Datenschutz in der Forschung

Verknüpfte Gesundheitsdaten und Datenschutz: (k)ein Widerspruch

Nicole Steck, Adrian Spoerri, Matthias Egger

Institut für Sozial- und Präventivmedizin (ISPM), Universität Bern

Mit der Verknüpfung von Gesundheitsdaten können in der medizinischen Forschung auf effiziente Weise wichtige Fragestellungen untersucht werden. Allerdings ist dies aus Datenschutzgründen häufig nicht möglich. Eine am Berner Institut für Sozial- und Präventivmedizin entwickelte Methode erlaubt die Verknüpfung sensibler Daten, ohne dass identifizierende Angaben ausgetauscht werden.

Das Erheben von neuen Gesundheitsdaten zu Forschungszwecken ist aufwendig, teuer und für die Teilnehmenden oft mit grossem zeitlichem Aufwand über viele Jahre und mit unangenehmen Untersuchungen verbunden. Es ist deshalb erstrebenswert, dass bestehende Gesundheitsdaten in der Forschung optimal genutzt und verknüpft werden. Damit können Qualität und Vollständigkeit der Daten verbessert und neue Fragestellungen untersucht werden. In longitudinalen Studien sind nach einigen Jahren oft nur noch ein Teil der ursprünglich in die Studie aufgenommenen Patienten unter Beobachtung, was zu Verzerrungen in den Studienresultaten führen kann (Selektionsbias). Dies kann verhindert werden, wenn der Vitalstatus aller ursprünglichen Patienten durch Verknüpfung mit den Sterblichkeitsdaten bestimmt wird. Eine Langzeitstudie der Wirksamkeit präventiver Massnahmen bei älteren Patienten in Hausarztpraxen in Solothurn konnte auf diese Weise den Vitalstatus von 98,2% der Studienteilnehmer eruieren [1]*. Ein anderes Beispiel für eine erfolgreiche *Record-Linkage*-Studie ist die Verknüpfung der Daten der Volkszählungen 1990 und 2000 mit dem Schweizerischen Kinderkrebsregister, die die Untersuchung eines allfälligen Zusammenhangs zwischen dem Wohnort in der Nähe von Kernkraftwerken und der Leukämie bei Kindern ermöglichte [2]. In diesem Artikel diskutieren wir die Verknüpfung von Datenbanken für Forschungszwecke in der Schweiz und stellen eine neue Methode vor, mit der sensible Gesundheitsdaten von einer unabhängigen Drittpartei, dem sogenannten *Trust Center*, verknüpft werden können, ohne dass identifizierende Angaben ausgetauscht werden müssen.

Verknüpfung von Datenbanken und Informed Consent

Bei der Verknüpfung der Daten gibt es grundsätzlich zwei Möglichkeiten. Beim einfacheren deterministischen Verknüpfen werden typischerweise Einträge gesucht, bei denen die eindeutige Identifikationsnummer übereinstimmt. In den skandinavischen Ländern ist die die Verknüpfung von verschiedenen Datensätzen über eine «Personennummer» möglich und erlaubt. Eine vor kurzem publizierte dänische Studie untersuchte zum Beispiel Antileptika während der Schwangerschaft als Risikofaktor für Aborte und Totgeburten [3]. Zu diesem Zweck wurde das Geburtsregister mit den Daten der Spitalaustritte und Medikamentenverschreibungen verlinkt.

Im Vergleich zu Skandinavien werden von Statistik Schweiz (Bundesamt für Statistik, BFS) weniger Daten zur Gesundheit erhoben. Verknüpfungen von BFS-Daten dürfen nur im Bundesamt für Statistik im Rahmen eines Datenschutzvertrags durchgeführt werden (siehe auch Datenverknüpfungsverordnung [4]). In der Forschung kann die 2008 eingeführte eindeutige Sozialversichertennummer in der Regel nicht zur Verknüpfung genutzt werden.

Grundsätzlich ist deterministisches Verknüpfen zwar auch mit Namen, Geschlecht, Postleitzahl und Geburtsdatum möglich. Dies ist aber schwierig, weil die Daten inkonsistent erfasst werden oder zum Teil fehlen. Deshalb wird bei Abwesenheit einer eindeutigen Identifikationsnummer häufig das sogenannte «probabilistische» Verknüpfen verwendet. Hier werden nicht nur genau übereinstimmende Einträge verknüpft, sondern die Wahrscheinlichkeit berechnet, dass zwei Einträge trotz Abweichungen von derselben Person stammen. Wenn zum Beispiel ein Geburtstag in einem Datensatz mit «12.02.1984» angegeben ist, wird beim probabilistischen Verknüpfen der «13.2.1984» oder der «2.12.1984» als ähnlicher eingestuft als etwa der «27.9.2001». Aufgrund der errechneten Wahrscheinlichkeiten kann dann bestimmt werden, ob Einträge als derselben Per-

* Die Literatur findet sich unter www.saez.ch → Aktuelle Ausgabe oder → Archiv → 2015 → 50/51.

Dialoggruppe Forschungsschwerpunkt Versorgungsforschung

Versorgungsforschung ist für die Ärzteschaft ein wichtiger und wegweisender Wissenschaftsbereich. In Zeiten des Umbruchs und der Veränderungen im Gesundheitswesen (neue Finanzierungs- und Versorgungsmodelle, demographische Veränderungen, sektorielle Verschiebungen usw.) ist eine akademisch verankerte Forschung im Bereich der ärztlichen Versorgung zwingend nötig. Um wissenschaftliche und von Partikulärinteressen unabhängige Grundlagen schaffen zu können, unterstützen die Verbindung der Schweizer Ärztinnen und Ärzte (FMH), die Konferenz der Kantonalen Ärztegesellschaften (KKA) sowie NewIndex (NI) gemeinsam den Forschungsschwerpunkt Versorgungsforschung am Institut für Sozial- und Präventivmedizin der Universität Bern.

Eine Dialoggruppe dient als Informations- und Austauschplattform: Vertreter der genannten Organisationen und der Forschungsgruppen diskutieren regelmässig die laufenden und geplanten Arbeiten im Bereich der Versorgungsforschung. Die Dialoggruppe verfolgt die Ziele, die Akzeptanz und Sensibilisierung innerhalb der Ärzteschaft für diesen Wissenschaftsbereich zu fördern und dabei den konkreten Nutzen aufzuzeigen, der mit der Versorgungsforschung für die Ärzteschaft resultiert.

Die Dialoggruppe steht ihrer Basis offen für Themen-, Diskussionsvorschläge sowie für weitere Fragen und Informationen. Die Abteilung Daten, Demographie und Qualität DDQ der FMH übernimmt die Koordination der Dialoggruppe und steht für weitere Informationen und Auskünfte gerne zur Verfügung: ddq[at]fmh.ch oder Tel. 031 359 11 11.

son zugehörig angesehen werden können. In der Schweiz können mit dem probabilistischen Verknüpfen aufgrund von Geburtsdatum, Geschlecht und Wohnort gute Ergebnisse erzielt werden [5]. Wenn zusätzlich Namen und Vornamen zur Verfügung stehen, sind die Resultate sogar ähnlich gut wie mit einer Identifikationsnummer. Name, Vorname und Geburtsdatum sind aber aus Datenschutzgründen problematisch und dürfen in der Regel nur dann verwendet werden, wenn die explizite Einwilligung *(Informed Consent)* der Patienten vorliegt. Es empfiehlt sich deshalb vor allem bei Langzeitstudien, einen entsprechenden Passus in die Einwilligungserklärung einzufügen. Ausnahmen betreffen Forschungsprojekte mit zum Beispiel durch Verschlüsselung anonymisierten, oder anonym erhobenen gesundheitsbezogenen Daten.

Die Verschlüsselung von Gesundheitsdaten

Auch bei Vorliegen des *Informed Consent* ist es im Interesse des Datenschutzes, die identifizierenden Angaben wie Name, Vorname und Geburtstag zur Verknüpfung der Daten zu nutzen, ohne dass dabei Personen identifiziert werden können. Dies kann mit einer Verschlüsselung der Daten erreicht werden. Dabei werden aus Namen oder Geburtsdaten nicht identifizierbare Buchstaben- oder Zahlenfolgen. Wird bei zwei Datensätzen der gleiche Schlüssel verwendet, können die Buchstaben- oder Zahlenfolgen einander zugeordnet werden. Unabdingbar ist, dass eine Umkehr der Verschlüsselung unmöglich ist. Leider sind die gängigen Verschlüsselungsprogramme für die Verknüpfung von Gesundheitsdaten nicht geeignet, da ein kleiner Unterschied in der Schreibweise in der Verschlüsselung zwei völlig unterschiedliche Werte ergibt. So sind «Emmenegger» und «Emmeneger» nach der Verschlüsselung nicht mehr als ähnliche Namen zu erkennen, ein Schreibfehler wirkt sich somit fatal aus. Da für das probabilistische Verknüpfen die Ähnlichkeit von Einträgen beurteilt werden muss, wurden entsprechende Verschlüsselungsmethoden gesucht. Dabei haben sich die sogenannten Bloom Filter bewährt [6]. Sie ermöglichen eine Berechnung der Ähnlichkeit auch in verschlüsselten Daten, so dass die Verknüpfung trotz Schreib- und anderen Fehlern möglich ist (Tab. 1).

Die P3RL Methode des ISPM Bern

Vor diesem Hintergrund hat das Institut für Sozial- und Präventivmedizin (ISPM) der Universität Bern Software für ein «Privacy Preserving Probabilistic Record Linkage» (P3RL) entwickelt, die eine Verknüpfung von verschiedenen Datenbanken mit individuellen Gesundheitsdaten unter Einhaltung des Datenschutzes erlaubt [7]. Die P3RL-Methode ist geeignet, wenn Gesundheitsdaten von mindestens zwei verschiedenen Zentren verknüpft werden sollen, die nicht über eine gemeinsame Identifikationsnummer verfügen und deren Datenschutzvorgaben die Verwendung von identifizierenden Variablen wie Name, Geburtstag, Todestag oder Adresse einschränken. Die Verknüpfung wird von einem unabhängigen Trust Center vorgenommen. Die P3RL-Methode setzt sich aus drei Schritten zusammen: der Vorbereitung der Daten, der Verschlüsselung und der Verknüpfung (Abb. 1).

1. Vorbereiten der Daten

In einem ersten Schritt werden Eigenheiten sowie Fehler in den zu verknüpfenden Datensätzen nach einheitlichen Regeln bereinigt. Zum Beispiel werden

Tabelle 1: Vergleich der Verschlüsselung von Namen mit verschiedenen Methoden. Die Ähnlichkeit der ursprünglichen Namen lässt sich nur mit einer Bloom Filter Verschlüsselung beurteilen.

Klartext	Konventionelle Verschlüsselung*	Beispiel einer Bloom Filter Verschlüsselung
Emmenegger	078f73ae3b2852b79e143a06aa573f21	11111111111111011101100111101110010101011110010101
Emmeneger	21783f44f4696323a2267a83a2f2dd7b	11111111111111011101100111101110010101001110010101
Meier	3399f3b498509a2f63b058db71a360f3	11011010111101000001001110110111111010011111111011

*unter Verwendung von MD5 Hash

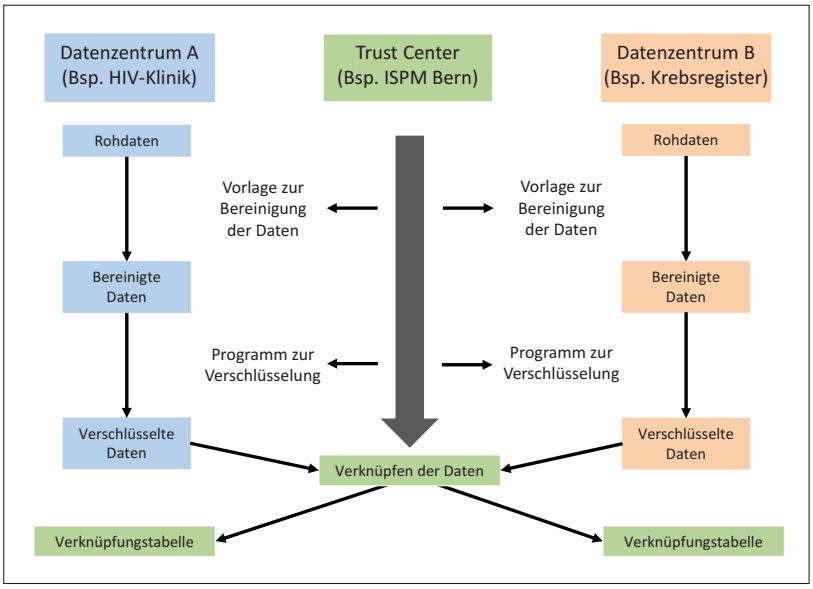

Abbildung 1: Ablauf und Verteilung der Arbeitsschritte bei der P3RL-Methode am Beispiel der Datensätze einer HIV-Klinik und eines Krebsregisters, die im Trust Center des ISPM in Bern verknüpft werden.

Sonder- und sprachspezifische Zeichen vereinheitlicht, vorangestellte Wörter («Frau» oder «Dr.») oder Anhänge («jun.») entfernt, pro Vorname eine Variable erstellt und Datumsangaben sowie fehlende Angaben einheitlich dargestellt. Das Trust Center stellt den anderen Zentren zu diesem Zweck eine Sammlung von einheitlichen Regeln zur Verfügung.

2. Verschlüsselung

Für die Verschlüsselung der Daten stellt das Trust Center ein Programm zur Verfügung, das auf den bereits erwähnten Bloom Filtern beruht. Das Programm erlaubt den Datenmanagern in den jeweiligen Zentren, ihre Daten zu verschlüsseln, ohne dass spezielle Kenntnisse vorhanden sein müssen. Der Schlüssel für das Verschlüsselungsprogramm wird von den beiden beteiligten Zentren in gegenseitiger Absprache definiert, ohne dass das Trust Center davon Kenntnis erhält.

3. Verknüpfung

Der letzte Schritt ist das probabilistische Verknüpfen der Daten, das vom Trust-Center vorgenommen wird. In einem ersten Schritt werden alle vollständig übereinstimmenden Einträge zusammengeführt. Anschliessend wird mit Hilfe von Wahrscheinlichkeiten, die für jedes Merkmal berechnet werden, die bestmögliche Verknüpfung bei Einträgen mit fehlenden oder leicht abweichenden Daten hergestellt. Dabei wird unter anderem geprüft, ob beispielsweise Vor- und Nachname vertauscht eingegeben wurden oder andere häufige Fehler wie vertauschter Tag und Monat vorliegen könnten. Die Verknüpfungstabelle sowie ein Bericht mit zusätzlichen Informationen zur Qualität der Verknüpfung werden den jeweiligen Zentren geliefert, bevor das Trust Center bei sich alle vorhandenen Daten löscht.

Simulationsstudie

Die Qualität der P3RL-Methode wurde in einer Simulationsstudie mit echten und fingierten Daten getestet. Die Simulationsstudie zeigte, dass bei der Verknüpfung unter Verwendung von Namen sehr gute Ergebnisse erzielt werden, auch wenn die Namen verschlüsselt sind. Entscheidend ist jedoch die Vorbereitung der Daten: Nur mit einheitlich bereinigten Daten liessen sich gute Ergebnisse erzielen. So waren bei der Verknüpfung mit verschlüsselten Namen 18% der gefundenen Verbindungen falsch, wenn die Daten vorher nicht bereinigt wurden, aber nur 0,7%, wenn die Daten bereinigt wurden.

Fazit und Ausblick

Die Simulationsstudie hat gezeigt, dass sich P3RL hervorragend eignet, um im Gesundheitswesen Daten verschiedener Quellen zu verknüpfen, ohne dass der Datenschutz verletzt wird. Auch wenn P3RL auf einer Kombination von technisch anspruchsvollen Lösungen beruht, ist die Methode in der Praxis auch von Zentren ohne Erfahrung und Spezialwissen in Verknüpfungstechniken anwendbar. Zudem kann bei der Bereinigung der Daten auf die Eigenheiten der verschiedenen Zentren eingegangen werden. Allerdings ist die P3RL-Methode zeitlich, personell und finanziell aufwendiger als die Verknüpfung von unverschlüsselten Daten. Zudem muss auch mit der Verwendung von P3RL für jedes Projekt individuell von der zuständigen Ethikkommission beurteilt werden, ob die Anforderungen des Datenschutzes erfüllt werden.

Nur mit einheitlich bereinigten Daten liessen sich gute Ergebnisse erzielen.

Nachdem die P3RL Methode entwickelt und erfolgreich getestet wurde, steht sie jetzt konkreten Projekten offen. Als erstes wird sie in einer Studie zum Krebsrisiko für HIV-infizierte Personen in der Schweiz angewendet. Für diese Studie soll die bevölkerungsbasierten Information aus dem Netzwerk der Schweizerischen Krebsregister (NICER) mit derjenigen der Schweizerischen HIV-Kohorten-Studie (SHCS) verknüpft werden. Um dabei den Datenschutz gewährleisten und eine qualitativ hochstehende Verknüpfung vollziehen zu können, wird die P3RL-Methode angewendet. Das Projekt wurde von der Kantonalen Ethikkommission Bern gutgeheissen.

Korrespondenz:
Prof. Dr. med.
Matthias Egger
ISPM Universität Bern
Finkenhubelweg 11
CH-3012 Bern

Literatur

1 Stuck A, Moser A, Morf U, Wirz U, Wyser J, Gillmann G et al. Effect of Health Risk Assessment and Counselling on Health Behaviour and Survival of Older People: Randomised Trial. PLoS Med. 12(10): e1001889. doi:10.1371/journal.pmed.1001889.

2 Spycher BD, Feller M, Zwahlen M, Roosli M, von der Weid NX, Hengartner H et al. Childhood cancer and nuclear power plants in Switzerland: a census-based cohort study. Int J Epidemiol. 2011;40: 1247–60. doi:10.1093/ije/dyr115.

3 Bech BH, Kjaersgaard MIS, Pedersen HS, Howards PP, Sørensen MJ, Olsen J et al. Use of antiepileptic drugs during pregnancy and risk of spontaneous abortion and stillbirth: population based cohort study. BMJ. 2014;349:g5159.

4 SR 431.012.13 Verordnung des EDI vom 17. Dezember 2013 über die Verknüpfung statistischer Daten (Datenverknüpfungsverordnung). Erhältlich unter https://www.admin.ch/opc/de/classified-compilation/20122208/index.html (herunter geladen am 28. September 2015).

5 Bopp M, Spoerri A, Zwahlen M, Gutzwiller F, Paccaud F, Braun-Fahrländer C et al. Cohort Profile: the Swiss National Cohort – a longitudinal study of 6.8 million people. Int J Epidemiol. 2009;38:379–84. doi:10.1093/ije/dyn042.

6 Schnell R, Bachteler T, Reiher J. Privacy-preserving record linkage using Bloom filters. BMC Med Inform Decis Mak. 2009;9:41. doi:10.1186/1472-6947-9-41.

7 Schmidlin K, Clough-Gorr KM, Spoerri A. Privacy preserving probabilistic record linkage (P3RL): a novel method for linking existing health-related data and maintaining participant confidentiality. BMC Med Res Methodol. 2015;15:46. doi:10.1186/s12874-015-0038-6.