Thieme

Sonderheft Methodische Aspekte der Sekundärdatenanalyse

Datenlinkage und spezifische Methoden der Sekundärdatenanalyse

Autoren
Enno Swart[1], Holger Gothe[2,3,4], Peter Ihle[5], Stefanie March[1], Ingrid Schubert[5], Christoph Stallmann[1], Falk Hoffmann[6]

Institute
1 Institut für Sozialmedizin und Gesundheitssystemforschung, Medizinische Fakultät, Otto-von-Guericke-Universität Magdeburg
2 IGES Institut, Berlin
3 Medizinische Fakultät, Lehrstuhl Gesundheitswissenschaften/Public Health, TU Dresden, Dresden
4 Department für Public Health, Versorgungsforschung und Health Technology Assessment, UMIT, Hall in Tirol, Austria
5 PMV forschungsgruppe, Universität zu Köln
6 Department für Versorgungsforschung, Fakultät Medizin und Gesundheitswissenschaften, Carl von Ossietzky Universität Oldenburg

Bibliografie
DOI https://doi.org/10.1055/a-1099-0789
Gesundheitswesen 2020; 82 (Suppl. 2): S91–S93

ISSN 0949-7013

Korrespondenzadresse
PD Dr. rer. biol. hum. Enno Swart
Medizinische Fakultät
Institut für Sozialmedizin und Gesundheitssystemforschung (ISMG)
Otto-von-Guericke-Universität Magdeburg
Leipziger Straße 44
39120 Magdeburg
enno.swart@med.ovgu.de

Mit dem vorliegenden zweiten Sonderheft zu Datenlinkage und spezifischen Methoden der Sekundärdatenanalyse ergänzen wir die neun Beiträge, die im vorangegangenen Sonderheft methodische Herausforderungen bei der Aufbereitung und Validierung verschiedener Sekundärdatenquellen thematisiert haben. Ein besonderes Augenmerk liegt auf dem individuellen Datenlinkage verschiedener Sekundärdatenquellen untereinander und/oder mit Primärdaten. Dieser methodische Zugang hat nicht erst seit der Bestandsaufnahme zum Datenlinkage in Deutschland [1] und der Veröffentlichung der Guten Praxis Datenlinkage [2] zunehmende Relevanz in der Epidemiologie und Versorgungsforschung erhalten. Daneben werden in diesem Heft aktuelle Ansätze zur Datenvalidierung, zur Kontrollbildung und zur Aufbereitung und Nutzung von Klartextangaben vorgestellt.

Die Validierung von Diagnoseangaben stellt eine permanente Herausforderung bei der wissenschaftlichen Nutzung der Abrechnungsdaten der gesetzlichen Krankenversicherung (GKV) dar. Die zur Validierung geeigneten Ansätze variieren nach dem Krankheitsbild und der Verfügbarkeit einer weiteren Datenquelle zur externen Validierung in Ergänzung interner Validierungsalgorithmen. Der von Behrendt et al. vorgestellte Ansatz beschreibt für Patienten mit peripherer arterieller Verschlusskrankheit eine externe Validierung unter Nutzung von Daten eines klinischen Krankheitsregisters (German VascRegister). Dabei wird keine individuelle Verlinkung (als selten erreichbarer Goldstandard) vorgenommen, sondern es werden die Ergebnisse separater multivariater Modellierungen in beiden Datenquellen verglichen, ebenso deskriptive Beschreibungen von Subgruppen. Obwohl sich dieser Ansatz zunächst nur für die beschriebene Patientenpopulation bewährt hat, zeigt er das Potenzial vergleichsweise einfacher Validierungen, an denen sich die Eignung der GKV-Daten für klinisch orientierte Forschungs- und Qualitätssicherungsprozesse evaluieren lässt.

Druschke et al. präsentieren Erfahrungen aus einem Versorgungsforschungsprojekt zu mittel- und langfristigen gesundheitlichen Outcomes bei Frühgeborenen, in dem drei Datenquellen individuell verlinkt wurden: Abrechnungsdaten der GKV, Daten einer schriftlichen Befragung von Eltern bzw. Betreuern dieser Kinder und Daten der Kindergarten- und Schuleingangsuntersuchungen sächsischer Gesundheitsämter. Der Beitrag beschreibt den organisatorischen und rechtlichen Rahmen dieses Projekts und die Validierungsschritte, mit denen Fehler bei der Verlinkung der Datensätze vermieden werden sollen. Die Autorinnen und Autoren leiten zusammenfassend Empfehlungen für zukünftige Verlinkungsstudien ab. Dazu gehören ausreichend zeitliche Ressourcen bei Planung und Durchführung der Datenerschließung und -verlinkung wie im anschließenden Prozess der Datenvalidierung. Dabei sollten auch klinische Experten eingebunden werden. Zu Validierungszwecken sollten in den zu verlinkenden Datenquellen mehr als nur die direkten Schlüsselvariablen vorhanden sein. Schließlich ist für Powerüberlegungen zu berücksichtigen, dass die finale Fallzahl sich durch mangelnde Response bzw. Mismatches deutlich gegenüber der Versichertenzahl in den GKV-Daten verringern kann.

GKV-Routinedaten weisen neben ihren bekannten Potenzialen eine Reihe von Limitationen auf. Diese müssen im Kontext konkreter Forschungsprojekte und klinischer Settings jeweils neu bzgl. der Auswirkung auf die interne und externe Validität bewertet werden. Diese Herausforderungen erläutern Brandl et al. am Beispiel der Schätzung inzidenter Ereignisse (post intensive care syndrom; PICS) bei Patienten nach Behandlung in Intensivstationen. Konkret gehören dazu die Differenzierung zwischen primärem Outcome (PICS) und kompetitivem, aber nicht unabhängigem Outcome

(Tod), die Operationalisierung einer komplexen medizinischen Entität wie des PICS über mehrere ICD-Codes und das Fehlen einer exakten Information zum Zeitpunkt des Auftretens des Zielereignisses. Für den Kontext der genannten Studie werden konkrete Lösungsansätze präsentiert. Die Übertragbarkeit auf andere Forschungskontexte muss im Einzelfall neu geprüft werden.

Weitere Limitationen, die die GKV-Abrechnungsdaten aufweisen, sind u. a. das Fehlen klinischer Informationen oder nur begrenzt belastbarer Informationen zum sozioökonomischen Status. Diese Limitationen versuchen epidemiologische Studien durch ein individuelles Datenlinkage zu überwinden. Dieses Linkage kann mit primär erhobenen Forschungsdaten erfolgen oder mit Sekundärdaten anderer Dateneigner. Ansätze zur Überwindung organisatorischer und rechtlicher Herausforderungen beim individuellen Linkage beschreiben Langner et al. am Beispiel der Verlinkung von GKV-Daten mit Informationen aus einem epidemiologischen Krebsregister zur Todesursache im Zusammenhang mit der Evaluation des Mammografie-Screening-Programms. Sie gehen dabei besonders auf die notwendigen Genehmigungsprozesse bei derartigen Linkage-Studien ein und beschreiben mögliche datenschutzkompatible Flow-Charts zwischen Dateneigner und Forschern unter Nutzung einer Vertrauensstelle. Die Autoren betonen die Notwendigkeit, bei derartigen Studien ausreichende zeitliche und finanzielle Mittel für den Datenlinkageprozess bereit zu stellen.

Mit Gothe et al. werfen wir einen Blick über die Landesgrenzen auf ein Datenlinkage-Projekt im Rahmen einer klinischen Studie zur Wirksamkeit eines Post-Stroke-Disease-Managementprogramms in Tirol. In dieser Studie wird über die Verlinkung von Studiendaten mit Abrechnungsdaten der Tiroler Gebietskrankenkasse auf der Basis eines individuellen informed consent ein mittelfristiges Follow-Up über die Akutversorgung hinaus ermöglicht. Erst die Verknüpfung der beiden Datensätze gestattet eine umfassende gesundheitsökonomische Evaluation, wie sie für die Entscheidung über die Implementierung von Versorgungsinnovationen in den Behandlungsalltag erforderlich ist und relevante Erkenntnisse für die Entscheidung über deren Erstattungsfähigkeit im sozialen Sicherungssystem liefern kann. Einzelheiten der Projektplanung und -durchführung werden beschrieben. Das Datenlinkage in Österreich wird dabei durch die Tatsache erleichtert, dass bundeslandbezogen die weit überwiegende Mehrheit der in einem Angestelltenverhältnis erwerbstätigen Einwohner bei der jeweiligen Gebietskrankenkasse (bzw. neuerdings der jeweiligen Landesstelle der Österreichischen Gesundheitskasse) versichert ist. Die Autorengruppe schlussfolgert aus diesem österreichischen Prototyp einer Datenlinkagestudie, in zukünftigen gesundheitsökonomischen Evaluationen verstärkt auf eine individuelle Verknüpfung von Informationen aus verschiedenen Datenquellen zu setzen.

Gewissermaßen in Fortführung des Beitrags von Langer et al. beschreiben Bartholomäus et al. die konkreten Schritte und Algorithmen einer individuellen Zusammenführung von Krebsregister- und GKV-Abrechnungsdaten. Dabei kommen eine separate Verschlüsselung personenidentifizierender Daten und eine parallele stufenweise Anonymisierung der Leistungsdaten zum Einsatz, womit die notwendige Anonymisierungstiefe in den Leistungsdaten erreicht werden soll, etwa durch Vergröberung des Geburts- (in Monat bzw. Quartal) oder Leistungsdatums (z. B. Quartal) oder der Postleitzahl (auf die ersten vier Ziffern verkürzt). In einem Datenaufbereitungszentrum werden über Zuordnungsnummern die Angaben aus verschiedenen Datenquellen individuell verknüpft, damit wird das Problem quasi-identifizierender Merkmale überwunden. Dieses mit der europäischen und deutschen Datenschutzgesetzgebung kompatible Verfahren eignet sich auch für andere Forschungskontexte mit angestrebtem individuellen Datenlinkage.

Beobachtungsstudien haben gegenüber randomisierten klinischen Studien den grundsätzlichen Nachteil, dass eine Strukturgleichheit zwischen zwei oder mehr Studienpopulationen a-priori nicht hergestellt werden kann. Insofern ist der Schluss von beobachteten Unterschieden zwischen den Gruppen auf einen kausalen Effekt, etwa unterschiedlicher Versorgungskonzepte, nicht prinzipiell gerechtfertigt. Jedoch versprechen verschiedene methodische Ansätze des Propensity-Score-Matchings (PSM) nachträglich eine mögliche Kontrolle potenzieller konfundierender Faktoren. Matschinger et al. untersuchen im Kontext einer Interventionsstudie bei chronisch kranken Versicherten zur Evaluation der Wirksamkeit eines individuellen Telefoncoachings zur Erhöhung der Gesundheitskompetenz die Performance verschiedener Verfahren des PSM, bei dem sich das entropy balancing als der Ansatz mit der potenziell geringsten Verzerrung herausstellt.

Abrechnungsdaten der GKV eignen sich potenziell zur Evaluation neuer Versorgungsmodelle wie z. B. der Disease-Management-Programme (DMP). Angesichts der bekannten Selektionsprozesse bei der freiwilligen Einschreibung in derartige Leistungsangebote ist die Generierung einer geeigneten Kontrollgruppe eine entscheidende Voraussetzung für eine belastbare Bewertung der mit dem neuen Versorgungsangebot verbundenen Outcomes. Das PSM hat sich in diesem Zusammenhang breit etabliert. Jacob et al. untersuchen in ihrem Beitrag, wie innerhalb eines PSM-Ansatzes zur Evaluation des DMP Asthma bronchiale potenziellen Kontrollen ein geeignetes virtuelles Einschreibedatum im Vergleich zum bekannten Einschreibedatum der DMP-TeilnehmerInnen zugewiesen werden kann, um eine Parallelität der Follow-Up-Zeiträume zu gewährleisten. In drei Szenarien erweist sich die zufällige Zuweisung eines virtuellen Einschreibedatums innerhalb eines Quartals als beste Lösung. Mit dieser Zuweisung kann dann anschließend das PSM angestoßen werden.

Die meisten Sekundärdaten enthalten primär quantitative und kurze qualitative Inhalte, ergänzt um Datumsinformationen. Gleichwohl existieren auch wissenschaftlich potenziell interessante personen- und fallbezogene Informationen in umfangreicherer Textform, z. B. aus Arztbriefen. Diese Informationen stehen solange de facto nicht einer systematischen Nutzung offen, solange es keine standardisierten Methoden zur Erfassung und Klassifizierung dieser Informationen gibt. Pokora et al. untersuchen existierende Ansätze bzgl. ihres Aufwands, ihrer Praktikabilität sowie Reliabilität und Validität. Dabei erweisen sich die manuelle Klassifizierung und die Nutzung vorgegebener Schlagworte als zu aufwändig bzw. zu fehlerbehaftet. Im Gegenzug bietet sich computergestütztes Textmining in Kombination mit maschinellem Lernen als zuverlässiges Klassifizierungsverfahren an. Eine bestehende Software weist lediglich eine gewisse Untererfassung klinisch auffälliger Befunde auf, bietet gleichwohl das Potenzial für die zukünftige Erschließung bislang ungenutzter Freitextangaben.

Die hier und im vorgehenden Sonderheft präsentierten Lösungsansätze für relevante und häufig auftretende methodische Herausforderungen bei Sekundärdatenanalysen können die Beschäftigung mit Grundlagenwerken und Standards der Sekundärdatenanalyse nicht ersetzen [3–5], sodass sie als Ergänzung dazu zu verstehen sind. Wir wünschen nun allen Leserinnen und Lesern eine erkenntnisreiche Lektüre, unabhängig davon, ob sie erfahren auf dem Feld der Sekundärdatenanalyse oder erst in jüngerer Zeit im Rahmen aktueller Forschungsprojekte mit diesen Datenquellen befasst sind.

Interessenkonflikt

Die Autorinnen/Autoren geben an, dass kein Interessenkonflikt besteht.

Literatur

[1] March S, Antoni M, Kieschke J et al. Quo vadis Datenlinkage in Deutschland? Eine erste Bestandsaufnahme. Gesundheitswesen 2018; 80: e20–e31. doi:10.1055/s-0043-125070

[2] March S, Andrich S, Drepper J et al. Gute Praxis Datenlinkage (GPD). Gesundheitswesen 2019; 81: 636–650. doi:10.1055/a-0962/9933

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

[4] Swart E, Gothe H, Geyer S et al. Gute Praxis Sekundärdatenanalyse (GPS): Leitlinien und Empfehlung, 3. Revision, Fassung 2012/2014. Gesundheitswesen 2015; 77: 120–126. doi:10.1055/s-0034-1396815

[5] Swart E, Bitzer EM, Gothe H et al. STandardisierte BerichtsROutine für SekundärdatenAnalysen (STROSA) – ein konsentierter Berichtsstandard für Deutschland, Version 2. Gesundheitswesen 2016; 78 (suppl 1): e145–e160. doi:10.1055/s-0042-108647