Z. Evid. Fortbild. Qual. Gesundh. wesen (ZEFQ) 149 (2019) 22–31

Contents lists available at ScienceDirect

Z. Evid. Fortbild. Qual. Gesundh. wesen (ZEFQ)

journal homepage: http://www.elsevier.com/locate/zefq

Versorgungsforschung / Health Services Research

Das Risiko von Re-Identifizierung bei der Auswertung medizinischer Routinedaten – Kritische Bewertung und Lösungsansätze

The risk of re-identification when analyzing electronic health records: a critical appraisal and possible solutions

Johannes Hauswaldt[a,*], Iris Demmer[a], Stephanie Heinemann[a], Wolfgang Himmel[a], Eva Hummers[a], Johannes Pung[b], Falk Schlegelmilch[a], Johannes Drepper[c]

[a] *Institut für Allgemeinmedizin, Universitätsmedizin Göttingen, Göttingen, Deutschland*
[b] *Institut für Medizinische Informatik, Universitätsmedizin Göttingen, Göttingen, Deutschland*
[c] *TMF - Technologie- und Methodenplattform für die vernetzte medizinische Forschung e.V., Berlin (TMF), Deutschland*

ARTIKEL INFO

Artikel-Historie:
Eingegangen: 25. Mai 2019
Revision eingegangen: 14. November 2019
Akzeptiert: 15. Januar 2020
Online gestellt: 10. März 2020

Schlüsselwörter:
Elektronische Patientenakten
Qualitätssicherung in der medizinischen Versorgung
Datenschutz
Datenanonymisierung
Statistische Datengrundlagen der ambulanten Primärversorgung
Allgemeinmedizin

ZUSAMMENFASSUNG

Hintergrund und Zielsetzung: Die sekundäre Nutzung medizinischer Routinedaten zum Beispiel aus Allgemeinarztpraxen könnte ein zentraler Baustein zukünftiger Versorgungsforschung werden; die Risiken und auch Vorbehalte dagegen sind aber nicht zu unterschätzen. Wir zeigen exemplarisch, vor welchen Problemen die Auswertung von Routinedaten im Sinne des Datenschutzes steht und wie technisch intelligente Lösungen aussehen können – auch im Sinne der Vertrauensbildung.
Methode: Basis des Projekts sind Routinedaten aus hausärztlichen Arztpraxisinformationssystemen. Diese Routinedaten, die zunächst der praxisinternen Dokumentation i. S. einer elektronischen Behandlungsakte und der Leistungsabrechnung dienen, werden über die standardisierte Behandlungsdaten-Transfer- (BDT-) Schnittstelle extrahiert, de-identifiziert und als verschlüsselter Datensatz in eine Forschungsdatenbank übertragen. Das Risiko einer Re-Identifizierung von Patienten bei sekundärer Nutzung solcher Routinedaten sollte anhand von 40 aus unserer Sicht besonders wichtigen Variablen aus einem BDT-Datensatz (dort als „Felder", mit „Feldkennungen" und „Feldinhalten" bezeichnet) beurteilt werden. Kriterien waren „Expertenwahrnehmung" (Rückschlüsse eines professionellen Beobachters von phänotypischen Besonderheiten einer Person auf Feldinhalte von BDT-Daten), „recherchierbares Zusatzwissen" (Kenntnis der Eigenschaften einer Person durch Informationen u.a. aus sozialen Netzwerken) und „statistische Häufigkeit" von BDT-Inhalten gemäß einer Diagnosen- und Medikamentenstatistik.
Ergebnisse: Diagnosen und Beratungsanlässe haben besonders identifizierende, weil persönlichkeitsbestimmende Eigenschaften, besonders deutlich bei „Adipositas" (ICD-10 E66) sowie „Schädlicher Tabakkonsum" (F17). Etwa die Hälfte aller ICD-Kodes aus einer Hausarztpraxis unterschreitet in ihrer absoluten Häufigkeit kritische Grenzwerte; dies ist erst recht problematisch, wenn solche Diagnosen phänotypisch erkennbar sind und damit Re-Identifizierungspotential besitzen. Bei Angaben zur Medikation hingegen ist dieses Potential eher gering, Medikation selbst jedoch kann durch ihre Anwendung re-identifizierend sein, z. B. Selbstinjektionen von Insulin oder Inhalatoren. Zeit- und Datumsinformationen sind dagegen besonders sensibel für die Re-Identifizierung einer Person. Informationen zum Geschlecht und Alter eines Patienten sind im Allgemeinen unproblematisch, ausgenommen das Alter bei sehr jungen und sehr alten Patienten, wenn diese praxisspezifisch selten sind.
Diskussion: Medizinische Routinedaten sind immer als sensible Daten anzusehen. Die genaue Kenntnis einzelner Felder und ihrer Inhalte in elektronischen Behandlungsakten aus Allgemeinpraxen und deren Bewertung erlauben eine zuverlässige Abschätzung des Risikos einer Re-Identifizierung von betroffenen Personen. Diagnosen, besonders als Dauerdiagnosen und/oder Langtext, Kalenderdaten von Kontakten und Behandlungen haben ein besonders hohes Risiko der Re-Identifizierung eines Patienten. Maßnahmen wie Datenentfernung, Wertmaskierung oder Kodierung dürften die Re-Identifikation erheblich erschweren. Ein Restrisiko wird bleiben und sollte offen und transparent diskutiert werden, um Befürchtungen gegenüber einem mangelnden Datenschutz und pauschaler Kritik an der Digitalisierung des Gesundheitswesens zu begegnen.

* Korrespondenzadresse: Dr. med. Johannes Hauswaldt, MPH, Institut für Allgemeinmedizin, Universitätsmedizin Göttingen, Humboldtallee 38, 37073 Göttingen, Deutschland.
E-mail: johannes.hauswaldt@med.uni-goettingen.de (J. Hauswaldt).

https://doi.org/10.1016/j.zefq.2020.01.002
1865-9217/

ARTICLE INFO

Article History:
Received: 25 May 2019
Received in revised form:
14 November 2019
Accepted: 15 January 2020
Available online: 10 March 2020

Keywords:
Electronic health records
Quality of healthcare
Privacy of patient data
Data anonymization
Statistics and numerical data in primary healthcare
Family practice

ABSTRACT

Background and objectives: The use of primary care data gathered from electronic health records in local practices could be an important building block for the future of health services research. However, the risks and reservations associated with using this data for research purposes should not be underestimated. We show the data protection and privacy problems that may arise through secondary analysis of routine primary care data and describe the technical solutions that are available to address these concerns – as a trust-building measure.

Methods: We screened 40 variables that are deemed important for documentation in the electronic health records of primary care physicians and rated the risk of patient re-identification when using these records from routine medical data for research purposes. The criteria used to rate the risk of re-identification were "expert perception" (inferences of a professional observer of phenotypical characteristics which are documented in the 40 variables), "researchable additional knowledge" (knowledge of characteristics of a person through publicly available information and social media networks), and "statistic frequency" according to diagnosis and medication statistics.

Results: Diagnoses and reasons for contacting a general practitioner can contain particularly identifiable characteristics such as "obesity" (ICD-10 E66) and "nicotine dependence" (F17). About half of all ICD codes documented in primary care fall below a critical threshold value in their absolute frequency; this is all the more problematic if diagnoses allow for re-identification due to phenotypical characteristics. Medication information holds little potential risk of re-identification of a person. However, the application of medications could be a source of re-identification, e. g., self-injections of insulin or use of inhalators. Information about times and dates are especially sensitive for the re-identification of a person. Sex and age of a patient generally pose no problems, except in the case of very young or very old individuals when these age groups are seldom represented in the practice.

Discussion: Routine health data are, in principle, sensitive data. Knowledge about the variables in primary care data gathered from electronic health records in local practices and the evaluation of this data allow us to more accurately estimate the risk of re-identification for the persons concerned. In particular, chronic diagnoses and/or diagnoses in long text, calendar dates for patient contacts and therapies bear a high risk of re-identification. Technical measures such as removing data, masking values and coding should make re-identification considerably more difficult. There will always be a remaining risk of re-identification which should be openly discussed to counteract concerns about a lack of data protection or a sweeping critique of digitization in healthcare.

Einleitung

Die Datenskandale der jüngsten Vergangenheit, z. B. die Versuche der Wahlbeeinflussung in Großbritannien durch Cambridge Analytica mittels Facebook-Daten oder Datenlecks bei sensiblen Patientendaten, haben auch die sog. Big Data-Forschung in der Medizin in der Öffentlichkeit oftmals diskreditiert, zumindest werden gegenüber den Chancen zunehmend häufiger die Risiken angesprochen. Zugleich ist das Thema Datenschutz mit der seit Mai 2018 anzuwendenden Europäischen Datenschutzgrundverordnung (DSGVO) in die öffentliche Aufmerksamkeit gerückt. Besonders kritisch wird die Verarbeitung von Gesundheitsdaten gesehen, darunter auch die international für Versorgungsforschung häufig genutzten medizinischen Routinedaten oder andere personenbezogene Routinedaten aus der Versorgung. Diese Datengruppe gilt als hochsensibel und unterliegt entsprechend auch besonderen datenschutzrechtlichen Bestimmungen [1]. Gleichzeitig sind jedoch diese medizinischen Routinedaten gut für die Versorgungsforschung und auch Qualitätssicherung in der Medizin geeignet, um z.B. klinisch-epidemiologische Fragen zu untersuchen oder Unter-, Fehl- und Überversorgung zu erkennen [2] – und das ohne weiteren Erhebungsaufwand. Dieser potentielle Konflikt zwischen Datenschutz und Versorgungsforschung oder Qualitätssicherung im Fall personenbezogener medizinischer Routinedaten soll zunächst genauer dargestellt werden, bevor sich daraus die spezifische Fragestellung unseres Projekts und dieses Artikels ergibt.

Um ihren primären Zweck bei der individuellen medizinischen Betreuung und Versorgung erfüllen zu können, enthalten elektronische Routinedaten regelmäßig personenidentifizierende Daten: sog. direkte Identifikatoren (*explicit identifiers*) wie beispielsweise Namensangaben. Weiterhin enthalten elektronische Routinedaten viele Variablen, die zwar zunächst nicht identifizierend erscheinen, jedoch in Abhängigkeit von ihrer Ausprägung bzw. in Kombination mit anderen Variablen einen Personenbezug herstellen lassen, z.B. durch einen Abgleich mit anderweitig zugänglichen Daten. Solche Variablen, wie etwa Alter, Geschlecht, Wohnort u.a., sind Quasi-Identifikatoren (*quasi identifiers*) [3–5].

Laut „Erwägungsgrund 26" der DSGVO [1] sollen die Grundsätze des Datenschutzes nicht für personenbezogene Daten gelten, „die in einer Weise anonymisiert worden sind, dass die betroffene Person nicht oder nicht mehr identifiziert werden kann" [6]. Um festzustellen, ob eine Person identifizierbar ist, sollten alle Mittel berücksichtigt werden, die hierzu nach allgemeinem Ermessen wahrscheinlich genutzt werden. Was nach allgemeinem Ermessen wahrscheinlich zur Identifizierung genutzt wird, soll nach objektiven Faktoren, wie den Kosten der Identifizierung und des dafür erforderlichen Zeitaufwands, ermittelt werden. Dabei sind die aktuell verfügbare Technologie und technologische Entwicklungen zu berücksichtigen. Nach diesen Rechtsvorgaben genügt es in der Regel nicht, allein die direkten Identifikatoren, wie etwa Name und Geburtsdatum, wegzulassen. Effektive Konzepte des Datenschutzes müssen vielmehr die unterschiedlichen Typen von Informationen, wie sie in medizinischen Daten abgelegt sind, und deren Potential, eine Identifizierung von Betroffenen, alleine oder in Kombination zu ermöglichen, berücksichtigen.

Wie verhalten sich nun die bekannten Verfahren der Pseudonymisierung und Anonymisierung zu diesen Re-Identifizierungsrisiken? Zunächst ist festzustellen, dass beide Verfahren typischerweise auch zur Umsetzung der Datensparsamkeit bzw. zur Datenminimierung, wie es in der DSGVO heißt, eingesetzt werden: Wissenschaftler sollen auf möglicherweise identifizierende Daten (*explicit identifiers*) keinen Zugriff haben, wenn sie diese nicht benötigen. Dies dient der Risikominimierung, bedeutet jedoch nicht, dass damit alle Risiken einer Re-Identifizierung beseitigt wären. Quasi-Identifikatoren sind

dabei von einer derartigen Pseudonymisierung in der Regel nicht betroffen. Wenn allerdings relevante Re-Identifizierungsrisiken bestehen bleiben, bedeutet das auch, dass die Daten weiterhin personenbeziehbar sind und dem Rechtsrahmen des Datenschutzes sowie unabhängig davon als Daten aus dem Behandlungskontext zusätzlich auch der ärztlichen Schweigepflicht unterliegen.

Die Nutzung von Daten ohne datenschutzrechtliche Grundlage, wie es etwa die explizite Einwilligung der Betroffenen nach entsprechender Aufklärung wäre, ist nur zulässig, wenn sie nach den oben genannten Kriterien der DSGVO nicht mehr identifizierbar sind [7,8]. Als Rechtsgrundlage jenseits einer Einwilligung können in bestimmten Fällen auch Forschungsklauseln aus dem Datenschutzrecht in Frage kommen, wie beispielsweise in § 27 Bundesdatenschutzgesetz (BDSG) formuliert. Diese datenschutzrechtliche Erlaubnisnorm gestattet auf der Grundlage einer sorgfältigen Abwägung die interne Nutzung der Daten zu einem anderen Zweck (der wissenschaftlichen Forschung), jedoch nicht die Übermittlung der personenbeziehbaren Daten aus dem mit der ärztlichen Schweigepflicht nach Berufs- und Strafrecht (§ 203 Strafgesetzbuch, StGB) besonders geschützten Bereich der Behandlung heraus. Somit unterscheidet sich die Rechtslage für die Sekundärnutzung der Daten aus Arztpraxen systematisch von der Situation zur Nachnutzung von Sozialdaten von den Krankenkassen mit der Erlaubnisnorm in §75 SGB X und der Nachnutzung von Behandlungsdaten aus Krankenhäusern, für die ggf. ein Landeskrankenhausgesetz eine spezifische Erlaubnisnorm mit einer Abwägung vorsieht.

Ein möglicher und rechtskonformer Lösungsansatz für die zentrale Auswertung von Daten aus vielen Arztpraxen ohne die informierte Einwilligung aller betroffenen Patienten ist dann die Nutzung der Daten in effektiv anonymer Form. Viele Forschungsfragen können auch mit Hilfe anonymer Daten beantwortet werden. Hierbei sind für jede verwendete Datensammlung eine sorgfältige Betrachtung des Re-Identifizierungspotentials und eine dazu passende Vergröberung oder Maskierung durchzuführen. Gleichzeitig jedoch ist der dabei entstehende Informationsverlust so in Grenzen zu halten, dass eine Beantwortung der ursprünglichen Fragestellung noch sinnvoll und möglich ist.

Dass die Bereitschaft von Entscheidungsträgern und Patienten, einer stärkeren Nutzung medizinischer Routinedaten zuzustimmen, nicht zuletzt vom wahrgenommenen Nutzen und ganz besonders vom Vertrauen in den Datenschutz abhängt, konnte eine kanadische Studie bereits 2011 feststellen [9]. Wie gestört das Vertrauen in Politik (und Wissenschaft!) gegenüber der Digitalisierung im Gesundheitswesen ist, zeigt exemplarisch ein Artikel aus der *Süddeutschen Zeitung* vom 25. Februar 2019, in dem es um die „elektronische Patientenakte“ geht, die mit dem „gläsernen Behandlungszimmer“ gleichgesetzt wird und hinter der „andere Interessen“, natürlich wirtschaftliche vermutet werden [10]. Noch deutlicher werden die Leserbriefe zu diesem Artikel am 7. März, in denen Ärzte und Patienten eine „öffentliche Debatte zur Datenethik“ vermissen, die „massiven Gewinne der IT-Firmen ohne verbesserte Patientenversorgung“ anprangern und zugleich sicher sind, dass die „zentrale Datensammlung aller Gesundheitsdaten . . . Begehrlichkeiten, zum Beispiel bei Wissenschaftlern“ weckt.

Dennoch fehlt dem Problem der Re-Identifizierungsrisiken bei sekundärer Nutzung von Routinedaten aus der Versorgung – national und international – wissenschaftliche Aufmerksamkeit. So kam kürzlich eine finnische Arbeitsgruppe in einem Review zur Strukturierung von elektronischen Patientendaten unter dem Aspekt ihrer sekundären Nutzung zum Ergebnis, dass Sicherheitsbedenken aus Patientensicht nur selten eine Rolle spielen [11]. Wenn überhaupt, geht es im Regelfall um Verschlüsselungsalgorithmen zum Schutz von Patientendaten gegen unrechtmäßige Nutzung wie Hackerangriffe oder Zugriffe von nicht berechtigtem Personal, also z. B. durch Administratoren ([12], [13]).

Ein Kernanliegen des Projekts „Routine Anonymized Data for Advanced Health Services Research“ (RADAR) ist es daher, nicht nur das unbestreitbare Potential von ambulanten Routinedaten für die Versorgungs- und Gesundheitssystemforschung (z. B. [14]) zu erschließen, sondern dabei auch besonderes Augenmerk auf technisch-organisatorische Aspekte und den Datenschutz zu legen. Das Projekt wird vom Göttinger Institut für Allgemeinmedizin durchgeführt – gemeinsam mit dem Göttinger Institut für Medizinische Informatik, dem Institut für Community Medicine der Universität Greifswald, der Gesellschaft für Wissenschaftliche Datenverarbeitung Göttingen (GWDG) und der Technologie- und Methodenplattform für die vernetzte medizinische Forschung (TMF) Berlin [15].

Die mit der sekundären Nutzung medizinischer Routinedaten verbundenen Gefahren und datenschutzrechtlichen Risiken sind nicht zu unterschätzen. Diese erfordern, auch und gerade um des Vertrauens willen von Patienten und anderen Betroffenen, die sorgfältige Planung, Umsetzung und Kontrolle umfassender Datenschutzmaßnahmen. Nachdem wir kürzlich über Mängel in der organisatorischen und technischen Bereitstellung von hausärztlichen Routinedaten berichtet haben [6], möchten wir in diesem Beitrag exemplarisch zeigen, vor welchen Problemen die anonyme Auswertung solcher Routinedaten bei Beachtung des Datenschutzes steht und wie technisch intelligente Lösungen aussehen können.

Das langfristige Ziel dabei ist es, medizinische Routinedaten und andere personenbezogenen Daten so für die Forschung zugänglich zu machen, dass ein Rückschluss auf konkrete Personen nicht möglich ist, zugleich aber die Daten trotz des unvermeidlichen Informationsverlusts noch wertvoll und wissenschaftlich verwertbar bleiben [16,17]. Die hier zur Diskussion gestellten Ergebnisse und Lösungsvorschläge verstehen sich auch als Maßnahmen der Vertrauensbildung durch Transparenz von Struktur und Prozess der Sekundäranalyse medizinischer Routinedaten aus der ambulanten Versorgung.

Methoden

Datenbasis

Basis des Projekts sind Routinedaten aus hausärztlichen Arztpraxisinformationssystemen. Unter medizinische Routinedaten werden hier Daten und Informationen verstanden, die primär zumeist aus der alltäglichen ambulanten Betreuung und Versorgung von Patienten vorhanden sind oder *de novo* entstehen und, patientenbezogen zusammengeführt, zunächst der praxisinternen Dokumentation i.S. einer elektronischen Behandlungsakte und der Leistungsabrechnung dienen. Diese Routinedaten können standardisiert über eine obligate Schnittstelle extrahiert werden, der Behandlungsdaten-Transfer- (BDT-) Schnittstelle. Forschungsrelevante Daten werden noch in der Praxis selektiert, de-identifiziert (Entfernung von Personennamen, Geburtsdaten, Kontaktdaten) und als BDT-Datensatz in eine Forschungsdatenbank verschlüsselt übertragen (Schutz vor unbefugtem Zugriff während der Übertragung).

Im hier zugrundeliegenden RADAR Projekt ist eine kleine Anzahl von Praxen ausreichend, um die Möglichkeit der datenschutzrechtlich sicheren Nutzung eines BDT-Datensatzes aus Routinedaten exemplarisch unter Beweis zu stellen („*proof of concept*“) – langfristig geht es um die Routinedaten aus großen und permanent bestehenden Forschungspraxen-Netzen bzw. gemäß angloamerikanischem Sprachgebrauch aus *Practice Based Research Networks* [18]. Wir überschauen gegenwärtig (Dezember 2019) Routinedaten von 100 Patienten mit dem Anwendungsfall „Orale Antikoagulation“, davon 46 weiblich, Geburtsjahr 1920 bis 1966,

aus 8 Hausarztpraxen für den Zeitraum 1. Quartal 2012 bis 3. Quartal 2019, extrahiert zwischen Mai und November 2019, mit 3 Variablen in 353.242 Datenzeilen.

Für unsere nachfolgenden Betrachtungen zu Re-Identifizierungsrisiken zogen wir zudem eine „historische" Sammlung hausärztlicher Routinedaten heran, die in vorangegangen Projekten von uns zwischen 2002 und 2007 ebenfalls mittels der BDT-Schnittstelle extrahiert wurden, von 581.015 Patienten, davon 51,8% weiblich, Geburtsjahr 1890 bis 2007, in 165 Praxen für den Zeitraum 1994 bis 2007, mit 8 Variablen in 20.294.205 Datenzeilen.

Definition von Re-Identifizierungsrisiken

Für eine Definition eines Re-Identifizierungs-Risikos ist zunächst eine Klärung möglicher Angriffsszenarien erforderlich, die nicht nur medizinische Daten betreffen. Eine unzulässige Re-Identifizierung kann entweder durch eine Person erfolgen, die regelhaft bzw. erlaubten Umgang mit der betreffenden Datenbasis hat, also von „innen" heraus, oder durch eine „externe" Person, die eigentlich keinen Zugang zu den entsprechenden Daten haben sollte, diesen aber zufallsweise (bei grober Verletzung der notwendigen Schutzmaßnahmen) oder durch einen systematischen Angriff erhält. Eine kritische Re-Identifizierung kann dann erfolgen, wenn ein interner oder externer Angreifer einen oder mehrere Datensätze der Datensammlung mit einer ihm bekannten Person – der Ziel-Person – eindeutig oder mit ausreichender Wahrscheinlichkeit verknüpfen kann. Die Ziel-Person kann sowohl aus dem engeren oder weiteren Bekanntenkreis des Angreifers entstammen als auch eine Person des öffentlichen Lebens sein.

Für die Abschätzung des Risikos ist zum einen die Wahrscheinlichkeit einer solchen Verknüpfung relevant und zum anderen der durch eine solche Verknüpfung entstehende Schaden. Eine eindeutige Verknüpfung mit extern verfügbaren Daten ist häufig schon mit Hilfe der Kombination weniger Einzeldaten wie etwa der Postleitzahl, dem Geschlecht und dem Geburtsdatum möglich [5]. Wenn diese Daten mit sensiblen Gesundheitsangaben wie z.B. Labordaten kombiniert verfügbar sind, ist schnell von einem hohen Schaden auszugehen. Insofern sind solche Angaben wie etwa der Wohnort und das Geburtsdatum im Regelfall zu vergröbern. Wenn dann aber die detaillierten Laborwerte unvergröbert im Datensatz verbleiben, könnten theoretisch auch diese aufgrund ihrer Eindeutigkeit für eine Re-Identifizierung genutzt werden. Allerdings bemisst sich der durch eine Re-Identifizierung entstehende Schaden danach, ob und in welchem Umfang sensible Informationen durch eine solche Verknüpfung offenbart werden. Wenn der Angreifer beispielsweise bereits Kenntnis eines komplexen und individuellen Musters von Laborwerten der Ziel-Person hat und dieses zur Verknüpfung mit der zu schützenden Datensammlung nutzt, so dürfte der Angreifer durch eben diese Verknüpfung kein relevantes zusätzliches Wissen über die Ziel-Person erhalten, da er schon über umfangreiche medizinische Fallinformation verfügt. Auch wenn die möglicherweise seltene Hauptdiagnose anhand eines pathognomonischen Zeichens von einem Angreifer erkannt wird, würden durch Zuordnung zu einer Datensammlung keine weiteren Diagnosen („Geheimnisse") offenbart werden, wenn es gar keine weiteren Diagnosen gibt.

Immerhin wird aber bei einer solchen Zuordnung einer Datensammlung die Anonymität der Daten aufgehoben, so dass eine Rechtsgrundlage für die Verarbeitung erforderlich ist. Die DSGVO statuiert in Art. 9 (2) e) die Rechtmäßigkeit der Verarbeitung besonderer Kategorien personenbezogener Daten – zu denen die Gesundheitsdaten zählen –, wenn die betroffene Person diese offensichtlich öffentlich gemacht hat. In Verbindung mit einer Rechtsgrundlage nach Art. 6 DSGVO, wie etwa dem öffentlichen Interesse an der Verarbeitung nach Abs. 2 e) wäre eine solche Verarbeitung damit auch in nicht anonymer Form (da öffentlich bekannt) legitimiert. Die für Abs. 2 e) zusätzlich erforderliche nationale Rechtsgrundlage könnte in § 27 BDSG gesehen werden. Eine unbefugte Offenbarung gemäß § 203 StGB besteht in diesen Fällen zudem nicht, da die Daten ja schon öffentlich gemacht wurden.

Für die Abschätzung der Wahrscheinlichkeit einer solchen Verknüpfbarkeit einer Datensammlung mit externem Wissen ist nun einerseits zu prüfen, welche Informationen aus der Sammlung mit potentiell verfügbarem, externem Wissen in Beziehung gesetzt werden können. Relevante Quellen für solches Zusatzwissen können sein:

- ein persönlicher Kontakt im öffentlichen oder auch privaten Umfeld,
- öffentlich bekannte Informationen über die Ziel-Person, da sie z.B. eine Person des öffentlichen Lebens ist,
- im Internet recherchierbare Daten inklusive solcher aus sozialen Netzwerken oder auch
- in ggf. zugänglichen bzw. käuflich erwerbbaren Datenbanken verfügbare Informationen (z.B. Register aller Art, SCHUFA, andere Forschungsdatenbanken usw.).

Andererseits sind neben dem extern verfügbaren Zusatzwissen auch statistische Eigenschaften der Datensammlung zu berücksichtigen, wenn die Wahrscheinlichkeit einer eindeutigen Verknüpfbarkeit abzuschätzen ist. Hier sind vor allem selten auftretende Variablen oder Ausprägungen kritisch, deren Vorkommen alleine oder in Kombination mit anderen Variablen oder ihren Ausprägungen in einem zu untersuchenden Datensatz, etwa nach Routinedatenextraktion aus einem Arztpraxisinformationssystem, eine zuvor festgelegte absolute Häufigkeit nicht erreicht. Beispielsweise könnte eine derartige *a priori* Festlegung sein, dass die Kombination der Ausprägungen aller für einen Abgleich relevanten Merkmale bzw. Variablen in mindestens 5 Datensätzen vollständig identisch sein muss. In diesem Fall spricht man von einer k-Anonymisierung mit $k = 5$ [5]. In der Literatur finden sich weitere Ansätze, um statistische Eigenschaften von Datensammlungen dahingehend zu normieren, dass individuelle Zuordnungen erschwert oder unmöglich werden [19].

Wichtig ist festzuhalten, dass nur die Kombination aus verknüpfbarem Zusatzwissen und Eindeutigkeit der zuzuordnenden Datensammlung zu einem relevanten Re-Identifizierungsrisiko führt. Umfangreiches Zusatzwissen spielt so lange keine Rolle, wie statistische Eigenschaften der Datensammlung dafür sorgen, dass eine eindeutige oder ausreichend wahrscheinliche Zuordnung zu einem identifizierenden Datensatz nicht möglich ist. Umgekehrt sind Datensätze mit hoch individueller Merkmalskombination kein Problem, wenn diese Merkmalskombination nicht für eine Zuordnung zu externem Wissen genutzt werden kann.

Datenauswertung

Das Risiko einer Re-Identifizierung bei sekundärer Nutzung hausärztlicher Routinedaten sollte anhand von 40 aus unserer Sicht besonders wichtigen Variablen aus BDT-Datensätzen (dort als „Felder", mit „Feldkennungen" und „Feldinhalten" bezeichnet) in qualitativer und dichotomer Form beurteilt werden. Diese 40 Variablen aus hausärztlichen Routinedaten werden zu insgesamt 11 semantischen Datengruppen bzw. -typen angeordnet, wie z. B. Diagnosen, Medikamente oder auch Stamm- und Dauerdaten eines Patienten (Tabelle 1). Diese semantischen Gruppen mit ihren Variablen entsprechen weitgehend den Modulen des Kerndatensatzes der Medizininformatik-Initiative (MI-I) [20], wie ebenfalls aus Tabelle 1 ersichtlich. Auf Variablen mit einer Freitextangabe oder einem Langtext zu einer kodierten Information (z. B. Diagnosetext zu einem ICD-Schlüssel) wurde dabei aus pragmatischen Gründen verzichtet, um die Analyse überschaubar zu halten.

Tabelle 1
Re-Identifizierungspotential von 40 ausgewählten BDT-Feldkennungen.

Gruppe	BDT- Feld	Inhalt	Modul* im Kerndatensatz der MI-Initiative	Re-Identifizierungspotential				Wichtigkeit**
				Feld- Kennung	Feld- Inhalt	insgesamt (geschätzt)	Abhilfe; z.B. durch	
Diagnosen	3650	Dauerdiagnosen ab Datum	Diagnosen	hoch	vereinzelt	mittel	ICDDreisteller; nur Kapitel	hoch
	6000	Abrechnungsdiagnose						
	6001	ICD-Schlüssel						
	6205	Aktuelle Diagnose						
Medikation	3652	Dauermedikamente	Medikation	mittel	vereinzelt	gering	ATC-Dreisteller	hoch
	6210	Medikament verordnet auf Rezept						
	6211	Außerhalb Rezept verordnetes Medikament						
	6215	Ärztemuster						
Labor-ergebnisse	8401	Befundart	Labor- befunde	gering	vereinzelt	gering	Maskierung von Werten	mittel
	8410	Test-Ident						
	8411	Testbezeichnung						
	8420	Ergebnis-Wert						
	8421	Einheit						
Befunde	6220	Befund	Prozeduren	gering	vereinzelt	gering	Maskierung von Werten	mittel
	6221	Fremdbefund						
	6222	Laborbefund						
	6225	Röntgenbefund						
Therapien	6260	Therapie	Prozeduren	gering	vereinzelt	gering	Maskierung von Werten	mittel
	6265	Physikalische Therapie						
Weitere Prozeduren	6280	Überweisung Inhalt	Prozeduren	gering	vereinzelt	gering	Maskierung von Werten	mittel
	6285	AU Dauer						
	6286	AU wegen						
	6290	Krankenhauseinweisung, Krankenhaus						
	6291	Krankenhauseinweisung wegen						
Zeit- und Datums-daten	3649	Dauerdiagnosen ab Datum	Falldaten	hoch	mittel	hoch	Inter-contact interval	hoch
	3651	Dauermedikamente ab Datum						
	4101	Quartal der Abrechnung						
	5000	Leistungstag						
	6250	Tag der Speicherung von Behandlungsdaten						
Stamm- und Dauerdaten	3103	Geburtsdatum des Patienten	Demographie	hoch	teilweise hoch	hoch	Alter in Jahren; Altersdekade	hoch
	3110	Geschlecht des Patienten						
	3656	Allergie (Dauerbemerkung)						
Kenndaten der Praxis	0202	Praxistyp	Personen	teilweise hoch	teilweise hoch	teilweise hoch	regionale Merkmale	mittel
	0204	Arztgruppe verbal						
	0206	PLZ und Ort der Praxis						
	0225	Anzahl Ärzte						
Kostenträger	4105	Geschäftsstelle	Personen (teilweise)	gering	teilweise mittel	gering	Feld-kennungen nicht nutzen	gering
	4107	Abrechnungsart (Schein)						
	4121	Gebührenordnung						
Abrechnung	5001	GNR/GNR-Ident	Entgelte	gering	mittel	gering	Maskierung von Werten	mittel

* Medizininformatik-Initiative [8]
** Wichtigkeit für die Versorgungsforschung

Das Risikopotential sollte beispielhaft in einzelnen der 40 Variablen, punktuell auch in der Kombination von Variablen anhand von drei Kriterien untersucht werden. Während mathematische bzw. statistische Eigenschaften eines Datensatzes wie etwa die k-Anonymität in der Literatur schon eine umfangreiche Würdigung erfahren haben (vergl. [5,19]), ist die Frage nach dem Beitrag einzelner Variablen zur Re-Identifizierbarkeit eines Datensatzes bislang kaum beachtet worden. Das hat einerseits damit zu tun, dass eine solche Betrachtung immer nur für einen spezifischen Datensatz mit spezifischen Variablen erfolgen kann. Andererseits ist die Frage, welche Variable ggf. für eine Zuordnung genutzt werden kann und damit als Quasi-Identifikator zu betrachten ist, jenseits der üblicherweise in der Literatur genannten demographischen Variablen (z. B. Alter, Geschlecht und Wohnort [5]) alles andere als trivial. Insofern wird im Folgenden zumindest eine erste Annäherung an die systematische Bestimmung von Quasi-Identifikatoren mit Hilfe zweier Kriterien vorgenommen. Das dritte hier herangezogene Kriterium zur Beurteilung des Beitrags einer Variable

zur Re-Identifizierbarkeit geht hingegen auf die statistisch erwartbare Häufigkeitsverteilung der Ausprägungen und den damit im Zusammenhang stehenden notwendigen Vergröberungen eines Datensatzes ein, wie auch anderweitig in der Literatur beschrieben (z. B. [19]).

Kriterium 1: Expertenwahrnehmung

Die ausgewählten Variablen des BDT-Datensatzes sollten daraufhin geprüft werden, inwieweit Informationen aus Beobachtungen bzw. Wahrnehmungen im persönlichen Kontakt und auch im öffentlichen Raum mit Inhalten („Ausprägungen“) einzelner Variablen eindeutig in Beziehung gesetzt werden können. Da ein Angreifer, wie eingangs festgestellt, auch von „innen“ kommen kann, sind wir bei diesem Kriterium bewusst von einem Beobachter mit profunder medizinischer Ausbildung und ärztlicher oder pflegerischer Berufserfahrung ausgegangen, der nicht nur wie medizinische Laien direkt beobachtbare Informationen, wie z.B. besondere Auffälligkeiten im Erscheinungsbild („Phänotyp“) erkennt. Ein solcher „professioneller“ Beobachter erkennt zusätzlich möglicherweise pathognomonische Details während einer Begegnung mit einer Ziel-Person im öffentlichen Raum, also Krankheitszeichen oder Diagnosen, die bereits für sich oder in Verbindung mit zufälliger Kenntnis und eventueller Verknüpfung von Variablen und/oder Variableninhalten eine Personenidentifizierung ermöglichen. Ob unter diesen Voraussetzungen eine Person erkennbar ist, hat einer der Autoren (JH, im Sinne eines „profunden medizinischen Beobachters“) geprüft und sich in Zweifelsfällen mit den anderen Autoren abgestimmt.

Kriterium 2: Recherchierbares Zusatzwissen

Neben den von außen wahrnehmbaren Eigenschaften einer Person, die ggf. für eine Re-Identifizierung genutzt werden können, sollten auch Eigenschaften berücksichtigt werden, die nicht direkt wahrnehmbar sind, zu denen aber dennoch Zusatzwissen bestehen bzw. erlangt werden kann, wie z. B. zum Alter einer Person bzw. deren Geburtsdatum. Hier wäre exemplarisch an die in sozialen Netzwerken reichhaltig zur Verfügung gestellten persönlichen Detailinformationen zu denken, auch an nicht öffentliche Datenbanken (s.o.) hinsichtlich möglicherweise abgleichbarer Zusatzinformation.

Kriterium 3: Statistische Häufigkeit

Werden gemäß der ersten beiden Kriterien Variablen als Quasi-Identifikatoren kategorisiert, sollte dann überprüft werden, ob eine Datensammlung anhand dieser Werte eindeutig identifiziert werden kann. Das ist z. B. dann der Fall, wenn eine bestimmte Kombination von Werten aus den Quasi-Identifikatoren nur einmal in der ganzen Datensammlung vorkommt. Das Kriterium der Angreifbarkeit einer Datensammlung aus ihren statistischen Eigenschaften heraus ist immer wieder aktuell für jede konkrete Datensammlung einzeln zu bestimmen. Zudem bezieht sich dieses Kriterium nicht nur auf einzelne Variablen einer Datensammlung, sondern immer auch auf die Kombination zumindest aller als Quasi-Identifikatoren zu behandelnden Variablen. Auch wenn sich insofern allgemeine Aussagen verbieten, werden wir – im Sinne einer ersten Problemsicht – verfügbare Daten zu den Häufigkeitsverteilungen der Variablen (BDT-Felder) für „Diagnosen“ aus zwei Quellen heranziehen: die „30 häufigsten Diagnosen in Prozent der Behandlungsfälle in Arztpraxen in Nordrhein (Rang und Anteil)“ [21] sowie Diagnosen nach der International Classification of Diseases (ICD) aus eigenen Praxisdaten [22]), verkürzt („trunkiert“) auf die führenden 3 ICD-Stellen (die sog. „ICD Gruppenebene“). So kann zumindest für die Diagnosen, die nach den ersten beiden Kriterien als potentiell quasi-identifizierende Werte kategorisiert wurden, eine erste Abschätzung erfolgen, ob diese notwendigerweise nur vergröbert in anonyme Datensätze übernommen werden dürfen. Die Häufigkeit von Medikamenten entnehmen wir aus zwei Quellen: der Anatomisch-Therapeutisch-Chemische Einordnung von Wirkstoffen und Arzneimitteln, sog. ATC-Kodes, nach WIdO [23] sowie ebenfalls eigene Praxisdaten [22] als ATC-Gruppen (2. Ebene). Im Umkehrschluss verdienen natürlich seltene Erkrankungen mit ihren Diagnosen bei der weiteren Behandlung eine besondere Aufmerksamkeit.

Wie wir anhand der drei Kriterien die 40 Variablen und die 11 semantischen Gruppen bewertet haben, ist im Appendix ausführlich dargestellt. Auf das dritte Kriterium (statistische Häufigkeit) können wir dabei aus den genannten Gründen nur näherungsweise eingehen. Zusammengefasst findet sich diese Bewertung im folgenden Ergebniskapitel und in Form einer Übersicht in Tabelle 1, einschließlich – wenn auch noch probatorisch – möglicher Abhilfemaßnahmen.

Ergebnisse

Die in 11 semantischen Gruppen zusammengefassten 40 ausgewählten Variablen (BDT-Felder) und ihre -inhalte (Ausprägungen) wurden jeweils unter den drei Kriterien (1) Expertenwahrnehmung, (2) recherchierbares Zusatzwissen und (3) statistische Häufigkeit ausgewertet, um das Risiko der Re-Identifizierung ihrer Träger (betroffene Personen) zu untersuchen und zu bewerten (Tabelle 1).

Diagnosen und Beratungsanlässe haben besonders persönlichkeitsbestimmende und damit -identifizierende Eigenschaften und sind deshalb als besonders zu schützendes Privatgeheimnis zu betrachten. Unter den 30 häufigsten Diagnosebezeichnungen in der ambulanten Versorgung sind hier zwei als vom äußeren Erscheinungsbild (phänotypisch) re-identifizierend zu bezeichnen, nämlich „Adipositas“ (ICD-10 E66) sowie ggf. „Schädlicher Tabakkonsum“ (F17). Zudem fanden wir, dass etwa die Hälfte aller ICD-Kodes aus einer Hausarztpraxis in ihrer absoluten Häufigkeit den von uns gesetzten Grenzwert ($k = 5$) unterschreitet. Dies gilt häufig auch für einzelne Diagnosen, die zudem auch phänotypisch re-identifizieren und deshalb *ex ante* explizit benannt werden müssen, um vor sekundärer Nutzung etwa maskiert zu werden. Zudem sind gerade in Variablen dieser semantischen Gruppe Ausprägungen mit einfach und solche mit schwer zu identifizierbarem Potential auffällig gemischt, dass also einzelne Diagnosen gut von außen erkennbar sind, andere aber nicht. Schließlich gilt gruppenübergreifend, besonders aber bei „Diagnosen“, dass eine einzigartige Kombination von Werten zu einer möglichen Identifikation führen kann.

Angaben zur Medikation haben wenig Potential zu Re-Identifizierung von Personen. Die absolute Häufigkeit von Angaben zur Medikation, untersucht auf der 2. Gruppenebene von 3-stelligen ATC-Kodes, unterschreitet in etwa einem Achtel bis einem Viertel der Fälle den Grenzwert von $k = 5$ nicht unterscheidbaren Datensätzen. Problematisch sind auch einzelne Medikamente, die durch ihre Anwendung re-identifizierend sein können, z. B. Selbstinjektionen von Insulin oder Atemwegsmedikamente über Inhalatoren.

Laborergebnisse sind Daten mit einem geringen Re-Identifizierungspotential, das allerdings mit der regelmäßig großen Anzahl von Labordaten bei einem Individuum und ihrer Kombination steigt. Auch Prozeduren in der Hausarztpraxis sind generell mit geringem Risiko behaftet, ausgenommen Krankenhauseinweisungen wegen ihrer Seltenheit, verbunden mit einem üblicherweise schwerwiegenden Anlass.

Zeit- und Datumsinformationen sind dagegen besonders sensible Daten für die Re-Identifizierung einer Person. Während

ein einzelnes Kalenderdatum allein und für sich kaum als problematisch anzusehen ist, verhält es sich mit einer Reihe von Kalenderdaten (Zeitreihe) eines Individuums vollkommen anders. Auch die Kombination von Datumsfeldern mit anderen Feldern kann möglicherweise leicht auf eine Person bezogen werden. Geburtstage von Prominenten sind im Internet verfügbar, z.B. https://geboren.am/. Aber auch Privatpersonen, die nicht im Rampenlicht stehen, veröffentlichen ihr Geburtsdatum in sozialen Netzwerken. Durch Verknüpfung mit Ortsdaten (Geodaten), siehe Gruppe 9, steigt das Risiko der Re-Identifizierung stark. Andererseits sind Versorgungsdaten ohne eine gewisse Möglichkeit des Zeitbezugs i. d. R. wissenschaftlich nicht verwertbar und sinnlos. Aus Gründen der Datensparsamkeit erscheint die Variable mit der BDT-Feldkennung 4101 (Quartal der Abrechnung) entbehrlich.

Informationen zum Geschlecht und Alter eines Patienten sind in der Hausarztpraxis im Allgemeinen in der Regel nicht identifizierend, ausgenommen Informationen zum Alter bei sehr jungen und bei sehr alten Patienten, wenn diese praxisspezifisch selten sind.

Kenndaten der Praxis wie Typ (Einzel- oder Gemeinschaftspraxis) [32] oder beteiligte Fachgebietsbezeichnungen können generell als unproblematisch angesehen werden, auch im Hinblick auf den Datenschutz von Praxismitarbeitern. Eine Ausnahme bildet die Variable mit der BDT-Feldkennung 0206 (PLZ und Ort der Praxis). Auch in der HIPAA-Liste werden Angaben für geografische Einheiten, die weniger als 20.000 Einwohner umfassen, als identifizierend angesehen. Angaben zur Geschäftsstelle des Kostenträgers sind in gleicher Weise identifizierend, weil ebenfalls Ortsdaten.

Die Kategorisierung der Abrechnungsart wird im Allgemeinen ein Individuum nicht identifizieren. Kenntnis von der Nutzung der Gebührenordnung für Ärzte (GOÄ) erlaubt zwar die Zuordnung eines Individuums als gesetzlich oder privat krankenversichert für den einzelnen Leistungsfall, erhöht aber nur für die kleine Patientengruppe mit allein nach GOÄ abgerechneten Leistungen das Re-Identifizierungsrisiko. Leistungsziffern und abgerechnete Gebührenordnungsnummern im Einzelnen haben nur wenig Re-Identifizierungspotential. Allerdings können sie, wenn in großer Zahl beobachtet, auf einen besonderen Betreuungsbedarf oder schweres Kranksein hindeuten.

Variablen- und gruppenübergreifend sollten Informationen erwähnt werden, die die betroffene Person selbst öffentlich gemacht hat und über die hinaus durch eine Zuordnung keine weitere, vorher geheime Information gewonnen werden kann (siehe oben Methodik). Dabei kann es sich z. B. um die Diagnose eines Patienten handeln, wenn es zu diesem Patienten keine weiteren Diagnosen gibt, die durch eine Zuordnung offenbart werden könnten.

Weitere Ergebnisse im Detail finden sich im Appendix; in Kurzform in Tabelle 1. Erste Vorschläge zu Abhilfemaßnahmen finden sich ebenfalls in Tabelle 1.

Diskussion

Gesundheitsdaten sind immer als sensible Daten anzusehen. Erst aber die genauere Kenntnis relevanter einzelner Variablen und ihrer Inhalte, hier gewonnen über die BDT-Schnittstelle – unter Berücksichtigung der Beobachterperspektive, recherchierbaren Zusatzwissens sowie der Häufigkeit kritisch zu bewertender Angaben bzw. von deren Kombination – erlaubt eine erheblich verbesserte qualitative und sogar diskriminative Abschätzung des Risikos einer Re-Identifizierung von betroffenen Personen (Patienten, Praxismitarbeiter, Dritte). Im Sinne der Re-Identifizierung mit hohem Risiko behaftete Gruppen von Routinedaten sind Diagnosen, Zeit- und Datumsdaten, Stamm- und Dauerdaten des Patienten, sowie Kenndaten der Praxis. Unter diesen als besonders sensibel anzusehen sind die Variablen: Diagnosen als Dauerdiagnosen sowie Diagnosen als Langtext oder in häufiger Wiederholung; Ortsdaten (Geodaten im weiteren Sinne); Kalenderdaten von Kontakten und Behandlungen, insbesondere in ihrer Reihung; permanente oder kaum veränderbare Personenmerkmale, darunter auch das Geburtsdatum.

Stärken und Schwächen der Studie

In einer Feinanalyse untersuchten wir das Risikopotential bei der wissenschaftlichen Auswertung von hausärztlichen Routinedaten aus Arztpraxisinformationssystemen. Das schafft Rationalität in einer oft irrationalen Debatte, die auf der einen Seite den Datenschutz bei Big Data-Auswertungen ausgehöhlt, auf der anderen Seite Datenschutz als Verhinderung wissenschaftlicher Analysen sieht. Unsere Ergebnisse zeigen, dass ein sensibler Abwägungsprozess unverzichtbar ist zwischen persönlichen Schutzinteressen von Patienten (und anderen Betroffenen, etwa Professionellen oder „Dritten“) einerseits, und wissenschaftlicher Verpflichtung zur Analyse der Qualität medizinischer Prozesse und Ergebnisse sowie Versorgungs- und Gesundheitssystemforschung andererseits. Diese Abwägung erhält nunmehr eine erste Grundlage, um verlässlicher beurteilen zu können, welche Daten tatsächlich – unter Beachtung verschiedener Szenarien – im Sinne der Re-Identifizierung von Personen mit hohem Risiko behaftet sind, auf welche dieser Daten ggf. verzichtet werden kann und wie mögliche zusätzliche Maßnahmen zur Sicherstellung des Datenschutzes aussehen könnten.

Um Daten aus Krankenversorgung und Forschung besser nutzbar zu machen, hat das Bundesministerium für Bildung und Forschung das Förderkonzept Medizininformatik initiiert: „Ziel des Förderkonzepts Medizininformatik ist die Verbesserung von Forschungsmöglichkeiten und Patientenversorgung durch IT-Lösungen. Diese sollen den Austausch und die Nutzung von Daten aus Krankenversorgung, klinischer und biomedizinischer Forschung über die Grenzen von Institutionen und Standorten hinweg ermöglichen.“ [24] Die in dieser Untersuchung getroffene Einordnung der 40 BDT-Variablen in 11 semantische Gruppen entspricht weitgehend der Informationszuordnung zu entsprechenden Modulen aus dem Kerndatensatz der Medizininformatik-Initiative [20], siehe Tabelle 1. Das spricht nicht nur für eine sinnvolle Gruppenbildung, sondern erleichtert auch die Diskussion zwischen Akteuren in der primären und sekundären Gesundheitsversorgung. Die Liste der 18 HIPAA-Identifikatoren [25] wurde dort zunächst für einen anderen Zweck, nämlich für die Übertragbarkeit und Verlässlichkeit von Gesundheitsdaten im US-amerikanischen Krankenversicherungskontext, erstellt und ist deshalb für die Fragestellung hier wenig erhellend, mit Ausnahme der Spezifizierungen zu Kalender- und zu Geodaten.

Der hier angestellte erste Versuch ist sicher noch unvollständig und diskussionsbedürftig. Die Folgen, die durch die Verknüpfung von Informationen aus den verschiedenen Gruppen entstehen, wurden hier nur punktuell und sicherlich nicht erschöpfend diskutiert. Auch gilt diese Risikoanalyse allein für die ausgewählten 40 Variablen und ihre möglichen Inhalte (Ausprägungen), wenn auch diese bereits bewusst aus Datensparsamkeit und unter Risikoaspekten ausgewählt wurden.

Ein weiteres Problem ist, dass je nach verwendeter Arztpraxissoftware die Daten einzelner Felder des BDT-Datensatzes auch Freitext-Angaben enthalten können. Das kann sogar dann der Fall sein, wenn eigentlich nur kategoriale Werte gemäß Spezifikation zu erwarten wären. Insofern können zusätzliche Vorprüfungen eines BDT-Datensatzes erforderlich werden. In zukünftigen Schnittstellenfestlegungen, die nach § 291d SGB V und mittels zeitgemäßer IT-Ressourcen, etwa FHIR der HL7-Gruppe, definiert werden, sind diese „handwerklichen“ Probleme wohl nicht mehr zu erwarten.

Identifizierungspotential verschiedener Variablen in Routinedaten

Anders als es globale Analysen zum Risikopotential von Routinedaten bzw. Big Data tun [26–28], zeigt unsere Analyse die Risiken einzelner Variablen und ihrer Inhalte im Detail auf. Die exemplarische Bewertung aus verschiedenen Blickwinkeln lässt Risiken in ihrer Vielfalt und konkreten Gestalt leichter erkennen.

Zum Beispiel ist in der Gruppe der Diagnosen „Adipositas“ eine für Laien leicht erkennbare Diagnose, eine Gonarthrose dagegen nur schwer. Letztere aber wäre wiederum für Medizin-Professionelle vergleichsweise leicht erkennbar. Berücksichtigt man dann noch, dass letztere Diagnose in vergleichsweise vielen Praxen seltener als fünfmal beobachtet wird, deutet sich hier ein Re-identifizierungsrisikopotenzial an (s. Appendix; Beispiel „Diagnosen“).

Wenn innerhalb einer Variablen Ausprägungen mit hoher und solche mit niedriger Risikobewertung auffällig gemischt sind, also z. B. einzelne Diagnosen gut von außen erkennbar sind, andere aber nicht, so empfiehlt sich möglicherweise eine Aufteilung dieser Ausprägungen auf zwei Variablen bzw. die damit einhergehende Erzeugung einer weiteren Variablen (einer sog. sekundären Variable), in welche Ausprägungen übertragen werden, die gemäß des hier beschriebenen Kriteriums gut anhand von Zusatzwissen, z. B. anhand der Wahrnehmung eines medizinischen Experten, zugeordnet werden können. Die neue Variable kann dann allerdings nur ausreichend vergröbert für die weitere Nutzung übernommen werden. In der originalen Variablen (der sog. primären Variable) werden diese Ausprägungen gelöscht, so dass diese Variable dann ggf. nicht mehr als Quasi-Identifikator betrachtet werden muss und entsprechend unvergröbert in einen anonymen Datensatz übernommen werden kann.

Ein weiterer Weg bestünde beispielsweise in der Vergröberung der Werte durch ein Zusammenfassen von Werten innerhalb eines Wertebereichs zu Gruppen oder auch in der Unterdrückung einzelner Werte. Kostenfreie Open-Source-Software zur Unterstützung dieser Prozessschritte steht zur Verfügung und hat sich in den letzten Jahren im Umfeld medizinischer Forschung bereits bewährt [18,19]. Im Regelfall unterstützen passende Softwaresysteme auch noch weitere Prozesse, um die Daten gegen unerlaubte Zuordnungen bzw. Offenlegung von Daten abzusichern.

Bei Medikamenten wäre zum Beispiel zu beachten, dass viele Verordnungen an sich nicht identifizierend sind, das Wissen um eine bestimmte Anwendung (Selbstmedikation zum Inhalieren oder zum subkutanen Injizieren) das Identifizierungsrisiko jedoch erhöht (s. Appendix; Beispiel „Medikamente“).

Implikationen für die weitere Arbeit mit Routinedaten

Es erscheint erforderlich, generell Diagnosen nicht als Langtext zu extrahieren und zu verarbeiten, sondern kodiert lediglich als „ICD-Dreisteller“ auf der sog. ICD Gruppenebene. Insbesondere die zwei Ausprägungen „Adipositas“ und „Schädlicher Tabakkonsum“ müssen in geeigneter Weise, beispielsweise durch alleinige Nennung des ICD-Kapitels (1. Stelle des ICD-Kodes), unkenntlich gemacht werden. ICD-Kodes aus einer Hausarztpraxis, die in ihrer absoluten Häufigkeit einen *a priori* konsentierten Grenzwert unterschreiten, müssen in gleicher Weise unkenntlich gemacht werden. Dies gilt auch für einzelne, phänotypisch re-identifizierende Diagnosen wie zum Beispiel das „Down-Syndrom“ (Trisomie 21).

Medikamente sollten nicht als Langtext kodiert und lediglich als „ATC-Dreisteller“ (2. Gruppenebene) extrahiert und verarbeitet bzw. ausgewertet werden. Unterschreiten Medikamentennennungen in ihrer absoluten Häufigkeit einen zuvor festgelegten kritischen Grenzwert, sollten sie in geeigneter Weise, beispielsweise durch alleinige Nennung des ATC-Kapitels (1. Stelle des ATC-Kodes), unkenntlich gemacht werden, ebenso Selbstinjektionen oder vom Patienten anzuwendende inhalative Medikamente.

Möglicherweise ist anstelle von Kalenderdaten das Generieren einer abgeleiteten Variablen sinnvoll. Als Beispiel sei die Berechnung des Interkontakt-Intervalls [22] (*inter-contact interval*, ICI) als Differenz in Tagen von zwei aufeinanderfolgenden Kontakten eines Patienten mit der Praxis oder dem Hausarzt genannt: während der ursprüngliche kalendarische Bezug damit nicht mehr besteht, bleibt ein wesentlicher Teil der enthaltenen Information verfügbar, der zudem größer und aussagekräftiger ist als beispielsweise Information aus Kontaktfrequenzen mit willkürlich gewähltem Bezugszeitraum.

Zeitliche Angaben könnten ggf. auch dahingehend ‚verfälscht‘ werden, dass sie um einen zufälligen Wert verschoben werden. Das kann den positiven Effekt haben, dass zeitliche Abläufe weiterhin in hoher Auflösung nachvollzogen werden können. Durch die zufällige und für einen Angreifer nicht nachvollziehbare Verschiebung wären diese zeitlichen Angaben dann nicht mehr Quasi-Identifikatoren.

Bei sehr jungen und sehr alten Patienten sollten Altersangaben lediglich in Altersdekaden erfolgen.

In Kenndaten der Praxis, insbesondere zum Ort, sollte auf den Langtext der Ortsangabe verzichtet werden. Wenn die Trunkierung auf die ersten drei (Kreisebene) oder auch nur auf die erste Stelle der PLZ nicht ausreichend de-identifiziert, sollte die PLZ algorithmisch durch eine globale Zuordnung des Praxisortes ersetzt werden, beispielsweise durch eine aus der PLZ abgeleitete Klassifizierung nach Anzahl der Einwohner der von der Praxis versorgten Gemeinde oder nach siedlungsstrukturellen Gesichtspunkten (z.B. Agglomerations-, städtische oder ländliche Region). Andere Ortsangaben, z. B. des Kostenträgers, wären für wissenschaftliche Untersuchungen leicht entbehrlich und sollten damit aus Gründen der Datensparsamkeit nicht genutzt werden.

Zumindest angedeutet sei hier das Problem, dass im Kontext der anonymen Verarbeitung von Daten die Betroffenen, anders als bei der Einholung einer informierten Einwilligung, nicht darauf aufmerksam gemacht werden, dass die eigene Veröffentlichung von Daten Einfluss auf das Re-Identifizierungsrisiko haben kann. Insofern werden – im Sinne des Kriteriums 2 – letztlich doch viele Variablen statistisch vergröbert werden müssen, um eindeutige Zuordnungen auszuschließen.

Die hier vorgestellten Methoden haben gemein, dass die Ausgangsdaten in allen Fällen verändert werden, sei es durch Löschung (Unterdrückung), Verallgemeinerung (Generalisierung) oder Verfälschung (Mikroaggregation). Die Grenze zwischen personenbezogenen und anonymen Daten ist fließend und erfordert einen Abwägungsprozess zwischen dem Datenschutzrisiko und einem zumutbaren Informationsverlust. Bei diesem Abwägungsprozess kann auf Softwareunterstützung (bspw. μ-ARGUS [29], sdcMicro [30] oder ARX [31]) zur Analyse und Reduktion des Re-Identifikationsrisikos zurückgegriffen werden. Die Bewertung eines Informationsverlustes als hinnehmbar muss letztlich jedoch immer vor dem Hintergrund der zu untersuchenden Fragestellung erfolgen.

Bei der Abschätzung des Re-Identifizierungspotentials einzelner Variablen wird methodisches und medizinisches Know-how vereint benötigt. Insofern ist die softwaregestützte und effektive Anonymisierung solcher Routinedaten derzeit noch eine Aufgabe für wenige und entsprechend geschulte Spezialisten. Unter Hinzuziehung ihres Fachwissens muss die softwaregestützte Anonymisierung dann zudem in der Praxis selbst erfolgen.

Die Art der digitalisierten Erfassung und Strukturierung medizinischer Behandlungsdokumentation steht aktuell vor großen Umbrüchen. Genannt sei als Stichwort die elektronische Patientenakte. Wenn diese tatsächlich in einer Form eingeführt wird, die einen strukturierten und einrichtungsübergreifenden

Zugriff auf längere Behandlungsverläufe unterstützt, dann werden die hier erfassten und bereitgestellten Daten auch neue und vielversprechende Möglichkeiten für die Nachnutzung in der Forschung bieten, gleichzeitig jedoch besonders schwer zu anonymisieren sein, wenn der wesentliche Informationsgehalt bewahrt werden soll. Insofern können die hier gezeigten Lösungsansätze durchaus auch auf die Nachnutzung von Daten aus elektronischen Patientenakten übertragen werden.

Schlussfolgerung

Das Risiko einer Re-Identifizierung von Personen anhand von quasi-identifizierenden medizinischen Daten existiert und muss bei der wissenschaftlichen Analyse von Routinedaten minimiert werden. Maßnahmen wie Datenentfernung, Wertmaskierung oder Kodierung dürften die Re-Identifikation erheblich erschweren, insbesondere für Personen mit großem Zusatzwissen (Forscher, Experten). Diese Schutzmaßnahmen können zudem zusätzlich wirksam werden bei nichtautorisierter Datenentschlüsselung oder der Verknüpfung mit anderen Datenbanken. Jedoch kann bei allen Anonymisierungsverfahren ein Restrisiko hinsichtlich der Re-Identifizierbarkeit kaum jemals vollständig ausgeschlossen werden. Das muss die Verwendung als anonyme Daten für die Forschung jedoch nicht verhindern, da auch die DSGVO in „Erwägungsgrund 26“ für die Anonymität von Daten nicht den vollständigen Ausschluss eines Re-Identifizierungsrisikos verlangt. Dies aber sollte viel offener und transparenter diskutiert werden. Ohne die oftmals pauschalen Vorbehalte gegenüber einer Digitalisierung im Gesundheitswesen zu teilen, sollten die eingangs genannten und berechtigten Befürchtungen gegenüber einem mangelnden Datenschutz berücksichtigt und in eine empirisch fundierte Nutzen-Risiko-Bewertung eingebunden werden.

Autorenbeteiligung

JH und JD, gemeinsam mit allen Autoren, entwickelten die Grundidee „11 semantische Gruppen”, verfeinerten sie gemeinsam mit WH und bezogen sie auf 40 BDT-Feldkennungen sowie die entsprechenden Module des Kerndatensatzes der Medizininformatik-Initiative. Alle Autoren in fortwährendem Austausch schufen erste Version und endgültige Ausformung des Manuskripts, wobei JP und FS besonders zu IT- und JD zu legalen Aspekten beitrugen. Alle Autoren lasen und billigten die Endversion. Das Manuskript ist ein Nebenprodukt des RADAR Projekts und seiner Finanzierung.

Finanzierung

Das RADAR Projekt wird gefördert von der Deutsche Forschungsgemeinschaft (DFG), Fördernummern HU 1587/2-1, HO 1937/7-1, RI 1000/7-1, YA 191/8-1, KR 1093/10-1 vom 29.04.2016

Danksagung

Die Autoren danken den übrigen Partnern im RADAR Projekt: Otto Rienhoff, Philipp Wieder, Ramin Yahyapour (alle Göttingen), Thomas Bahls, Arne Blumentritt, Wolfgang Hoffmann (alle Greifswald), Valérie Kempter, Sebastian Claudius Semler, Jonas Steinmann (alle Berlin).

Interessenkonflikt

Die Autoren erklären, dass kein Interessenkonflikt besteht.

Appendix A. Zusätzliche Daten

Zusätzliche Daten verbunden mit diesem Artikel finden sich in der Online-Version unter: doi:10.1016/j.zefq.2020.01.002.

References

[1] Datenschutzgrundverordnung (DSGVO). https://dsgvo-gesetz.de/.(last accessed on 2 January 2018).
[2] Holzer K, Gall W. Utilizing IHE-based Electronic Health Record systems for secondary use. Methods Inf Med 2011;50(4):319–25.
[3] OECD: Glossary of Statistical Terms: Quasi-identifier. http://stats.oecd.org/glossary/detail.asp?ID=6961.(last accessed on 18 April 2018).
[4] Petrlic R, Sorge C:. Datenschutz.: Einführung in technischen Datenschutz Datenschutzrecht und angewandte Kryptographie. Wiesbaden: Springer Fachmedien; 2017.
[5] Sweeney L. k-anonymity: a model for protecting privacy Int. J. Unc. Fuzz. Knowl. Based Syst 2002;10(05):557–70.
[6] Hauswaldt J, Himmel W, Kempter V, Hummers E. Hindernisse bei der sekundären Nutzung hausärztlicher Routinedaten. Gesundheitswesen 2018.
[7] Monika Wójtowicz MC. Anonymisierung nach der DSGVO. Privacy in Germany 2017:186–92. PinG 05.17.
[8] David, Seiler. Überblick zur Datenverarbeitung im medizinischen Bereich unter der DSGVO: Unter Berücksichtigung der Novellierung des § 203 StGB. Privacy in Germany 2018. PinG 01.18.
[9] McGinn CA, Grenier S, Duplantie J, et al. Comparison of user groups' perspectives of barriers and facilitators to implementing electronic health records: A systematic review. BMC Med 2011;9:46.
[10] Meißner A. Das gläserne Behandlungszimmer. Süddeutsche Zeitung 2019, 25 February 2019 https://www.sueddeutsche.de/politik/aussenansicht-das-glaeserne-behandlungszimmer-1.4344293.
[11] Vuokko R, Mäkelä-Bengs P, Hyppönen H, Lindqvist M, Doupi P. Impacts of structuring the electronic health record: Results of a systematic literature review from the perspective of secondary use of patient data. Int J Med Inform 2017;97:293–303.
[12] Lu C, Wu Z, Liu M, Chen W, Guo J. A patient privacy protection scheme for medical information system. J Med Syst 2013;37(6):9982.
[13] Fernández-Alemán JL, Señor IC, Lozoya PÁO, Toval A. Security and privacy in electronic health records: A systematic literature review. J Biomed Inform 2013;46(3):541–62.
[14] Deutsche Gesellschaft für Allgemeinmedizin und Familienmedizin: Positionspapier der Deutschen Gesellschaft für Allgemeinmedizin und Familienmedizin (DEGAM) zur obligaten Einrichtung und Unterhaltung einer Wissenschaftlichen Datentransferschnittstelle in Arztpraxisinformationssystemen. http://www.allgemeinmedizin.med.uni-goettingen.de/de/media/DEGAM_Positionspapier_Praxissoftware.pdf.(last accessed on 10 October 2017).
[15] Langarizadeh M, Orooji A, Sheikhtaheri A. Effectiveness of Anonymization Methods in Preserving Patients' Privacy: A Systematic Literature Review. Stud Health Technol Inform 2018;248:80–7.
[16] Lee H, Kim S, Kim JW, Chung YD. Utility-preserving anonymization for health data publishing. BMC Med Inform Decis Mak 2017;17(1):104.
[17] Hauswaldt J, Himmel W. RADAR - Anonymisierte Routinedaten aus der ambulanten Versorgung für die Versorgungsforschung. http://www.allgemeinmedizin.med.uni-goettingen.de/de/content/forschung/510_591.html.(last accessed on 15 September 2017).
[18] Dolor RJ, Campbell-Voytal K, Daly J, et al. Practice-based Research Network Research Good Practices (PRGPs): Summary of Recommendations. Clin Transl Sci 2015;8(6):638–46.
[19] Prasser F, Kohlmayer F, Lautenschläger R, Kuhn KA. ARX–A Comprehensive Tool for Anonymizing Biomedical Data. AMIA Annu Symp Proc 2014;2014:984–93.
[20] Ganslandt T, Boeker M, Löbe M, Prasser F, Schepers J, Semler SC, et al. Der Kerndatensatz der Medizininformatik-Initiative.: Ein Schritt zur Sekundärnutzung von Versorgungsdaten auf nationaler Ebene. Forum der Medizin-Dokumentation und Medizin-Informatik (mdi) 2018;20(1):17–21.
[21] Zi-ADT-Panel, Zentralinstitut für die kassenärztliche Versorgung in der Bundesrepublik Deutschland: Häufigste Diagnosen in Prozent der Behandlungsfälle in Arztpraxen in Nordrhein. http://www.gbe-bund.de/oowa921-install/servlet/oowa/aw92/dboowasys921.xwdevkit/xwd_init?gbe.isgbetol/xs_start_neu/&p_aid=3&p_aid=3179591&nummer=638&p_sprache=D&p_indsp=-&p_aid=49995736 (last accessed on 22 April 2018).
[22] Hauswaldt J, Himmel W, Hummers-Pradier E. The inter-contact interval: a new measure to define frequent attenders in primary care. BMC Fam Pract 2013;14:162.
[23] GKV-Arzneimittelindex, Wissenschaftliches Institut der AOK: Therapeutische Arzneimittel, die zu Lasten der gesetzlichen Krankenversicherung verordnet wurden.(last accessed on 22 April 2018).
[24] Medizininformatik-Initiative - Ziele. http://www.medizininformatik-initiative.de/de/ueber-die-inititative/ziele.(last accessed on 21 July 2018).
[25] Czajka J, Schneider C, Sukasih A, Collins K. Minimizing Disclosure Risk in HHS Open Data Initiatives: Final Report. September 2014;29, https://aspe.hhs.gov/system/files/pdf/77196/rpt_Disclosure.pdf (last accessed on 7 June 2018).

[26] Weichert T. Big Data im Gesundheitsbereich.: Gutachten erstellt im Projekt Assessing Big Data (ABIDA). http://www.abida.de/de/blog-item/gutachten-big-data-im-gesundheitsbereich.
[27] Mayer-Schönberger V. Big Data - Eine Revolution, die unser Leben verändern wird. Bundesgesundheitsblatt Gesundheitsforschung Gesundheitsschutz 2015;58(8):788–93.
[28] Drepper J. Übersicht – Big-Data-Anwendungen und der Datenschutz. Tumor-Diagn u Ther 2016;37(06):316–9.
[29] Franconi L, Polettini S. Individual Risk Estimation in μ-Argus: A Review. In: Kanade T, Kittler J, Kleinberg JM, Mattern F, Mitchell JC, Naor M, et al, editors. Privacy in Statistical Databases. Berlin. Heidelberg: Springer Berlin Heidelberg; 2004. p. 262–72.
[30] Templ M. Statistical Disclosure Control for Microdata Using the R-Package sdcMicro. Journal Transactions on Data Privacy 2008;1(2):67–85.
[31] Kohlmayer F, Prasser F, Kuhn KA. The cost of quality: Implementing generalization and suppression for anonymizing biomedical data with minimal information loss. J Biomed Inform 2015;58:37–48.
[32] Osterloh F. Medizinische Versorgungszentren: Eine Alternative, keine Konkurrenz. Dtsch Arztebl 2017;114(42):A1901–3.