Anreicherung eines GKV-Datensatzes mit amtlichen Todesursachen über einen Abgleich mit dem Epidemiologischen Krebsregister Nordrhein-Westfalen: Machbarkeitsstudie und Methodenvergleich

Ingo Langner, Volker Krieg, Oliver Heidinger, Hans-Werner Hense, Hajo Zeeb

DOI
10.1055/s-0043-124669

Published in
Das Gesundheitswesen

Document version
Accepted manuscript

Online publication date
1 February 2018

Corresponding author
Ingo Langner

Citation
Langner I, Krieg V, Heidinger O, Hense H-W, Zeeb H. Anreicherung eines GKV-Datensatzes mit amtlichen Todesursachen über einen Abgleich mit dem Epidemiologischen Krebsregister Nordrhein-Westfalen: Machbarkeitsstudie und Methodenvergleich. Das Gesundheitswesen. 2019;81(08/09):629-35.

TITEL

Anreicherung eines GKV-Datensatzes mit amtlichen Todesursachen über einen Abgleich mit dem Epidemiologischen Krebsregister Nordrhein-Westfalen: Machbarkeitsstudie und Methodenvergleich

TITLE

Enrichment of claims data with official causes of death using a record linkage with the epidemiological cancer registry of North Rhine-Westphalia: Feasibility study and comparison of procedures

Autoren und Adressen:

Ingo Langner[1], Volker Krieg[2], Oliver Heidinger[2], Hans-Werner Hense[2,3], Hajo Zeeb[1,4]

[1] Leibniz-Institut für Präventionsforschung und Epidemiologie - BIPS

[2] Landeskrebsregister Nordrhein-Westfalen

[3] Institut für Epidemiologie und Sozialmedizin, Westfälische Wilhelms-Universität Münster

[4] Wissenschaftsschwerpunkt Gesundheitswissenschaften, Universität Bremen

Ziel-Journal: Gesundheitswesen (25.000 Zeichen, inklusive Leerzeichen, Literaturverzeichnis und Zusammenfassung; max. 5 Tabellen oder Abbildungen)

Schlüsselwörter: Sekundärdaten, Krankenkassendaten, Datenabgleich, Todesursachen

Key words: secondary data, claims data, record linkage, cause of death

Zusammenfassung

Ziel der Studie: Für die Evaluation von Krebsvorsorgeprogrammen stellen Daten der Gesetzlichen Krankenversicherung (GKV) eine wichtige Datenquelle dar, die jedoch nicht die benötigten Daten zum genauen Sterbedatum und zur Todesursache enthält. Diese Studie sollte prüfen, ob eine diesbezügliche Anreicherung individueller GKV-Daten über einen Abgleich mit einer geeigneten externen Datenquelle erfolgen kann.

Methodik: In der pharmako-epidemiologischen Forschungsdatenbank GePaRD identifizierten wir eine Versichertenstichprobe von 25.528 Frauen, die laut den Angaben in GePaRD im Zeitraum 2006 bis 2013 verstorben waren und ihren Wohnsitz in Nordrhein-Westfalen (NRW) hatten. Datum und Ursache des Todes aller Einwohner von NRW seit 2005 liegen im Epidemiologischen Krebsregister von NRW vor. In Kooperation mit zwei gesetzlichen Krankenkassen wurde mit einem probabilistischen bzw. deterministischen Abgleichverfahren versucht, jeder Verstorbenen der Stichprobe einen Todesfall aus NRW und damit eine Todesursache zuzuordnen.

Ergebnisse: Für 94,72% der Verstorbenen der Versichertenstichprobe konnte probabilistisch und für 93,36% deterministisch ein Todesfall aus NRW zugeordnet werden.

Schlussfolgerung: Das probabilistische und das deterministische Verfahren erreichten vergleichbar hohe Trefferquoten. Nicht erfolgte Zuordnungen sind vermutlich größtenteils auf Fehler bei der Erfassung der Personendaten zurückzuführen. Aufgrund des geringeren technischen Aufwands erscheint das deterministische Verfahren als die Methode der Wahl für die Anreicherung von GKV-Daten mit amtlichen Todesursachen aus geeigneten externen Datenquellen.

Abstract

Background: Claims data of the statutory health insurance (SHI) are an important data source for the evaluation of cancer prevention programs. However, the relevant information for cause of death is lacking in this source. This study examined, whether an enrichment of individual claims data with the required cause of death using record linkage procedures with suitable external data sources.

Methods: In the German pharmacoepidemiologic research database (GePaRD) we identified a sample of 25,528 deceased female residents of North Rhine Westphalia (NRW) who, according to GePaRD information, died between 2006 and 2013. Date and cause of all deaths among inhabitants of NRW since 2005 were available in the epidemiological cancer registry of NRW. In cooperation with two SHI companies we tried to match each individual of the sample with a case of death in NRW and the corresponding cause of death using a probabilistic and, alternatively, a deterministic linkage procedure.

Results: Of the study sample, 94.72% were successfully matched by the probabilistic and 93.36% by the deterministic method.

Conclusions: The probabilistic and the deterministic record linkage approach produced comparably high matching rates. Cases without matches are probably due to errors occurring at the stage of personal data entry. Given the lower technical efforts, the deterministic approach appears to be the method of choice for the enrichment of claims data with cause of death information from suitable external data sources in Germany.

Einleitung

Abrechnungsdaten der Gesetzlichen Krankenversicherungen (GKV) bieten eine wichtige Datenressource für die epidemiologische Forschung. Swart und Ihle [1] heben hervor, dass diese Daten, prinzipiell jederzeit und zeitnah zur Verfügung stehen und dass sie aktuell, kostengünstig und schnell zugänglich sind. Die Nutzung der routinemäßig erhobenen Abrechnungsdaten der GKV ermöglicht daher in relativ einfacher Weise die Planung und Durchführung von retrospektiven, aber auch prospektiven Studienansätzen zu aktuellen Fragestellungen. Dies ist besonders im Zusammenhang mit langfristig wirkenden Einflussfaktoren hilfreich, da medizinische Leistungen für individuelle Patienten in standardisierter, gesetzlich geregelter Art und Weise (Sozialgesetzbuch V §295, § 301) über lange Zeiträume dokumentiert werden. Hinzu kommt, dass die Vielfalt der individuellen Diagnosen, Behandlungen und Heilverfahren auch noch rückwirkend erfasst werden kann: dies gilt für Einflussfaktoren wie Zielereignisse gleichermaßen. Andererseits unterliegen die Datensätze der GKV aber auch einigen inhaltlichen Beschränkungen. So werden die Daten beim primären Erhebungsanlass ohne direkten Bezug zur späteren wissenschaftlichen Fragestellung dokumentiert, wichtige klinische Begleitinformationen, zum Beispiel Laborresultate oder Befunde der Bildgebung, liegen nicht vor und die Validität der Kodierungen kann gegebenenfalls ebenfalls unzureichend sein [1].

Auch die Todesursache (TU) als ein wichtiger Endpunkt der epidemiologischen Forschung ist in den Daten der gesetzlichen Krankenkassen nicht enthalten. Im Rahmen der Machbarkeitsstudie zur Evaluation der Brustkrebs-assoziierten Sterblichkeit im deutschen Mammographie-Screening-Programm [2] prüften wir deshalb, ob es möglich ist, die Daten von verstorbenen Versicherten mit den amtlichen TU aus einer anderen Datenquelle anzureichern. Dabei wurden zwei verschiedene Abgleichsverfahren verglichen.

Mit einer erfolgreichen Ergänzung von TU in Kassendaten könnten diese auch in Studien zu spezifischen Mortalitätsendpunkten genutzt werden.

Methodik

Datenquellen

Die Forschungsdatenbank GePaRD (German pharmaco-epidemiological research data base) enthält Routinedaten der Gesetzlichen Krankenversicherung (GKV) und wird bisher primär in der pharmakoepidemiologischen Forschung und in der Versorgungsforschung genutzt [3-6]. Die in GePaRD enthaltene Versichertenstichprobe von vier gesetzlichen Krankenkassen erfasst etwa 15% der gesamten deutschen Bevölkerung und enthält Angaben zu Geschlecht, Geburtsjahr, Wohnregion, Tätigkeitscode und Versicherungszeiten sowie Informationen über erbrachte ambulante und stationäre medizinische Leistungen und die zugrunde liegenden Erkrankungen. Die von Krankenhäusern an die Kassen übermittelten Daten enthalten neben der Entlassungshauptdiagnose den Entlassungsgrund sowie das Entlassungsdatum der behandelten Person. Zudem halten die gesetzlichen Krankenkassen bei Beendigung einer Versicherungsperiode eines Versicherten den Grund für die Beendigung – soweit bekannt - fest.

Das Versterben des Versicherten ist ein möglicher Austrittsgrund, allerdings wird die TU dabei nicht erfasst. Diese kann in GePaRD nicht ohne weiteres mit Informationen aus einer externen Datenbank ergänzt werden, da die direkt personenidentifizierenden Daten wie Namen, Wohnadresse, Geburtsdatum und Versichertennummer nur bei den jeweiligen Krankenkassen vorliegen. Die zwei größten der in GePaRD einbezogenen Krankenkassen, die Techniker Krankenkasse (TK) und die DAK-Gesundheit, erklärten sich bereit, im Rahmen der Machbarkeitsstudie an einem Datenabgleich mit dem EKR NRW teilzunehmen und hierfür datenschutzkonform Identitätsdaten zur Verfügung zu stellen.

Nach § 75 des SGB X mussten die zuständigen Aufsichtsbehörden (Bundesversicherungsamt; Senatorin für Wissenschaft, Gesundheit und Verbraucherschutz der Hansestadt Bremen; Niedersächsisches Ministerium für Soziales, Gesundheit und Gleichstellung) der Nutzung der Sozialdaten der gesetzlichen Krankenkassen für die Machbarkeitsstudie zustimmen. Die behördliche Genehmigung legt dabei den Zweck, die Art und Dauer der Nutzung dieser Daten fest. Für die Datenflüsse der hier beschriebenen Abgleiche wurde eine Genehmigung beim zuständigen Bundesversicherungsamt eingeholt.

Das Epidemiologische Krebsregister NRW (EKR NRW; seit dem 1. April 2016 Teil des integrierten epidemiologisch-klinischen Landeskrebsregisters NRW, LKR NRW) erfasst neben den Daten zu inzidenten Krebsneuerkrankungen auch Informationen zu allen Sterbefällen einschließlich der zugehörigen TU (kodiert nach ICD-10-GM) in der Wohnbevölkerung von

NRW, was auch Personen ohne vorherige Krebserkrankung mit einschließt. Dazu sind sämtliche 396 Meldeämter sowie die zentrale Statistikstelle des Landes (IT.NRW) an das elektronische Meldenetzwerk des Krebsregisters angeschlossen [7]. Bei den obligat elektronischen Meldungen wurden die personenidentifizierenden Daten bei der Datenübermittlung durch eine doppelte Verschlüsselung in sogenannte Patientenpseudonyme überführt und getrennt von den medizinischen Daten an das Krebsregister gesandt. Die eingesetzten deterministischen Verschlüsselungsverfahren ermöglichen die einzelfallbezogene Verknüpfung von Informationen verschiedener Datenhalter im Krebsregister auch auf Basis verschlüsselter Identitätsdaten , da bei dieser Art der Verschlüsselung der gleiche Klartext immer auf das gleiche Chiffrat abgebildet wird. Das vormalige Krebsregistergesetz (KRG) NRW (gültig bis 31.03.2016) regelte, dass die Todesfallmeldungen aus den Einwohnermeldeämtern und die amtlichen TU aller in NRW verstorbenen Personen zum Zweck der krebsbezogenen Forschung herangezogen werden dürfen. Nach § 10 des KRG NRW war das Einholen einer Einwilligung der Patientinnen und Patienten zur Nutzung ihrer Daten im Rahmen von Forschungsvorhaben nicht gefordert, wenn lediglich das Sterbedatum und die amtliche TU einer verstorbenen Person übermittelt werden.

Studienpopulation

Die aus GePaRD für diese Studie selektierte Stichprobe umfasste 25- bis 80-jährige, in Nordrhein Westfalen wohnhafte, weibliche Versicherte der TK und der DAK, bei denen 'Tod' als Grund für die Beendigung des Versicherungsverhältnisses oder als Grund für die Entlassung aus einer stationären Krankenhausbehandlung im Zeitraum 2006 bis 2013 dokumentiert wurde. Basierend auf den Daten von GePaRD wurde als Sterbedatum zunächst das Datum der Beendigung der Versicherungsperiode respektive das Krankenhausentlassungsdatum gewählt. Unter 259.585 Frauen, die laut GePaRD Informationen aus NRW stammten und ein zwischen 2006 und 2013 endendes Versicherungsverhältnis aufwiesen, befanden sich 25.647 Frauen, die als verstorben gekennzeichnet waren; von diesen wurden auf Basis der vollständigeren Adressinformationen bei den Kassen 25.528 Frauen mit tatsächlichem Wohnsitz in NRW selektiert.

Datenfluss und Verschlüsselung

Zur Ergänzung der Daten in GePaRD mit Informationen zur TU waren Datenflüsse zwischen verschiedenen Institutionen erforderlich: neben dem Institut für Informations-, Gesundheits- und Medizinrecht (IGMR), das als Vertrauensstelle für die Krankenkassen und GePaRD dient, waren dies die beiden teilnehmenden Kassen, das EKR NRW sowie das Leibniz-Institut für Präventionsforschung und Epidemiologie (BIPS) in Bremen (Abbildung 1). Die Datenflüsse umfassten nur die für den Abgleich notwendigen Daten, d.h. für die aus GePaRD selektierte Stichprobe wurden lediglich das GePaRD-Pseudonym der Versicherten, das Austrittsdatum, den Austrittsgrund sowie eine Kommunikations-ID (KID), die als fortlaufende Nummerierung der Teilnehmerinnen erzeugt wurde, weitergegeben. In der Vertrauensstelle am IGMR wurde das GePaRD-Pseudonym gegen ein anderes, spezifisch von den gesetzlichen Krankenkassen verwendetes Pseudonym ausgetauscht. Dieses wurde danach bei der jeweiligen Krankenkasse durch die Versichertennummer ersetzt. Über die Versichertennummer konnten bei der Krankenkasse aus den Stammdaten dann jeweils Nachname, Vorname, Geburtsdatum und Wohnadresse dem Datensatz hinzugefügt werden; anschließend wurden das Krankenkassen-Pseudonym sowie die Versichertennummer wieder gelöscht.

Da die Identitätsdaten im EKR NRW in doppelt verschlüsselter Form als Patientenpseudonyme vorliegen, mussten die Identitätsdaten der Studienkohorte in gleicher Weise verschlüsselt werden. Hierzu stellte das EKR die Datenübermittlungs- und -verschlüsselungssoftware "EpiCan" bereit. Bei diesem Verfahren verließen personenidentifizierende Merkmale, wie Name, Vorname, Geburtsname, frühere Namen und Titel, der Tag des Geburtsdatums, Straße und Hausnummer der Wohnanschrift zum Zeitpunkt der Diagnosestellung (PID-1) zu keinem Zeitpunkt den Hoheitsbereich der Krankenkassen, sondern wurden zuvor von EpiCan zerlegt, normiert und mithilfe des MD5-Verfahrens zu maximal 31 sogenannten Personenkryptogrammen (PKG) einwegverschlüsselt. Dieses deterministische Kryptographierungsverfahren ist auf rechnerischem Wege zwar nicht umkehrbar, kann jedoch durch Probeverschlüsselung nachvollzogen werden. Daher wurden die PKG mit einem weiteren Chiffrierverfahren überverschlüsselt, was gemäß KRG NRW von einem speziell eingerichteten

Pseudonymisierungsdienst (PSD) durchgeführt wurde. Der PSD nahm die PKG von den Kassen entgegen, überverschlüsselte diese deterministisch unter Verwendung eines geheimen Schlüssels in Patientenpseudonyme (PSN) [8] und übermittelte diese an das EKR. Der PSD löschte anschließend sowohl die PKG als auch die PSN vollständig. Ein Rückbezug der beim EKR vorliegenden PSN zu den Ausgangsdaten (PID-1) ist nach heutigem Kenntnisstand ausgeschlossen. Die zusätzlichen Personendaten wie Geschlecht, Geburtsmonat und -jahr, PLZ und Wohnort (PID-2), wurden auf getrenntem Weg und separat von den PSN im Klartext an das EKR übermittelt. Ein von der Software EpiCan vergebener Zeitstempel ermöglichte dann die Zusammenführung der PID-1 und PID-2 beim EKR.

Nach erfolgtem Abgleich (siehe nachfolgenden Abschnitt) wurden die PSN beim EKR gelöscht und Informationen zur TU wurden allein mit der Kommunikations-ID (KID) an GePaRD zurückgesandt (Abbildung 1).

Abgleichverfahren

Die zwei im Rahmen unserer Studie beim EKR NRW verwendeten Abgleichverfahren basierten zum Teil auf verschlüsselten (PID-1) und zum Teil auf unverschlüsselten Identitätsdaten (PID-2). Beide Verfahren unterschieden sich in Anzahl und Auswahl der PID sowie in den zum Abgleich eingesetzten Verfahren.

Das **probabilistische** Verfahren nutzte Vor-, Nach- und frühere Namen und deren phonetische Entsprechungen sowie Tag des Geburtsdatums, Straßenname und Hausnummer in verschlüsselter Form (19 PSN zu den PID-1: siehe Tabelle 1). Hinzu kamen die 4 Klartextmerkmale Geschlecht, Monat und Jahr des Geburtsdatums sowie die aus Postleitzahl und Wohnort gebildete Gemeindekennziffer (8-stellig) der Wohnregion (PID-2). Die PSN wurden – wie oben beschrieben - durch deterministische Verschlüsselungsverfahren gebildet. Allerdings ist dabei zu beachten, dass (anders als beim Abgleich mit Klartextdaten) ähnliche Originalausprägungen (wie z.B. Meyer und Meier) nicht zu ähnlichen PSN führen. Um speziell Schreibfehlern zu begegnen, die bei der Dokumentation von Merkmalen von einer gleich klingenden Aussprache der Schreibvarianten herrühren, werden zusätzlich Phoneme der Namen gebildet und als separate Pseudonyme verschlüsselt. Damit

Meldungen zu einer Person, die von verschiedenen Datenhaltern stammen und möglicherweise nicht vollständig identisch sind, auch mit Hilfe verschlüsselter Zeichenketten korrekt zusammenführt werden können, erlaubt das probabilistische Record Linkage eine Zuordnungstoleranz, so dass nicht alle zu vergleichenden Zeichenketten für eine Zuordnung übereinstimmen müssen. Für jedes abzugleichende Paar wurde dazu ein Übereinstimmungsgewicht aus einem Vergleich der 19 PSN sowie den PID-2 berechnet. Dabei werden die Häufigkeitsverteilungen der Identitätsdaten berücksichtigt [8,9], z.B. ob ein Name selten ist oder ob Meldungen aus einer kleinen oder einer großen Gemeinde kommen. Obere und untere Grenzwerte für die Übereinstimmungsgewichte erlauben dann eine automatisierte Entscheidung, ob zwei Datensätze als zusammengehörend oder als unterschiedlich betrachtet werden. Fälle mit Übereinstimmungsgewichten im Bereich zwischen diesen Grenzwerten erfordern eine manuelle Nachbearbeitung und Entscheidung über die Zuordnung.

Dieses beim EKR NRW etablierte, probabilistische Abgleichverfahren hat sich in verschiedenen Studien bewährt [10-13] und erlaubt eine sehr zuverlässige Verknüpfung von Datensätzen zu einer Person. Bezogen auf 150.000 Meldungen lagen die Synonymfehlerrate (d.h., die Datensätze aus verschiedenen Quellen zu einer Person werden nicht als zusammengehörend erkannt) bei 0,18% und die Homonymfehlerrate (d.h., die Datensätze verschiedener Personen werden fälschlicherweise einer Person zugeordnet) bei lediglich 0,015% [14].

Das **deterministische** Abgleichsverfahren nutzte dagegen das Sterbedatum, Monat und Jahr des Geburtsdatums, das Geschlecht und die Kreiskennziffer (5-stellig) des Wohnortes im Klartext und den Tag des Geburtsdatums in verschlüsselter Form. Für eine Zuordnung von Datensätzen aus den jeweiligen Datenquellen musste eine exakte Übereinstimmung all dieser Merkmale vorliegen. Eine Zuordnungstoleranz mit Übereinstimmungsgewichten wie beim probabilistischen Abgleichverfahren gab es nicht.

Statistische Analysen

Für die Analysen der Übereinstimmung zwischen den Verstorbenen aus GePaRD und allen Todesfällen, so genannte ‚Treffer' bzw. ‚Matches', wurden ausschließlich deskriptive Statistiken verwendet. Alle Auswertungen erfolgten mit SAS Version 9.3.

Ergebnisse

Aus der Stichprobe von 25.528 Frauen im Alter von 25 bis 80 Jahren, die auf Basis der Kassendaten als verstorben klassifiziert wurden, konnte 24.180 Frauen (94,72%) mit dem probabilistischen Abgleichverfahren aufgrund hoher Übereinstimmungsgewichte ein Todesfall aus dem Datensatz des EKR NRW automatisiert zugeordnet werden (Tabelle 2). Es traten keine manuell nachzubearbeitenden Zuordnungen auf. Allerdings stimmte bei 892 der Frauen mit automatisierter Zuordnung das in den beiden Datenbanken angegebene Sterbedatum nicht überein, wobei nur 39 dieser Fälle eine Abweichung von über einem Monat aufwiesen.

Unter allen Frauen, die im probabilistischen Abgleich eine tagesgenaue Übereinstimmung des Todesdatums aufwiesen (N = 23,288), wurden N = 23.034 (98,91%) im deterministischen Verfahren genauso zugeordnet wie im probabilistischen Ansatz. Weitere 3 Fälle erhielten jeweils zwei Treffer von verschiedenen beim EKR gemeldeten Verstorbenen und 251 Fälle erhielten trotz übereinstimmendem Todesdatum keinen Treffer im deterministischen Verfahren. Keinem der Fälle mit Abweichungen zwischen Krankenkassendaten-basierten und amtlichen Todesdatum (siehe unten) wurde im deterministischen Verfahren ein Datensatz des EKR zugeordnet.

Insgesamt wurden 23,034 (90,23%) Verstorbene aus GePaRD mit beiden Verfahren dem gleichen Todesfall im EKR NRW zugeordnet, weitere 799 (3,13%) erhielten nur im deterministischen und 892 (3,49%) nur im probabilistischen Verfahren einen Treffer. Nur für 549 (2,15%) Frauen der Stichprobe fand sich mit keinem der beiden Verfahren ein Todesfall aus dem EKR.

Insgesamt erreichten beide Abgleichverfahren ähnlich hohe Trefferquoten (TQ): probabilistisch=94,72%, deterministisch=93,36%.

Diskussion

Die in der epidemiologischen Forschung, vor allem in der Versorgungsforschung, zunehmend häufiger verwendeten Abrechnungsdaten der Gesetzlichen Krankenkassen (GKV) umfassen keine genauen Informationen zur TU. Die vorliegende Studie zeigt, dass mit zwei verschiedenen Abgleichverfahren hohe eindeutige TQ zwischen einer GKV-Datenbank und einem Mortalitätsregister beim EKR NRW erreicht werden können. Diese betrugen 94,72% bei dem probabilistischen Verfahren, das sich weitgehend an die Modalitäten des pseudonymisierten Record Linkage-Verfahrens der deutschen EKR [8] orientierte. Bei dem vereinfachten deterministischen Verfahren, das nur vier Klartextmerkmale und lediglich den Tag des Geburtsdatums verschlüsselt zum Abgleich nutzte, wurde eine TQ von 93,36% erreicht. Damit konnte bei fast allen, laut Informationen der GKV-Daten vermuteten ‚Todesfällen' der Tod nicht nur bestätigt werden, sondern die eindeutige Zuordnung erlaubte auch die Verknüpfung mit einer amtlichen TU. Die Studie zeigt exemplarisch am Abgleich von GePaRD und EKR NRW, dass eine Anreicherung von GKV-Daten mit den Mortalitätsinformationen aus Epidemiologischen Krebsregistern grundsätzlich eine realistische Option ist.

Hierbei gilt es allerdings, verschiedene Aspekte angemessen zu berücksichtigen. So ist bekannt, dass die amtliche TU keine fehlerfreie oder stets valide Information darstellt. Besonders bei außerhalb des Krankenhauses Verstorbenen gilt die Erfassung der zum Tode führenden Umstände als unsicher oder unvollständig [15, 16]. Andererseits ist die amtliche TU gerade für regionale oder strukturelle Vergleiche, auch internationaler Art, oft durchaus informativ. Insbesondere bei der TU Krebs, und insbesondere bei der im Rahmen der hier durchgeführten Machbarkeitsstudie im Fokus stehenden Brustkrebsmortalität, ist davon auszugehen, dass die Erkrankung aufgrund der Vorlaufzeit bekannt und die Zuordnung der TU wegen der Diagnose und Therapie der fortgeschrittenen Tumorstadien, vor allem der Metastasierung, von hinreichend großer Validität sein dürfte.

Durch die Einbeziehung sämtlicher Meldeämter von NRW in das elektronische Meldenetzwerk des EKR NRW kann von einer vollständigen Erfassung von Sterbefällen mit Wohnsitz in NRW ausgegangen werden. Dies betrifft auch außerhalb von NRW in Deutschland bzw. im Ausland verstorbene Fälle, da das für den Wohnsitz zuständige Meldeamt die jeweilige Sterbeurkunde erhält.

Es ist weiterhin davon auszugehen, dass in den Sterbeurkunden die Personen-identifizierenden Merkmale wie z.B. Vor- und Nachnamen nicht durchgehend fehlerfrei erfasst werden und es ist zu vermuten, dass auch eine gewisse Fehlerrate bei den Stammdaten der Krankenkassen vorliegt. Aufgrund dieser Datenfehler war a priori nicht zu erwarten, dass mit den Abgleichverfahren eine vollständige TQ erreicht würde, so dass die erreichte TQ von deutlich über 90% nahe dem potentiell in diesem Abgleich erreichbaren Maximum lag.

Einer der wichtigsten Unterschiede zwischen den beiden Abgleichverfahren bestand darin, dass im deterministischen Ansatz das Identifizierungsmerkmal 'Datum des Todes' genutzt wurde, während diese Angabe im probabilistischen Verfahren nicht gefordert war. Bei 892 Frauen der Abgleichstichprobe (etwa 3,5%) mit einem probabilistischer Treffer ergab sich kein Treffer mit dem deterministischen Verfahren: in allen diesen Fällen wich das Todesdatum aus den GKV-Daten vom amtlichen Todesdatum ab. Weiterhin traten nur 3 Fälle mit doppelten Zuordnungen auf, obwohl auf Seiten des EKR eine Datenquelle mit einer Vollerfassung sämtlicher Sterbefälle von NRW für den Abgleich zur Verfügung stand. Dies unterstreicht die hohe Selektivität der für die deterministische Methode gewählten Identifikatoren, so dass z.B. keine Treffer unter den Verstorbenen aus NRW gefunden wurden, wenn das Todesdatum inkorrekt war. Andererseits traten auch Fälle auf, bei denen trotz Übereinstimmung des Todesdatums im deterministischen Verfahren keine Zuordnung erfolgte, wohl aber im probabilistischen Verfahren (251 Fälle): dies lässt sich durch die höhere Anzahl genutzter Merkmale und die Zulassung von Toleranzen in der Übereinstimmung bei diesem Verfahren erklären. Demgegenüber sind Treffer mit dem deterministischem Verfahren, die aber probabilistisch nicht entdeckt wurden (799 Fälle), durch fehlende Übereinstimmungen in Merkmalen, die nur im probabilistischen Ansatz verwendet wurden, erklärbar. Hierbei kann es sich z.B. um seltene Ausprägungen dieser Merkmale handeln, die dann eine deutliche Verringerung der Übereinstimmungsgewichte bewirken und so zu keiner automatisierten Zuordnung führten.

Die Möglichkeit einer methodisch bedingten TQ-Senkung durch das probabilistische Verfahren ist als gering einzustufen, da eine unabhängige Evaluation des Verfahrens eine Synonymfehlerrate (fälschliche Nicht-Zuordnung) von 0,18% ermittelt hat [14]. Für die andere Fehlerart, bei der Datensätze fälschlich einer Person zugeordnet werden

(Homonymfehler), wurde eine Rate von nur 0,015% bestimmt. Dies legt nahe, dass es sich bei den 892 diskrepanten Fällen im probabilistischen und deterministischen Verfahren allenfalls zu einem sehr geringen Teil um Homonymfehler gehandelt hat. Vielmehr stellt das Austrittsdatum aus der Versicherung nur eine näherungsweise Angabe (‚Proxy') für das Sterbedatum bei verstorbenen Versicherten dar, da es keine amtlichen Meldungen von Sterbefällen oder eine Weitergabe des Sterbedatums an die Kassen gibt. Deshalb sind die Abweichungen im Todesdatum wahrscheinlich eher auf die unsichere Informationsbasis der Kassendaten als auf mögliche Homonymfehler beim Abgleich zurückzuführen.

Eine mögliche Modifizierung des deterministischen Verfahrens dahingehend, dass für eine Zuordnung auch Toleranzen in der Übereinstimmung z.B. des Todesdatums erlaubt sind, erscheint uns aber insgesamt nicht sinnvoll. Wenngleich bei einem größeren Teil der Fälle mit Abweichungen zwischen dem Todesdatum von Kassen und EKR nur geringe Differenzen vorliegen, würde die TQ durch die Modifikation potentiell nur gering verbessert werden, während die Homonymfehlerrate aber vermutlich steigen würde. Interessanterweise lag bei dem - hier von uns erstmalig verwendeten - deterministischen Verfahren die Homonymfehlerrate mit 0,012% (3 Fälle aus 25528) in der gleichen Größenordnung wie beim probabilistischen Verfahren.

In Deutschland wären eine Anwendung der von uns vorgestellten Verfahren für die Ergänzung von TU in Datenbanken mit Abrechnungsdaten wie z.B. GePaRD über einen Abgleich mit EKR derzeit grundsätzlich nur in den weiteren Bundesländern Rheinland-Pfalz, Berlin, Mecklenburg-Vorpommern, Bremen und NRW möglich, da es nur hier erste Strukturen eines integrierten Mortalitätsregisters gibt. Außerhalb von NRW wären darüber hinaus spezifische Anpassungen an die regionalen Gegebenheiten in den Datenflüssen der Verfahren erforderlich.

Ein probabilistisches Abgleichverfahren weist gegenüber einem deterministischen den Vorteil auf, dass auch nicht perfekt übereinstimmendende Datensätze einer Person zugeordnet werden können. Allerdings muss dafür im probabilistischen Verfahren auch eine größere Anzahl von Merkmalen genutzt werden, um eine hohe TQ kombiniert mit geringer Homonymfehlerrate zu erreichen. Das probabilistische Verfahren war in unserem Setting, in dem viele verschiedene Institutionen in den Datenfluss eingebunden waren, mit einem relativ hohen Aufwand bei Genehmigungen und Datenverarbeitung verbunden. Zum einen

besitzen Datenschutzaspekte durch die Vielzahl einzubeziehender Merkmale eine höhere Relevanz, zum anderen ist die Extraktion, doppelte Verschlüsselung und Zusammenführung, die für jede Kasse und jedes Register jeweils neu aufgesetzt und durchgeführt werden muss, wesentlich aufwändiger. Zudem ist das in dieser Studie genutzte Datenexport-Tool EpiCan nicht außerhalb von NRW implementiert. Das deterministische Verfahren bietet dagegen bei einer wesentlich geringeren Anzahl von erforderlichen, Personen-identifizierenden Merkmalen sowie einfacheren technischen Voraussetzungen eine vergleichbar hohe TQ. Nach den Ergebnissen dieser Studie ist es im Rahmen des von uns zu nutzenden Settings als Verfahren der Wahl für die Anreicherung von Todesdaten in GKV-Datensätzen vorzuschlagen.

Literatur

1. Swart E, Ihle P. Sekundärdatenanalyse: Aufgaben und Ziele. In: Swart E, Ihle P, Gothe H, Matusiewicz D, Hrsg. Routinedaten im Gesundheitswesen. Handbuch Sekundärdatenanalyse: Grundlagen, Methoden und Perspektiven. 2., vollst. überarb. Aufl. Bern: Verlag Hans Huber; 2014: 16-18

2. Fuhs A, Bartholomäus S, Heidinger O, Hense HW. [Evaluation of the impact of the mammography screening program on breast cancer mortality: feasibility study on linking several data sources in North Rhine-Westphalia]. Bundesgesundheitsblatt Gesundheitsforschung Gesundheitsschutz 2014; 57(1):60-67

3. Jobski K, Kollhorst B, Garbe E, Schink T. The Risk of Ischemic Cardio- and Cerebrovascular Events Associated with Oxycodone-Naloxone and Other Extended-Release High-Potency Opioids: A Nested Case-Control Study. Drug Saf 2017;40(6):505-515

4. Dörks M, Langner I, Timmer A, Garbe E. Treatment of paediatric epilepsy in Germany: antiepileptic drug utilisation in children and adolescents with a focus on new antiepileptic drugs. Epilepsy Res 2013;103(1):45-53

5. Mikolajczyk RT, Kraut AA, Garbe E. Evaluation of pregnancy outcome records in the German Pharmacoepidemiological Research Database (GePaRD). Pharmacoepidemiol Drug Saf 2013;22(8):873-80

6. Mikolajczyk RT, Schmedt N, Zhang J, Lindemann C, Langner I, Garbe E. Regional variation in caesarean deliveries in Germany and its causes. BMC Pregnancy Childbirth 2013;13:99

7. Bertram H, Heidinger O, Kajüter H, Khil L, Krieg V, Kühling L, Mattauch V. Jahresbericht 2016 - Krebsgeschehen in Nordrhein-Westfalen 2014. Münster: Landeskrebsregister NRW gGmbH (Hrsg.) 2017

8. Meyer M. Kontrollnummern und Record Linkage. In: Hentschel S, Katalinic A, Hrsg. Das Manual der epidemiologischen Krebsregistrierung. München, Wien, New York: W. Zuckerschwerdt Verlag 2008:57–68.

9. Fellegi IP, Sunter AB. A theory for record linkage. Journal of the American Statistical Association 1969;64(328):1183-1210

10. Kajüter H, Batzler WU, Krieg V, Heidinger O, Hense HW. [Linkage of secondary data with cancer registry data on the basis of encrypted personal identifiers – results from a pilot study in North Rhine-Westphalia]. Gesundheitswesen 2012;74(8-9):e84-e89

11. Heidinger O, Batzler WU, Krieg V, Weigel S, Biesheuvel C, Heindel W, Hense HW. [The incidence of interval cancers in the German mammography screening program – results from the population-based cancer registry in North Rhine-Westphalia.] Dtsch Arztebl Int 2012;109(46):781-787

12. Kajüter H, Geier AS, Wellmann I, Krieg V, Fricke R, Heidinger O, Hense HW. [Cohort study of cancer incidence in patients with type 2 diabetes: record linkage of encrypted data from an external cohort with data from the epidemiological Cancer Registry of North Rhine-Westphalia]. Bundesgesundheitsbl 2014;57:52-59

13. Kampfenkel T, Arning A, Heidinger O, Jürgens H, Koch R. [Treatment of colorectal cancer in certified bowel cancer centers. A retrospective observational study with innovative data acquisition]. Onkologe 2016;22:984-991

14. Schmidtmann I, Sariyar M, Borg A, Gerold-Ay A, Heidinger O, Hense HW, Krieg V, Hammer GP. Quality of record linkage in a highly automated cancer registry that relies on encrypted identity data. GMS Med Inform Biom Epidemiol 2016;12(1):Doc2

15. Schröder AS, Wilmes S, Sehner S, Ehrhardt M, Kaduszkiewicz H, Anders S. Post-mortem external examination: competence, education and accuracy of general practitioners in a metropolitan area. Int J Legal Med 2017. doi: 10.1007/s00414-017-1559-9

16. Schubert-Fritschle G, Eckel R, Eisenmenger W, Hölzel D. Qualität der Angaben von Todesbescheinigungen - Ist die Todesursachenstatistik zu Krebserkrankungen besser als ihr Ruf? Dtsch Arztebl 2002;99(1–2):50-55

Abbildungslegenden

Abbildung 1: Datenflussschema für die Anreicherung der GKV-Daten (GePaRD) am BIPS mit Todesursachen vom Mortalitätsregister des Epidemiologischen Krebsregisters in Nordrhein Westfalen (EKR NRW) (weitere verwendete Abkürzungen: GePaRD = Pharmako-epidemiologische Forschungsdatenbank; KID = Kommunikations-ID; BIPS = Leibniz-Institut für Präventionsforschung und Epidemiologie; IGMR = Institut für Informations-, Gesundheits- und Medizinrecht; PID = Personen-identifizierende Daten; PKG = Personenkryptogramme; PSN = Patientenpseudonyme; KVWL = Kassenärztliche Vereinigung Westfalen-Lippe; PSD = Pseudonymisierungsdienst)

Tabellenlegenden

Tabelle 1: Beim Abgleich genutzte Merkmale im probabilistischen und im deterministischen Verfahren (bei den verschlüsselt verwendeten Merkmalen ist jeweils angegeben, in wie viele Patientenpseudonyme (PSN) der Quelltext des Merkmals bei der Verschlüsselung umgewandelt wurde)

Tabelle 2: Anzahlen und Quoten für korrekte, automatisiert erfolgte Zuordnungen beim Abgleich zwischen GePaRD-Datensätzen und EKR NRW-Datensätzen für das probabilistische und das deterministische Verfahren

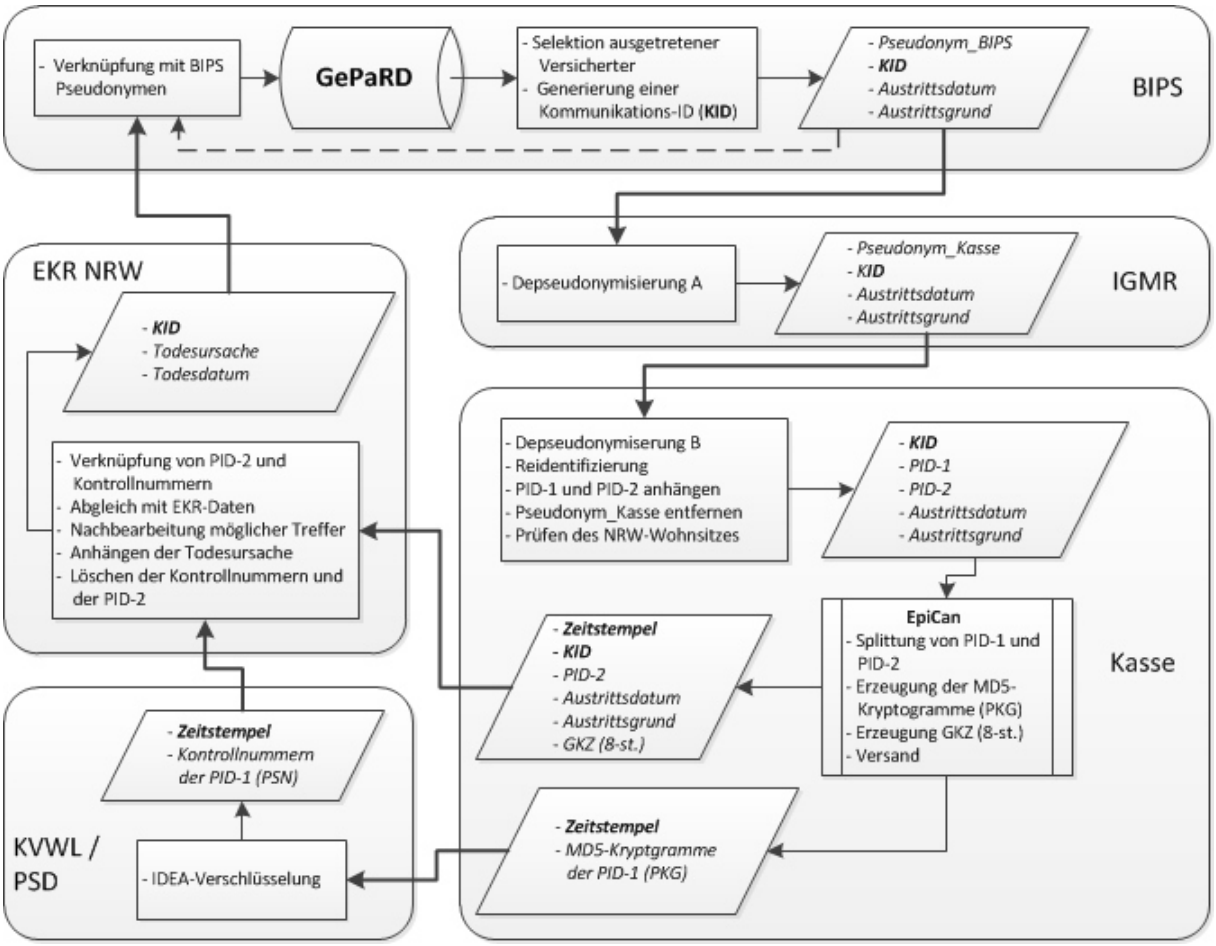
- Verknüpfung mit BIPS Pseudonymen
GePaRD
- Selektion ausgetretener Versicherter
- Generierung einer Kommunikations-ID (KID)
- Pseudonym_BIPS
- KID
- Austrittsdatum
- Austrittsgrund
BIPS
- Depseudonymisierung A
- Pseudonym_Kasse
- KID
- Austrittsdatum
- Austrittsgrund
IGMR
EKR NRW
- KID
- Todesursache
- Todesdatum
- Verknüpfung von PID-2 und Kontrollnummern
- Abgleich mit EKR-Daten
- Nachbearbeitung möglicher Treffer
- Anhängen der Todesursache
- Löschen der Kontrollnummern und der PID-2
- Depseudonymisierung B
- Reidentifizierung
- PID-1 und PID-2 anhängen
- Pseudonym_Kasse entfernen
- Prüfen des NRW-Wohnsitzes
- KID
- PID-1
- PID-2
- Austrittsdatum
- Austrittsgrund
EpiCan
- Splittung von PID-1 und PID-2
- Erzeugung der MD5-Kryptogramme (PKG)
- Erzeugung GKZ (8-st.)
- Versand
- Zeitstempel
- KID
- PID-2
- Austrittsdatum
- Austrittsgrund
- GKZ (8-st.)
Kasse
- Zeitstempel
- MD5-Kryptgramme der PID-1 (PKG)
- Zeitstempel
- Kontrollnummern der PID-1 (PSN)
KVWL / PSD
- IDEA-Verschlüsselung

Tabelle 1: Beim Abgleich genutzte Merkmale im probabilistischen und im deterministischen Verfahren (bei den verschlüsselt verwendeten Merkmalen ist jeweils angegeben, in wie viele Patientenpseudonyme (PSN) der Quelltext des Merkmals bei der Verschlüsselung umgewandelt wurde.)

Merkmal	**Genutzt im Verfahren**	
	Probabilistisch	**Deterministisch**
Nachname (zerlegt in bis zu 3 Segmente) verschlüsselt → 3 PSN	Ja	Nein
Phonem (standardisiert) des Nachnamens verschlüsselt → 1 PSN	Ja	Nein
Vorname(n) (zerlegt in bis zu 3 Segmente) verschlüsselt → 3 PSN	Ja	Nein
Phonem (standardisiert) des Vornamens verschlüsselt → 1 PSN	Ja	Nein
Geburtsname (zerlegt in bis zu 3 Segmente) verschlüsselt → 3 PSN	Ja	Nein
Phonem (standardisiert) des Geburtsnamens verschlüsselt → 1 PSN	Ja	Nein
Früherer Name (zerlegt in bis zu 3 Segmente) verschlüsselt → 3 PSN	Ja	Nein
Phonem (standardisiert) des früheren Namens verschlüsselt → 1 PSN	Ja	Nein
Titel (zerlegt in bis zu 2 Segmente) verschlüsselt → 2 PSN	Ja	Nein
Geschlecht	Ja	Ja
Jahr des Geburtsdatums	Ja	Ja
Monat des Geburtsdatums	Ja	Ja
Tag des Geburtsdatums verschlüsselt	Ja	Ja
Gemeindekennziffer (8-stellig)	Ja	Nein
Kreiskennziffer (5-stellig)	Nein	Ja
Sterbedatum	Nein	Ja

Tabelle 2: Anzahlen und Quoten für korrekte Zuordnungen beim Abgleich zwischen GePaRD-Datensätzen und EKR NRW-Datensätzen für das probabilistische und das deterministische Verfahren

Art der Zuordnung von EKR zu GePaRD Datensatz	**Abgleichverfahren**			
	probabilistisch		**deterministisch**	
	N	**%**	**N**	**%**
Eindeutiger Treffer zwischen GePaRD und EKR bei übereinstimmendem Todesdatum	23288	91,23	23833	93,36
Eindeutiger Treffer zwischen GePaRD und EKR bei nicht übereinstimmendem Todesdatum	892	3,49	0	0,00
GePaRD-Datensatz mit doppeltem Treffer von verschiedenen Personen beim EKR			3	0,01
GePaRD-Datensatz ohne Treffer	1348	5,28	1692	6,63
Gesamte Stichprobe	25528	100,00	25528	100,00