Thieme

Methodik und Zuordnungserfolg eines Linkage von Daten klinischer Krebsregister mit Abrechnungsdaten gesetzlicher Krankenkassen

Methodology and Attribution Success of a Data Linkage of Clinical Registry Data with Health Insurance Data

Autorinnen/Autoren
Christoph Bobeth[1], Kees Kleihues-van Tol[2], Martin Rößler[1], Veronika Bierbaum[1], Michael Gerken[3], Christian Günster[4], Patrik Dröge[4], Thomas Ruhnke[4], Monika Klinkhammer-Schalke[2, 3], Jochen Schmitt[1], Olaf Schoffer[1]

Institute
1 Zentrum für Evidenzbasierte Gesundheitsversorgung, Universitätsklinikum und Medizinische Fakultät Carl Gustav Carus an der Technischen Universität Dresden, Dresden, Germany
2 Arbeitsgemeinschaft Deutscher Tumorzentren e.V. (ADT), Berlin, Germany
3 Tumorzentrum Regensburg, Institut für Qualitätssicherung und Versorgungsforschung, Universität Regensburg, Regensburg, Germany
4 Wissenschaftliches Institut der AOK, Berlin, Germany

Schlüsselwörter
Datenlinkage, Klinische Krebsregisterdaten, GKV-Routinedaten, indirekte Identifikatoren, Linkagevalidierung

Key words
clinical cancer registry data, public health insurance data, data linkage, indirect identifiers, validation of linkage

online publiziert 2023

Bibliografie
Gesundheitswesen 2023; 85 (Suppl. 2): S154–S161
DOI 10.1055/a-1984-0085
ISSN 0949-7013

Georg Thieme Verlag, Rüdigerstraße 14,
70469 Stuttgart, Germany

Korrespondenzadresse
Christoph Bobeth
Universitätsklinikum Carl Gustav Carus Dresden,
Zentrum für Evidenzbasierte Gesundheitsversorgung
Fetscherstraße 74
01307 Dresden
Germany
christoph.bobeth@uniklinikum-dresden.de

Zusätzliches Material zu diesem Beitrag finden Sie unter https://doi.org/10.1055/a-1984-0085

ZUSAMMENFASSUNG

Hintergrund Das vom Innovationsfonds geförderte Projekt „Wirksamkeit der Versorgung in onkologischen Zentren“ (WiZen) ist ein breit angelegtes Projekt zur Erforschung der Wirksamkeit von Zertifizierungen in der Onkologie. Im Rahmen des Projektes werden bundesweite Daten der AOKen und Daten Klinischer Krebsregister aus verschiedenen Bundesländern für die Jahre 2006–2017 verwendet. Zur Kombination der Stärken beider Datenquellen werden diese für acht verschiedene Krebsentitäten datenschutzkonform miteinander verknüpft.

Methoden Das Datenlinkage erfolgte dabei anhand indirekter Identifikatoren und wurde mittels der Krankenversichertennummer als direktem Identifikator und Goldstandard validiert. Dies ermöglicht die Quantifizierung von Potenzial und Qualität verschiedener Linkage-Varianten. Als Kriterien zur Bewertung der Zuordnungen wurden Sensitivität und Spezifität sowie Treffergenauigkeit und Treffergüte genutzt. Die durch das Linkage resultierenden Verteilungen relevanter Variablen wurden anhand der ursprünglichen Verteilungen in den Einzeldatensätzen validiert.

Ergebnisse Je nach Kombination indirekter Identifikatoren ergab sich eine Bandbreite von 22.125 bis 3.092.401 Linkage-Treffern. Eine nahezu perfekte Verknüpfung der betrachteten Daten konnte durch die Kombination von Informationen zu Entitätsart, Geburtsdatum, Geschlecht und Postleitzahl der Personen erreicht werden. Insgesamt wurden mit den genannten Merkmalen 74.586 eineindeutige Verknüpfungen und für die verschiedenen Entitäten eine mediane Treffergüte von mehr als 98 % erreicht. Die Alters- und Geschlechtsverteilungen der verschiedenen Datenquellen sowie die verknüpften Sterbedaten wiesen zudem eine hohe Übereinstimmung auf.

Diskussion und Schlussfolgerung GKV- und Krebsregisterdaten lassen sich mit hoher interner und externer Validität auf Individualdatenebene verknüpfen. Die stabile Verknüpfung ermöglicht durch den simultanen Zugang zu Variablen beider Datensätze („das Beste aus beiden Welten“) gänzlich neue Ana-

lysemöglichkeiten: Für einzelne Personen stehen nun sowohl Informationen zum UICC-Stadium der Erkrankung aus den Registern als auch Komorbiditäten aus den GKV-Daten zur Verfügung. Durch die Verwendung gut verfügbarer Linkagevariablen und den hohen Verknüpfungserfolg ist das Verfahren vielversprechend für künftige Linkages in der Versorgungsforschung.

ABSTRACT

Background The aim of the project "Effectiveness of care in oncological centres" (WiZen), funded by the innovation fund of the federal joint committee, is to investigate the effectiveness of certification in oncology. The project uses nationwide data from the statuory health insurance AOK and data from clinical cancer registries from three different federal states from 2006–2017. To combine the strengths of both data sources, these will be linked for eight different cancer entities in compliance with data protection regulations.

Methods Data linkage was performed using indirect identifiers and validated using the health insurance's patient ID („Krankenversichertennummer") as a direct identifier and gold standard. This enables quantification of the quality of different linkage variants. Sensitivity and specificity as well as hit accuracy and a score addressing the quality of the linkage were used as evaluation criteria. The distributions of relevant variables resulting from the linkage were validated against the original distributions in the individual datasets.

Results Depending on the combination of indirect identifiers, we found a range of 22,125 to 3,092,401 linkage hits. An almost perfect linkage could be achieved by combining information on cancer type, date of birth, gender and postal code. A total of 74,586 one-to-one linkages were achieved with these characteristics. The median hit quality for the different entities was more than 98 %. In addition, both the age and sex distributions and the dates of death, if any, showed a high degree of agreement.

Discussion and conclusion SHI and cancer registry data can be linked with high internal and external validity at the individual level. This robust linkage enables completely new possibilities for analysis through simultaneous access to variables from both data sets ("the best of both worlds"): Information on the UICC stage that stems from the registries can now be combined, for instance, with comorbidities from the SHI data at the individual level. Due to the use of readily available variables and the high success of the linkage, our procedure constitutes a promising method for future linkage processes in health care research.

Einführung

In Krebsregistern werden sowohl Personencharakteristika wie Alter und Geschlecht als auch erkrankungsspezifische Variablen wie Tumorcharakteristika und Therapien gemäß gesetzlichen Regelungen erfasst [1]. Aufgrund der Meldepflicht bieten die Daten gesetzlicher Krebsregister grundsätzlich eine solide Grundlage zur Darstellung des Erkrankungsgeschehens und zur validen Schätzung epidemiologischer Kennzahlen von Krebserkrankungen.

Eine weitere Datenquelle zur Erforschung von Krebserkrankungen sind die Abrechnungsdaten der gesetzlichen Krankenversicherungen (GKV). Diese Daten bieten individuelle Informationen zu diagnostizierten Erkrankungen sowie zur Inanspruchnahme von Leistungen des Gesundheitssystems – im Idealfall mit mehrjähriger personenbezogener Historie. So können insbesondere Komorbiditäten und Versorgungswege in Analysen berücksichtigt werden. Da GKV-Daten jedoch nicht primär zu Forschungszwecken erhoben werden, stehen einige relevante Informationen nicht oder nur eingeschränkt zur Verfügung - für Krebserkrankungen beispielsweise Informationen zur histologischen Sicherung und zum UICC-Stadium. Unschärfen in der Kodierung können insbesondere epidemiologische Analysen auf Basis von GKV-Daten erheblich erschweren [2, 3].

In dieser Studie wurde die Machbarkeit einer Verknüpfung (Record-Linkage) von Daten Klinischer Krebsregister (KKR) mit GKV-Daten untersucht, der Linkageerfolg verschiedener Linkagevarianten quantifiziert und die Validität der erhaltenen Datenbasis überprüft.

Methoden

WiZen-Projekt

Das vom Innovationsfonds des G-BA geförderte WiZen-Projekt ist eine Studie zur Quantifizierung der „Wirksamkeit der Versorgung in onkologischen Zentren" (Förderkennzeichen 01VSF17020). Im Projekt werden Abrechnungsdaten der AOKs sowie Daten klinischer Krebsregister genutzt. Hauptziel der Studie ist ein Vergleich der Krankenhäuser mit und ohne Zertifikat hinsichtlich des Überlebens. Die Studie betrachtet Brustkrebs, kolorektales Karzinom, gynäkologische Tumoren, Kopf- und Halstumoren, Lungenkrebs, neuroonkologische Tumoren, Bauchspeicheldrüsenkrebs und Prostatakrebs. Der hier betrachtete Projektteil betrifft die Verknüpfung von Abrechnungsdaten der AOKs und KKR-Daten mit Ziel, ein geeignetes Linkageverfahren zu etablieren und zu validieren.

Datenquellen

Die im Linkage genutzten Routinedaten der KKR Dresden, Erfurt und Regensburg der Jahre 2006–2017 (Datenstand 08/2020) umfassen die Einzugsgebiete Thüringen, Ostsachsen, Niederbayern und die Oberpfalz und schließen Fälle mit histologischer Sicherung (keine DCO-Fälle) ein. Sie beinhalten Personeninformationen (Geburtsdatum, Geschlecht, Postleitzahl, Vitalstatus, verschlüsselte Krankenversichertennummer KV-Nr) sowie erkrankungsspezifische Daten (Diagnosedatum, Tumorstadium nach TNM und UICC, Therapien, Nachsorgedaten) und wurden für acht Entitäten bereitgestellt.

Die vom Wissenschaftlichen Institut der AOK (WIdO) an die Vertrauensstelle übermittelten bundesweiten Daten AOK-versicherter Personen enthalten potentiell inzidente Fälle der Jahre 2009–2017

für acht Entitäten: Im Datensatz befinden sich nur Patient:innen, die in den Jahren 2006–2008 keine stationäre ICD-10-Ziffer einer Entität erhalten haben. Die Festlegung der Definition „Entität" erfolgte durch ein Gremium von klinischen Experten (Spl.Tab. 1 für GKV und KKR, online verfügbar). Enthalten sind personenbezogene Informationen (Geburtsdatum, Geschlecht, Todesdatum, verschlüsselte (KV-Nr)), sowie erkrankungsspezifische Daten (z. B. Diagnosen nach ICD10-GM und Prozeduren in Form von OPS-Codes).

Datenschutz und Ethik

Die Daten der AOKs wurden mit den Daten der DKG zur Zertifizierung und den Daten der strukturierten Qualitätsberichte verknüpft und personen- und krankenhausseitig pseudonymisiert. Die Vertrauensstelle bei der Arbeitsgemeinschaft Deutscher Tumorzentren (ADT) verfuhr entsprechend mit den Daten der KKR. Diese Daten wurden verschlüsselt über gesicherte Austauschlaufwerke als CSV-Dateien übermittelt. Die zum Abgleich verwendete Krankenversicherungsnummer wurde auf Seiten des WIdO und auf Seiten der KKR mit demselben Verfahren verschlüsselt, und die Chiffrate der Vertrauensstelle zur Verfügung gestellt.

In der Vertrauensstelle bei der ADT wurde die eigentliche Datenverknüpfung (Linkage) durchgeführt und das Resultat in Form einer Treffertabelle an die Auswertestelle beim Zentrum für evidenzbasierte Gesundheitsversorgung (ZEGV) übermittelt. Es wurden folgende Datenbereinigungen durchgeführt: Die Vertrauensstelle vereinheitlichte das Datumsformat aller Quellen. Personen ohne vollständiges Geburtsdatum/Geschlecht (n = 13, n = 4) wurden aus dem Datensatz entfernt. Unvollständige Datumsangaben jenseits des Geburtsdatums (z. B. das OP-Datum) wurden seitens der ADT um Stichtage ergänzt und entsprechend markiert. Das Linkage auf Geburtsdatum und Geschlecht bei der Vertrauensstelle fand nur anhand gültiger Angaben statt. Die Vertrauensstelle führte keine weiteren Ausschlüsse durch. Die zugehörigen Datenflüsse und das methodische Vorgehen beim Linkage sind in ▶ **Abb. 1** zusammengefasst.

Für das WiZen-Projekt liegt ein Ethikvotum der Ethikkommission der TU Dresden vor (EK95022019). WiZen wurde bei ClinicalTrials.gov registriert (Identifier: NCT04334239). Die Datenverarbeitung und -analyse erfolgte in Übereinstimmung mit der Deklaration von Helsinki und der General Data Protection Regulation der Europäischen Union.

Datenlinkage

Das Datenlinkage erfolgt als Abgleich personenidentifizierender Merkmale, sogenannter Identifikatoren. Direkte Identifikatoren kennzeichnen ein Individuum eindeutig, wodurch eine sichere Verknüpfung von Individualdaten möglich ist. Indirekte Identifikatoren kennzeichnen bestimmte Charakteristika von Individuen. Da verschiedene Individuen aber teilweise identische Charakteristika haben (z. B. Geburtsdatum, Geschlecht), ist eine diesbezügliche Verknüpfung nur mit Unsicherheit möglich. Hier wurde ein Linkage anhand indirekter Identifikatoren (kurz indirektes Linkage) realisiert und anhand des direkten Identifikators validiert. Als Goldstandard hinsichtlich des Datenlinkage wurde die Zuordnung über KV-Nr definiert, ein sog. exaktes Linkage [4, 5].

Als indirekte Identifikatoren wurden Kombinationen von Geburtsdatum (tagesgenau), Geschlecht, 5-stelliger Postleitzahl (PLZ), Diagnosejahr und -quartal (von stationäre(n) Diagnose(n)), Krankenhausidentifikation (behandelndes Krankenhaus), ICD-Ziffer (ICD 3-Steller der stationären Hauptdiagnose) und OP-Datum (tagesgenau) und als direkter Identifikator die verschlüsselte KV-Nr genutzt. Darüber hinaus kamen Bundesland und Kreis (Wohnort) als einseitiger Filter auf der GKV-Kohorte zum Einsatz.

Bei der Auswertestelle wurde zwecks Datenschutz und -sparsamkeit lediglich eine Treffertabelle erstellt. Die Validierung des Linkage erfolgte anhand dieser Treffertabelle. Für das beste Linkage-Ergebnis wurden zusätzliche Informationen zum Abgleich der entsprechenden Verteilungen übermittelt.

Identifikation der besten Linkage-Variablenkombination

Die Bestimmung der Zuordnungsgüte für das Linkage erfolgte anhand einer von der Vertrauensstelle erstellten Treffertabelle (▶ **Abb. 1**).

Diese beinhaltete das kartesische Produkt der jeweiligen Ausgangspopulationen aus der GKV- und der KKR-Datenquelle – eingeschränkt auf mögliche Zuordnungen („hits") mit übereinstimmendem Geschlecht und Geburtsdatum, getrennt nach Entität. Diese beiden Variablen bilden daher eine „feststehende" Ausgangsbedingung, die sog. Grundmenge über Geburtsdatum und Geschlecht. Für jeden Identifikator kennzeichnete eine boolesche Variable die Übereinstimmung der zugehörigen Information in beiden Datensätzen. Diese booleschen Treffermarker wurden für folgende Variablen erstellt: verschlüsselte KV-Nr, Postleitzahl, Diagnosejahr/-quartal, Krankenhauspseudonym, Region, Kreis, ICD-3-Steller, OP-Datum.

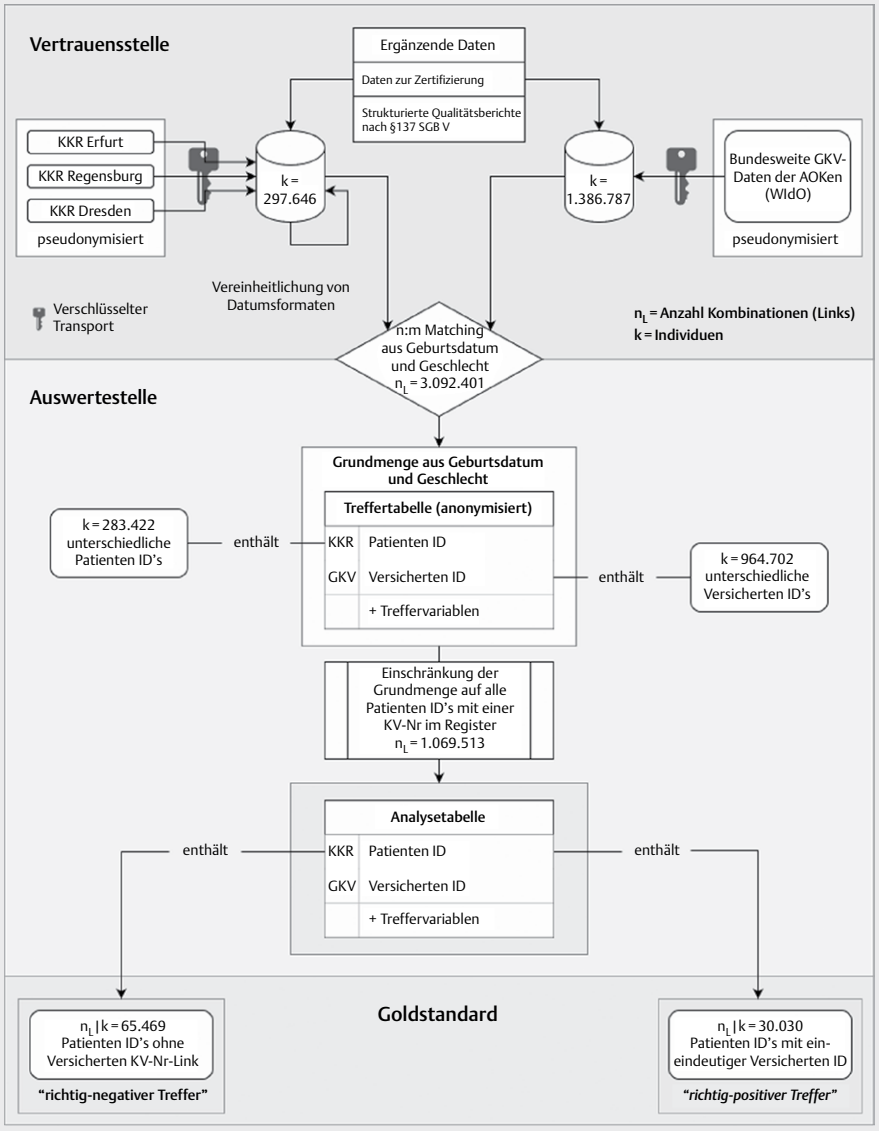
Zum Abgleich der verschiedenen Variablenkombinationen mit dem Goldstandard umfasste die sog. Analysepopulation alle Fälle mit einer dokumentierten KV-Nr in den KKR (zugehörige Populationen in Spl. **Abb. 1**, online verfügbar). Da in den GKV-Daten die KV-Nr vollständig erfasst sind, enthält diese Population *echt positive* (KV-Nr Register- und GKV-seitig) und *echt negative* (KV-Nr nur registerseitig, definitiv nicht im GKV-Datensatz) Zuordnungen. Dubletten hinsichtlich der KV-Nr (n = 35, Spl.Tab. 2, online verfügbar) innerhalb der Analysepopulation wurden ausgeschlossen.

Statistische Analyse

Zur Quantifizierung der Zuordnungsgüte im indirekten Linkage wurden die dabei erreichten Zuordnungen in richtig positive (rp), falsch positive (fp), falsch negative (fn) und richtig negative (rn) bezüglich des Goldstandards KV-Nr unterteilt. Daraus wurden die gebräuchlichen Gütekriterien Sensitivität, Spezifität und Korrektklassifikationsrate (Treffergenauigkeit, ACC) abgeleitet. Ergänzend wurde der Gilbert-Skill-Score (GSS) verwendet [6]. Dieser ist ein fähigkeitskorrigiertes Maß der Prädiktionsgüte (Formel 1), das die Anzahl der Zufallstreffer (z) (Formel 2) berücksichtigt.

$$GSS = \frac{rp - z}{rp + fp + fn - z} \quad (1)$$

$$z = \frac{(rp + fp)(rp + fn)}{rp + fp + fn + rn} \quad (2)$$



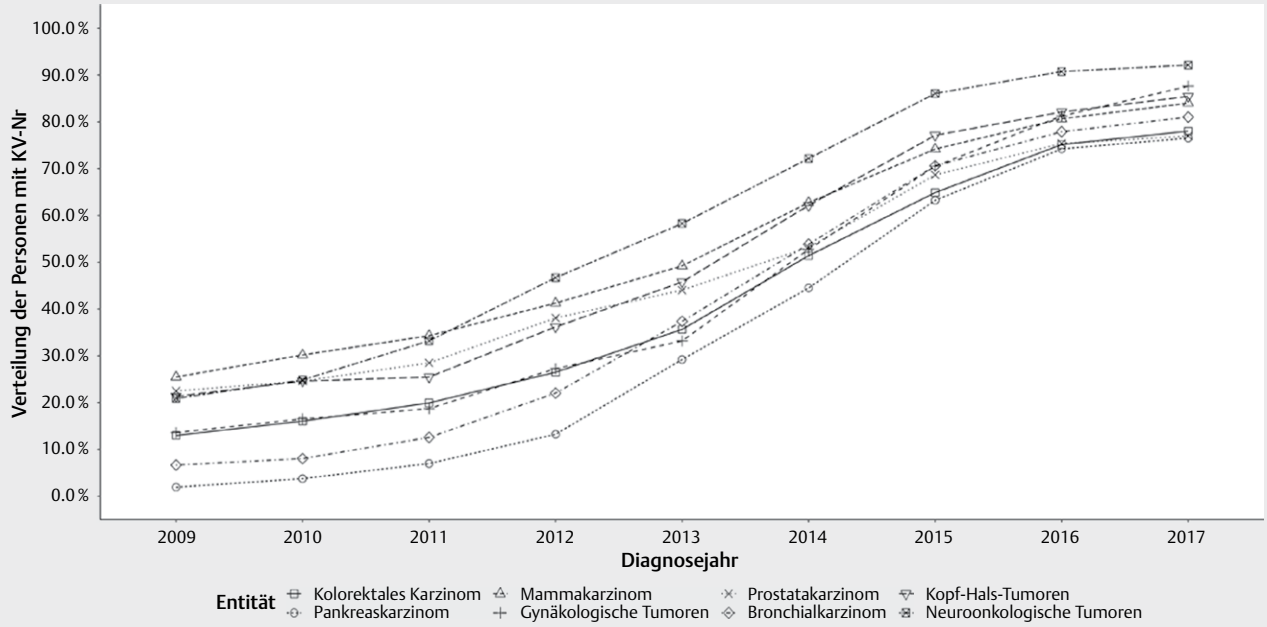
▶ **Abb. 1** Datenflussdiagramm und Linkageverfahren mit Darstellung ausgewählter Fallzahlen für die Ausgangsdatensätze, die Treffertabelle für die Grundmenge über Geburtsdatum und Geschlecht sowie Darstellung der Auswahl des Goldstandards.

Zudem wurde die Qualität des Linkage anhand der Häufigkeiten von Ein- bzw. Mehrdeutigkeiten bei Zuordnung beurteilt. (Ein-) Eindeutigkeit (1:1) lag vor, wenn ein Personenidentifikator aus der GKV-Datenquelle genau einem Personenidentifikator aus der KKR-Datenquelle zugeordnet wurde und dies auch umgekehrt galt. Mehrdeutigkeit lag vor, wenn ein Personenidentifikator entweder in der GKV- oder KKR-Datenquelle (1:m, n:1) oder in beiden Quellen (n:m) mehrfach zugeordnet wurde.

Als weitere Validierungsstufe unabhängig von der verschlüsselten KV-Nr wurden für alle eineindeutig indirekt gelinkten Individuen innerhalb der Population, die sich aus der Grundmenge über Geburtsdatum und Geschlecht ergibt, die Übereinstimmung der Sterbeinformation überprüft. Das Sterbedatum ist hierfür die einzig mögliche Variable, da sie in beiden Datensätzen identisch definiert ist. Dies ist für alle weiteren indirekten Variablen nicht der Fall. Des Weiteren wurde die Strukturgleichheit der Populations-Charakteristika für die GKV-Daten betrachtet.

In den KKR war das Sterbedatum mindestens monatsgenau, in den GKV-Daten tagesgenau dokumentiert. Individuen ohne dokumentiertes Sterbedatum wurden als „lebend" bezeichnet. Zunächst wurden vier mögliche Zustände festgelegt: 1) lebend (beide), 2) lebend/verstorben (GKV/KKR), 3) verstorben/lebend (GKV/KKR) und 4) verstorben (beide). In mindestens einer der Datenquellen als verstorben dokumentierte Zuordnungen wurden zusätzlich nach Sterbejahr dargestellt. Unterschieden wurde dabei hinsichtlich des Sterbedatums 1.) exakte Übereinstimmung, 2.) Übereinstimmung von Monat und Jahr, 3.) Abweichung in Monat und/oder Jahr 4./5.) nur in GKV bzw. in KKR als verstorben dokumentiert.

Als weitere Validierungsstufe erfolgte der Abgleich der deskriptiven Eigenschaften der Ausgangs- und der gelinkten Population



▶ **Abb. 2** Prozentuale Verteilung der Personen aus den klinischen Krebsregistern mit KV-Nr über das Jahr der Diagnose.

▶ **Tab. 1** Gegenüberstellung ausgewählter Variablenkombinationen der Analysepopulation hinsichtlich der Gütekriterien und Trefferanzahlen.

Ausgewählte Variablenkombinationen	Linkage-Kombination aus Geburtsdatum, Geschlecht und			
	Krankenhausinstitutskennzeichen und Diagnosequartal {hit_ik,hit_diag_jq}	Kreis und Diagnosejahr {hit_kr,hit_diagjahr}	OP-Datum und Region {hit_op,hit_reg}	Postleitzahl {hit_plz}
Anzahl in der Grundmenge Geburtsdatum und Geschlecht N (%)				
1:1 Verknüpfungen	46.905 (96,3 %)	38.813 (44,3 %)	38.343 (95,1 %)	74.586 (97,9 %)
1:n Verknüpfungen	1.573 (3,2 %)	35.562 (40,6 %)	1.808 (4,5 %)	1.266 (1,7 %)
n:m Verknüpfungen	234 (0,5 %)	13.172 (15,0 %)	148 (0,4 %)	322 (0,4 %)
Gesamtverknüpfungen	48.712 (100 %)	87.547 (100 %)	40.299 (100 %)	76.174 (100 %)
Anzahl innerhalb der Analysepopulation (KV-Nr im Register) N (%)				
1:1 Verknüpfungen	20.753 (96,9 %)	22.968 (58,2 %)	19.416 (95,6 %)	33.019 (98,9 %)
1:n Verknüpfungen	595 (2,8 %)	12.113 (30,7 %)	844 (4,2 %)	294 (0,9 %)
n:m Verknüpfungen	76 (0,4 %)	4.413 (11,2 %)	59 (0,3 %)	82 (0,2 %)
Gesamtverknüpfungen	21.424 (100 %)	39.494 (100 %)	20.319 (100 %)	33.395 (100 %)
Gütekriterien innerhalb der KV-Nr-Analysepopulation (KV-Nr im Register) Median (Min,Max)				
Sensitivität	71,5 % (42,2 %; 85,8 %)	92,1 % (82,2 %; 95,7 %)	57,0 % (25,4 %; 81,6 %)	99,5 % (99,4 %; 99,6 %)
Spezifität	99,9 % (99,8 %; 100,0 %)	99,0 % (98,8 %; 99,3 %)	99,9 % (99,7 %; 99,9 %)	99,9 % (99,8 %; 100,0 %)
Positiver Vorhersagewert	97,1 % (93,0 %; 98,8 %)	84,8 % (56,2 %; 94,1 %)	95,7 % (91,2 %; 98,3 %)	98,8 % (97,6 %; 99,3 %)
Treffergenauigkeit	98,2 % (96,8 %; 99,4 %)	98,6 % (98,0 %; 99,1 %)	97,6 % (90,1 %; 99,4 %)	99,9 % (99,8 %; 100,0 %)
Gilbert-Skill-Score	68,4 % (40,5 %; 82,5 %)	75,9 % (49,4 %; 88,4 %)	53,6 % (24,4 %; 76,8 %)	98,1 % (97,1 %; 98,8 %)

identisch mit der Vorgehensweise in [7]. Um im Rahmen dieser Validierung Abweichungen durch regionale Variation (z. B. in der Altersverteilung) auszuschließen, wurde der nicht gelinkte bundesweite GKV-Datensatz auf die Bundesländer Bayern, Sachsen und Thüringen eingeschränkt. Die resultierenden Verteilungen von Elixhauser-Komorbiditäten [8], Zentrenzugehörigkeit, Alter und Geschlecht wurden beispielhaft für das Pankreaskarzinom verglichen.

Ergebnisse

Die Datengrundlage bildeten bundesweit 1.386.811 AOK-Versicherte im GKV-Datensatz und 297.646 Individuen aus den KKR Erfurt, Dresden und Regensburg. Von diesen konnten 964.702 (GKV) bzw. 283.422 (KKR) über Geburtsdatum und Geschlecht zugeordnet werden, wodurch sich 3.092.401 Verknüpfungen ergaben (▶ **Abb. 1**, Spl.Tab. 3, online verfügbar). Diese Gruppe weist 98.499

eindeutige KKR-Individuen mit einer KV-Nr auf (Analysepopulation), von denen 33.030 eine Übereinstimmung mit der KV-Nr eines GKV-Versicherten hatten.

Der Anteil an Personen mit dokumentierter KV-Nr in den KKR nahm im Zeitverlauf bezüglich Diagnosejahr deutlich zu (▶ **Abb. 2**). So wiesen im Jahr 2009 zwischen 1,9 % (Pankreaskarzinom) und 25,5 % (Mammakarzinom) eine KV-Nr auf. Im Jahr 2017 lagen diese Anteile zwischen 76,5 % (Pankreaskarzinom) und 92,1 % (Neuroonkologische Tumoren).

Güte des Linkage

Die in ▶ **Tab. 1** dargestellten Variablenkombinationen erzielten im indirekten Linkage im Abgleich mit dem Goldstandard bezüglich Spezifität, positivem Vorhersagewert und Treffergenauigkeit Werte von mehr als 95 % (weitere ausgewählte Variablenkombinationen sind in Spl.Tab. 4 enthalten, online verfügbar). Es ist davon auszugehen, dass die Erfassung von Geburtsdatum, Geschlecht, PLZ und den anderen Kennziffern sich über den Untersuchungszeitraum nicht signifikant geändert hat.

Unterschiede zwischen den verschiedenen Linkage-Varianten wurden insbesondere anhand des GSS deutlich. Während das Linkage anhand Geburtsdatum, Geschlecht und PLZ für alle Entitäten GSS-Werte ≥ 97,1 % erreichte, lagen diese für alle Variablenkombination ohne PLZ-Bezug bei ≤ 88,4 %. Die PLZ-Linkage-Kombination bleibt mit einer Streuung von 1,7 % über alle Entitäten als einzige stabil.

Die hohe Güte des indirekten Linkages anhand Geburtsdatum, Geschlecht und PLZ spiegelte sich auch in der Anzahl von 74.586 eineindeutigen Verknüpfungen („1:1“), welche alle anderen Varianten übertraf wider (▶ **Tab. 1**). Mehrdeutige Links traten bei dieser Linkage-Variante selten auf („n:1“ bzw. „1:m“: 1.266 Links und „n:m“: 322 Links).

Die Verfügbarkeit der KV-Nr nimmt in den KKR über die Zeit zu (Spl.Tab. 2, online verfügbar). In einer Sensitivitätsanalyse wurde daher das Linkage eingeschränkt auf den Zeitraum 2014–2017. Hier waren die Sensitivität und GSS gegenüber dem Gesamtzeitraum etwas erhöht und die übrigen Gütekriterien vergleichbar (Spl. Tab. 5, online verfügbar).

Validierung

Nachfolgend wurden ausschließlich die 74.586 eineindeutigen Links über die Kombination mit der besten Güte, also Geburtsdatum, Geschlecht und PLZ betrachtet (unabhängig von der Existenz einer KV-Nr). Von diesen waren gemäß GKV-Daten 36.546 (49 %) und gemäß KKR-Daten 32.974 (44,2 %) als verstorben dokumentiert (▶ **Tab. 2**). 70.924 (95,1 %) wurden in beiden Datenquellen übereinstimmend als lebend bzw. verstorben klassifiziert. Für 3.617 Personen war in den GKV-Daten ein Sterbedatum dokumentiert, während in den KKR-Daten keine Sterbeinformation hinterlegt war. Umgekehrt wurden nur 45 Personen in den GKV-Daten als lebend und in den KKR-Daten als verstorben geführt.

Für die 36.591 Personen mit dokumentiertem Sterbedatum in wenigstens einer der Datenquellen zeigten sich zeitliche Trends in der Übereinstimmung der Sterbedaten (▶ **Abb. 3**). Der Anteil von Übereinstimmungen in Jahr und Monat reduzierte sich von 94 % im Jahr 2013 auf 75 % im Jahr 2017. Gleichzeitig stieg der Anteil ohne Sterbeinformation in den KKR-Daten von ≤ 4,6 % auf 24 %. Der Anteil in beiden Datenquellen dokumentierter aber abweichender Sterbedaten lag in allen Jahren zwischen 0,7 und 1,1 %.

Die Charakteristika des gelinkten Pankreaskarzinom-Datensatzes hinsichtlich Geschlecht und den Elixhausergruppen wichen von denen der ungelinkten GKV-Daten in 14 von 18 Fällen weniger als 3 %-Punkte ab (Spl.Tab. 6, online verfügbar).

▶ **Tab. 2** Vergleich der Sterbeinformation im GKV- und Registerdatensatz.

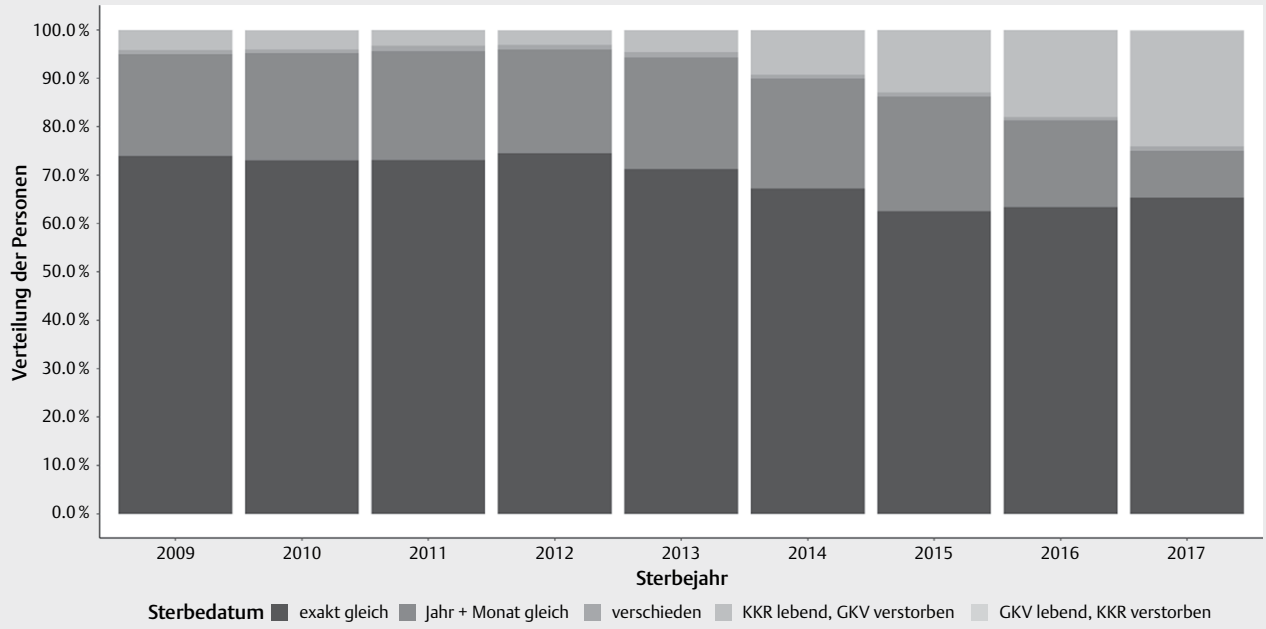
N (%)	GKV lebend	GKV verstorben	Gesamt
KKR lebend	37.995 (50,94 %)	3.617 (4,85 %)	41.612 (55,79 %)
KKR verstorben	45 (0,06 %)	32.929 (44,15 %)	32.974 (44,21 %)
Gesamt	38.040 (51,00 %)	36.546 (49,00 %)	74.586 (100,00 %)

Diskussion

Im Rahmen des WiZen-Projektes wurden erfolgreich zwei umfangreiche Datensätze mit komplementären Inhalten verknüpft – die Daten der AOKs sowie die Daten dreier KKR. Mit der Kombination von PLZ, Geburtsdatum und Geschlecht als indirekten Identifikatoren wurde für jede der betrachteten Krebsentitäten im Rahmen des Abgleichs mit dem direkten Identifikator KV-Nr eine hohe Treffergüte von mindestens 97,1 % erreicht, womit sie sich unter allen Varianten als überlegen erwies. Die sehr gute Verknüpfungsleistung bestätigte sich auch im Abgleich von Sterbedatumsangaben beider Quellen. Die Verteilungen personenbezogener Merkmale stimmten zwischen Linkage-Datensatz und GKV-Ausgangsdatensatz gut überein, sodass von einem geringen Verzerrungspotenzial der statistischen Eigenschaften der Kohorte aufgrund des Linkage auszugehen ist.

Vorarbeiten zum indirekten Linkage wiesen Erfolgsquoten von über 80 % [9–11] bzw. über 97 % [12] auf. Das vorliegende Linkage erreichte vergleichbare bzw. für einzelne Entitäten höhere Werte der Gütekriterien. Epidemiologische Krebsregister erreichen mittels Linkage über sogenannte Kontrollnummern [13] ebenfalls sehr valide Zuordnungen [14]. Ein solches Vorgehen für das Linkage von KKR- und GKV-Daten erfordert die Bereitschaft der Datenhalter, diese Kontrollnummern zu erzeugen. Insofern ist es als großer Erfolg zu werten, dass das vorliegende Linkage entitätsspezifisch mit nur drei indirekten Identifikatoren mit sehr geringer Fehlerrate umsetzbar ist. Dieses Ergebnis bestätigt frühere Evidenz, dass bereits mit wenigen Variablen eine valide Zuordnung möglich ist [15]. Zudem sind die hier verwendeten Identifikatoren generische Variablen, welche in vielen (anderen) Datenquellen enthalten sind, womit eine gute Übertragbarkeit des Vorgehens gegeben ist.

Durch das Linkage entstehen im Vergleich zu den Ursprungsdatensätzen eine Reihe analyserelevanter Vorteile. So sind für personenbezogene Analysen gleichzeitig gesicherte Angaben zu klinischen Charakteristika der Tumoren (Quelle: KKR) und relevanten Komorbiditäten (Quelle: GKV) verfügbar. Zudem können die oft aktuelleren und vollständigeren Angaben zu durchgeführten diagnostischen und therapeutischen Maßnahmen sowie zum Lebendstatus aus dem GKV-Datensatz den KKR-Datensatz insbesondere



▶ **Abb. 3** Abgleich der Sterbeinformation nach Jahr.

für prognostische Längsschnittbetrachtungen aufwerten. Andererseits erlaubt die Art der Dokumentation in KKR eine sehr zuverlässige Zuordnung des Datums der Neuerkrankung und erhöht so die Validität von Analysen durch die zuverlässige Unterscheidung inzidenter von prävalenten Erkrankungen gegenüber der alleinigen Nutzung von GKV-Daten.

Stärken und Limitationen

Die Daten wiesen eine hohe Qualität bezüglich der genutzten indirekten Identifikatoren auf. Das Geburtsdatum wurde in beiden Quellen auch als interne Prüfvariable genutzt und die Postleitzahl lag in den GKV-Daten historisiert vor. Seitens der KKR wird nur eine Postleitzahl erfasst, zu der kein explizites Datum vorliegt. Durch den Abgleich mit allen Postleitzahlen der GKVen wird hier die Wahrscheinlichkeit eines Treffers im Vergleich zu einzelnen stichtagsbezogenen PLZ in beiden Datensätzen erhöht. Zudem war die Erfassung von Geburtsdatum, Geschlecht und Postleitzahl in beiden Datenquellen vollständig. Somit konnten durch das Linkage die besten Informationen aus beiden Datenquellen valide miteinander kombiniert werden, sodass entweder fehlende oder unvollständige Angaben aufgefüllt oder gänzlich neue Sachverhalte aus der jeweils komplementären Quelle hinzugefügt werden konnten. Jedoch ist nicht immer eindeutig, welche Quelle bei überlappenden Variablen die zuverlässigere Angabe bietet. Bei der Verwendung der KV-Nr als Goldstandard hinsichtlich eines Linkage ist zu beachten, dass einerseits die Verfügbarkeit der KV-Nr in den KKR über die Zeit zugenommen hat. Andererseits kann ein Unterschied zwischen Individuen mit und ohne KV-Nr hinsichtlich der Belastbarkeit von Angaben der Identifikatoren zu einer eingeschränkten Übertragbarkeit der vorliegenden Ergebnisse auf die Gesamtpopulation führen. In einer diesbezüglichen Sensitivitätsanalyse wurde jedoch nur ein moderater Einfluss des Diagnosezeitpunkte auf den Erfolg des Datenlinkage festgestellt. Zudem ist das Problem nicht vorliegender KV-Nr in den KKR hauptsächlich für historische Daten relevant.

Der Linkage-Datensatz stellt die Schnittmenge beider Einzeldatensätze dar und weist somit eine geringere Fallzahl als diese auf. Das ist insbesondere relevant, wenn die Populationen beider Einzeldatenquellen sich nur teilweise überlappen (KKR: Einschränkung auf Einzugsgebiete/GKV: Einschränkung auf Versicherte der jeweiligen Krankenkasse). Dem Zugewinn an Informationen steht also eine Reduktion der Fallzahl entgegen, welche durch eine hohe Zuordnungsrate nur teilweise kompensiert werden kann.

Aufgrund von Datenschutzerfordernissen ist die Verknüpfung nur unter Einbezug einer Vertrauensstelle möglich. Die Verwendung indirekter Identifikatoren impliziert zudem eine gewisse Unsicherheit der Verknüpfung, wodurch das Verfahren nur für Anwendungen geeignet ist, welche keine absolute Sicherheit der Zuordnung erfordern. Die Fehlerquote ist allerdings so gering, dass sich der resultierende Datensatz für Kohortenstudien sehr gut eignet. Die Güte des Linkage kann zwischen verschiedenen Bundesländern variieren, beispielsweise aufgrund unterschiedlichen Mobilitätsverhaltens. Je nach Datenbasis könnte daher das Hinzuziehen weiterer Identifikatoren oder der Aufbau eines mehrstufigen Linkage erforderlich werden. Zudem ist zu beachten, dass das entitätsspezifische Vorgehen implizit als Linkage mit der Krebsentität als zusätzlichem indirekten Identifikator anzusehen ist.

Schlussfolgerung und Ausblick

Im Rahmen des WiZen-Projekts wurde mit wenigen indirekten Identifikatoren ein hoher Zuordnungserfolg des personenbezogenen Linkage erreicht und die Validität des erhaltenen Linkage-Da-

tensatzes bestätigt. Die Zusammenführung komplementärer Informationen aus verschiedenen Datenquellen auf Individualdatenebene bietet dabei ein hohes Potenzial für tiefergehende versorgungsbezogene Analysen gegenüber den Einzeldatenquellen. Das vorgestellte Verfahren unterstreicht somit das Potential dieser Verknüpfungsart für künftige Linkage-Verfahren in der Versorgungsforschung.

Inwiefern sich die zusammengeführten komplementären Informationen auf Analyseergebnisse auswirken und welche Angaben sich für die gelinkten Personen jeweils ergänzen lassen, wird in einer separaten Auswertung untersucht.

Danksagung

Wir danken Herrn Dr. med Udo Altmann für die Bereitstellung eines Moduls zur vereinfachten Datenlieferung auf Seiten der Krebsregister sowie Frau Dr. Anne Neumann für wertvolle Diskussionen zum Thema Datenschutz und Linkage.

Fundref Information

Innovationsfonds des gemeinsamen Bundesausschusses (G-BA) — 01VSF17020

Interessenkonflikt

CB, VB, JS und OS arbeiten an einem Universitätsklinikum mit zertifizierten Krebszentren, und MR hat in der Vergangenheit dort gearbeitet. Darüber hinaus erhielten sie während der Durchführung der Studie eine Förderung durch den Innovationsausschuss des Gemeinsamen Bundesausschusses. Unabhängig von dieser Studie erhielt JS institutionelle Zuschüsse für die Investigator-initiierte Forschung vom GBA, dem BMG, BMBF, der EU, dem Bundesland Sachsen, Novartis, Sanofi, ALK und Pfizer. Außerdem nahm er als bezahlter Berater für Sanofi, Lilly und die ALK an Advisory Board Meetings teil. Unabhängig von dieser Studie war OS als bezahlter Berater für Novartis tätig. Er ist außerdem Mitglied der Zertifizierungskommission „Hautkrebszentren" der Deutschen Krebsgesellschaft und Mitglied des Expertengremiums im Projekt „Entwicklung von Kriterien zur Bewertung von Zertifikaten und Qualitätssiegeln nach § 137a Abs. 3 Satz 2 Nr. 7 SGB V" für das Institut für Qualitätssicherung und Transparenz im Gesundheitswesen (IQTIG). Die anderen Autoren erklären, dass sie keinen Interessenkonflikt haben.

Literatur

[1] Möslein G, Haier J, Schlag PM. Klinische und epidemiologische Krebsregister. Der Onkologe 2013; 19: 1022–1024. doi:10.1007/s00761-013-2515-z

[2] Horenkamp-Sonntag D, Schneider U, Engel S et al. Validität von GKV-Routinedaten: In welchem Umfang muss bei der wissenschaftlichen Nutzung von Sekundärdaten die Daten-Qualität geprüft werden? Zeitschrift für Palliativmedizin 2014; 15:. doi:10.1055/s-0034-1374464

[3] Schoffer O, Roessler M, Datzmann T et al. Medical Care and Survival of Soft-Tissue and Bone Sarcoma Patients: Results and Methodological Aspects of a German Subnational Cohort Study Based on Administrative Healthcare Data. Oncol Res Treat 2021; 44: 103–110. doi:10.1159/000513178

Schubert I, Ihle P, Köster I et al. Datengutachten für das Deutsche Institut für Medizinische Dokumentation und Information (DIMDI). Gutachten: Daten für die Versorgungsforschung. Zugang und Nutzungsmöglichkeiten. In. Köln: PMV Forschungsgruppe; 2014

[4] March S, Andrich S, Drepper J et al. Good Practice Data Linkage. Gesundheitswesen 2019; 81: 636–650. doi:10.1055/a-0962-9933

[5] Schaefer JT. The Critical Success Index as an Indicator of Warning Skill. Weather and Forecasting 1990; 5: 570–575. doi:10.1175/1520-0434(1990)005<0570:Tcsiaa>2.0.Co;2

[6] Roessler M, Schmitt J, Bobeth C et al. Is treatment in certified cancer centers related to better survival in patients with pancreatic cancer? Evidence from a large German cohort study. BMC Cancer 2022; 22: 621. doi:10.1186/s12885-022-09731-w

[7] Elixhauser A, Steiner C, Harris DR et al. Comorbidity measures for use with administrative data. Med Care 1998; 36: 8–27. doi:10.1097/00005650-199801000-00004

[8] Hammill BG, Hernandez AF, Peterson ED et al. Linking inpatient clinical registry data to Medicare claims data using indirect identifiers. Am Heart J 2009; 157: 995–1000. doi:10.1016/j.ahj.2009.04.002

[9] Maier B, Wagner K, Behrens S et al. Deterministic record linkage with indirect identifiers: data of the Berlin Myocardial Infarction Registry and the AOK Nordost for patients with myocardial infarction. Gesundheitswesen 2015; 77: e15–e19. doi:10.1055/s-0034-1395642

[10] March S, Antoni M, Kieschke J et al. Quo Vadis Data Linkage in Germany? An Initial Inventory. Gesundheitswesen 2018; 80: e20–e31. doi:10.1055/s-0043-125070

[11] Rothe U, Müller G. Evaluation eines Strukturvertrages zur Inzidenz des Gestationsdiabetes auf der Basis von Sekundärdaten. Diabetologie und Stoffwechsel 2013; 8:. doi:10.1055/s-0033-1341725

[12] Hinrichs H. Bundesweite Einführung eines einheitlichen Record Linkage-Verfahrens in den Krebsregistern der Bundesländer nach dem KRG, Abschlußbericht des Projekts. In. Oldenburg 1999

[13] Thoben W, Appelrath HJ. Verschlüsselung personenbezogener und Abgleich anonymisierter Daten durch Kontrollnummern. In Verläßliche IT-Systeme. Wiesbaden: Vieweg + Teubner Verlag; 1995: 193–206. doi:10.1007/978-3-322-91094-3_13

[14] Rocher L, Hendrickx JM, de Montjoye YA. Estimating the success of re-identifications in incomplete datasets using generative models. Nat Commun 2019; 10: 3069. doi:10.1038/s41467-019-10933-3