- oder Nebendiagnose	Herkunft der Diagnose
Aufnahmegrund	Z. B. Notfall	Grund der Aufnahme
Entlassungsgrund	Reguläre Entlassung, Entlassung auf Patientenwunsch, Verlegung, Tod des Patienten	Grund der Entlassung
Art der stationären Behandlung	Vollstationär, teilstationär, ambulante OP im KH, Sonstige	Art der stationären Behandlung
Beatmungsstunden	Numerischer Ausdruck in Stunden	Dauer einer künstlichen Beatmung in Stunden
Aufnehmende Fachabteilung	Innere Medizin etc.	Codierung der Fachabteilung, welche den Patienten aufnimmt
Entlassende Fachabteilung	Kardiologie etc.	Codierung der Fachabteilung, welche den Patienten entlässt
IK-Nummer des Krankenhauses	Zehnstellige Nummer	Anonymisierte Identifikationsnummer der Institution (ergänzt durch die Art der Institution und regionale Zuordnung)

Quelle: eigene Darstellung aus § 301 SGB V; Vauth (2010); Zeidler und Braun (2012); Grobe (2005); Müller-Bergfort und Fritze (2007); GKV-Datenaustausch (d)

2.4.4 Arzneimitteldaten

Die Apotheken rechnen nicht unmittelbar mit den Krankenkassen ab, sondern können für die elektronische Übermittlung der Daten Rechenzentren in Anspruch nehmen. In § 300 SGB V werden die Apotheken verpflichtet, die Verordnungsblätter oder die elektronischen Verordnungsdatensätze an die Krankenkassen weiterzuleiten. Diese enthalten Informationen zur Facharztgruppe des verordnenden Arztes, zum Ausstellungsdatum der Verordnung und zum Abgabedatum des Präparates sowie zur siebenstelligen Pharmazentralnummer (PZN). Die PZN in Kombination mit dem Anatomisch-Therapeutisch-Chemischen (ATC) Klassifikationssystem ermöglicht die Ergänzung weiterer relevanter Informationen wie z. B. Packungsgröße, Darreichungsform, definierte Tagesdosen (DDD), Hersteller sowie Handelsnamen (§ 300 SGB V, Vauth 2010). Des Weiteren existiert in den Arzneimitteldaten eine Kenn-

zeichnung für Hilfsmittel. Hilfsmittel können auch von der Apotheke abgegeben werden und treten dadurch gelegentlich in den Arzneimitteldaten auf. Die Variable Hilfsmittelkennzeichen kann diese identifizieren und zur einer besseren Trennung von Arznei- und Hilfsmitteln beitragen. Der Weg eines Arzneimittels von der Verordnung durch den Arzt auf dem Rezeptblatt bis zur Vergütung verläuft bundeseinheitlich und bei allen Krankenkassen gleich (Nink et al. 2005).

In den Arzneimitteldaten der Krankenkassen sind lediglich Informationen zu verschreibungspflichtigen Arzneimitteln enthalten. Nichtapothekenpflichtige sowie ohne Rezept, d. h. privat in der Apotheke erworbene Arzneimittel, die sogenannten OTC-Arzneimittel (over the counter), werden nicht erfasst. Weiterhin sind in den GKV-Routinedaten die Verordnungen zulasten der privaten Krankenversicherung sowie in Krankenhäusern abgegebene Arzneimittel, sofern diese nicht über eine OPS codiert werden können, nicht verzeichnet.

Eine sinnvolle Ergänzung der Arzneimittelabrechnungsdaten bilden der GKV-Arzneimittelindex des WIdO sowie die LAUER-Taxe (WIdO; LAUER-Taxe). Diese beiden Datenbanken können herangezogen werden, wenn einzelne Informationen nicht vollständig bzw. fehlerhaft übermittelt wurden. So können beispielsweise anhand der PZN der dazugehörige ATC-Code, die Packungsgröße und Dosierung sowie die Darreichungsform über die LAUER-Taxe ermittelt und fehlende Informationen in der Routinedatenbank ergänzt werden. Über die LAUER-Taxe können auch weitere Hintergrundinformationen generiert werden; so ist beispielsweise der Beipackzettel mit allen wichtigen Patienteninformationen dort individuell für jede PZN hinterlegt.

In den Arzneimitteldaten der gesetzlichen Krankenkassen sind auch Informationen zu den entstandenen Kosten zu finden. Bei Kostenanalysen ist jedoch darauf zu achten, welche Art von Kosten die Krankenkasse übermittelt hat. Sogenannte Bruttokosten des Arzneimittels spiegeln den Apothekenabgabepreis wider. Diese berücksichtigen allerdings keine Rabatte oder Zuzahlungen der Versicherten. Für eine Kostenermittlung aus der Perspektive der Krankenkassen müssten diese Komponenten allerdings noch abgezogen werden, um sogenannte Nettokosten zu kalkulieren. Aus wettbewerbsstrategischen Gründen erhalten Forscher meist lediglich die Bruttokosten. Wie mit Zuzahlungen umgegangen werden kann, wird in Kapitel 5.4 näher erläutert. Bei Rabatten kann zwischen einheitlichen Zwangsrabatten der Arzneimittelhersteller und kassenindividuellen Rabattverträgen differenziert werden. Kassenindivi-

duelle Arzneimittel-Rabattverträge sind vertragliche Vereinbarungen zwischen einzelnen Krankenkassen bzw. deren Verbänden und einzelnen Arzneimittelherstellern über die exklusive Abgabe einzelner Arzneimittel des Herstellers an die Versicherten.

Eine Neuerung ergibt sich seit dem 01.01.2013. So sind alle PZN ab diesem Zeitpunkt (Abgabedatum ab dem 01.01.2013) achtstellig zu formatieren und nicht wie vorher lediglich siebenstellig (GKV-Datenaustausch (b); GKV-Datenaustausch (f)). Bei Auswertungen vor und nach diesem Datum ist diese Änderung zu beachten. Des Weiteren können veraltete PZN neu vergeben werden. Bei langen Untersuchungszeiträumen könnte dies eine Herausforderung bei der Analyse darstellen und PZN-Doppelungen zur Folge haben.

Detaillierte Analysen von Verordnungsprofilen sind auf Basis der Arzneimitteldaten sowie durch die Verknüpfung von Verordnungs- und Leistungsdaten aus anderen Leistungsbereichen möglich. So können das Einnahmeverhalten und indikationsspezifische Medikationsmuster aus dem Alltagsgeschehen untersucht werden.

Einen Überblick über die relevanten Variablen im Arzneimittelbreich gibt nachfolgend Tabelle 5.

Tabelle 5: Variablenbeschreibung der Arzneimitteldaten

Variable	Mögliche Ausprägungen	Erläuterung
Identifikationsnummer	Ziffernfolge oder Buchstaben-/ Zahlenkombination	Anonymisierte Versicherten-ID; dient als Primärschlüssel und zur personenbezogenen Zuordnung der Leistungen
ATC-Codierung	Alphanumerische Systematik	Code nach ATC-Klassifikation (z. B. N05AA01)
Pharmazentralnummer (PZN)	Ehemals siebenstellige, seit dem 01.01.2013 achtstellige numerische Systematik	Offizielle Pharmazentralnummer
Datum der Ausstellung	TT.MM.JJJJ	Datum der Verordnungsblatt-Ausstellung
Datum der Abgabe	TT.MM.JJJJ	Datum der Verordnungsblatt-Abgabe bei der Apotheke
Anzahl	Numerischer Ausdruck	Verordnete Menge je Einzelzeile (in der Regel entspricht dies der Anzahl der Packungen)
Bruttokosten	Euro-Betrag	Kosten in € (Apothekenabgabepreis)
Nettokosten	Euro-Betrag	Kosten aus Perspektive der Krankenversicherung in €
DDD	Tage	Defined Daily Doses, die angenommene mittlere Tagesdosis bei Arzneimitteln
Hilfsmittelkennzeichen	ja/nein	Angabe, ob es sich um ein Hilfsmittel handelt
Zuzahlungen	Euro-Betrag	Höhe der Zuzahlungen des Patienten

Quelle: eigene Darstellung aus § 300 SGB V; Grobe und Ihle (2005); Vauth (2010); Zeidler und Braun (2012); GKV-Datenaustausch (b)

2.4.5 Heil- und Hilfsmitteldaten

Heil- und Hilfsmittel fallen unter die Abrechnung sonstiger Leistungserbringer, die in § 302 SGB V geregelt sind. In § 302 SGB V Abs. 1 wird der Leistungserbringer verpflichtet, den Krankenkassen die erbrachten Leistungen per elektronischer Datenübertragung oder auf einem Datenträger zu melden. Hierbei müssen folgende Informationen grundsätzlich geliefert werden: Art, Menge und Preis der Leistung sowie Datum der Verordnung des Arztes, Tag der Leistungserbringung bzw. -bereitstellung

sowie die Arztnummer. Diese Informationen finden sich einheitlich auf dem Verordnungsblatt, das der Versicherte von dem verordnenden Arzt erhält (Schröder et al. 2005). Neben den gelieferten Variablen existiert häufig – ähnlich wie bei den ambulanten Daten – eine Verknüpfungsvariable, die es ermöglicht mehrere Tabellen bzw. Datenblätter miteinander zu verknüpften. Zum Tragen kommt diese Variable, wenn beispielsweise Leistungs- und Verschreibungsdaten separat verwaltet werden.

Häufig werden Heil- und Hilfsmittel in einem gemeinsamen Datawarehouse geführt und auch gebündelt an den jeweiligen Forscher übermittelt. Dennoch bestehen erhebliche Unterschiede in der Form der Leistungserbringung und im Abrechnungsprozess zwischen den beiden Leistungsarten. So werden Heilmittel durch Therapeuten und meist in mehreren Sitzungen abgegeben. Hilfsmittel hingegen werden von der Apotheke oder vom Sanitätsfachhandel ausgehändigt. Diese Heterogenität ist auch aus methodischer Sicht zu beachten. So sind beispielsweise die Heilmittelpositionsnummern fünfstellig und die Hilfsmittelpositionsnummern zehnstellig, was bei gemeinsamen Auswertungen Beachtung finden muss. Des Weiteren ist es möglich, dass Hilfsmittel, die nicht im GKV-Hilfsmittelverzeichnis aufgeführt sind, von den gesetzlichen Krankenkassen übernommen wurden, hierfür sind dann keine bzw. ausschließlich Sonderhilfsmittelpositionsnummern verfügbar. Das GKV-Hilfsmittelverzeichnis dient lediglich als eine Orientierungs- und Auslegungshilfe. Anzumerken ist weiterhin, dass Hilfsmittel nicht budgetrelevant für den behandelnden Arzt sind und es keine Richtgrößen, d. h. Geldwerte der Hilfsmittel die ein Arzt pro Quartal und Patient verordnen kann, gibt (REHADAT).

Der Abrechnungsweg von Heilmitteln kann wie folgt beschrieben werden: von den jeweiligen Leistungserbringern (Physiotherapeut, Masseur o. Ä.) wird das Verordnungsblatt als Abrechnungsbeleg – entweder unmittelbar oder über ein entsprechendes Abrechnungszentrum – an die jeweilige Krankenkasse weitergeleitet. Bei Inanspruchnahme des Heilmittels fügt der Leistungserbringer weitere Informationen, wie das Institutionskennzeichen, die Kosten des Heilmittels und Zuzahlungen des Versicherten, hinzu (Schröder et al. 2005). Alle diese Variablen können auch in den Heil- und Hilfsmitteldaten der GKV wiedergefunden werden. Siehe hierzu nachfolgend Tabelle 6:

Tabelle 6: Variablenbeschreibung der Heil- und Hilfsmitteldaten

Variable	Mögliche Ausprägungen	Erläuterung
Identifikationsnummer	Ziffernfolge oder Buchstaben-/ Zahlenkombination	Anonymisierte Versicherten-ID; dient als Primärschlüssel und zur personenbezogenen Zuordnung der Leistungen
Datum der Verordnung	TT.MM.JJJJ	Datum der Verordnungsaus-stellung
Datum der Abgabe/Erbringung	TT.MM.JJJJ	Datum der Leistungserbrin-gung bei Heilmitteln und Re-zepteinlösung bei Hilfsmitteln
Positionsnummer	Numerischer Ausdruck z. B. 5-stellige Heilmittelpositi-onsnummer oder 10-stellige Hilfsmittelpositionsnummer	Art des Heil-/Hilfsmittels
Anzahl	Numerischer Ausdruck	Menge je Einzelzeile
Kosten	Euro-Betrag	Kosten aus Perspektive der Krankenversicherung in € (Nettokosten)
Arzt-/Behandlernummer	Numerischer Ausdruck	Arztnummer und KV-Behandlernummer werden in einer Variablen zusammenge-fasst
Kennzeichen Hilfsmittel	Z. B. Neulieferung, Reparatur, Wiedereinsatz, Miete, Nachlie-ferung, Zurichtung, Abgabe eines von der Verordnung abweichenden, höherwertigen Hilfsmittels etc.	Hilfsmittelerbringer haben bei der maschinellen Abrechnung über den Datenträgeraus-tausch nach § 302 SGB V das Feld "Kennzeichen Hilfsmittel" auszufüllen, wenn ein be-stimmter Sachverhalt für die Leistungserbringung zutrifft. Dieses Kennzeichen ist für das jeweilige Hilfsmittel dem Vertrag zu entnehmen
Verknüpfungsvariable	Numerischer Ausdruck	Verordnungs-ID zum Verknüp-fen der Tabellen

Quelle: eigene Darstellung aus §§ 92, 124, 139, 302 SGB V; Grobe und Ihle (2005); Vauth (2010) sowie Zeidler und Braun (2012)

2.4.6 Arbeitsunfähigkeitsdaten und Krankengeld

Die Krankenkassen erhalten Informationen über die Arbeitsunfähigkeit (AU) des Versicherten auf Grundlage des Entgeltfortzahlungsgesetzes (EntgFG 2012). Die vom behandelnden Arzt ausgestellte AU-Bescheinigung wird umgehend an die Krankenkasse weitergeleitet und enthält Informationen über den Befund und die voraussichtliche Krankheitsdauer (§ 295 Abs. 1, Nr. 1 SGB V). Die Inhalte und Form der AU-Daten sind explizit nicht per Gesetz geregelt. Jedoch existieren Musterformulare, die überwiegend verwendet werden (Bödeker 2005).

Die Arbeitsunfähigkeitsdaten sind ein wichtiger Indikator für fehlzeitenbedingte Produktivitätsausfälle oder können beispielsweise für Untersuchungen des betrieblichen Gesundheitsmanagements genutzt werden. Sie beinhalten das Anfangs- und Enddatum der Krankschreibung, die Anzahl an Krankengeldtagen sowie die Höhe des gezahlten Krankentagegeldes. Zu beachten ist jedoch, dass nicht für alle Personen Arbeitsunfähigkeiten gemeldet werden müssen. Hierzu zählen Freiberufler und Selbständige, Schüler, Studenten, Rentner sowie Arbeitslose. Auch Kurzzeitarbeitsunfähigkeit bis zu einer Dauer von drei Tagen müssen nicht zwangsläufig gemeldet werden, da manche Arbeitgeber bei Krankheit von weniger als drei Tagen keine Bescheinigung verlangen (Bödeker 2005; siehe ebenfalls Kapitel 6).

In der Realität werden häufig viele AU-Diagnosen codiert. Diese stehen gleichwertig nebeneinander, sodass nicht ersichtlich ist, welche Diagnose primär zur AU bzw. zum AU-Fall geführt hat. Die Erstellung einer Hierarchie ist hierbei nicht möglich. Des Weiteren ist selten ein genaues Datum der Krankengeldzahlung vorhanden. Eine Zuordnung zu Zeiträumen, wenn das Krankengeld über mehrere Monate gezahlt wird, ist daher schwierig.

Tabelle 7 gibt einen Überblick über die relevanten Variablen in den Daten zur Arbeitsunfähigkeit und zum Krankengeld.

Tabelle 7: Variablenbeschreibung der Arbeitsunfähigkeitsdaten und des Krankengeldes

Variable	Mögliche Ausprägungen	Erläuterung
Identifikationsnummer	Ziffernfolge oder Buchstaben-/ Zahlenkombination	Anonymisierte Versicherten-ID; dient als Primärschlüssel und zur personenbezogenen Zuordnung der Leistungen
Beginn Arbeitsunfähigkeit	TT.MM.JJJJ	Anfangsdatum der Arbeitsunfähigkeit
Ende Arbeitsunfähigkeit	TT.MM.JJJJ	Enddatum der Arbeitsunfähigkeit
Anzahl Arbeitsunfähigkeitstage	Numerischer Ausdruck (in Tagen)	Länge der Arbeitsunfähigkeit in Tagen
Diagnosen	Drei- bis fünfstellige alphanumerische Systematik (z. B. M5419)	ICD-10-Diagnose
Beginn Krankengeld	TT.MM.JJJJ	Anfangsdatum der Krankengeldzahlung
Ende Krankengeld	TT.MM.JJJJ	Enddatum der Krankengeldzahlung
Krankengeldtage	Numerischer Ausdruck (AU-Tage z. B. 2)	Tage des Krankengeldbezugs
Betrag Krankengeld	Geldbetrag in €	Kosten aus Perspektive der Krankenversicherung in €
Facharztgruppe	Alphanumerischer Ausdruck (z. B. G06 Innere Medizin)	Facharztgruppe des jeweiligen behandelnden Arztes
KV-Behandlernummer	Numerischer Ausdruck	Lebenslange Arztnummer

Quelle: eigene Darstellung aus § 300 SGB V, Grobe und Ihle (2005); Vauth (2010) sowie Zeidler und Braun (2012)

2.4.7 Rehabilitation

Die medizinische Rehabilitation verfolgt das Ziel, durch frühzeitige Einleitung der gebotenen Maßnahmen Behinderungen einschließlich chronischer Krankheiten abzuwenden, zu beseitigen, zu mindern, auszugleichen oder eine Verschlimmerung zu verhüten. Darüber hinaus ist das Ziel der medizinischen Rehabilitation, Einschränkungen der Erwerbsfähigkeit und Pflegebedürftigkeit zu vermeiden, zu überwinden, zu mindern, eine Verschlimmerung zu verhüten sowie den vorzeitigen Bezug von laufenden Sozialleistungen zu vermeiden oder laufende Sozialleistungen zu mindern (SGB IX § 26 Abs. 1 Nr. 1 und Nr. 2) (SGB IX 2012).

Die Leistungen zur medizinischen Rehabilitation werden in Deutschland durch verschiedene Sozialleistungsträger finanziert. Zu den sogenannten Rehabilitationsträgern zählen insbesondere die gesetzlichen Krankenkassen, die gesetzliche Rentenversicherung und die gesetzliche Unfallversicherung. Die Gliederung des Systems der Rehabilitation ist historisch gewachsen. Die Zuständigkeit der verschiedenen Kostenträger ist nach dem sogenannten Prinzip der Risikozuordnung geregelt (Tiedt 1996). Gemäß diesem Prinzip ist derjenige Sozialleistungsträger für die Finanzierung einer Rehabilitationsmaßnahme zuständig, der das finanzielle Risiko eines Scheiterns der Rehabilitationsleistung zu tragen hätte. Denn gerade dieser Träger hat ein besonderes Interesse daran, eine Rehabilitationsmaßnahme erfolgreich abzuschließen, um weitere Leistungsansprüche zu vermeiden. Die gesetzliche Krankenversicherung ist nach diesem Prinzip vor allem für Kinder und Jugendliche, nicht berufstätige Erwachsene und Rentner der zuständige Leistungsträger. Die Abrechnung der entstandenen Kosten erfolgt in der Regel direkt zwischen der Krankenversicherung und der Rehabilitationsklinik. Die Kostenerstattungsbeträge werden dabei häufig individuell zwischen der Krankenkasse und der Rehabilitationsklinik verhandelt.

Die gesetzliche Krankenversicherung verfügt nur bei Rehabilitationsmaßnahmen über detaillierte Informationen, für deren Finanzierung sie auch zuständig ist. Bei Rehabilitationsmaßnahmen, die beispielsweise durch die Rentenversicherung finanziert werden, kann die Krankenkasse hingegen in der Regel auf keine bzw. nur sehr eingeschränkte Abrechnungsinformationen zurückgreifen (siehe auch Kapitel 6). Nach Antragseingang klären die Leistungsträger untereinander die Zuständigkeit ab. Ist der zuerst angesprochene Leistungsträger nicht zuständig, leitet dieser den Antrag innerhalb einer Frist von 14 Tagen an den Zuständigen weiter. Sofern der Erstantrag über die Krankenkasse gestellt wurde, liegen dort zumindest die Antragsinformationen vor, auch wenn die Leistung letztendlich über einen anderen Kostenträger finanziert wird. Der zuständige Kostenträger ist in diesen Fällen über die Variable „Kostenträger der Rehabilitationsmaßnahme“ dokumentiert. Weitere Informationen, wie beispielsweise die Art und Dauer der Rehabilitation sowie die damit verbundenen Kosten, sind in diesen Fällen in der Regel nicht dokumentiert.

Bei der Art der Rehabilitation kann zwischen der Anschlussrehabilitation und der weiterführenden Rehabilitation unterschieden werden. Die Anschlussrehabilitation wird in Form von Heilverfahren in Rehabilitationsfachkliniken durchgeführt, die unmittelbar

an eine Krankenhausbehandlung anschließen oder zumindest in einem engen zeitlichen Zusammenhang stehen (maximal 14 Tage nach der Entlassung). In der Systematik der Rentenversicherung wird dieses Verfahren auch als Anschlussheilbehandlung (AHB) bezeichnet. Die Anschlussrehabilitation wurde entwickelt, um bei akuten Erkrankungen oder Gesundheitsstörungen einen möglichst nahtlosen Übergang vom Akutkrankenhaus in die Rehabilitation zu gewährleisten. Die weiterführende Rehabilitation wird in Form von Heilverfahren bei Patienten mit chronischen Erkrankungen durchgeführt. Die gesetzliche Krankenversicherung zielt dabei auf die Verbesserung der Lebensqualität chronisch Kranker oder die Vermeidung von Pflegebedürftigkeit nach dem Grundsatz „Rehabilitation vor Pflege“ ab (Gutenbrunner und Glaesener 2007).

Sowohl die Anschlussrehabilitation als auch die weiterführende Rehabilitation kann als ambulante oder stationäre Rehabilitationsmaßnahme durchgeführt werden. Der Unterschied der ambulanten zur stationären Rehabilitation liegt dabei ausschließlich in der täglichen Rückkehr des Patienten in sein häusliches Umfeld, in der Wohnortnähe der Maßnahme und in der damit gegebenen Nutzung lokaler Ressourcen. Der Anteil ambulanter Rehabilitation ist in den letzten Jahren stark gewachsen. Bei GKV-Routinedatenstudien sollten daher in der Regel auch ambulante Rehabilitationen einbezogen werden.

Sowohl für ambulante als auch für stationäre Rehabilitationsmaßnahmen, die durch die gesetzliche Krankenversicherung finanziert wurden, liegen in den Abrechnungsdaten Informationen zur Diagnose, Aufenthaltsdauer sowie den entstandenen Kosten vor. Bei ambulanten Rehabilitationen sollte zur Bestimmung der Rehabilitationsdauer auf die Variable „Anzahl Tage der Rehabilitation“ zurückgegriffen werden, da die Differenz zwischen Beginn und Ende der Rehabilitation aufgrund von Unterbrechungen und flexiblen Behandlungsalgorithmen nicht zwangsläufig der tatsächlichen Rehabilitationsdauer entsprechen muss.

Einen Überblick über die relevanten Variablen in den Rehabilitationsdaten gibt nachfolgend Tabelle 8.

Tabelle 8: Variablenbeschreibung der Rehabilitationsdaten

Variable	Mögliche Ausprägungen	Erläuterung
Identifikationsnummer	Ziffernfolge oder Buchstaben-/ Zahlenkombination	Anonymisierte Versicherten-ID; dient als Primärschlüssel und zur personenbezogenen Zuordnung der Leistungen
Beginn der Rehabilitation	TT.MM.JJJJ	Datum des Rehabilitationsbeginns
Ende der Rehabilitation	TT.MM.JJJJ	Datum des Rehabilitationsendes
Anzahl Tage der Rehabilitation	Numerischer Ausdruck	Dauer der Rehabilitation in Tagen
Diagnose	Drei- bis fünfstellige alphanumerische Systematik	ICD-10-Diagnose
Art der Rehabilitation	AR: Anschlussrehabilitation WR: Weiterführende Rehabilitation SR: Stationäre Rehabilitation AR: Ambulante Rehabilitation	Rehabilitationssetting
Kosten	Euro-Betrag	Kosten aus Perspektive der Krankenversicherung in €
Kostenträger der Rehabilitation	Krankenkasse, BFA: Bundesversicherungsanstalt für Angestellte, Rentenkasse, Unfallversicherung	Kostenträger
Kurgangskategorie	Ein- bis zweistelliger Code (z. B. 24 = Ambulante kardiologische Rehabilitation)	Spezifizierung der Rehabilitation
IK-Nummer der Rehabilitationsklinik	Neunstellige Nummer	Identifikationsnummer der Institution
Arzt-/Behandlernummer	Numerischer Ausdruck	Arztnummer und KV-Behandlernummer des einweisenden Arztes

Quelle: eigene Darstellung aus § 301 (4) SGB V; Grobe und Ihle (2005); Vauth (2010); Zeidler und Braun (2012); GKV-Datenaustausch (e)

2.4.8 Disease-Management-Programme

Im Rahmen der Disease-Management-Programme (DMP) werden detaillierte Daten für chronisch kranke Patienten erhoben und dokumentiert. Grundsätzlich ist die Teilnahme eines Patienten an diesen Programmen freiwillig. Ärzte müssen ihre Teilnahme gegenüber der Kassenärztlichen Vereinigung erklären und nach einer Prüfung der Strukturvoraussetzungen die Teilnahme nochmals bestätigen. Eine (elektronische) Teilnahmeerklärung und die Erstdokumentation werden gemeinsam vom Arzt und Patienten ausgefüllt. Die erhobenen Daten werden vollständig sowohl der jeweiligen Krankenkassen als auch der KV übermittelt.

Diese strukturierten Behandlungsprogramme liefern ergänzende und weiterführende Daten zu den eingeschriebenen Versicherten wie z. B. Körpergröße, Gewicht und Raucherstatus. Forschungseinrichtungen sind nach § 137f SGB V befugt, als externe Sachverständige diese Programme zu evaluieren. Dennoch zeigt sich, dass die Daten derzeit noch unzureichend gepflegt sind (Horenkamp-Sonntag und Linder 2012). Dies könnte daran liegen, dass zunächst die Dokumentation lediglich in Papierform vorlag und erst im späteren Verlauf auf die elektronische Datenverarbeitung umgestellt wurde. Da nicht alle Patienten in diesen Programmen eingeschrieben sind, können die ergänzenden Informationen lediglich für die teilnehmende Subgruppe des Versicherungsbestandes genutzt werden.

Tabelle 9 gibt einen Überblick über die relevanten Variablen in den Daten zu den Disease-Management-Programmen.

Tabelle 9: Variablenbeschreibung der Daten der Disease-Management-Programme

Variable	Mögliche Ausprägungen	Erläuterung
Identifikationsnummer	Ziffernfolge oder Buchstaben-/ Zahlenkombination	Anonymisierte Versicherten-ID; dient als Primärschlüssel und zur personenbezogenen Zuordnung der Leistungen
Beginndatum der DMP-Teilnahme	TT.MM.JJJJ	Datum des DMP-Teilnahmebeginns
Enddatum der DMP-Teilnahme	TT.MM.JJJJ	Datum des DMP-Teilnahmeendes
Art des DM-Programms	Brustkrebs, Diabetes mellitus Typ II, Koronare Herzkrankheit	Indikation des DM-Programms

Variable	Mögliche Ausprägungen	Erläuterung
	(KHK), Diabetes mellitus Typ I, Chronisch obstruktive Atemwegserkrankungen (COPD), Asthma bronchiale	
Körpergewicht des Versicherten	Numerischer Ausdruck in kg	Körpergewicht des Versicherten
Körpergröße des Versicherten	Numerischer Ausdruck, Größe in cm	Körpergröße des Versicherten
Raucherstaus	J = Ja, N = Nein	Raucherstaus des Versicherten
ACE-Hemmer	J = Ja, N = Nein, K = Kontraindikation, NK = Nein und Kontraindikation	Angaben zur Einnahme von ACE-Hemmern
Betablocker	J = Ja, N = Nein, K = Kontraindikation, NK = Nein und Kontraindikation	Angaben zur Einnahme von Betablockern
HMG-CoA-Reduktase-Hemmer	J = Ja, N = Nein, K = Kontraindikation, NK = Nein und Kontraindikation	Angaben zur Einnahme von HMG-CoA-Reduktase-Hemmern
Thrombozytenaggregations-hemmer	J = Ja, N = Nein, K = Kontraindikation, NK = Nein und Kontraindikation	Angaben zur Einnahme von Thrombozytenaggregations-hemmern
Modulteilnahme – Chronische Herzinsuffizienz	J = Ja, N = Nein	Modulteilnahme des Versicherten – Chronische Herzinsuffizienz
Serum-Kreatinin	Serumkreatinin in mg/dl	Angaben zum Serumkreatinin-Gehalt im Urin
Serum-Elektrolyte	B = Bestimmt, N = Nicht bestimmt, leer = keine Informationen im Dokumentationsdatensatz übermittelt (nur bei gleichzeitiger Teilnahme am Modul Herzinsuffizienz ist diese Angabe verpflichtend)	Laborparameter; gibt Auskunft über die Konzentrationen von Elektrolyten im Blut wider
Serum-KR_mol	Serumkreatinin in µmol/l	Serumkreatinin ist ein Laborparameter, der zur groben Abschätzung der Nierenfunktion bestimmt wird

Quelle: Reinhold et al. 2011a

2.4.9 Daten der Institutsambulanzen

Psychiatrische Fachkrankenhäuser sowie psychiatrische Abteilungen an Allgemeinkrankenhäusern sind gemäß § 118 SGB V zur Einrichtung psychiatrischer Institutsambulanzen (PIA) ermächtigt. Versicherte, die wegen Art, Schwere oder Dauer ihrer Erkrankung oder wegen zu großer Entfernung zu geeigneten Fachärzten auf eine ambulante psychiatrische oder psychotherapeutische Behandlung angewiesen sind, können diese in einer PIA in Anspruch nehmen. Die PIA-Behandlung ist bei chronischen oder chronisch rezidivierenden psychischen Krankheiten indiziert, zu denen insbesondere Schizophrenien, affektive Störungen und schwere Persönlichkeitsstörungen sowie Suchtkrankheiten mit Komorbidität und gerontopsychiatrische Krankheiten gehören (Melchinger 2008). Der Patientenzugang erfolgt durch die Überweisung einer psychiatrischen Abteilung oder eines niedergelassenen Vertragsarztes. Das Leistungsspektrum der PIA umfasst das gesamte Spektrum psychiatrisch-psychotherapeutischer Diagnostik und Therapie. In Deutschland besteht eine fast flächendeckende Versorgung mit PIA (Melchinger 2008).

Die Leistungen der PIA werden gemäß § 120 SGB V unmittelbar von den Krankenkassen vergütet. Die Vergütung erfolgt außerhalb des vertragsärztlichen Gesamtbudgets, wobei drei unterschiedliche Vergütungsmodelle zum Einsatz kommen können (Melchinger 2008). Dabei sind sowohl Quartalspauschalen, eine Vergütung nach besonderen Komplexleistungen als auch eine Vergütung nach EBM möglich. Für Versicherte, die in pauschaliert vergüteten PIA behandelt werden, liegen in den Routinedaten der Krankenkassen häufig keine detaillierten Informationen zu der PIA-Behandlung vor. In diesen Fällen sind weder ICD-Diagnosen noch Informationen zu den im Detail erbrachten Leistungen dokumentiert. Es können lediglich Informationen zu eingegangenen Rechnungen sowie das Buchungsdatum extrahiert werden (siehe Kapitel 6). Diese Informationen können jedoch zumindest einen Hinweis darauf geben, wie viele Versicherte durch PIA behandelt wurden und zu welchen Kosten diese Behandlung geführt hat.

Neben den PIA existieren im deutschen Gesundheitssystem eine Reihe weiterer Ambulanzen. Als Beispiel können geriatrische Institutsambulanzen, die zu einer strukturierten und koordinierten ambulanten geriatrischen Versorgung von Versicherten ermächtigt sind, genannt werden (§ 118a SGB V). Bei Studien, die Krankheitsbilder beinhalten, die in Institutsambulanzen behandelt werden können, sollte mit dem

Dateneigner die Datenverfügbarkeit abgestimmt werden. Potenzielle Informationsdefizite sind bei der Studienplanung zu berücksichtigen.

Tabelle 10 gibt einen Überblick über die relevanten Variablen in den Daten zu den Institutsambulanzen.

Tabelle 10: Variablenbeschreibung der Institutsambulanzen

Variable	Mögliche Ausprägungen	Erläuterung
Identifikationsnummer	Ziffernfolge oder Buchstaben-/ Zahlenkombination	Anonymisierte Versicherten-ID; dient als Primärschlüssel und zur personenbezogenen Zuordnung der Leistungen
Buchungsdatum	TT.MM.JJJJ	Datum der Rechnung
Buchungsbetrag	Kosten in €	Kosten aus Perspektive der Krankenversicherung in €

Quelle: eigene Darstellung

Allgemeine Empfehlungen zu den Datenkategorien

- Bei der Studienplanung sollten alle relevanten Variablen definiert und die Verfügbarkeit mit der Krankenkasse abgestimmt werden
- Der Prozess der Datenerhebung und -übermittlung muss bei der qualitativen Beurteilung der einzelnen Variablen berücksichtigt werden
- Die Aussagekraft und Validität der einzelnen Variablen muss gemeinsam mit der Krankenkasse im Hinblick auf die Forschungsfragen evaluiert und gesichert werden
- Die Ergänzung fehlender Information aus weiteren Datenquellen sollte geprüft werden
- Dem Grundsatz der Datensparsamkeit ist Rechnung zu tragen
- Die Limitationen der Variablen sind zu berücksichtigen (Siehe Kapitel 6)

3 Studiendesigns

Die Studiendesigns für die Analyse von GKV-Routinedaten sind vielfältig und hängen von der Fragestellung der jeweiligen Studien ab, wobei diese sowohl ökonomische als auch medizinische und epidemiologische sowie viele weitere Fragestellungen adressieren. Grundsätzlich eignet sich diese Datenquelle als Grundlage für viele unterschiedliche Beobachtungsstudien. Im Folgenden werden verschiedene Analysemöglichkeiten vorgestellt und Stärken sowie Schwächen diskutiert. Ein Anspruch auf Vollständigkeit kann aufgrund der Vielfalt an Fragestellungen jedoch nicht erhoben werden.

3.1 Gesundheitsökonomische Analysen

Der Bedarf an standardisierten gesundheitsökonomischen Bewertungen, die Informationen über die Kosten und Effekte medizinischer Verfahren bereitstellen, hat in den letzten Jahren in Deutschland stetig zugenommen. Großes Interesse besteht insbesondere an Informationen zum tatsächlichen ökonomischen Einfluss medizinscher Verfahren unter Realbedingungen im Versorgungsalltag. GKV-Routinedaten haben sich als eine sinnvolle Grundlage für gesundheitsökonomische Studien erwiesen (Reinhold et al. 2011b; Zeidler und Braun 2012). Dies ist vor allem auf ihre originäre Zweckbestimmung zurückzuführen. Da GKV-Routinedaten für Abrechnungszwecke erhoben werden, umfassen sie nahezu alle Ressourcenverbräuche, die einen Erstattungsanspruch an die GKV beinhalten. Daher sind diese Daten besonders geeignet für ökonomische Analysen aus der Perspektive der GKV (Reinhold et al. 2011b).

Die in gesundheitsökonomischen Studien zu berechnenden Kostenkomponenten lassen sich in direkte, indirekte und intangible Kosten unterscheiden (Greiner und Damm 2012). Unter direkten Kosten werden Ressourcenverbräuche subsumiert, die für medizinische Leistungen in der Prävention, Diagnostik, Behandlung, Rehabilitation und Palliativmedizin (direkte medizinische Kosten) sowie für nicht-medizinische Leistungen (direkte nicht-medizinische Kosten), wie beispielsweise Kosten für Krankentransporte, aufgewendet werden. Sofern die direkten Kostenkomponenten einen Erstattungsanspruch an die GKV besitzen, lassen sie sich in der Regel umfassend durch GKV-Routinedatenanalysen abbilden. Dies gilt beispielsweise für die Kosten stationärer Krankenhausaufenthalte oder die Verordnung erstattungsfähiger Arzneimittel. Direkte Kosten von Leistungen, die durch andere Kostenträger, wie beispiels-

weise von der Rentenversicherung verwaltete Rehabilitationsmaßnahmen, finanziert werden, lassen sich hingegen auf dieser Datengrundlage nicht abbilden (Holle et al. 2005). Gleiches gilt für Leistungen, die, wie beispielsweise individuelle Gesundheitsleistungen (IGeL), durch die Patienten privat finanziert werden.

Neben den direkten Kosten einer Leistung können auch indirekte Wirkungen bei ökonomischen Studien berücksichtigt werden. Indirekte Kosten erfassen den volkswirtschaftlichen Produktionsverlust aufgrund von krankheitsbedingter Abwesenheit vom Arbeitsplatz, Invalidität oder vorzeitigem Tod. Auf Basis von GKV-Routinedaten lassen sich Informationen aus den Arbeitsunfähigkeitsdaten zur approximativen Berechnung indirekter Kosten heranziehen. Durch eine entsprechende Bewertung können Fehlzeiten, z. B. mittels Humankapitalansatz (für weiterführende Informationen siehe Greiner und Damm 2012), in indirekte Kosten überführt werden (Reinhold et al. 2011b). Zur Berechnung von Produktivitätsverlusten wird gemäß den aktuellen deutschen Empfehlungen zur gesundheitsökonomischen Evaluation die folgende Formel vorgeschlagen (Greiner und Damm 2012; Graf von der Schulenburg et al. 2007):

$$\textit{Indirekte Kosten} = \textit{Arbeitsunfähigkeitstage} \cdot \frac{\textit{Arbeitnehmerentgelt in Deutschland pro Jahr}}{\textit{Arbeitnehmer} \cdot \textit{365 Tage}}$$

Die Arbeitsunfähigkeitstage können direkt den Arbeitsunfähigkeitsdaten der Krankenkassen diagnosebezogen entnommen werden. Zur monetären Bewertung der entstandenen Produktivitätsverluste wird das Arbeitnehmerentgelt herangezogen, das den offiziellen Statistiken des Statistischen Bundesamts entnommen werden kann.

Als dritte Kostenkomponente können intangible Kosten genannt werden, die Faktoren wie Schmerz, Freude oder physische Einschränkungen bezeichnen. Diese Effekte sind per Definition kaum einer monetären Berechnung zu unterziehen und können daher in der Regel nicht mit GKV-Routinedaten abgebildet werden. Zur Berechnung intangibler Kosten müssen daher andere Datenquellen, beispielsweise Informationen aus Patientenbefragungen, genutzt werden (Greiner und Damm 2012).

Die Zurechnung von Kosten auf bestimmte Leistungen hängt von der Perspektive der Untersuchung ab. Neben der Krankenkassenperspektive können hier insbesondere

die gesellschaftliche Perspektive, die eine Bewertung aus Sicht der gesamten Volkswirtschaft umfasst, die Perspektive der Leistungserbringer (Ärzte, Krankenhäuser etc.) und die Patientenperspektive genannt werden. Da die Perspektive der Krankenkasse im Wesentlichen direkte Kosten enthält, die sich mit GKV-Routinedaten besonders gut berechnen lassen, bietet sich diese Datenquelle insbesondere für Studien aus der Kostenträgerperspektive an. Sofern indirekte Kosten anhand der Arbeitsunfähigkeitsinformationen berechnet werden, lassen sich die dokumentierten Daten jedoch auch auf die gesellschaftliche Ebene extrapolieren. Auch die Abbildung der Patientenperspektive ist möglich, da anhand der GKV-Routinedaten teilweise z. B. auf die Höhe der Patientenzuzahlungen geschlossen werden kann. Anhand der Informationen aus den Leistungssektoren lässt sich darüber hinaus die Perspektive einzelner Leistungserbringer abbilden.

Bei gesundheitsökonomischen Analysen kann zwischen verschiedenen Studienformen unterschieden werden (Schöffski 2012). Dabei wird zwischen Studien mit vergleichendem und ohne vergleichenden Charakter differenziert. Zu den nicht vergleichenden Studientypen zählen Kostenanalysen und Krankheitskostenanalysen. Bei Kostenanalysen werden die mit einer Intervention verbundenen Ressourcenverbräuche einer monetären Bewertung unterzogen. In Krankheitskostenanalysen werden hingegen die Kosten von Erkrankungen und Ereignissen sowie die Einflussfaktoren der Kosten einzelner Erkrankungen untersucht. Diese Analysen können Informationen bereitstellen, wie stark eine Volkswirtschaft durch bestimmte Krankheiten und deren Folgen belastet wird. Krankheitskostenanalysen dienen somit als Instrument zur Entscheidungsfindung für die Politik, da eine größenmäßige Schätzung der ökonomischen Konsequenzen verschiedenerer Krankheiten die Grundlage rationaler Allokationsprozesse und Prioritätensetzung darstellt (Reis 2005). Sowohl Kostenanalysen als auch Krankheitskostenanalysen sind in der Regel auf Basis von GKV-Routinedaten durchführbar (Reinhold et al. 2011b; Zeidler und Braun 2012). Dabei lassen sich sowohl Querschnittsanalysen (Prävalenzansatz) als auch Längsschnitt- oder Longitudinalanalysen (Inzidenzansatz) umsetzen.

Bei den vergleichenden Studien kann zwischen Kosten-Kosten-, Kosten-Nutzen-, Kosten-Wirksamkeits- und Kosten-Nutzwert-Analysen unterschieden werden (Abbildung 3). Die Wahl der Analyseform hängt vom Untersuchungsgegenstand und dem Zweck der Studie ab.

Abbildung 3: Systematik gesundheitsökonomischer Evaluationen

vergleichend				nicht vergleichend	
Kosten-Kosten-Analyse	Kosten-Nutzen-Analyse	Kosten-Wirksamkeits-Analyse	Kosten-Nutzwert-Analyse	Kosten-Analyse	Krankheits-kosten-Analyse

Quelle: in Anlehnung an Schöffski (2012)

Für die gesundheitsökonomische Evaluation werden häufig vergleichende Studiendesigns eingesetzt. Die einfachste Form stellen Kosten-Kosten-Analysen dar. Hierbei handelt es sich im Prinzip um zwei separate Kosten-Analysen von zwei oder mehr alternativen Maßnahmen mit dem Ziel, die kostengünstigste Alternative zu ermitteln. Dieser Methode liegt die Annahme zugrunde, dass die beiden Maßnahmen zu einem identischen Behandlungsergebnis bzw. Outcome führen. Unter dieser Voraussetzung kann die Beurteilung der Vorteilhaftigkeit auf einen reinen Kostenvergleich reduziert werden. Die Situation der gleichen Wirksamkeit ist im Gesundheitswesen jedoch selten gegeben; als praktisches Beispiel können Generika genannt werden, die in der Regel wirkungsgleich wie das Originalpräparat sind (Greiner und Damm 2012). Sofern eine gleiche Wirksamkeit sichergestellt werden kann, eignen sich GKV-Routinedaten hervorragend für Kosten-Kosten-Analysen. So haben beispielsweise Zeidler et al. mit diesem Studiendesign einen Kostenvergleich der ambulanten und stationären Rehabilitation durchgeführt (Zeidler et al. 2008a; Zeidler et al. 2008b). Dieses Verfahren war möglich, da eine äquivalente Wirksamkeit der ambulanten und stationären Rehabilitation bereits in mehreren Studien nachgewiesen werden konnte.

Die klassische Form von ökonomischen Evaluationen, insbesondere in Bereichen außerhalb des Gesundheitswesens, ist die Kosten-Nutzen-Analyse (Greiner und Damm 2012). Bei diesem Verfahren werden sämtliche Kosten und Nutzen der zu evaluierenden Maßnahmen in Geldeinheiten bewertet. Die gesundheitsökonomische Bewertung von Arzneimitteln kann mit einer Kosten-Nutzen-Analyse beispielsweise erfolgen, indem aus den GKV-Routinedaten zunächst die relevante Zielpopulation identifiziert wird. Dies kann anhand bestimmter Diagnosen sowie weiteren Patienteneigenschaften, wie z. B. Geschlecht, Alter, Region, erfolgen (Reinhold et al. 2011b; siehe Kapitel 4.1). Anschließend kann auf Basis von ATC-Codes oder der Pharma-

zentralnummer eine bestimmte Medikamentengruppe oder ein konkretes Medikament identifiziert werden. Häufig wird bei gesundheitsökonomischen Analysen auch eine geeignete Kontrollgruppe identifiziert, die das zu untersuchende Medikament nicht eingenommen hat. Die möglichen monetären Nutzeneffekte ergeben sich bei Kosten-Nutzen-Analysen dann durch einen Vergleich der Kosten mit der Vergleichsgruppe. Die monetären Nutzeneffekte werden dabei mit den unterschiedlichen Kostenaufwendungen verrechnet, die zur Realisierung der Intervention erforderlich sind. Kosten-Nutzen-Analysen sind jedoch in Deutschland umstritten, da die Nutzenkomponenten auch intangible Effekte beinhalten, deren Bewertung in Geldeinheiten mit großen Herausforderungen verbunden ist (Greiner und Damm 2012). Auch wenn in den letzten Jahren substanzielle methodische Weiterentwicklungen auf dem Gebiet der monetären Bewertung von Nutzenkomponenten verzeichnet werden konnten, sind GKV-Routinedaten ohne die Ergänzung von Primärdaten nur selten für Kosten-Nutzen-Analysen geeignet.

Kann keine monetäre Bewertung der möglichen mit einer Therapie verbundenen Nutzeneffekte vorgenommen werden, können die Therapieeffekte in Form naturalistischer Parameter gemessen werden. Dabei kommen sogenannte Kosten-Wirksamkeits-Analysen zum Einsatz, welche die nicht unmittelbar in monetäre Einheiten bewertbaren Effekte in naheliegenden natürlichen Einheiten messen. Auf Basis von GKV-Routinedaten können mit diesem Verfahren beispielsweise folgende Outcomes erfasst und berechnet werden (Reinhold et al. 2011b):

- Anhand von ICD-10-Codes
    - Kosten pro vermiedenem Event
    - Kosten pro vermiedenem Rezidiv
    - Kosten pro vermiedener Neuerkrankung
- Anhand der Information „Austrittsgrund: Tod“
    - Kosten pro vermiedenem Todesfall (gerettetem Menschenleben)
- Anhand Arzt/Klinikkontakte
    - Kosten pro vermiedenem Krankenhausaufenthalt
    - Kosten pro vermiedenem Arztkontakt

Aufgrund der unzureichenden Dokumentation naturalistischer Parameter wie z. B. in Form medizinischer Messwerte (Blutdruckwerte, Tumorstadien, Daten zur Lebens-

qualität etc.) ist das Potenzial für die Durchführung von Kosten-Wirksamkeits-Analysen auf Basis von GKV-Routinedaten jedoch ebenfalls eingeschränkt (Reinhold et al. 2011b; siehe auch Kapitel 6). Dies gilt auch für Kosten-Nutzwert-Analysen, welche die Effekte auf die Lebensqualität und die Lebenserwartung des Patienten anhand standardisierter Nutzwerte, wie beispielsweise mittels des QALY-Konzepts, erfassen. QALY-Werte lassen sich nicht auf Basis von GKV-Routinedaten berechnen, sodass eine Durchführbarkeit nur über die Ergänzung von Primärdaten, insbesondere aus Versichertenbefragungen, erreicht werden kann. Die Vor- und Nachteile von Primär- und Sekundärdaten werden durch Schreyögg und Stargardt systematisch gegenübergestellt sowie konkrete Beispiele für routinedatenbasierte ökonomische Studien gegeben (Schreyögg und Stargardt 2012). Diese Systematik kann als Entscheidungshilfe bei der Auswahl der Datengrundlage dienen.

Bei allen vergleichenden Analyseformen, unabhängig davon, welche Methode konkret gewählt wurde, besteht die Gefahr eines Selektionsbias. Dies bedeutet, dass es aufgrund von Patienteneigenschaften, wie dem Vorliegen und der Schwere einer Erkrankung, zu einem Selektionseffekt bestimmter Patienten zu bestimmten Therapiemaßnahmen kommen kann (Reinhold et al. 2011b). Dies kann zu stark verzerrten Ergebnissen und entsprechenden Fehlinterpretationen führen. Bei Beobachtungsstudien, welche die beabsichtigten Effekte von therapeutischen Maßnahmen untersuchen, kann es sogar zu besonders stark ausgeprägtem Confounding kommen (sogenanntem Confounding by indication). Bei Untersuchungen von unerwarteten und unbeabsichtigten Effekten ist hingegen mit einem deutlich niedrigeren Potenzial für Verzerrungen durch Confounding zu rechnen.

Unabhängig von der gewählten Methode ist daher bei vergleichenden Studien die Vergleichbarkeit der zu evaluierenden Patientengruppen sicherzustellen. Einen guten Überblick über mögliche vergleichende Studiendesigns bieten Zeidler und Braun (2012). Eine Lösung kann in der Verwendung entsprechender Adjustierungsmethoden, beispielsweise einem Kontrollgruppendesign, liegen. Die Versicherten der Kontrollgruppe sollten sich von der Interventionsgruppe möglichst in allen relevanten Eigenschaften nicht unterscheiden; dies kann beispielsweise durch ein Matching nach Risikofaktoren (wie beispielsweise Propensity Score Matching) sichergestellt werden. Der Propensity Score ist definiert als die Wahrscheinlichkeit bei gegebenen Kovariaten einer der Vergleichsgruppen zugehörig zu sein. Matching-Verfahren beinhalten

viele heterogene Methoden (Zeidler und Braun 2012). Die Wahl des geeignetsten Verfahrens hängt von der konkreten Forschungsfrage ab.

Allgemein ist jedoch zu berücksichtigen, dass die Möglichkeiten zur statistischen Adjustierung bei GKV-Routinedaten eingeschränkt sind, da nur eine begrenzte Anzahl patientenrelevanter Parameter vorliegt. Zur möglichst effektiven Nutzung aller durch GKV-Routinedaten abgebildeten Confounder-Informationen wird die Bildung von sogenannten high-dimensional Propensity Scores vorgeschlagen (Reinhold et al. 2011b). Bei diesem Verfahren werden nicht nur die durch den Wissenschaftler als relevant angesehenen Confounder berücksichtigt, sondern es wird mit einem empirischen Vorgehen automatisch nach weiteren wichtigen Confoundern gesucht. Jedoch kann selbst bei einer maximal effizienten Ausnutzung der vorhandenen Informationen weiterhin ein Risiko für Verzerrungen bestehen, da nicht zwangsläufig alle Confounder in den GKV-Routinedaten erfasst sein müssen. Eine Möglichkeit, den Einfluss von ungemessenen Confoundern zu evaluieren, kann in der Nutzung von Sensitivitätsanalysen liegen (Reinhold et al. 2011b). So könnte berechnet werden, wie stark ein hypothetisch ungemessenes Confounding sein müsste, um das beobachtete Studienergebnis zu erklären. Unter Kenntnis dieser Information ließe sich abschätzen, ob das Vorliegen eines solchen, bislang unbekannten Confounders überhaupt realistisch erscheint.

Gesundheitsökonomische Analysen werden häufig in Form von Modellierungsstudien praktisch umgesetzt. Um qualitativ hochwertige Entscheidungsmodelle zu entwickeln, sind eine Reihe von Inputvariablen, die sowohl die Kosten als auch den Nutzen unterschiedlicher Gesundheitstechnologien systematisch beschreiben, erforderlich. In diesem Zusammenhang stellen Patientenflussanalysen eine vermutlich in Zukunft an Bedeutung gewinnende Analyseform dar (Reinhold et al. 2011b). Ziel dieser Analyse ist die Abbildung der Patientenwege durch das Versorgungssystem. Dabei wird eine vorab definierte Patientengruppe hinsichtlich der in Anspruch genommenen Ressourcen analysiert. Die Ableitung von Wahrscheinlichkeiten kann in Kombination mit den Daten einer Kostenanalyse als Grundlage für die Erstellung gesundheitsökonomischer Stochastikmodelle, wie beispielsweise Entscheidungsbaumanalysen oder Markov-Modellen, dienen. Beispielsweise haben Frey et al. 2013 die Kosteneffektivität verschiedener Antipsychotika zur Behandlung von Schizophrenien anhand eines

Markov-Modells evaluiert, das mit Inputfaktoren aus GKV-Routinedaten spezifiziert worden ist (Frey et al. 2013).

Obwohl die Qualität der GKV-Routinedaten für ökonomische Analysen in den Hauptkostenbereichen recht gut ist, wird diese Datenquelle für gesundheitsökonomische Evaluationen in Deutschland noch relativ selten eingesetzt (Schreyögg und Stargardt 2012). Dies liegt zum einen daran, dass vielen Forschern die Breite und Tiefe der bei den größeren Krankenkassen liegenden Datenbestände noch nicht bekannt ist. Zum anderen werden für die Erstattungsentscheidungen des Instituts für Qualität und Wirtschaftlichkeit (IQWiG) primär Ergebnisse aus randomisierten klinischen Studien herangezogen. Aktuell werden GKV-Routinedaten daher vornehmlich für Krankheitskostenanalysen, Kosten-Kosten-Analysen und Kosten-Wirksamkeits-Analysen verwendet (Schreyögg und Stargardt 2012). Aufgrund der offensichtlichen Potenziale ist jedoch in Zukunft insgesamt ein vermehrter Einsatz dieser Datenquelle im Rahmen von ökonomischen Studien zu erwarten.

Methoden zur Berechnung indikationsspezifischer Ressourcenverbräuche

Unabhängig von der Wahl des generellen Studiendesigns ist bei jeder ökonomischen Studie zu entscheiden, wie die relevanten Kosten anhand der GKV-Routinedaten konkret berechnet werden sollen. Eine besondere methodische Herausforderung ergibt sich bei der Kalkulation indikationsspezifischer Ressourcenverbräuche, d. h. bei der Identifikation derjenigen Kosten, die auf die Zielerkrankung sowie die damit zusammenhängenden Komorbiditäten zurückzuführen sind. Die alleinige Betrachtung der Gesamtkosten würde zu einer Überschätzung der Behandlungskosten führen, da in den GKV-Routinedaten weitgehend alle Ressourcenverbräuche eines Patienten, unabhängig von der in einer Studie zu untersuchenden Zielerkrankung, erfasst sind. Um eine Überschätzung der Kosten zu vermeiden, sind daher bei gesundheitsökonomischen Studien in der Regel indikationsspezifische Kosten anzugeben. Die Kosten der Zielerkrankung sind also sorgfältig von den Kosten anderer Erkrankungen abzugrenzen. In der Gesundheitsökonomie existieren verschiedene Methoden zur Identifikation der indikationsspezifischen Ressourcenverbräuche (Zeidler et al. 2013). In den folgenden Abschnitten werden diese unterschiedlichen Methoden dargestellt.

Expertengestützte Methode

Bei der expertengestützten Methode werden die gesamten Leistungsausgaben um diejenigen Kosten gemindert, die nicht der Zielerkrankung zugeordnet werden können (Zeidler et al. 2013). Hierzu werden anhand standardisierter Klassifikationsinstrumente, wie z. B. der ICD-Klassifikation, der ATC-Klassifikation oder dem EBM, alle Leistungen identifiziert, die mit der Zielerkrankung in Zusammenhang stehen. Für die Identifikation relevanter Leistungen muss entsprechendes Expertenwissen zur Verfügung stehen, da nur eine vollständige Identifikation aller relevanten Leistungen eine valide Kalkulation der Behandlungskosten ermöglicht. Daher kann die Einbeziehung eines Mediziners oder Abrechnungsexperten bei vielen Analysen sinnvoll sein. Im Anschluss an die expertengestützte Definition relevanter Leistungen werden die Kosten dieser spezifischen Abrechnungsvorgänge ermittelt und der Zielerkrankung zugeordnet. Welche Leistungen im Einzelfall konkret einer Erkrankung zugeteilt werden müssen, kann nur im Hinblick auf die jeweilige Forschungsfrage und Zielindikation entschieden werden. Im Folgenden können daher nur Beispiele für mögliche methodische Herangehensweisen gegeben und die generellen Vor- und Nachteile der eingesetzten Methoden diskutiert werden.

Zeidler et al. haben die expertengestützte Methode für die Kalkulation der Kosten der Aufmerksamkeitsdefizit-/Hyperaktivitätsstörung (ADHS) genutzt (Zeidler et al. 2013). Für die Kalkulation der Krankheitskosten wurden Krankenhaus- und Rehabilitationsaufenthalte, ambulante Versorgungsdaten, Arznei- und Heilmitteldaten sowie Arbeitsunfähigkeitsdaten berücksichtigt. Dabei wurden alle Krankenhausaufenthalte, Rehabilitationsmaßnahmen und Krankengeldzahlungen als ADHS-spezifisch definiert, die auf die Diagnose F90.- (Hyperkinetische Störungen) zurückzuführen sind (Zeidler et al. 2013). Bei stationären Krankenhausaufenthalten wurden hierfür sowohl Haupt- als auch Nebendiagnosen berücksichtigt. Bei Rehabilitationsmaßnahmen liegt eine eindeutige Diagnose vor und bei Krankengeldzahlungen wurde für jeden Bezugszeitraum geprüft, ob mindestens eine ADHS-Diagnose in den Arbeitsunfähigkeitsdaten vorlag. Zur Identifikation der ambulanten Ressourcenverbräuche wurde ein zweistufiges Verfahren eingesetzt, da eine direkte Verknüpfung zwischen Diagnosen und einzelnen ambulanten Leistungen (EBM-Ziffern) aufgrund der quartalsweisen Diagnosedokumentation nicht möglich war. Daher wurde zunächst für jeden Patienten individuell geprüft, ob in dem jeweiligen Quartal eine gesicherte ADHS-

Diagnose vorgelegen hat. War dies der Fall, wurden alle EBM-Ziffern als indikationsspezifisch definiert, die auf ausgewählte Gebührenordnungspositionen von spezifischen Fachärzten (Kinder- und Jugendpsychiater, Psychiater) zurückzuführen waren. Da für andere Leistungsbereiche in der Regel keine ICD-Diagnosen vorliegen, wurden weitere Klassifikationsinstrumente zur Selektion der indikationsspezifischen Kosten genutzt. Indikationsspezifische Arzneimittel wurden anhand der ATC-Klassifikation und Heilmittel anhand des Heilmittelpositionsnummernverzeichnisses identifiziert. Diesem Verfahren liegt die Annahme zugrunde, dass die Leistungen ausschließlich zur Behandlung der Zielerkrankung sowie der damit zusammenhängenden Komorbiditäten eingesetzt wurden.

Die expertengestützte Methode besitzt den Vorteil, dass die krankheitsrelevanten Leistungen eindeutig durch medizin-theoretische Vorüberlegungen und Expertenbefragungen strukturiert werden können. Außerdem kann dem Grundsatz der Datensparsamkeit am besten Rechnung getragen werden, da keine Daten von Kontrollgruppenpatienten erforderlich sind, sondern nur Daten für die Patienten mit der Zielerkrankung extrahiert werden müssen. Ein Nachteil dieser Methode ergibt sich aus der Tatsache, dass beispielsweise in den Arznei-, Heil- und Hilfsmitteldaten keine Diagnosen gespeichert sind. Daher kann keine Aussage darüber getroffen werden, ob die betrachteten Leistungen tatsächlich ausschließlich zur Behandlung der Zielerkrankung und der damit assoziierten Komorbiditäten eingesetzt wurden oder ob eine davon unabhängige Erkrankung behandelt werden sollte. Zur Verfeinerung der Identifikationsmethode ist jedoch bei Leistungen, zu denen keine explizite Diagnosedokumentation vorliegt, eine Verknüpfung mit ambulanten und/oder stationären Diagnosen denkbar. So könnten spezifische Arzneimittel oder Heilmittel nur dann einer Zielerkrankung zugeordnet werden, wenn in einem bestimmten Zeitraum vor der Verordnung eine entsprechende ambulante oder stationäre Diagnose vorgelegen hat. Auch eine Berücksichtigung der Fachgruppe des verordnenden Arztes wäre zur weiteren Spezifizierung denkbar.

Bei Leistungsbereichen, in denen ICD-Diagnosen verfügbar sind, entstehen immer dann Unschärfen, wenn mehrere Diagnosen gleichwertig nebeneinander stehen. Dies gilt insbesondere für ambulante Diagnosen und Arbeitsunfähigkeitsdiagnosen, bei denen eine Leistung häufig nicht eindeutig einem einzelnen Diagnoseschlüssel zugeordnet werden kann. Im Bereich der ambulant-ärztlichen Leistungen wird diese

Limitation durch die quartalsweise Diagnosecodierung, die keine eindeutige Verknüpfung mit den tagesgenau erfassten Leistungsdaten erlaubt, zusätzlich verschärft. Auch im Krankenhausbereich existieren Herausforderungen bei der Zuordnung von Diagnosen. So stellt sich beispielsweise die Frage, ob nur Haupt- oder auch Nebendiagnosen der Zielerkrankung zugeordnet werden. In der Literatur finden sich sowohl Studien, die sich ausschließlich auf Primärdiagnosen stützen, als auch Studien, die zusätzlich Sekundärdiagnosen einbeziehen (für eine systematische Übersicht siehe Zeidler et al. 2013). Wesentliche Nachteile dieser Methode sind daher mögliche Unschärfen bei der Kausalität einzelner Leistungen (eine Herausforderung, die sich auch mit den im Folgenden vorgestellten Methoden nicht vollständig lösen lässt) sowie die eingeschränkte Abbildbarkeit von Komorbiditäten. Außerdem können bei komplexen Krankheitsbildern, die durch viele unterschiedliche Leistungen behandelt werden können, die Definition relevanter Leistungen und die Kostenkalkulation sehr aufwendig oder gar unmöglich sein.

Die Zuschlüsselung von Komorbiditäten ist mit der expertengestützten Methode bei komplexen Krankheitsbildern kaum möglich. Daher werden bei nur wenigen Studien, die auf diesem Verfahren basieren, Komorbiditäten abgebildet. Hier kann sich ein Kontrollgruppenvergleich anbieten, der im Idealfall die Kosten von den mit der Zielerkrankung assoziierten Komorbiditäten genau zu dem Anteil abbildet, in dem sie den normalen Anteil in der nicht an der Zielerkrankung leidenden Durchschnittsbevölkerung übersteigen.

Kontrollgruppenvergleich

Beim Kontrollgruppenansatz werden die gesamten Leistungsausgaben von Patienten, die an der Zielerkrankung leiden, mit denen einer geeigneten Kontrollgruppe ohne diese Zielerkrankung verglichen. Die indikationsspezifischen Kosten ergeben sich bei diesem inkrementellen Ansatz rechnerisch aus der Differenz der jeweiligen Gesamtkosten (Holle et al. 2005). Beim Kontrollgruppenansatz werden unterschiedliche Matching-Verhältnisse eingesetzt. Das bedeutet, dass den einzelnen Patienten mit der Zielerkrankung eine zu definierende Anzahl an Kontrollgruppenpartnern zugeordnet werden kann. Üblich ist hier ein 1:1- oder 1:3-Matching (Zeidler et al. 2013), d. h. jedem Patienten mit der Zielerkrankung stehen im Verhältnis 1 bzw. 3 Kontrollgruppenmitglieder gegenüber. Teilweise wird auch ein 1:2- oder 1:5-Matching sowie eine deutlich größere Kontrollgruppe verwendet. Als übliche Matchingvariablen wer-

den beispielsweise das Alter, Geschlecht, die Versicherungsart, der Wohnort, der Erwerbsstatus, die ethnische Zugehörigkeit, Vorjahreskosten, Komorbiditätsscores oder ausgewählte Komorbiditäten verwendet. Eine internationale Übersicht und ein konkretes Beispiel zu den krankheitsspezifischen Kosten der Aufmerksamkeitsdefizit- / Hyperaktivitätsstörung (ADHS) findet sich bei Zeidler et al. (2013).

Ein zentraler Vorteil des Kontrollgruppenvergleichs ist die Möglichkeit, Komorbiditäten automatisch zu erfassen und abzubilden. Dies ist darauf zurückzuführen, dass die inkrementellen Kosten alle Kosten der Komorbiditäten sowie Komplikationen der untersuchten Zielkrankheit mit einschließen (Holle et al. 2005). Ist die geeignete Kontrollgruppe erst einmal definiert und extrahiert, kann der Aufwand der Kostenkalkulation als sehr überschaubar bezeichnet werden, da nur noch das Inkrement zwischen den Gesamtkosten der Zielpopulation und der Kontrollgruppe berechnet werden muss. Ein Nachteil des Kontrollgruppenansatzes liegt jedoch in dem schwierigen Zugang zu einer Kontrollgruppe, da diese in der Regel durch die Krankenkasse identifiziert und extrahiert werden muss. Aufgrund der begrenzten personellen Kapazitäten ist daher die Extraktion einer Kontrollgruppe nicht immer möglich, da das Matching aus Datenschutzgründen bei der Krankenkasse durchgeführt wird und dort entsprechende Kapazitäten einplant werden müssen. Darüber hinaus ist der Grundsatz der Datensparsamkeit zu beachten (§ 3a BDSG). Außerdem sind unter Umständen nicht alle für eine exakte Adjustierung erforderlichen Variablen in GKV-Routinedaten erfasst. Zur Vermeidung von Verzerrungen ist jedoch die Berücksichtigung aller kritischen Unterschiede zwischen der Zielpopulation und der Kontrollgruppe zwingend erforderlich. Als weitere Einschränkung des Kontrollgruppenvergleichs ist seine begrenzte Eignung für kleine Stichproben und seltene Krankheiten zu nennen, da für valide Ergebnisse eine ausreichend große Stichprobe zur Verfügung stehen muss.

Regressionsverfahren

Als weiteres Verfahren zur Identifikation indikationsspezifischer Ressourcenverbräuche können Regressionsmethoden eingesetzt werden. Bei einer Regression wird der Zusammenhang zwischen einer abhängigen Variable (hier: Kosten) und einer oder mehreren unabhängigen Variablen (hier: Indikatorvariable mit den Ausprägungen „an der Zielerkrankung erkrankt“ und „nicht an der Zielerkrankung erkrankt“) ermittelt. Neben einer ausreichenden Anzahl von Versicherten, die an der Zielerkrankung leiden, ist daher auch immer ein Datensatz mit Patienten erforderlich, welche die Er-

krankung nicht haben. Mit diesem Verfahren können diejenigen Kosten, die auf die Zielerkrankung zurückzuführen sind, unter Berücksichtigung aller relevanten Einflussfaktoren, wie z. B. Alter, Geschlecht und Komorbiditäten, berechnet werden. Bei der Nutzung von Regressionsverfahren kommen häufig Generalisierte Lineare Modelle (GLM) zum Einsatz (Zeidler et al. 2013). Diese werden teilweise in ein zweistufiges Verfahren integriert, um den spezifischen Limitationen von Kostendaten gerecht zu werden (schiefe Verteilung, Nullkostenfälle sowie Verletzung der Homoskedastizitätsannahme). Dabei wird zunächst auf Basis einer logistischen Regression ermittelt, welche Personen Kosten größer null haben. Anschließend wird dann die eigentliche Regression zur Ermittlung der indikationsspezifischen Ressourcenverbräuche durchgeführt. Mit der GLM-Regression können klassische, oben genannte Limitationen der Methode der kleinsten Quadrate (OLS) vermieden werden, die aufgrund der spezifischen Eigenschaften von Kostendaten auch nach einer Transformation (z. B. mittels Smearing-Estimation) zu nicht effizienten Schätzern führen kann. Als Herausforderung kann jedoch bei GLM-Regressionen die Ermittlung einer angemessenen Link-Funktion genannt werden.

Weitere Verfahren

Weder der Kontrollgruppenvergleich noch die Regressionsverfahren können den Einfluss von Ungleichheiten bei unbeobachtbaren Variablen adjustieren (Zeidler et al. 2013). Hierfür muss mit Instrumenten-Variablen oder einem Differenz-von-Differenzen-Ansatz gearbeitet werden. Darüber hinaus wird der Vorher-Nachher-Vergleich, auch als Prä-/Post-Vergleich bezeichnet, bei GKV-Routinedatenanalysen eingesetzt. Bei diesem Verfahren stellt die Zielpopulation ihre eigene Kontrollgruppe dar, und die Kosten vor und nach dem erstmaligen Auftreten der Zielerkrankung werden gemessen sowie verglichen. Der Vorteil dieses Studiendesigns liegt in dem vergleichsweise geringen Kalkulationsaufwand, der leichten Verständlichkeit der Ergebnisse sowie der Datensparsamkeit. Als Nachteil dieses Verfahrens kann die Kritik genannt werden, dass jeweils vor und nach den Messungen verschiedene zeitbezogene Effekte und Veränderungen, welche die Outcomevariable beeinflussen, auftreten können. Als Beispiel kann das fortschreitende Alter der beobachteten Personen im Zeitverlauf genannt werden. Außerdem ist bei diesem Verfahren eine Begrenzung auf inzidente Fälle erforderlich, was nicht bei jeder Krankheitskostenanalyse zielführend ist.

Als weiterer methodischer Ansatz zur Identifikation indikationsspezifischer Ressourcenverbräuche kann der Vergleich mit standardisierten Vergleichswerten genannt werden. So könnten dem Risikostrukturausgleich relevante Referenzwerte entnommen und mit den eigenen Daten verglichen werden. Diese Methode wurde beispielsweise durch Bowles et al. zur Berechnung der Kosten von angeborenen Neuralrohrdefekten genutzt (Bowles et al. 2014). Vorteilhaft an dieser Vorgehensweise ist die öffentliche Verfügbarkeit und Transparenz der Referenzwerte. Nachteile können sich jedoch durch methodische Veränderungen bei der Ermittlung der Referenzwerte im Zeitablauf sowie durch eine unzureichende Berücksichtigung regionaler und kassenspezifischer Besonderheiten ergeben.

In der Forschungspraxis werden sowohl die expertengestützte Methode als auch der Kontrollgruppenvergleich sowie Regressionsansätze regelmäßig eingesetzt (Zeidler et al. 2013). Der Unterschied zwischen den Ergebnissen kann je nach der gewählten Methode erheblich sein. Bei vielen Studien werden daher auch die verschiedenen Methoden miteinander kombiniert.

So ist beispielsweise bei bestimmten Leistungen eine Kombination der expertengestützten Methode mit dem Kontrollgruppenansatz möglich. Im Rahmen eines zweistufigen Verfahrens könnten zunächst relevante Leistungen ausgewählt, z. B. alle zur Behandlung der Zielerkrankung eingesetzten Arzneimittel, und anschließend speziell für diese Leistungen die Differenz der Kosten zwischen der Zielpopulation und der Kontrollgruppe gebildet werden.

Empfehlungen

- Bei der Berechnung direkter Kostenkomponenten ist zu prüfen, ob ein Erstattungsanspruch gegenüber der GKV besteht
- Zur Approximation von indirekten Kosten sind geeignete Verfahren einzusetzen
- Die eingeschränkte Abbildbarkeit intangibler Kosten ist bei der Studienplanung zu berücksichtigen
- Die Studienperspektive ist eindeutig zu definieren
- Bei der Wahl der Studienform sind die Vor- und Nachteile von Primär- und Sekundärdaten im Hinblick auf die Abbildbarkeit der Zielgrößen zu prüfen
- Unabhängig von der gewählten Methode ist die Vergleichbarkeit der zu evaluierenden Patientengruppen sicherzustellen
- Bei krankheitsspezifischen Analysen sollten die indikationsspezifischen Kosten berechnet werden, um einer Überschätzung der Behandlungskosten zu vermeiden
- Für die Berechnung indikationsspezifischer Kosten sind geeignete Verfahren einzusetzen und deren Vor- und Nachteile im Kontext der Forschungsfrage abzuwiegen

3.2 Regionale Auswertungen mit GKV-Routinedaten

GKV-Routinedaten bieten aufgrund ihrer umfangreichen Datenbasis auch die Möglichkeit von regionalen Auswertungen. Insbesondere aufgrund der häufig diskutierten Unterschiede in der Gesundheitsversorgung von städtischen und ländlichen Gebieten wird der Bedarf für derartige Analysen in Zukunft weiter zunehmen und an Relevanz gewinnen. So können beispielsweise Fragen zur regionalen Versorgungsqualität auf Grundlage von evidenzbasierten Versorgungleitlinien ermittelt und die regionale Epidemiologie verglichen werden. Die GKV-Routinedaten können insbesondere auch Hinweise auf Determinanten der Inanspruchnahme auf Angebotsebene geben, d.h. beispielsweise Indizien für angebotsinduzierte Nachfrage im stationären Sektor liefern. Dadurch ergeben sich Ansatzpunkte für eine gezielte Strukturentwicklung und zur Verringerung von Über-, Unter- und Fehlversorgung (Swart 2005b).

Für die Auswertung existieren in den GKV-Routinedaten verschiedene Variablen, die für eine regionale Differenzierung genutzt werden können. Häufig stehen explizite Informationen zum Wohnort der Versicherten aus datenschutzgründen allerdings nur in sehr grober Form (z. B. Bundesland) in GKV-Routinedaten zur Verfügung (bezüglich der datenschutzrechtlichen Aspekte und Zugangswege vergleiche Kapitel 2.2 und 2.3). Grundlegende Auswertungen auf Bundeslandebene oder zwischen Ost- und Westdeutschland sind damit zwar möglich, detailliertere Betrachtungen allerdings regelmäßig nicht. Darüber hinaus sind Informationen zur Postleitzahl des Wohnortes der Versicherten aufgrund von datenschutzrechtlichen Bestimmungen beschnitten und häufig nur drei- oder vierstellig verfügbar (siehe Kapitel 2.4.1). Krankenkassen besitzen in ihren Daten allerdings sowohl für niedergelassene Ärzte und Krankenhäuser sowie Rehabilitationseinrichtungen als auch für Versicherte Angaben zu ihrer individuellen Kreiskennziffer, anhand derer sie regional verortet werden können.

Die Kreiskennziffern teilen das Bundesgebiet zum jetzigen Zeitpunkt (Stand 05.2013) in 402 Landkreise (295) und kreisfreie Städte (107) ein. Für diese Ebene liegt weiterhin eine Reihe von Informationen (z. B. Bildung, Einkommen, Umwelt usw.) von Seiten des Bundesinstituts für Bau-, Stadt- und Raumforschung (BBSR) und des Statistischen Bundesamtes (Einwohnerdichte) vor (Bundesinstitut für Bau-, Stadt- und Raumforschung (BBSR) 2011a). In diesem Zusammenhang stellt sich die Frage, welche zusätzlichen Informationen mit Hilfe von Kreiskennziffern genutzt werden können. Möglich ist beispielsweise, dass Daten der KVen und/oder der Kassenärztliche Bundesvereinigung (KBV) genutzt werden, um pro Kreis die Anzahl und Dichte von bestimmten Arztgruppen zu ermitteln. Anschließend wäre es dann möglich, Aussagen über die spezifische Versorgungssituation im jeweiligen Kreis der Versicherten zu treffen. Häufig kann es auch sinnvoll sein, Kreiskennziffern weiter zu aggregieren, um diese besser zu Analysezwecken zu nutzen. Durch Aggregation lassen sich mit Hilfe von Informationen des BBSR, z. B. Aussagen darüber treffen, ob es sich um ländliche oder städtische Kreise handelt. Detailliertere Abgrenzungen bietet hingegen der siedlungsstrukturelle Kreistyp, der Kreise anhand von dem Bevölkerungsanteil in Groß- und Mittelstädten und der Einwohnerdichte in die vier Gruppen (1) kreisfreie Großstädte, (2) städtische Kreise, (3) ländliche Kreise mit Verdichtungsansätzen und (4) dünn besiedelte ländliche Kreise untergliedert. Geeignet scheint auch die Systematik „Raumtypen 2010:Lage“ des BBSR ((Bundesinstitut für Bau-, Stadt- und

Raumforschung (BBSR) 2011b). Die Kreise werden hierbei anhand eines Zentralitäts-Indexes in Abhängigkeit von der Nähe zu Konzentrationen von Bevölkerung und Arbeitsplätzen, die sich durch ein Angebot an Beschäftigungsmöglichkeiten und Versorgungseinrichtungen auszeichnen, in die vier Lagetypen einteilt: (1) sehr peripher, (2) peripher, (3) zentral und (4) sehr zentral. Mit Hilfe der zuvor genannten Kategorisierungen wäre es denkbar, Versicherte anhand der jeweiligen regionalen Besonderheiten und der Nähe zu medizinischen Versorgungseinrichtungen zu untergliedern. Fragen zur regionalen Versorgungsqualität wären somit unter Umständen hochwertiger und spezifischer zu beantworten.

Zu berücksichtigen ist dabei, dass aufgrund von Kreisreformen in den letzten Jahren eine Reihe von Änderungen vorgenommen und Kreise z. B. zusammengelegt wurden. Bei Analysen mit Daten aus mehreren Jahren ist also besondere Sorgfalt geboten, um keine falschen Schlüsse zu ziehen. Auch beachtet werden muss, dass die individuelle Mobilität von Versicherten damit nicht berücksichtigt werden kann und die Kreiskennziffer anhand des Wohnortes der Versicherten festgelegt wird. Darüber hinaus scheinen Auswertungen mit einer regionalen Differenzierung nur dann richtig sinnvoll, wenn eine ausreichend große Anzahl an Versicherten in verschiedenen Kreisen in den jeweiligen GKV-Routinedaten zur Verfügung steht.

Publizierte Artikel für regionale Analysen mit deutschen GKV-Routinedaten finden sich bisher nur vereinzelt. So zeigen Swart et al., wie kleinräumige Analysen im stationären Bereich anhand von 4-stelligen Postleitzahlen durchgeführt werden können (Swart et al. 2008). Melchior et al. untersuchen hingegen mit Hilfe der Kreiskennziffern regionalen Unterschiede in der Behandlung und Diagnostik von Depressionen (Melchior et al. 2014). Im Rahmen eines Methodenworkshops der AGENS wurden weiterhin verschiedene Projekte mit einem regionalen Fokus präsentiert (Zentralinstitut für die kassenärztliche Versorgung in der Bundesrepublik Deutschland 2013). Hierzu zählen Analysen zu Unterschieden in der Prävalenz und Pharmakotherapie von Multipler Sklerose und zur Feststellung regionaler Besonderheiten bei der Bedarfsplanung. Darüber hinaus werden methodische Aspekte der kleinräumigen Versorgungsforschung durch Swart diskutiert (Swart 2005a).

Empfehlungen

- Datenschutzrechtliche Aspekte sind bei der Auswahl geeigneter Regionalisierungsparameter zu berücksichtigen
- Änderungen aufgrund von Kreisreformen sind insbesondere bei mehrjährigen Analysen zu prüfen
- Es ist zu prüfen, ob die Stichprobengröße in den jeweiligen Kreisen ausreichend ist um valide Ergebnisse zu erhalten

3.3 Ereigniszeitanalysen mit GKV-Routinedaten

Eine zentrale Zielgröße bei vielen medizinischen Fragestellungen ist die Zeit bis zum Auftreten eines bestimmten Ereignisses (Zwiener et al. 2011). Bei klinischen Studien im onkologischen Bereich wird beispielsweise die Zeit von der Erstdiagnose bis zum Tod gemessen. Daher werden solche Analysen häufig als Überlebenszeitanalysen bezeichnet. Prinzipiell lässt sich mit den gleichen methodischen Ansätzen, wozu beispielsweise das Kaplan-Meier-Verfahren und die Cox-Regressionen gehören, auch die Zeit von der ersten Fraktur bis zu einer möglichen Folgefraktur oder die Zeit ab dem Behandlungsbeginn bis zum Therapieerfolg analysieren. Allgemeine methodische Aspekte der Überlebenszeitanalyse werden vertiefend durch Ziegler und Doblhammer (2009) diskutiert.

Im Rahmen von klinischen Studien beruhen diese Analysen auf Primärdaten. Grundsätzlich können aber auch GKV-Routinedaten für sogenannte Ereigniszeitanalysen genutzt werden. Hierzu findet sich bereits eine Reihe von Beispielen in der Literatur. L'hoest und Marschall untersuchen z. B. mithilfe von Daten der Barmer GEK den Einfluss von der Größe des Transplantationszentrums auf die Überlebenszeiten von Patienten nach einer Transplantation (L'hoest und Marschall 2013). Die Überlebensraten infolge von verschiedenen Rehabilitationsmaßnahmen im Alter vergleichen Meinck et al. mithilfe von GKV-Routinedaten der AOK (Meinck et al. 2014). Hendricks et al. evaluieren ein Case-Management-Programm für Patienten mit chronischer Herzinsuffizienz hinsichtlich Mortalität, Krankenhauseinweisungen und -kosten (Hendricks et al. 2014).

Zentrale Größen einer Ereigniszeitanalyse stellen regelmäßig die beiden relevanten Zeitpunkte – Anfangs- und Endzeitpunkt – dar. Anhand dieser beiden Zeitpunkte wird

die individuelle ereignisfreie Zeitspanne ermittelt. Tritt das Ereignis im Untersuchungszeitraum nicht ein, werden die jeweiligen Beobachtungen als zensierte Daten mit in die Analysen aufgenommen. Prinzipiell sind aber auch Analysen möglich, bei denen ein Ereignis mehrmals auftreten kann, z. B. wiederkehrende Lungenentzündungen bei Kindern oder multiple rezidivierende Frakturen. Der Anfangszeitpunkt kann z. B. den Zeitpunkt einer Operation (OPS), den Zeitpunkt der Entlassung aus dem Krankenhaus oder einer Diagnose (ICD) widerspiegeln. Der Zeitpunkt sollte dabei möglichst eindeutig und genau feststellbar sein. Diese Forderung gilt umso mehr, wenn die erwarteten Analysezeiträume sehr kurz sind, wie es z. B. bei Überlebenszeitanalysen von bestimmten onkologischen Erkrankungen der Fall ist. Probleme treten dann auf, wenn der Zeitpunkt nur sehr ungenau erfasst ist. Dies kann beispielsweise der Fall sein, wenn ein bestimmter Krankenhausaufenthalt als Startzeitpunkt gewählt wurde und zwischen Aufnahme- und Entlassungsdatum mehrere Wochen liegen. Noch problematischer sind ICD-Diagnosen aus dem ambulanten Sektor, da dort aufgrund der quartalsweisen Abrechnung standardmäßig kein konkretes Datum zugeordnet werden kann (Lösungsmöglichkeiten siehe Kapitel 3.6). Ähnliche Probleme können sich auch bei der Definition des Endzeitpunktes ergeben. Wie zuvor bereits erwähnt, kann jeder sinnvolle Zeitpunkt gewählt werden, der eindeutig und genau ist. Bei tatsächlichen Überlebenszeitanalysen ist selbstverständlich der Zeitpunkt des Todes aus den Daten zu ermitteln. Dieser ist aufgrund der obligatorischen amtlichen Todesmeldung vollständig und valide in den GKV-Routinedaten erfasst (WIdO 2007). Hierfür findet sich daher in der Regel eine separate Variable in den Daten. Zu beachten ist allerdings, dass teilweise auch das Ende des Versicherungszeitraumes – was nicht gleichbedeutend mit dem Versterben ist – in derselben Variablen verzeichnet ist und eine weitere Variable den jeweiligen Grund des Versicherungsendes codiert. Die GKV-Routinedaten enthalten jedoch keine Informationen über die Todesursache. Es ist daher häufig nur sehr schwer nachzuvollziehen, welche Erkrankungen oder Gründe einen Einfluss auf das Versterben hatten. Erste Ansätze zur indirekten Ermittlung von Todesursachen unter Zuhilfenahme der umfangreichen Informationen von Routinedaten der Krankenkassen existieren bereits, diese zielen allerdings auf Krankheiten mit einer hohen Letalität ab (Ohlmeier et al. 2012). Prinzipiell besteht auch die Möglichkeit einer Verlinkung mit anderen Daten, z. B. aus Krebsregistern, um die Todesursachen zu klären. Derartige Verknüpfungen erfordern

allerdings umfangreiche datenschutzrechtliche Abklärungen und Zugangsvoraussetzungen.

Im Rahmen von Überlebenszeitanalysen mit sehr kurzen Analysezeiträumen wird mitunter nicht die explizite Zeit bis zu einem Ereignis analysiert, sondern nur, ob das Ereignis (z. B. Tod) überhaupt eintritt. Im engeren Sinne sollte hierbei allerdings nicht mehr von „Überlebenszeit“ gesprochen werden. Heller et al. untersuchen so z. B. das Sterblichkeitsrisiko von Neugeborenen mit sehr niedrigem Geburtsgewicht anhand von Krankenhausabrechnungsdaten und analysieren diese anhand einer logistischen Regression (Heller et al. 2007).

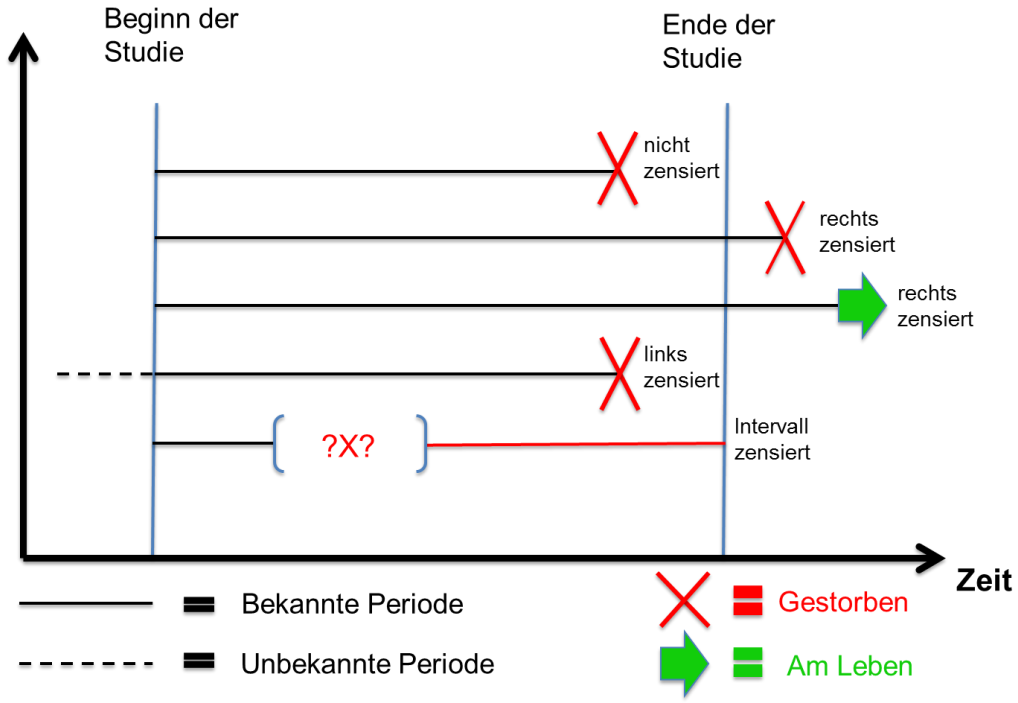
Empfehlungen

- Bei Ereigniszeitanalysen müssen Start- und Endzeitpunkt möglichst zweifelsfrei und genau definierbar sein
- Bei Überlebenszeitanalyse ist darauf zu achten, dass i.d.R. kein Todesgrund in den Daten verzeichnet ist

3.4 Die Bedeutung zensierter Daten

GKV-Routinedaten können z. B. durch Tod, Krankenkassenaustritt bzw. -wechsel oder durch das Ende des Studienzeitraums zensiert sein. Die methodische Herausforderung hierbei liegt darin, dass nach solchen Ereignissen keine Möglichkeit der Nachbeobachtung einzelner Personen besteht (Grobe und Ihle 2005). Dies stellt eine rechtsseitige Zensierung dar, da die Versicherten nicht bis zum Ende des Studienzeitraumes beobachtet werden können. Dieser – in epidemiologischen Studien – sogenannte „Lost to follow up“ muss bei der Auswertung berücksichtigt werden. Weiterhin existieren auch linksseitig zensierte Daten. Von linkszensierten Daten wird gesprochen, wenn das Ereignis zu einem unbekannten Zeitpunkt in der Vergangenheit eingetreten ist oder für die Versicherten keine Daten vorliegen, da sie erst neu in den Datensatz aufgenommen werden (z. B. Neugeborene und Neuversicherte). Zensierte Daten werden auch trunkiert, gestutzt oder (englisch) censored, truncated data genannt.

Abbildung 4: Mögliche Szenarien für zensierte Daten



Quelle: eigene Darstellung

Verglichen mit einer Primärdatenstudie existiert bei GKV-Routinedatenanalysen kaum die Möglichkeit, mittels entsprechender Studiendesigns diesen Effekten entgegenzuwirken. Dies gilt insbesondere für Längsschnittanalysen, die hierdurch verzerrt werden können. Bei kurzen Beobachtungsintervallen bestehen bezüglich dieser Effekte hingegen lediglich geringe Probleme.

Eine Möglichkeit einer Zensierung, insbesondere bei Längsschnittanalysen, zumindest partiell entgegenzuwirken, ist es nur diejenigen Patienten einzuschließen, die kontinuierlich innerhalb des Studienzeitraums bei der entsprechenden Krankenkasse versichert waren. Dieses in der Vergangenheit häufig verwendete Einschlusskriterium garantiert die zeitliche Konstanz der Versichertenzeiten und einen einheitlichen Beobachtungszeitraum. Nachteil dieser Vorgehensweise ist jedoch der Ausschluss von Kassenwechslern oder Verstorbenen. Dies kann je nach Fragestellung zu einer Über- oder Unterschätzung bzw. Verzerrung der Ergebnisse führen. Kassenwechsler sind häufig besonders junge und gesunde Versicherte, die z. B. nur geringe Kosten für die Krankenversicherung verursachen (Zok 2011). Wird ein Teil dieser Versicherten aufgrund der Selektionskriterien aber ausgeschlossen, führt das zu einer Überschätzung der durchschnittlichen Kosten pro Versicherten. Diese Problematik gilt es daher bei der Gestaltung des Studiendesigns zu berücksichtigen.

Wissenschaftler verzichten daher immer häufiger auf das Einschlusskriterium einer durchgängigen Versicherung. Vielmehr wird zunehmend ein Ansatz verwendet, bei dem die Ergebnisse auf Basis von Versicherungszeiten bzw. Ereignishäufigkeiten je Versicherungsjahr ausgewiesen werden. Dieses aus der Epidemiologie stammende Vorgehen gleicht die Unterschiede in der Populationsgröße zu unterschiedlichen Zeitpunkten aus und ist personenbezogenen Angaben vorzuziehen (Grobe und Ihle 2005).

Empfehlungen

- Bei jeder Analyse sollte geprüft werden, ob eine Zensierung vorliegt
- Im Falle einer Zensierung sind entsprechende Methoden zur Vermeidung von Verzerrungen einzusetzen
- Alternativen zum klassischen Kriterium der durchgängigen Versicherung sind zu prüfen

3.5 Compliance- und Persistence-Messung

Eine weitere Analysemöglichkeit unter Zuhilfenahme der GKV-Routinedaten ist die Analyse von Verschreibungsmustern im Arzneimittelbereich. Besonders relevant sind in diesem Zusammenhang Analysen, die die Compliance der Patienten beschreiben. Unter „Compliance“ wird im englischen Sprachgebrauch das konsequente Einhalten der ärztlichen Anweisung verstanden (Cramer et al. 2008). Im deutschen Sprachgebrauch wird synonym der Begriff „Therapietreue“ verwendet. Darüber hinaus wird häufig auch der Begriff „Adherence“ genutzt. Über die genaue Abgrenzung zwischen den Begriffen Compliance und Adherence gibt es konträre Meinungen, häufig werden diese beiden Begriffe jedoch synonym eingesetzt (Cramer et al. 2008). Compliance beinhaltet nicht nur die Einnahme von Medikamenten, sondern auch Änderungen des Lebensstils und andere Maßnahmen. Die Auslegung der Arzt-Patienten-Beziehung ist hierbei von Bedeutung. Mit „Non-Compliance“ wird ein abweichendes Verhalten des Patienten in Hinblick auf die therapeutischen Anweisungen des Arztes verstanden (Pirk und Schöffski 2012). Grundsätzlich können laut Pirk und Schöffski bei einer Primärdatenerhebung folgende Abweichungen auftreten:

- Unterdosierung (eingenommene Menge ist niedriger als die therapeutisch empfohlene bzw. allgemein übliche Dosis),

- Überdosierung (eingenommene Menge ist höher als die therapeutisch empfohlene bzw. allgemein übliche Dosis),
- zu kurze Einnahmedauer,
- zu lange Einnahmedauer,
- nicht zeitgerechte Einnahme,
- bedarfsweise Einnahme (und nicht nach Verordnung),
- unregelmäßige Einnahme sowie
- vollständig unterlassene Einnahme.

Anhand der GKV-Routinedaten kann die Einnahme der Medikamente durch den Patienten nicht unmittelbar analysiert werden. Dennoch existieren methodische Vorgehensweisen, um mittels der Arzneimittelverschreibungen das Einnahmeverhalten approximativ zu erklären. In den ambulanten Arzneimittelverordnungen liegen genaue Informationen zum Wirkstoff, zur Packungsgröße sowie zum Datum der Rezeptausstellung sowie zur Rezepteinlösung vor. Mithilfe dieser Informationen können der Zeitraum der Einnahme sowie Lücken in der medikamentösen Behandlung untersucht werden.

Die Compliance wird häufig als Prozentzahl ausgedrückt und in Form der sogenannten „medication possession ratio“ (MPR) berechnet. Dabei wird die Anzahl der verordneten Dosierungen in Relation zur geplanten Therapiedauer gesetzt. Die Anzahl der Tage einer Versorgung mit dem Medikament errechnet sich dabei in der Regel anhand einer definierten Tagesdosis (engl.: Defined Daily Dose, DDD). Unter der Annahme, dass die verordneten Tagesdosen durch den Patienten auch tatsächlich in vollem Umfang eingenommen wurden, kann dann im Sinne der MPR die Anzahl der verordneten Dosierungen in Relation zur geplanten Therapiedauer approximiert werden. Hat ein Patient beispielsweise innerhalb eines einjährigen Therapiezeitraums insgesamt 250 DDD verordnet bekommen, so beträgt seine MPR 0,69 bzw. 69 %. Häufig wird von einer „guten“ Compliance gesprochen, wenn im Beobachtungszeitraum mindestens 80 % der geplanten Dosierung eingenommen wurden. Abweichende Definitionen, wie beispielsweise 50 % oder 75 %, sind möglich (Cramer et al. 2008). Teilweise werden von den möglichen Therapietagen noch diejenigen Tage abgezogen, welche der Patient im Krankenhaus verbracht hat. Mit diesem Ansatz soll eine Unterschätzung der Compliance durch nicht abgebildete stationäre Verordnungen vermieden werden (siehe Kapitel 6 Limitationen). Unter Berücksichtigung der

Krankenhausaufenthalte kann die Formel zur Berechnung der MPR für ein Therapiejahr anhand von DDD folgendermaßen aussehen:

$$MPR = \frac{\textit{Anzahl Tagesdosen der Zielmedikation (DDD)}}{\textit{365 Tage – Anzahl Tage im Krankenhaus}}$$

Neben der Compliance wird bei Arzneimittelstudien häufig auch die Persistence analysiert. Die Persistence ist definiert als die Dauer zwischen dem Therapiebeginn und dem Abbruch der Therapie mit einer Zielmedikation (Cramer et al. 2008). Der Beobachtungszeitraum wird dabei in der Regel eindeutig definiert, sodass für jeden Patienten individuell ermittelt werden kann, ob es innerhalb dieser Periode zu einem Abbruch oder einer Unterbrechung der Therapie gekommen ist und wie viele Tage der Pateinten mit dem Medikament vorsorgt worden ist. So kann beispielsweise auf der Grundlage eines Indexereignisses (Verschreibung der ersten Medikation oder erste Dokumentation einer Indikation) mithilfe der DDD berechnet werden, wie lange die jeweilige Verschreibung dem Patienten zur Einnahme hypothetisch zur Verfügung steht. Ein Medikamentenabbruch kann dann als das erste Auftreten einer Lücke von mehr als einer vorab definierten Anzahl an Tagen bezeichnet werden. Das heißt, wenn zwischen Abgabedatum addiert mit der tagesbezogenen, theoretischen Einnahmedauer und der Folgeverschreibung eine Lücke von mehr als der vorab definierten Anzahl an Tagen vorliegt, wird von einem Abbruch gesprochen. Die Definition der zulässigen Lücke zwischen aufeinanderfolgenden Verordnungen hängt von den Dosierungsvorgaben des Medikamentes und dem konkreten Krankheitsbild ab. Je nach Indikation und Fragestellung ist daher individuell über die Länge der zulässigen Lücke bzw. Medikamentenabstinenz zu diskutieren. Darüber hinaus sollten Sensitivitätsanalysen bezüglich der Variation der zulässigen Lücke durchgeführt werden, um den Effekt unterschiedlicher Lückendefinitionen analysieren zu können. Auch die Berücksichtigung von stationären Krankenhausaufenthalten ist analog zur Compliance-Berechnung möglich.

Da es sich bei den DDD um hypothetische mittlere Tagesdosen handelt, kann die empfohlene oder verschriebene Dosis des jeweiligen Arztes anders ausfallen. Welche ärztliche Intention einer Leistungsverordnung zugrunde lag bzw. wie die konkrete Applikation durch den Patienten erfolgte, geht aus den GKV-Routinedaten nicht hervor (Zeidler und Braun 2012). Des Weiteren könnte auch die Änderung in der Thera-

pie aufgrund von Nebenwirkungen ein Grund für einen Abbruch sein. Die Gründe für einen Therapieabbruch oder eine Therapieunterbrechung lassen sich mit GKV-Routinedaten häufig nicht abbilden.

Sowohl für die Berechnung der Compliance als auch der Persistence existieren viele unterschiedliche Methoden, welche sich konzeptionell in spezifischen Details unterscheiden. Frey und Stargardt haben zehn unterschiedliche Methoden systematisch verglichen und deren Prognosequalität in Hinblick auf die Hospitalisierung von Schizophreniepatienten analysiert (Frey und Stargardt 2012). Eine umfassende Analyse von insgesamt 216 unterschiedlichen Studiendesigns zur Bestimmung von Non-Adherence bei Typ 2 Diabetes Mellitus findet sich bei Wilke et al. (2013).

Empfehlungen

- Es ist bei der Studienkonzeption zu beachten, dass die Bestimmung von Compliance bzw. Persistence mit GKV-Routinedaten nur approximativ möglich ist
- Je nach Indikationsgebiet und Applikationsform sind geeignete Methoden einzusetzen und Sensitivitätsanalysen durchzuführen
- Es müssen bei der Interpretation der Ergebnisse die vielfältigen Gründe für Therapieabbrüche oder Dosierungsänderungen berücksichtigt werden, welche sich jedoch nur begrenzt aus GKV-Routinedaten kausal ableiten lassen

3.6 Überprüfbarkeit von Leitlinienempfehlungen

Leitlinien spiegeln die aktuelle wissenschaftliche sowie medizinische Evidenz wider und spielen in der Medizin eine zentrale Rolle. Sie sollen den behandelnden Ärzten Handlungsempfehlungen geben und bieten auch bei der juristischen Beurteilung von Komplikationen nach einer medizinischen Behandlung Orientierung. Durch die Potenziale zur Abbildung des Versorgungsalltages ist die Analyse der Leitlinienadhärenz in den letzten Jahren zunehmend zu einem Anwendungsgebiet von GKV-Routinedaten geworden.

Um die Leitlinienkonformität zu überprüfen, müssen jedoch unterschiedliche Voraussetzungen bezüglich der GKV-Routinedaten, des Indikationsgebiets und der Leitlinien erfüllt sein. So können manche Empfehlungen gut und umfassend abgebildet

werden und andere lediglich unter gewissen Annahmen oder Limitationen. Leilinienempfehlungen, die gar nicht über GKV-Routinedaten abbildbar sind, existieren ebenfalls.

Im Bereich der Datengrundlage ist es wichtig, dass die Indikation mittels ICD-10-Codierung abbildbar sein muss, da sonst die Identifizierung der Zielpopulation anhand der GKV-Routinedaten schwierig bzw. sogar unmöglich ist. Die zu untersuchenden Leitlinienempfehlungen müssen mittels EBM-Ziffern, ATC- und/oder OPS-Codes sowie ICD-10-Codierung abbildbar und speziell für das Indikationsgebiet beschrieben sein. Da in den GKV-Routinedaten keine klinischen Parameter erfasst sind, sollten die zu überprüfenden Handlungsempfehlungen darüber hinaus nicht auf entsprechenden Informationen, beispielsweise zur Krankheitsschwere, basieren.

Eine hohe Inzidenz der Erkrankung sowie eine gewisse Größe der kooperierenden Krankenkasse ermöglicht eine ausreichend große Stichprobe, um die breite Versorgungspraxis mithilfe von GKV-Routinedaten abzubilden. Konkrete zeitliche und mengenmäßige Angaben wie z. B. „jährlich“ sind erforderlich, um die Empfehlungen mithilfe der GKV-Routinedaten abbilden zu können. Angaben wie „regelmäßig“ oder „stabile Patienten“ sind zu unpräzise und nicht ausreichend definiert, um sie mit GKV-Routinedaten abzubilden.

Auch der Schweregrad der Erkrankung sollte als Unterkategorie oder mittels eigener ICD-10-Codierung abbildbar sein. Diese Forderung ist jedoch nur in seltenen Fällen erfüllt. So lassen sich Informationen zum Schweregrad von Diabetes und Herzinsuffizienz in der ICD-10-Systematik finden (Eberhard 2013). Bei vielen Krankheitsbildern fehlen allerdings solche Systematiken, die sich mithilfe von GKV-Routinedaten analysieren lassen. Unterschiedliche Behandlungsmaßnahmen sollten für unterschiedliche Schweregrade vorliegen und beschrieben sein. Die Ermittlung der Anzahl der Arztbesuche sowie die Reihenfolge der unterschiedlichen Arztbesuche (z. B. Hausarzt/ Facharzt) innerhalb eines Quartals sind lediglich unter bestimmten Annahmen möglich. Ein Beispiel hierfür ist eine Verknüpfung aller Leistungsdaten (ambulant, Arzneimittel, AU-Bescheinigungen etc.). Grund hierfür ist, dass in diesen Leistungssektoren datumsgenaue Angaben zu finden sind, die dann dem jeweiligen Arztbesuch zugeschlüsselt werden können.

Die Pharmakotherapie lässt sich mithilfe der GKV-Routinedaten sehr gut abbilden, da die ATC-Codes, die DDD etc. meist vollständig in den GKV-Routinedaten vorliegen. Auch Kontraindikationen können durch eine ICD-10-Codierung in den ambulanten und stationären Diagnosedaten aufgedeckt und abgebildet werden. Die Medikationsmuster sollten abhängig vom Schwergrad sowie von Kontraindikationen formuliert und mithilfe entsprechender ATC-Klassifikationssysteme darstellbar sein. Dosierungsempfehlungen pro Körpergewicht können nur annahmebasiert bzw. gar nicht abgebildet werden. Impfschutzempfehlungen in den Leitlinien können nur teilweise abgebildet werden, da KV-spezifische Sonderziffern für Impfungen existieren. Um Impfschutzempfehlungen zu überprüfen, müssen zudem die Impfintervalle in der Leitlinie explizit angegeben werden.

Für Subgruppenanalysen sollten die Empfehlungen der Leitlinie nach unterschiedlichen Patientengruppen (Schweregrad der Krankheit, Kinder/Erwachsene etc.) unterteilt sein. Familienversicherte, z. B. Kinder oder Ehepartner, können in den GKV-Routinedaten dem Mitglied zugeschlüsselt werden. Voraussetzung ist jedoch, dass beide bei derselben Krankenkasse versichert sind. Da andere Familienbeziehungen außerhalb der Familienversicherung nicht abbildbar sind, kann auch die familiäre Disposition nicht nachvollzogen werden. Auch Vorerkrankungen sind nur teilweise erkennbar, da die GKV-Routinedaten meist lediglich für einen fünf-Jahreszeitraum zugänglich sind. Dennoch kann ein Indexereignis oder eine sogenannte Baseline definiert werden, mit dessen/deren Hilfe analysiert wird, welche weiteren Erkrankungen vor dem Indexereignis oder in der Baseline vorlagen. Definiert ist das Indexereignis als erstmaliges Auftreten eines Events bzw. Ereignisses im Beobachtungszeitraum (beispielsweise erstmals dokumentierte ICD-10-Diagnose oder Arzneimittelverschreibung). Ultsch et al. entwickelten für die Indikation Herpes zoster einen Algorithmus zur Identifikation des initialen Diagnosedatums. Das Datum der initialen Diagnose wurde als das früheste Datum definiert, das sich mit folgendem Schema ermitteln ließ (Ultsch et al. 2013):

- Verschreibungsdatum eines indikationsspezifischen Arzneimittels;
- Datum einer Hospitalisierung aufgrund der Zielindikation (Aufnahmedatum);
- Datum des Beginns einer zielindikationsspezifischen Arbeitsunfähigkeit;

- Datum des ambulanten Arztkontaktes, wenn es sich hierbei um die einzige Leistungsinanspruchnahme oder einzige Diagnose im Quartal der initialen Diagnose handelt;
- Datum des ersten ambulanten Kontaktes innerhalb des Quartals der initialen Diagnose mit dem Arzt, der die Diagnose zuerst dokumentiert hat, falls keine der zuvor erwähnten vier Bedingungen zutrifft.

Dieser innovative Ansatz könnte auch für andere GKV-Routinedatenstudien bezüglich der jeweiligen Indikation angepasst werden, auch wenn diese Systematik ursprünglich für die Indikation Herpes zoster entwickelt wurde. Häufig dient dieses Indexdatum dazu für alle Versicherten einen einheitlichen Nachbeobachtungszeitraum sicherzustellen. Die Baseline wird definiert als eine Vorlaufzeit, die entweder als Ausgangspunkt gesehen wird, um beispielsweise das Alter zu messen, oder als ein symptomfreier Zeitraum, um beispielweise einen inzidenten Patienten abbilden zu können.

Nicht darstellbar sind beispielsweise krankheitsspezifische Präventionsangebote. Ziel dieser Maßnahmen ist es, die ausgewählte Erkrankung gar nicht erst auftreten zu lassen. Somit existiert auch keine ICD-10-Diagnose für den Aufgriff und die Identifikation der Studienpopulation. Auch die Kommunikation zwischen Arzt und Patient sowie die Beratung, die häufig eine große Bedeutung in den Leitlinien einnimmt, ist anhand von GKV-Routinedaten nicht zu veranschaulichen. Ähnliches gilt für die Mitarbeit des Patienten. Die Empfehlung eines Arztes, z. B. mehr Sport zu treiben, besitzt keine spezifische Abrechnungsziffer und ist somit auch nicht analysierbar. Gleiches gilt für Empfehlungen zur Ernährung und zum Gewicht der Versicherten. Die DMP-Dokumentation schließt diese Lücke etwas, da hier teilweise diese Merkmale (z. B. Raucherstatus) erfasst werden; dennoch sind diese derzeit noch schlecht dokumentiert. Trotzdem sind Empfehlungen, z. B. in Bezug auf eine Ernährungsumstellung oder körperliches Training, nicht in den GKV-Routinedaten darstellbar.

Bislang existieren nur wenige Studien, die sich mit der Überprüfung der Anwendung und Anwendbarkeit von Leitlinien in der Versorgungspraxis auseinandergesetzt haben. So untersuchten Swart und Willer die Arthrose-Leitlinien verschiedener Fachgesellschaften. Sie stellten fest, dass soweit die untersuchten Leitlinien sich anhand von GKV-Routinedaten operationalisieren lassen, ihnen weitgehend gefolgt wird. Dennoch sahen sie Herausforderungen in der Nutzung von Routinedaten der Kran-

kenkassen, da wesentliche Elemente der Leitlinien nicht abgebildet werden konnten, die Beobachtungszeiträume häufig kurz waren oder die Validität der Abrechnungsdaten partiell für die Beantwortung der Fragestellung unzureichend war (Swart und Willer 2012). Laux et al. betrachteten die Qualitätsindikatoren der Nationalen Versorgungsleitlinie (NVL) für chronische Herzinsuffizienz; andere Empfehlungen fanden in ihrer Studie keine Berücksichtigung (Laux et al. 2011; NVL 2012). Auch sie sahen Herausforderungen in der Abbildbarkeit der Leitlinienkonformität, da für fünf von den insgesamt neun Indikatoren notwendige Informationen in den GKV-Routinedaten nicht enthalten waren (Laux et al. 2011). In einem Beitrag von Eberhard werden diagnostische und therapeutische Aspekte einer leitliniengerechten Versorgung von Patienten mit arterieller Hypertonie überprüft. Sie kommt zu dem Ergebnis, dass der Anteil und ob die Patienten leitlinienadäquat behandelt werden, pauschal nicht mit GKV-Routinedaten analysiert werden kann (Eberhard 2013).

Empfehlungen

- Vorab geprüft werden sollte, ob sich die Leitlinienempfehlungen anhand von GKV-Routinedaten valide operationalisieren lassen
    - Konkrete zeitliche und mengenmäßige Angaben sind erforderlich, um die Empfehlungen mithilfe der GKV-Routinedaten abbilden zu können
    - Alle Empfehlungen sollten mit Klassifikationssystemen (ICD, EBM, OPS, ATC etc.) abbildbar sein
    - „Weiche“ Faktoren, beispielsweise die Arzt-Patienten.-Kommunikation, sind häufig nicht nachvollziehbar
    - Krankheitsspezifische Präventionsangebote sind nicht darstellbar
- Es muss berücksichtigt werden, dass in der Regel für eine umfassende Überprüfung von Empfehlungen unterschiedliche Schweregrade abbildbar sein müssen

4 Datenextraktion und Validierung

Die GPS fordert eine begleitende Qualitätssicherung als unabdingbaren Bestandteil jeder Sekundärdatenanalyse (AGENS 2012). Dies ist aufgrund des Sekundärdatencharakters erforderlich, da auf Daten zurückgegriffen wird, die primär zu einem anderen Zweck und von anderen Personen erhoben wurden. Auf die primäre Datenerhebung und die Qualität der Dokumentation hat der Sekundärdatennutzer somit keinen Einfluss. Gerade bei den vertragsärztlichen Diagnosen kommt es laut Bundesversicherungsamt zu zahlreichen inkonsistenten Diagnosestellungen (IGES Institut GmbH 03.12.2012). Zur Qualitätssicherung sind daher unter anderem Validierungsverfahren einzusetzen, um die Vorhersagequalität zu optimieren. Für die Überprüfung der Validität von GKV-Routinedaten existieren unterschiedliche Verfahren. Bei der internen Validierung wird die Konsistenz anhand des vorliegenden Datensatzes geprüft. Im Rahmen der externen Validierung erfolgt die Überprüfung hingegen anhand eines externen Goldstandards (Hoffmann et al. 2008). Im Folgenden werden unterschiedliche Validierungstechniken vorgestellt und ihre jeweiligen Vor- und Nachteile diskutiert.

4.1 Datenextraktion und Aufgreifkriterien

Für eine systematische und zielgerichtete Datenextraktion existieren verschiedene Aufgreifkriterien. Für den zweckmäßigen Datenaufgriff ist es wichtig, einen geeigneten Selektionsalgorithmus zu wählen. Hierbei ist es entscheidend, die Kriterien so zu gestalten, dass möglichst alle relevanten Fälle eingeschlossen und gleichzeitig dabei unzureichend gesicherte oder nicht korrekte Fälle ausgeschlossen werden (Hoffmann und Glaeske 2011). Auch der Grundsatz der Datensparsamkeit sollte bei dem Datenaufgriff berücksichtigt werden, d. h., dass bei der Erhebung, Verarbeitung und Nutzung personenbezogener Daten so wenig wie möglich personenbezogene Daten extrahiert werden sollten (§ 3a BDSG). Im Folgenden werden einige ausgewählte Aufgreifalgorithmen dargestellt.

Falls eine Indikation präzise mittels ICD-Codes abbildbar ist, könnten diese Codes allein genutzt werden, um relevante Fälle bzw. Patienten zu identifizieren. Hierzu muss die jeweilige Indikation bzw. der jeweilige ICD-Code vorgelegen haben und in dem relevanten Zeitraum abgerechnet worden sein. Zu diskutieren ist an dieser Stelle die Diagnosesicherheit. Wie bereits erwähnt, existieren sowohl im ambulanten als

auch im stationären Leistungsbereich unterschiedliche Kennzeichen in der Diagnosestellung. Im stationären Bereich wird zwischen Aufnahmediagnose, Einweisungsdiagnose, Entlassungsdiagnose, Verlegungsdiagnose, Hauptdiagnose und Nebendiagnose unterschieden. In GKV-Routinedatenstudien werden üblicherweise die Hauptdiagnose und die Nebendiagnosen als Aufgreifkriterium verwendet, da diese die validesten Diagnosen darstellen und auch abrechnungsrelevanten Nutzen haben. Im ambulanten Versorgungssektor erhalten die ICD-Codes für diese Behandlungsdiagnose das Zusatzkennzeichen A, G, V und Z (siehe Kapitel 2.4.2). Häufig wird im ambulanten Bereich die gesicherte Diagnose als Aufgreifkriterium gewählt. Dennoch kann je nach Fragestellung die Verdachtsdiagnose oder die „Zustand nach"-Diagnose als zusätzlicher Selektionsparameter dienen. Den Aufgriff lediglich durch eine V- oder Z-Diagnose vorzunehmen, ist eher unüblich. Dennoch können diese Qualitätsschlüssel für die Subgruppenanalyse genutzt und beispielsweise bei Kosten-Vergleichsanalysen oder Vorher-Nachher-Vergleichen berücksichtigt werden. Gerade im ambulanten Leistungsbereich wird zuweilen der Vorwurf erhoben, dass ein ICD-10-Code das klinische Krankheitsbild, wie es sich vor allem dem ambulant tätigen Arzt darstellt, nur ungenügend beschreibt (Schubert et al. 2010). Um sicherzustellen, dass die zu analysierende Indikation auch tatsächlich vorliegt, kann auch ein mehrfaches Auftreten einer ambulanten Diagnose gefordert werden. Je nach Indikation und Fragestellung muss – wie auch bei den Morbi-RSA gefordert – eine ambulante Diagnose im Folgequartal bestätigt werden (DIMDI 2013b).

Oftmals empfiehlt es sich, eine ICD-Diagnose mit einem krankheitsspezifischen Arzneimittel zu verknüpfen. In diesem Fall werden Versicherte in die Studienpopulation aufgenommen, wenn für diese die jeweilige ICD-Diagnose und zusätzlich nach der ATC-Klassifikation codierte Arzneimittelverordnung abgerechnet wurde bzw. vorlag (GKV-Spitzenverband 2012). Auch die Verknüpfung mit indikationsbezogenen OPS oder anderen Leistungsbereichen kann als Aufgreifkriterium genutzt werden.

Eine weitere methodische Frage zur Identifizierung der Studienpopulation ist die Frage nach dem Alter der Patienten. Für manche Fragestellungen kann es sinnvoll sein, z. B. Kinder und Jugendliche von der Analyse auszuschließen, da für diese Subgruppe andere Behandlungsempfehlungen vorliegen.

Auch die Bedingung einer durchgängigen Versicherung wird bei vielen GKV-Routinedatenanalysen zugrunde gelegt. Hierbei werden lediglich Individuen betrachtet, die

im gesamten Studienzeitraum durchgängig versichert waren. Dies kann zu einer Unterschätzung der tatsächlichen Anzahl bzw. der tatsächlichen Kosten führen, da beispielsweise Verstorbene unberücksichtigt bleiben. Es existieren jedoch unterschiedliche statistische Methoden, beispielsweise die Ereigniszeitanalysen (siehe Kapitel 3.3), die auch zensierte Daten berücksichtigen.

Grundsätzlich sind, sowohl im Studienplan als auch während der Analysen, die Aufgreifkriterien und das Extraktionskonzept schriftlich zu fixieren.

Empfehlungen

- Der Aufgreifalgorithmus muss definiert und vor der Studie festgelegt werden. Mögliche Kriterien beim Aufgriff sind:
  - Lediglich der ICD-Code; hier Kennzeichen in der Diagnosestellung beachten
  - Kombination aus ICD-Diagnose und krankheitsspezifischen Arzneimitteln
  - Kombination aus ICD-Diagnose und krankheitsspezifischen Prozeduren
- Es ist zu prüfen welche Altersklassen mit einbezogen werden
- Die durchgängige Versicherung der Studienpopulation ist zu diskutieren
- Das Extraktionskonzept ist schriftlich zu fixieren
- Der Grundsatz der Datensparsamkeit ist zu beachten

4.2 Vollständigkeit

Nach der Datenlieferung bzw. -extraktion ist zunächst die Vollständigkeit der Daten zu prüfen, um die Qualität der GKV-Routinedaten beurteilen und überprüfen zu können. Im ersten Schritt sollte daher untersucht werden, ob alle für die Studie relevanten Variablen aus allen Leistungsbereichen übermittelt wurden. Liegen Informationen zu allen relevanten Variablen vor, muss geprüft werden, ob die datenliefernde Institution alle erforderlichen Informationen vollständig erfasst hat oder ob es zu Unterbrechungen im Datenfluss gekommen ist (Hoffmann et al. 2008). Die Suche nach auffälligen Mustern oder Schwankungen im Zeitablauf kann dabei wichtige Hinweise auf Inkonsistenzen liefern. So können beispielsweise Krankenhauseinweisungen, Verschreibungen, Diagnosen oder OPS-Codes im Zeitablauf (z. B. tages-, monats- oder

quartalsbezogen) dargestellt und anhand einer grafischen Aufbereitung mögliche Unterbrechungen im Datenfluss identifiziert werden. Auffällige Schwankungen können dann einen Hinweis auf Inkonsistenzen im Datenfluss geben. Hier kann sich im Bereich der ambulanten Diagnosen auch eine nach KVen stratifizierte Analyse anbieten, um Lücken im Datenfluss bei einzelnen KVen zu identifizieren (Hoffmann et al. 2008). Eine Herausforderung stellt bei dieser Vorgehensweise jedoch die Abgrenzung zwischen natürlichen (saisonalen) Schwankungen wie beispielsweise dem „Dezemberknick“, d. h. einer geringen Zahl an Hospitalisierungen zu den Weihnachtsfeiertagen sowie zum Jahreswechsel und tatsächlichen Datenlücken dar. Auch können Schwankungen im Arzneimittelbereich auf nachträgliche Ergänzungen zurückzuführen sein, da die Apothekenabrechnungszentren und Krankenkassen fehlende oder fehlerhafte Datumsangaben häufig auf den 5., 15. oder 25. Tag eines Monats (Hoffmann et al. 2008) oder den letzten Tag eines Monats setzen.

Liegen für ein Projekt Daten mehrerer Krankenkassen vor, können datumsbezogene Ereignisraten im Zeitablauf dargestellt und miteinander verglichen werden (Hoffmann et al. 2008). Hoffmann et al. haben ein Verfahren vorgeschlagen, bei dem zunächst die Anzahl an Versicherten jeder liefernden Krankenkasse zeitbezogen bestimmt wird, z. B. pro Monat, und in Relation zu den Ereignisraten gesetzt wird (Hoffmann et al. 2008). Dann können beispielsweise die monatlichen Hospitalisierungsraten je 1.000 Versichertenmonate der verschiedenen Krankenkassen miteinander verglichen und auffällige Abweichungen identifiziert werden. Auch hier wird eine grafische Aufbereitung der Ergebnisse empfohlen. Mit diesem vergleichenden Verfahren können Unterbrechungen im Datenfluss von saisonalen Schwankungen abgegrenzt werden, da sich saisonale und extern bedingte Ausschläge bei allen Krankenkassen zeigen dürften (Hoffmann et al. 2008).

Bei der Vollständigkeitsprüfung sollte darüber hinaus überprüft werden, ob und wie häufig bei einzelnen Variablen leere Datenfelder enthalten sind. Dies kann durchaus vorkommen, beispielsweise ist in den Arzneimitteldaten nicht zu jeder Verordnung auch ein ATC-Code angegeben. Dies ist beispielsweise darauf zurückzuführen, dass zu einzelnen Arzneimitteln gar kein ATC-Code existiert, wie dies z. B. bei den in der Apotheke individuell angefertigten Zytostatika der Fall ist. Eine ungewöhnliche Häufung leerer Felder kann jedoch einen Hinweis auf eine Unterbrechung im Datenfluss geben.

Eine Verknüpfung verschiedener Datenbereiche kann ebenfalls ein sinnvoller Bestandteil der Vollständigkeitsprüfung sein. So ist beispielsweise zu prüfen, ob auch zu jedem Patienten, der Leistungen in Anspruch genommen hat, entsprechende Stammdaten geliefert wurden. Wenn in der Arzneimitteldatenbank beispielsweise Individuen mit einer entsprechenden Arzneimittelverordnung aufgeführt sind, zu deren Pseudonymen sich jedoch keine Stammdaten finden lassen, kann dies ein Hinweis auf Unvollständigkeit sein. Hier ist dann unter Rücksprache mit dem Dateneigner zu klären, ob diese Personen bei der Extraktion der Stammdaten übersehen wurden oder ob es sich um Personen handelt, bei denen die Krankenkasse in Vorleistung gegangen ist und daher kein eindeutiges Pseudonym existiert. Dies kann der Fall sein, wenn ein Versicherter lediglich pflege-, renten- und arbeitslosenversichert ist oder unter das Bundessozialhilfegesetz (BSHG 1999) fällt. Bei diesen Personen geht die Krankenkasse zunächst in Vorleistung, d. h. die Person verursacht Kosten und erscheint als Abrechnungs- bzw. Kostenfall in den Versorgungs- und Leistungsbereichen, diese Kosten werden jedoch rückwirkend zurückerstattet. Diese Fälle müssen von weiteren Analysen ausgeschlossen werden, da sie nicht bei der Krankenkasse versichert sind und daher keine expliziten Stammdaten vorliegen.

Empfehlungen

- Nach der Datenlieferung muss überprüft werden, ob:
  - Alle relevanten Variablen aus allen Leistungsbereichen übermittelt wurden
  - Die erforderlichen Informationen vollständig erfasst wurden oder Unterbrechungen im Datenfluss existieren
  - Auffällige Muster oder Schwankungen im Zeitablauf vorhanden sind
  - zu jedem Patienten, der Leistungen in Anspruch genommen hat, auch entsprechende Stammdaten geliefert wurden
  - Und wie häufig bei einzelnen Variablen leere Datenfelder existieren

4.3 Interne Diagnosevalidierung

Einen Schritt weiter als die Vollständigkeitsprüfung geht die interne Validierung. Bei der internen Validierung wird die Konsistenz anhand des vorliegenden Datensatzes geprüft. Ein besonders wichtiger Bestandteil der internen Validierung ist die interne

Diagnosevalidierung. Da ICD-Diagnosen das klinische Krankheitsbild, insbesondere bei Symptomen und Beschwerden mit einem unspezifischen oder multifaktoriellen sowie psychosomatischen Hintergrund, teilweise nur unzureichend beschreiben, stellt sich bei jeder Studie die Frage, ob die Codierung valide ist (Schubert et al. 2010). Ziel der Diagnosevalidierung ist daher die Bestätigung einer Diagnose anhand weiterer Charakteristika aus den GKV-Routinedaten (Garbe 2008). Dabei soll eine Abgrenzung zwischen sicheren und unsicheren Diagnosen, d. h. zwischen Verdachts-/Ausschlussdiagnosen und gesicherten Diagnosen, vorgenommen werden (Hoffmann et al. 2008). Außerdem soll zwischen akut auftretenden und bereits länger zurückliegenden historischen Ereignissen differenziert werden. Die Abgrenzung zwischen akuten und historischen Ereignissen ist bei Krankenhausdiagnosen in der Regel jedoch nicht notwendig, da diese stets akute Erkrankungen erfassen. Im ambulanten Bereich ist dies hingegen erforderlich, da beispielsweise die Praxissoftware Einfluss auf die Codierweise nehmen kann (Schubert et al. 2010). Die Identifikation von gesicherten Diagnosen wird seit dem 01.01.2004 erleichtert, da seit diesem Zeitpunkt die Diagnosesicherheit verpflichtend in den ambulanten Daten differenziert wird (Hoffmann et al. 2008) (siehe auch Kapitel 2.4.2). In einer Untersuchung von Hoffmann et al. zeigte sich, dass im Jahr 2006 der Anteil gesicherter Diagnosen bei rund 90 % lag und sich nur in 1,6 bis 3,8 % keine Angaben zur Diagnosesicherheit finden ließen (Hoffmann et al. 2008). Die Anteile der ausgeschlossenen Diagnosen, Verdachtsdiagnosen bzw. „Zustand nach“-Diagnosen lagen bei jeweils rund 3 %.

Für die interne Validierung existiert bisher kein festgelegter Standard, die Vorgehensweise ist im Hinblick auf den jeweiligen Auswertungsinhalt festzulegen (Schubert et al. 2010). Beispielsweise können Arzneimittelinformationen, ärztliche ambulante Leistungen (EBM), verordnete Sachleistungen (z. B. Heil- und Hilfsmittel) und Prozeduren im Krankenhaus (OPS) für die Sicherung einer Diagnose verwendet werden (Schubert et al. 2010). In der Literatur finden sich verschiedene Beispiele zur internen Diagnosevalidierung. Hoffmann et al. stellen ein Verfahren vor, bei dem das akute venöse thromboembolische Ereignis, d. h. die tiefe Beinvenenthrombose bzw. Lungenembolie, anhand von Coumarin-Verschreibungen, Informationen zum Versterben sowie Krankenhauseinweisungen validiert wird (Hoffmann et al. 2008). Als weiteres Beispiel kann die Hinzuziehung von Antidiabetikaverordnungen bei der Identifizierung von Diabetespatienten genannt werden.

Der bisher umfassendste methodische Vorschlag zur internen Diagnosevalidierung bei chronischen Erkrankungen findet sich bei Schubert et al. (2010). Hierbei wurden Kriterien für die Einschätzung der Validität bei den drei Erkrankungen Herzinsuffizienz, Demenz und Tuberkulose zur Prävalenzschätzung entwickelt. Bei diesem Verfahren werden zunächst alle ICD-Diagnosen zusammengetragen, welche die Zielerkrankung vollständig beschreiben. Danach werden alle Versicherten selektiert, bei denen mindestens eine der relevanten ICD-Diagnosen als ambulante Diagnose und/ oder stationäre Diagnose (Aufnahme-, Entlassungs-, Haupt- und Nebendiagnosen) dokumentiert wurde. Im nächsten Schritt werden dann mithilfe eines Ausschlussverfahrens alle Patienten selektiert, die potenziell an der definierten Zielerkrankung leiden. Ausgeschlossen werden alle Patienten, die ausschließlich ambulante Diagnosen mit dem Zusatz „ausgeschlossene Diagnose“ oder „Verdachtsdiagnose“ haben. In einem weiteren Schritt wird untersucht, ob die relevante Diagnose ausschließlich im ambulanten oder stationären Sektor oder in beiden Sektoren vorgelegen hat. Dann wird für jeden Bereich geprüft, inwieweit vorab definierte Kriterien zur Diagnosesicherung erfüllt sind. Dabei können die im Folgenden näher beschriebenen Kriterien herangezogen werden.

Diagnose in mehreren Quartalen

Bei chronischen und schwerwiegenden Erkrankungen ist von einer regelmäßigen Wiederholung der Diagnosedokumentation im ambulanten und/oder stationären Sektor auszugehen. Wird die Diagnose nur einmal dokumentiert, kann bei chronischen und regelmäßig behandlungsbedürftigen Erkrankungen davon ausgegangen werden, dass es sich um eine Verdachtsdiagnose oder Fehldiagnose handelt.

Diagnose durch verschiedene Ärzte

Die Dokumentation einer Diagnose durch mehrere Ärzte oder Einrichtungen kann ebenfalls ein aussagekräftiges Kriterium zur Absicherung einer Diagnose darstellen. So kann eine Reihe unterschiedlicher Ursachen zu einer Dokumentation durch mehrere Ärzte führen (z. B. fortgeschrittenes Stadium der Erkrankung, Abklärung einer Erstdiagnose durch einen weiteren Arzt, eine stationäre Behandlung oder eine Urlaubsvertretung des Arztes).

Unterschiedlich differenzierte ICD-Diagnosen

Herausgeber von ICD-10-GM und OPS ist das DIMDI. Änderungen im Krankheitsspektrum und der medizinisch-technische Fortschritt werden durch jährliche Anpassungen der Klassifikationen berücksichtigt. Deswegen kann die Dokumentation verschiedener ICD-Diagnosen für die Zieldiagnose ebenfalls als Kriterium zur Validierung herangezogen werden. Unterschiedliche Spezifizierungen der ICD-Diagnose im Zeitablauf, z. B. durch Angaben zur Topografie, zur Stadieneinteilung oder zu Komplikationen, können die Validität der Diagnose erhärten. Dieses Kriterium erfüllen Patienten, für die sich unterschiedlich differenzierte Diagnosen innerhalb der Zieldiagnose finden (z. B. Diabetes und diabetesbedingte Komplikationen).

Medikation

Auch medikamentöse Verordnungen können zur Bestätigung einer Diagnose genutzt werden, sofern das Medikament für eine gut eingrenzbare, enge Indikation (z. B. Insulin) zugelassen ist. Bei chronischen Erkrankungen muss die Arzneimittelverordnung nicht zwangsläufig im Quartal der Diagnosestellung liegen. Bei weniger spezifischen Verordnungen kann eine Verknüpfung zwischen dem die Diagnose codierenden und dem die Verordnung ausstellenden Arzt, jeweils im selben Quartal, vorgenommen werden. Verordnungen anderer Ärzte, die eventuell zur Behandlung anderer Erkrankungen ausgestellt wurden, können mit diesem Verfahren ausgeschlossen werden und die Wahrscheinlichkeit, dass eine unspezifische Medikation mit der Zieldiagnose in Zusammenhang steht, kann erhöht werden.

Als weiteres Verfahren zur internen Validierung ist die Reproduktion eines bekannten Zusammenhangs möglich. Hoffmann et al. haben anhand der Verordnung eines Psychostimulans (Methylphenidat, Pemolin, Fenetyllin, Amphetamin-Rezepturen) bei Kindern und Jugendlichen auf die Diagnose F90 (Hyperkinetische Störung) geschlossen (Hoffmann et al. 2008). Anschließend wurde der Zusammenhang zwischen einer Hyperkinetischen Störung und Unfällen untersucht. Bei Kindern mit Stimulanzienverordnung war das relative Risiko, wegen Verletzungen oder Vergiftungen im Krankenhaus behandelt zu werden, signifikant erhöht.

Versterben im unmittelbaren zeitlichen Zusammenhang zur Diagnose

Bei Erkrankungen mit einer hohen Wahrscheinlichkeit zu versterben, ist keine Diagnosewiederholung zu erwarten. In diesen Fällen schlagen Schubert et al. folgende Definition zur Diagnosesicherung vor: a) der Patient verstirbt während eines Krankenhausaufenthaltes (Sterbedatum ist gleich dem Krankenhauseinweisungsdatum und die Aufnahme-, Entlassungs-, Haupt- oder Nebendiagnosen ist die Zieldiagnose), b) Patient mit Zieldiagnose verstirbt vier Wochen nach dem Krankenhausaufenthalt (Aufnahme-, Entlassungs-, Haupt- oder Nebendiagnosen ist die Zieldiagnose), c) bei einmaliger ambulanter Diagnose: Patient verstirbt im Diagnosequartal (Schubert et al. 2010). Bei Erkrankungen mit einer entsprechend hohen Mortalität kann das Versterben zur Diagnosesicherung herangezogen werden.

Weitere Kriterien

Je nach Forschungsfrage und Datengrundlage können weitere Kriterien verwendet werden. Hier seien beispielsweise EBM-Leistungen, Heil- und Hilfsmittelverordnungen oder OPS-Leistungen genannt, die im Zusammenhang mit der Zielerkrankung stehen.

Nach der Festlegung geeigneter Kriterien sind diese für einen genau definierten Zeitraum für jeden einzelnen Versicherten zur Diagnosesicherung zu prüfen. Zur Beurteilung der Validität wird zunächst eine getrennte Betrachtung nach dem ambulanten und stationären Sektor empfohlen, da hierdurch Unterschiede in der Codierweise zwischen den Sektoren deutlich werden. Wird eine Zusammenführung erforderlich, werden die stationären Diagnosen ebenfalls beispielsweise Quartalen zugeordnet, um einen einheitlichen Definitionszeitraum für ambulante und stationäre Diagnosen zu erhalten. Aufnahme- und Nebendiagnosen werden dabei dem Quartal der Aufnahme und Hauptentlassungsdiagnosen dem Quartal der Entlassung zugeordnet (Schubert et al. 2010). Eine Beschränkung des Validierungszeitraums auf den Prävalenzzeitraum, d. h. den Zeitraum für den eine Prävalenzschätzung vorgenommen werden soll, wird nicht empfohlen. Dies würde keine Beurteilung von Patienten erlauben, die im Beobachtungsjahr versterben, deren Erkrankung endet, bei denen die Diagnose erstmalig gestellt wird oder die unregelmäßig einen Arzt aufsuchen. Für die interne Diagnosevalidierung wird daher ein Zeitraum von drei Quartalen vor und drei Quartalen nach dem Prävalenzzeitraum empfohlen. Die Validierung erfolgt dann für

einen Patienten mit Zieldiagnose für jedes einzelne Diagnosequartal im Prävalenzzeitraum getrennt. Zur Validierung der Quartalsdiagnose wird ein Zeitraum von drei Quartalen vor und nach diesem Quartal herangezogen. Dann wird nacheinander für die vier sich ergebenden Zeitfenster geprüft, ob die relevanten Validierungskriterien erfüllt sind. Ist ein Kriterium in mindestens einem Zeitfenster erfüllt, so gilt das Kriterium für das beobachtete Diagnosequartal als bestätigt. Dieses Verfahren wird für jedes Quartal des Prävalenzzeitraumes wiederholt. Wird bei einem Patienten die Zieldiagnose in mindestens einem Quartal als valide eingestuft, so gilt die Bestätigung der Diagnose für den gesamten Prävalenzzeitraum.

Kritisch ist im Hinblick auf die interne Validierung anzumerken, dass die Diagnosen mit diesem Verfahren anhand von Charakteristika derselben Datenquelle überprüft werden (Hoffmann et al. 2008). Das Verfahren zur internen Validierung von Schubert et al. kann nicht identifizieren, ob eine Diagnose zurechtgestellt und der Patient richtig behandelt wurde; vielmehr kann nur die interne Konsistenz der Angaben geprüft werden (Schubert et al. 2010). Außerdem ist das Verfahren vom Inanspruchnahmeverhalten der Versicherten abhängig. Dies kann dazu führen, dass die Diagnosevalidität von Erkrankungen, die mit einer unregelmäßigen Inanspruchnahme verbunden sind, unterschätzt wird.

Empfehlungen

- Diagnosevalidität ist beispielsweise durch folgende Kriterien zu prüfen:
  - Diagnose in mehreren Quartalen
  - Diagnose durch verschiedene Ärzte
  - Unterschiedlich differenzierte ICD-Diagnosen
  - Indikationsabhängige Medikation
  - das Versterben bei Erkrankungen mit einer hohen Mortalität
- Es muss berücksichtigt werden, dass die Diagnosevalidität von der untersuchten Erkrankung abhängig ist
- Die Validierung sollte in enger Abstimmung mit Medizinern und Kassenvertretern stattfinden
- Der Zeitraum zur Diagnosesicherung muss definiert werden
- Empfohlen wird eine getrennte Betrachtung des ambulanten und stationären Sektors

4.4 Externe Validierung

Bei der externen Validierung erfolgt die Überprüfung anhand eines externen „Goldstandards“ (Hoffmann et al. 2008). Als Goldstandard kommt dabei beispielsweise die Patientenakte des Hausarztes, die Krankenhausakte oder auch eine Patientenbefragung in Betracht. Die externe Validierung von Diagnosen erfolgt meist anhand einer kleinen Stichprobe an Patienten, da sie mit einem erheblichen Aufwand verbunden ist. Auch eine Rezeptsichtung, d. h. ein Abgleich zwischen den Originalrezepten und den GKV-Routinedaten, kann wichtige Hinweise geben (Hoffmann et al. 2008). Die externe Validierung ist jedoch, wie erwähnt, mit einem erheblichen Aufwand verbunden und muss aus datenschutzgründen durch die Krankenkasse oder eine Vertrauensstelle durchgeführt werden. Aufgrund datenschutzrechtlicher Restriktionen ist eine externe Validierung vorhandener Diagnosen häufig nicht möglich. Umfassende externe Validierungsstudien sind für Deutschland nicht bekannt (Schubert et al. 2010). Insgesamt liegt in Deutschland daher wenig Wissen über die Diagnosevalidität vor. Die externe Validierung liefert jedoch den größten Zugewinn an Informationen zur Güte einer Diagnose (Hoffmann et al. 2008).

Empfehlungen

- Wenn möglich, sollte eine Validierung anhand externer Quellen durchgeführt werden

4.5 Plausibilität

Bei der Plausibilitätsprüfung sollen unlogische und falsche Informationen identifiziert werden. Hier kann beispielsweise überprüft werden, ob in der Datenbank falsche Datumsinformationen vorliegen. So wird bei zeitraumbezogenen Informationen wie der Verweildauer im Krankenhaus geprüft, ob zwischen dem Aufnahme- und Entlassungsdatum eine negative Differenz liegt, d. h. die Entlassung vor der Aufnahme erfolgte. Auch das Vorliegen negativer Kosten kann einen Hinweis auf Unstimmigkeiten geben. Dies muss jedoch nicht zwangsläufig der Fall sein, da negative Kosten auch auf nachträgliche Umbuchungen, Regresse oder Gutschriften zurückzuführen sein können. Hier muss im Einzelfall eine Abstimmung mit dem Dateneigner stattfinden, um die Plausibilität negativer Werte einschätzen zu können.

Bei der Plausibilitätsprüfung kann auch nach nicht plausiblen Altersangaben gesucht werden. Ein Alter von mehr als 124 Jahren kann einen eindeutigen Hinweis auf einen Erfassungsfehler geben. Kerek-Bodden et al. definieren Altersangaben von einem Tag bis 110 Jahre als plausibel (Kerek-Bodden et al. 2005). Altersangaben von mehr als 110 Jahren werden auf k. A. (keine Angabe) gesetzt. Auch bei anderen Variablen kann nach unrealistischen Ausreißern gesucht werden. Als Beispiele seinen unrealistisch lange oder kurze Liegezeiten im Krankenhaus oder auch Hochkostenfälle genannt.

Eine weitere Auffälligkeit stellen im Zeitablauf wechselnde Geschlechtsinformationen dar. Auch wenn dies in Einzelfällen durchaus vorkommen kann, sollten diese Fälle überprüft werden. Sonst kann es bei Subgruppenanalysen, beispielsweise zu Genderunterschieden, zu Zuordnungsproblemen in die jeweilige Subgruppe kommen. Um dieser Auffälligkeit entgegenzuwirken kann beispielsweise nach geschlechtsspezifischen Erkrankungen oder nach für einen Geschlechtswechsel typischen Arzneimittelverordnungen gesucht werden. Des Weiteren sind geschlechtsspezifische Diagnoseschlüssel zu prüfen. So gelten Daten als nicht plausibel, wenn für einen Mann die Diagnose O81 „Geburt“ codiert wird. Wie mit solchen Unplausibilitäten umzugehen ist, wird im Kapitel 5 erläutert.

Als weitere Möglichkeit zur Plausibilitätsprüfung kann die zeitliche Konstanz der Versichertenzeiten überprüft werden. Sofern beim Dateneigner ausschließlich Informationen zu durchgängig versicherten Personen angefordert wurden, stellen Unterbrechungen bei den Versichertenzeiten einen Hinweis auf Fehler dar.

Bei Kosteninformationen sind ebenfalls Plausibilitätsprüfungen möglich. Erstens kann ein zumindest stichprobenartiger Abgleich der erfassten Kosten mit öffentlich zugänglichen Gebührenordnungen durchgeführt werden. Größere Abweichungen können hierbei einen Hinweis auf Fehler geben. Zweitens sollte die Währungseinheit überprüft werden, d. h. ob es sich um Euro- oder Cent-Werte handelt.

Bei Längsschnittanalysen sind mögliche Änderungen der Datenerhebung und -erfassung sowie die Gültigkeitsdauer der Schlüssel zur Merkmalcodierung im zeitlichen Verlauf zu prüfen (Grobe 2005; Grobe und Ihle 2005). Dies ist erforderlich, da sich die verwendeten Merkmalcodierungen und Klassifikationssysteme im Zeitablauf ändern können. Dies ist unter anderem bei der ATC-Klassifikation der Fall, wo sich

beispielsweise die ATC-Codes der TNF-α-Hemmer Adalimumab, Etanercept und Infliximab im Jahr 2008 geändert haben. Der bis zum Jahr 2007 bei Adalimumab gültige ATC-Code L04AA17 wurde dabei auf den ATC-Code L04AB04 geändert. Derartige Änderungen sind bei der Studienplanung unbedingt zu beachten, da andernfalls eine lückenhafte Erfassung für die betroffenen Zeiträume droht. Zusätzlich können gesundheitspolitische Entscheidungen zu Trends und Sprungstellen bei den Leistungsdaten führen (Holle et al. 2005). Als Beispiel kann die vorgezogene Inanspruchnahme von Leistungen als Effekt einer Ankündigung von Zuzahlungserhöhungen, Leistungseinschränkungen oder Leistungsausschlüssen genannt werden. Zur Vermeidung von Fehlinterpretationen ist eine Berücksichtigung derartiger zeitlicher Effekte dringend zu empfehlen.

Die GPS fordert eine Überprüfung der Anonymisierung bzw. Pseudonymisierung, da Fehler bei diesem Prozessschritt den Rohdatensatz verändern und Auswertungsergebnisse verfälschen können (AGENS 2012). Dies kann beispielsweise durch eine stichprobenartige Prüfung und Sichtung der Ausgangsdatensätze sowie der anonymisierten bzw. pseudonymisierten Datensätze oder durch Pseudonymisierung von Testdaten geschehen.

Das Auftreten zeitlicher Besonderheiten sollte ebenfalls überprüft werden. Ein Beispiel wäre hier die Anzahl abgerechneter EBM im ambulanten Bereich. So sollte montags bis freitags eine höhere Anzahl an abgerechneten Leistungen anfallen als am Wochenende. Im Gegenzug sind dadurch, dass häufig die ambulanten Praxen am Wochenende geschlossen sind, die Krankenhauseinweisungen höher. Sollten große Peaks oder Einbrüche erkennbar sein, sind die Daten genauer zu überprüfen.

Empfehlungen

- Die Plausibilität der Daten muss anhand von zweckdienlichen Variablen überprüft werden. Beispiele hierfür sind:
  - falsche Datumsinformationen,
  - negative Kosten,
  - nicht plausible Altersangaben,
  - wechselnde Geschlechtsinformationen,
  - lange oder kurze Liegezeiten im Krankenhaus oder Hochkostenfälle,
  - Konstanz der Versichertenzeiten,
  - Merkmalcodierungen und Klassifikationssysteme,
  - Überprüfung der Anonymisierung bzw. Pseudonymisierung,
  - Auftreten zeitlicher Besonderheiten
- Änderungen der Datenerhebung und -erfassung sowie die Gültigkeitsdauer der Schlüssel zur Merkmalcodierung im zeitlichen Verlauf sind zu prüfen
- Eine enge Abstimmung mit dem Dateneigner ist erforderlich, um die Plausibilität unterschiedlicher Werte einschätzen zu können

5 Datenaufbereitung und -analyse

Nach der Prüfung der Vollständigkeit und Validität sind die Datenaufbereitung und -analyse die nächsten Schritte einer GKV-Routinedatenstudie. Die Datenaufbereitung dient dazu, die Abrechnungsdaten für wissenschaftliche Auswertungen nutzbar zu machen. Aufgrund des Sekundärdatencharakters sind häufig umfassende Aufbereitungsschritte erforderlich, um die GKV-Routinedaten im Hinblick auf die zu beantwortenden Fragestellungen auswerten zu können. Im Folgenden werden Empfehlungen gegeben, wie beispielsweise mit Ausreißern, Nullkosten und anderen Datenauffälligkeiten umgegangen werden kann. Jegliche Schritte der Datenaufbereitung müssen dokumentiert werden, sodass die methodische Vorgehensweise immer nachvollziehbar und transparent ist. Eine Möglichkeit zur systematischen Dokumentation ist die Erstellung eines Datenaufbereitungsprotokolls. In diesem Protokoll sollten die Anzahl und die Struktur der übermittelten Datensätze, das Erhebungs- und Lieferdatum, Codierungsänderungen und gegebenenfalls Referenzlisten schriftlich vermerkt sein (AGENS 2012).

5.1 Allgemeines Vorgehen

Die Reproduzierbarkeit der einzelnen Prozessschritte – beispielsweise die Aufbereitung und die Bereinigung der Daten – ist zu jedem Zeitpunkt sicherzustellen. Als Referenzgröße kann dabei auch der noch nicht modifizierte Originaldatensatz dienen. Hierfür wird empfohlen, vom Originaldatensatz eine Sicherheitskopie zu erstellen. Dieser Ausgangsdatensatz ist getrennt von den Auswertungsdatensätzen aufzubewahren und es sind die Vereinbarungen zu den Aufbewahrungsfristen des Datenschutzkonzeptes zu berücksichtigen. Während aller Auswertungs- und Aufbereitungsschritte sind Kontrollen hinsichtlich der Plausibilität durchzuführen. So muss z. B. die Anzahl der Datensätze bei einer Zusammenfügung mehrerer Tabellen der Summe der Einzeltabellen entsprechen. Genauer gesagt, ist darauf zu achten, dass beim Zusammenfügen kein Individuum unberücksichtigt bleibt sowie Einzeldaten wie beispielsweise einzelne Verordnungen bzw. Diagnosen etc. wegfallen.

Der Aufwand für die Aufbereitung der GKV-Routinedaten sollte nicht unterschätzt werden, insbesondere wenn Daten aus unterschiedlichen Leistungsbereichen verwendet werden. Bei der Auswertung größerer Datensätze sind Kontrollen von einzelnen Beobachtungen kaum möglich und anschließende individuellen Korrekturen ein-

zelner Beobachtungen nicht effizient umsetzbar. Selbst geringe Fehlerquoten von unter einem Prozent können, je nach Datenbankgröße, bereits mit mehreren Tausend Fehleinträgen verbunden sein. Daher müssen in der Regel Aufbereitungsroutinen, die relevante Fehler ohne Einzelfallprüfung automatisiert korrigieren bzw. entsprechende Merkmalsausprägungen löschen, programmiert werden (Grobe und Ihle 2005).

Empfehlungen

- Die Reproduzierbarkeit der einzelnen Prozessschritte ist zu jedem Zeitpunkt sicherzustellen
- Es wird empfohlen eine Sicherheitskopie des Originaldatensatzes zu erstellen
- Vereinbarungen des Datenschutzkonzeptes müssen bei der Datenaufbereitung berücksichtigt werden
- Der Aufwand für die Aufbereitung der GKV-Routinedaten sollte nicht unterschätzt werden und muss in die zeitliche Planung einkalkuliert werden

5.2 Datenauffälligkeiten

Bei GKV-Routinedatenanalysen werden häufig ganz unterschiedliche Datenauffälligkeiten im Rahmen der Validierung und der deskriptiven Analyse ersichtlich. Zur Identifikation relevanter Auffälligkeiten können verschiedene Instrumente eingesetzt werden. So sind bei einer überschaubaren Datenmenge beispielsweise Ausreißer schnell mittels eines Boxplots gesondert dargestellt bzw. identifiziert. Auch die Ermittlung von Minimum und Maximum der jeweiligen Variablen kann Aufschluss über Datenauffälligkeiten geben. Streudiagramme und Häufigkeitstabellen stellen eine weitere Möglichkeit zur Identifikation auffälliger Muster dar.

Eine enge Zusammenarbeit und Rücksprache mit dem Datenhalter ist hierbei sehr wichtig, um die Plausibilität auffälliger Muster im Kontext der Datenentstehung mit den zuständigen Fachabteilungen diskutieren zu können. Sind die Abweichung der Daten auf einen Datenfehler zurückzuführen, empfiehlt es sich, die Daten gegebenenfalls neu anzufordern. Fehler in den Daten können bei unterschiedlichen Vorgängen entstehen. Unterschieden werden können Fehler bei der eigentlichen Erfassung der Daten bei den Leistungserbringern, Fehler bei der Extraktion der Daten aufseiten

der Krankenkasse und Fehler bei der Formatierung bzw. beim Einlesen der Daten beim Forscher. Darüber hinaus gibt es Auffälligkeiten, die nur fälschlicherweise als Datenfehler interpretiert werden. Gegebenenfalls müssen auch das Aufgreifkriterium, die Zeiträume oder zusätzliche Variablen neu definiert werden.

Nach einer systematischen Identifizierung aller Datenauffälligkeiten existieren unterschiedliche methodische Möglichkeiten mit diesen umzugehen. Ein simples, aber wissenschaftlich recht zweifelhaftes Vorgehen ist, diese Auffälligkeiten zu ignorieren. Das heißt, die auffälligen Ausprägungen der Variablen werden in ihrem Rohzustand gelassen und nicht weiter aufbereitet. Der Vorteil dieser Herangehensweise ist die leichte Umsetzbarkeit und die minimale Manipulation der Originaldaten. Nachteilig ist jedoch, dass dieses Vorgehen womöglich zu Verzerrungen führen kann. Genauer gesagt, könnten z. B. aus einer deskriptiven Angabe von Minimum und Maximum falsche Schlüsse gezogen werden, wenn die Werte falsch oder unrealistisch sind. Beispielsweise könnte bei einer Kosten-Minimumangabe ein negativer Wert berichtet werden, der für den Leser erstmal als unrealistisch eingestuft bzw. nicht nachvollziehbar ist. Dieses Vorgehen ist daher in der Regel nicht zu empfehlen. Handelt es sich jedoch um redundante Datensätze, sind diese zu löschen. Dieser Ausschluss ist gut zu dokumentieren.

Eine weitere Möglichkeit, aber auch die drastischste Maßnahme mit Datenauffälligkeiten umzugehen, ist ein Ausschluss auffälliger Merkmalsausprägungen. Werden Unplausibilitäten bei der Validierung entdeckt, können diese Datensätze auch gelöscht werden. Dieses Vorgehen ist jedoch lediglich bei einer größeren Datenmenge zu empfehlen, da durch das Löschen relevante Informationen verloren gehen.

Wenn ein eigenständiges Korrigieren der Daten valide möglich ist, sollte dieses Vorgehen allen anderen oben genannten Möglichkeiten vorgezogen werden. Ein Beispiel hierfür ist die Ergänzung der Arzneimittelabrechnungsdaten durch den GKV-Arzneimittelindex des WIdO (WIdO). Wenn einzelne Informationen nicht vollständig bzw. fehlerhaft übermittelt wurden, kann diese Datenbank herangezogen werden. So kann anhand der PZN der dazugehörige ATC-Code ermittelt und fehlende Informationen in der Routinedatenbank ergänzt werden.

Auch eine Umcodierung der Daten kann hierbei sinnvoll sein. Ändern sich beispielsweise die Merkmalcodierungen und Klassifikationssysteme im Zeitablauf, so kann es

sinnvoll sein, diese dann umzucodieren. Als Beispiel ist die ATC-Klassifikation der TNF-α-Hemmer Adalimumab, Etanercept und Infliximab im Jahr 2008 noch einmal aufzugreifen. Um die Auswertungen zu vereinfachen, kann der veraltete ATC-Code durch den neuen ATC-Code ersetzt werden. Nachfolgend muss dann lediglich nach einem ATC-Code ausgewertet werden. Ein weiteres Beispiel für eine mit dieser Methode korrigierbare Datenauffälligkeit liegt im Bereich der Stammdaten. Kerek-Bodden et al. schlagen als plausible obere Grenze für das Alter 110 Jahre vor. Sie begründen diese Obergrenze damit, dass Menschen, die älter als 110 Jahre sind, extrem selten vorkommen (Kerek-Bodden et al. 2005). Zum einen könnte das Alter der Versicherten älter als 110 Jahre auf diese Grenze gesetzt werden. Eine andere Empfehlung wäre alle Versichertendaten bezüglich eines Alters von über 110 Jahren auf „keine Angabe“ zu setzen, um das mittlere Alter nicht zu verzerren (Kerek-Bodden et al. 2005).

Weiterhin ist es möglich, dass sich verwendete Merkmalcodierungen und Klassifikationssysteme im Zeitablauf ändern (siehe Kapitel 4.5). Um die Auswertung effizienter zu gestalten, können beispielsweise alte Ausprägungen durch neue ersetzt werden. Aber auch die Zusammenfassung von Variablenausprägungen zu aussagefähigen Gruppen, beispielsweise die Transformation des Alters in Altersgruppen, kann für eine spätere Auswertung nützlich sein.

Weiterhin können Angaben aus dem Sozialgesetzbuch oder anderen öffentlichen Quellen genutzt werden, um realistische Werte für verschiedene Variablen zu definieren. So kann laut SGB V eine Arbeitsunfähigkeitsbescheinigung „wegen derselben Krankheit für maximal 78 Wochen (546 Kalendertage) innerhalb von je drei Jahren ab Beginn der Arbeitsunfähigkeit“ ausgestellt werden (siehe § 48 Abs. 1 SGB V). AU-Zeiträume größer als diese 78 Wochen sind genauer zu untersuchen und gegebenenfalls unter Absprache mit dem Dateneigner zu löschen oder zu korrigieren.

Im Arzneimittelbereich existieren zwei Datumsangaben: das Verordnungs- und das Abgabedatum. Wird hier die Datumsdifferenz gebildet, kann es ebenfalls zu Auffälligkeiten kommen. Das sogenannte Kassenrezept ist grundsätzlich drei Monate lang gültig. Die jeweilige Krankenkasse bezahlt die verschriebenen Arzneimittel allerdings lediglich bei Einlösung innerhalb eines Monats. Danach, d. h. die übrigen zwei Monate, erhält der Versicherte die verschriebenen Medikamente zwar noch, er muss jedoch den vollen Preis selbst tragen. Rezepte für den Akne-Wirkstoff Vitamin-A-Säure

und damit verwandte Substanzen bilden eine Ausnahme bei der Gültigkeit und müssen innerhalb einer Woche eingelöst werden (Kelm 2012).

Bei Privatrezepten, die gesetzlich Versicherte für verschreibungspflichtige, aber nicht erstattungsfähige Arzneimittel erhalten, gilt ebenfalls eine dreimonatige Gültigkeit. Dadurch, dass der Versicherte den vollen Preis des verschriebenen Medikaments in der Apotheke selbst entrichtet, werden diese Arzneimittel jedoch nicht den GKV-Abrechnungsdaten zugeführt (Köster et al. 2011).

Weiterhin kann der Arzt Arzneimittel, die unter das Betäubungsmittelgesetz (BTMG 2013) fallen, verordnen. Hierzu zählen beispielsweise Drogenersatzstoffe wie Methadon, aber auch starke Schmerzmittel oder Medikamente gegen ADHS. Die entsprechenden BTM-Rezepte sind lediglich sieben Tage lang gültig, da sich bei Missbrauch gefährliche Wirkungen zeigen können.

Bei Auswertungen im Arzneimittelbereich ist weiterhin darauf zu achten, dass seit dem 01.01.2013 die Pharmazentralnummern von siebenstellig auf achtstellig umgestellt wurden. So wurde vor die siebenstellige PZN eine Null vorweggestellt. Diese wird je nach Codierung der Variablen, d. h. als String oder numerische Variable, mit angezeigt oder systembedingt gelöscht. Bei Auswertungen ist also die Ausprägung bzw. das Format der Variable zu beachten und die relevante PZN in der korrekten Schreibweise zu analysieren. Des Weiteren können veraltete PZN neu vergeben werden.

Ähnlich wie bei der Arbeitsunfähigkeit existieren auch bei der Länge der Rehabilitation gesetzliche Rahmenbedingungen. Nach § 40 SGB V, Abs. 3, Satz 2 sollen „Leistungen nach Absatz 1 für längstens 20 Behandlungstage, Leistungen nach Absatz 2 für längstens drei Wochen erbracht werden, es sei denn, eine Verlängerung der Leistung ist aus medizinischen Gründen dringend erforderlich.“ Problematisch ist hierbei allerdings, dass mögliche medizinische Gründe in den GKV-Routinedaten nicht nachvollzogen werden können.

Die Umcodierung und Bildung neuer Variablen muss vollständig dokumentiert werden (AGENS 2012). Weiterhin ist der Effekt der Anpassungen auf die jeweiligen Analyseergebnisse zu überprüfen. Quantifizieren lassen sich die Auswirkungen insbesondere durch Sensitivitätsanalysen. Eine Sensitivitätsanalyse sagt allgemein aus,

wie sehr Abwandelungen der Ausgangsbedingungen, in diesem Fall die GKV-Routinedaten bzw. Variablen, das Ergebnis beeinflussen, also wie sensitiv bzw. empfindlich ein System reagiert (Frank 1976).

Empfehlungen

- Mit Datenauffälligkeiten kann wie folgt umgegangen werden:
  - Ignorieren
  - Ausschluss von Datenfällen
  - Eigenständiges Korrigieren
  - Umcodieren
- Sofern eine valide Korrektur von Fehlern möglich ist, sollte dieses Verfahren den anderen Methoden vorgezogen werden
- Zur Abschätzung des Einflusses von Auffälligkeiten sollte eine enge Absprache mit dem Datenhalter stattfinden
- Die Verwendung von Sensitivitätsanalysen wird empfohlen
- Jegliche Korrekturen der Datenauffälligkeiten sind stets zu dokumentieren

5.2.1 Ausreißer

Ausreißer sind Extremwerte, die nicht in eine erwartete Messreihe passen oder allgemein nicht dem Streuungsbereich um den Erwartungswert entsprechen (Müller-Benedict 2007). Wenn Ausreißer vorhanden sind, muss geprüft werden, wie diese entstanden sind und ob es sich um reguläre Abweichungen oder um Datenfehler handelt. Ob Werte überhaupt als Ausreißer bezeichnet werden können, lässt sich anhand verschiedener statistischer Tests ermitteln. Einen guten Überblick geben Rousseeuw und Leroy (Rousseeuw und Leroy 1987). Ausreißer können einen großen Effekt auf beispielsweise statistische Parameter wie den Mittelwert haben. So kann ein Hochkostenfall den Mittelwert einer Kostenschätzung, insbesondere bei einer geringen Stichprobengröße, verzerren. Ausreißer sind in Analysen also mit entsprechender Vorsicht zu bewerten. Eine mögliche Lösung, einer solchen Verzerrung entgegenzuwirken, ist die Nutzung des Medians, der weniger anfällig für Ausreißer ist (Lange und Bender 2007). Des Weiteren kann der Median bei schiefen und unsymmetrische Verteilungen - beispielsweise Laborwerte – oder bei der Betrachtung von Überlebenszeiten (siehe Kapitel 3.3) besser Interpretiert werden (Lange und Bender

2007). Für einige Fragestellung ist der Median jedoch kaum geeignet. So lassen sich beispielsweise mit dem Wissen über die Stichprobengröße und den Mittelwert Rückschlüsse auf die Gesamtkosten schließen, dies gelingt dagegen mit dem Median nicht. Das arithmetische Mittel hingegen ist weiterhin ein sinnvolles Lagemaß, wenn der Ausreißer einen plausiblen Wert einer Stichprobe darstellt.

Eine weitere Möglichkeit die Ausreißer in Analysen mit berücksichtigen zu können, ist die Berechnung eines getrimmten Mittelwerts (engl.: trimmed oder truncated mean). Hierbei werden die Daten „getrimmt“, d. h. ein bestimmter Prozentsatz der Randdaten wird entfernt und bleibt unberücksichtigt. Aus den verbleibenden Werten des Datensatzes wird dann das arithmetische Mittel errechnet. Von einem um 5 % getrimmten Mittel wird somit gesprochen, wenn 5 % der höchsten Werte und 5 % der niedrigsten Werte bei der Berechnung des Mittelwerts entfernt werden. Dennoch verbleiben diese Daten im Datensatz. Möglich sind auch andere Grenzen, beispielsweise ein um 10 % getrimmtes Mittel. Nachteilig bei dieser Methode ist, dass ein bestimmter Anteil der Daten unberücksichtigt bleibt. Andererseits bietet sie die Möglichkeit, Verzerrungen des Mittelwertes durch Ausreißer entgegenzuwirken.

Die Ausreißer miteinzubeziehen wäre das simpelste, aber methodisch diskutierbarste, Vorgehen. Hierbei werden alle Beobachtungen inklusive Ausreißern in die Analyse miteinbezogen. Dieses Vorgehen ist jedoch lediglich zu empfehlen, wenn es sich bei dem Extremwert um einen plausiblen Wert der Stichprobe handelt, da so die Einbeziehung aller Beobachtungen die Realität widerspiegeln kann. Nachteilig ist jedoch, dass dieses Vorgehen womöglich zu Verzerrungen beispielsweise des Mittelwerts führen und hierdurch falsche Schlussfolgerungen gezogen werden kann. Handelt es sich bei dem Ausreißer um einen nicht plausiblen Wert einer Stichprobe kann der Ausschluss des Datensatzes eine mögliche Datenaufbereitungsstrategie sein. Da durch dieses Vorgehen jedoch relevante Informationen verloren gehen, ist dieses lediglich bei einer großen Datenmenge oder bei einer drastischen Verzerrung zu empfehlen.

Empfehlungen

- Ausreißer können folgendermaßen adressiert werden:
  - Getrimmtes Mittel
  - „Abschneiden“ bzw. festlegen von Unter- und Obergrenzen
  - Ausschluss von Datenfällen
  - Miteinbeziehung von Datenfällen

5.2.2 Negative Werte

Auch negative Werte können die statistischen Ergebnisse verfälschen oder zu falschen Aussagen führen, beispielsweise bei der Untergrenze möglicher Ausprägungen bzw. den Minimumangaben. Häufig sind negative Werte im Bereich der Kosteninformationen zu finden. So verringert ein negativer Kostenfall den Mittelwert einer Kostenanalyse und die Durchschnittskosten werden möglicherweise unterschätzt. Das Vorliegen negativer Kosten kann, ähnlich wie bei den Ausreißern, einen Hinweis auf Unstimmigkeiten bzw. Datenfehler geben. Dies muss jedoch nicht zwangsläufig der Fall sein. Negative Kosten können auch auf nachträgliche Umbuchungen, Regresse oder Gutschriften zurückzuführen sein. Allerdings können diese Werte auch durch Datenfehler verursacht sein. Dies gilt es ähnlich wie bei den Ausreißern zu prüfen. Im Einzelfall ist eine Abstimmung mit dem Dateneigner durchzuführen, um die Plausibilität negativer Werte einschätzen zu können.

Negative Werte können auch bei Berechnungen von Datumsdifferenzen auftreten. Negative Datumsangaben sind häufig, im Gegensatz zu negativen Kosten, nicht plausibel erklärbar. Durch die Digitalisierung und automatische Verarbeitung der Rezepte kann es beispielsweise im Arzneimittelbereich zum einen zu Einlese- bzw. Übermittlungsfehlern kommen und zum anderen können z. B. handschriftlich eingetragene Datumsangaben nicht richtig erfasst worden sein. Ist dies der Fall, so wird das „nicht lesbare“ Datum auf das Ende des Monats gesetzt, sodass es hier durchaus vorkommen kann, dass das Abgabedatum vor dem Verschreibungsdatum liegt. Hierbei entsteht bei Bildung der Datumsdifferenz eine negative Zeitspanne.

Negative Tagesangaben bei einem Krankenhausaufenthalt lassen sich meist nicht plausibel erklären. Sie entstehen im stationären Bereich, wenn die Entlassung vor

der Aufnahme erfolgte bzw. wenn das Entlassungsdatum vor dem Aufnahmedatum liegt, dies ist jedoch nicht plausibel.

Für den Umgang mit negativen Werten sind unterschiedliche Vorgehensweisen möglich. Die drastischste Lösung dieser Datenauffälligkeit ist auch hier das Löschen der Daten. Wie bereits in Kapitel 5.2 erwähnt, ist dies jedoch mit einem Informationsverlust verbunden, sodass hiervon möglichst Abstand genommen werden sollte.

Eine weitere Möglichkeit, die speziell bei Kostendaten zum Einsatz kommen kann, ist das Bilden des Aggregats der jeweiligen Kosten auf Patientenebene. Oft gleichen sich positive und negative Kosten aus, sodass Rückbuchungen entsprechend dem realen Abrechnungsgeschehen ausgeglichen werden können. Ergibt die Summe der Kosten pro Versicherten weiterhin einen negativen Wert, so ist zu überlegen, diesen entweder auf null zu setzen, um das Minimum nicht zu verfälschen und dennoch den Fall in der Analyse zu berücksichtigen. Zum anderen wäre das Löschen bzw. das Ignorieren solcher Fälle möglich, um die durchschnittlichen Kosten nicht zu unterschätzen.

Wenn es sich um einen Datenfehler handelt, wäre das Vertauschen beider Datumsangaben eine Möglichkeit zur Korrektur. Dies ist jedoch lediglich mit beispielsweise Einsicht in die Originaldaten bzw. nach Rücksprache mit dem Dateneigner zu empfehlen. Andernfalls könnte dieses Vorgehen auch eine starke Manipulation der Daten darstellen.

Empfehlungen

- Mit negativen Werten kann wie folgt umgegangen werden:
  - Ausschluss von Datenfällen
  - Auf Patientenebene summieren
  - Auf null setzen
  - Vertauschen beider Datumsangaben
  - Ignorieren von Datenfällen

5.2.3 Nullkosten

Auch Nullkosten können bei Kostenanalysen die Durchschnittskosten verzerren. Als Nullkosten werden alle Datensätze bezeichnet, die ausschließlich eine Null als Datenwert aufweisen, daher einen regulären Fall darstellen, jedoch keine positiven Kosten beinhalten. Sofern der Fall in die Analyse mit einfließt, kann dieser bei einer Kostenanalyse den Mittelwert und damit die Durchschnittskosten verzerren. Dies führt zu einer Unterschätzung der Kosten und damit eventuell zu falschen Folgerungen aus den Ergebnissen.

Gerade in den Rehabilitationsdaten treten solche Unregelmäßigkeiten häufiger auf, da in Deutschland traditionell unterschiedliche Kostenträger für die Erstattung der Maßnahme zuständig sind (siehe Kapitel 2.4.7). In der Regel ist bei jüngeren und erwerbsfähigen Versicherten die Rentenversicherung für die Finanzierung und Koordination der Rehabilitationsmaßnahme zuständig, sodass für diesen Personenkreis keine detaillierten Aussagen zum Rehabilitationsgeschehen anhand von GKV-Routinedaten möglich sind. Wenn der Antrag auf Rehabilitation bei der Krankenkasse eingereicht wird, jedoch ein anderer Kostenträger für diese zuständig ist, weist das Datawarehouse zwar den Fall der Rehabilitation aus, jedoch mit den bereits genannten Nullkosten. Um die Durchschnittskosten eines Rehabilitationsfalls aus Sicht der Krankenkasse zu ermitteln, empfiehlt es sich daher, alle Nullkostenfälle aus der Analyse auszuschließen.

Ebenfalls ist der Einfluss der Nullkosten je Variable, d. h. die Häufigkeit des Auftretens im jeweiligen Datenbereich, zu ermitteln. Weist eine Variable eine hohe Anzahl an Nullkosten auf, so kann dies ein Zeichen für einen Datenfehler sein. Empfohlen wird in solchen Fällen, Rücksprache mit dem Dateneigner zu halten und gegebenenfalls eine neue Datenlieferung zu veranlassen.

Grundsätzlich können Nullkosten nahezu in allen Leistungsbereichen vorkommen, jedoch sind die Gründe dafür häufig unbekannt. Ein Lösungsansatz könnte, wie bereits beschrieben, das Löschen der jeweiligen Datensätze sein. Eine Unterschätzung der Kosten kann vermieden werden. Eventuell ergibt sich jedoch durch das Löschen eine Überschätzung bezogen auf die mittleren Kosten, da die Gesamtkosten durch eine geringere Fallzahl dividiert werden. Nachteilig ist hingegen, dass der Fall bei anderen Analysenschritten evtl. auch unberücksichtigt bleibt, beispielsweise bei der Berechnung von Durchschnittstagen (Anzahl der Tage im Krankenhaus). So sollte

nicht der gesamte Datensatz des jeweiligen Versicherten gelöscht werden, sondern lediglich für die jeweilige Kostenanalyse unberücksichtigt bleiben.

Eine weitere Möglichkeit ist es, die Daten unverändert zu lassen und im vollen Umfang in die Analyse mit einzubeziehen. Ein Vorteil dieser Vorgehensweise ist, dass die Daten nicht „willkürlich“ manipuliert werden. Wiederum wäre eine mögliche Unterschätzung der Ergebnisse denkbar, da die Datensätze mit dem Wert null eingehen würden.

Als dritte, jedoch aufwendigste Vorgehensweise sei die Bewertung der Kosten über Standardkosten zu nennen. Bei dieser Methode wird beispielsweise im Krankenhaus und Rehabilitationsbereich mit einem Durchschnittssatz (Geldeinheiten pro Tag) gerechnet (Prenzler et al. 2010). Im Arzneimittelbereich könnten die Preise der LAUER-Taxe entnommen werden und mit diesen die Kosten kalkuliert werden (LAUER-Taxe). Diese Verfahrensweise bildet die Kosten bestmöglich ab und ergänzt somit die Daten. Jedoch kann es z. B. durch kassenindividuelle Rabattverträge zu einer Überschätzung der Ergebnisse aus der Perspektive der GKV kommen. Des Weiteren stellt dies eine umfassende Manipulation der Daten dar. Eine weitere Möglichkeit Nullkosten zu adressieren ist das Ersetzen der Nullwerte durch den Mittelwert der übrigen Datensätzen. Hier gelten die gleichen Vor- und Nachteile wie zuvor.

Empfehlungen

- Mit Nullkosten kann wie folgt umgegangen werden:
  - Ausschluss bzw. Löschung von Datenfällen
  - Daten unverändert lassen
  - Berechnungen mit Durchschnitts- bzw. Standardsätzen
  - Substitution der Nullwerte durch den Mittelwert der übrigen Datensätzen

5.2.4 Fehlende Werte

Von fehlenden Werten, engl. missing values, wird gesprochen, wenn kein Datenwert für die jeweilige Variable während einer Beobachtung vorliegt. Solche Ereignisse treten häufig in Datenerhebungen bzw. -auswertungen auf, können jedoch einen erheblichen Einfluss auf die statistischen Analysen und deren Schlussfolgerungen haben. Anders als bei Nullkosten kann der jeweilige Datenfall aufgrund der fehlenden konkreten Ausprägung nicht mitberücksichtigt werden. Dies kann zu einer geringen Fallzahl und gegebenenfalls zu einer Unterschätzung der Fallzahl führen.

Die Gründe für das Fehlen von Daten können vielschichtig sein und müssen überprüft werden. Unvollständigkeit von Sekundärdaten sowie Codierungs- und Übertragungsfehler der Daten sind die Hauptgründe für das Fehlen von Daten bei einer Sekundärdatenanalyse. Dies kann beispielsweise der Fall sein, wenn einzelne Variablen im Rahmen des Extraktionsprozesses nicht vollständig extrahiert wurden.

Zunächst ist zu prüfen, ob es sich bei den fehlenden Werten um einen Datenfehler bzw. eine Fehlübermittlung handelt. Liegt eine Extraktions- bzw. ein Übermittlungsfehler vor, muss dies mit den bereits in Kapitel 5.2 erwähnten Maßnahmen korrigiert werden.

Unterschieden wird zwischen systematisch fehlenden Werten, die z. B. nicht im Datensatz erfasst sind und nicht zufällig fehlen, sowie unsystematisch fehlenden Werten, die tatsächlich fehlerhaft codiert wurden bzw. zufällig fehlen (Runte 1999). Rubin unterscheidet drei Arten von fehlenden Werten:

- Missing at random (MAR), wenn das Fehlen der Daten unabhängig von der Merkmalsausprägung selbst ist,
- Oberserved at random (OAR), wenn das Fehlen der Daten unabhängig von den anderen Merkmalsausprägungen ist und
- Missing completely at random (MCAR), wenn sowohl MAR und OAR zutreffen

(Runte 1999; Rubin 1976).

Ein Beispiel für systematisch fehlende Daten ist der Tätigkeitsschlüssel bei Familienversicherten. Dieser wird in den GKV-Routinedaten nicht erfasst und tritt somit zwingenderweise als ein fehlender Wert auf. Dies ist im Studiendesign zu berücksichtigen und bereits vorher sorgfältig zu planen. Muschik und Jaunzeme haben in einem Bei-

trag die Übertragbarkeit des Bildungsstandes von Versicherungsmitgliedern, der anhand des Tätigkeitsschlüssels abbildbar ist, auf die mitversicherten Familienversicherten diskutiert (Muschik und Jaunzeme 2014). In der epidemiologischen Forschung wird über den Ehepartner oder die Familie versucht, den sozialen Status einer Person zu erfassen (Baxter 1994). In der Studie wurde überprüft, ob diese Übertragung auch in einer GKV-Routinedatenanalyse möglich ist. Diese Herangehensweise birgt jedoch einige Risiken und ist mit massiven Annahmen verbunden, sodass diese für die meisten Studien nicht zu empfehlen ist.

Auswertungen auf Basis des Mittelwertes könnten durch fehlende Werte verzerrt werden. Um Aufschluss über die Art des Ausfallmechanismus bzw. der Missing-Value-Struktur zu bekommen, kann die Durchführung einer Strukturanalyse hilfreich sein. Ziel ist es hierbei, unsystematische Ausfallmechanismen aufzudecken. Als mögliche Vorgehensweisen sind die deskriptive, explorative sowie die induktive Analyse zu nennen (Bankhofer 1995). Hierbei werden bei der deskriptiven Analyse sogenannte Missing-Data-Maße berechnet, die das Verhältnis von fehlenden und existierenden Werten ermitteln (Rubin 1976). Bei der explorativen Analyse werden Abhängigkeiten bzw. Zusammenhänge innerhalb der Daten analysiert und aufgedeckt. Bei der induktiven Analyse wird auf Konzentrationen von missing values und/oder unsystematischen Mechanismen getestet (Bankhofer 1995).

Auf Basis dieser Strukturanalysen können unterschiedliche Strategien zum Umgang mit fehlenden Werten angewendet werden. Die Literatur unterscheidet meist zwischen drei Methoden: Eliminierungsverfahren bzw. Ausschlussverfahren, Imputationsverfahren und Parameterschätzverfahren (Schwab 1991).

Bei der ersten Methode werden unvollständige Fälle oder Variablen bewusst von der Analyse ausgeschlossen und aus dem Datensatz entfernt; respektive werden ausschließlich vollständige Fälle für die weitere Analyse verwendet (complete-case analysis). Eine weitere Methode ist die sogenannte available-case analysis, bei der partiell Variablen bzw. Merkmale ausgeschlossen werden. Hier stehen die Daten für weitere Auswertungen noch weiterhin zur Verfügung. Ein zentraler Vorteil dieser Methoden ist ihre einfache Anwendbarkeit. Es wird mit einer vollständigen Datenmatrix weitergearbeitet und gewährleistet, dass die Ergebnisse mit univariaten Analysen verglichen werden können (Little und Rubin 2002). Liegt ein systematischer Ausfallmechanismus vor, kann es jedoch bei diesem Verfahren zu schwerwiegenden Verzerrun-

gen kommen (Runte 1999). Des Weiteren ist darauf zu achten, dass die Stichprobe durch die Eliminierung der Fälle nicht zu gering ausfällt und damit eine valide Interpretation der Daten unmöglich wird. Somit ist bei entsprechend großen Stichproben, bei einer geringen Anzahl von fehlenden Werten und beim Vorliegen von MCAR das Eliminierungsverfahren zu empfehlen. Zwischen den Informations- bzw. Datenverlust durch den Ausschluss der Daten und den Vorteilen, die aus der Reduktion der fehlenden Werte entstehen können, ist hier abzuwägen. Als „wenige fehlende Werte“ können schätzungsweise weniger als 5 % der Gesamtzahl an Fällen definiert werden. Wenn die fehlenden Werte dann auch noch als zufällig fehlend betrachtet werden können, also das Fehlen eines Werts unabhängig von anderen Werten ist, dann ist die Methode des listenweisen Löschens relativ sicher zu empfehlen, da sie dann keinen zu starken Informationsverlust hervorruft (Bühl 2012).

Eine weitere Möglichkeit, mit missing values umzugehen, besteht darin, diese durch verschiedene induktive und statistische Verfahren zu ersetzen (Imputationsmethode). Ein solches Ersetzen ist jedoch lediglich unter bestimmten Voraussetzungen möglich und dient dazu, dem Informationsverlust des Eliminierungsverfahrens entgegenzuwirken. Dies geschieht dadurch, dass die fehlenden Werte ersetzt und somit die Daten vervollständigt werden. Von induktiven Ersatzwertverfahren wird gesprochen, wenn die fehlenden Werte ohne Berechnungen und auf der Basis von anderen, teilweise externen Informationen ersetzt werden. Hierzu zählen beispielsweise das Nachbeobachten und Nachfassen bei Non-Response, wobei dies bei Sekundärdaten in der Regel nicht machbar ist und darüber hinaus bei Zufallsstichproben die Repräsentativität gefährdet. Externe Quellen (Cold-Deck-Technik) oder Daten aus vorangegangenen Studien lassen sich alternativ auch als Konstanten für fehlende Werte verwenden (Reinboth 2006). Ein Beispiel für die GKV-Routinedaten wäre im Bereich der Arzneimittel zu nennen. So können fehlende Werte bei den ATC-Codes auftreten. Falls dies der Fall ist, können mithilfe des GKV-Arzneimittelindex des WIdO und der PZN u. a. die dazugehörigen ATC-Codes und die DDD-Angaben ergänzt werden (WIdO). Auch die LAUER-Taxe kann in diesem Bereich als externe Datenquelle genutzt werden (LAUER-Taxe). So können die Arzneimittel mithilfe der PZN beispielsweise in Depotmedikation oder Nicht-Depotmedikation eingeteilt oder Markennamen ergänzt werden. Ergibt sich der Fall, dass kein ATC-Code vorhanden ist, weil es sich um ein Heilmittel handelt, ist zu empfehlen, diesen Datensatz im Heil- und Hilfsmittel-

sektor zu verorten. Gegebenenfalls sind anfallende Kosten auch dem jeweiligen Sektor zuzuordnen.

Zu den statistischen Ersatzverfahren gehört beispielsweise das Ersetzen der fehlenden Werte durch statistische Maße. Voraussetzung hierfür ist, dass der fehlende Wert zufällig ausgefallen ist (MCAR). Je nach Skalenniveau kann der Mittelwert, der Median oder der Modus als Imputationsschätzer eingesetzt werden. So wird aus den vorhandenen Werten das jeweilige statistische Maß errechnet und die entsprechenden fehlenden Werte werden durch dieses ergänzt. Weitere Variationen des Mittelwertersatzes sind ferner der Einsatz des Medians der Nachbarpunkte, die Berechnung eines Zeitreihen-Mittelwerts (wo Zeitreihen-Daten vorliegen) oder die lineare Interpolation. Die Vorteile dieser Verfahren liegen in der einfachen Umsetz- und Anwendbarkeit. Jedoch können diese die Verteilung der Daten, die Varianz der Variablen und eventuell auftretende Korrelationen in den Daten verzerren (Fahrmeir 2010; Bühl 2012).

Eine Spezialform des statistischen Werteersatzes ist der Einsatz eines linearen Trendmodells. Dieses kann eingesetzt werden, wenn für die gültigen Werte ein klarer linearer Trend erkennbar ist. Da sich jedoch durch das Ersetzen die Varianz der Variablen verringert, könnten vorhandene Regelmäßigkeiten verstärkt werden. Weitere eingängige/simple Imputationstechniken sind der Einsatz eines Verhältnisschätzers (Ford 1976), einer Zufallsauswahl (Schnell 1986) und eines Expertenratings (Bankhofer 1995; Little und Rubin 2002). Komplexere Verfahren sind multivariate Imputationstechniken. Bei diesen gibt es allerdings viele Abwandlungen. Einen Überblick gibt Bankhofer (1995).

Im Allgemeinen können als Vorteile aller Imputationsverfahren die Vermeidung von Informationsverlust und die vollständige Datenmatrix genannt werden. Dennoch können diese Verfahren unter bestimmten Voraussetzungen zu Verzerrungen führen.

Bei den Parameterschätzverfahren werden die fehlenden Werte durch geeignete Methoden, z. B. Faktoren- und Diskriminanzanalysen, geschätzt. Im Unterschied zu den Imputationstechniken werden bei der Schätzung der Konstanten Korrekturen durchgeführt, die einer Verzerrung entgegenwirken sollen. Auch hier ist der Vorteil, dass dieses Verfahren zu keinem Informationsverlust durch das Löschen von Daten führt. Des Weiteren gelten hierbei weniger restriktive Voraussetzungen als bei den Imputa-

tionstechniken. Nachfolgende Analysen sind jedoch lediglich dann anzuwenden, wenn sie auf den ermittelten Parametern beruhen.

Eine weitere Möglichkeit, eine nicht vollständig ausgefüllte Datenmatrix zu analysieren, ist das Missing-Value-Linkage-Verfahren (Schader und Gaul 1992). Diese Vorgehensweise gehört zu den multivariaten Analyseverfahren und bietet den Vorteil, dass keine künstlichen Daten erzeugt werden und somit der unvollständigen Datenmatrix Rechnung getragen wird. Nachteilig ist jedoch, dass nur auf die vorhandenen Daten zurückgegriffen werden und dies zu Verzerrungen führen kann.

Gezeigt wurde, dass fehlende Daten ein Problem bei der statistischen Datenanalyse darstellen. Grundsätzlich ist darauf zu achten, welcher Ausfallmechanismus vorliegt und wie sich die fehlenden Daten auf die Analyse auswirken. Mit dieser Kenntnis können unterschiedliche Verfahren herangezogen werden, um den Herausforderungen, die in Verbindung mit den missing values entstehen, entgegenzuwirken. Jedoch existiert kein universell geeignetes Verfahren. So muss je nach Datenlage, Zielsetzung und Abwägung der Vor- und Nachteile der jeweiligen Verfahren eine individuelle Entscheidung getroffen werden.

Empfehlungen

- Es existieren unterschiedliche Verfahren fehlende Werte in Analysen zu berücksichtigen, beispielsweise:
  - Eliminierungsverfahren bzw. Ausschlussverfahren
  - Ausschluss von Datenfällen
  - Eliminierungsverfahren: bei großen Stichproben und Vorliegen von MCAR
  - Teilweiser Ausschluss von Datenfällen
  - Imputationsverfahren
  - Parameterschätzverfahren

5.3 Zuordnungsproblematik

Zu Anfang jeder Studie sollte der Studienzeitraum klar definiert und sorgfältig ausgewählt werden. Dies ist insbesondere bei GKV-Routinedatenanalysen zu beachten, da in den Rohdaten jeder Ressourcenverbrauch unabhängig von der Versicherungszeit abgebildet wird (Jaunzeme und Muschik 2014). Abhängig von der Fragestellung muss entschieden werden, ob z. B. nur durchgängig versicherte Personen inkludiert oder auch unterjährige Versichertenzeiten berücksichtigt werden sollen. Durch den Einschluss lediglich von durchgängig Versicherten könnten in Bezug auf den gesamten Studienzeitraum die in Anspruch genommenen Leistungen überschätzt werden, da nicht durchgängig Versicherte einen kürzeren Beobachtungszeitraum haben und somit auch weniger Leistungen in Anspruch nehmen können. Des Weiteren wird durch den Ausschluss der nicht durchgängig Versicherten die Fallzahl unterschätzt.

Des Weiteren existieren Herausforderungen bei der Zuordnung der Leistungen zu den jeweiligen Zeiträumen. So stehen beispielsweise bei Längsschnittanalysen mehrere Jahre für die Analyse zur Verfügung, und die Zuordnung von Kosten, Diagnosehäufigkeiten sowie erbrachten Leistungen bzw. Verordnungen stellt eine Herausforderung dar, wenn diese über den Jahreswechsel gehen.

Eine Möglichkeit für eine Zuordnung zu einem Betrachtungszeitraum (beispielsweise ein Jahr) ist eine rechtsseitig zensierte Selektion. Hierbei werden nur die Fälle in die Analyse mit eingeschlossen, die im jeweiligen Betrachtungszeitraum abgeschlossen sind (die sogenannten Einstrahler). Das heißt, einstrahlende Fälle werden mit dieser Methode berücksichtigt, ausstrahlende jedoch nicht (Bödeker 2005). Dieses Verfahren wird beispielsweise bei der Krankheitsartenstatistik des BKK Bundesverbandes verwendet. Hierbei wird eine rechtszensierte Selektion bei der Berechnung von AU-Fällen genutzt. Diese Vorgehensweise hat zur Folge, dass AU-Fälle, die vor dem Berichtszeitraum begonnen haben, berücksichtigt werden. Dieses vom Bundesministerium für Gesundheit vorgeschriebene Auswertungsverfahren zur Berichtserstattung ist für eine möglichst genaue Abbildung der tatsächlichen AU-Dauer vorzuziehen. Wird sich für die genauen angefallenen AU-Tage im jeweiligen Beobachtungszeitraum interessiert, wäre eine Berücksichtigung ausschließlich der Tage, die im Beobachtungszeitraum liegen, viel genauer. Selbstverständlich kann die Vorgehensweise der Rechtszensierung auch bei Krankenhausaufenthalten, Krankengeldzahlungen und Versichertenzeiten angewendet werden. Nachteilig an dieser Methode ist

jedoch, dass Fälle, die nicht im jeweiligen Betrachtungszeitraum abgeschlossen sind, unberücksichtigt bleiben und somit ein Informationsdefizit über diese Fälle vorliegt. Dennoch besitzt die Rechtszensierung gerade bei Kostenanalysen eine gute Rationale, da alle Vorgänge bei diesem Verfahren abgeschlossen sind, hinreichend viele Informationen über Diagnosen und Maßnahmen verfügbar sind und nach dem Prinzip des Rechnungseingangs vorgegangen werden kann.

Grobe berichtet bei der Berechnung von Behandlungsfallhäufigkeiten über eine zeitliche Zuordnung über das Aufnahmedatum im Krankenhaus (Grobe 2005). Im Gegensatz zu der oben genannten rechtsseitig zensierten Selektion werden dabei lediglich Fälle in der Analyse berücksichtigt, die im Untersuchungszeitraum begonnen haben; das End- bzw. Entlassungsdatum bleibt jedoch zunächst unberücksichtigt. Bei dieser Linkszensierung würden beispielsweise vermehrte Krankenhauseinweisungen aufgrund einer Grippeepidemie am Jahresende in das jeweilige Kalenderjahr fallen. Als Vorteil ist die genaue Abbildbarkeit und Analyse der tatsächlichen, in dem jeweiligen Jahr angefangenen Krankenhauseinweisungen oder der AU-Fälle zu nennen. Jedoch werden hier einstrahlende Fälle von der Analyse ausgeschlossen. So kommt es ebenfalls, ähnlich wie bei den Einstrahlern zuvor, zu einem gewissen Informationsverlust, da in diesen Fällen eine Linkszensierung vorliegt (Grobe 2005). Des Weiteren liegt das Aufnahmedatum weiter entfernt von der Kostenentstehung bzw. -abrechnung, welches bei Kostenanalysen ein Wiederspruch zur zuvor empfohlenen Methode darstellt.

Eine weitere Vorgehensweise ist die Berücksichtigung von sowohl ein- als auch ausstrahlenden Fällen. Hier werden alle Datensätze berücksichtigt, die vor dem Untersuchungsintervall begonnen haben, sofern sie innerhalb des Zeitraumes enden (Einstrahler), sowie Datensätze, die im Beobachtungszeitraum beginnen, jedoch erst nach dem Untersuchungsende enden (Ausstrahler). Ein Vorteil dieser Methode ist die vollständige Abbildbarkeit aller Fälle, die im Betrachtungszeitraum vorlagen. Jedoch werden mit dieser Methode die Ergebnisse, beispielsweise die durchschnittlichen Tage einer Gesundheitsleistung oder die damit verbundenen Kosten, überschätzt, da sowohl Fälle hineinzählen, die vor dem Beobachtungszeitraum angefangen haben, als auch solche, die über dieses Jahr hinausgehen.

Wahlweise können auch lediglich im jeweiligen Bezugsjahr begonnene und zusätzlich auch abgeschlossene Fälle in den Analysen berücksichtigt werden. Genauer ge-

sagt, müsste hierfür sowohl das Anfangsdatum als auch das Enddatum im jeweiligen Berichtsjahr liegen. Ein Vorteil wäre, dass die Daten nicht künstlich zensiert werden. Auf der anderen Seite würde dieses Vorgehen möglicherweise zu einer Unterschätzung der tatsächlichen Fälle führen, da Ein- und Ausstrahler unberücksichtigt blieben.

Wenn der genaue Umfang der angefallenen AU-Tage beispielsweise in einem Kalenderjahr festgestellt werden soll, dann kann auch eine weitere in der Krankheitsartenstatistik und den Gesundheitsberichten des AOK-Verbandes genutzte Methode angewendet werden. Diese, auch in der betrieblichen und innungsspezifischen Gesundheitsberichtserstattung der IKK angewendete Vorgehensweise selektiert jene AU-Fälle, die im jeweiligen Betrachtungszeitraum gemeldet werden. Bei der untersuchungszeitbezogenen Aufbereitung wird die AU-Dauer jeweils vom Beobachtungsbeginn und -ende zensiert. Das wiederum heißt, dass der AU-Beginn bzw. das AU-Ende vor bzw. nach dem Untersuchungszeitraum liegen kann, diese Tage aber nicht in die Analyse mit eingeschlossen werden. Ein Vorteil dieser Methode ist, dass langwierige AU-Fälle die Auswertungen nicht verzerren, da diese „abgeschnitten“ und künstlich zensiert werden. Auch ein anteiliges Verrechnen bzw. Aufteilen der Kosten ist möglich. Sofern der Fall beispielsweise im Bereich der Krankengeldzahlungen außerhalb des definierten Betrachtungszeitraums liegt, werden die Kosten anteilig den Zeiträumen zugeschlüsselt. Liegt zum Beispiel der Zahlungsbeginn bzw. das Zahlungsende außerhalb dieses Betrachtungszeitraums, werden nur die Kosten berücksichtigt, die im relevanten Zeitraum bezüglich ihrer Tage anfallen. Die Methode der tageweisen Aufschlüsselung ist jedoch im stationären Sektor nicht zu empfehlen, da der Ressourcenverbrauch im Krankenhaus tageweise sehr unterschiedlich ausfallen kann. Auch die Umstellung im Jahre 2004 von den bisher tagesgleichen Pflegesätzen hinzu den DRG-Fallpauschalen bestärkt dieses Argument. Des Weiteren dürfen die Ressourcenverbräuche bei der Berechnung von Durchschnittskosten pro Krankenhausaufenthalt bzw. der durchschnittlichen Länge eines Krankenhausaufenthalts nicht aufgesplittet werden, sondern es empfiehlt sich den gesamten Krankenhausfall zu einer exakten Periode zuzuschlüsseln.

Ein einheitlicher Standard für die Zurechnung der Leistungen auf den jeweiligen Beobachtungszeitraum existiert jedoch nicht (Bödeker 2005). So ist es von der Fragestellung abhängig, für welches methodische Vorgehen sich der Routinedatennutzer

entscheidet. Weitere Regelungen bezüglich Stichtag oder Versichertendauer müssen vorher bei der Studiendesignplanung getroffen werden.

Die Wahl der Studienperspektive gehört zu den grundlegenden Entscheidungen einer Evaluationsstudie. Die Perspektivenwahl für die Bewertung von Ressourcenverbräuchen im Gesundheitswesen hat einen entscheidenden Einfluss beispielsweise auf die Höhe der zu ermittelnden Kosten. So können die Ergebnisse der Untersuchung sehr unterschiedlich ausfallen. Meist werden drei Perspektiven unterschieden: die Perspektive des Kostenträgers (GKV), die Perspektive der Patienten und Angehörigen sowie die gesellschaftliche Perspektive. Der breiteste Ansatz stellt die soziale bzw. gesamtwirtschaftliche Perspektive dar. Diese bezieht sämtliche Kosten (und den Nutzen, der jedoch nicht mit GKV-Routinedaten abgebildet werden kann) mit ein, ohne zu berücksichtigen, bei wem diese entstehen. Eine andere, auch bei GKV-Routinedatenstudien weit verbreitete Perspektive ist die Kostenträgerperspektive. Aus Krankenkassensicht ist der reine Netto-Zahlbetrag der Krankenkasse die relevante Maßgröße. Diese sollte z. B. von der Zuzahlung durch den Versicherten bereinigt werden. Eine enge Zusammenarbeit mit dem Datenhalter ist auch hier notwendig, um den Zahlbetrag der Krankenkasse klar von dem Rechnungsbetrag abgrenzen zu können. In diesem Zusammenhang stehen unterschiedliche Variablen wie beispielsweise Zuzahlung, Nettobeträge etc. zur Verfügung. Je nach Fragestellung ist die relevante Variable bei dem Datenhalter nachzufragen.

Empfehlung

- Mögliche Strategien für die Leistungszuordnung
  - Einstrahler (Enddatum muss im Untersuchungszeitraum liegen)
  - Ausstrahler (Anfangsdatum muss im Untersuchungszeitraum liegen)
  - Ein- und Ausstrahler (Anfangs- und Enddatum können sowohl vor als auch nach dem Untersuchungszeitraum liegen)
  - Abschneiden (Der Untersuchungszeitraum beschneitet die Datenfälle)
  - Anteilig verrechnen (Nur die Tage die in den Untersuchungszeitraum werden berücksichtig)

Bezugsgrößen für die jeweilige Zuschlüsselung

Die Problematik der ein- und ausstrahlenden Fälle betrifft sowohl die Kosten als auch generell jede zeitliche Zuordnung in den GKV-Routinedaten. Zur Identifikation der relevanten Leistungskosten kann entweder der Beginn oder das Ende der Leistungserbringung bzw. das Entlassungsdatum als relevante Bezugsgröße für die oben genannten Methoden herangezogen werden. Eine Analyse anhand des Entlassungsdatums ist von Vorteil, da sich dieses Datum, zeitlich gesehen, „näher“ an der Kostenentstehung für den Kostenträger (in diesem Fall die Krankenkasse) und der jeweiligen Abrechnung befindet. Des Weiteren sind mit dem abgeschlossenen Fall hinreichend viele Informationen über Diagnose und Maßnahmen verfügbar. Werden beispielsweise Leistungen ein Jahr vor bzw. ein Jahr nach einem Indexereignis untersucht, findendie Fälle Beachtung, deren Aufnahmedatum außerhalb des festgelegten Zeitraums und ihr Entlassungsdatum innerhalb des Zeitraums liegen, Berücksichtigung. Andererseits werden Fälle nicht einbezogen, deren Aufnahmedatum im Beobachtungszeitraum läge, aber deren Entlassungsdatum außerhalb der Grenzen liegt. Häufig kann davon ausgegangen werden, dass sich die Einstrahler- und Ausstrahlergegebenheiten größtenteils ausgleichen und sie somit eine akzeptable Limitation darstellen. Wenn jedoch mit einem Indexereignis (beispielsweise das erstmalige Auftreten einer Krankheit) gearbeitet wird, können vor und nach diesem Ereignis unterschiedliche Gegebenheiten vorherrschen, sodass eine Vergleichbarkeit und somit ein Ausgleichen der Ein- und Ausstrahler nicht angenommen werden kann. Es ist somit stets zu prüfen, ob die Annahme der sich ausgleichenden Ein- und Ausstrahler Bestand hat.

Bei gesundheitsökonomischen Analysen ergeben sich weiterhin einige Herausforderungen durch die unterschiedlichen Zeitpunkte der Abrechnung bzw. Datumsdokumentation und der realen Ressourceninanspruchnahme (Reinhold et al. 2011a). So setzt beispielsweise die Abrechnung eines Arzneimittelrezeptes an einem konkreten Zeitpunkt an, obwohl davon auszugehen ist, dass der Patient das Medikament über einen gewissen Zeitraum einnimmt. Im Arzneimittelbereich existieren zwei unterschiedliche Datumsangaben: das Verschreibungs- und das Abgabedatum (siehe Kapitel 2.4.4). Je nach Fragestellung kann das eine oder das andere Datum gewählt werden. Wird die Leitlinienadhärenz oder das Verhalten von Ärzten untersucht, ist das Verschreibungsdatum die optimale Referenz. Werden Dosierungsanalysen

durchgeführt, wäre das Abgabedatum als relevantes Bezugsdatum zu wählen. Die Rationale dafür ist, ähnlich wie bereits im Krankenhausbeispiel zuvor, die zeitliche Komponente. So lange beispielsweise der Versicherte das Verordnungsblatt in der Apotheke noch nicht eingelöst hat, kann er das Medikament auch nicht einnehmen. Somit liegt das Datum näher an der tatsächlichen Einnahme des Medikamentes.

Bei der Zuordnung von Arbeitsunfähigkeitstagen kann eine Herausforderung darin bestehen, dass die Krankengeldzahlungen lediglich für den gesamten Zeitraum und nicht monatsgenau erfasst werden. Zwar sind die einzelnen Auszahlungsbeträge in den GKV-Routinedaten abgebildet, eine exakte zeitraumbezogene Zuordnung ist jedoch aufgrund der fehlenden zeitlichen Angabe nicht möglich. Entsprechend ist die Zuordnung bei jahresübergreifenden Analysen auf die einzelnen Jahre eventuell problematisch.

Darüber hinaus ist die Berechnung des Alters der Studienpopulation zu diskutieren. Da in den GKV-Routinedaten meist lediglich das Geburtsjahr des jeweiligen Versicherten zur Verfügung steht, muss definiert werden, zu welchem Bezugszeitpunkt das Alter ermittelt bzw. errechnet werden soll. Grobe und Ihle schlagen zur Berechnung des Alters bei der Auswertung von Daten zu einzelnen Kalenderjahren folgende Formel vor (Grobe und Ihle 2005):

Alter = Beobachtungsjahr – Geburtsjahr

Dieses Verfahren kann auch dann angewendet werden, wenn ein exakter Geburtstag vorliegt (Grobe und Ihle 2005). Wie das Beobachtungsjahr jedoch definiert wird, ist von der Zielaussage, die getroffen werden soll, abhängig. So könnte das Alter zum Indexereignis oder zur Baseline eine mögliche Bezugsgröße sein. Je nach Fragestellung ist es wichtig, wie alt der Versicherte zum Ereignis (z. B. Ausbruch der Erkrankung) ist oder wie alt der Versicherte zur Baseline ist, d. h. wie die Baseline-Charakteristika der jeweiligen Studienpopulation sind.

Viele Studien untersuchen anhand von GKV-Routinedaten auch die Inzidenz und Prävalenz von Indikationen. Abbas et al. haben hierbei untersucht, wie lang die Baseline - das heißt der krankheitsfreie Zeitraum - sein muss, um valide Ergebnisse bei derartigen Fragestellungen zu erhalten (Abbas et al. 2012). Hierbei untersuchten sie den Einfluss von verschieden langen vorangehenden krankheitsfreien Intervallen an-

hand von drei ausgewählten Erkrankungen (Diabetes mellitus, Kolorektalkarzinom und Herzinsuffizienz). Sie kamen zu dem Ergebnis, dass es, verglichen mit einem acht-jährigen krankheitsfreien Vorlauf, bei einem ein-jährigen freien Vorlauf zu einer Überschätzung der Inzidenz von 40 %, 23 %, und 43 % für Diabetes, Darmkrebs und Herzversagen kommt. Bei der Annahme eines fünf-jährigen krankheitsfreien Zeitraum kam es hingegen zu einer Überschätzung von 5 %, 9 % und 5 %. Daraus lässt sich schließen, dass Vorsicht geboten ist bei der Verwendung von kurzen krankheitsfreien Perioden für die Inzidenzschätzungen. Die inzidenten Patienten können durch zu kurze krankheitsfreie Zeiträume extrem überschätzt werden.

Da sich bei längeren Beobachtungszeiträumen einzelne Merkmale der Stammdaten im zeitlichen Verlauf ändern können, ist auch hier zu überlegen, welche Ausprägung für die Analysen gewählt wird (Grobe und Ihle 2005). Die Beitragsgruppe und der Tätigkeitsschlüssel zählen zu den Parametern, die sich beispielsweise durch einen Arbeitsplatzwechsel ändern können. Grundsätzlich ist es von der Fragestellung und dem Ziel der Auswertungen abhängig, in welchem Umfang die Daten dafür bearbeitet und genutzt werden. Werden vorranging Eintrittsrisiken nach einem bestimmten Ereignis ausgewertet, müssen die Variablen innerhalb dieses Zeitraumes bzw. der letzten dokumentierten Ausprägung gewählt werden. Wird die jeweilige Variable bzw. werden ihre Ausprägungen jedoch für Subgruppenanalysen berücksichtigt, sind auch Änderungen der Parameter zu allen Zeitpunkten höchst relevant. Ein Beispiel hierfür wäre die Analyse von berufsgruppenspezifischen AU-Fehlzeiten (Grobe und Ihle 2005). Unterschiedliche Bezugsgrößen können herangezogen werden. Der erste Status im Beobachtungszeitraum könnte dabei eine sinnvolle Möglichkeit zur Statusdefinition darstellen. Der Vorteil wäre eine leichte Identifikation und zum anderen würde dieser Status in der Baseline liegen, falls dies in der Analyse so vorgesehen ist. Auch der letzte Fall im Studienzeitraum wäre ein mögliches Auswahlkriterium für sich im Zeitablauf ändernde Variablen. Dieser Fall ist ebenfalls leicht zu identifizieren, jedoch liegt dieser meist zeitlich am Ende des Studienzeitraumes und somit zu weit entfernt von der Baseline bzw. dem Indexereignis um den relevanten Status abbilden zu können. Häufig wird auch die Ausprägung zum Zeitpunkt des Indexereignisses gewählt. Wird mit einem Initialereignis gearbeitet, so ist diese Ausprägung zum Indexereignis am besten geeignet, um den zu diesem Zeitpunkt geltenden Status abzubilden. Des Weiteren können von dieser Ausgangssituation auch Veränderungen im Krankheitsverlauf und Zeitablauf analysiert werden.

Empfehlungen

- Es ist das Datum zu wählen, das näher am Ereignis bzw. an der relevanten Bezugsgröße liegt
- Beispiele hierfür sind:
  - Erstes Auftreten
  - Letzter Fall
  - Ausprägung zum Indexereignis

5.4 Zuzahlungen

Die Versicherten der GKV werden an den Kosten bestimmter Leistungen beteiligt, um ein kostenbewusstes und verantwortungsvolles Inanspruchnahmeverhalten zu fördern. Der Eigenanteil bzw. die Zuzahlungen der Versicherten im Arzneimittel- und Hilfsmittelbereich umfasst/umfassen meist 10 % der Kosten, mindestens jedoch 5 € und höchstens 10 €. Die Zuzahlungen überschreiten dabei jedoch nie die Kosten des jeweiligen Mittels. Im Krankenhaus- und Rehabilitationsbereich existieren keine prozentualen Zuzahlungssätze, hier werden meist Tagespauschalen berechnet.

Zuzahlungen sind auch für die Arbeit mit den GKV-Routinedaten relevant. Informationen zu Kosten werden in der Regel in GKV-Routinedaten als sog. Bruttokosten abgebildet. Diese umfassen – wie der Name bereits andeutet – Komponenten, die über die tatsächlichen Kosten (Nettokosten) aus Sicht der Krankenkasse hinausgehen. Hierzu zählen insbesondere die Zuzahlungen der Versicherten und mögliche Einsparungen durch kassenindividuelle Arzneimittel-Rabattverträge. Angaben zu Nettokosten sind häufig nicht verfügbar, da geheime, kassenindividuelle Rabattverträge eine bedeutende Wettbewerbskomponente im GKV-Markt darstellen.

Problematisch wird es, wenn z. B. die Kosten aus Sicht der GKV ermittelt werden sollen, allerdings nur Bruttokosten vorliegen, die auch Patientenzuzahlungen enthalten. Auch bei Analysen aus gesellschaftlicher Perspektive wäre es wünschenswert, die Kosten der GKV und die Eigenanteile der Patienten separat auswerten zu können.

Bisher mangelt es an Literatur und Erkenntnissen, wie mit dieser Problematik bei der Analyse von GKV-Routinedaten umgegangen werden kann. Im Folgenden werden

daher Vorschläge präsentiert, welche die Verzerrung durch Zuzahlungen verringern können.

Grundsätzlich scheint es aufgrund der transparenten Zuzahlungsregelungen möglich, diese generell entsprechend der gesetzlichen Vorgaben von den Kosten in allen relevanten Bereichen der GKV mit Zuzahlungsregelungen abzuziehen. Insbesondere bei Krankenhausbehandlungen, ambulanten und stationären Reha-Maßnahmen und Anschlussrehabilitationen fallen - abweichend von der 10 %-Zuzahlungsregelung - 10 € Zuzahlungen pro Kalendertag an. Die Länge der jeweiligen Behandlung ist dabei in der Regel hinreichend exakt aus den GKV-Routinedaten ermittelbar. Somit wäre es z. B. bei einem siebentägigen Krankenhausaufenthalt denkbar, 70 € an Zuzahlungen von den Kosten in den Daten für den Fall abzuziehen, so lange der Versicherte die Belastungsgrenze noch nicht erreicht hat bzw. grundsätzlich von Zuzahlungen befreit ist. Deutlich schwieriger ist hingegen die Ermittlung der Zuzahlungshöhe bei Arzneimitteln in den GKV-Routinedaten. Hierbei liegt die Höhe der Zuzahlungen bei 10 % des Apothekenabgabepreises, aber bei mindestens 5 € und bei maximal 10 €, wobei die Zuzahlungen nicht höher liegen als die tatsächlichen Kosten des Arzneimittels. Dabei ist zu beachten, dass der GKV-Spitzenverband bestimmte Arzneimittel von Zuzahlungen befreien kann, deren Abgabepreis 30 % niedriger als der jeweils gültige Festbetrag liegt. Genauere Informationen hierzu finden sich beim GKV-Spitzenverband. Problematischer für die Ermittlung der Zuzahlungen sind allerdings kassenindividuelle Zuzahlungsbefreiungen oder -reduzierungen, die Krankenkassen im Rahmen von Rabattverträgen ihren Versicherten gewähren können (§ 31 Abs. 3 S. 5 SGB V). Hierzu liegen, wie oben bereits angesprochen, in der Regel keine Informationen vor (GKV-Spitzenverband 2014a).

Generell ist bei der Feststellung darauf zu achten, dass Belastungsgrenzen für Zuzahlungen pro Kalenderjahr in der GKV existieren. Diese liegen bei 2 % der zu berücksichtigenden Bruttoeinnahmen zum Lebensunterhalt bzw. bei 1 % bei chronisch kranken Versicherten. Wird die individuelle Belastungsgrenze während eines Kalenderjahres erreicht, haben Versicherte Anspruch auf einen Befreiungsbescheid durch die Krankenkassen. Berücksichtigt werden hierbei sämtliche Zuzahlungen zu Leistungen der GKV. In 2010 waren rund 7 Mio. GKV-Versicherte (10 %) zuzahlungsbefreit aufgrund des Erreichens der Belastungsgrenze. Für 90 % davon war die Belastungsgrenze von 1 % maßgebend (Deutscher Bundestag). Darüber hinaus existieren

besondere Regelungen für Personengruppen, die z. B. die Hilfe zum Lebensunterhalt oder die Grundsicherung im Alter erhalten. Die Ermittlung der tatsächlichen Zuzahlungen in der GKV anhand von Routinedaten gestaltet sich daher in der Praxis als schwierig und es bleibt fragwürdig, ob hierzu verlässliche Schätzungen durchgeführt werden können.

Empfehlungen

- Bei der Ermittlung von Zuzahlungen sollte sich an den gesetzlichen Bestimmungen orientiert werden
- Es muss überprüft werden, welche Kosten von dem Dateneigner übermittelt wurden (Brutto- oder Nettokosten)
- Die Belastungsgrenzen bei Zuzahlungen sind zu beachten

5.5 Standardisierung

In der Regel werden GKV-Routinedatenanalysen aktuell anhand von Datensätzen einzelner Krankenkassen vollzogen. Die Ergebnisse dieser Analysen und die damit einhergehenden Implikationen sollen aber regelmäßig repräsentativ für z. B. die deutsche Gesamtbevölkerung oder die GKV-Versichertengemeinschaft ausgewiesen werden. Eine Problematik ergibt sich hierbei dadurch, dass die Versichertenkollektive der einzelnen Kassen mitunter z. B. hinsichtlich der Variablen Alter und Geschlecht deutlich von der Struktur der deutschen Gesamtbevölkerung abweichen können.

Um für diese Diskrepanzen zu adjustieren und allgemeingültige Aussagen zu treffen, können Verfahren zur Standardisierung der Ergebnisse genutzt werden. Neben verhältnismäßig einfachen und anschaulichen direkten Standardisierungsverfahren existiert auch eine Reihe von Modellen, die mithilfe statistischer Verfahren die ermittelten Ergebnisse standardisieren (Bajekal et al. 2004). Diese Verfahren kommen in der Regel dann zum Einsatz, wenn das interessierende Phänomen nicht mit ausreichender Genauigkeit gemessen werden kann, beispielsweise auf regionaler Ebene, da keine oder nur sehr wenige Fälle in dem jeweiligen Datensatz verfügbar sind. Bei kleinzelligen regionalen Analysen z. B. auf Kreisebene (aktuell existieren mehr als 400 Kreise in Deutschland) treten diese Probleme gehäuft auf, weshalb diese Verfahren auch manchmal unter dem Begriff „Small Area Estimation“ subsumiert werden (Heady et al. 2003; Fay und Herriot 1979). Melchior et al. haben vor diesem Hinter-

grund ihre Ergebnisse zu den regionalen Unterschieden in der Behandlung und Diagnostik von Depressionen mithilfe eines Small-Area-Verfahrens standardisiert und dabei weitere Hilfsvariablen wie die Arbeitslosenrate, das Einkommen, den Anteil von Personen ohne Schulabschluss und die Einwohnerdichte verwendet (Melchior et al. 2014).

Bei einer direkten Standardisierung werden regelmäßig nur potenzielle Unterschiede hinsichtlich der verschiedenen Alters- und Geschlechtsstrukturen ausgeglichen und somit Häufigkeiten (Raten) eines bestimmten Phänomens von einer Stichprobenpopulation auf eine Standard- oder Referenzpopulation abgeleitet. Die Referenzpopulation für deutsche GKV-Routinedatenstudien stellt dabei häufig die deutsche Gesamtbevölkerung oder GKV-Versichertengemeinschaft dar. Daten zur Alters- und Geschlechtsverteilung der deutschen Gesamtbevölkerung stehen beim Statistischen Bundesamt zur Verfügung. Die Gesundheitsberichterstattung des Bundes (gbe-bund) bietet hingegen Daten zur Struktur der GKV-Versicherten an (Bundesministerium für Gesundheit 2013).

Das konkrete methodische Vorgehen bei einer direkten Standardisierung ist wie folgt: Zunächst muss die zu standardisierende Rate für den vorliegenden Datensatz (Stichprobe) für die relevanten Alters- und Geschlechtsgruppen, z. B. im jeweiligen Kreis, ermittelt werden. Dazu ist es notwendig, dass ausreichend Daten vonseiten der jeweiligen Krankenkasse vorliegen (Alter und Geschlecht der Versicherten in den Kreisen). Die somit erhobene Rate je Altersgruppe und Geschlecht pro Kreis wird anschließend mit der Anzahl der Einwohner der Referenzpopulation (z. B. deutsche Gesamtbevölkerung, GKV-Versichertengemeinschaft) in der jeweiligen Alters- und Geschlechtsgruppe pro Kreis multipliziert und abschließend mit der gesamten Einwohnerzahl des Kreises dividiert (Melchior et al. 2014). Die resultierenden Raten sind nun hinsichtlich der Unterschiede in der Alters- und Geschlechtsstruktur zwischen der Stichprobenpopulation und einer Referenzpopulation standardisiert.

Empfehlungen

- Die Ergebnisse müssen mithilfe von geeigneten Verfahren standardisiert werden, um repräsentative Aussagen für die gesamtdeutsche Bevölkerung treffen zu können

6 Limitationen

GKV-Routinedaten weisen spezifische Limitationen auf, die bei der Studienplanung zu berücksichtigen sind (Zeidler und Braun 2012). In diesem Kapitel werden wesentliche Limitationen dargestellt, um mögliche Fehlinterpretationen zu vermeiden und Grenzen von GKV-Routinedatenanalysen aufzuzeigen. Ein Anspruch auf Vollständigkeit kann aufgrund der vielen möglichen Einschränkungen, die sich in der Regel nur umfassend im Kontext der jeweiligen Forschungsfrage beurteilen lassen, jedoch nicht erhoben werden.

Eine wesentliche Einschränkung ergibt sich aus der breit gefächerten Finanzierung von Gesundheitsleistungen in Deutschland. Neben den Krankenkassen sind weitere Sozialversicherungsträger und Institutionen, wie beispielsweise die Renten- und Unfallversicherung, aber auch die Versicherten selbst an der Finanzierung der Kosten einzelner Therapieoptionen beteiligt. Mit GKV-Routinedaten können nur Leistungen erfasst werden, die auch über die GKV abgerechnet wurden (Schubert et al. 2008). Die Aussagekraft und Vollständigkeit von GKV-Routinedaten ist daher immer dann eingeschränkt, wenn medizinische Leistungen nicht über die GKV abgerechnet werden (Zeidler und Braun 2012). Als nicht durch die GKV erfasste Leistungen können folgende Beispiele genannt werden:

- nicht abrechnungsfähige ambulante Arztkontakte wie IGeL-Leistungen oder Leistungen, die der Arzt aus unterschiedlichen Gründen nicht dokumentiert (Kerek-Bodden et al. 2005),
- nicht verschreibungspflichtige Arzneimittel wie Schmerzmittel, Vitamine oder Nahrungsergänzungsmittel (Weiß et al. 2010),
- Rehabilitationsmaßnahmen, die durch andere Sozialversicherungsträger (Rentenversicherung, Unfallversicherung etc.) finanziert werden.

Darüber hinaus schränken pauschalierte Vergütungssysteme wie das im stationären Sektor verwendete DRG-System, das ganze Leistungsbündel mit einer Fallpauschale vergütet, eine detaillierte Abbildung des Leistungsgeschehens ein (Bowles et al. 2011). Dies hat zur Folge, dass während eines Krankenhaus- oder Rehabilitationsaufenthaltes abgegebene Arznei-, Heil- und Hilfsmittel in der Regel nicht erfasst werden. Längsschnittanalysen zur Medikationsstrategie sind daher nur annahmenbasiert durchführbar. Eine Ausnahme bilden jedoch Leistungen, die explizit im OPS-Katalog

abgebildet sind und eine stationäre Verschreibung, auf Basis dieser OPS-Codes, an die Krankenkasse übermittelt wird. Dies gilt beispielsweise für die Applikation von Medikamenten wie die TNF-α-Hemmer Adalimumab, Etanercept und Infliximab.

Als weiteres Informationsdefizit von GKV-Routinedaten ist das Fehlen klinischer Informationen zu nennen. Befund- und Labordaten oder Daten zum Blutdruck der Patienten sind in den Abrechnungsdaten der Krankenkassen nicht erfasst (Schubert et al. 2008). Auch Informationen über den Schweregrad einer Erkrankung oder zur Lebensqualität der Patienten sind in GKV-Routinedaten in der Regel nicht abgebildet (Icks et al. 2010). Bei einzelnen Krankheitsbildern können jedoch anhand der ICD-Diagnosen Rückschlüsse über die Krankheitsschwere gezogen werden. So werden beispielsweise bei der Herzinsuffizienz anhand von NYHA-Stadien verschiedene Schweregrade systematisch als ICD-Diagnosen erfasst. Auch eine approximative Abbildung der Krankheitsschwere anhand spezifischer Leistungen, z. B. bestimmter Arzneimittelverordnungen oder Krankenhauseinweisungen, kann bei einzelnen Studien möglich sein. Weiterhin sind persönliche Informationen der Versicherten wie die Körpergröße, das Gewicht, die Lebensgewohnheiten und entsprechende Kontextfaktoren sowie die familiäre Disposition nicht standardmäßig in den GKV-Routinedaten erfasst. Diese Informationen sind bei einzelnen Analysen jedoch relevant, da beispielsweise einige Medikamente in Bezug auf das Gewicht des Patienten dosiert werden. Eine Möglichkeit zur zumindest partiellen Lösung dieser Herausforderung liegt in der Nutzung von DMP-Informationen, die in der Regel Informationen zur Körpergröße, zum Gewicht oder auch zum Raucherstatus der Versicherten enthalten. Diese Informationen liegen jedoch lediglich bei Versicherten vor, die auch in ein oder mehrere DMP-Programme eingeschrieben sind. Des Weiteren ist die Validität der DMP-Informationen umstritten (Horenkamp-Sonntag und Linder 2012; Horenkamp-Sonntag et al. 2012).

Im Bereich der ambulanten Arzneimittelverordnungen kann als Informationsdefizit insbesondere das Fehlen von Dosierungsinformationen genannt werden. So liegen in den GKV-Routinedaten keine Informationen darüber vor, über welchen Zeitraum und in welcher Dosierung der Patient das Medikament eingenommen hat. Daher können nur Aussagen über die Einlösung verschriebener Rezepte, jedoch nicht über die tatsächliche Adhärenz der Patienten, d. h. die Frage, ob die durch den behandelnden Arzt intendierte Behandlungsstrategie tatsächlich umgesetzt wurde, überprüft werden

(Weiß et al. 2010). Informationen über Rezepte, die durch den Arzt verschrieben, aber nicht durch den Patienten in der Apotheke eingelöst wurden, liegen den Krankenkassen nicht vor. Bei Analysen zur Einhaltung medikamentöser Leitlinien können daher nur eingeschränkte Aussagen über die Ursachen von empfohlenen, aber nicht durchgeführten Therapien getroffen werden, d. h. es bleibt unklar, ob der Arzt keine Verordnung ausgestellt hat oder ob diese nicht durch den Patienten eingelöst wurde.

Da die GKV-Routinedaten von einer Fülle an beteiligten Personengruppen (Ärzten, Apothekern, Krankenhäusern, Sanitätshäusern etc.) erhoben werden, ergeben sich naturbedingt Inkonsistenzen und fehlende Daten (Reinhold et al. 2011a). Dies kann sowohl Fehler zu Beginn der Dokumentationskette (z. B. ein fehlerhaft ausgestelltes Rezept) als auch in späteren Phasen der Datenerfassung beinhalten. Im Umgang mit fehlenden Daten ist generell eine Einzelprüfung zu empfehlen, d. h. es sollte geprüft werden, ob z. B. ein Datenersatz stattfinden soll (siehe auch Kapitel 5).

Die Diagnosen aus der ambulanten Versorgung werden nur quartalsbezogen dokumentiert, eine datumsgenaue Zuordnung von Diagnosen zu konkreten Behandlungsanlässen ist daher in der Regel nicht möglich (Schubert et al. 2008). Ein chronologischer Bezug zwischen Diagnosen und Leistungsvorgängen kann somit nicht immer abgebildet werden (Bowles et al. 2011). Die kausale Zuordnung einzelner Leistungen zu einer spezifischen Erkrankung ist dadurch erheblich eingeschränkt. Dies gilt insbesondere für Arzneimittel sowie Heil- und Hilfsmittel, die einen breiten Anwendungskontext besitzen. Aber auch eine Rekonstruktion der Abfolge der Besuche unterschiedlicher Fachärzte kann hierdurch eingeschränkt sein. Ursache dieser Einschränkung ist, dass diesen Leistungen in den GKV-Routinedaten standardmäßig keine expliziten Diagnosen zugeordnet werden.

Eine Unsicherheit besteht außerdem bezüglich der Validität von Diagnosen und Prozeduren (Swart und Ihle 2008). Die bei der Krankenkasse hinterlegten Informationen zum Gesundheitszustand einer Person sind in hohem Maße von der Diagnosecodierung der Leistungserbringer sowie von den zugrunde liegenden Informationssystemen abhängig. Sowohl Über-, Unter- als auch Fehlcodierungen sind möglich. Auch muss bei der Analyse und Interpretation der Daten immer berücksichtigt werden, dass die Dokumentation aufgrund des primären Abrechnungszwecks ökonomischen Anreizen der zugrunde liegenden Honorierungssysteme folgen kann (Reinhold et al. 2011a). Ist beispielsweise die Abrechnung einer spezifischen EBM-Ziffer aus Sicht

des behandelnden Arztes oder die Abrechnung einer spezifischen DRG aus Sicht des Krankenhauscontrollings ökonomisch sinnvoll, kann dies unter Umständen zu einer Fehlinterpretation der wissenschaftlichen Analyseergebnisse führen. Zusätzlich sind mögliche Reformen der Abrechnungs- und Honorarsysteme zu berücksichtigen, da diese häufig von ökonomischen Anreizen und Fehlanreizen begleitet sind. Bei jeder Studie sollte daher über eine Validierung der zugrunde liegenden Diagnosen nachgedacht werden (Schubert et al. 2010; siehe Kapitel 4).

Eine zusätzliche Limitation in Zusammenhang mit den Abrechnungsdiagnosen ist die unspezifische Diagnosecodierung (Hoffmann et al. 2008). Im ambulanten Bereich werden häufig unspezifische Diagnoseschlüssel verwendet. Hoffmann et al. konnten anhand von Demenzpatienten zeigen, dass in 59,7 % der Fälle der Schlüssel „Nicht näher bezeichnete Demenz“ (ICD-10: F03) abgerechnet wurde (Hoffmann et al. 2008). Als weitere Einschränkung kann die Vielfalt der Codierungsmöglichkeiten genannt werden (Hoffmann et al. 2008). In diesen Fällen kann die Lösung dann häufig nur in einer externen Validierung bzw. der Verknüpfung mit Primärdaten liegen.

Bei einer Nutzung von Arbeitsunfähigkeitsdaten ergeben sich ebenfalls Einschränkungen. Erstens müssen Arbeitsunfähigkeiten nur von Personen gemeldet werden, die sozialversicherungspflichtig beschäftigt sind (Vauth 2010). Rentner, Kinder und Jugendliche sowie Familienversicherte werden daher in der Regel nicht durch die Arbeitsunfähigkeitsdaten erfasst. Empfänger von Arbeitslosengeld müssen hingegen den Agenturen für Arbeit eine Arbeitsunfähigkeit melden und sind damit in den GKV-Routinedaten erfasst. Zweitens besteht keine Meldepflicht für Kurzzeitarbeitsunfähigkeit bis zu einer Dauer von drei Tagen (Bödeker 2005).

Das Datum einzelner Heilmittelsitzungen ist aus den GKV-Routinedaten in der Regel nicht detailliert ersichtlich. Es kann zwar erfasst werden, zu welchem Datum ein Heilmittel verordnet wurde, aber nicht, an welchen Tagen die Leistung tatsächlich durch den Versicherten in Anspruch genommen wurde.

Eine weitere Einschränkung ergibt sich bei der Abbildung von Todesfällen. Zum einen ist in den GKV-Routinedaten die Todesursache in der Regel nicht erfasst. Zum anderen wird bei vielen gesetzlichen Krankenkassen für Familienversicherte der Austrittsgrund „Tod“ überhaupt nicht dokumentiert, sodass eine Abgrenzung zu einem anderweitigen Austrittsgrund, wie z. B. einem Wechsel der Krankenkasse, nicht mög-

lich ist (Reinhold et al. 2011a). Daher empfiehlt sich bei der Verwendung dieser Information eine zusätzliche Datenvalidierung (z. B. mittels nationaler Sterbedaten).

Längsschnittanalysen können durch Versicherungswechsel zensiert und verzerrt sein, da seit der Öffnung fast aller Krankenkassen Eintritte und Austritte nahezu jederzeit möglich sind. Daher sollten alle Personen identifiziert werden, die im Beobachtungszeitraum die Krankenkasse verlassen haben. Anschließend kann die Personenzeit dieser Versicherten ermittelt werden und bei den Analysen entsprechend berücksichtigt werden (Reinhold et al. 2011a; siehe Kapitel 5). Dieses Vorgehen ist zur Vermeidung eines Selektionsbias erforderlich, da häufig nicht klar ist, warum diese Patienten die Krankenkasse verlassen haben. In Fällen, wo eine Kombination aus der Variable Kassenaustritt und dem Austrittsgrund „Tod“ vorzufinden ist, kann davon ausgegangen werden, dass die Ursache des Kassenaustritts auf das Versterben zurückgeführt werden kann (Reinhold et al. 2011a).

Eine weitere Einschränkung von GKV-Routinedaten liegt in der zeitlich begrenzten Datenverfügbarkeit. Häufig können aus datenschutzrechtlichen Gründen maximal Daten für einen Zeitraum von fünf Jahren zur Verfügung gestellt werden. Bei einzelnen Studien kann dieser Zeitraum zu kurz für eine umfassende und valide Beantwortung der Forschungsfragen sein.

Bei der Nutzung von Routinedaten einzelner Krankenkassen ist die Repräsentativität bezogen auf die GKV bzw. die Gesamtbevölkerung kritisch zu hinterfragen. Die Versichertenstruktur einer Krankenkasse kann im Hinblick auf die Alters- und Geschlechterverteilung sowie den Sozialstatus von der Gesamtbevölkerung abweichen (Icks et al. 2010). Die Entwicklung einer solchen Stammklientel bei einzelnen Krankenkassen ist historisch bedingt (Grobe und Ihle 2005). Die Übertragbarkeit und Generalisierbarkeit von Ergebnissen auf Basis der Routinedaten einzelner Krankenkassen kann daher eingeschränkt sein. Die Anwendung geeigneter statistischer Verfahren, wie beispielsweise eine direkte Alters- und Geschlechtsadjustierung, wird daher empfohlen (Grobe und Ihle 2005; Kapitel 5.5).

Empfehlungen

- Die vielfältigen Limitationen müssen bei GKV-Routinedatenstudie Berücksichtigung finden

Literatur

Abbas, S.; Ihle, P.; Köster, I.; Schubert, I. (2012): Estimation of Disease Incidence in Claims Data Dependent on the Length of Follow-Up: A Methodological Approach. In: *Health Services Research Journal* 47 (2), S. 746–755.

AGENS (2012): Gute Praxis Sekundärdatenanalyse (GPS). Leitlinien und Empfehlungen. 3. Fassung. Arbeitsgruppe Erhebung und Nutzung von Sekundärdaten (AGENS) der Deutschen Gesellschaft für Sozialmedizin und Prävention (DGSMP) und der Deutschen Gesellschaft für Epidemiologie (DGEpi).

Bajekal, M.; Scholes, S.; Pickering, K.; Purdon, S. (2004): Synthetic estimation of healthy lifestyles indicators: Stage 1 report. London.

Bankhofer, U. (1995): Unvollständige Daten- und Distanzmatrizen in der multivariaten Datenanalyse. Bergisch Gladbach, Köln: Eul (Reihe: quantitative Ökonomie, 64).

Barmer GEK (2010-2014): Report Krankenhaus. Hrsg. v. Barmer GEK. Online verfügbar unter https://presse.barmer-gek.de/barmer/web/Portale/Presseportal/-Subportal/Infothek/Studien-und-Reports/Report-Krankenhaus/Einstieg-Report-Krankenhaus.html?w-cm=CenterColumn_t261002.

Baxter, J. (1994): Is Husband's Class Enough? Class Location and Class Identity in the United States, Sweden, Norway, and Australia. In: *American Sociological Review* 59 (2), S. 220–235.

BDSG (2099): Bundesdatenschutzgesetz in der Fassung der Bekanntmachung vom 14. Januar 2003 (BGBl. I S. 66), das zuletzt durch Artikel 1 des Gesetzes vom 14. August 2009 (BGBl. I S. 2814) geändert worden ist.

Bödeker, W. (2005): Gesundheitsberichterstattung und Gesundheitsforschung mit Arbeitsunfähigkeitsdaten der Krankenkassen. In: Swart, E. und Ihle, P. (Hrsg.):Routinedaten im Gesundheitswesen – Handbuch Sekundärdatenanalyse: Grundlagen, Methoden und Perspektiven. 1. Aufl. Bern: Verlag Hans Huber, S. 57–78.

Bowles, D.; Damm, O.; Greiner, W. (2011): Gesundheitsbezogene Versorgungsforschung mit GKV-Routinedaten - Grenzen am Beispiel der Prophylaxe venöser Thromboembolien in der Hüft- und Kniegelenkendoprothetik. In: *Gesundheitsökonomie und Qualitätsmanagement* 16 (2), S. 96–107.

Bowles, D.; Wasiak, R.; Kissner, M.; van Nooten, F.; Engel, S.; Linder, R. et al. (2014): Economic burden of neural tube defects in Germany. In: *Public Health* 128 (3), S. 274–281.

BSHG (1999): Bundessozialhilfegesetz in der Fassung der Bekanntmachung vom 23. März 1994 (BGBl. I S. 646, 2975), zuletzt geändert durch Art. 4 des Gesetzes zur Familienförderung vom 22. Dezember 1999 (BGBl. I S. 2552).

BTMG (2013): Betäubungsmittelgesetz in der Fassung der Bekanntmachung vom 1. März 1994 (BGBl. I S. 358), das zuletzt durch Artikel 2 Absatz 20 u. Artikel 4 Absatz 7 des Gesetzes vom 7. August 2013 (BGBl. I S. 3154) geändert worden ist.

Bühl, A. (2012): SPSS 20. Einführung in die moderne Datenanalyse.

Bundesinstitut für Bau-, Stadt- und Raumforschung (BBSR) (2011a): Laufende Raumbeobachtung - Raumabgrenzungen. Kreise und Kreisregionen. Unter Mitarbeit von P. Kuhlmann. Bonn. Online verfügbar unter https://www.bbsr.bund.de/-BBSR/DE/Raumbeobachtung/Raumabgrenzungen/Kreise_Kreisregionen/kreise.html.

Bundesinstitut für Bau-, Stadt- und Raumforschung (BBSR) (2011b): Laufende Raumbeobachtung - Raumabgrenzungen. Raumtypen 2010. Unter Mitarbeit von T. Pütz. Bonn. Online verfügbar unter http://www.bbsr.bund.de/BBSR/DE/-Raumbeobachtung/Raumabgrenzungen/Raumtypen2010_vbg/-Raumtypen2010_alt.html.

Bundesministerium für Gesundheit (2013): Informationen rund um Mitglieder und Versicherte der GKV. GKV-Mitglieder, mitversicherte Angehörige, Beitragssätze und Krankenstand. Online verfügbar unter http://www.bmg.bund.de/krankenversicherung/zahlen-und-fakten-zur-krankenversicherung.html.

Cramer, J. A.; Roy, A.; Burrell, A.; Fairchild, C. J.; Fuldeore, M. J.; Ollendorf DA et al. (2008): Medication compliance and persistence. Terminology and definitions. In: *Value in health: the journal of the International Society for Pharmacoeconomics and Outcomes Research* 11 (1), S. 44–47.

DAK Forschung; IGES Institut GmbH (2013): DAK-Gesundheitsreport 2013. Unter Mitarbeit von Kordt, M.. Hamburg. Online verfügbar unter www.dak.de/dak/-download/Vollstaendiger_bundesweiter_Gesundheitsreport_2013-1318306.pdf.

Damm, K.; Lange, A.; Zeidler, J.; Braun, S.; Graf von der Schulenburg, J.-M. (2012): Einführung des neuen Tätigkeitsschlüssels und seine Anwendung in GKV-Routinedatenauswertungen. In: *Bundesgesundheitsblatt* 55 (2), S. 238–244.

DaTraGebV (2014): Datentransparenz-Gebührenverordnung vom 30. April 2014 (BGBl. I S. 458).

DaTraV (2012): Verordnung zur Umsetzung der Vorschriften über die Datentransparenz - Datentransparenzverordnung vom 10. September 2012 (BGBl. I S. 1895).

Deutscher Bundestag: Drucksache 17/8722. Bericht des Spitzenverbandes Bund der Krankenkassen zur Evaluation der Ausnahmeregelungen von der Zuzahlungspflicht. Online verfügbar unter dip21.bundestag.de/dip21/btd/17/087/1708722.pdf.

Deutscher Bundestag (1995): Wirkungen des Chipkarten-Einsatzes im Gesundheitswesen. Antwort der Bundesregierung auf die Kleine Anfrage der Abgeordneten Marina Steindor und der Fraktion BÜNDNIS 90/DIE GRÜNEN. Online verfügbar unter http://dipbt.bundestag.de/doc/btd/13/030/1303001.asc.

Deutsches Ärzteblatt (Hrsg.) (2011): Ambulante Kodierrichtlinien: Diagnosensicherheit und Seitenlokalisation. 108(6): A-271 / B-216 / C-216. Online verfügbar unter http://www.aerzteblatt.de/archiv/80824/Ambulante-Kodierrichtlinien-Diagnosensicherheit-und-Seitenlokalisation.

Deutsches Ärzteblatt (Hrsg.) (2014): Techniker Krankenkasse überrundet Barmer-GEK. Online verfügbar unter http://www.aerzteblatt.de/nachrichten/57202/Techniker-Krankenkasse-ueberrundet-Barmer-GEK.

DIMDI: ICD-10-GM. Hrsg. v. Deutsche Institut für Medizinische Dokumentation und Information. Online verfügbar unter https://www.dimdi.de/static/de/klassi/icd-10-gm/.

DIMDI (2013a): Informationssystem Versorgungsdaten (Datentransparenz). Hrsg. v. Deutsche Institut für Medizinische Dokumentation und Information. Online verfügbar unter http://www.dimdi.de/static/de/versorgungsdaten/index.htm, zuletzt aktualisiert am 12.02.14.

DIMDI (2013b): Morbi-RSA und Gesundheitsfonds. Hrsg. v. Deutsche Institut für Medizinische Dokumentation und Information. Online verfügbar unter https://www.dimdi.de/static/de/klassi/icd-10-gm/anwendung/zweck/morbi-rsa/index.htm, zuletzt aktualisiert am 22.08.2013.

DIMDI (2014a): Datensatzbeschreibung. Hrsg. v. Deutsche Institut für Medizinische Dokumentation und Information. Online verfügbar unter http://www.dimdi.de/static/-de/versorgungsdaten/datensatzbeschreibung/index.htm, zuletzt aktualisiert am 17.02.14.

DIMDI (2014b): G-DRG-System - Fallpauschalen in der stationären Versorgung. Hrsg. v. Deutsche Institut für Medizinische Dokumentation und Information. Online verfügbar unter http://www.dimdi.de/static/de/klassi/icd-10-gm/anwendung/zweck/g-drg/index.htm, zuletzt aktualisiert am 11.03.14.

DIMDI (2014c): Gebührenverordnung. Hrsg. v. Deutsche Institut für Medizinische Dokumentation und Information. Online verfügbar unter http://www.dimdi.de/static/-de/versorgungsdaten/gebuehrenverordnung.htm, zuletzt aktualisiert am 17.02.14.

Eberhard, S. (2013): Lassen sich GKV-Routinedaten nutzen, um auf leitliniengerechte Versorgung zu schließen? Eine Analyse am Beispiel der arteriellen Hypertonie. In: *Gesundheits- und Sozialpolitik: Zeitschrift für das gesamte Gesundheitswesen* 67, S. 29–36.

Elm, E. von; Altmann, D. G.; Egger, M.; Pocock, S. C.; Gøtzsche, P. C.; Vandenbroucke, J. P. (2008): Das Strengthening the Reporting of Observational Studies in Epidemiology (STROBE-) Statement. In: *Internist* 49 (6), S. 688–693.

EntgFG (2012): Entgeltfortzahlungsgesetz vom 26. Mai 1994 (BGBl. I S. 1014, 1065), das zuletzt durch Artikel 1a des Gesetzes vom 21. Juli 2012 (BGBl. I S. 1601) geändert worden ist.

Fahrmeir, L. (2010): Statistik. Der Weg zur Datenanalyse. 7., neu bearb. Aufl. Berlin, Heidelberg: Springer (Springer-Lehrbuch).

Fay, R. E.; Herriot, R. A. (1979): Estimates of Income for Small Places: An Application of James-Stein Procedures to Census Data. In: *Journal of the American Statistical Association 74* (366), S. 269–277.

Ford, B. L. (1976): Missing data procedures: A comparative study. Hrsg. v. Statistical Reporting Service, U. S. Department of Agriculture. Washington DC.

Frank, P. M. (1976): Empfindlichkeitsanalyse dynamischer Systeme. Eine einführende Darstellung. München [u.a.]: Oldenbourg (Methoden der Regelungstechnik).

Frey, S.; Linder, R.; Juckel, G.; Stargardt, T. (2013): Cost-effectiveness of long-acting injectable risperidone versus flupentixol decanoate in the treatment of schizophrenia: a Markov model parameterized using administrative data. In: *The European Journal of Health Economics* 15 (2), S. 133-42. Epub 2013 Feb 19.

Frey, S.; Stargardt, T. (2012): Performance of Compliance and Persistence Measures in Predicting Clinical and Economic Outcomes Using Administrative Data from German Sickness Funds. In: *Pharmacotherapy* 32 (10), S. 880–889.

Garbe, E. (2008): Nutzung von Sekundärdaten für ein Versorgungsmonitoring: zur Notwendigkeit einer Validierung. In: Fuchs, C.; Kurth, B. M. und Scriba, P. C. (Hrsg.): Report Versorgungsforschung. 1 Band. Köln, S. 49–56.

gbe-bund (2012): Haupt- und Nebendiagnose. Online verfügbar unter http://www.gbe-bund.de/gbe10/abrechnung.prc_abr_test_logon?p_uid=gasts&p_aid=&p_knoten=FID&p_sprache=D&p_suchstring=11115::nebendiagnose, zuletzt aktualisiert am 24.10.2012.

GG (2012): Grundgesetz für die Bundesrepublik Deutschland in der im Bundesgesetzblatt Teil III, Gliederungsnummer 100-1, veröffentlichten bereinigten Fassung, das zuletzt durch Artikel 1 des Gesetzes vom 11. Juli 2012 (BGBl. I S. 1478) geändert worden ist.

GKV-Datenaustausch (a): Online verfügbar unter http://www.gkv-datenaustausch.de/startseite/startseite.jsp.

GKV-Datenaustausch (b): Apotheken. Online verfügbar unter http://www.gkv-datenaustausch.de/leistungserbringer/apotheken/apotheken.jsp.

GKV-Datenaustausch (c): Ärzte. Online verfügbar unter http://www.gkv-datenaustausch.de/leistungserbringer/aerzte/aerzte.jsp.

GKV-Datenaustausch (d): Krankenhäuser. Online verfügbar unter http://www.gkv-datenaustausch.de/leistungserbringer/krankenhaeuser/krankenhaeuser.jsp.

GKV-Datenaustausch (e): Reha-Einrichtungen. Online verfügbar unter http://www.gkv-datenaustausch.de/leistungserbringer/reha_einrichtungen/reha_einrichtungen.jsp.

GKV-Datenaustausch (f): Einführung der 8-stelligen PZN. Online verfügbar unter http://www.gkv-datenaustausch.de/media/dokumente/leistungserbringer_1/-apotheken/technische_anlagen_aktuell/Einfuehrung_der_8-stelligen_-PZN_1_3_0.pdf.

GKV-Modernisierungsgesetz (GMG) (2003): Gesetz zur Modernisierung der gesetzlichen Krankenversicherung.

GKV-Spitzenverband (2012): Morbiditätsorientierter RSA (Morbi-RSA). Online verfügbar unter http://www.gkv-spitzenverband.de/krankenversicherung/krankenversicherung_grundprinzipien/finanzierung/rsa/rsa.jsp, zuletzt aktualisiert am 15.06.2012.

GKV-Spitzenverband (2014a): Befreiungsliste Arzneimittel. Online verfügbar unter http://www.gkv-spitzenverband.de/service/versicherten_service/zuzahlungen_-und_befreiungen/befreiungsliste_arzneimittel/befreiungsliste_arzneimittel.jsp, zuletzt aktualisiert am 06.01.2014.

GKV-Spitzenverband (2014b): Grafik zu den Veränderungen bei der Krankenkassenanzahl. Online verfügbar unter http://www.gkv-spitzenverband.de/krankenversicherung/krankenversicherung_grundprinzipien/alle_gesetzlichen_krankenkassen/alle_gesetzlichen_krankenkassen.jsp, zuletzt aktualisiert am 06.01.2014.

Graf von der Schulenburg, J.-M.; Greiner, W.; Jost, F.; Klusen, N.; Kubin, M.; Leidl, R. et al. (2007): Deutsche Empfehlungen zur gesundheitsökonomischen Evaluation - dritte und aktualisierte Fassung des Hannoveraner Konsens. In: *Gesundheitsökonomie und Qualitätsmanagement* 12 (5), S. 285–290.

Greiner, W.; Damm, O. (2012): Die Berechnung von Kosten und Nutzen. In: Schöffski, O. und Graf von der Schulenburg, J.-M. (Hrsg.): Gesundheitsökonomische Evaluationen. 4. Aufl. Berlin, Heidelberg: Springer-Verlag, S. 23–42.

Grobe, T. G. (2005): Stationäre Versorgung - Krankenhausbehandlungen. In: Swart, E. und Ihle, P. (Hrsg.): Routinedaten im Gesundheitswesen – Handbuch Sekundärdatenanalyse: Grundlagen, Methoden und Perspektiven. 1. Aufl. Bern: Verlag Hans Huber, S. 79–98.

Grobe, T. G. (2008): Arbeiten mit Daten der Gmünder Ersatzkasse. In: *Bundesgesundheitsblatt* 51 (10), S. 1106–1117.

Grobe, T. G.; Ihle, P. (2005): Versichertenstammdaten und sektorübergreifende Analyse. In: Swart, E. und Ihle, P. (Hrsg.): Routinedaten im Gesundheitswesen – Handbuch Sekundärdatenanalyse: Grundlagen, Methoden und Perspektiven. 1. Aufl. Bern: Verlag Hans Huber, S. 17–34.

Gutenbrunner, C.; Glaesener, J.-J. (2007): Rehabilitation, Physikalische Medizin und Naturheilverfahren; mit 57 Tabellen. [Online-Ausg.]. Heidelberg: Springer (SpringerLink: Springer e-Books).

Harnischmacher, U.; Ihle, P.; Berger, B.; Goebel, J. W.; Scheller, J. (2006): Checkliste und Leitfaden zur Patienteneinwilligung. Grundlagen und Anleitung für die klinische Forschung. Berlin: Med.-Wiss. Verl.-Ges (Schriftenreihe der Telematikplattform für Medizinische Forschungsnetze, Bd. 3).

Hase, F. (2011): Forschung mit Sozialdaten. In: *Datenschutz und Datensicherheit* 35 (12), S. 875–878.

Heady, P.; Clarke, P.; Brown, P.; Ellis, K.; Heasman, D.; Hennell, S.; Mitchell, B. (2003): Model- Based Small Area Estimation Series No. 2: Small Area Estimation Project Report. Norwich: National Statistics UK.

Heller, G.; Günster, C.; Misselwitz, B.; Feller, A.; Schmidt, S. (2007): Jährliche Fallzahl pro Klinik und Überlebensrate sehr untergewichtiger Frühgeborener (VLBW) in Deutschland - Eine bundesweite Analyse mit Routinedaten. In: *Zeitschrift für Geburtshilfe und Neonatologie* 211 (3), S. 123–131.

Hendricks, V.; Schmidt, S.; Vogt, A.; Gysan, D.; Latz, V.; Schwang, I. et al. (2014): Case Management Program for Patients With Chronic Heart Failure: Effectiveness in Terms of Mortality, Hospital Admissions and Costs. In: *Deutsches Ärzteblatt international* 111(15): S. 264-270.

Hennessy, S. (2006): Use of Health Care Databases in Pharmacoepidemiology. In: *Basic and Clinical Pharmacology and Toxicology* 98 (3), S. 311–313.

Hoffmann, F. (2009): Review on use of German health insurance medication claims data for epidemiological research. In: *Pharmacoepidemiology and Drug Safety* 18 (5), S. 349–356.

Hoffmann, F.; Andersohn, F.; Giersiepen, K.; Scharnetzky, E.; Garbe, E. (2008): Validierung von Sekundärdaten. Grenzen und Möglichkeiten. In: *Bundesgesundheitsblatt* 51 (10), S. 1118–1126.

Hoffmann, F.; Glaeske, G. (2011): Analyse von Routinedaten. In: Pfaff, H. (Hrsg.): Lehrbuch Versorgungsforschung. Systematik - Methodik - Anwendung; mit 19 Tabellen. Stuttgart: Schattauer, S. 317–322.

Hoffmann, F.; Icks, A. (2012): Unterschiede in der Versichertenstruktur von Krankenkassen und deren Auswirkungen für die Versorgungsforschung: Ergebnisse des Bertelsmann-Gesundheitsmonitors. In: *Gesundheitswesen* 74 (5), S. 291–297.

Hoffmann, W.; Maaz, A.; Nordheim, J.; Winter, M.; Kuhlmey, A. (2004): Chronisch krank werden im Alter – zur Abschätzung von Inzidenz und Prävalenz mittels Routinedaten einer Betriebskrankenkasse. In: *Gesundheitswesen*, S. 66–80.

Holle, R.; Behrend, C.; Reitmeir, P.; John, J. (2005): Methodenfragen der Nutzung von GKV-Routinedaten für Kostenanalysen. In: Swart, E. und Ihle, P. (Hrsg.): Routinedaten im Gesundheitswesen – Handbuch Sekundärdatenanalyse: Grundlagen, Methoden und Perspektiven. 1. Aufl. Bern: Verlag Hans Huber, S. 301–318.

Horenkamp-Sonntag, D.; Linder, R. (2012): Untersuchungen zur externen Validität der DMP-Dokumentation. In: Roski, R. (Hrsg.): Disease Management Programme. Statusbericht 2012; MVF-Fachkongresse "10 Jahre DMP" und "Versorgung 2.0". Bonn: eRelation Content in Health (Schriftenreihe Monitor Versorgungsforschung), S. 227–231.

Horenkamp-Sonntag, D.; Linder, R.; Ahrens, S.; Verheyen, F. (2012): Externe Validität von DMP-Doku-Bögen im Abgleich mit GKV-Routinedaten: Wie valide werden Arzneimittel-Therapien und stationäre Notfalleinweisungen von DMP-Ärzten dokumentiert? Online verfügbar unter http://www.tk.de/tk/vortraege/vortraege-aktuell/448568.

Icks, A.; Chernyak, N.; Bestehorn, K.; Brüggenjürgen, B.; Bruns, J.; Damm, O. et al. (2010): Methoden der gesundheitsökonomischen Evaluation in der Versorgungsforschung. In: *Gesundheitswesen* 72 (12), S. 917–933.

IGES Institut GmbH (03.12.2012): Bewertung der Kodierqualität von vertragsärztlichen Diagnosen. Eine Studie im Auftrag des GKV-Spitzenverbands in Kooperation mit der BARMER GEK. Berlin.

Ihle, P. (2008): Datenschutzrechtliche und methodische Aspekte beim Aufbau einer Routinedatenbasis aus der Gesetzlichen Krankenversicherung zu Forschungszwecken. In: *Bundesgesundheitsblatt* 51 (10), S. 1127–1134.

Ihle, P.; Köster, I.; Herholz, H.; Rambow-Bertram, P.; Schardt, T.; Schubert, I. (2005): Versichertenstichprobe AOK Hessen/KV Hessen - Konzeption und Umsetzung einer personenbezogenen Datenbasis aus der Gesetzlichen Krankenversicherung. In: *Gesundheitswesen* 67 (08/09), S. 638–645.

IMVR; WINEG: Projektdatenbank Versorgungsforschung Deutschland. Instituts für Medizinsoziologie, Versorgungsforschung und Rehabilitationswissenschaft der Universität zu Köln; Wissenschaftlichen Instituts der TK für Nutzen und Effizienz im Gesundheitswesen. Online verfügbar unter http://www.versorgungsforschung-deutschland.de/.

Institut des Bewertungsausschusses: Einheitlicher Bewertungsmaßstab (EBM). Online verfügbar unter http://www.institut-des-bewertungsausschusses.de/ba/ebm.html.

Jaunzeme, J.; Muschik, D. (2014): Stichtag oder Versicherungsdauer als Selektionskriterium der Versicherten für die Analyse von GKV-Daten. AGENS-Mehtodenworkshop 2014. Hannover, 13.02.2014. Online verfügbar unter www.mh-hannover.de/fileadmin/institute/med_soziologie/Dokumente/AGENS2014_-Abstractband.pdf.

KBV (2008): BAR-Schlüsselverzeichnis, Anlage 35. Zweistellige Fachgruppencodierung für die 8. und 9. Stelle der LANR. Kassenärztliche Bundesvereinigung (KBV). Online verfügbar unter http://applications.kbv.de/keytabs/ita/schluesseltabellen.asp?page=S_BAR2_WBO_V1.07.htm.

KBV (2011a): Ambulante Kodierrichtlinien: Diagnosensicherheit und Seitenlokalisation. In: *Deutsches Ärzteblatt international* 108 (6), S. A-271-A-273. Online verfügbar unter http://www.aerzteblatt.de/int/article.asp?id=80824.

KBV (2011b): ICD-10-GM: Wesentliche Regeln für den vertragsärztlichen Bereich. Online verfügbar unter http://www.kbv.de/html/2007.php.

Kelm, S. (2012): Wie lange ist mein Rezept gültig? Hrsg. v. Apotheken-Umschau. Online verfügbar unter http://www.apotheken-umschau.de/Medikamente/Wie-lange-ist-mein-Rezept-gueltig-192477.html.

Kerek-Bodden, H.; Heuer, J.; Brenner, G.; Koch, H.; Lang, A. (2005): Morbiditäts- und Inanspruchnahmeanalysen mit personenbezogenen Abrechnungsdaten aus Arztpraxen. In: Swart, E. und Ihle, P. (Hrsg.): Routinedaten im Gesundheitswesen – Handbuch Sekundärdatenanalyse: Grundlagen, Methoden und Perspektiven. 1. Aufl. Bern: Verlag Hans Huber, S. 35–55.

KHG (2013): Krankenhausfinanzierungsgesetz in der Fassung der Bekanntmachung vom 10. April 1991 (BGBl. I S. 886), das zuletzt durch Artikel 5c des Gesetzes vom 15. Juli 2013 (BGBl. I S. 2423) geändert worden ist. In: BGBl. I S. 886.

Köster, I.; Ihle, P.; Schubert, I. (2011): Zwischenbericht 2004-2008 für Gesundes Kinzigtal GmbH hier: LKK-Daten. PMV Forschungsgruppe. Köln. Online verfügbar unter http://www.gesundes-kinzigtal.de/media/documents/KIT-PMV-%C3%9CUF_-LKK-fin-2011-08-10.pdf.

Krüger-Brand, H. E. (2013): Datentransparenz: Einblick ins Versorgungsgeschehen. In: *Deutsches Ärzteblatt international* 110 (4), S. A-120-A-121. Online verfügbar unter http://www.aerzteblatt.de/int/article.asp?id=134211.

KV Berlin: Einheitlicher Bewertungsmaßstab EBM. Online verfügbar unter http://www.kvberlin.de/20praxis/30abrechnung_honorar/10ebm/.

L'hoest, H.; Marschall, U. (2013): Ist häufiger besser und weniger teurer? Eine Datenanalyse zur Organtransplantation. In: Repschläger, U.; Schulte, C. und Osterkamp, N. (Hrsg.): Gesundheitswesen aktuell 2013. Beiträge und Analysen. 1. Aufl. Wuppertal: Barmer GEK, S. 248–269. Online verfügbar unter http://www.barmer-gek.de/barmer/web/Portale/Versicherte/Rundum-gutversichert/Infothek/-Wissenschaft-Forschung/Publikationen/Gesundheitswesen-aktuell-2013/-Gesundheitswesen-aktuell-2013-Marschall-Organtransplantation,property=Data.pdf.

Lange, S.; Bender, R. (2007): Median oder Mittelwert? In: *Deutsche medizinische Wochenschrift* 132 (S 01), S. e1.

LAUER-Taxe. Online verfügbar unter http://www.lauer-fischer.de/lf/Seiten/WEBAPO-Lauer-Taxe/WEBAPO-Lauer-Taxe-demo.aspx.

Laux, G.; Nothacker, M.; Weinbrenner, S.; Störk, S.; Blozik, E.; Peters-Klimm, F. et al. (2011): Nutzung von Routinedaten zur Einschätzung der Versorgungsqualität: Eine kritische Beurteilung am Beispiel von Qualitätsindikatoren für die „Nationale Versorgungsleitlinie Chronische Herzinsuffizienz". In: *Zeitschrift für Evidenz, Fortbildung und Qualität im Gesundheitswesen* 105 (1), S. 21–26.

Little, R. J. A.; Rubin, D. B. (2002): Statistical analysis with missing data. 2nd ed. Hoboken, N.J.: Wiley (Wiley series in probability and statistics).

Majeed, R.; Corvinus, U.; Weismüller, K.; Röhrig, R.; Harnischmacher, U.; Ihle, P. (2007): Computerunterstützte Erstellung von Patienteneinwilligungen – eine webbasierte Navigation durch die Checkliste Patienteneinwilligung. Kongress Medizin und Gesellschaft. Augsburg, 17.–21.9.2007. Hrsg. v. German Medical Science GMS Publishing House. Düsseldorf (Doc 07gmds619). Online verfügbar unter http://www.egms.de/static/de/meetings/gmds2007/07gmds619.shtml.

Mansky, T.; Robra, B.-P.; Schubert, I. (2012): Vorhandene Daten besser nutzen Für die sektorübergreifende Zusammenführung medizinischer Routinedaten sollten die Krankenkassen zur Lieferung bereits vorliegender Daten verpflichtet werden. In: *Deutsches Ärzteblatt* 109 (21), S. A1082-A1085.

Meinck, M.; Lübke, N.; Polak, U. (2014): Rehabilitation vor Pflegebedürftigkeit im Alter: eine Analyse anhand von Routinedaten. In: *Rehabilitation* 53 (2), S. 74–80.

Melchinger, H. (2008): Strukturfragen der ambulanten psychiatrischen Versorgung. unter besonderer Berücksichtigung von Psychiatrischen Institutsambulanzen und der sozialpsychiatrischen Versorgung außerhalb der Leistungspflicht der Gesetzlichen Krankenversicherung. Medizinische Hochschule Hannover. Hannover.

Melchior, H.; Schulz, H.; Härter, M. (2014): Faktencheck Gesundheit Regionale Unterschiede in der Diagnostik und Behandlung von Depressionen. Unter Mitarbeit von Walker, J. und Ganninger, M.. Hrsg. v. Bertelsmann Stiftung. Online verfügbar unter https://faktencheck-gesundheit.de/fileadmin/daten_fcd/Dokumente/faktencheck_-depression_studie.pdf.

Müller, W. (2012): Informationssystem "Datentransparenz" bei DIMDI im Aufbau. Arbeitsgemeinschaft der Wissenschaftlichen Medizinischen Fachgesellschaften (AWMF). Düsseldorf. Online verfügbar unter http://www.egms.de/static/en/journals/-awmf/2012-9/awmf000268.shtml.

Müller-Benedict, V. (2007): Grundkurs Statistik in den Sozialwissenschaften. Eine leicht verständliche, anwendungsorientierte Einführung in das sozialwissenschaftlich notwendige statistische Wissen. 4., überarb. Aufl. Wiesbaden: VS, Verl. für Sozialwiss. (Lehrbuch).

Müller-Bergfort, S.; Fritze, J. (2007): Diagnose- und Prozedurendaten im deutschen DRG-System. In: *Bundesgesundheitsblatt* 50 (8), S. 1047–1054.

Muschik, D.; Jaunzeme, J. (2014): Übertragung des Bildungsstandes von Haupt- auf Familienversicherte bei der Analyse von GKV-Daten. AGENS-Mehtodenworkshop 2014. Hannover, 13.02.2014. Online verfügbar unter www.mh-hannover.de/-fileadmin/institute/med_soziologie/Dokumente/AGENS2014_Abstractband.pdf.

Nink, K.; Schröder, H.; Schubert, I. (2005): Arzneimittel. In: Swart, E. und Ihle, P. (Hrsg.): Routinedaten im Gesundheitswesen – Handbuch Sekundärdatenanalyse: Grundlagen, Methoden und Perspektiven. 1. Aufl. Bern: Verlag Hans Huber, S. 99–122.

NVL (2012): Nationale Versorgungsleitlinie Chronische Herzinsuffizienz – Langfassung. Unter Mitarbeit von Bundesärztekammer (BÄK), Kassenärztliche Bundesvereinigung (KBV) und Arbeitsgemeinschaft der Wissenschaftlichen Medizinischen Fachgesellschaften (AWMF).

Ohlmeier, C.; Niemeyer, M.; Garbe, E.; Mikolajczyk, R. (2012): Identifizierung von Todesursachen in Daten der Gesetzlichen Krankenversicherung am Beispiel des Lungen- und Pankreaskrebs. 57. Jahrestagung der Deutschen Gesellschaft für Medizinische Informatik, Biometrie und Epidemiologie e.V. (GMDS)., 16.-20.09.2012. Düsseldorf. Hrsg. v. German Medical Science GMS Publishing House. Braunschweig (Doc12gmds177). Online verfügbar unter http://www.egms.de/static/en/meetings/-gmds2012/12gmds177.shtml.

Pirk, O.; Schöffski, O. (2012): Primärdatenerhebung. In: Schöffski, O. und Graf von der Schulenburg, J.-M. (Hrsg.): Gesundheitsökonomische Evaluationen. 4. Aufl. Berlin, Heidelberg: Springer-Verlag, S. 197–242.

Prenzler, A.; Zeidler, J.; Braun, S.; Graf von der Schulenburg, J.-M. (2010): Bewertung von Ressourcen im Gesundheitswesen aus der Perspektive der deutschen Sozialversicherung. In: *PharmacoEconomics German Research Articles* 8 (1), S. 47–66.

REHADAT: Hilfsmittel - Versorgungsablauf. Online verfügbar unter http://www.rehadat-hilfsmittelportal.de/de/infothek/versorgungsablauf/index.html.

Reinboth, C. (2006): Multivariate Analyseverfahren in der Marktforschung. Hochschule Harz.

Reinhold, T.; Andersohn, F.; Hessel, F.; Brüggenjürgen, B.; Willich, S. N. (2011a): Die Nutzung von Routinedaten der gesetzlichen Krankenkassen (GKV) zur Beantwortung gesundheitsökonomischer Fragestellungen – eine Potenzialanalyse. In: *Gesundheitsökonomie und Qualitätsmanagement* 16 (3), S. 153–159.

Reinhold, T.; Lindig, C.; Willich, S. N.; Brüggenjürgen, B. (2011b): The costs of atrial fibrillation in patients with cardiovascular comorbidities--a longitudinal analysis of German health insurance data. In: *Europace* 13 (9), S. 1275–1280.

Reis, A. (2005): Krankheitskostenanalysen. In: Swart, E. und Ihle, P. (Hrsg.): Routinedaten im Gesundheitswesen – Handbuch Sekundärdatenanalyse: Grundlagen, Methoden und Perspektiven. 1. Aufl. Bern: Verlag Hans Huber, S. 291–300.

Rousseeuw, P. J.; Leroy, A. M. (1987): Robust Regression and Outlier Detection. Hoboken, NJ, USA: John Wiley & Sons, Inc.

Rubin, D. B. (1976): Inference and missing data. In: *Biometrika* 63 (3), S. 581–592.

Runte, M. (1999): Missing Values. Konzepte und statistische Literatur. Kiel.

Schader, M.; Gaul, W. (1992): The MVL (Missing Values Linkage) Approach for Hierarchical Classification when Data are Incomplete. In: Bock, H. H.; Opitz, O. und Schader, M. (Hrsg.): Analyzing and Modeling Data and Knowledge. Berlin, Heidelberg: Springer Berlin Heidelberg (Studies in Classification, Data Analysis, and Knowledge Organization), S. 107–115.

Scharnetzky, E.; Busch, H.; Wobbe, S.; Rebscher, H. (2013): Versorgungsforschung aus der Perspektive einer Gesetzlichen Krankenkasse. In: *Gesundheitsökonomie und Qualitätsmanagement* 18 (6), S. 290–294.

Schnell, R. (1986): Missing-Data-Probleme in der empirischen Sozialforschung. Bochum.

Schöffski, O. (2012): Grundformen gesundheitsökonomischer Evaluationen. In: Schöffski, O. und Graf von der Schulenburg, J.-M. (Hrsg.): Gesundheitsökonomische Evaluationen. 4. Aufl. Berlin, Heidelberg: Springer-Verlag, S. 43–70.

Schreyögg, J.; Stargardt, T. (2012): Gesundheitsökonomische Evaluation auf Grundlage von GKV-Routinedaten. In: *Bundesgesundheitsblatt* 55 (5), S. 668–676.

Schröder, H.; Schwinger, A.; Waltersbach, A. (2005): Heilmittel. In: Swart, E. und Ihle, P. (Hrsg.): Routinedaten im Gesundheitswesen – Handbuch Sekundärdatenanalyse: Grundlagen, Methoden und Perspektiven. 1. Aufl. Bern: Verlag Hans Huber, S. 123–135.

Schubert, I.; Ihle, P.; Köster, I. (2010): Interne Validierung von Diagnosen in GKV-Routinedaten: Konzeption mit Beispielen und Falldefinition. In: *Gesundheitswesen* 72 (6), S. 316–322.

Schubert, I.; Köster, I.; Küpper-Nybelen, J.; Ihle, P. (2008): Versorgungsforschung mit GKV-Routinedaten. In: *Bundesgesundheitsblatt* 51 (10), S. 1095–1105.

Schwab, G. (1991): Fehlende Werte in der angewandten Statistik. Wiesbaden: Dt. Univ.-Verl. (DUV: Wirtschaftswissenschaft).

SGB X (2013): Das Zehnte Buch Sozialgesetzbuch – Sozialverwaltungsverfahren und Sozialdatenschutz – in der Fassung der Bekanntmachung vom 18. Januar 2001 (BGBl. I S. 130), das zuletzt durch Artikel 6 des Gesetzes vom 25. Juli 2013 (BGBl. I S. 2749) geändert worden ist.

SGB IX (2012): Das Neunte Buch Sozialgesetzbuch – Rehabilitation und Teilhabe behinderter Menschen – (Artikel 1 des Gesetzes vom 19. Juni 2001, BGBl. I S. 1046, 1047), das zuletzt durch Artikel 3 des Gesetzes vom 14. Dezember 2012 (BGBl. I S. 2598) geändert worden ist. Online verfügbar unter http://www.gesetze-im-internet.de/sgb_5/.

SGB V (2014): Das Fünfte Buch Sozialgesetzbuch – Gesetzliche Krankenversicherung – (Artikel 1 des Gesetzes vom 20. Dezember 1988, BGBl. I S. 2477, 2482), das zuletzt durch Artikel 1 des Gesetzes vom 27. März 2014 (BGBl. I S. 261) geändert worden ist. Online verfügbar unter http://www.gesetze-im-internet.de/sgb_5/.

Statistisches Bundesamt (2012): Bevölkerungsstand: Bevölkerung nach Geschlecht, regionale Tiefe: Kreise und krfr. Städte. Stichtag 31.12. Online verfügbar unter https://www.regionalstatistik.de/genesis/online;jsessionid=E427AC1486DD49D7A00D3F94D2979E62?sequenz=tabelleErgebnis&selectionname=173-01-4.

SVR (2002): Gutachten 2000/2001: Bedarfsgerechtigkeit und Wirtschaftlichkeit. Band I: Zielbildung, Prävention, Nutzerorientierung und Partizipation. 1. Aufl. Baden-Baden: Nomos-Verl.-Ges. (Gutachten / Sachverständigenrat für die Konzertierte Aktion im Gesundheitswesen, 2000/01).

Swart, E. (2005a): Kleinräumige Versorgungsforschung mit GKV-Routinedaten. In: Swart, E. und Ihle, P. (Hrsg.): Routinedaten im Gesundheitswesen – Handbuch Sekundärdatenanalyse: Grundlagen, Methoden und Perspektiven. 1. Aufl. Bern: Verlag Hans Huber, S. 243–252.

Swart, E. (2005b): Über-, Unter- und Fehlversorgung in der stationären Versorgung – Welche Rückschlüsse lassen sich aus GKV-Routinedaten ziehen? In: Swart, E. und Ihle, P. (Hrsg.): Routinedaten im Gesundheitswesen – Handbuch Sekundärdatenanalyse: Grundlagen, Methoden und Perspektiven. 1. Aufl. Bern: Verlag Hans Huber, S. 253–262.

Swart, E.; Deh, U.; Robra, B.-P. (2008): Die Nutzung der GKV-Daten für die kleinräumige Analyse und Steuerung der stationären Versorgung. In: *Bundesgesundheitsblatt* 51 (10), S. 1183–1192.

Swart, E.; Ihle, P. (2008): Der Nutzen von GKV-Routinedaten für die Versorgungsforschung. In: *Bundesgesundheitsblatt* 51 (10), S. 1093–1094.

Swart, E.; Schmitt, J. (2014): STROSA - Ein Berichtsstandard für Sekundärdatenanalyse. AGENS-Mehtodenworkshop 2014. Hannover, 13.02.2014. Online verfügbar unter www.mh-hannover.de/fileadmin/institute/med_soziologie/Dokumente/-AGENS2014_Abstractband.pdf.

Swart, E.; Willer, C. (2012): Lässt sich die Umsetzung ärztlicher Leitlinien anhand von GKV-Routinedaten überprüfen? In: *Gesundheitswesen* 74 (08/09).

Tiedt, G. (1996): Rechtliche Grundlagen der Rehabilitation. In: Delbrück, H. und Haupt, E. (Hrsg.): Rehabilitationsmedizin: Therapie- und Betreuungskonzepte bei chronischen Krankheiten. München, Wien, Baltimore: Urban & Schwarzenberg, S. 27–50.

Ultsch, B.; Köster, I.; Reinhold, T.; Siedler, A.; Krause, G.; Icks, A. et al. (2013): Epidemiology and cost of herpes zoster and postherpetic neuralgia in Germany. In: *The European Journal of Health Economics* 14 (6), S. 1015–1026.

Vauth, C. (2010): Gesundheitsökonomische Sekundärforschung: Das Beispiel der Bewertung stark wirksamer Analgetika in der chronischen Schmerztherapie. 1. Aufl. Baden-Baden: Nomos Verlagsgesellschaft mbH & Co. KG (Beiträge zum Gesundheitsmanagement, 29).

Völzke, H.; Alte, D.; Schmidt, C. O.; Radke, D.; Lorbeer, R.; Friedrich, N. et al. (2011): Cohort profile: the study of health in Pomerania. In: *International Journal of Epidemiology* 40 (2), S. 294–307.

Weiß, F.; Vietor, C.; Hecke, T. L. (2010): Verwendung von Routinedaten zu Evaluationszwecken in Krankenkassen – die Evaluation des TK-Patientendialog. In: *Gesundheitswesen* 72 (6), S. 371–378.

Werner, A.; Reitmeir, P.; John, J. (2005): Kassenwechsel und Risikostrukturausgleich in der gesetzlichen Krankenversicherung -- empirische Befunde der Kooperativen Gesundheitsforschung in der Region Augsburg (KORA). In: *Gesundheitswesen* 67 Suppl 1, S. S158-66.

WIdO: GKV-Arzneimittelindex. Online verfügbar unter http://wido.de/amtl_atc-code.html.

WIdO (2007): Qualitätssicherung der stationären Versorgung mit Routinedaten (QSR). Abschlussbericht. 1. Aufl. Unter Mitarbeit von S. Sollmann. AOK-Bundesverband FEISA HELIOS Kliniken WIdO. Bonn.

Wilke, T.; Groth, A.; Mueller, S.; Reese, D.; Linder, R.; Ahrens, S.; Verheyen, F. (2013): How to use pharmacy claims data to measure patient nonadherence? The example of oral diabetics in therapy of type 2 diabetes mellitus. In: *The European Journal of Health Economics* 14 (3), S. 551–568.

WINEG: Homepage des Wissenschaftliches Institut der TK für Nutzen und Effizienz im Gesundheitswesen. Online verfügbar unter http://www.tk.de/tk/wineg/118306.

Wulffen, M. von; Schütze, B. (2014): SGB X. Sozialverwaltungsverfahren und Sozialdatenschutz. In: *SGB X*.

Zeidler, J.; Braun, S. (2012): Sekundärdatenanalysen. In: Schöffski, O. und Graf von der Schulenburg, J.-M. (Hrsg.): Gesundheitsökonomische Evaluationen. 4. Aufl. Berlin, Heidelberg: Springer-Verlag, S. 243–274.

Zeidler, J.; Lange, A.; Braun, S.; Linder, R.; Engel, S.; Verheyen, F.; Graf von der Schulenburg, J.-M. (2013): Die Berechnung indikationsspezifischer Kosten bei GKV-Routinedatenanalysen am Beispiel von ADHS. In: *Bundesgesundheitsblatt* 56 (3), S. 430–438.

Zeidler, J.; Mittendorf, T.; Vahldiek, G.; Graf von der Schulenburg, J.-M. (2008a): Kostenvergleichsanalyse der ambulanten und stationären kardiologischen Rehabilitation. In: *Herz* 33 (6), S. 440–447.

Zeidler, J.; Mittendorf, T.; Vahldiek, G.; Zeidler, H.; Merkesdal, S. (2008b): Comparative cost analysis of outpatient and inpatient rehabilitation for musculoskeletal diseases in Germany. In: *Rheumatology* 47 (10), S. 1527–1534.

Zentralinstitut für die kassenärztliche Versorgung in der Bundesrepublik Deutschland: AGENS-Methodenworkshop 2013. Online verfügbar unter http://www.zi.de/cms/-veranstaltungen/agens-methodenworkshop-2013/.

Ziegler, U.; Doblhammer, G. (2009): Prävalenz und Inzidenz von Demenz in Deutschland – Eine Studie auf Basis von Daten der gesetzlichen Krankenversicherungen von 2002. Rostocker Zentrum – Diskussionpapier Nr. 24. Rostocker Zentrum zur Erforschung des Demografischen Wandels.

Zok, K. (2011): Reaktionen auf Zusatzbeiträge in der GKV. Ergebnisse einer Repräsentativ-Umfrage. 1. Aufl. Hrsg. v. WIdO.

Zwiener, I.; Blettner, M.; Hommel, G. (2011): Überlebenszeitanalyse. Teil 15 der Serie zur Bewertung wissenschaftlicher Publikationen. In: *Deutsches Ärzteblatt international* 108 (10), S. 163–169. Online verfügbar unter http://www.aerzteblatt.de/archiv/-81171/Ueberlebenszeitanalyse-Teil-15-der-Serie-zur-Bewertung-wissenschaftlicher-Publikationen?src=series.