Zur De-Identifizierung von Feldinhalten in hausärztlichen Routinedaten

Hauswaldt J[1], Groh R[2], Kaulke K[3], Schlegelmilch F[1], Zarei A[2], Hummers E[1].
[1] Institut für Allgemeinmedizin, Universitätsmedizin Göttingen, Göttingen
[2] Gesellschaft für wissenschaftliche Datenverarbeitung Göttingen, Göttingen
[3] Technologie- und Methodenplattform für die vernetzte medizinische Forschung (TMF) e.V., Berlin

Hintergrund Sekundäre Nutzung von hausärztlichen Routinedaten ist technisch und organisatorisch rechtskonform machbar [1]. Potentiell identifizierende Feldinhalte (PIF), insbesondere Freitexteinträge, behindern die „faktische Anonymisierung“ eines wissenschaftlich genutzten Sekundärdatensatzes (SDS).
Ziel Schrittweises und systematisches Erkennen von PIF in einem exemplarischen SDS aus strukturierten Routinedaten einer hausärztlichen Praxis, extrahiert mittels der Behandlungsdatentransfer (BDT)-Schnittstelle. Ergebnisbewertung im Sinne einer Datenschutz-Folgenabschätzung (DSFA).
Methodische/s Kernproblem/e Untersucht wird auf den Ebenen (a) der Feldkennungen (Variablen, Attribute), (b) ihrer Kombinationen, (c) ihrer Feldinhalte (Ausprägungen, Werte) und (d) des gesamten Datensatzes.
Instrumente sind für (a) und (b) Feldtyp, relative Häufigkeiten, Kategorien, und hausärztliche Expertise, (c) TextCrawler [2], (d) ARX [3]. Bewertung als Abschätzen des Zusammentreffens von Schwere eines möglichen Schadens mit seiner Eintrittswahrscheinlichkeit.
Lösungsansätze Ein SDS aus einer hausärztlichen Praxis, 1993 bis 2017, von 14.285 Patienten, vorliegend als .csv-Datei mit 5.918.321 Datenzeilen (224 MB) und drei Variablen (Reihenfolge, Feldkennung, Feldinhalt), wurde untersucht.
PIF wurden v.a. in den Feldern „Dauerbemerkungen“ und „Befunde“ erkannt und als „Namen“, „Ortsnamen“, „Telefonnummern“, „Funktions-“ und „Berufsbezeichnungen“ kategorisiert. „Sterbedatum“ wird als hoher Schaden mit mittlerer Eintrittswahrscheinlichkeit angesehen – Abhilfe: Umwandlung in Sterbejahr. Die Kombination von BDT-typischer temporaler Reihung, pseudonymisierter Patientenzuordnung und einzelnen Feldinhalten erhöht das Re-Identifizierungsrisiko im SDS als Ganzem.
Diskussion Untersuchungen zu PIF müssen an einem konkreten, abgeschlossen vorliegenden SDS durchgeführt werden. Sie setzen fach- und sachspezifische Kenntnisse über Entstehung und Rahmenbedingungen der Rohdaten in Hausarztpraxen sowie Metainformationen über die Primärdaten voraus.
Schlussfolgerungen Mit vertretbarem Aufwand können PIF in einem abgeschlossenen SDS immer nur unvollständig erkannt werden. Erkennen und Bewerten von PIF sind Voraussetzung für de-identifizierende Maßnahmen.

Literatur

[1] Hauswaldt J, Bahls T, Blumentritt A, Demmer I, Drepper J, Groh R, Heinemann S, Hoffmann W, Kempter V, Pung J et al. (2021): Sekundäre Nutzung von hausärztlichen Routinedaten ist machbar – Bericht vom RADAR Projekt. Gesundheitswesen 83, S130-S138
[2] TextCrawler Free 3.0, DigitalVolcano Software. https://www.digitalvolcano.co.uk/tcdownloads.html letzter Zugriff 23.11.2021
[3] ARX Data Anonymization Tool. https://arx.deidentifier.org/ letzter Zugriff 23.11.2021

Korrespondenzadresse des Erstautors
Institut für Allgemeinmedizin, UMG, Humboldtallee 38, 37073 Göttingen
johannes.hauswaldt@med.uni-goettingen.de